NEWTON'S NOTEBOOK

The Haverford School STEM Journal

Issue VIII
Spring 2024

Letters from the Editors

Dear readers,

Welcome to the XVIII issue reprint of the *Notebook*.

With topics ranging from a contemplation of the universe's end to the Fast Inverse Square Root Algorithm, this edition of the Notebook is perhaps our most diverse yet. We are so excited for you all to dive into the diverse works of writing created by Haverford students and curated here.

Milan Varma '25

Fellas,

Thank you to all the contributors and faculty members that have made this issue of *Newton's Notebook* possible. I am extremely proud of your interest in STEM topics and your ability to conduct research and share it with the community. I urge you to continue to look into these interests throughout the summer months.

To whoever picked up this edition, I hope you enjoy this edition and learn something new!

Kevin Li '25

Hey Everyone,

It's been a great year for Newton's Notebook so far, and I only have three things to say! For those who contributed, great job! Your interest in STEM and your willingness to learn is the fuel we run on. For those who have just picked up a copy, sit back and enjoy! There's a lot of interesting content in here, that I'm sure you'll find fascinating. Lastly, to all of the teachers who have made this process possible, thank you so much!

Nicholas Lu '25

Table of Contents

Mission Statement

Newton's Notebook: The Haverford School STEM Journal is designed to enhance the interests, talents, and achievements of individuals in mathematics and science to promote the work of those passionate about the disciplines. The following articles were written by members of the Haverford community and edited by the *Notebook* staff. We hope these articles inspire readers to further discover the universally beautiful realm of STEM exploration.

Staff

Advisor

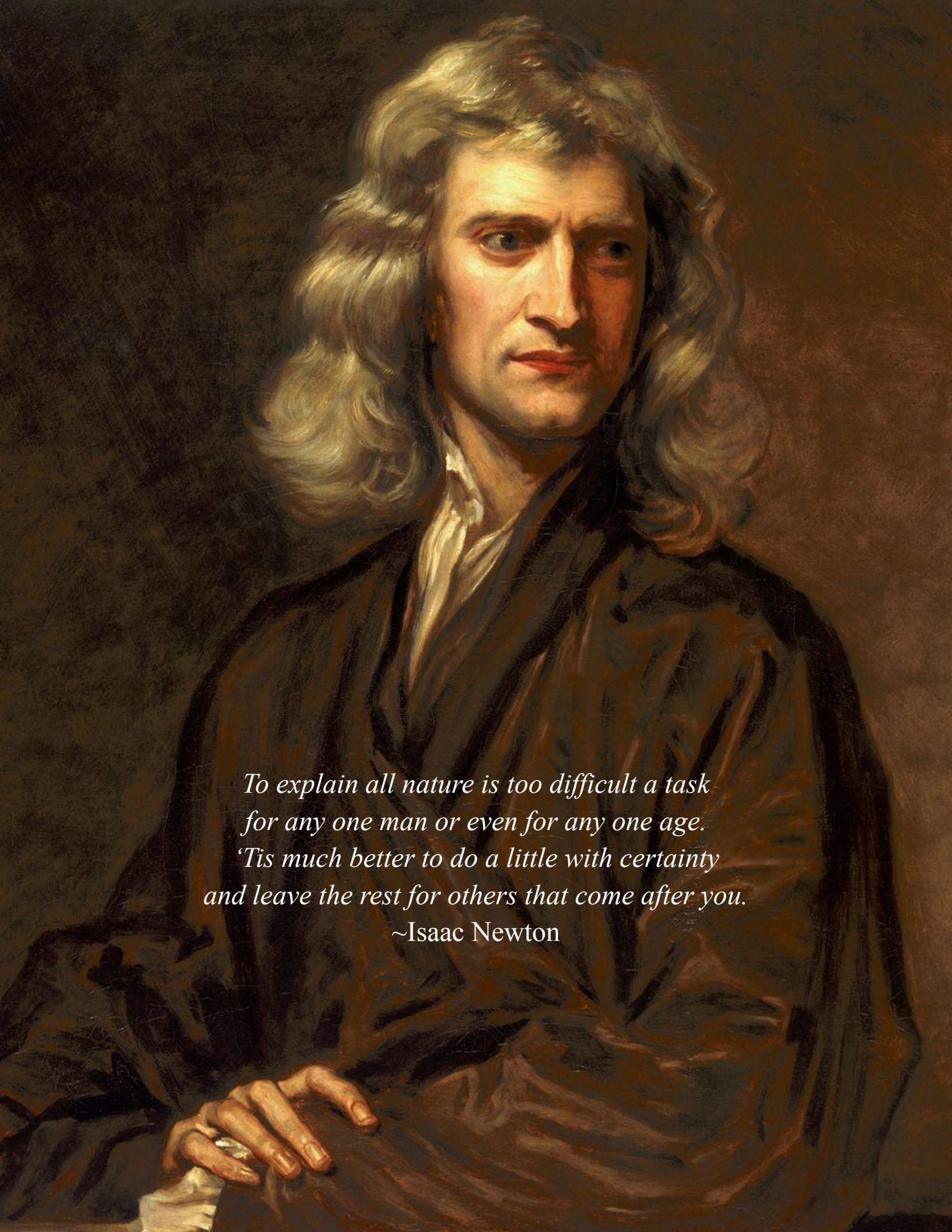
Acknowledgments

The *Notebook* staff thanks the following:
Dr. Gottlieb
for his guidance and mentorship throughout the school year;

The Haverford School Math Department faculty members
for their encouragement and support;

Lulu Publishing
for its press resources;

All of our contributors and editors
for their hard work and pursuit of mathematical and scientific excellence.


To explain all nature is too difficult a task
for any one man or even for any one age.
'Tis much better to do a little with certainty
and leave the rest for others that come after you.
~Isaac Newton

Black-Scholes Equation Derivation

Nicholas Lu '25

Applied Mathematics

Let's consider a financial market with two investment options:

- A risky asset, typically a stock, whose price at time t is denoted by S(t).

- A risk-free bond with a constant risk-free interest rate r.

The goal is to price a European call option on the stock, which gives the holder the right (but not the obligation) to buy the stock at a predetermined price E (the strike price) at a specified future time T (the expiration date). The price of the option at time t is denoted by C(S(t), t). To model the stock price dynamics, we assume it follows a geometric Brownian motion with a constant drift μ and volatility σ. This can be expressed as:

$$dS(t) = \mu S(t)dt + \sigma S(t)dW(t)$$

where dW(t) represents the increment of a Wiener process (or Brownian motion). To work with the option pricing function C(S(t), t), we apply Ito's lemma, which is a fundamental tool for dealing with stochastic processes. For a function f(x, t) where x follows a stochastic process, Ito's lemma gives the differential:

$$df(x,t) = \left(\frac{\partial f}{\partial t} + \mu\frac{\partial f}{\partial x} + \frac{1}{2}\sigma^2 x^2 \frac{\partial^2 f}{\partial x^2}\right)dt + \sigma x \frac{\partial f}{\partial x} dW(t)$$

Applying Ito's lemma to C(S(t), t), we get:

$$dC = (\frac{\partial C}{\partial t} + \mu S\frac{\partial C}{\partial S} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 C}{\partial S^2})dt + \sigma S\frac{\partial C}{\partial S}dW(t)$$

Consider a portfolio Π composed of holding one option and short selling Δ units of the stock. The value of this portfolio is:

$$\Pi = C - \Delta S$$

The change in the portfolio value over a small time interval dt is given by:

$$d\Pi = dC - \Delta dS$$

Substituting dC from Ito's lemma and dS from the stock's geometric Brownian motion, we have:

$$d\Pi = (\frac{\partial C}{\partial t} + \mu S\frac{\partial C}{\partial S} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 C}{\partial S^2})dt + \sigma S\frac{\partial C}{\partial S}dW(t) - \Delta(\mu Sdt + \sigma SdW(t))$$

By choosing $\Delta = \frac{\partial C}{\partial S}$, we make dΠ risk-free as it eliminates the dW(t) term, leading to:

$$d\Pi = (\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 C}{\partial S^2})dt$$

Since dΠ is risk-free, it must grow at the risk-free rate r to avoid arbitrage opportunities, implying:

$$d\Pi = r\Pi dt = r(C - \Delta S)dt$$

Substituting $\Delta = \frac{\partial C}{\partial S}$ and equating it with the expression for dΠ derived above, we get:

$$r(C - S\frac{\partial C}{\partial S})dt = (\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 C}{\partial S^2})dt$$

Simplifying the above equation and rearranging terms, we obtain the Black-Scholes PDE:

$$\frac{\partial C}{\partial t} + rS\frac{\partial C}{\partial S} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 C}{\partial S^2} - rC = 0$$

Which simplifies to

$$C = N(d_1)S_t - N(d_2)Ke^{-rt}$$

Beyond Numbers: Unraveling Credit Card Attrition with Socioeconomic Insights

Nicholas Lu '25
Applied Mathematics

1. Introduction

Imagine a world in which financial institutions are able to anticipate the behavior of their credit card users. In today's era of data-driven insights, such a feat is not merely a pipe dream - it is an actionable reality that has the potential to redefine how credit card issuers engage with their customers.

Credit card attrition, defined by advisory firm R.K. Hammer as the percentage of issuers' total number of accounts who voluntarily or involuntarily have their card accounts closed, prior to netting those closures from the new card accounts added during the same reporting period[1], presents a formidable challenge for financial institutions. Though transactions are often considered the most powerful data in financial services, attrition also wields significant influence in the ever-evolving landscape of financial services.

As the digital age empowers organizations with vast amounts of data, the ability to harness this information has never been more crucial. The exponential growth in data generation is a hallmark of our society - approximately 328.77 million terabytes of data are created each day, and it's estimated that 90% of the world's data was generated in the last two years alone.[2] This data explosion has revolutionized industries, offering

[1] "Card Member Attrition Rates Softening." *CUInsight*, Card Knowledge Factory, www.cuinsight.com/press-release/card-member-attrition-rates-softening/#:~:text=Gross%20attrition%20is%20defined%20as,during%20the%20same%20reporting%20period. Accessed 19 Aug. 2023.

[2] Duarte, Fabio. "Amount of Data Created Daily (2023)." *Exploding Topics*, Exploding Topics, 3 Apr. 2023, explodingtopics.com/blog/data-

unprecedented opportunities for informed decision-making and innovation across sectors.

Against this backdrop, predictive analytics stands as a beacon of possibility. It holds the promise of early intervention, enabling credit card issuers to tailor their strategies for customer retention and satisfaction. By identifying customers at risk of attrition, institutions can offer tailored incentives, personalized services, and improved customer experiences.

However, while the concept of credit card attrition is well-established, the exploration of predictive models incorporating socioeconomic dimensions remains relatively uncharted. Existing studies often focus on financial aspects, leaving a gap in our understanding of how socioeconomic factors interact with attrition tendencies. This research paper aims to bridge the gap by constructing predictive models that integrate both traditional financial variables and essential socioeconomic indicators to shed light on this intricate relationship, enriching the accuracy and depth of such predictions.

This paper first presents a literature review to contextualize the gaps in current research, then delves into an overview of the dataset, revealing initial insights into the interplay of socioeconomic variables and attrition patterns. Then, the robust analytical methodology encompassing data preprocessing, decision trees and clustering, regression analysis, and predictive modeling is presented. Finally, the results and conclusions section unveils the implication of these research findings.

Examining the nuances of credit card attrition prediction through the lens of socioeconomic insight contributes valuable knowledge to the intersection of financial behavior and data analytics, equipping credit card issuers with the tools they need to navigate the complex landscape of customer retention.

generated-per-day.

2. Literature Review

Credit card providers face the persistent challenge of attrition due to various influences, from competitive incentives to customer service concerns. While the subject has been the focus of many studies, there is noticeable lack of in-depth research targeting the precise prediction of credit card attrition, particularly those employing advanced methodologies.

Some analyses have already gauged the rate and reasons of credit card attrition in general[3] and examined the overarching landscape and costs associated with it. Others have delved deeper, studying specific methods of maximizing customer retention, analyzing factors such as interest rates and GDP trends[4]. Similarly, research has been conducted on the impact of promotional offers, such as refreshment samples, offering an additional perspective on the dynamics of credit card attrition[5].

A main challenge in this field of research is the variety in benefits offered - each card offers unique benefits and as such, attrition patterns vary widely. Certain credit cards make use of exclusive marketing campaigns, such as leverage endorsements from celebrities, and tend to gain more users of a certain demographic around promotional events[6]. Conversely, other cards may face denser attrition rates due to hidden fees or policies, causing them to lose a significant portion of their user base swiftly[7].

[3] Lin, Wei, et al. "Cardholder Attrition Analysis and Treatments Framework." *2016 EMC Proven Professional Knowledge Sharing*, 2016.

[4] Hamilton, Robert, and Barry J. Howcroft. "A practical approach to maximizing customer retention in the Credit Card Industry." *Journal of Marketing Management*, vol. 11, no. 1–3, 1995, pp. 151–163, https://doi.org/10.1080/0267257x.1995.9964335.

[5] Chen, Heng, et al. "Retail payment innovations and cash usage: Accounting for attrition by using refreshment samples." *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 180, no. 2, 2016, pp. 503–530, https://doi.org/10.1111/rssa.12208.

[6] Woo, Ka-shing, et al. "An analysis of endorsement effects in Affinity Marketing: The case for affinity credit cards." *Journal of Advertising*, vol. 35, no. 3, 2006, pp. 103–113, https://doi.org/10.2753/joa0091-3367350307.

[7] Agarwal, Sumit, Souphala Chomsisengphet, and Chunlin Liu. "The importance of adverse selection in the credit card market: Evidence from Randomized Trials of credit card solicitations." *Journal of Money, Credit and Banking*, vol. 42, no. 4, 2010, pp. 743–754, https://doi.org/10.1111/j.1538-4616.2010.00305.x.

Minimal research has been conducted in the field of predicting credit card attrition, with the vast majority of work focused primarily on other incidents or conditions, such as predicting churn incidence events[8] or understanding the correlation between relationship accounts and user retention[9], with credit card attrition as a secondary concern. Of the existing research that does attempt to predict attrition, the vast majority focuses on highly specific models or datasets. For example, one paper utilizes data mining and the Recency, Frequency, and Monetary Value (RFM) analysis to understand customer spending behaviors, subsequently enabling providers to cater to evolving customer preferences and potentially mitigate attrition.[10]

Consequently, this paper aims to devise a novel method capable of predicting credit card attrition with heightened precision, specifically by examining socioeconomic factors.

3. Context & Implications

At the heart of this inquiry lies the fundamental question: how can one predict credit card attrition with a comprehensive understanding of the socioeconomic landscape? To answer this question, it is imperative to contextualize credit card attrition within the larger framework of financial behavior - attrition is not only an outcome but also an indicator of customers' changing needs, preferences, and life circumstances.

The dynamics of credit card attrition resonate deeply within the financial sector. With 77% of U.S. adults owning at least one credit card[11] and the average American juggling 3.84 credit cards[12], the intricacies of attrition have a far-reaching impact. Predicting and

[8] Van den Poel, Dirk, and Bart Larivière. "Customer attrition analysis for financial services using proportional hazard models." European Journal of Operational Research, vol. 157, no. 1, 2004, pp. 196–217, https://doi.org/10.1016/s0377-2217(03)00069-9.

[9] Agarwal, Sumit, Souphala Chomsisengphet, Chunlin Liu, et al. "Benefits of relationship banking: Evidence from consumer credit markets." SSRN Electronic Journal, 2009, https://doi.org/10.2139/ssrn.1647019.

[10] Chen, Ruey-Shun, et al. "Data mining application in Customer Relationship Management of Credit Card Business." *29th Annual International Computer Software and Applications Conference (COMPSAC'05)*, 2005, https://doi.org/10.1109/compsac.2005.67.

[11] Greene, Claire, and Oz Shy. "Payment Card Adoption and Payment Choice." *Policy Hub*, Federal Reserve Bank of Atlanta, 10 Nov. 2022, www.atlantafed.org/-/media/documents/research/publications/policy-hub/2022/07/11/10--payment-card-adoption-and-payment-choice.pdf.

[12] Stolba, Stefan Lembo. "What Is the Average Number of Credit Cards per US Consumer?" *Experian*, Experian, 17 Nov. 2022, www.experian.com/blogs/ask-experian/average-number-of-credit-cards-a-person-has/.

managing attrition is a significant endeavor, with 441 million open credit card accounts in the US as of Q3 2022[13], and the average American's household's credit card balance of $5,910 in 2022[14] underscores the financial stakes involved.

Beyond the balance sheets, credit card attrition holds societal implications that intersect with economic inequality and access to financial services. The attrition patterns of different demographic segments may underscore disparities in financial literacy, economic stability, and access to credit. Understanding how socioeconomic dimensions influence attrition can shed light on these disparities and inform strategies for fostering equitable financial inclusion.

At the crossroads of financial and social dimensions, attrition becomes a lens through which researchers can understand the interplay between economic factors, individual behavior, and broader societal trends. As U.S. consumers used their credit cards to complete 28% of transactions in 2021[15], attrition's intricate interplay with financial habits becomes evident. Predicting attrition with socioeconomic insight not only empowers credit card issuers to make informed decisions but also contributes to a more equitable financial landscape by addressing the multifaceted challenges of attrition.

[13] *Credit Card Market Monitor*, American Bankers Association, May 2021, www.aba.com/-/media/documents/reports-and-surveys/2020-q4-credit-card-market-monitor.pdf?rev=89ec658360e1413d9723da005ed574b0&hash=2FEB609D7A8A4634D298A4509187E3FB.

[14] Horymski, Chris. "Credit Scores Steady as Consumer Debt Rises in 2022." *Credit Scores Steady as Consumer Debt Balances Rise in 2022*, Experian, 24 Feb. 2023, www.experian.com/blogs/ask-experian/consumer-credit-review/.

[15] "Survey and Diary of Consumer Payment Choice." *Consumer Payments*, Federal Reserve Bank of Atlanta, 3 Aug. 2023, www.atlantafed.org/banking-and-payments/consumer-payments/survey-and-diary-of-consumer-payment-choice.aspx.

4. Data Overview & Analysis

4.1 Dataset Overview

Data for the present study are acquired from Leaps Analyttica. They were provided by the manager of a bank concerned about the increased number of customers churning from their credit card services who wants to understand the factors behind this significant customer departure in order to mitigate it and provide more helpful services to their client base.

The dataset comprises 10,000 customers each characterized by 21 features that can be broadly categorized into demographic and product-based attributes. These attributes collectively form the bedrock of this paper's exploration into credit card attrition.

Demographic-Based Attributes
The demographic-based features include variables that provide insights into customers' personal characteristics and socioeconomic context. These attributes include Customer Age, Gender, Dependent Count, Education Level, Marital Status, and Income Category.

Product-Based Attributes
Complementing the demographic features are product-based attributes that delve into customers' interaction with credit products. These attributes encompass the Client Identifying Number, Attrition Flag, Card Category, Months on Book, Total Relationship Count, Months Inactive, Contacts Count, Credit Limit, Total Revolving Balance, Total Amount Change, Total Transaction Amount, Total Count Change, and Average Utilization Ratio.

4.2 Attrition Rate

Within this dataset, a noteworthy 16% of customers have experienced attrition, signifying the prevalence and importance of this phenomenon. This attrition rate serves as a baseline against which predictive models are evaluated, enabling the assessment of the accuracy and effectiveness of these predictions in real-world scenarios.

4.3 Exploratory Data Analysis

To gain initial insights into the Exploratory Data Analysis (EDA) was conducted, focusing on demographic and product-based attributes. This analysis serves to reveal trends, disparities, and potential relationships that will inform the subsequent predictive modeling.

Percentage of Existing vs. Attrited Customers

Understanding the distribution of existing and attrited customers is essential for contextualizing this analysis. It was determined that among the 100,000 customers, approximately 84% are existing customers while the remaining 16% have experienced attrition. This breakdown provides a fundamental understanding of the composition of this dataset and the prevalence of attrition.

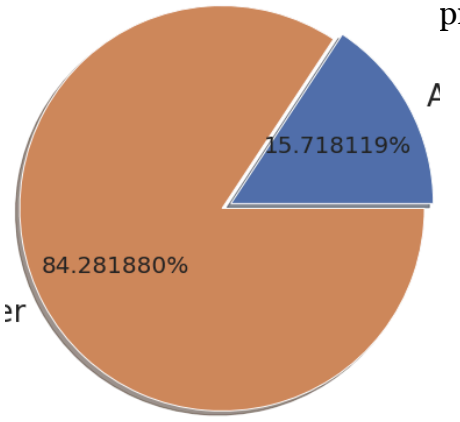


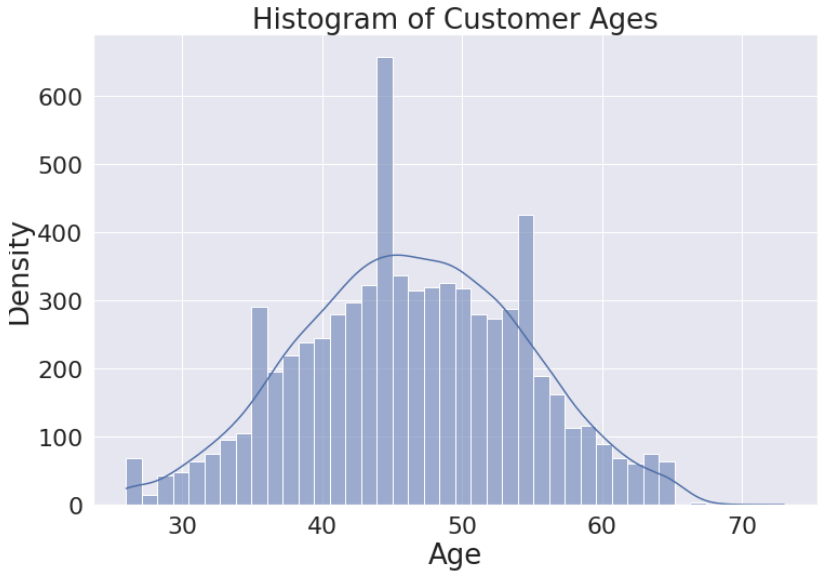
Figure 1

Existing & Attrited Customers in Data

Customer Age Distribution

Next, the distribution of customer ages across both existing and attrited customers was examined. The below histogram of customer ages provides an overview of the age composition within the dataset, highlighting any age-related trends that may contribute to attrition dynamics.

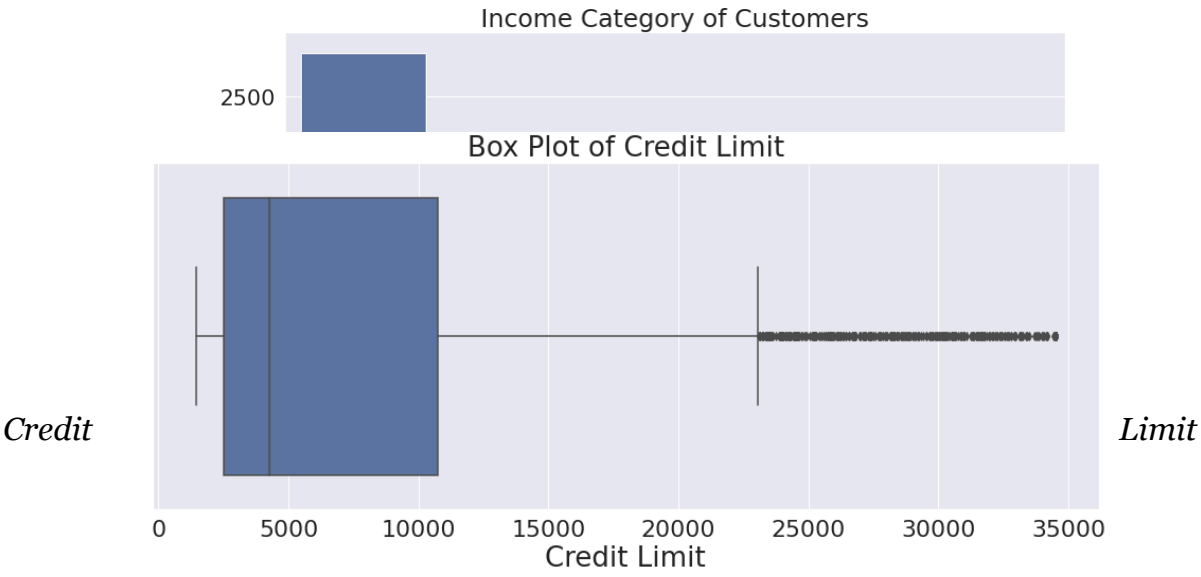
Figure 2



Income Category Overview

Analyzing the income categories of customers offers insights into the socioeconomic diversity of the dataset. By comparing income distributions between existing and attrited customers, potential disparities that may contribute to attrition tendencies can be uncovered.

Figure 3



Credit Limit

Variation

The variation in credit limits among customers can shed light on their credit utilization habits. Box plots illustrating credit limit distributions for both groups reveal insights into how credit limits are managed and whether these behaviors differ between existing and attrited customers.

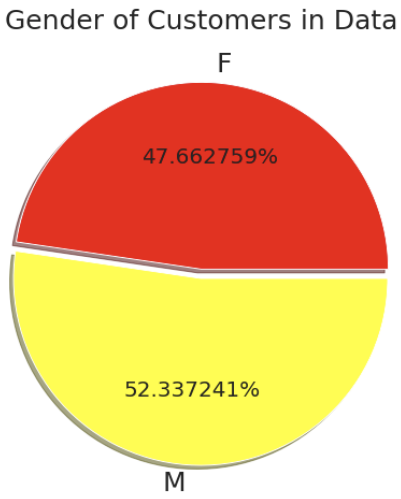
Education Level Breakdown

In the examination of education levels among customers, the distribution of educational backgrounds within each group were uncovered. This analysis reveals potential correlations between education and attrition.

Gender Analysis

A similar exploration of gender distribution among customers provides insights into the demographic composition. Comparing gender proportions between existing and attrited customers aids in identifying gender-related trends in attrition.

Figure 4



This exploratory data analysis has laid the groundwork for understanding the initial patterns and relationships within the dataset. These insights will inform subsequent predictive modeling efforts, enabling the construction of predictive models that

encompass both demographic and product-based attributes, enriched with socioeconomic context.

5. Methodology

This section outlines the methodologies employed to predict credit card attrition by integrating both traditional financial metrics and essential socioeconomic indicators are outlined. This comprehensive approach involves data preprocessing, decision trees, clustering, regression analysis, and predictive modeling techniques.

5.1 Data Preprocessing | Cleaning and Formatting Data for Analysis

Data preprocessing constitutes a pivotal initial phase in our predictive modeling framework. Raw data is subjected to a sequence of transformative steps to ensure its compatibility with subsequent analyses. Missing values are then addressed through methods such as imputation, prioritizing approaches that align with variable distribution. Categorical variables are encoded using techniques like one-hot encoding to convert them into numerical representations, facilitating model compatibility. Numerical variables are standardized or normalized to achieve uniform scales across features, mitigating undue influence from variables with larger magnitudes. The resultant dataset undergoes further scrutiny for anomalies, outliers, and inconsistencies that could skew model performance, necessitating data cleansing to uphold data integrity.

5.2 Clustering & Decision Trees | Grouping and Identifying Characteristics

This paper's approach aims to segment customers into distinct groups and identify characteristics that contribute to attrition tendencies.

Clustering

Clustering is a pivotal strategy within this analysis framework, discovering distinctive groupings within our dataset. The core of this endeavor lies in partitioning algorithms, notably the k-means algorithm, which facilitates the detection of customer behavior variances. The optimal number of clusters (k) is determined using silhouette scores and cluster attrition percentages, enabling the pinpointing of abnormal clusters with significant attrition deviations. By comparing cluster centers' features against the entire population and other clusters, insights into unique attrition patterns are gained.

Given a dataset D, of n objects, and k as the number of clusters formed, a partitioning algorithm organizes the objects into k partitions ($k \leq n$), where each partition represents a cluster. These clusters are formed to optimize objectioning partitioning such that the objects within a cluster are "similar" to each other and "dissimilar" to objects in other clusters. Formally, suppose a dataset D contains n objects which can be represented in Euclidean space. k-means partitioning methods distribute data points in D into K clusters, $C_1, \ldots, C_K$ such that $C_i \subset D$ and $C_i \cap C_J = \otimes$ for $1 \leq j \leq K, j \neq i$.

This centroid-based technique uses the centroid of a cluster, C_i to represent that cluster. The centroid is theoretically the center point of a cluster, but can be defined in various ways - for example, the mean or medoid of the objects assigned to the cluster. The difference between an object $p \in C_i$ and c_i, the representative of the cluster, is measured by the Euclidean distance $dist(p, c_i)$ between the two points p and c_i. The quality of cluster C_i is measured by the sum of the squared error between all objects in C_i and the centroid c_i, the within-cluster variation defined below as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2$$

where E is the sum of the squared error for all objects in the data set, p is the point in space representing a given object, and c_i is the centroid of cluster C_i. This objective function tries to make the resulting k clusters as compact and as separate as possible - the lower the sum, the higher the quality of clustering.
We then evaluate our clustering by employing an intrinsic method to assess clustering quality. Calculating the silhouette coefficient allows us to examine how well our clusters are separated and how compact the clusters are.

Suppose a data set D comprised of n objects is partitioned into k clusters, $C_1, \ldots, C_K$. For each object $o \in D$, we calculate $a(o)$ as the average distance between o and all other objects in the cluster to which o belongs and $b(o)$ as the minimum average distance between o and all other objects in the cluster to which o does not belong. Formally, for $o \in C_i$ $(1 \leq i \leq k)$,

$$a(o) = \frac{\sum_{o' \in C_i, o \neq o'} dist(o, o')}{|C_i| - 1}$$

and

$$b(o) = min_{C_j : 1 \leq j \leq k, j \neq 1} \left\{ \frac{\sum_{o' \in C_j} dist(o, o')}{|C_j|} \right\}.$$

The silhouette coefficient of o is then defined as the average of the coefficients of every data point,

$$s(o) = \frac{b(o) - a(o)}{max\{a(o), b(o)\}}$$

Decision Trees

Decision tree analysis is another core element of this predictive modeling framework, providing a systematic approach to explore customer data and identify influential features. It centers on the deployment of the Gini index, a pivotal metric of impurity used in tandem with binary decision splits. This methodology constructs hierarchical tree-like structures, offering an expedited data-driven framework for decision making.

Binary decision splits serve as the foundation of decision tree construction and the Gini index, a fundamental component of the Classification and Regression Trees (CART) algorithm introduced by Breiman in 1984, evaluates the potential of attribute-based binary splits to minimize impurity in a node, enhancing class segregation. It is performed through the following steps:

1. **Attribute Selection:** For each candidate attribute, the algorithm assesses binary splits to identify the optimal threshold that effectively segregates instances into child nodes

2. **Gini Index Calculation:** The Gini index is computed for both resulting child nodes after the split, quantifying impurity based on class distributions. It measures the impurity of D, a data partition or set of training tuples, as

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2,$$

where p_i is the probability that a tuple in D belongs to class C_i and is estimated by $\frac{|C_{i,D}|}{|D|}$. The sum is computed over m classes.

3. **Impurity Reduction:** The impurity reduction resulting from the attribute-based split is determined by comparing the Gini index of the parent node to the weighted sum of the Gini indices of the child nodes. For a binary split on A partitions D into D_1 and D_2, the Gini index of D given that partitioning is

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

For each attribute, each of the possible binary splits is considered. The reduction in impurity incurred by a binary split is

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

4. **Selecting the Optimal Split:** Among all candidate attributes and their associated splits, the attribute and threshold that yield the greatest impurity reduction are selected as the optimal decision point.

This project exhaustively evaluates candidate attributes and thresholds to find the split that offers the most significant split in Gini impurity, considering all possible thresholds for a given attribute and calculating the impurity reduction for each potential split. The optimal split is chosen based on the highest reduction in Gini impurity, leading to nodes that effectively differentiate between classes.

The binary splitting process is then repeated recursively as the tree branches out. The algorithm identifies decision nodes and branches that yield the most substantial reduction in impurity, creating a hierarchical structure that segments data based on the most discriminative attributes.

5.3 Regression Analysis | Identifying Impactful Features

Logistic regression is another instrumental strategy, an advanced statistical technique that models the probability of a binary outcome - in this case, attrition. The core of logistic regression lies in the sigmoid function, defined below

$$P(Y = 1 \mid X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p)}}$$

where $P(Y = 1 \mid X)$ signifies the probability of attrition ($Y = 1$) given the feature vector X, while $\beta_0, \beta_1, \ldots, \beta_p$ represent the coefficients to be estimated.

The most critical aspect of performing logistic regression lies in estimating the model coefficients that maximize the likelihood of observing the given data, which are solved through Maximum Likelihood Estimation (MLE). The likelihood function calculates the probability of observing the actual outcomes Y given the corresponding feature matrix X and the estimated coefficients β,

$$L(\beta) = \prod_{i=1}^{n} \left[P(Y_i = 1 \mid X_i)\right]^{Y_i} \left[1 - P(Y_i = 1 \mid X_i)\right]^{1-Y_i}$$

Once the set of β values that maximize the log-likelihood function are identified through the iterative model of Gradient Descent, the resultant β values reflect the best-fitting coefficients that align the model's predicted probabilities with the observed outcomes, encapsulating the essence of the dataset's attrition behavior.

5.4 Predictive Modeling | Creating Optimal Predictive Models

The final phase of analysis involves predictive modeling to construct a robust attrition prediction model. A solution is derived through simple ensemble learning, a technique that utilized the combined predictive insights of diverse algorithms.

The logic behind ensemble learning is that a collective of models surpasses the results of the contribution of its parts. Basic ensemble learning amalgamates predictions from various models in order to create a more robust prediction of attrition patterns.

Predictions from the distinct models of decision trees, logistic regression, and neural networks are combined to identify the most insightful features of attrition prediction, defined mathematically as

$$Ensemble\ Prediction\ =\ \frac{1}{N}\sum_{i=1}^{N} Individual\ Model\ Prediction_i.$$

These predictions undergo rigorous evaluation using metrics such as accuracy, specificity, and precision, defined below

$Precision = \frac{TP}{TP + FP}$ $Sensitivity = \frac{TP}{TP + FN}$ $Specificity = \frac{TN}{TN + FP}$

Through the lenses of data preprocessing, clustering, decision trees, regression analysis, and predictive modeling, the above analysis aims to unravel the intricate tapestry of customer behavior.

By meticulously preparing the data, a solid foundation for in-depth analyses was established. The application of clustering and decision trees unveiled distinct customer segments, offering glimpses into the attributes that steer attrition dynamics. Simultaneously, the logistic regression analysis illuminated the unique influence of variables, uncovering the pivotal contributors to attrition tendencies. Finally, a diverse array of algorithms converges to forecast customer attrition in predictive modeling.

Transitioning from methodology to insight, the path ahead promises empirical revelations that intertwine data patterns with real-world implications. The synthesis of these analytical findings has the potential to reshape customer retention strategies, empowering stakeholders with actionable insights to bridge the gap between academic inquiry and practical application, transforming knowledge into informed decisions.

6. Results & Conclusions

This section presents the outcomes of data-driven approaches across clustering, decision trees, regression, and predictive modeling. Each technique contributes a unique perspective to the understanding of credit card attrition dynamics, providing insights into customer behavior, influencing features, and the predictive power of models.

Clustering analysis revealed distinct customer segments with varying attrition tendencies. Low usage customers and high rollers emerged as indicative of attrition, suggesting that both inactive users and those with high credit utilization are likely to attrite. Interestingly, working class mothers exhibited the lowest attrition rates, highlighting a potential demographic group with strong retention potential.

Regression analysis identified features that hold the most substantial impact on the likelihood of attrition. Dependents, income, gender, and marital status emerged as pivotal variables influencing attrition behavior. These insights provide a nuanced view of the socioeconomic and personal factors that drive customer attrition.

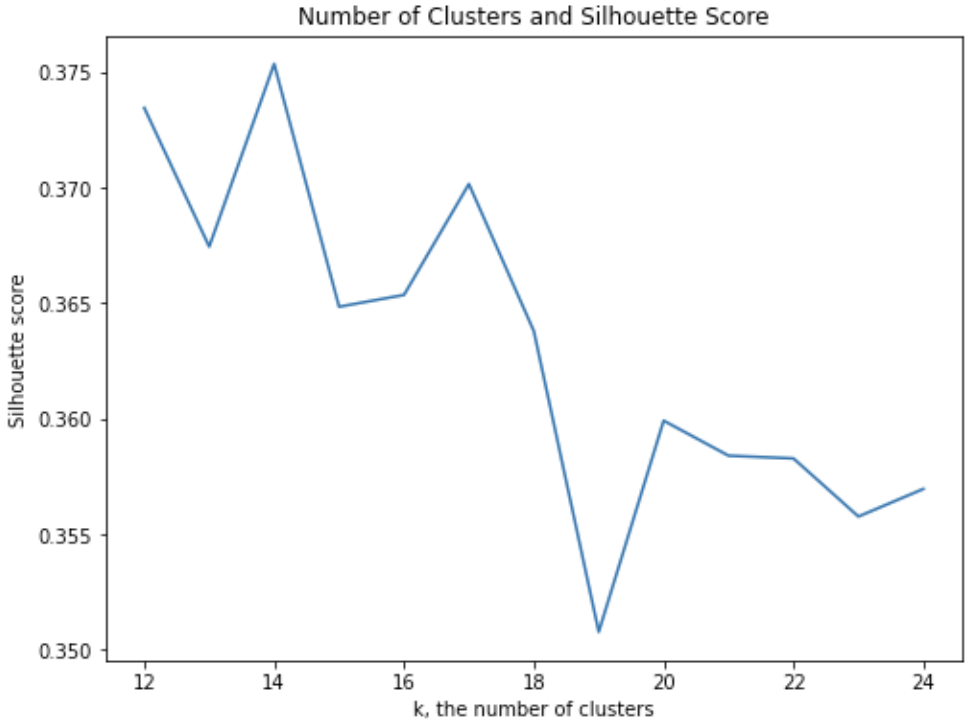
Finally, predictive modeling integrated insights to create accurate models for attrition prediction, demonstrating enhanced accuracy, particularly in predicting customers who remain loyal. However, the predictive performance was relatively less successful for customers who exhibit attrition tendencies.

6.1 Clustering

Cluster Number

For k-means clustering analysis, determining the optimal number of clusters is vital. Thorough evaluation found that k = 14 yielded the highest silhouette score. This value signifies meaningful clusters where attrition within the cluster is significantly different from the overall attrition rate of 16%. In contrast, k values below 12 failed to generate clusters with substantial differentiation.

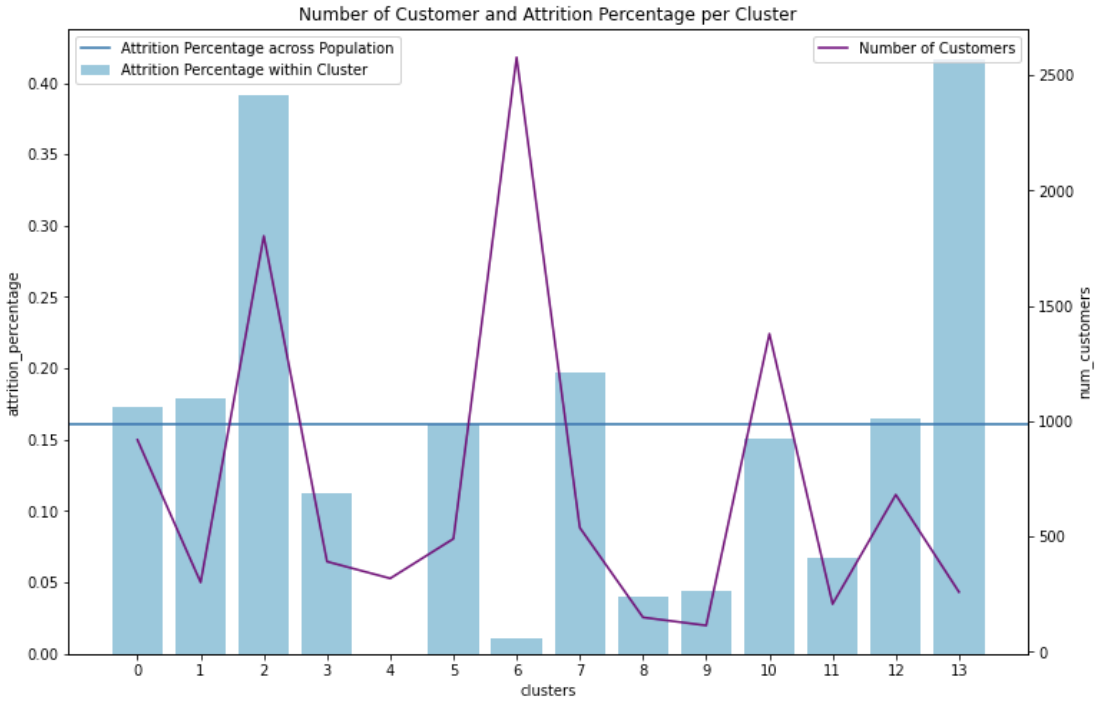
Figure 5



Abnormal Clusters

Among the 14 clusters, three emerged as particularly influential due to their attrition percentages and customer counts. Specifically, clusters 2, 6, and 13 stood out. Cluster 2, with a high attrition rate (n=1803), and clusters 6 (n=2578) and 13 (n=259), with low attrition rates stood out due to their statistical significance and the insights they could offer.

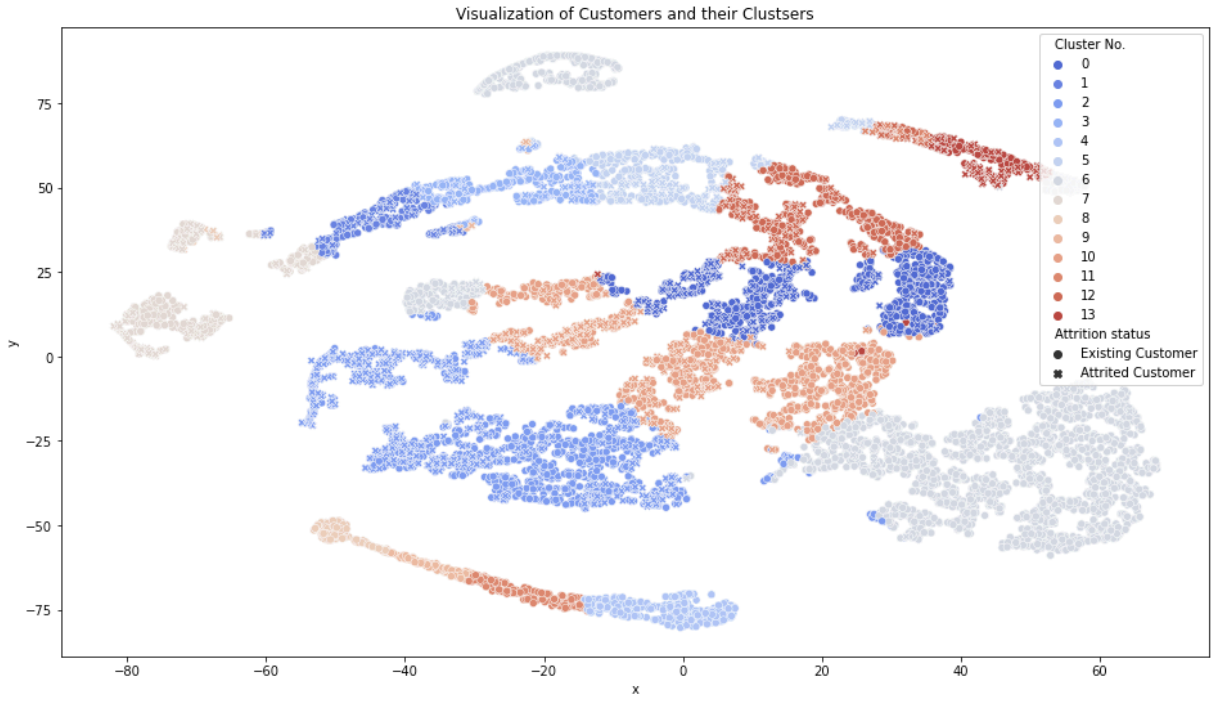
Figure 6



Cluster Significance

These three clusters collectively account for 50% of all attrited users. Cluster 13 represents the high attrition group, while cluster 6 showcases low attrition. The presence of these clusters emphasizes their substantial influence on the overall attrition landscape.

Figure 7



Summary of Clustering Results

In examining these clusters, specific features emerge as most characteristic, providing valuable insights into the unique behaviors of each group:

- **Cluster 2:** Low Usage Users
    - These users exhibit a 35% lower transaction count compared to the entire population.

    - On average, their transaction amount is 50% lower than the population average.
    - A higher likelihood of being married, at 57% within this cluster compared to 46% overall.
- **Cluster 6:** Working Class Mothers
    - Predominantly female (80%) compared to the overall average (53%).
    - Income distribution skewed toward the <$40K bracket (60% compared to 35% overall).
    - Displaying an impressive 17% more transactions than the population average, akin to Cluster 13's high rollers.
    - Demonstrates 6% more dependents than average, which is 24% more than the median.
- **Cluster 13:** High Rollers
    - Significantly higher credit limit and transaction amount, up by 35% and 107%, respectively, compared to the median.
    - 13% more of this cluster has an income above $60K compared to Cluster 2.
    - Holds 41% fewer products than the customer average, reflecting a potentially specialized product utilization strategy.

Clustering analysis has unearthed distinctive customer segments, each revealing distinct behaviors and characteristics. This granular understanding serves as a foundation for subsequent analyses, enabling the identification of drivers and formulation of strategies that cater to the diverse needs and preferences of these segments.

6.2 Decision Trees

Then, decision trees were utilized to uncover key features that might serve as strong indicators of attrition. However, upon rigorous exploration, it was determined that decision trees do not meet the accuracy thresholds required for accurate feature

selection in predictive models. High Gini scores and limited predictive accuracy compelled the exploration of alternative methods for identifying pivotal attributes.

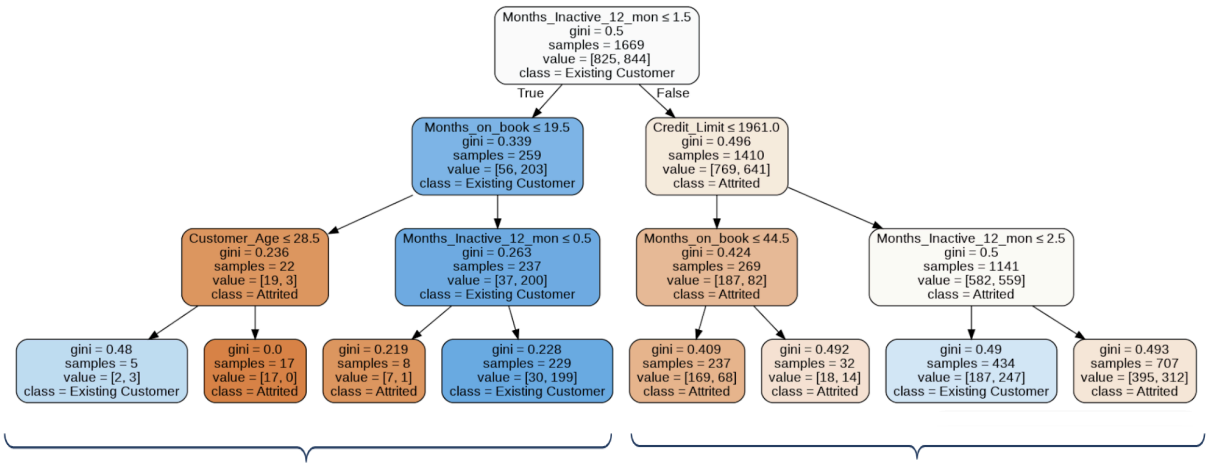
Overview of Decision Tree Analysis

With a focus on decision trees as a preliminary analysis, a dataset containing 1,669 samples was analyzed. The decision tree model was configured with a maximum depth of 3 and a minimum leaf size of 5, aiming to strike a balance between model complexity and interpretability. This configuration allowed for meaningful splits in the data without overfitting.

Performance Metrics and Insights

On the validation dataset, the decision tree achieved an accuracy of approximately 61%. While this level of accuracy is informative, it did not meet the stringent requirements for robust feature selection. Despite achieving a reasonable accuracy rate, the decision tree's Gini scores indicated that the predictive power of individual features was not sufficient to support feature selection in our comprehensive attrition prediction models.

Figure 8



- Splits have lower gini scores

- Customer_Age_Months_Inactive _12_mon & Months_on_book are better splitting points
- But only contain 259 data points out of 1669
- Splits with higher gini scores, quite ineffective
- Months_on_book & Credit_Limit are better splitting points
- Containing majority 1410 data points out of 1669

The decision tree analysis, although insightful in providing a preliminary understanding of feature importance, falls short of serving as an effective proxy for feature selection. Consequently, alternative methods that better integrate the insights gained from clustering, regression, and predictive modeling to achieve enhanced predictive accuracy and a deeper understanding of the driving factors behind attrition behavior were used.

6.3 Regression

The regression analysis unveiled pivotal factors that significantly influence the likelihood of attrition. Notably, dependents, income, gender, and marital status stand out as key drivers of attrition behavior.

Table 2: Regression Results

Variable	Coefficient
Dependent_Count	-144e-01
Total_Trans_Amt	-1.88e-04
Total_Trans_Ct	4.45e-02

Credit_Limit	2.97e-06
Avg_Utilization_Ratio	1.68e-01
Gender_F	-1.59e-01
Gender_M	4.10e-02
Income_Category_$120K +	-2.14e-02
Income_Category_$40K-$60K	5.55e-02
Income_Category_$60K-$80K	2.99e-02
Income_Category_$80K-$120K	-3.15e-03
Income_Category_Less_than_$40K	-1.29e-01
Marital_Status_Divorced	-3.33e-02
Marital_Status_Married	7.85e-02
Marital_Status_Single	-1.63e-01

Effect of Dependents

Increasing the number of dependents by 1, such as children, results in a substantial 14.4% decrease in the likelihood of attrition. This finding underscores the significance of familial responsibilities in shaping customers' decisions regarding their credit card engagement.

Impact of Income

Customers with an income level below $40K are notably less likely to attrite, showcasing a 12.9% reduction in the attrition likelihood. This observation hints at the relationship between financial stability and the propensity to remain engaged with credit services.

Role of Gender

Gender emerges as a noteworthy influencer. Women exhibit a significantly lower attrition likelihood, with a reduction of 15.9% compared to the baseline. In contrast, men are 4% more likely to attrite. These gender-based disparities in attrition likelihood provide valuable insights into nuanced behavioral dynamics.

Influence of Marital Status

Marital status also proves to be a significant factor. Individuals who are single exhibit a 16.3% reduced likelihood of attrition. This finding suggests that marital status plays a role in customers' credit card engagement decisions, potentially linked to financial responsibilities or risk perceptions.

Statistical Significance Consideration

It is important to note that while these coefficients have been determined to be significant in influencing the likelihood of attrition, the absence of standard error values from sklearn does not facilitate the confirmation of the statistical significance of these coefficients. Thus, further analysis is required to fully validate these results.

Conclusion of Regression Results

In sum, the regression analysis unveils a multifaceted interplay between personal, socioeconomic, and demographic factors that collectively shape the likelihood of attrition. These insights provide financial institutions with a deeper understanding of the intricate drivers that impact customer retention and inform the development of targeted strategies for reducing attrition rates.

6.4 Prediction Modeling

The pursuit of the most accurate predictive model led to the conclusion that a simple ensemble approach holds the highest potential for accurately predicting credit card

attrition. This approach harnesses the strengths of various algorithms while mitigating their respective weaknesses, ultimately yielding enhanced accuracy in attrition predictions.

Table 3: Prediction Modeling Results

Algorithm	Training Score	Validation Score	Notes
Decision Trees	1.0	0.91	
Decision Trees with Grid Search	0.93	0.93	Better accuracy than plain decision trees as it's a comprehensive sweep of hyperparameter space
Random Forest	0.94	0.61	Overfitting
Neural Nets	0.84	0.85	solver = lbfgs, activation = logistic, hidden_layers = 20, max_iter = 1000
Ensemble	**0.96**	**0.92**	**Simple Ensemble**

Insights into Random Forest Model

While the random forest model exhibited promising results, a closer examination raised concerns about potential overfitting. The nature of the random forest, which consists of sections of decision trees, can exacerbate overfitting when faced with imbalanced datasets. Given that only 16% of customers within the dataset have experienced attrition, the number of attrited customers when broken down into multiple decision trees becomes insufficient for accurate model learning.

Initial Ensembling Attempt

The initial foray into predictive modeling focused on factors assumed to play a crucial role in influencing attrition rates. Guided by intuition and informed speculation, variables were selected that held the promise of unveiling nuanced relationships between customer attributes and attrition behavior. The chosen variables included Customer Age, Gender, Dependent Count, Education Level, Marital Status, Income Category, and Credit Limit.

Employing a straightforward ensemble approach that leveraged the predictive power of these selected variables achieved an accuracy rate of approximately 80%, reflecting an impressive degree of predictive strength.

Post-Clustering Ensembling Attempt

Building upon the insights gleaned from our clustering analysis, an ensembling approach that harnessed the power of factors empirically determined to be most predictive of attrition was formulated. Informed by the clustering outcomes, variables that carried significant weight in defining the attrition dynamics within distinct customer segments were identified. The chosen variables included Dependent Count, Total Transaction Amount, Total Transaction Count, Gender, Income Category, Marital Status, Credit Limit, and Average Utilization Ratio.

The post-clustering ensembling approach yielded an impressive accuracy rate of 91%. This outcome significantly surpasses the accuracy level achieved in the initial ensembling attempt and underscores the potency of incorporating insights from clustering analysis into our predictive models.

Both ensembling attempts contribute valuable insights. The initial ensembling attempt showcased the potential of the chosen variables, while the post-clustering ensembling approach exemplified the transformational influence of data-driven insights.

Collectively, these endeavors empower stakeholders within the credit card industry to make informed decisions that mitigate attrition, foster customer retention, and navigate the complex terrain of customer behaviors.

6.5 Error Metrics

Error metrics offer a comprehensive lens through which the efficacy of predictive models is gauged. The models showcase commendable precision in predicting existing users, yielding an accuracy rate of 92%. This proficiency extends to models' sensitivity, which stands at an impressive 98%. This high sensitivity underscores their capability to identify true positives within the realm of attrited customers. Conversely, the models reveal a distinct pattern of specificity, registering at 59%, which highlights the challenge of accurately forecasting attrition instances. To delve deeper, precision signifies the accuracy in predicting attrition, sensitivity pertains to predicting individuals who will churn, while specificity gauges our proficiency in anticipating customers who won't churn.

6.6 Approach Improvements

In the quest for predictive precision, a spectrum of techniques aimed at bolstering the robustness of predictive models were explored. Among these, strategic improvements that address key aspects of the model-building process were employed.

Oversampling
Oversampling the under-represented class emerged as a strategic approach to redress the imbalance inherent in the dataset. Rather than simply reducing the over-represented class, this paper amplified the under-represented class through strategic sampling. In doing so, the emergence of more refined predictive patterns within the minority class was promoted, ultimately enhancing the model's capability to predict attrition instances accurately.

Bagging

The adoption of bagging, a technique of generating multiple bootstrap training samples with replacement, surfaced as a potent tool for mitigating overfitting. The diversification introduced by these samples curtails the model's propensity to over-learn from specific instances, thus enabling it to generalize better to unseen data points.

Cluster the Abundant Class

This exploration delved into cluster-based refinement of the abundant class, particularly to address the intricacies of the dataset's class imbalance. By clustering the abundant class into groups, retaining medoids for each group, and training the model on rare class instances alongside the medoids, a balance was struck between model complexity and accuracy in addressing the imbalanced landscape.

Choose Models Based on Different Metrics

This paper advocates for a nuanced approach to model selection, transcending accuracy as the sole proxy for effectiveness. Instead, our paper proposes evaluating models based on a range of metrics, including specificity, recall, and the area under the ROC curve (AUC). This diversified approach empowers the selection of models that excel in various aspects of predictive performance, ensuring a more comprehensive understanding of model efficacy.

Through these strategic refinements, our paper endeavors to construct predictive models that are not only accurate but also resilient, capable of navigating the complexities of credit card attrition prediction with a holistic and robust outlook.

6.7 Implications

The insights derived from the predictive modeling endeavors bear significant implications for financial institutions aiming to proactively address credit card attrition.

Equipped with a nuanced understanding of customer behavior, banks can now tailor strategies that not only curtail attrition rates but also foster lasting customer engagement.

Support & Incentives

The clustering analysis sheds light on distinct customer segments that are more likely to attrite, such as the high rollers and low usage customers. Armed with this insight, banks can allocate targeted support, loyalty rewards, and incentives to these clusters. By recognizing the specific needs and preferences of these segments, financial institutions can fortify customer loyalty and engender a sense of valued partnership.

Customized Service

The regression features with the most pronounced impact – dependents, income, gender, and marital status – can serve as compasses for banks seeking to optimize resource allocation. Therefore, banks can prioritize services tailored to these specific customer attributes, creating a customized approach that empowers institutions to streamline efforts and channel resources where they are most impactful, engendering a customer-centric paradigm.

Qualitative Insight

While these data-driven methodologies provide valuable insight into which customers are likely to leave, the underlying motivations often remain enigmatic. To bridge this gap, financial institutions should consider adopting a qualitative approach. By coupling these predictive models with qualitative surveys, banks can unveil the underlying motivations, needs, and grievances that steer customers towards attrition and subsequently design holistic retention strategies that address the root causes and foster enduring customer relationships.

These implications serve as guideposts for institutions to navigate the terrain of customer retention with data-backed strategies that foster loyalty, enhance service, and forge a profound understanding of customer behavior.

7. Appendix

Table 4 - Feature Descriptions

CLIENTNUM	Client number; unique identifier for the customer holding the account
Attrition_Flag	Internal event (customer activity) var – 1 if the account is closed, else 0
Customer_Age	Demographic var – customer's age in years
Gender	Demographic var – M = Male, F = Female
Dependent_Count	Demographic var – # of dependents
Education_Level	Demographic var – educational qualifications of the account holder (e.g. high school, masters)
Marital_Status	Demographic var, e.g. Married, Single, Divorced
Income_Category	Demographic var, e.g. $40K - $60K
Card_Category	Product var, type of card (e.g. blue, silver, gold, platinum)
Months_on_Book	Period of relationship with bank
Total_Relationship_Count	Total # of products held by the customer
Months_Inactive_12_mon	# of months inactive in the last 12 months

Contacts_Count_12_mon	# of contacts in the last 12 months
Credit_Limit	Credit limit on the credit card
Total_Revolving_Bal	Total revolving balance on the credit card
Avg_Open_To_Buy	Open to buy credit line (average of last 12 months)
Total_Amt_Chng_Q4_Q1	Change in transaction amount (Q4 over Q1)
Total_Trans_Amt	Total transaction amount (last 12 months)
Total_Trans_Ct	Total transaction count (last 12 months)
Total_Ct_Chng_Q4_Q1	Change in transaction count (Q4 over Q1)
Avg_Utilization_Ratio	Average Card Utilization Ratio

8. References

Agarwal, Sumit, Souphala Chomsisengphet, Chunlin Liu, et al. "Benefits of relationship banking:

Evidence from consumer credit markets." SSRN Electronic Journal, 2009, https://doi.org

/10.2139/ssrn.1647019.

Agarwal, Sumit, Souphala Chomsisengphet, and Chunlin Liu. "The importance of adverse selection in the

credit card market: Evidence from Randomized Trials of credit card solicitations." Journal of

Money, Credit and Banking, vol. 42, no. 4, 2010, pp. 743–754, https://doi.org/10.1111

/j.1538-4616.2010.00305.x.

"Card Member Attrition Rates Softening." *CUInsight*, Card Knowledge Factory, www.cuinsight.com/

press-release/card-member-attrition-rates-softening/#:~:text=Gross%20attrition%20is%20defined

%20as,during%20the%20same%20reporting%20period.

Chen, Heng, et al. "Retail payment innovations and cash usage: Accounting for attrition by using

refreshment samples." Journal of the Royal Statistical Society Series A: Statistics in Society, vol.

180, no. 2, 2016, pp. 503–530, https://doi.org/10.1111/rssa.12208.

Chen, Ruey-Shun, et al. "Data mining application in Customer Relationship Management of Credit Card

Business." 29th Annual International Computer Software and Applications Conference

(COMPSAC'05), 2005, https://doi.org/10.1109/compsac.2005.67.

Credit Card Market Monitor, American Bankers Association, May 2021, www.aba.com/-/media/

documents/reports-and-surveys/2020-q4-credit-card-market-monitor.pdf?rev=89ec658360e1413d

9723da005ed574b0&hash=2FEB609D7A8A4634D298A4509187E3FB.

Duarte, Fabio. "Amount of Data Created Daily (2023)." *Exploding Topics*, Exploding Topics, 3 Apr.

2023, explodingtopics.com/blog/data-generated-per-day.

Gre ene, Claire, and Oz Shy. "Payment Card Adoption and Payment Choice." *Policy Hub*, Federal Reserve

Bank of Atlanta, 10 Nov. 2022, www.atlantafed.org/-/media/documents/research/publications/

policy-hub/2022/07/11/10--payment-card-adoption-and-payment-choice.pdf.

Hamilton, Robert, and Barry J. Howcroft. "A practical approach to maximizing customer retention in the

Credit Card Industry." Journal of Marketing Management, vol. 11, no. 1–3, 1995, pp. 151–163,

https://doi.org/10.1080/0267257x.1995.9964335.

Han, Jiawei, et al. Data Mining: Concepts and Techniques. 3rd ed., Elsevier Science, 2011.

Horymski, Chris. "Credit Scores Steady as Consumer Debt Rises in 2022." *Credit Scores Steady as*

Consumer Debt Balances Rise in 2022, Experian, 24 Feb. 2023, www.experian.com/blogs/

ask-experian/consumer-credit-review/.

Lin, Wei, et al. "Cardholder Attrition Analysis and Treatments Framework." 2016 EMC Proven

Professional Knowledge Sharing, 2016.

Mourtas, Spyridon D., et al. "Credit card attrition classification through Neuronets." European

Proceedings of Computers and Technology, 2023, https://doi.org/10.15405/epct.23021.11.

Stolba, Stefan Lembo. "What Is the Average Number of Credit Cards per US Consumer?" *Experian*,

Experian, 17 Nov. 2022, www.experian.com/blogs/ask-experian/average-number-of-credit-cards-

a-person-has/.

"Survey and Diary of Consumer Payment Choice." *Consumer Payments*, Federal Reserve Bank of

Atlanta, 3 Aug. 2023, www.atlantafed.org/banking-and-payments/consumer-payments/

survey-and-diary-of-consumer-payment-choice.aspx.

Van den Poel, Dirk, and Bart Larivière. "Customer attrition analysis for financial services using

proportional hazard models." European Journal of Operational Research, vol. 157, no. 1, 2004,

pp. 196–217, https://doi.org/10.1016/s0377-2217(03)00069-9.

Woo, Ka-shing, et al. "An analysis of endorsement effects in Affinity Marketing: The case for affinity

credit cards." Journal of Advertising, vol. 35, no. 3, 2006, pp. 103–113,

https://doi.org/10.2753/joa0091-3367350307.

Cantor's Infinities

Phineas Manogue '24

Pure Mathematics

Infinity. A measurement "without limits". A concept that has taunted the minds of humanity for eons. Since the conceptualization of apeiron (the boundless) in 6th century B.C. by Greek philosopher Anaximander, infinity has grown to be a topic of much debate and controversy. By the late 19th century, Russian mathematician and philosopher Georg Cantor had revolutionized the discussion.

To understand Cantor's theories, two fundamentals must be established. The first of these fundamental concepts that must be accepted is that infinity is a totality. That is, infinity is the whole of something – the sum.

The second of these rudiments is one-to-one correspondence. As Cantor said, one-to-one correspondence is when "Two sets M and N are equivalent… if it is possible to put them, by some law, in such a relation to one another that to every element of each one of them corresponds one and only one element of the other". In other words, two sets are the same when a pairing of the two sets leaves no element unpaired. An example of this would be if I take all of the fingers in my left hand and match them with all of the fingers in my right hand. Each finger is paired up with exactly one finger from the other hand so I must have the same amount of fingers on each hand. In this way, one-to-one correspondence is much simpler than counting as I had no need to specifically mention that I have five fingers on each hand. When two data sets have the same number of elements, they have the same *cardinality*.

Using these concepts, Cantor began his theory by diving into the realm of countable sets. For now, we can think of a set to be countable when it is possible to put it into a list e.g. {1, 2, 3, 5, 9}. This means that all finite sets, sets with a limited number of elements, are countable.

Cantor furthered this definition of a countable set by taking all of the natural numbers, N. The natural numbers include all positive integers. Since it is possible to list all of the natural numbers, e.g. {1, 2, 3, 4, 5…}, N is a countable set. Next, Cantor took a set with all of the even integers, E, e.g. {2, 4, 6, 8, 10…}. Using one-to-one correspondence, sets N and E can be put into relation with each other: 1 pairs with 2, 2 pairs with 4, 3 with 6, and so on. This relationship can be put simply as element n from N pairs with element $2n$ from E. Although one might assume that set E would be double that of set N, one-to-one correspondence tells us that the two sets are equal.

For a set with all of the integers, Z, however, a pairing of all of the lowest values would not work as Z extends forever in both the positive and negative directions. To

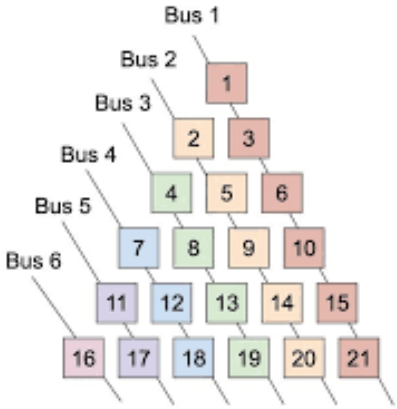
account for this, the integers can be put in order of their magnitude or absolute value, e.g. {0, 1, -1, 2, -2, 3, -3…}. Now Z can be put into one-to-one correspondence with N as 1 pairs with 0, 2 pairs with 1, 3 pairs with -1, and so on.

This means that the cardinality of N is equal to the cardinality of E which is equal to the cardinality of Z. Cantor labeled this cardinality, the cardinality of countable infinities, as aleph-null or aleph-nought ($\aleph_0$). Years later, German mathematician David Hilbert explored Cantor's theory of aleph-null with a paradox that would become known as Hilbert's Grand Hotel.

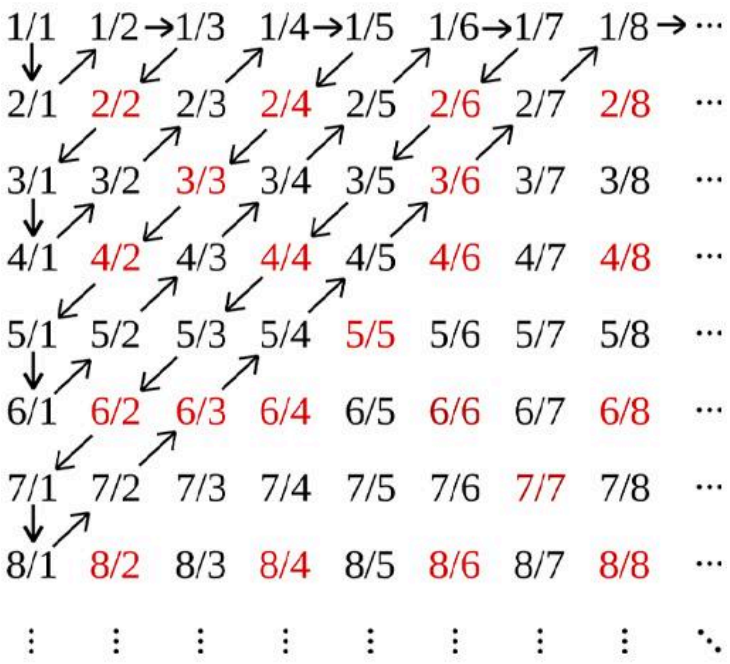
Assume that there is a hotel with a countably infinite number of rooms that is currently holding a countably infinite number of people. The hotel is full, right? No. When one more person comes to the hotel looking for a room, the manager has an idea. He tells all of the people who already have a room to move down one room. The person in room 1 moves to room 2, the person in room 2 moves to room 3, the person in room 12 moves to room 13, and so on. This would work for all finite numbers of people as the manager could just ask everybody in a room to move down however many are needed to accommodate for the new customers. This means that $\aleph_0 + 1 = \aleph_0$.

Then, a bus with a countably infinite number of people arrives. The manager can't just ask everybody in a room to move down a countably infinite number of rooms, so what can he do? The manager recalls Cantor's theory about one-to-one correspondence where the number of natural numbers (n) is equal to the number of even numbers (2n) and he tells all of the people residing in a room to look at whatever room number they are currently in, double it, and then move to that room. The person in room 1 moves to room 2, the person in room 2 moves to room 4, the person in room 127 moves to room 254, and so on. This means that $\aleph_0 + \aleph_0 = \aleph_0$.

Just when the manager thinks that he is in the clear, a countably infinite number of buses arrive each with a countably infinite number of people. After some thinking, the manager comes up with a way to serve all of these people: rearrange the rooms into the shape of a triangle. He then tells all of the people who currently have a room to take the rightmost room of each row. That is, the person in room 1 will take room 1, room 2 will take room 3, room 3 will take room 6, 4 to 10, 5 to 15, and so on. The manager knows that because the triangle is infinite, removing the rightmost rooms from every row will leave a similarly infinite triangle. He then tells the people in the first bus to take the free right most room in each row. He does this with each bus so that each bus is assigned the free right most room of each column. $\aleph_0 \cdot \aleph_0 = \aleph_0$.



Hilbert's Hotel does not just work because the numbers are discrete or have gaps. Fractions have no discrete gaps. Between each fraction is another fraction that is attainable by taking the average of the two fractions. Fractions, too, can be put into one-to-one correspondence with N. Although it is hard to write fractions out as a list, a way has been devised to do so. By writing out all of the natural numbers in column (or row), e.g. {1, 2, 3, 4…}, and then all of the halves in a column next to that, e.g. $\{\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, 3\ldots\}$, and all the thirds next to that and so on, and connecting each number with a line that crosses diagonally until it reaches a side and must "snake around", all of the fractions will be displayed. To turn this display into a list, simply follow the snaking line and write out each number that it goes through, skipping those that have already been listed, e.g. $\{1, 2, \frac{1}{2}, \frac{1}{3}, 3, 4, \frac{3}{2}, \frac{2}{3}\ldots\}$.



The red numbers are repeated and are therefore not listed.

With all of the fractions listed in such a way, it is possible to put them into one-to-one correspondence with the natural numbers and therefore all of the fractions create a countable infinity. The fractions cardinality is $\aleph_0$. Going back to Hilbert's Hotel, if a bus with all of the fractions were to arrive, the manager would be able to devise a way to accommodate all of the new customers.

Unfortunately for Hilbert's Hotel, they can not accommodate all numbers. Despite the fractions seemingly continuous strand, a great amount of irrational numbers lie in between them. Irrational numbers are those that can not be represented by an integer divided by another integer. Before Cantor, it was widely believed that there were more rational numbers than irrational numbers – if there are an infinite number of integers then shouldn't most numbers be able to be represented by a quotient? To begin testing this theory, Cantor made a set such as the following.

0.109982438...
0.234567892...
0.382346780...
0.461002331...

Each of these being a real number, Cantor then added a new number to the set by first taking the first decimal from the first number and choosing a new digit other than that number (1), say 3. Cantor then took the second decimal from the second number and chose a new digit other than that number (3), say 8. Cantor could then continue this trend with the third, fourth, and all other numbers in the set to create an entirely new number. Using this list of all real numbers between 0 and 1, Cantor then tried to put it to one-to-one correspondence with N such that N paired with X_n, an infinitely long decimal between 0 and 1. Though some of the numbers within the list of real numbers between 0 and 1 could be matched with the integers, using Cantor's strategy of replacing a digit for every number in the list would then create another number that would need to be matched. Since this could be done forever as adding a new number to the list would allow for another number to be created in which the last decimal place would be replaced with another digit, Cantor determined that the set of real numbers thus has a greater cardinality than the set of integers and the two sets could not be matched. Cantor labeled this larger, uncountable infinity as c.

Now that Cantor had proved that there was an uncountably infinite number of real numbers and a countably infinite number of rational numbers, he was ready to return to the irrational number debate. Using his knowledge of sets, he knew that since all real numbers were uncountably infinite, they must contain an uncountably infinite set within them. As the rational numbers are countably infinite, the irrational numbers must therefore be uncountably infinite and therefore larger in cardinality.

Set	Description	Cardinality
Natural numbers	1, 2, 3, 4, 5...	$\aleph_0$
Integers	..., -3, -2, -1, 0, 1, 2, 3...	$\aleph_0$
Rational numbers or fractions	All the decimals which terminate or repeat	$\aleph_0$
Irrational numbers	All the decimals which don't terminate or repeat	c
Real numbers	All decimals	c

Bibliography

https://www.gresham.ac.uk/watch-now/cantors-infinities

https://ui.adsabs.harvard.edu/abs/2011A%26AT...27..153T/abstract

Georg Cantor's life - Britannica

Totality definition

David Hilbert - Britannica.

https://byjus.com/maths/irrational-numbers/

Podcast: The Infinity Hotel (stuff to blow your mind)

Millikan Oil Drop Experiment

Derivation

Nicholas Lu '25

Applied Mathematics

To derive the charge of an electron (q) from the Millikan oil drop experiment, we use the equilibrium condition where the electrical force on a charged oil droplet is balanced by the gravitational force acting on it. This balance allows us to solve for q, the charge on the oil droplet.

$$qE = mg$$

Where:

- q is the charge of the oil droplet we want to find,
- E is the electric field strength between the plates ($E = \frac{V}{d}$, where V is the applied voltage and d is the distance between the plates),
- m is the mass of the oil droplet, and
- g is the acceleration due to gravity.

The mass m of the oil droplet can be related to its volume and density (ρ). Given that the volume V of a spherical droplet is $\frac{4}{3}\pi r^3$ and assuming the density of the oil is known, the mass m can be calculated as:

$$m = \frac{4}{3}\pi r^3 \rho$$

The electric field strength between the plates is given by:

$$E = \frac{V}{d}$$

Substituting m and E into the equilibrium equation gives:

$$q\frac{V}{d} = \frac{4}{3}\pi r^3 \rho g$$

Rearranging for q, the charge on the droplet, gives:

$$q = \frac{\frac{4}{3}\pi r^3 \rho g d}{V}$$

This equation allows us to calculate q for a single droplet. However, to find the charge of an electron, Millikan observed the charges on many droplets and noted that q was always a multiple of the smallest charge measured, which is the charge of a single electron (e). Through his experiments, he found that:

$$e = 1.6 \cdot 10^{-19}\, C$$

Predictive Analysis of Invasive Plants

Nicholas Lu '25

Applied Mathematics

1. Introduction

The regions under study are Eastern and Southern Africa, along with California, whose ecosystems, while officially protected, are currently exposed to invasion by invasive plant species (Obiri 2011). In Eastern and Southern Africa, the spread of invasive species is impacting the livelihoods of the rural poor who are dependent on natural resources for income and survival, as well as undermining foreign development investment in the area. In California, the largest agricultural state is threatened. Given that their dependency on natural resources has increased more significantly over the last decade (Shackleton et al. , 2000, and Witt et al. 2010), efforts to understand the root causes of the growth and spread of invasive species are needed to be proactive about their current and future survival. However, there has been a limited study on what promotes or prevents the growth of these invasive species in those regions. Therefore, using the methods to create various models, this study investigates how the time elapsed and the type of the invasive species impacts their own spread, then uses this information to predict the spread of those invasive species.

2. Literature Review

Plants spread by dispersing their seeds using everything from fire to animals. Some studies have already measured the quantity and quality of seed dispersal by animals in general (Schupp 1993), and others focusing on measuring more specific animals such as birds (Gosper et al. 2005). Similarly, studies have been written surrounding the spread of invasive plants in relation to fire, specifically fire's ability to create new habitats (Maringer et al. 2012), which further presents another way plants can spread. Invasive species are known to spread in different ways. More specifically, some invasive species have particularly aggressive root systems that spread distances from a single plant and release chemicals to harm the seeds of other plants (Ens et al. 2008), whereas some other invasive species may grow so densely that they smother the species around them (Burke et al. 2011).

Minimal studies have attempted to precisely model and predict the growth or spread of invasive plants, specifically in regards to using methods such as Machine Learning (Schneider et al. 2021). Those that do often only use Machine Learning for the sole purpose of classifying potential invasive plants through aerial imagery (Baron et al. 2018). The vast majority of studies concerning invasive plants focus on one specific variable or location, such as attempting to predict events or scenarios that may come as a result of invasive plants (Faccenda et al. 2021), determining the relationship between temperature on growth of invasive plants (Skálová et al. 2015), or predicting the distribution of invasive plants in either a highly general or specific area (Ahmed et al. 2020, and Jones et al. 2010). As a result, this paper seeks to create a method capable of predicting the spread of invasive plants to a highly precise degree.

Similarly, there have also been minimal studies attempting to understand and model the invasion curves of invasive species, specifically those of individual species. These invasion curves are vital to managing the spread of current invasive plants, as they allow for greater understanding of how the plant spreads at each stage, specifically during the so-called lag phase or beginning phase. Studies that attempt to model lag phases and invasion curves are often created manually through records of herbarium data (Antunes and Schamp 2017, and Larkin 2011), which is time consuming and inefficient.

Due to the vast amount of damage caused by late stage growth of invasive plants, a tremendous amount of resources have been put into understanding and preventing the spread of invasive plants at later stages. However, little work has been done on early detection, prediction, and eradication of invasive plants, which this study hopes to change.

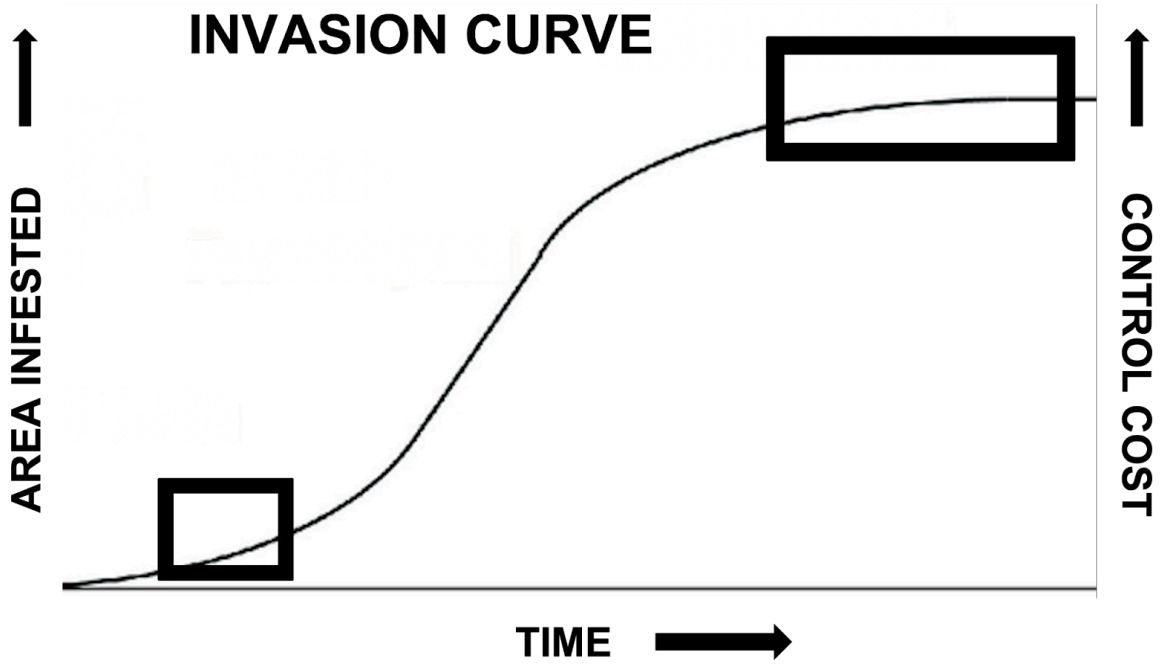


Figure 1: Invasive Species Invasion Curve

3. Methodology

a. Tools

The construction and training of the models in this study was implemented using R and Python, specifically the Keras, Numpy, and Tensorflow Computational Libraries. The method was run on an NVIDIA Tesla K30 GPU hosted by Google Cloud Collaboratory for optimal runtime.

b. Project Research

In this study, Random Forest, Multivariate Multiple Regression, Support Vector Machines, and Neuroevolution through neural networks were used to create and refine models capable of predicting the spread of invasive plants. The programming language Python was used for machine learning computing and graphics. In particular, the packages Pandas and Tensorflow were used.

Two dependent variables were used to determine the plant spread, being the X and Y coordinate locations a plant was most likely to spread to. Numerous independent variables were used, such as a categorical variable indicating a total of 75 different types of species of invasive plants, or the length in time from when each plant was first discovered to the day the current plant was cataloged, among others, and variables relating to the environmental conditions and factors were included, in order to best optimize the predictive method.

Thus, the hypothesis was:

Is it possible to create an accurate, early stage, predictive method for determining the spread of invasive plants by utilizing Machine Learning techniques based on historical data correlating with environmental variables? And furthermore, is this method applicable to multiple geo locations and situations?

c. Data Collection

The Centre for Agriculture and Biosciences International (CABI) Africa Invasive and Alien Species dataset was used, which holds over 77,00 records of over 70 invasive alien species from Ethiopia to Uganda. The data were originally collected by a group of

researchers from the University of Chicago through roadside surveys over the course of fieldwork trips over a 43-year period from 1974 to 2017, leading to a total of 77,075 observations of invasive plants which they wrote a study about (Witt et al. 2018).

The second dataset used was collected by the Center for Invasive Species and Ecosystem Health at the University of Georgia. The dataset consists of over 99,000 records of more than 120 invasive alien species, collected from 2008 to 2021, specifically from the state of California, US. To collect their data, the Center for Invasive Species and Ecosystem Health used a combination of field surveys and GIS mapping to determine the locations and density of invasive plant species. Field surveys involved manually collecting data on the various characteristics of invasive plants, such as species presence, abundance, and distribution, from a variety of locations. GIS mapping complemented the previous method by also utilizing information such as land cover type and soil type to best record density and location.

Additionally, numerous records of environmental data and factors were collected from Visual Crossing, a cloud-based data-visualization platform that collects and stores environmental data from a variety of sources. The data collected provides insight into environmental trends and variables such as temperature, humidity, and precipitation to name a few, and would help optimize the predictive models.

Finally, the Biodiversity Intactness satellite biodiversity data from Impact Observatory. The data is a global dataset used to calculate, map, and monitor vegetation and biodiversity coverage on land surfaces. They can be used to assess the impact of climate change, agricultural practices, and land use on land and forest cover change.

Most of the datasets were sorted in different ways. The first two datasets were initially sorted into subsets based upon the species of plant, to group plants of the same species together. Then, the subsets were further sorted by the date the plant was first discovered. The Visual Crossing environmental data was sorted to overlay with the pre-existing plant data, with the Biodiversity Intactness data providing validation and accuracy.

d. Data Preparation

To ensure that the data meet the minimum assumptions of parametric statistics, several tests were performed. First, the histograms (Figures 2 and 3) of variable "time" and "spread" show that they are reasonably normally distributed. Skewness statistic was found to be between -2 and +2 and Kurtosis statistic was between -7 and +7, which suggest that the data do not deviate significantly from normal distribution. A multicollinearity issue was inspected using Tolerance and VIF (variance inflation factor), and scores of both showed that multicollinearity was not a significant issue in the analysis. Furthermore, the data was initially split into 70% training, 20% testing, and 10% validation chronologically in order to determine the accuracy of the method.

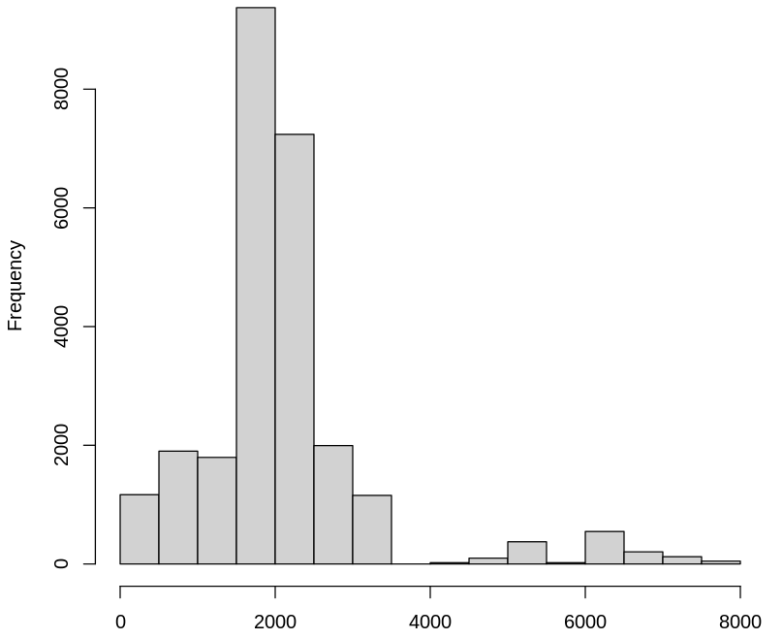


Figure 2: Histogram of Time

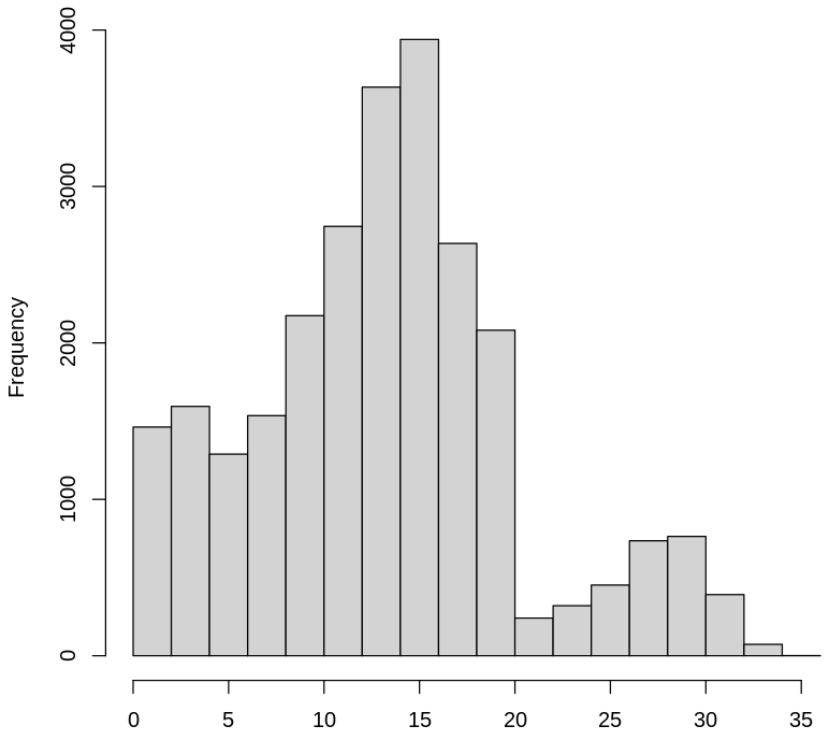


Figure 3: Histogram of Distance Spread

4. Methods

In this study, multiple data sources were used to train and test the models, described below. The utilization of various data sources and ML techniques allows for a thorough understanding of the current ecology of invasives and prediction of potential areas in which they may become established.

a. Random Forest (RF)

Random Forest is a supervised machine learning algorithm that relies on combining the predictions of multiple decision trees to arrive at a more accurate model.

Particularly useful when dealing with larger datasets, Random Forest utilizes a bootstrap aggregation called the bagging technique. After splitting the existing data into multiple random subsets, each subset is used to train individual decision tree models which are then used to make predictions. Random Forest then takes the average of the predictions from all of the decision trees to make a final model that is more accurate than any single decision tree model.T his allows the model to reduce variance and prevent overfitting, especially when faced with noisy data.

Due to many potential characteristics impacting the spread of invasive plants, along with the massive size of the two datasets, Random Forest was chosen as a preliminary predictive model. Before the two more complex models, explained below, were used.

b. Multivariate Multiple Linear Regression

Despite being based on one of the simpler models, Multivariate Multiple Linear Regression is no less effective, with its general formula shown below. While Linear Regression seeks to establish relationships between an independent and dependent variable, Multivariate Multiple Linear Regression takes this a step further by using n independent variables, which are X_1 to X_n, in order determine a one or more dependent variables, which is represented by $\widehat{Y}_1$ to $\widehat{Y}_n$, shown in the equation below.

$$\widehat{Y}_0 + \widehat{Y}_1 + \ldots + \widehat{Y}_n = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

From all of the independent variables, the Multivariate Multiple Linear Regression determines a line of best fit which is used to determine the single dependent variable.

As the data utilizes multiple instances of independent variables and the goal is to determine two dependent variables, it makes sense to use Multivariate Multiple Linear Regression as a basis for the predictions. Potential benefits of Multivariate Multiple Linear Regression include that it is easily able to identify correlation between the independent variables and the dependent variables, and is able to identify and ignore major outliers which may have skewed the data. Thus, a Multivariate Multiple Linear Regression Model was fitted onto the dataset. The MMLR model works with initial locations such as longitudes and latitudes along with other factors such as the time elapsed or type of plant species in order to determine possible resulting locations.

c. Support Vector Machines (SVM)

Support Vector Machines (SVM) are models for classification and prediction that are able to solve both linear and nonlinear tasks. SVM usually begins by plotting all data points on an N-dimensional space. Afterwards, the model utilizes hyperplanes (equation shown below), which in N-dimensional space acts as N-1 dimension subsets of that space that work to divide said space into two separate parts, in order to separate the data points into various groups, categorizing them.

$$x \cdot \overline{w} + b = 0 \text{ (Equation for Hyperplane)}$$

When attempting to optimize hyperplanes in order to better categorize data, we can create margin lines to separate the different categories of data.

When attempting to split two categories, you can use the following equations to determine the optimal hyperplane.

$$x \cdot \overline{w} + b \leq -1 \text{ (Category 1)}$$

$$x \cdot \overline{w} + b \geq -1 \text{ (Category 2)}$$

We can further measure the distance between the two categories, using the following equation.

$$(x_2 - x_1) \cdot \frac{\overline{w}}{||w||} \text{ (Equation for distance between categories)}$$

$$\Rightarrow \frac{x_2 \cdot \overline{w} - x_1 \cdot \overline{w}}{||w||},$$

Since we know that $x_1\overline{w} = 1 - b$ and $x_2\overline{w} = -1 - b$, then the previous equation becomes

$$\frac{(1-b)+(b+1)}{||w||} = \frac{2}{||w||} = d$$

When attempting to include margin lines, we attempt to maximize the margin between potential points and have the Largest Margin, M, between the two categories. Thus, we can use the equation below to maximize the Margins, as a hyperspace is maximized when the values of the margin are maximized

$$y(x \cdot \overline{w} + b) \geq M \text{ (Equation for Margin of Hyperplanes)}$$

Support Vector Machines work well when there are large amounts of variables, sometimes more so than data values. Additionally, the model type often works well in cases where the data points result in larger margins. So, in order to further increase the accuracy of the prediction model, a Support Vector Machine Model was fitted as the third model, as it applies classification algorithms on labeled training data to categorize new data. The SVM utilizes hyperplanes and divides the data into separate categories in order to better spread according to characteristics such as the type of plant or length traveled.

d. Neuroevolution

Neuroevolution is a method of machine learning which relies on using principles of evolution to create and optimize neural networks. Particularly useful when dealing with complex datasets, neuroevolution begins with the creation of multiple neural networks, one-hundred in this study, with unique architectures and randomized weighting. Afterwards, the performance of each neural network is measured and ranked, with the top 10% of networks advancing to the next generation. Rank-Based Selection is then used to select parent networks, with those of a higher rank having a higher probability of being chosen. Parent networks are paired and undergo crossover, meaning parts of the parent's architectures or weights are taken and combined to form a new neural network with new architectures and weights. Afterwards, the new networks are added to the next generation of the process, which repeats itself until a definitive final neural network emerges.

Due to the many potential environments and conditions impacting the spread of invasive plants, neuroevolution was chosen as a method of creating a final predictive network through its ability to work with complex and noisy data.

5. Results

The various models and methods were used to determine potential locations of each invasive species as a result of factors such as initial longitude and latitude of the plants and the time elapsed. Using a data split of 70% training, 20% testing, and 10% validation, the models waweres constantly improved and refined.

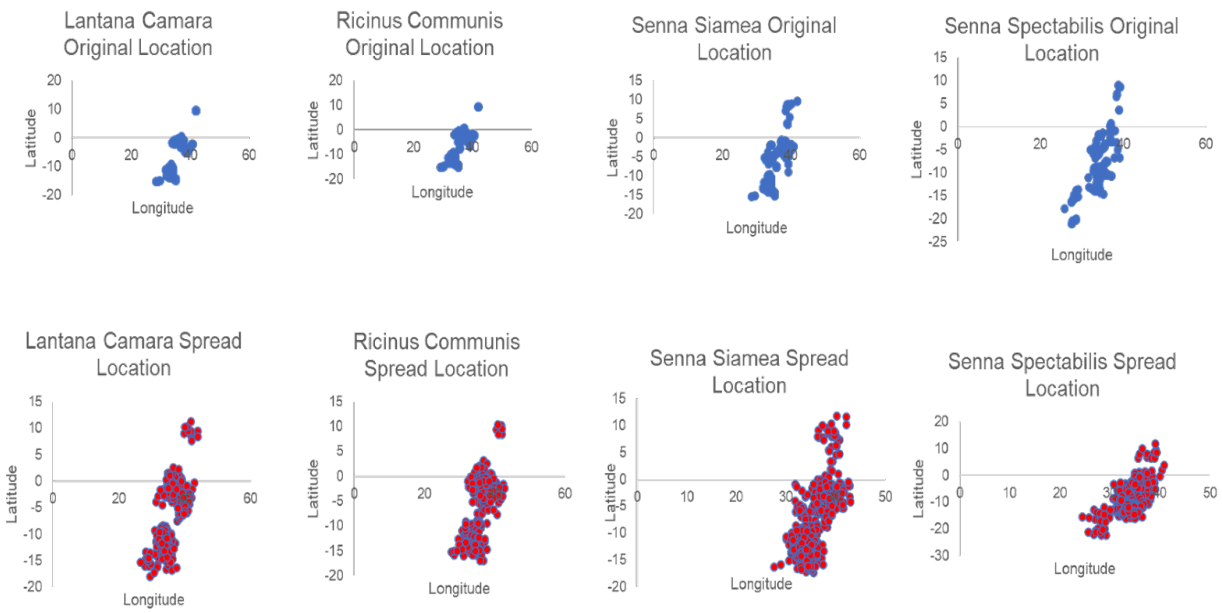
Accuracy of Models				
	Random Forest	Multivariate Multiple Regression	Support Vector Machine	Neuroevolution
Africa	65.2%	82.1%	97.3%	98.8%
US California	60.9%	91.5%	89.7%	95.74%

Table 1: Accuracy of Resulting Method

Using data from the Africa Invasive and Alien Species dataset along with the Invasive Species and Ecosystem Health dataset from the University of Georgia, several predictive machine learning models were created and ran. As shown in the table above, each of their accuracies are displayed.

When compared against the test data, the Neural Network fared remarkably well, with a 98.8% accuracy on the Africa dataset and a 95.74% accuracy on the California dataset.

Below, the graphs of the original and predicted locations of the most invasive plant species are shown. Specifically, Graph 1 shows the original and predicted locations of the invasive plants from the South Africa dataset, with the most invasive species being Lantana Camara, Ricinus Communis, Senna Siamea, Senna Spectabilis, and Tithonia Diversifolia.

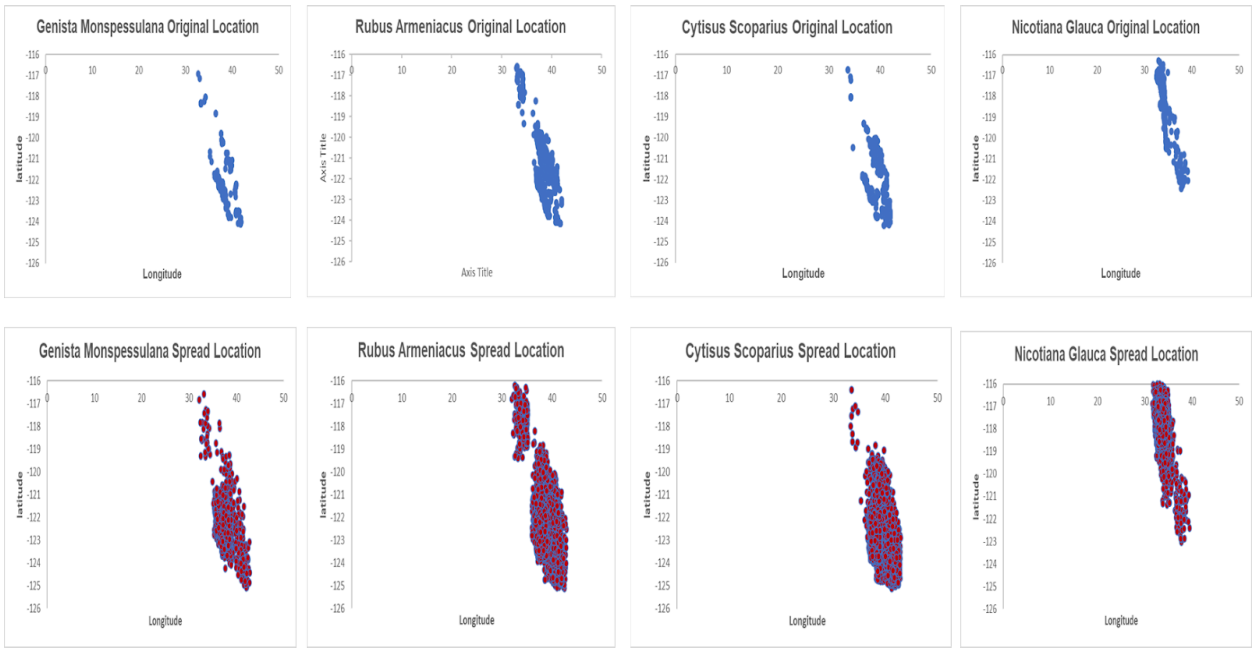


Graph 1: Predicted Location

The graphs above show the original and predicted locations of five most invasive plants from the Africa dataset. The graphical representation shows a medium to high amount of change over the time, with the plants spreading $0.0003\ km$ to $0.06\ km$ in a day. The original graphs oversee the time period from January 14, 2014 to April 5th, 2017, while the predictive graphs predict the spread of invasive plants from April 5th, 2017 to January 8th, 2020.

Below, the graphs of the original and predicted locations of the most invasive plants of the California dataset are shown. Specifically, Graph 2 shows the original and

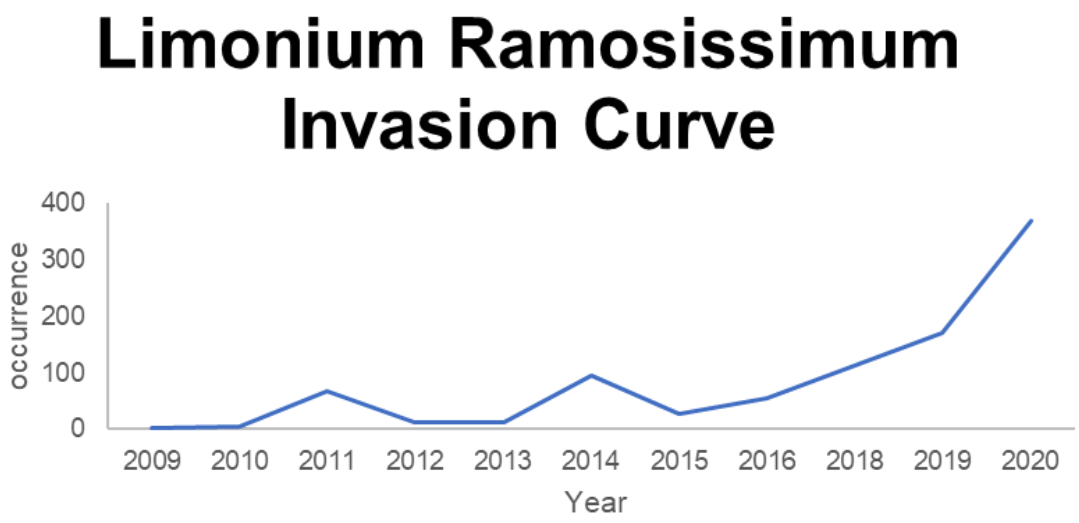
predicted location of the plants Genista Monspessulana, Rubus Armeniacus, and Cytisus Scoparius, Nicotiana Glauca.



Graph 2: Predicted Location

The graphs above show the original and predicted locations of five most invasive plants from the California dataset. The graphical representation shows a medium amount of change over the time, with the plants spreading $0.0002\ km$ to $0.05\ km$ in a day. The original graphs oversee the time period from February 12, 2008 to December

18th, 2021, while the predictive graphs predict the spread of invasive plants from December 18th, 2021 to May 21st, 2022.



Graph 3: Invasion Curve

While invasion curves at the lag phase are often poorly understood, our predictive models were able to accurately model said curves to a higher degree than ever before. These curves, especially at an early stage (also called the lag phase), are one of the essential components of effective invasive plant management. The application of these curves is two-fold, not only they are able to model the spread of newly arrived invasive plants, but also they can be used to model the spread of plants that have previously been treated. As shown in Graph 3 above, these invasion curves allow us to

understand the ways and rate at which invasive plants spread, highly vital for future modeling.

6. Conclusion

Analysis of the data collected from the CABI Africa Invasive and Alien Species dataset resulted in the creation of numerous models capable of predicting possible future locations of invasive plants. Using data such as the initial location of plants along with other factors such as time elapsed and the species, the potential longitudes and latitudes of spread and location of invasive plants were determined, allowing their spread to be identified. The prediction-based approach was able to directly pinpoint species that have yet to invade, and areas not yet invaded but have a high likelihood of being invaded in the future. This method, applicable to any other taxa, represents an unparalleled opportunity to implement timely and proactive management strategies against biological invasions.

Furthermore, this study presents a machine learning-based approach to understanding invasion growth. This would allow a better understanding of the growth spread dynamics of newly detected as well as newly treated invasive plants. With numerous applications, from greater management of invasive plants to modeling their growth, the invasion curve widely expands the applications of this study.

Due to resource limitations, specifically in terms of computing power, this study did not fully incorporate the full range of satellite and environmental data, which will be explored and refined in further methods.

The study contributes to the Early Detection and Eradication strategies of invasive species by providing a machine learning based method to accurately predict the distribution of invasive species at an early stage. Specifically, the method created the ability to predict potential locations of invasive plants with extreme accuracy and reliability, with the neuro-evolved neural network being able to predict the spread of invasive plants in Africa with a 98.8% accuracy. The method is further able to determine the distribution patterns and potential locations of invasive plants depending on the species and geographic location, with the accuracy for spread and location of invasive species in California having an accuracy of over 95.78%. Most significantly, this method is both scalable and applicable in a multitude of different areas, from diseases to even people, allowing for a novel approach to detecting and predicting any types of spread before they become a significant issue.

7. References

Ahmed, N., Atzberger, C., & Zewdie, W., Integration of remote sensing and bioclimatic data for

prediction of invasive species distribution in data-poor regions: a review on challenges and opportunities. Environmental Systems Research, 9(1),2020. https://doi.org/10.1186/s40068-020-00195-0

Antunes, Pedro M, and Brandon Schamp. "Constructing Standard Invasion Curves From

Herbarium Data – Toward Increased Predictability of Plant Invasions." *Cambridge Core*, Cambridge University Press, 28 Dec. 2017, https://www.cambridge.org/core/journals/invasive-plant-science-and-management/article/constructing-standard-invasion-curves-from-herbarium-datatoward-increased-predictability-of-plant-invasions/342A9CEFF2EC149427B9D6D0FD420F32.

Baron, Jackson, et al. "Combining Image Processing and Machine Learning to Identify Invasive

Plants in High-Resolution Images." *Taylor & Francis*, 3 Jan. 2018, https://www.tandfonline.com/doi/abs/10.1080/01431161.2017.1420940.

Burke, J. M., & DiTommaso, A., Corallita (Antigonon leptopus): Intentional Introduction of a

Plant with Documented Invasive Capability. Invasive Plant Science and Management, 4(3), 265–273, 2011. https://doi.org/10.1614/ipsm-d-10-00088.1

Coutts, S. R., van Klinken, R. D., Yokomizo, H., & Buckley, Y. M., What are the key drivers of

spread in invasive plants: dispersal, demography or landscape: and how can we use this knowledge to aid management? Biological Invasions, 13(7), 1649–1661,2010. https://doi.org/10.1007/s10530-010-9922-5.

Ebeling, S. K., Welk, E., Auge, H., & Bruelheide, H., Predicting the spread of an invasive plant: combining experiments and ecological niche model. Ecography, 31(6), 709–719,2008. https://doi.org/10.1111/j.1600- 0587.2008.05470.x

Ens, E. J., Bremner, J. B., French, K., & Korth, J., Identification of volatile compounds released

by roots of an invasive plant, bitou bush (Chrysanthemoides monilifera spp. rotundata), and their inhibition of native seedling growth. Biological Invasions, 11(2), 275–287, 2008. https://doi.org/10.1007/s10530-008-9232-3

Faccenda, K., & Daehler, C. C., A screening system to predict wildfire risk of invasive plants.

Biological Invasions, 24(2), 575–589, 2021. https://doi.org/10.1007/s10530-021-02661-x

Gosper, C. R., Stansbury, C. D., & Vivian-Smith, G., Seed dispersal of fleshy-fruited invasive

plants by birds: contributing factors and management options. Diversity Distributions, 11(6), 549–558, 2005. https://doi.org/10.1111/j.1366-9516.2005.00195.x

Hastings, A., Cuddington, K., Davies, K. F., Dugaw, C. J., Elmendorf, S., Freestone, A., Harrison, S., Holland, M., Lambrinos, J., Malvadkar, U., Melbourne, B. A., Moore, K., Taylor, C., & Thomson, D., The spatial spread of invasions: new developments in theory and evidence. Ecology Letters, 8(1), 91–101, 2004. https://doi.org/10.1111/j.1461-0248.2004.00687.x

Jones, C. C., Acker, S. A., & Halpern, C. B., Combining local- and large-scale models to predict

the distributions of invasive plant species. Ecological Applications, 20(2), 311–326, 2010. https://doi.org/10.1890/08-2261.1

Larkin, Daniel J. "Lengths and Correlates of Lag Phases in Upper-Midwest Plant Invasions -

Biological Invasions." *SpringerLink*, Springer Netherlands, 15 Oct. 2011, https://link.springer.com/article/10.1007/s10530-011-0119-3.

Maringer, J., Wohlgemuth, T., Neff, C., Pezzatti, G. B., & Conedera, M., Post-fire spread of alien

plant species in a mixed broad-leaved forest of the Insubric region. Flora - Morphology, Distribution, Functional Ecology of Plants, 207(1), 19–29, 2012. https://doi.org/10.1016/j.flora.2011.07.016

MORISON, J. I. L., & LAWLOR, D. W., Interactions between increasing CO2 concentration and

temperature on plant growth. Plant, Cell and Environment, 22(6), 659–682, 1999. https://doi.org/10.1046/j.1365- 3040.1999.00443.x

Obiri, J. F., Invasive plant species and their disaster-effects in dry tropical forests and rangelands

of Kenya and Tanzania. Jàmbá: Journal of Disaster Risk Studies, 3(2), 2011. https://doi.org/10.4102/jamba.v3i2.39

Schneider, K., Makowski, D., & van der Werf, W. (2021). Predicting hotspots for invasive

species introduction in Europe. Environmental Research Letters, 16(11), 2021. 114026. https://doi.org/10.1088/1748-9326/ac2f19

Schupp, E. W., Quantity, quality and the effectiveness of seed dispersal by animals. Vegetation,

107–108(1), 15–29, 1993. https://doi.org/10.1007/bf00052209

Shackleton, C. M., McGarry, D., Fourie, S., Gambiza, J., Shackleton, S. E., & Fabricius, C.

(2006). Assessing the Effects of Invasive Alien Species on Rural Livelihoods: Case Examples and a Framework from South Africa. Human Ecology, 35(1), 113–127. https://doi.org/10.1007/s10745-006-9095-0

Skálová, H., Moravcová, L., Dixon, A. F. G., Kindlmann, P., & Pyšek, P. (2015). Effect of temperature and nutrients on the growth and development of seedlings of an invasive plant. AoB PLANTS, 7https://doi.org/10.1093/aobpla/plv044

Sultan, S. E., Horgan-Kobelski, T., Nichols, L. M., Riggs, C. E., & Waples, R. K., A resurrection

study reveals rapid adaptive evolution within populations of an invasive plant. Evolutionary Applications, 6(2), 266–278, 2012. https://doi.org/10.1111/j.1752-4571.2012.00287.x

Witt, A. B. R. (2010). Biofuels and invasive species from an African perspective - a review. GCB

Bioenergy, 2(6), 321–329. https://doi.org/10.1111/j.1757-1707.2010.01063.x

Witt, A., Beale, T., & van Wilgen, B. W. (2018). An assessment of the distribution and potential

ecological impacts of invasive alien plant species in eastern Africa. Transactions of the Royal Society of South Africa, 73(3), 217–236,2018. https://doi.org/10.1080/0035919x.2018.1529003

Proof that the Speed of Light is Constant

Nicholas Lu '25
Applied Mathematics

To demonstrate that the speed of light is constant, we can use two of Maxwell's equations in vacuum, where there are no charges (ρ = 0) and no currents (J =0). The two relevant Maxwell's equations are Faraday's law of induction and Ampere's law with Maxwell's addition. In a vacuum, these equations can be written as:

1. Faraday's Law of Induction:

$$\nabla \times E = -\frac{\partial B}{\partial t}$$

2. Ampere's Law (with Maxwell's addition):

$$\nabla \times B = \mu_0 \epsilon_0 \frac{\partial E}{\partial t}$$

where:

- E is the electric field

- B is the magnetic field

- μ_0 is the permeability of free space ($4\pi \cdot 10^{-7}\ N/A^2$)

- ϵ_0 is the permittivity of free space

- $\nabla \times$ denotes the curl of a vector field, and

- t represents time.

Steps to Derive the Speed of Light

Step 1: Take the curl of both sides of Faraday's law

$$\nabla \times (\nabla \times E) = \nabla \times (-\frac{\partial B}{\partial t})$$

Step 2: Use a vector calculus identity

The vector calculus identity for the curl of the curl of a vector field A is:

$$\nabla \times (\nabla \times A) = \nabla(\nabla \cdot A) - \nabla^2 A$$

Since in vacuum, $\nabla \cdot E = 0$ and $\nabla \cdot B = 0$, this simplifies to:

$$\nabla \times (\nabla \times E) = -\nabla^2 E$$

Step 3: Apply Ampere's Law to Faraday's Law

By substituting Ampere's Law into the derivative of Faraday's Law:

$$-\nabla^2 E = \nabla \times (-\frac{\partial B}{\partial t})$$

$$-\nabla^2 E = \mu_0 \epsilon_0 \frac{\partial^2 E}{\partial t^2}$$

This equation describes how electromagnetic waves propagate through a vacuum.

Step 5: Compare to the standard wave equation

The standard form of the wave equation for a wave propagating with speed v is:

$$\nabla^2 E = \frac{1}{v^2} \frac{\partial^2 E}{\partial t^2}$$

Step 6: Derive the speed of light

By comparing coefficients, we see that:

$$\frac{1}{v^2} = \mu_0 \epsilon_0$$

Solving for v, the speed of electromagnetic waves (including light) in a vacuum, we find:

$$v = \frac{1}{\sqrt{\mu_0 \epsilon_0}}$$

Step 7: Calculate the speed of light

Using the values of μ_0 and ε_0, we can calculate the speed of light:

$$c = \frac{1}{\sqrt{(4\pi \cdot 10^{-7})(8.854 \cdot 10^{-12})}}$$

Stefan-Boltzmann Derivation

Nicholas Lu '25

Applied Mathematics

Planck's law for black-body radiation gives the energy radiated per unit area of the emitting body per unit time per unit solid angle per unit frequency, which is defined as:

$$B(v,T) = \frac{2hv^3}{c^2}\frac{1}{e^{\frac{hv}{kT}}-1}$$

where:

- B(ν, T) is the spectral radiance,
- h is Planck's constant ($6.62607015 \times 10{-}34\ m^2kg/s$),
- ν is the frequency of the electromagnetic radiation,
- c is the speed of light in a vacuum (3.00 × 108 m/s),
- k is the Boltzmann constant (1.380649 × 10−23 J/K),
- T is the absolute temperature of the black body in Kelvin.

Our goal is to integrate Planck's law over all frequencies to find the total power emitted per unit area, which leads to the Stefan-Boltzmann law. To find the total emitted power per unit area (j*) we integrate B(ν, T) over all frequencies from 0 to infinity:

$$j^* = \int_0^\infty B(v,T)dv$$

Substituting B(ν, T) into the integral:

$$j^* = \int_0^\infty \frac{2h\nu^3}{c^2} \frac{1}{e^{\frac{h\nu}{kT}} - 1} d\nu$$

To simplify the integral, we perform a change of variables. Let:

$$x = \frac{h\nu}{kT}$$

Then, differentiate both sides with respect to ν to find dν:

$$dx = \frac{h}{kT} d\nu$$

$$d\nu = \frac{kT}{h} dx$$

Substituting x and dν into the integral:

$$j^* = \int_0^\infty \frac{2h}{c^2} \left(\frac{kT}{h}\right)^4 x^3 \frac{1}{e^x - 1} \frac{kT}{h} d\nu$$

The integral simplifies to:

$$j^* = \frac{2k^4T^4}{c^2h^3} \int_0^\infty \frac{x^3}{e^x - 1} dx$$

Which simplifies to $\frac{\pi^4}{15}$.

Now,

$$j^* = \frac{2k^4T^4}{c^2h^3} \frac{\pi^4}{15}$$

$$j^* = \frac{2\pi^4k^4}{15c^2h^3} T^4$$

The Stefan-Boltzmann constant σ is thus identified as:

$$\sigma = \frac{2\pi^4k^4}{15c^2h^3}$$

Tachyons

Ayush Varma '27

Physics Possibilities

Is light the fastest velocity an object can travel?

The speed of light, clocking in at about 300 million meters/second, is supposedly the fastest velocity possible, or is it? Tachyons are hypothetical particles that can travel faster than the speed of light. A scientist named Gerald Feinberg has a theory that tachyons arise from a quantum field with imaginary mass. No one has ever proved that tachyons actually exist though. Tachyons disobey the laws of physics, but is that possible? If so, how do tachyons accomplish this feat, and how fast can they actually go?

When the Big Bang occurred, it made things with mass and things with no mass such as tachyons. Tachyons were created moving at the speed of light. Because tachyons have no mass, they are not able to slow down, which allows tachyons to go faster than the speed (of light) (Kelley, n.d.). However, tachyons get faster and gain more velocity only as the energy decreases ("Tachyon" 2023).

According to a new theory, Tachyons move so fast that they can travel in time. This has made more people question their existence, as to current knowledge, nothing has ever gone back in time. Tachyons can travel back in time if they travel faster than the speed of light. This is due to Einstein's theory of relativity which states that time and space are curved ("Tachyon" 2023). This can lead to tachyons being on 2 points at the same time. It is said that with a tachyon transmitter, messages could be sent back in time. However, this ability has yet to be proven or disproven as no evidence of tachyons has been found.

Einstein's first postulate on special relativity (not general relativity) says the rules of physics apply to all objects, no matter the time, and frames of reference. E=MC Squared is the equation, but it cannot be used because there is no mass on tachyons ("Tachyon" 2023). A tachyon can not get to infinite speed, because while there, its energy approaches zero. It would phase into its antiparticle when it hits a photon near infinite speed. It gets closer to infinite speed which is also pure energy because when it's infinite speed, it has no energy so it is not a thing. (Kelley, n.d.)

x2 – c2t2=0 and x=ct are the base equations of a postulate on a reference frame. With math and rephrasing of the equation, this equation is created: x'=±(x-vt)/(v2/c2-1)½. But for that equation, the velocity has to be faster than the speed of light at all times because when there's a number over a bigger number, the answer is less than one ("Tachyon" 2023). However, because one is subtracted, it would be negative, and the square root of a negative number does not exist.

A photon carries small electromagnetic fields with charge. Therefore, they interact with tachyons because they also have charge. If tachyons approach quantum vacuum, antimatter phases into existence, closer to pure energy, where they collide, at nothing but infinite speed (I think) (Wikipedia Contributors 2020). It would require energy to slow down, but the energy is minimal due to speed they travel at, so their only option is to lower their energy and travel faster (Wikipedia Contributors 2020).

This relates to velocity and acceleration. It relates to these concepts because it covers an object that can travel faster than the speed of light and has trouble decelerating to a speed below. Tachyons are also traveling faster than the speed of light the moment they are created.

Bibliography

(Kelley, n.d.) - Kelley, Leonard. n.d. "What Are Tachyons?" Owlcation. https://owlcation.com/stem/What-Are-Tachyons.

("Tachyon" 2023)- "Tachyon." 2023. Wikipedia. September 11, 2023. https://en.wikipedia.org/wiki/Tachyon#:~:text=Tachyons%20would%20exhibit%20the%20unusual.

(Wikipedia Contributors 2020)- Wikipedia Contributors. 2020. "Antimatter." Wikipedia. March 29, 2020. https://en.wikipedia.org/wiki/Antimatter.

Special recognition to Form III student Grant Itin who wrote 4 sentences about the recent news of Tachyons.

Why Would the Universe End?

Ayush Varma '27

Physics Possibilities

Disclaimer: These are all theories without hard-proof evidence.

Firstly, to know why the universe will end, which it will because nothing lasts forever, we have to look back in time to when the universe started.

The most popular theory for the beginning of the universe is that the universe undertook a massive expansion, "doubling in size at least 80 times in a fraction of a second" ("What Came before the Big Bang? UB Physicist's New Popular Science Book Explains One Leading Theory," n.d.). Many believe that the universe was an infinitely dense and extremely small ball of matter. Then, 4 fundamental forces came together inside this ball of matter and created a bang, leading to expansion. This expansion of the universe continues today. The growth of the universe still occurring hints at the idea that time is infinite in the future. But was it infinite in the past? Before the Big Bang, was there any space at all? Was there any time? If nothing moves at all, did time pass or did it freeze? According to Newton's Third Law, everything has an equal and opposite reaction; could the universe implode at any moment?

This now brings me to my main point, could the universe go back to the tiny dense particle at any time?

Considering Newton's Third Law, everything has an equal and opposite reaction. The universe could go back into the singularity (where the laws of physics break down) that it once was. This idea is called the "Big Crunch," but it is very unlikely to matter though. "The Big

Crunch scenario hypothesized that the density of matter throughout the universe is sufficiently high that gravitational attraction will overcome the expansion which began with the Big Bang" ("Big Crunch" 2023). One day, if the universe stops expanding, it will start moving again. However, this time it will go inward into a super dense and hot singularity, like how it started. This will be the opposite reaction of the Big Bang, but is not likely to occur (Betz 2021). Instead, a plausible ending of the universe could be the big freeze. The big freeze is possible if the universe keeps expanding, causing the energy across the universe to spread out and leading to absolute zero temperature.

The last possible way the universe ends is the big rip. The big rip is possible if the universe keeps expanding, and everything becomes separated and atoms are destroyed (very unlikely). The atoms will be destroyed by electrons flying off atoms and with the nuclei separating. This reaction would cause atomic bombs to go off and would undoubtedly result in the end of the universe ("How the Big Crunch Theory Works" 2009).

Bibliography

April 2020, Paul Sutter 26. 2022. "What Happened before the Big Bang?" Space.com. February 12, 2022. https://www.space.com/what-came-before-big-bang.html.

Betz, Eric. 2021. "The Beginning to the End of the Universe: The Big Crunch vs. the Big Freeze." Astronomy Magazine. January 31, 2021. https://www.astronomy.com/science/the-beginning-to-the-end-of-the-universe-the-big-crunch-vs-the-big-freeze/.

"Big Crunch." 2023. Wikipedia. November 26, 2023. https://en.wikipedia.org/wiki/Big_Crunch#:~:text=The%20Big%20Crunch%20scenario%20hypothesized.

"How the Big Crunch Theory Works." 2009. HowStuffWorks. March 2, 2009. https://science.howstuffworks.com/dictionary/astronomy-terms/big-crunch.htm#:~:text=What%20is%20the%20theory%20of.

Lewis, Ralph. 2020. "How Do We Know What Is Real? | Psychology Today." Www.psychologytoday.com. August 6, 2020. https://www.psychologytoday.com/us/blog/finding-purpose/202008/how-do-we-know-what-is-real.

"What Came Before the Big Bang? UB Physicist's New Popular Science Book Explains One Leading Theory." n.d. Www.buffalo.edu. https://www.buffalo.edu/news/releases/2022/03/033.html#:~:text=Prior%20to%20the%20Big%20Bang.

"Would Time Exist If There Was Nothing?" 2023. The Philosophy Forum. September 29, 2023. https://thephilosophyforum.com/discussion/14683/would-time-exist-if-there-was-nothing#:~:text=Time%20is%20a%20concept%2C%20a.

Solar Sails

Ayush Varma '27

Physics Possibilities

When you think of sailing, you think of boats in the ocean moving by the wind. You would've never thought you could sail through space until you met solar sails. Solar sails are new cutting-edge technology that propel a spaceship to space. It works by capturing the momentum of the light from the sun to the earth and using that to push the ship out. It could also be used to push humans, which would revolutionize space travel.

Solar sails function by light hitting solar sails in space, then the photons hit the sail, and photons' momentum is transferred to the sail, giving it a push. The solar sail is made of mirror-like surfaces, and the photons bounce off the sail after transferring its momentum.

Ships can be attached to the huge sail to get to space. Some pros are that solar sails do not need fuel, making them very efficient. They are also very light but big, so they can carry big things such as ships. They are also very quiet, so if a stealth mission is desired, they would be an adequate choice. They also have a long lifeline and can be used in deep space with a low cost.

Some cons to these sails are that the initial acceleration is slow compared to fuel spaceships. They are also not good for missions that need quick departures, because there are not a lot of photons hitting the sail near the Earth. They also need the sun to fuel them which is unreliable due to solar eclipses, bad weather, and other phenomenons. It is also incredibly hard to keep them pointing in the right direction because they are so big. Lastly, any damage the sails succumb to cannot be fixed unless people fly up to the sail which in itself is a lot of work. Damages affect the sail's direction a lot, which is another reason why it is hard to control.

In 1608 Johannes Kepler shared his idea for solar sails with his friend, Galileo Galilei. Then, in 1873, James Clerk Maxwell first demonstrated that sunlight exerts a small amount of

pressure as photons bounce off a reflective surface. This led to Echo-1 in 1960. Echo-1 felt these solar pressure effects loudly and clearly. "Photon pressure played orbital soccer with the Echo-1 thin-film balloon in orbit.... The shards were flung far and wide by sunlight" (Coulter). In 1974, NASA had a good experience with solar sailing when their rocket was running low on altitude control gas. In the 1970s Dr. Louis Friedman, then at NASA's Jet Propulsion Laboratory, led a project to try the first solar sail flight. In 1993, the Russian Space Agency launched a 20-meter diameter spinning mirror called Znamya 2, hoping to beam solar power back to the ground. In the Early 2000s, the Planetary Society helped with the construction of the Cosmos 1 spacecraft and utilized solar sail technology. This ended with the Japanese deploying solar sail materials sub-orbitally from a sounding rocket in 2004. Although it was not a demonstration of a free-flying solar sail that could be used for deep-space exploration, the deployment was nevertheless "a valuable milestone".

As stated by NASA, in April, a next-generation solar sail technology – known as the Advanced Composite Solar Sail System – will launch aboard Rocket Lab’s Electron rocket from the company’s Launch Complex 1 in Māhia, New Zealand. The technology could advance future space travel and expand our understanding of our sun and solar system. Solar Sails will be crucial in guiding interplanetary travel. In 2019, The LightSail 2 completed its first flight aided by only photons (NASA Next-Generation Solar Sail Boom Technology Ready for Launch - NASA).

The concept of these sails include topics many Haverford students are learning about right now. They include:

Photons: The photons power the sail by giving their momentum to it.

Momentum: The photons pass their momentum to the sail, propelling it.

Mirror-like surfaces: A large ultra-thin mirror is used in the sail so the mirror reflects the photons and takes in their momentum.

Acceleration: The sail accelerates very slowly, to begin with due to the lack of photons hitting it but accelerates when it gets closer to the sun where there are more photons.

Force: The photon interacts with electrons by imparting momentum onto it. The electron then breaks away from the center of its nucleus and moves in the direction the photon is when it hits the electron. That electron does the same action to its neighbor atoms, pushing the sail.

Bibliography

Coulter, Dauna. "A Brief History of Solar Sails." *Phys.org*, phys.org/news/2008-08-history-solar.html#:~:text=In%201960%2C%20Echo%2D1%20felt. Accessed 13 Apr. 2024.

"How Do Photons Move Solar Sails?" Physics Stack Exchange, physics.stackexchange.com/questions/677865/how-do-photons-move-solar-sails#:~:text=Photons%20do%20not%20interact%20with%20electrons. Accessed 10 Apr. 2024.

NASA Next-Generation Solar Sail Boom Technology Ready for Launch - NASA. 10 Apr. 2024, www.nasa.gov/general/nasa-next-generation-solar-sail-boom-technology-ready-for-launch/. Accessed 13 Apr. 2024.

"Solar Sails: A New Way to Explore Space." Www.linkedin.com, www.linkedin.com/pulse/solar-sails-new-way-explore-space-harishraj-d#:~:text=No%20fuel%20required%3A%20Solar%20sails. Accessed 10 Apr. 2024.

"Understanding Solar Sails: Advantages and Disadvantages." New Space Economy, 4 Mar. 2024, newspaceeconomy.ca/2024/03/04/understanding-solar-sails-advantages-and-disadvantages/#:~:text=Disadvantages%20of%20Solar%20Sail&text=The%20initial%20acceleration%20provided%20by. Accessed 10 Apr. 2024.

"What Is Solar Sailing?" The Planetary Society, www.planetary.org/articles/what-is-solar-sailing#:~:text=When%20light%20hits%20a%20solar.

Satellites for Crop Disease Prevention

Ryan Shams '26

Natural Science

Introduction:

Crop diseases are a detriment to the global agriculture sector and the global economy. It is estimated that these diseases result in 220 billion dollars in damages each year. The outbreaks are often sporadic, but large outbreaks that aren't prevented at early stages have caused catastrophic problems in the past. For example, a fungi outbreak in coffee plants in central america resulted in a loss of 1 billion dollars and 500 thousand jobs.

How can these diseases be prevented? Proper sanitation and crop rotations are two ways that minimize the risk of crop disease outbreaks. Specifically, clean planting sites without remains from the previous year's crops are used and the same soil is used for different types of crops.

However, the surveillance aspect for prevention is mostly lacking. For many farms, identifying crop diseases is a laborious process, involving analyzing patterns of the host plant, determining whether the problem was caused by a living organism or not, and looking for signs of disease to confirm a diagnosis.

It is essential that disease outbreaks are identified at an early stage. This is where the solution of satellite imaging and remote sensing technology plays a significant role, as it can help

farmers monitor their crops and diagnose diseases quickly. This article explores how this advanced technology can have a significant impact on the surveillance of crop disease.

Crop Disease:

Even with proper preventative measures and crop health being maximized, crop diseases are bound to occur. Some types of diseases are abiotic(non infectious), while others are biotic(infectious). Crop diseases only occur with specific conditions that allow the source of the disease to thrive, and when the host is susceptible to the disease. Sources of disease include bacteria, fungi, parasites, and viruses.

Satellite Imaging:

Satellite imaging is one component of a larger field called precision agriculture. This field uses technology and data analysis to assist in farm management. Satellites have the ability to capture high resolution images from space. Farmers can efficiently detect crop diseases by analyzing these detailed images.

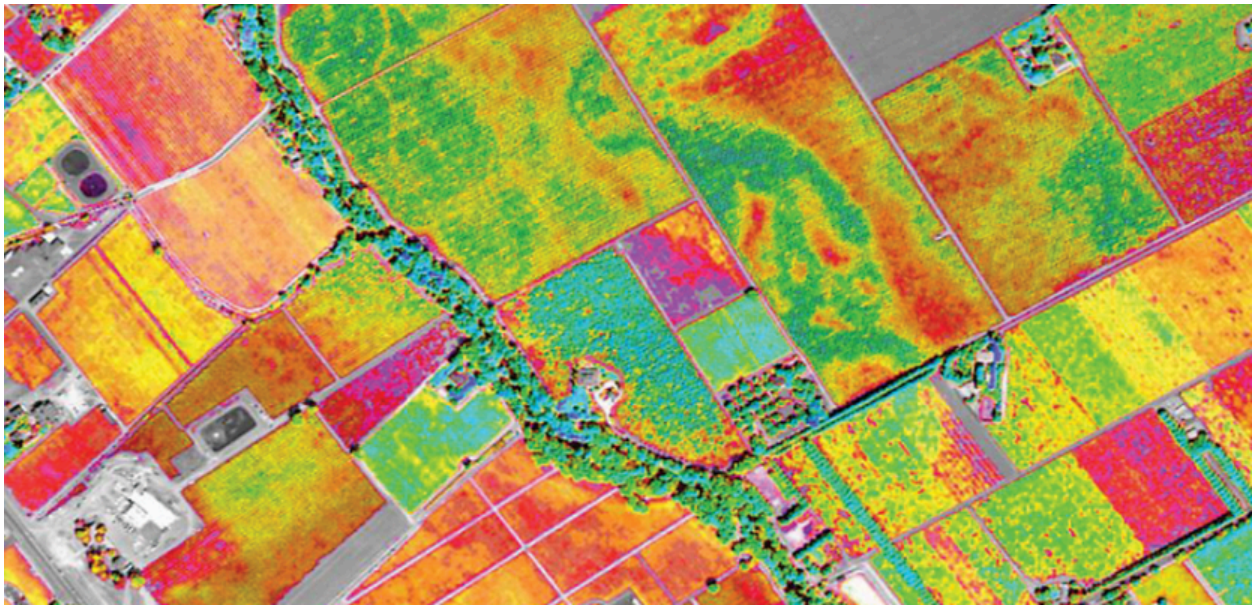


Image 1: High resolution image of farmland from a satellite

Satellites can cover large areas of farmland, and help collect data on crop health and areas where disease is spreading. The images from satellites can be used to make disease distribution maps, which highlight regions prone to specific diseases. Knowing this information can help farmers prepare management strategies if the diseases were to appear.

Remote Sensing Technology:

How are satellites able to collect this important data? These satellites are equipped with remote sensing technology, which is sensors that can distinguish different wavelengths in the electromagnetic spectrum, including visible and infrared radiation.

Image 2: Photograph of satellite equipped with remote sensing technology

Distinguishing the different wavelengths results in the remote sensing satellites being able to determine crop health. One way of doing this is by studying near-infrared radiation, which is located just beyond visible light in the electromagnetic spectrum. Just like visible light, plants absorb or reflect near-infrared radiation.

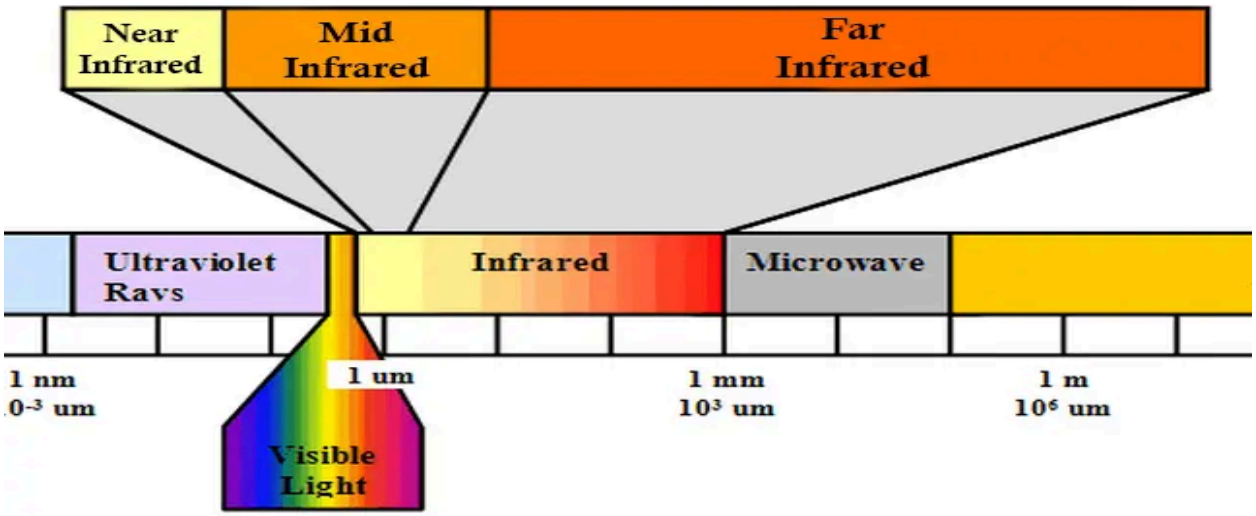


Image 3: Electromagnetic spectrum diagram

Chlorophyll is a compound that allows plants to absorb light energy and convert it into chemical energy for photosynthesis. Healthy plants that produce more chlorophyll reflect more near-infrared radiation than unhealthy plants that produce less chlorophyll. Since remote sensing technology can analyze the amount of near infrared reflection, it can be used to determine crop health.

Conclusion:

Using high resolution images and remote sensing technology, satellites have a substantial impact on the modern agricultural sector. Satellites detect and monitor crop diseases at early stages, analyze crop health, and identify risks in certain areas of farmland. These abilities help prevent potential widespread outbreaks.

As crop diseases are inevitable on farms, satellites will remain as a key component for agriculture. As technology advances, satellites will become even more effective for managing vast regions of crops.

References

Crop Diseases: How To Identify, Control, And Prevent. (2021, December 6). Eos.com. https://eos.com/blog/crop-diseases/

How can satellite imaging assist in detecting crop diseases? – Agriculture.Gov.Capital. (2023, September 8). https://agriculture.gov.capital/how-can-satellite-imaging-assist-in-detecting-crop-diseases

Major crop disease outbreaks. (n.d.). 2Blades Foundation. https://2blades.org/major-crop-disease-outbreaks/

Preventing and Managing Plant Diseases. (n.d.). Extension.missouri.edu. https://extension.missouri.edu/publications/mg13

shustova, alexandra. (2022, May 3). *How Remote Sensing Satellites Works*. Dragonfly Aerospace. https://dragonflyaerospace.com/how-remote-sensing-satellites-works/#:~:text=They%20observe%20and%20gather%20data

The Artificial Production of Plastic Degrading Enzymes

Milan Varma '25

Natural Science

In humanity's never-ending quest for strong and easily manufacturable materials, we have allowed materials like plastics to take over our land and oceans. Millions of animals are killed by plastics every year, such as birds, fish, and other marine organisms. To solve this problem, we looked to the one thing humans have always relied on; nature.

It is no surprise that we are facing a plastic and pollution crisis. The question is, how can we solve this problem without using massive amounts of energy, fossil fuels, or valuable time? The answer lies in nature.

More specifically, we looked at one of the most unknown but interesting creatures in nature, wax worms. In a study published by the journal *Nature Communications* in 2022, it was found that wax worms could degrade polyethylene, one of the most common forms of plastic found across the world. The first discovery of wax worms was in fact an accident that occurred in 2017 when Federica Bertocchini, a study author and amateur beekeeper, was tending to her hives. Bertocchini says that she came across these worms when she cleaned out her beehives and placed them into plastic bags, only to notice that holes began to appear. Rather than her initial suspicion that the worms were chewing the bags, it turned out to be a chemical reaction induced by the worms' saliva. After further research, it was discovered that the worms' saliva contained

The enzymes found in said saliva were an arylphorin and a hexamerin named Demetra and Ceres, respectively. These enzymes, when they come into contact with polyethylene-based plastic, perform a reaction that oxidizes the plastic and breaks it down after a few hours in room temperature conditions. An article posted by *Nature Communications* outlines an experiment into how exactly these enzymes work. The investigation commences with an initial observation that plastic waste emerges when a polyethylene film comes into contact with the recently formed wax worm cocoon and mouth secretions. Subsequently, they conduct an in-depth examination of the wax worm's saliva (referred to as GmSal), uncovering its ability to oxidize and break down polyethylene within a few hours. This effect is verified through Gel Permeation Chromatography (GPC) analysis of GmSal-treated polyethylene, which indicates the fragmentation of long hydrocarbon chains into smaller molecules. Utilizing Gas Chromatography-Mass Spectrometry (GC-MS), the team identifies degradation byproducts like small oxidized aliphatic chains, further confirming the polymer's breakdown into shorter molecules. Proteomic analyses of GmSal reveal the presence of a limited number of enzymes belonging to the hexamerin/prophenoloxidase family, including Demetra and Ceres.

Polyethylene is notoriously difficult to degrade and stays in the environment for long periods of time, polluting ecosystems and our lives. Therefore, there are huge amounts of polyethylene plastic in landfills and oceans around the world. Simply placing wax worms or collecting their saliva will not have a big enough impact on this global problem. Due to this, we will have to use a few methods in biology to dig deeper into these enzymes and replicate them.

In a hypothetical research scenario, we could begin by collecting wax worms and obtaining samples of their saliva. The objective would be to investigate and identify the gene or genes responsible for encoding a specific enzyme capable of degrading plastic. To achieve this,

we could employ genetic analysis techniques to pinpoint the target gene(s). Upon successful identification, we could use Polymerase Chain Reaction (PCR) technology to multiply and generate ample copies of the relevant gene(s), enabling us to have sufficient material for further experimentation. Subsequently, we could introduce the multiplied gene(s) into a plasmid, a circular DNA molecule commonly used as a vector in genetic engineering. For precise gene insertion, restriction enzymes could be utilized, allowing us to accurately cut and join specific DNA sequences. Lastly, we could explore the possibility of using bacteria as agents to degrade the plastic instead of relying on the worms. By inserting the modified plasmids carrying the plastic-degrading gene(s) into the bacteria, we could harness their natural metabolic processes to potentially break down the plastic waste. Through these steps, we aim to explore the potential of bacteria as a sustainable and efficient solution for managing plastic waste.

In this way, we utilize nanotechnology to provide a more efficient solution while looking to nature rather than fully artificial methods. Innovative nanotechnology offers promising avenues to enhance the enzymatic breakdown of plastic by wax worms. First, through nano-encapsulation, nanoparticles can be tailor-made to shield the wax worms' enzymes, safeguarding them from degradation or denaturation. This encapsulation not only ensures their stability but also boosts their efficiency in breaking down plastic waste. Additionally, nanotechnology presents the opportunity for enzyme enhancement, wherein the enzyme's structure can be modified or its catalytic properties improved. This approach has the potential to further enhance the effectiveness of the enzymes in degrading plastics. Moreover, nanotechnology enables targeted delivery of the enzymes, with nanoparticles serving as carriers to transport the enzymes specifically to plastic waste sites. This targeted delivery approach enhances the precision of the process, making it more efficient and environmentally friendly. By

harnessing these three cutting-edge strategies, we can explore novel and sustainable solutions to combat the global plastic pollution challenge.

References:

"The MSDS HyperGlossary: Biodegradable." Ilpi.com, 2013, www.ilpi.com/msds/ref/biodegradable.html#:~:text=Biodegradable%20materials%20are%20generally%20plant,microbes%20to%20do%20their%20work. Accessed 26 July 2023.
"A New Enzyme Breaks down PET Waste for Reuse. - ASME." Asme.org, 2022, www.asme.org/topics-resources/content/a-new-enzyme-eats-plastics#:~:text=Functional%2C%20active%2C%20stable%20and%20tolerant,parts%2C%20a%20process%20called%20depolymerization. Accessed 26 July 2023.
Wikipedia Contributors. "PETase." Wikipedia, Wikimedia Foundation, 18 July 2023, en.wikipedia.org/wiki/PETase. Accessed 26 July 2023.
"Engineering a Plastic-Eating Enzyme." University of Portsmouth, 31 Aug. 2022, www.port.ac.uk/news-events-and-blogs/news/engineering-a-plastic-eating-enzyme. Accessed 26 July 2023.
Parker, Laura. "The World's Plastic Pollution Crisis Explained." Environment, National Geographic, 7 June 2019, www.nationalgeographic.com/environment/article/plastic-pollution#:~:text=Millions%20of%20animals%20are%20killed,caused%20by%20entanglement%20or%20starvation. Accessed 26 July 2023.
Admin. "What Is Polyethylene? Monomer, Types, Properties, Applications (with FAQs)." BYJUS, BYJU'S, 22 May 2020, byjus.com/chemistry/polyethylene/#:~:text=Polyethylene%2C%20also%20known%20as%20polythene,synthetic%20polymers%20is%20in%20packaging. Accessed 26 July 2023.
Admin. "What Is Polyethylene? Monomer, Types, Properties, Applications (with FAQs)." BYJUS, BYJU'S, 22 May 2020, byjus.com/chemistry/polyethylene/#:~:text=Polyethylene%2C%20also%20known%20as%20polythene,synthetic%20polymers%20is%20in%20packaging. Accessed 26 July 2023.
"Differences between HDPE Plastic and Polyethylene Plastic." Sciencing, 2018, sciencing.com/differences-hdpe-plastic-polyethylene-plastic-6807965.html. Accessed 26 July 2023.
Wikipedia Contributors. "Polyethylene." Wikipedia, Wikimedia Foundation, 14 July 2023, en.wikipedia.org/wiki/Polyethylene. Accessed 26 July 2023.
Business Bliss FZE. "Polyethylene Chemistry and Molecular Structure." Ukessays.com, UK Essays, 21 Mar. 2023, www.ukessays.com/essays/chemistry/polyethylene-chemistry-and-molecular-structure.php. Accessed 26 July 2023.
Germain, Jacquelyne. "Wax Worm Saliva Is the Unlikely Hero of Fighting Plastic Waste." Smithsonian Magazine, Smithsonian Magazine, 7 Oct. 2022, www.smithsonianmag.com/smart-news/wax-worm-saliva-is-the-unlikely-hero-of-fighting-plastic

-waste-180980908/#:~:text=Wax%20worms%2C%20or%20moth%20larvae,hours%2C%20according%20to%20the%20study. Accessed 26 July 2023.
"Plastics Engineering - January 2015 - Plastic Failure through Molecular Degradation." Nxtbook.com, 2015, read.nxtbook.com/wiley/plasticsengineering/january2015/consultantscorner.html. Accessed 26 July 2023.
natrue. "Beating Plastic Pollution - World Environment Day 2018 - NATRUE." NATRUE, 27 Feb. 2019, natrue.org/de/beating-plastic-pollution-world-environment-day-2018/. Accessed 26 July 2023.
Rozsa, Matthew. "Ocean Microplastic Pollution Is Filtering up into the Seafood We Eat." Salon, Salon.com, 8 Mar. 2023, www.salon.com/2023/03/08/microplastics-ocean-fish-seafood/. Accessed 26 July 2023.
Kreutzer, Laura. "Growth Firms See Plastic Waste as an Investment Opportunity." Penews.com, Private Equity News, 6 July 2021, www.penews.com/articles/growth-firms-see-plastic-waste-as-an-investment-opportunity-20210706. Accessed 26 July 2023.

AI in Banking/Finance

Ryan Shams '26

Computer Science

Introduction:

Artificial Intelligence in banking and financial companies is a rapidly growing field that will likely have a large impact on the future of the financial sector. One factor for this growth is due to the increased usage of applications and websites for a variety of activities. For example, it is estimated that online banking usage increased by an estimated 20 to 50 percent during the first few months of coronavirus.

Currently, almost 60 percent of financial service companies have implemented an artificial intelligence based system. These types of systems help to optimize the efficiency that these companies are able to complete tasks, and allow them to have more services and personalization for customers. These systems, if implemented properly, have the potential to increase customer satisfaction and revenue, which will help companies against their competition. It is estimated that automation and artificial intelligence in banking will increase the sector's value by over one trillion dollars per year.

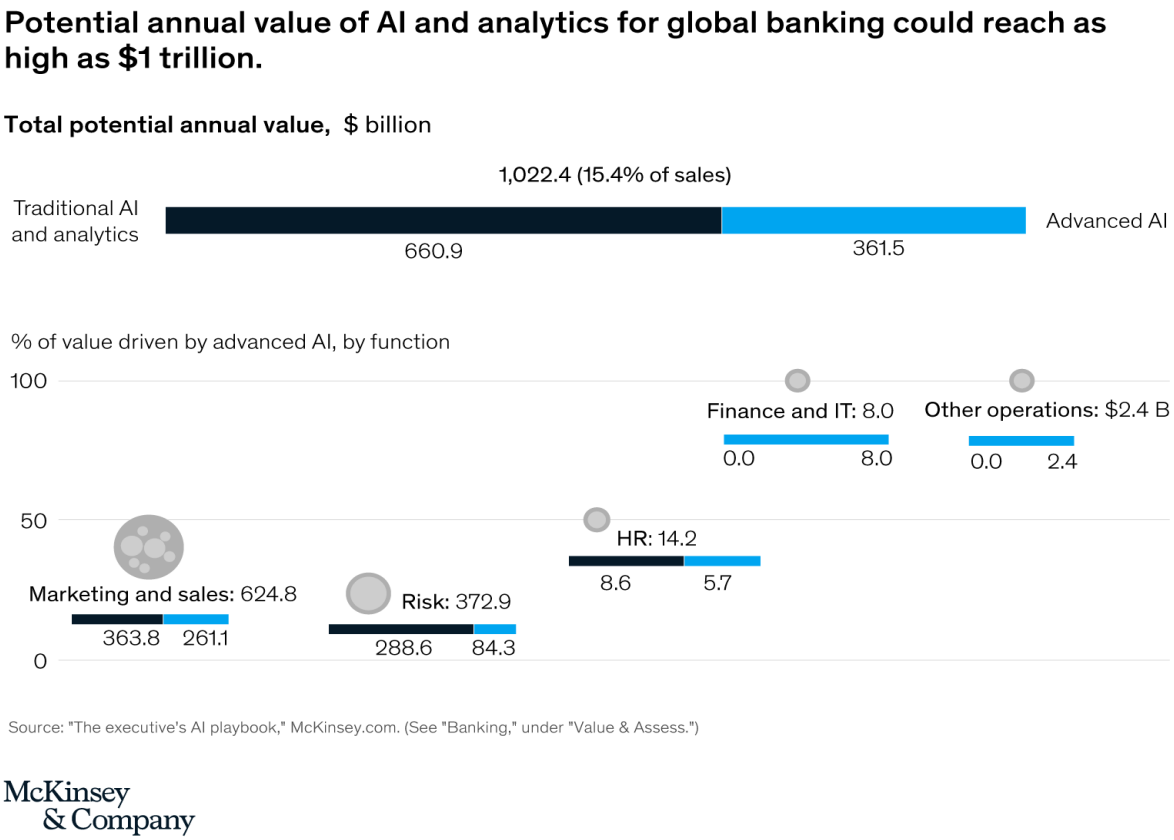


Figure 1: Graph showing the potential annual value for the global banking AI sector

However, there are some limitations to consider. Many of these systems are still in an experimentation stage due to a lack of clear goals and a limited level of technology. Security standards and safety regulations must be followed and prioritized. Furthermore, the current data backbone of many financial service companies must be changed, and rigorous testing of artificial intelligence models will need to take place for the systems to be used at a larger scale. There is also the obstacle of managing and securing these systems. This article explores the different uses of these systems, how they function, and the benefits and disadvantages of them.

Fraud Protection:

Artificial Intelligence fraud protection systems can help secure consumers' accounts and identities. Efficient and effective fraud protection is important to gain trust with consumers. However, fraud has become a growing issue recently, as reported fraud losses increased by 70% from 2020 to 2021, to a total amount of 5.8 billion dollars. Correspondingly, the use of advanced fraud prevention systems with automation and artificial intelligence increased from 10% to 31%. These systems can help prevent types of fraud such as money laundering, identify theft, and credit card fraud through the analysis of transactions. Additionally, loan and mortgage fraud can be prevented by analyzing and classifying the loan and mortgage documents.

These systems can accurately detect and flag fraudulent activity by analyzing past transaction history of a consumer. Furthermore, this data could help detect money laundering. This type of fraud prevention is also more effective than the traditional rule based system, which categorizes transactions based on a series of rules to detect fraud. An artificial intelligence based system can use machine learning to adapt to changes in consumer behavior and improve by using the transactions as training data. As a result, there are less normal transactions flagged. Paired with a fraud protection team, flagged transactions are then checked and efforts can be made to notify the consumer of fraud and prevent future fraudulent activity from the account.

There are companies that have already created and implemented these types of systems. For example, a data science company, Feedzai, has developed a machine learning system to identify fraud by inspecting transactions. This company has partnered with Citibank to help them catch abnormal behavioral changes among their customers before transactions are completed.

Similarly, the company DataVisor has examined transactions with a machine learning based method to prevent 15 million dollars worth of fraud losses.

Using the training data, a model is able to recognize patterns of both normal and fraudulent transactions. In comparison with the traditional rule based model, this model is more adaptive as it recognizes patterns for each individual consumer and pairs it with its training on fraudulent transactions. It also changes according to a shift in a consumer's spending patterns. The model is trained to recognize three types of anomalies to accomplish these tasks.

Collective: The machine learning model collects similar data points that are considered abnormal only when compared to the rest of the data set. For example, two transactions occurring at one time at different locations is a collective anomaly because it is normal for there to be two transactions at the same time, but when at different locations it is clear that there is abnormal and possibly fraudulent behavior.

Conditional: A conditional anomaly is when multiple values or variables greatly differ from the rest of the dataset. For example, if there are many withdrawal attempts in a short period of time, it would be flagged because it is a large shift in consumer behavior.

Point: A point anomaly is a single data value that is greatly different from the rest. For example, if there is a sudden large transfer to an account that normally receives smaller transfers, this could be flagged as potentially fraudulent because one transfer is much larger than the previous ones.

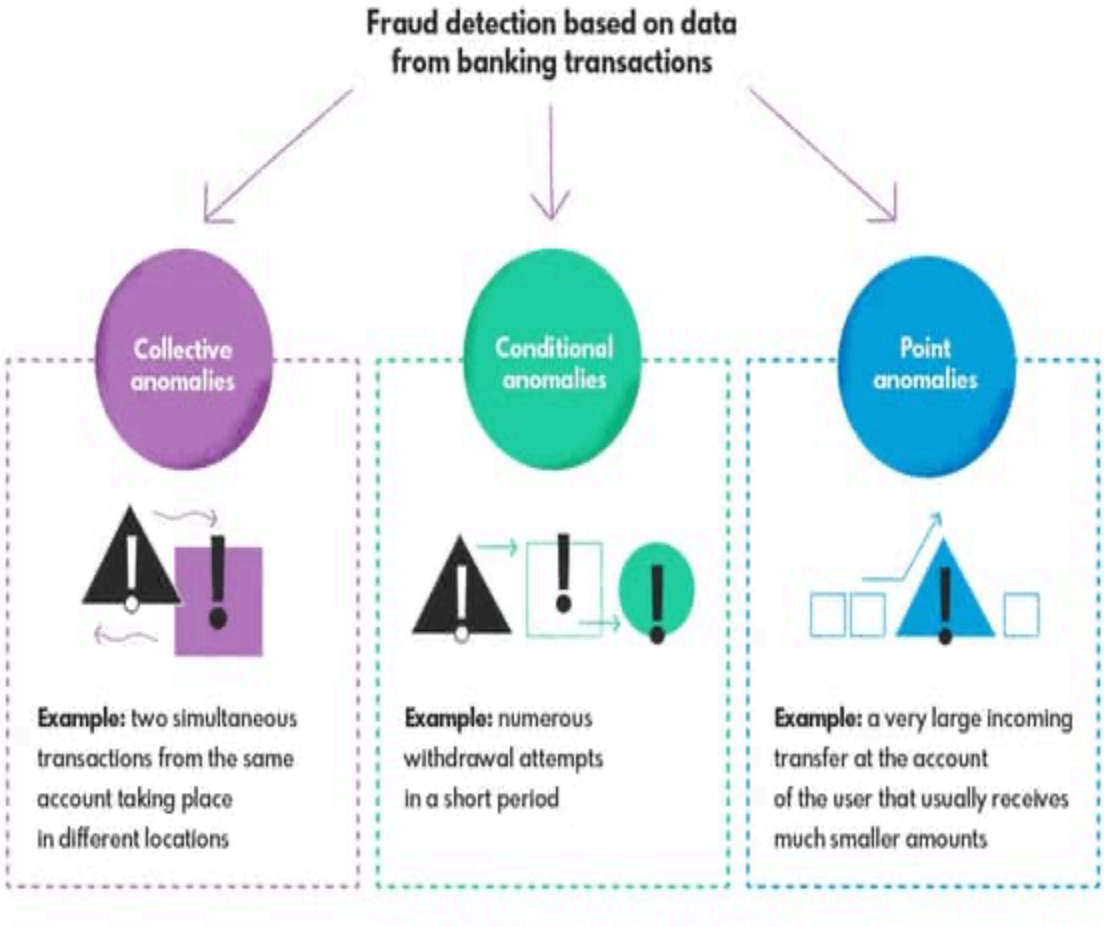


Figure 2: Fraud detection types with examples

These models can be trained using two different methods: supervised learning and unsupervised learning. Supervised learning is when models are trained on previously tagged outputs, and the amount of training data and the level of organization of the data determines the reliability of the model. Unsupervised learning is more advanced, as it is able to identify similarities and differences within the data, and a model is created based on the pattern recognition.

There are three common algorithms that models can use: logistic regression, decision tree, and neural networks. Logistic regression is the most common, and it graphs the relationship between the dependent variable(probability of fraud) and the independent variables(transaction amount, frequency, and location). The result is a sigmoid/ S shaped graph with the x value being the independent variables and the y value being the probability of fraud. After being trained on previous outputs, a model has the ability to recognize patterns of both normal and abnormal transactions, and based on this it is able to plot each transaction on the logistic regression graph. Doing this helps assess the risk of fraud for each transaction, from 0(0% likelihood) to 100(100% likelihood).

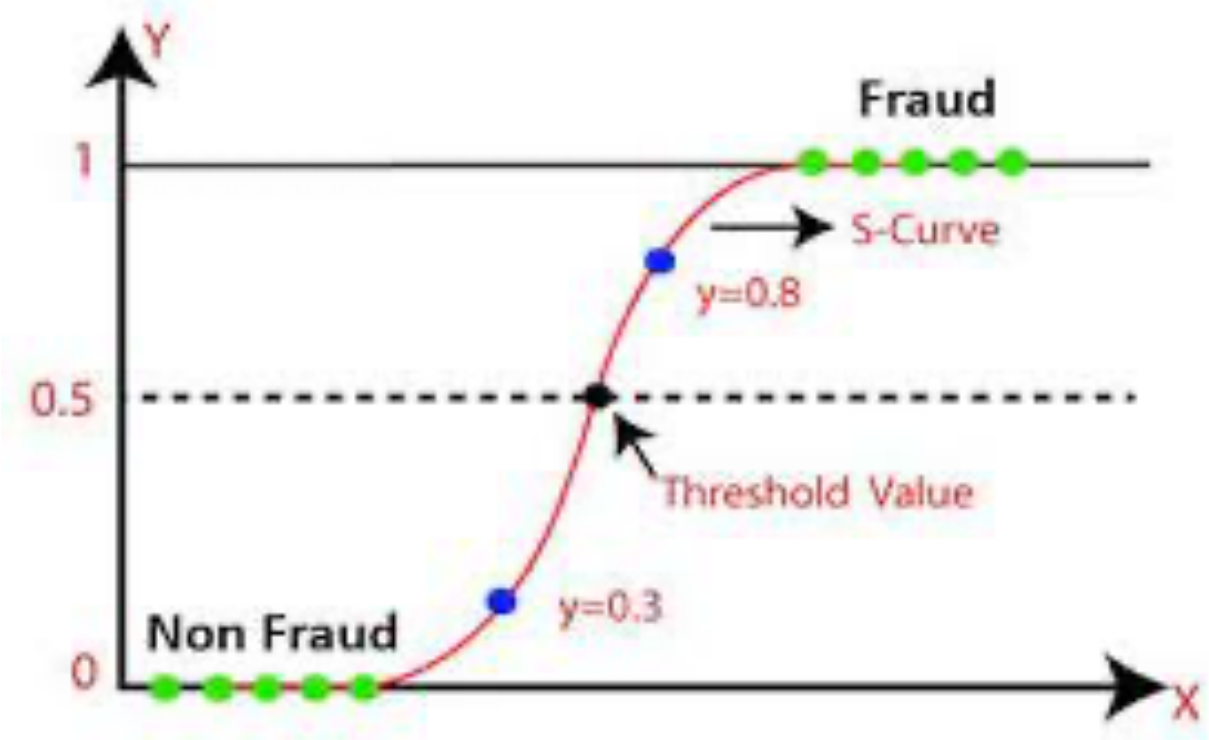


Figure 3: Sigmoid logistic regression graph for fraud detection

Another method is a decision tree, which involves a transaction being classified into numerous groups and subgroups, and the probability of fraud is based on its classification. Location, transaction amount, and relation to past transaction history are examples of variables that would influence the classification.

The third method is neural networks, which uses algorithms such as pattern recognition and data mining to classify transactions. The design of neural networks is inspired by the human brain, as the models are built to try to process data in a similar way to how the brain does it. This system is able to improve over time and analyze complex patterns of various transactions.

A secondary way that artificial intelligence and machine learning have been used is with the review of loan and mortgage applications. This type of use is currently in an early developmental stage. For example, a model was developed that uses Natural Language Processing(a type of machine learning for understanding text) to classify loan and mortgage applications. There were two main challenges with creating the model: a small difference between fraud and non-fraud activity and an imbalance between fraud and non fraud applications. To resolve these problems, the model was also trained on company policy and statistics, and more data was sampled using different sampling methods. The model was trained to review these applications and identify fraud.

Bank Account Management/Data Analysis:

Artificial Intelligence based systems can also be used to help manage the bank accounts of consumers and help make personalized financial recommendations for them. These systems can assist customers with completing loan applications or transferring funds between accounts.

One use of these systems in this sector is to help give credit scores. Machine learning models have the ability to analyze existing debts, payment history, and use of credit to assess a consumer. These models analyze and use a wide range of data, more than the traditional algorithmic approach that is used to calculate many common credit scores, such as the FICO score.

Regulations for financial companies are also very significant, and artificial intelligence systems can help analyze the lengthy financial reports by using natural language processing and identifying the key variables/information. This process is known as extractive document summarization, and it can be used to extract essential information and create a summary with it. The key information from a passage is selected and then condensed, so that just the important information can then be reviewed. Firstly, specific key sentences are selected from different passages that can convey the main idea of the document, which are sentences that are scored as "high". The high-scored sentences are kept, which creates a shorter summary of the entire document.

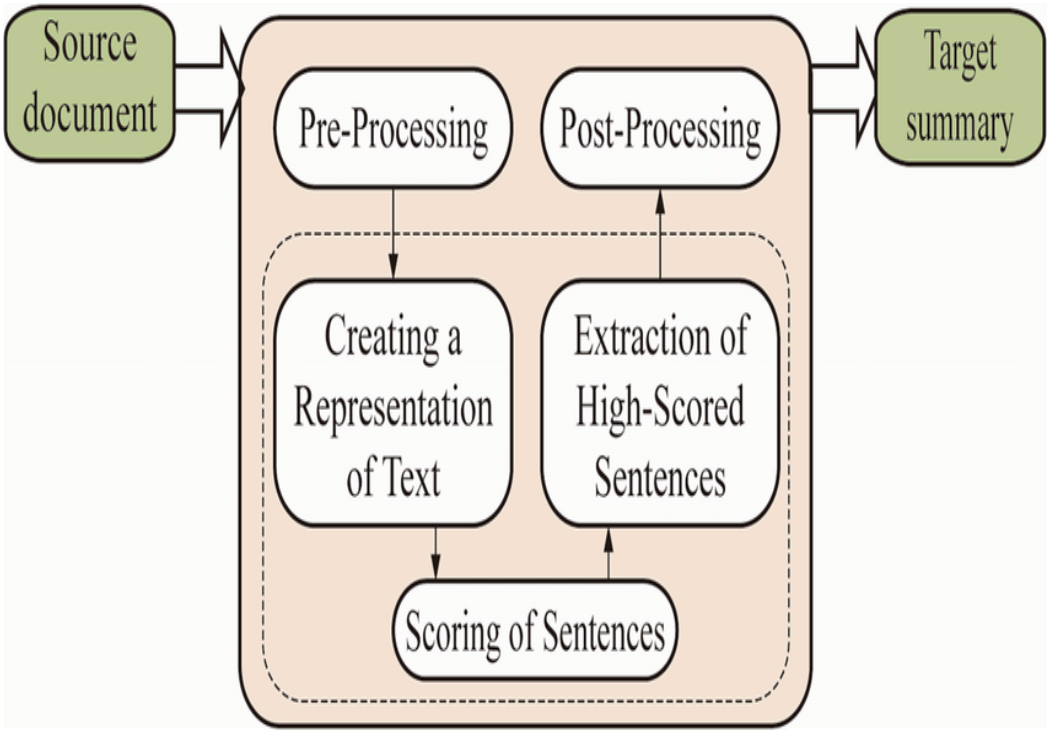


Figure 4: Diagram showing the process of extractive document summarization

Any sort of business could use sentiment analysis, a branch of machine learning, to review and process criticism data. The parts of the text that express opinion or sentiment can be taken from unstructured data and then can be classified as positive, neutral, or negative. The sentiment polarity (the rating on how positive or negative the review is) can also be determined. Since larger companies often receive thousands of reviews on varieties of platforms, it is impossible to analyze all of them, and a system that can help understand the overall view of a company is critical.

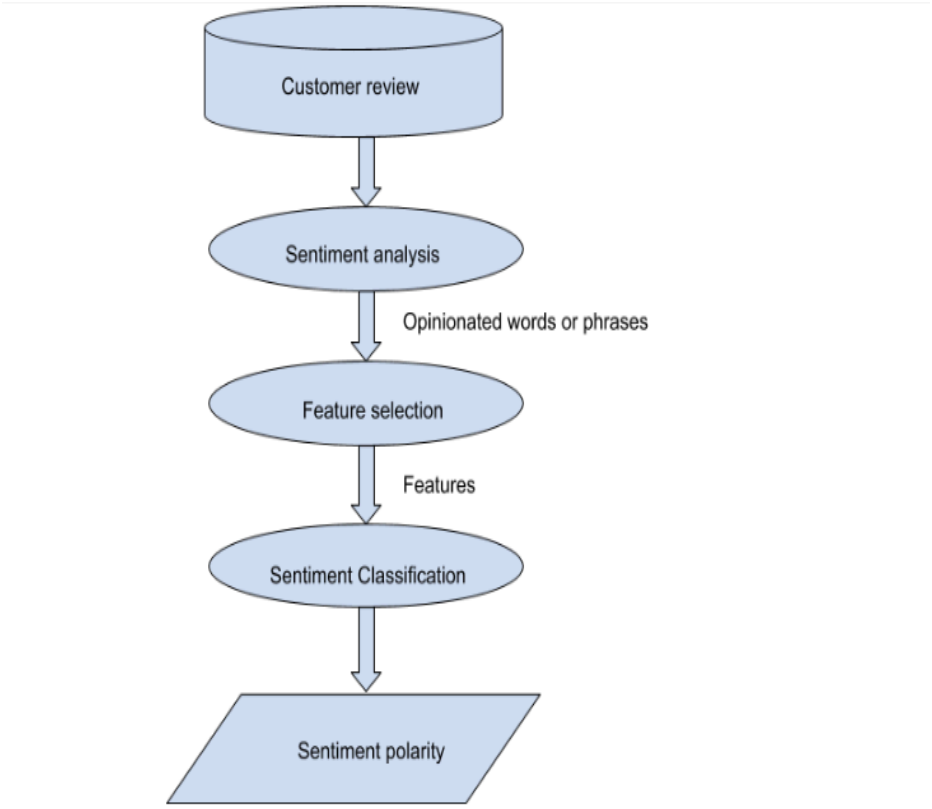


Figure 5: Basic Diagram of Sentiment Analysis

These different systems are currently in a development stage in certain companies. For example, the credit card company Capital One has created an algorithm to check credit data of consumers, improving the efficiency of loans and lending. The banking corporation JP Morgan has used extractive document summarization to review important legal documents. Additionally, the analytics company Kensho Technologies has provided a machine learning based system to financial companies to help their consumers answer financial questions.

Conclusion:

This article explored the uses of artificial intelligence in finance, specifically for fraud protection, bank account management, and data analysis. In order to assess how to implement these systems and to what extent, it is important to discuss the advantages and disadvantages, as well as the limitations and obstacles that must be overcome.

One significant disadvantage is the high likelihood of failure after first implementing these systems. Managing and analyzing vast amounts of data with artificial intelligence based models is prone to errors, which is why it is essential that financial companies do not become overly reliant on these systems for any specific use.

Another limitation is the difficulty involved in making sure regulations are followed when implementing these systems, and that data is kept private and secured. There are regulations regarding the use of data for companies, and it is important that it is the main priority for companies. The upfront development costs such as hiring and training staff to develop these systems and purchasing the technology can range from tens to hundreds of millions of dollars depending on the scale of the company. The potential for job loss could also be another concern, because it is possible that certain roles could be replaced.

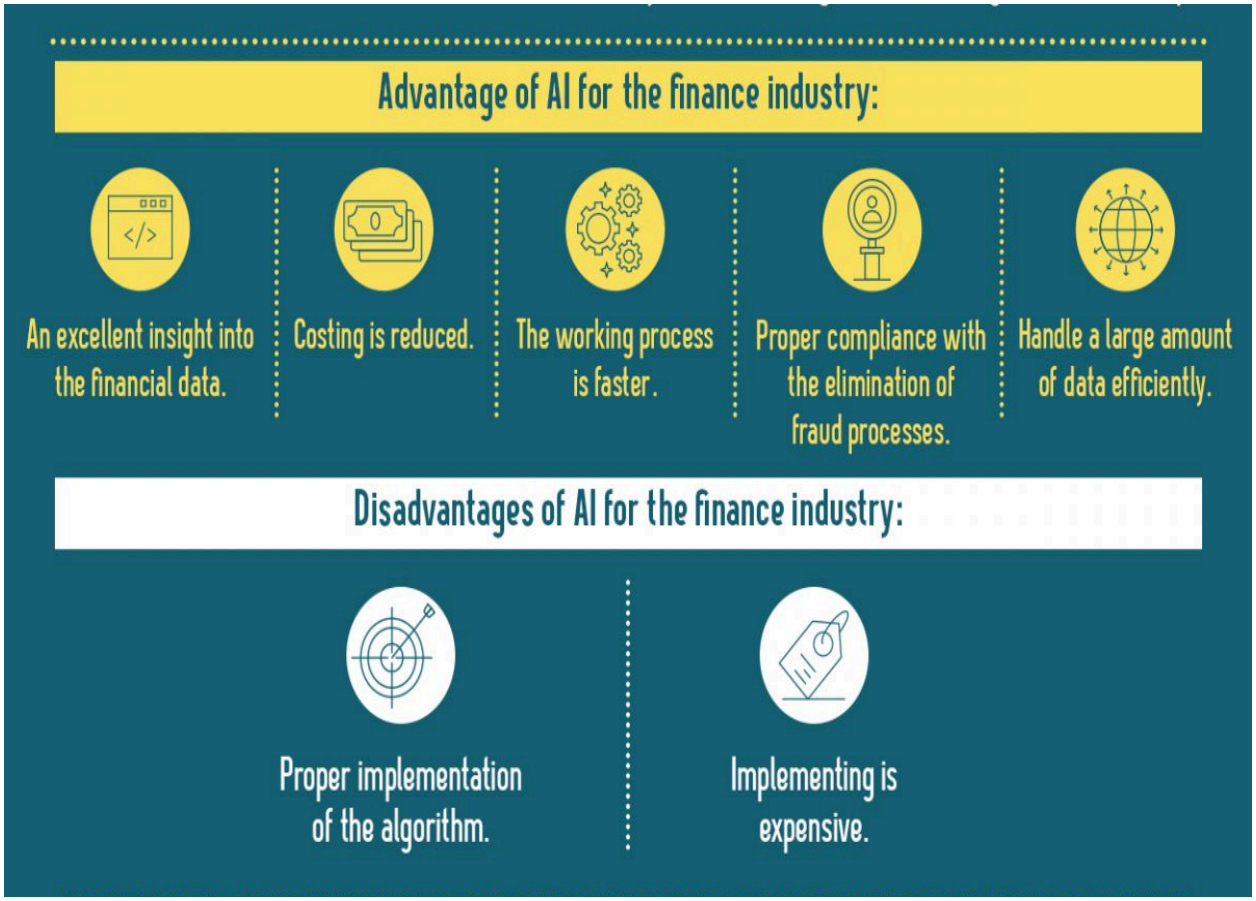


Figure 6: Display of advantages and disadvantages

What are the advantages of artificial intelligence in finance? One of the most significant advantages is fraud protection, due to the rising amount of fraud losses, which was estimated to be 5.8 billion dollars in 2021. Fraud protection systems using logistic regression, decision trees, or neural networks can categorize transactions and identify fraud accurately. Another significant advantage is analyzing vast amounts of data beyond the amount that employees can analyze. Processes such as document summarization, sentiment analysis, and credit scoring are complex processes that can be done using artificial intelligence systems.

To conclude, artificial intelligence in finance is a continuously growing field that has the potential to resolve problems within the industry. If used properly, it can be used to make numerous processes in the sector more efficient.

References

AI in Banking – How Artificial Intelligence is Used in Banks. (n.d.). Appinventiv.com. https://appinventiv.com/blog/ai-in-banking/amp/

Artificial Intelligence in Bank Fraud Detection and Prevention - SQN Banking Systems. (2022, October 7). https://sqnbankingsystems.com/blog/artificial-intelligence-in-bank-fraud-detection-and-prevention/#:~:text=The%20AI%20gets%20to%20know

Biswas, S., Carson, B., Chung, V., Singh, S., & Thomas, R. (2020, September 19). *AI in banking: Can Banks Meet the challenge?*McKinsey. https://www.mckinsey.com/industries/financial-services/our-insights/ai-bank-of-the-future-can-banks-meet-the-ai-challenge

Fraud Detection Algorithms | Fraud Detection using Machine Learning. (2019, September 12). Intellipaat Blog. https://intellipaat.com/blog/fraud-detection-machine-learning-algorithms/?US

Gossett, S. (2019). *AI bankability: 10 ways artificial intelligence is transforming banking*. Built In. https://builtin.com/artificial-intelligence/ai-in-banking

Larcelet-Prost, J.-R. (2023, May 1). *5 Ways AI Prevents Fraud in Banking and Fintech*. ReadSpeaker. https://www.readspeaker.com/blog/anti-fraud-ai/#:~:text=Increasingly%2C%20they%27re%20using%20artificial

Loan Application Fraud Detection. (2022, February 28). Nexocode. https://nexocode.com/case-studies/loan-application-fraud-detection/

Fast Inverse Square Root

Jack Ford '26

Computer Science

Introduction

During the 1990s, 3D video games were in their earliest stages. The visuals in these games were heavily restricted by the hardware of the time, which could barely perform mathematical calculations quickly enough to display the game at suitable frame-rates. This led programmers to come up with innovative solutions to work around these hardware restrictions. One such solution was the fast-inverse-square-root algorithm, or Q_sqrt, used by id software in their 1999 game: *Quake III Arena*. This algorithm sought to solve a seemingly simple problem: what is the inverse square-root of a 32-bit floating point number?

Why?

Vectors are used everywhere in 3D graphics. When simulating physics, reflections, and lighting, these vectors need to be normalized to have a maximum value of 1 and a minimum value of 0. To do this, one can simply divide the vector by its magnitude, which is $\sqrt{x^2 + y^2 + z^2}$. This is also the equivalent of multiplying by $1/\sqrt{x^2 + y^2 + z^2}$. Calculating the inverse square-root of a number is wildly important to 3D games; this calculation is performed millions of times per second. However, the two necessary mathematical operations, division and square-root, are very slow on older computers. Therefore, any small performance increase to this calculation has significant improvements to the game's overall performance.

Code Overview

```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 1.5F;

    x2 = number * 0.5F;
    y  = number;
    i  = * ( long * ) &y;                       // evil floating point bit level hacking
    i  = 0x5f3759df - ( i >> 1 );               // what the fuck?
    y  = * ( float * ) &i;
    y  = y * ( threehalfs - ( x2 * y * y ) );   // 1st iteration
//  y  = y * ( threehalfs - ( x2 * y * y ) );   // 2nd iteration, this can be removed


    return y;
}
```

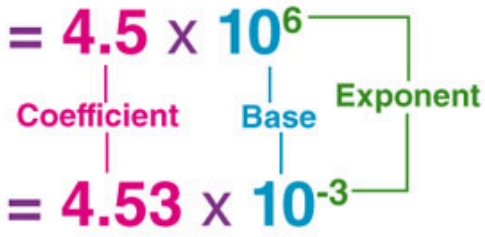
Original (unedited) algorithm code with preprocessor directives removed (Wikimedia)

While the algorithm may seem daunting at first, this paper will only focus on the most crucial step: $i = 0x5f3759df - (i >> 1)$. This simple line performs the large majority of the calculation; all other lines are used to allocate memory or refine the initial approximation. In this code, the variable 'i' is essentially a 32-bit floating point number. In addition, '0x5f3759df' is simply a hexadecimal representation of a constant number, which will be derived later in this

paper. To understand the math behind this algorithm, it is first imperative to understand how floats are stored in the C programming language.

Floating-Point Numbers

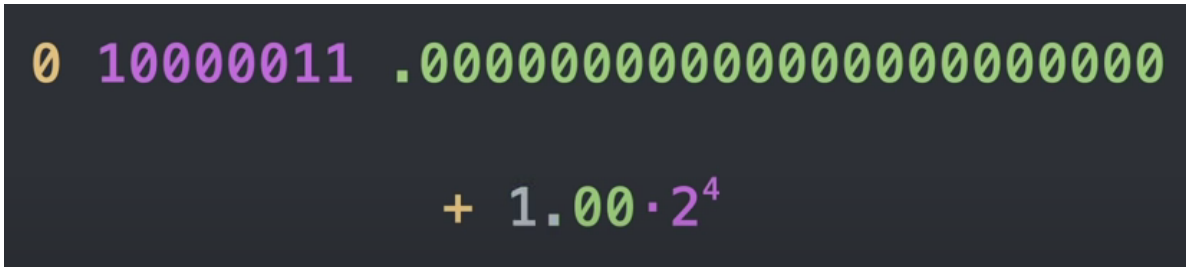
In C, floating-point numbers are represented using the IEEE-754 standard. Single precision numbers, the ones used in the Q_rsqrt algorithm, are made up of 32 bits. The first bit is used to determine whether the number is positive or negative, which is known as the number's sign. However, because it is impossible to take the square root of a negative number, the Quake developers assume that all inputs are positive and simply ignore this bit. This leaves 31 bits that can be used to represent precise decimal numbers. To maximize the precision and size of these numbers, the developers of IEEE-754 employed scientific notation. In standard scientific notation, there is a coefficient, a base, and an exponent.



Example of base-10 scientific notation (Noon Learning)

The coefficient can be split into two parts: the single-digit whole number and the decimal values that precede it, known as the mantissa. The range of the coefficient is simply [1, base). For instance, in binary's base-2 number system, the coefficient has the range [1, 2). Because of this limited range, the formula to get a full coefficient in binary is simply 1 + M, where M is the

mantissa. Because the 1 remains constant no matter the number, it is only necessary to store the mantissa and not the full coefficient. Despite only saving 1 bit of storage, this clever discovery enables the mantissa to store an extra 4,194,304 numbers. In IEEE-754, the last 23 bits are used to store this mantissa, while the remaining 8 bits are used to store the exponent. While 8 bits can usually represent values between 0 and 255, in order to account for negative numbers, the range is altered to values between -127 and 128. The formula used to convert an exponent E to scientific notation is: 2^{E-127}.



Color-coded representation of bits and their base-10 counterpart (Nemean)

The Key Insight

Using mantissa-exponent form, a formula can be created to represent the bits of a number: $M + 2^{23} * E$, where E is the exponent and M is the mantissa. This is because the exponent needs to be shifted left by 23 bits in order to prepend the mantissa, which is the same as multiplying by 2^{23}. The same idea applies to the mantissa, which needs to be divided by 2^{23} in order to create the decimal part of the coefficient. Combining all previous steps, this formula can be created to convert a floating-point binary number into a base-10 decimal number:

$$x = (1 + \frac{M}{2^{23}}) * 2^{E-127}$$

The key insight of this algorithm comes from the realization that the log of this complicated equation can be manipulated to include the bits of the number. First, take the log_2 of the entire equation and use exponent rules to simplify.

$$log_2(x) = log_2(1 + \frac{M}{2^{23}}) + E - 127$$

Second, because the $log_2(1 + x) \approx x$ for small values of x, then $log_2(1 + \frac{M}{2^{23}}) \approx \frac{M}{2^{23}}$. However, in order to get a more precise approximation, the constant μ is added as an optimization of the error between $f(x) = log_2(x)$ and $f(x) = x$ between 0 and 1. Further rearranging simplifies the equation to:

$$log_2(x) \approx \frac{(M+E*2^{23})}{2^{23}} + \mu - 127$$

As stated before, $M + E * 2^{23}$ is the formula for the bits of the number. The above equation contains this formula shifted and stretched by some constants. Therefore, when dealing with binary numbers, the bit representation of a number is essentially its own logarithm.

Bit Shifting

In computer science, a bit shift moves each digit in a binary number left or right. For instance, when shifting left, the number loses its most significant bit and a 0 is inserted on the other end. The << operator denotes a left shift, while the >> operator denotes a right shift.

$$0010 << 1 = 0100$$

Left bit shift of 0010 (Interview Cake)

Because binary works in a base-2 system, shifting a number left multiplies its value by 2 while shifting a number right divides it by 2. In the Q_rsqrt algorithm, the log of $\frac{1}{\sqrt{x}}$ can be rewritten as $-\frac{log_2(x)}{2}$. Because bit shifting is computationally faster than division, a bit shift is used to replace the division by 2 and speed up the program. This is where $-(i >> 1)$ is derived from; it replaces the division in $\frac{1}{\sqrt{x}}$ by shifting all of the number's bits right by 1.

0x5f3759df

The final mystery of this algorithm is the constant 0x5f3759df. As previously stated, the log of a number's value is roughly equal to its bits. However, there are some shifts and stretches to these bits that must be accounted for in the final computation. 0x5f3759df is a magic number that accounts for these shifts and is represented by this equation. To derive this equation, remember that the calculation for $y = \frac{1}{\sqrt{x}}$ is based on the identity:

$$log_2(y) = -\frac{1}{2} log_2(x)$$

Substituting the approximation of the logarithm for both x and y, a formula can be constructed where B represents the bits of the number:

$$\frac{1}{2^{23}} * B_y + \mu - 127 \approx -\frac{1}{2}\left(\frac{1}{2^{23}} * B_x + \mu - 127\right)$$

To find the inverse square root of x, simply rearrange this equation and solve for B_y:

$$B_y \approx \frac{3}{2} * 2^{23} * (127 - \mu) - \frac{1}{2}B_x$$

As previously mentioned $-\frac{1}{2}B_x$ is computed using a right bit shift. This written in the in code as: $-(i >> 1)$. The rest of the calculations are stored in the magic number 0x5f3759df, which is formally defined as:

$$\frac{3}{2} * 2^{23} * (127 - \mu)$$

This number is the remnants of the constant μ, the scaling factor, and the shifting. Combining these two steps, $0x5f3759df - (i >> 1)$ is used to quickly calculate the inverse square root of any positive number.

Conclusion

The fast-inverse-square-root algorithm is a perfect example of the many optimizations used by developers in the late 1990s and early 2000s. Strict hardware limitations made it very difficult to generate complex computer graphics or perform large mathematical operations, especially at the speeds required by gamers, scientists, and enthusiasts. Although now obsolete due to better hardware, Q_sqrt was approximately four times faster than other methods and led to astronomical performance improvements for *Quake III: Arena*. Without clever developers and ingenious algorithms like Q_sqrt, the computer industry would fail to achieve the remarkable feats of efficiency and innovation that we see today.

From Transplant to Regeneration: Recent Advances in the Functional Restoration of the Upper Extremity Amputee

Abdullah Kanchwala '25

Biology

Three hundred fifty thousand people in the United States suffer from the loss of an upper extremity and now live as amputees (Rawan ElAbd, 2023). While there have been many advances in the care of these patients, no current treatment options fully restore patients to their pre-injury state. Dating back to Hammurabi's code in 1755 BCE, amputation of the hands was a punishment for severe crimes because of its devastating social and economic consequences. The importance of hand function can be measured by the number of neurons dedicated to the hand in the sensory and motor cortices of the brain. As seen in the homunculus (Appendix 1), a stylized representation of the brain, the hand occupies large regions of the sensory and motor cortices. The coordinated movement of 54 bones, myriad forearm muscles, and two sets of tendons requires extensive space in the motor cortex. Similarly, the density of sensory receptors in the hand is rivaled only by the lips (Matters 2022). Total functional restoration after an amputation requires both motor and sensory restoration. The purpose of this paper is to examine advances in existing treatment modalities for upper extremity amputees and emerging concepts in bioelectricity that hold the potential for total limb regeneration.

While there are many different ways to treat patients with an upper-extremity amputation, none provide a complete restoration of both form and function of the original limb. Traditional prosthetic hands have primarily served as "helper" hands and have assisted a patient's other functioning hand. These simple devices lack fine motor ability and often have only rudimentary grasping ability (Godwin 2022). More modern myoelectric prostheses can provide patients with much more nuanced motor ability. Such prosthetics can be further enhanced by total muscle reinnervation (TMR), a surgical procedure invented in 2002 at Northwestern University by Dr. Gregory Dumanian (Medicine, N. 2023). His procedure involves taking nerves from the amputated extremity and reconnecting them to new motor nerves in the stump of the amputation. After healing, nerves from the severed extremity now cause micro-movements within the muscles of the stump (Davis 2023). When the patient imagines moving a missing wrist joint, for example, nerves that had previously carried out those signals now cause micromovements in the forearm muscles (Appendix 3). Sensors in modern myoelectric prostheses can detect these new outputs, allowing the patient to use the prosthesis in a more intuitive way (Rawan ElAbd, 2023). Unfortunately, while this does restore the motor function of the hand, there is little to no sensory feedback.

A prosthetic hand, particularly a more modern one, can work well in patients with a single amputation. However, patients with bilateral hand amputation (often the result of devastating infections) have a harder time with recovery due to the lack of sensory input. For some patients with bilateral hand amputations, advances in hand transplantation can provide a more fundamental solution. Hand transplantation is similar to the transplantation of human organs, such as the kidney or the heart. However, due to the fact that the hand is a composite group of tissues – bone, tendon, muscle, and skin– there is a greater risk of rejection of the donor

hand (Appendix 2)(Takeo, M., Lee, W., & Ito, M, 2015). Transplantation, unlike prosthetics, allows for nerve regrowth and restores both sensory and motor function after a prolonged period of rehabilitation and healing. A significant drawback of this treatment, however, is the requirement of lifelong immunosuppression similar to patients receiving a solid organ transplant such as a kidney or heart (Montagna 2019). Among the many risks of immunosuppression is a higher incidence of cancer due to the immune system's important role in clearing abnormal cells, as well as the risk of infections.

As a result, while there have been many advances in both prosthetics and hand transplantation, neither one of these approaches provides a solution for both the sensory and motor aspects of hand function. Limb regeneration, if possible, would allow for the ideal restoration because it would allow sensory feedback without requiring immunosuppression. In the last several years, researchers have begun investigating the differences in regenerative capacity between humans and simpler organisms such as flatworms and starfish (Matters, 2022).

Limb regeneration in the natural world has been seen as an extraordinary phenomenon that only select animal species can access, the most prominent being Salamanders and Axolotls, which can regrow limbs through epimorphic regeneration, which is characterized by the formation of a mass of undifferentiated cells called a blastema. Epimorphic regeneration allows certain reptiles to de-differentiate cells to regrow lost limbs, joints, and even organs (Arenas, 2021). The blastema forms a microenvironment around the injury in the reptile, and different signaling proteins, such as fibroblast growth factor (FGF) and bone morphogenetic protein (BMP), drive cellular proliferation in the blastema (Joven, A., Elewa, A., & Simon, A, 2019). In these organisms, the role of nerve signaling and guidance cues from the nervous system shows the interconnectedness of cellar and neural elements in limb regeneration.

According to new research published by the University of Chicago, it appears that the key to unlocking this regenerative potential in humans involves understanding the concept of bioelectric potential. At its core, bioelectricity is the electrical networks in the body that send and process information. Bioelectricity is created through ion-channel proteins found on the surface of most cells, allowing each to communicate through a voltage potential with neighboring cells (Levin, 2009). When a person has a thought, that thought is sent throughout the body through electrical synapses to elicit a physical response, for example, walking upstairs. Bioelectricity is the layer between the mind and the chemistry of the body. It allows your cognitive processes to affect the molecules in your muscles and other body parts. The body's cells and bioelectricity can be compared to the hardware and software of a computer – the cellular structures represent hardware, while bioelectricity is the software (Thorp, 2022).

Work in studying the planarian worm has helped understand the role that bioelectricity plays in limb regeneration. The worm can grow from cut limbs, being on record to regrow a full copy of itself from 1/273rd of its original body. In 2013, Michael Levin, a professor at Tufts University, examined pieces of a planarian worm and found electrical patterns embedded into the worm's tissue; these patterns were instructions to rebuild the entirety of the worm, except they were found in tail segments in this case. The bioelectrical pattern allows the planarian worm to rebuild its head and organs just from the tissue from the tail. Levin's team was able to exploit this, using different drugs that controlled the opening and closing of the ion channels that made the bioelectric potential and were able to rewrite the pattern. Levin's team applied this methodology to frogs in the lab, which yielded successful results; they could regenerate a frog's leg after an amputation (Levin, n.d).

Under normal circumstances, the human body responds to injury through scar formation or fibrosis. This process involves myofibroblasts, which rebuild and remodel the extracellular matrix environment in the injured area to protect the body further. As fibroblasts deposit ECM, the structure of the matrix is altered and becomes stiff, causing scarring. Applying bioelectricity to human limb regeneration would allow reprogramming of the scar response and direct the body toward regeneration (Madhusoodanan, 2018). Electrical synapses would tell the cells to retread to the kinds of paths that they took during embryogenesis to rebuild the damaged structure rather than the scar formation. The regeneration of an entire limb takes massive amounts of biological resources, and the human body doesn't have this energy budgeted for regeneration. Levin's team has partially solved this problem by attaching wearable bioreactors, which can form an aqueous microenvironment around the amputated limb. This method allows energy to be directly and continuously fed to the regenerating cells. Research is being continued on this "second stage" of human regeneration. If a sustainable method is found, it would mean great things for many patients, including spinal cord injuries and brain tissue regeneration (Uchicago, 2023).

Advances in regenerative medicine may lead to a more complete solution for the hand amputee. If such a mechanism could be controlled, it could allow for the regeneration of one's hand rather than using either a prosthesis or a transplant (Fell, 2018). Such regeneration would provide the ideal sensory and motor component and be the ideal recovery course for an amputee patient. Total limb regeneration may currently seem to be a domain not of science but of science fiction. That being said, hand transplantation, a technique that has now been replicated in many centers in the world, was similarly viewed as fiction in the mid-1800s when Mary Shelley wrote her famous novel *Frankenstein*. Hopefully, the wait until functional limb regeneration is a clinical reality will be shorter.

References

Adams, D. S. (2019). What Is Bioelectricity? *Bioelectricity*, *1*(1), 3–4. https://doi.org/10.1089/bioe.2019.0005

Arenas Gómez, C. M., & Echeverri, K. (2021). Salamanders: The molecular basis of tissue regeneration and its relevance to human disease. *Current Topics in Developmental Biology*, *145*, 235–275. https://doi.org/10.1016/bs.ctdb.2020.11.009

Chicago, U. (2023, May 4). *Can bioelectricity regrow limbs and organs?* Futurity. https://www.futurity.org/bioelectricity-regrow-limbs-organs-2914742-2/

Davis, M. (2023, October 23). *TMR Improves Functionality After Upper Limb Amputation*. Physician Resource. https://uvaphysicianresource.com/tmr-prosthetics-improves-functionality/

Fell, A. (2018, November 7). *Exploring the Role of Redox and Bioelectric Players in Tissue Regeneration*. College of Biological Sciences. https://biology.ucdavis.edu/news/exploring-role-redox-and-bioelectric-players-tissue-regeneration

Godwin, J. (2022, February 28). *Regeneration of human limbs and organs – science fact or science fiction? – MDI Biological Laboratory*. Mdibl.org. https://mdibl.org/in-the-media/regeneration-of-human-limbs-and-organs-science-fact-or-science-fiction/

Joven, A., Elewa, A., & Simon, A. (2019). Model systems for regeneration: salamanders. *Development*, *146*(14), dev167700. https://doi.org/10.1242/dev.167700

Kückelhaus, M., Rothoeft, T., Teig, N., Jacobsen, F., Lehnhardt, M., Pellegrini, G., De Luca, M., & Hirsch, T. (2018). 355 Regeneration of the Entire Human Epidermis by Transgenic

Epidermal Stem Cell Transplants and its Implications for treating Burns. *Journal of Burn Care & Research*, *39*(suppl_1), S148–S148. https://doi.org/10.1093/jbcr/iry006.277

Levin, M. (n.d.). *How bioelectricity could regrow limbs and organs: Big Brains podcast with Michael Levin | University of Chicago News*. News.uchicago.edu. Retrieved October 10, 2023, from https://news.uchicago.edu/how-bioelectricity-could-regrow-limbs-and-organs

Levin, M. (2009). Bioelectric mechanisms in regeneration: Unique aspects and future perspectives. *Seminars in Cell & Developmental Biology*, *20*(5), 543–556. https://doi.org/10.1016/j.semcdb.2009.04.013

Madhusoodanan, J. (2018, June 13). *Bioelectricity's Potential*. Www.pbs.org. https://www.pbs.org/wgbh/nova/article/bioelectric-potential/

Matters, T. (2022, August 10). *Regenerating a new limb or, you know, the entire body*. American Chemical Society. https://www.acs.org/pressroom/tiny-matters/regenerating-a-new-limb-or-you-know-entire-body.html

Medicine, N. (2023, August). *What Is Targeted Muscle Reinnervation (TMR)?* Northwestern Medicine. https://www.nm.org/healthbeat/medical-advances/new-therapies-and-drug-trials/What-Is-Targeted-Muscle-Reinnervation-TMR-Infographic

Montagna, W. (2019). Human skin | anatomy. In *Encyclopædia Britannica*. https://www.britannica.com/science/human-skin

Rawan ElAbd, Dow, T., Sinan Jabori, Becher Alhallabi, Lin, S. J., & Dowlatshahi, S. (2023). Pain and Functional Outcomes Following Targeted Muscle Reinnervation: A Systematic

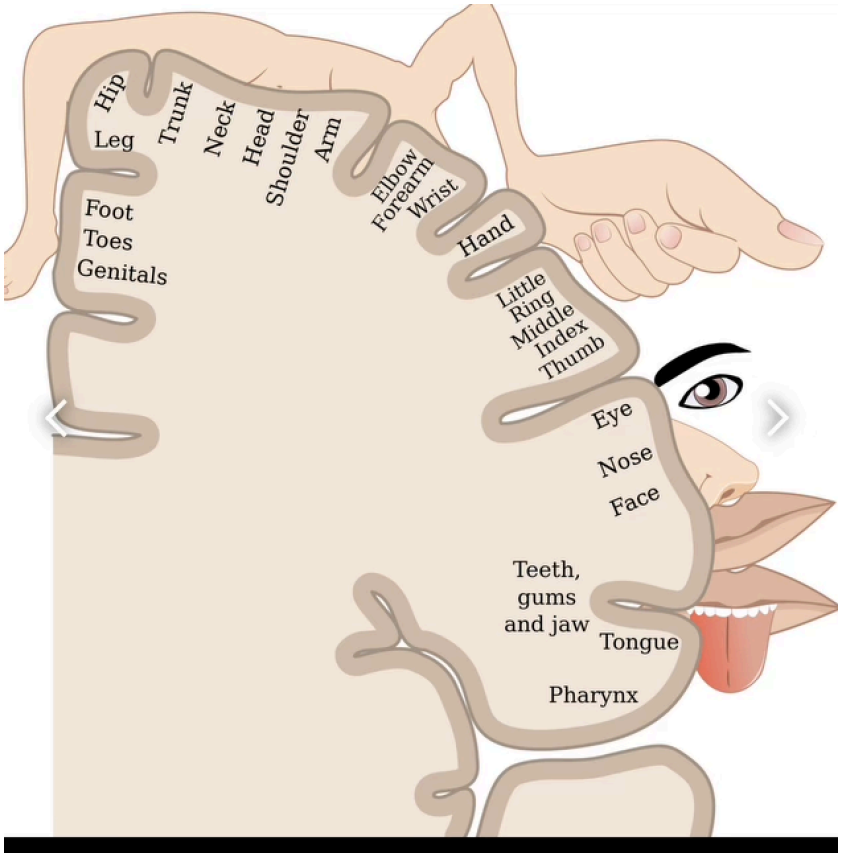
Review. *Plastic and Reconstructive Surgery*, *Publish Ahead of Print*. https://doi.org/10.1097/prs.0000000000010598

Takeo, M., Lee, W., & Ito, M. (2015). Wound Healing and Skin Regeneration. *Cold Spring Harbor Perspectives in Medicine*, *5*(1). https://doi.org/10.1101/cshperspect.a023267

Thorp, L. (2022, June 22). *Intelligent Bioenergetics - Bioelectricity: How Our Cells Talk to Each Other*. Intelligent Bioenergetics. https://intellbio.com/bioelectricity-how-our-cells-talk-to-each-other/

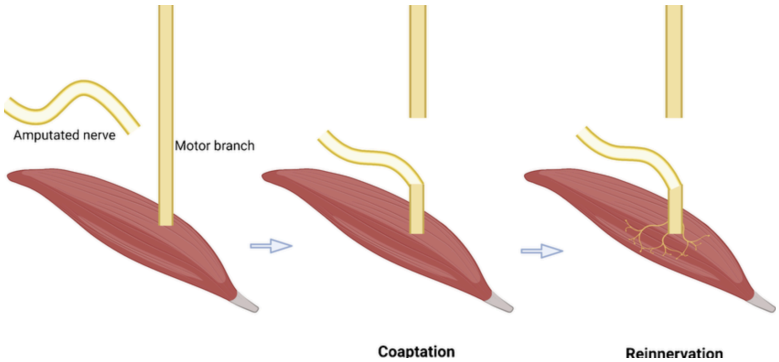
\

Appendix



Appendix 1 - Sensory and Motor Homunculus of the human body

-https://www.britannica.com/science/homunculus-biology



Appendix 3 - Schematic overview of the key steps of targeted muscle reinnervation (TMR) - Researchgate.net

Epigenetic Pathogenesis of Type 2 Diabetes and Current Therapeutic Epigenetic Modulators

Liam French '25

Biology

It was once thought that one's genetic code was the sole determiner of cellular phenotype. It is now known that epigenetics, defined as a series of heritable changes in gene expression that occur without a change in DNA sequence, also contributes to phenotype and cellular function. Epigenetic modifications, such as DNA methylation and histone acetylation, can be triggered by environmental stimuli, alter the expression of genes, and thus change cellular phenotype (Suárez et al., 2023). DNA methylation condenses gene sequences into heterochromatin, suppressing gene expression. Histone acetylation unwinds DNA into euchromatin, increasing gene expression. As a result, epigenetic modifications help determine what genes are transcribed and, thus, the resulting cellular phenotype (Ling, Bacos, Rönn, 2022).

Translational researchers have shown the broad impact epigenetics has on human health and disease. Epigenetics allows for cells with the same genetic code to differentiate into different cell types with specialized functions. Epigenetic modifications have also shown to be causative in the development of cancer, autoimmune diseases, neurologic conditions, and metabolic diseases (Kaimala et al., 2021). These adverse health risks can be inherited from the epigenome of one's parents, predisposing one to a disease or disorder. Type 2 diabetes (T2D) is an example

of a metabolic disease in which epigenetics contributes to its pathogenesis. The scope of this paper focuses on the epigenetic alterations that lead to the development of T2D and its complications, as well as current potential therapeutic epigenetic modulators for T2D.

Insulin is a hormone that regulates blood glucose levels by enhancing glucose uptake in target tissues. Pancreatic islet beta cells (β cells) produce and secrete insulin in response to regulating blood glucose levels. In T2D, there is a decrease in β cell weight, impaired insulin secretion, and insulin resistance in targets such as the liver, muscle, and adipose cells resulting from impaired β cell function (Kaimala et. al, 2021). These histologic changes lead to hyperglycemia (chronically elevated blood glucose levels), as a result of decreased insulin secretion and insulin resistance (Ling, Rönn 2019). Chronic hyperglycemia leads to microvascular complications, most prominently in the kidneys, eyes, peripheral nervous system, and macrovascular complications in large arteries. T2D affects approximately 11% of the world's population, and it is estimated to lead to 4 million deaths annually worldwide (Kowluru, Mohammad, 2022). The American Diabetes Association estimated that the total annual cost of diabetes care in the US accounts for 1 out of every 4 health care dollars spent (Parker et al., 2022). The cause of T2D is multifactorial. Unhealthy diets and inactive lifestyles are known risk factors for T2D (Ling, 2020). Additionally, it is now known that epigenetic changes in the insulin-producing pancreatic beta cells affect transcription of genes regulating insulin production and section (Marushchak, Krynytska, 2021).

Epigenetic modifications were first discovered to have a role in the pathogenesis of T2D in 2008 (Ling, Rönn, 2019). Since that time, a growing number of researchers have contributed to the current understanding of how epigenetic alterations contribute to the development of T2D. Researchers at Lund University Diabetes Centre in Malmö, Sweden found that β cells in

individuals with T2D, compared to nondiabetic controls, had increased DNA methylation in key insulin regulating genes, PDX1 and INS, causing decreased insulin secretion (Ling, Bacos, Rönn, 2022). Pancreatic and duodenal homeobox-1 (PDX1), a β cell transcription factor, encodes key regulators of β cell development and function (Kaimala et al., 2022). Hypermethylation of this gene found in patients with T2D results in decreased expression and abnormalities in insulin regulation. INS is the gene that encodes the insulin hormone. When INS is differentially methylated at 4 CpG sites, its expression and resulting insulin production are repressed (Kowluru, Mohammad, 2022). As INS regulates glucose homeostasis, its differential methylation may result in β cell failure, mitochondrial failure, microvascular, and macrovascular complications (Kaimala et al., 2021). Differential methylation of the growth factor receptor bound protein 10 (GRB10) gene, which regulates insulin signaling, and the potassium voltage-gated channel subfamily Q member 1 (KCNQ1) gene, which encodes potassium channels that regulate insulin secretion, are associated with increased risk of T2D (Dhawan, Natarajan, 2020).

Epigenetic modifications can also contribute to complications associated with T2D. Diabetic retinopathy, which can cause vision abnormalities and even blindness, is a microvascular disease that is caused by the bathing of the retina in a high glucose solution. High glucose concentrations increase the binding of DNA methyltransferase (DNMT) at retinal mitochondrial DNA (mtDNA) (Kowluru, Mohammad, 2022). The hypermethylation of mtDNA results in a faulty electron transport chain and leaky mitochondrial membrane. This, in turn, reduces the mitochondria's efficiency of producing ATP for the highly metabolically active retinal cell, resulting in its dysfunction.

Diabetic nephropathy is the most common cause of acquired end-stage kidney disease. A 2022 publication on epigenetic modifications of diabetes reported that diabetic patients show significantly higher global DNA methylation levels in blood mononuclear cells compared to healthy controls (Kowluru, Mohammad, 2022). Mononuclear cells are the predominant cell type in the immune system and play an important role in inflammation, cell proliferation, and apoptosis. Their altered methylation patterns appear to be causative in cellular injury to the kidneys. Additional studies report differential methylation in the gene Unc-13 Homolog B (UNC13B) induces renal cell apoptosis under hyperglycemic conditions, resulting in renal failure (National Library of Medicine, 2024).

Atherosclerosis is a macrovascular complication of T2D characterized by the buildup of atheromatous plaque in large arteries. Research supports that chronic hyperglycemia disrupts DNA methylation patterns in primary vascular endothelial cells, disrupting the expression of the RELA gene which is responsible for regulating inflammatory gene products. Abnormal regulation of inflammatory signals causes mononuclear cells to attack the vascular endothelium, contributing to plaque development (Kowluru, Mohammad, 2022).

The known changes in DNA methylation profiles of certain genes are also helping researchers develop novel therapeutics for T2D to alter methylation patterns, which improve β cell insulin production and secretion, and insulin sensitivity. Research has focused on the use of histone acetyltransferase (HDAC) inhibitors, epigenetic biomarkers, and CRISPR-Cas9 technology to counteract the altered methylation patterns that predispose one to T2D.

HDAC inhibitors have been associated with improved glucose homeostasis by improving β cell proliferation and function and decreasing β cell apoptosis through alterations of chromatin structure (Kaimala et al., 2022). HDAC inhibitors have also been shown to control insulin

receptor substrate (IRS) dependent insulin signaling by repressing the acetylation of an IRS promoter, resulting in the repression of dysregulated gene expression (Marushchak, Krynytska, 2021).

There has been ongoing research in the development of biomarkers for T2D. The ideal blood-based biomarker would be used to detect disease pathogenesis and its progression (Slieker, et al., 2023). Cell-free DNA in peripheral blood shows promise of being used as a biomarker to detect specific epigenetic changes in disease conditions (Dhawan, Natarajan, 2020). Through blood based biomarkers, scientists may be able to predict risk for T2D and employ therapeutic responders to help guide choice of treatment for optimal care.

One such therapeutic responder, the CRISPR/Cas9 system, has been employed to precisely target epigenetic patterns, such as altered DNA methylation or demethylation or chromatin structure, within specific genomic regions (Dhawan Natarajan, 2020). CRISPR/Cas9 is used to drive proliferation of β cells by altering the methylation pattern of cell-cycle inhibitor gene CDKN1C. Furthermore, CRISPR/Cas9 technology has been used to demethylate and increase expression of PDX1, a β cell transcription factor that encodes regulators of β cell development and function.

Environmental factors and the human epigenome, an extensive and interconnected web of genes and epigenetic modifiers, contribute to the pathogenesis of T2D. Growing research continues to reveal how differential methylation of various insulin regulatory genes alters glucose homeostasis and contributes to vascular complications. T2D impacts approximately 11% of the world's population and is a leading cause of healthcare expenditure, morbidity, and mortality worldwide (Parker et al., 2023). Fortunately, research on potential epigenetic modulators and their use as therapeutic agents against T2D offers hope. Epigenetic modulators

such as HDAC inhibitors and biomarkers are able to target differentially methylated genes that arise in T2D offering avenues for personalized medicine and targeted interventions.

References

Dhawan, S., & Natarajan, R. (2019). Epigenetics and type 2 diabetes risk. *Current Diabetes Reports*, *19*(8). https://doi.org/10.1007/s11892-019-1168-8

Domingo-Relloso, A., Gribble, M. O., Riffo-Campos, A. L., Haack, K., Cole, S. A., Tellez-Plaza, M., Umans, J. G., Fretts, A. M., Zhang, Y., Fallin, M. D., Navas-Acien, A., & Everson, T. M. (2022). Epigenetics of type 2 diabetes and diabetes-related outcomes in the strong heart study. *Clinical Epigenetics*, *14*(1). https://doi.org/10.1186/s13148-022-01392-7

Ethiopian journal of health sciences. (n.d.).

Marushchak, M., & Krynytska, I. (2021). Insulin Receptor Substrate 1 Gene and Glucose Metabolism Characteristics in Type 2 Diabetes Mellitus with Comorbidities. *Ethiopian journal of health sciences*, *31*(5), 1001–1010. https://doi.org/10.4314/ejhs.v31i5.12

Kaimala, S., Kumar, C. A., Allouh, M. Z., Ansari, S. A., & Emerald, B. S. (2022). Epigenetic modifications in pancreas development, diabetes, and therapeutics. *Medicinal Research Reviews*, *42*(3), 1343-1371. https://doi.org/10.1002/med.21878

Kowluru, R. A., & Mohammad, G. (2022). Epigenetic modifications in diabetes. *Metabolism*, *126*, 154920. https://doi.org/10.1016/j.metabol.2021.154920

Ling, C., Bacos, K., & Rönn, T. (2022). Epigenetics of type 2 diabetes mellitus and weight change — a tool for precision medicine? *Nature Reviews Endocrinology*, *18*(7), 433-448. https://doi.org/10.1038/s41574-022-00671-w

Ling, C., & Rönn, T. (2019). Epigenetics in human obesity and type 2 diabetes. *Cell Metabolism*, *29*(5), 1028-1044. https://doi.org/10.1016/j.cmet.2019.03.009

National Library of Medicine. (2024, April 11). UNC13B Unc-13 homolog B [Homo sapiens (human)] [Fact sheet]. *National Library of Medicine*. Retrieved May 5, 2024, from https://www.ncbi.nlm.nih.gov/gene/10497

nature communications. (n.d.).

Slieker, R.C., Donnelly, L.A., Akalestou, E. *et al.* Identification of biomarkers for glycaemic deterioration in type 2 diabetes. *Nat Commun* **14**, 2533 (2023). https://doi.org/10.1038/s41467-023-38148-7

Parker, E. D., Lin, J., Mahoney, T., Ume, N., Yang, G., Gabbay, R. A., ElSayed, N. A., & Bannuru, R. R. (2023). Economic costs of diabetes in the U.S. in 2022. *Diabetes Care*, *47*(1), 26-43. https://doi.org/10.2337/dci23-0085

Suárez, R., Chapela, S. P., Álvarez-Córdova, L., Bautista-Valarezo, E., Sarmiento-Andrade, Y., Verde, L., Frias-Toral, E., & Sarno, G. (2023). Epigenetics in obesity and diabetes mellitus: New insights. *Nutrients*, *15*(4), 811. https://doi.org/10.3390/nu15040811

Ye, S., Zhang, M., Tang, S. C. W., Li, B., & Chen, W. (2024). PGC1-α in diabetic kidney disease: Unraveling renoprotection and molecular mechanisms. *Molecular Biology Reports*, *51*(1). https://doi.org/10.1007/s11033-024-09232-y

Colophon

Front cover title is self-designed; front cover subtext is self-designed; spine text is Futura Condensed Medium 10 pt; back cover text is Futura Condensed Medium 20 pt; table of contents title is Montserrat 24 pt; table of contents entry, categories, and authors are Georgia 14pt and 12 pt; article titles are Montserrat 24 pt; article section titles are Montserrat 18 pt; article subsection titles are Montserrat 14 pt; article authors and article categories are Open Sans 12 pt; article headers are Open Sans 10 pt; article footers (page numbers) are Open Sans 12 pt; article body texts are Georgia 12 pt; article reference texts are Georgia 12 pt. The softwares utilized are Google Docs and Adobe Illustrator 27.5.

If you are interested in contributing to the 2023-2024 edition of Newton's Notebook: The Haverford School STEM Journal, please contact the Issue XVIII Editors-in-Chief: Milan Varma '25, Kevin Li '25, & Nicholas Lu '25

Dedicated to the class of 2024, whose selfless pursuit of excellence continues to inspire the next generation of mathematicians, scientists, and innovators.