Mastering the Data Science Landscape:

A Journey through Syracuse's Applied Data Science Program

Syracuse University

Winter 2023

M.S. Applied Data Science

Sintia Stabel

SUID 467548556

GitHub - https://github.com/SintiaStabel/SyracuseCapstonePortfolio

Table of Contents

Introduction:

Data Science is a dynamic field that empowers organizations across diverse industries to extract actionable insights, inform decision-making, and foresee future trends through the analysis of data. In this ever-evolving landscape, Artificial Intelligence (AI) stands as a transformative force, reshaping the fabric of our world at an unprecedented pace. As data scientists, we not only confront the challenges presented by this rapid evolution but also seize the opportunities it offers to harness the power of data science and analytics for the greater good.

This portfolio serves as a comprehensive testament to how the learning objectives outlined in the Master's in Applied Data Science program at Syracuse University have not only been met but thoroughly mastered. In the forthcoming sections, I will meticulously present compelling evidence and practical examples drawn from the diverse array of courses I have completed. These examples will illuminate not only my acquisition of essential data science skills but also my ability to refine and apply these skills across an extensive spectrum of problems and real-world scenarios.

Having journeyed through this program, I stand confident in my proficiency to adeptly collect, manage, analyze, and interpret data. I recognize that data is the linchpin of my problem-solving toolkit, enabling me to address a wide gamut of business inquiries and real-world challenges. Throughout this portfolio, my aim is to vividly demonstrate my capacity to leverage data science techniques for strategic decision-making, prominently featuring predictive modeling as a powerful tool in my arsenal. Moreover, I will exemplify my adeptness in conveying my insights through impactful data visualizations, coupled with the flexibility to tailor my communication to suit diverse audiences.

In the ensuing sections, I will provide compelling evidence of my extensive deployment of quantitative analytics. This encompasses a rich repertoire of Statistical, Machine Learning, Deep Learning, and Natural Language Processing (NLP) techniques. These methodologies have been applied with precision to tackle a range of tasks, including predictive analytics, risk assessment, and classification. Furthermore, I will underscore my proficiency in wielding a variety of big data management and automation tools, such as Tableau, PowerBI, in addition to commanding programming languages and libraries like SQL, R, and Python. These include essential tools like Pandas, NumPy, Matplotlib, Scikit-learn, NLTK, TensorFlow, PyTorch, and Spark. This portfolio thus serves as a testament to the breadth and depth of my technical expertise, cultivated throughout the duration of this program.

The field of data science continues to evolve, presenting us with challenges and opportunities alike. As data experts, we are poised to navigate this exciting terrain, harnessing the power of AI, and contributing to the betterment of society through data-driven insights and innovations.

Section 1: Data Collection and Management:

IST-652 Scripting for Data Science

Data collection and management encompass a wide array of techniques and strategies. In the IST-652 Scripting for Data Science course, I delved into various methods for effectively managing, analyzing, and transforming both structured and unstructured data. This learning experience equipped me with practical skills that have proven invaluable in my data science journey.

One noteworthy project from this course involved scraping Twitter (now Twitter X) messages and saving the raw JSON data in a MongoDB instance, demonstrating the versatility of data collection methods. Using Python's powerful Pandas library, I efficiently transformed the raw JSON file into a clean, structured format. This transformation laid the foundation for diverse analytical approaches, ranging from descriptive to predictive methods.

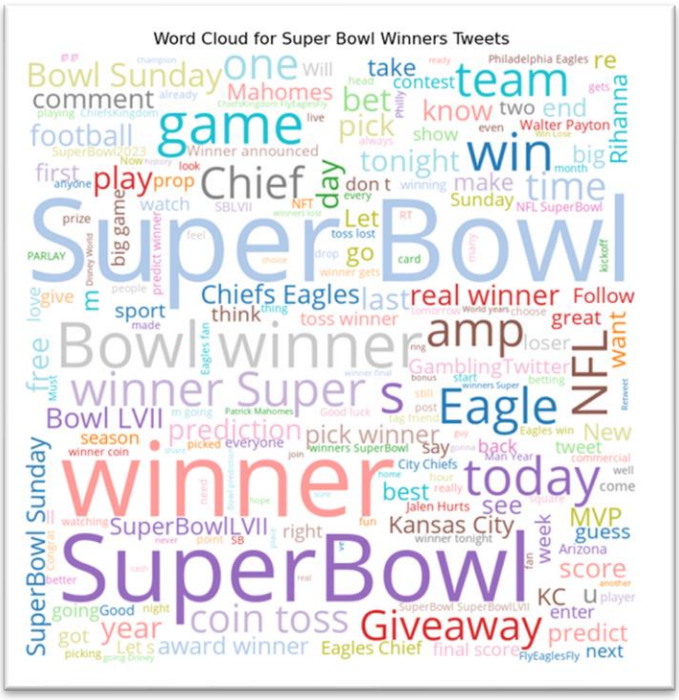
An intriguing experiment within this project involved utilizing sentiment analysis to assess whether tweets related to Super Bowl 2023 could be predictive of the game's outcome. To facilitate this analysis, I leveraged the Social Network Site (SNS) platform to collect a substantial dataset of 4,000 public tweets on the day of Super Bowl 2023.

The data collection process was two-fold. Firstly, I gathered the initial 2,000 tweets using the query "Philadelphia Eagles." Subsequently, I collected an additional set of tweets using the keywords "Super Bowl 2023." This comprehensive dataset formed the basis for subsequent analyses.

With data cleaning and preprocessing completed, the Super Bowl dataset revealed deeper insights, which I showcased through engaging WordCloud visualizations. A WordCloud offers a succinct visual representation of the dataset's most prominent words, with word size corresponding to its frequency.

Figure 1 presents a WordCloud focusing on distinct topics or labels identified within the dataset. Notably, the WordCloud for "super bowl winner" tweets exhibited a neutral sentiment, with team names appearing as smaller words. Conversely, larger and more frequent words included "Super Bowl," "winner," "Bowl Winner," "Eagle," "Chief," "Giveaway," "NFL," "team," "game," "Rihanna," among others. These results underscored the absence of a clear preference for a game winner in the sentiment expressed within the tweets.

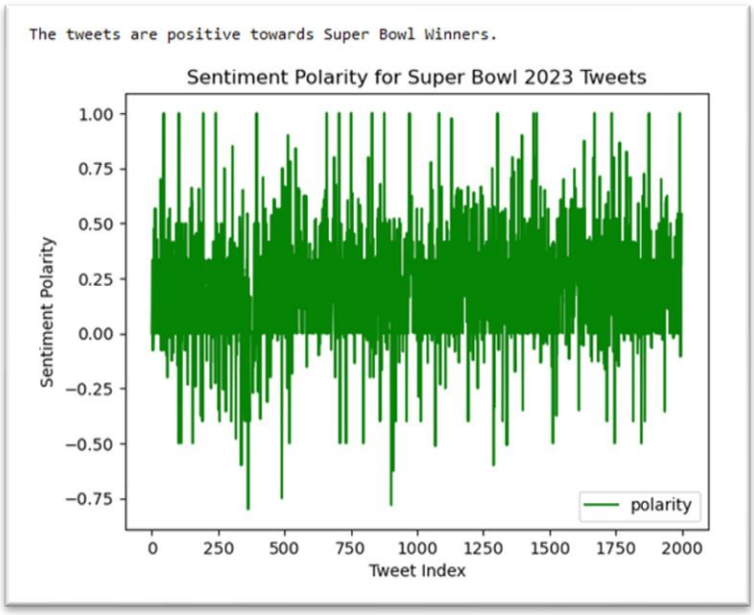
Figure 1 – Word Cloud of “Super Bowl Winner” Tweets



In my analysis of Super Bowl 2023 tweets, I harnessed the power of the TextBlob Python library as the foundation for sentiment analysis. This library provided a valuable baseline for measuring the polarity of sentiments within the dataset.

Notably, when I examined the sentiment analysis results from the tweets related to "Super Bowl 2023," a compelling trend emerged. As evidenced by the chart shown in Figure2, the frequency of positive polarity sentiments significantly outweighed the negative frequencies. These findings underscored a notable absence of a decisive preference for a Super Bowl game winner within the sentiments expressed within the tweets. In essence, the sentiment analysis revealed a nuanced and multifaceted perspective from the Twitter community, rather than a resounding endorsement of one team over the other.

Figure 2 – Sentiment Polarity of “Super Bowl 2023” Tweets



This project stands as a testament to my proficiency in collecting, managing, and analyzing data with precision and rigor. It not only showcases my ability to harness data effectively but also highlights the intricate and multidimensional nature of data science applications. By navigating the complexities of sentiment analysis within this real-world context, I have demonstrated my capacity to uncover valuable insights and contribute to the evolving field of data science.

IST-736 Text Mining

During my enrollment in IST-736 Text Mining, I acquired proficiency in web data retrieval techniques. This included both contemporary approaches such as utilizing APIs and conventional web scraping methodologies employing libraries such as Beautiful Soup. This diverse skill set allowed me to collect a corpus of news articles centered around topics of interest, thereby broadening my understanding of data collection and management in the context of text mining.

For this project, I employed these techniques to gather news messages related to specific themes—namely, "wildfire," "heatwaves," and "droughts." During severe weather events, emergency management authorities require up-to-date and comprehensive information to assess the situation and allocate resources effectively. By categorizing recent news articles, these authorities can gain valuable insights to prioritize response efforts, implement evacuation plans, and deploy resources where they are most needed.

The data was sourced from NewsAPI, a versatile REST API renowned for providing access to a vast repository of current and historical news articles from over 80,000 sources worldwide. Following a well-defined query, I amassed a dataset encompassing 300 rows and 5 columns, converting the API's JSON output into a structured CSV format.

Figure 3 – Raw data in JSON format from NewsAPI


```
<Response [200]>
{'status': 'ok', 'totalResults': 3381, 'articles': [{'source': {'id': None, 'name': 'Lifehacker.com'}, 'author': 'Eliza
beth Yuko', 'title': 'Use This Phone Number to Find a Cooling Center Near You', 'description': 'We're not even a full m
onth into summer, and much of the United States has already dealt with record-setting heat—not to mention the intermitt
ent wildfire smoke from Canada. Between climate change and El Niño, this type of weather is expected to get worse du
r…', 'url': 'https://lifehacker.com/use-this-phone-number-to-find-a-cooling-center-near-you-1850641814', 'urlToImage':
'https://i.kinja-img.com/gawker-media/image/upload/c_fill,f_auto,fl_progressive,g_center,h_675,pg_1,q_80,w_1200/c3be802
7600a57f72bc8352f391747a0.jpg', 'publishedAt': '2023-07-15T15:00:00Z', 'content': 'Were not even a full month into summ
er, and much of the United States has already dealt with record-setting heatnot to mention the intermittent wildfire sm
oke from Canada. Between climate change and … [+1728 chars]'}, {'source': {'id': 'bbc-news', 'name': 'BBC News'}, 'auth
or': None, 'title': 'Italy: Drone spots suspected wildfire arsonist', 'description': 'A man is detained in Italy after
he was seen at the scene of a wildfire in Calabria.', 'url': 'https://www.bbc.co.uk/news/av/world-europe-66343164', 'ur
lToImage': 'https://ichef.bbci.co.uk/news/1024/branded_news/160AF/production/_130578209_p0g3kd5k.jpg', 'publishedAt': '
2023-07-28T17:21:56Z', 'content': 'A man has been detained after he was spotted by a drone near a recently started wild
```

Subsequently, rigorous data cleaning and preparation procedures were executed. These steps were essential to ensure the dataset's suitability for classification analysis. Notably, the "Headline" column underwent text preprocessing, which was pivotal for accurate modeling. The following transformations were applied:

- Special Characters and Punctuation Removal: Utilizing regular expressions (regex), non-alphanumeric characters, including numerical digits, were expunged from the text.
- Text to Lowercase Conversion: All text was transformed to lowercase to standardize text representation, mitigating case sensitivity issues during analysis.
- Tokenization: Employing NLTK's word_tokenize function, the text was tokenized, fragmenting it into individual words or tokens.
- Stopwords Removal: Common but semantically insignificant words, such as “the”, “and”, and “is” were pruned from the text using NLTK's English stopwords set. This step retained only meaningful words.
- NaN Value Elimination: Rows containing missing values were expunged, as they could impede the model's learning process.

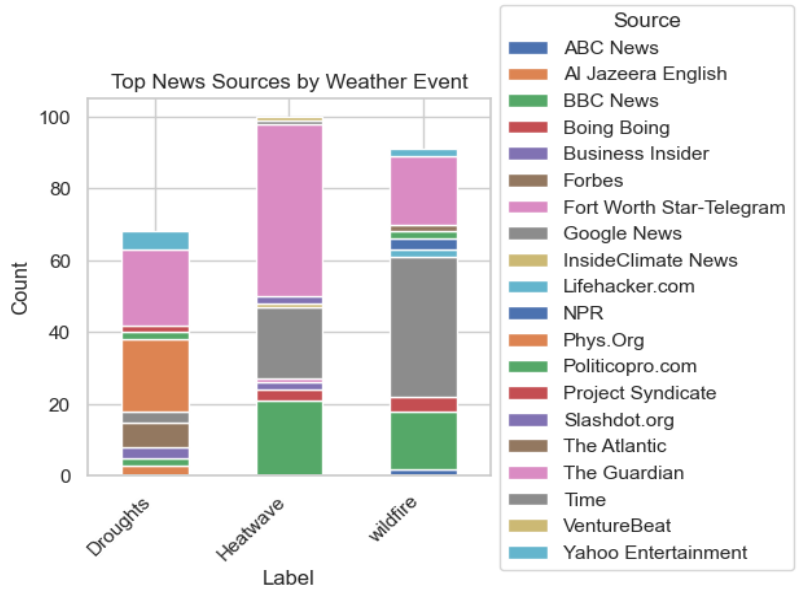
Figure 4 - New Column with Preprocessed and Clean Data: “df_processed”

	LABEL	Date	Source	Title	Headline	df_processed
295	Droughts	7/20/2023	Forbes	Rivers As Weapons Of War And Warnings Of Clima...	millennia armies have used rivers weapons Whil...	millennia armies used rivers weapons intention...
296	Droughts	7/21/2023	Science Daily	Climate science is catching up to climate chan...	Africa climate change impacts experienced extr...	africa climate change impacts experienced extr...
297	Droughts	7/15/2023	Home.blog	Loops and Arcs	Here tools been using lately better understand...	tools using lately better understand functiona...
298	Droughts	7/24/2023	Phys.Org	Algeria fires fanned by winds extreme heat kill	Wildfires raging across Algeria during blister...	wildfires raging across algeria blistering hea...
299	Droughts	8/3/2023	Phys.Org	Geostationary satellite reveals widespread mid...	western particularly Southwest experienced not...	western particularly southwest experienced not...

Beyond article snippets and titles, the dataset also encompassed the sources from which these articles originated, providing valuable context. The ensuing stacked bar chart unveiled prominent patterns in source distribution across the different weather-related topics.

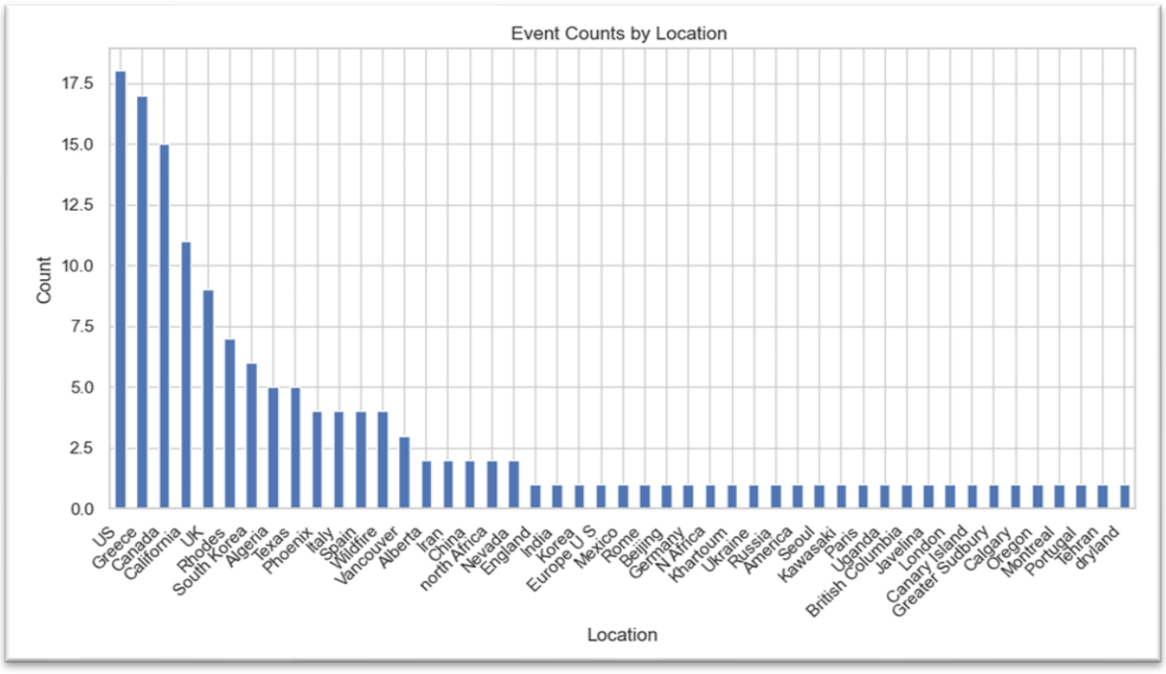
Notably, for articles concerning wildfire and heatwaves, Google News and Fort Worth Star-Telegram emerged as predominant sources, closely followed by BBC News. Conversely, drought-related articles predominantly originated from Al Jazeera English and Fort Worth Star-Telegram, with Forbes serving as a secondary source. This visualization underscored the influential prominence of specific news outlets within each distinct weather-related category.

Figure 5 - Top News Sources Distribution By Weather Event



To further explore the dataset, I leveraged spaCy, an NLP library integrated within NLTK, to systematically extract locations mentioned within the article titles. This effort aimed to provide comprehensive insights into the geographical regions most frequently associated with the unfolding weather events.

Figure 6 - Weather event count frequency by location



The aforementioned examples vividly illustrate my adeptness in harnessing Python scripts and API technology to procure data. Furthermore, they showcase my proficiency in the indispensable art of data cleaning and preparation—a skill crucial for unearthing nuanced insights that often lie concealed beneath the surface.

IST 659 Data Administration Concepts and Database Management

In the realm of data collection and management, proficiency in working with relational databases is of paramount importance. Among the various database types, the relational database format stands out as the most ubiquitous and widely adopted. My enrollment in IST 659 Data Administration Concepts and Database Management was instrumental in equipping me with a comprehensive understanding of this foundational aspect of data science.

Throughout this course, I delved deep into the intricate intricacies of data management concepts, gaining mastery over the complexities of SQL querying. The curriculum went beyond the basics and ventured into advanced SQL concepts, covering a broad spectrum of topics. I grappled with challenging concepts, including but not limited to nested queries, intricate multi-table joins, and the art of data transformation through casting functions.

Moreover, the course placed a strong emphasis on error handling through techniques like "try-catch," ensuring data integrity and reliability even in the face of unforeseen issues. Additionally, I acquired insights into indexing techniques designed to optimize the data retrieval process, ultimately making it more efficient and responsive.

Figure 7 - Query for Overall Average Seller Rating and Average User Rating

```
with avg_per_user as (
select user_firstname, user_lastname, rating_for_user_id, rating_value,
    avg(rating_value) over() as avg_overall_rating,
    avg(rating_value)over(partition by rating_for_user_id)  as avg_seller_rating
      from vb_user_ratings
     JOIN vb_users  ON rating_for_user_id = user_id
group by  rating_for_user_id, rating_value, user_firstname, user_lastname
)
select rating_for_user_id, rating_value, user_firstname, user_lastname, avg_overall_rating,avg_seller_rating,
cast(avg_seller_rating as decimal (3,2)) as average_seller_rating,
cast(avg_overall_rating as decimal (3,2)) as average_overall_rating,
avg_overall_rating - avg_seller_rating as avg_difference
    from avg_per_user
group by  rating_for_user_id, rating_value, user_firstname, user_lastname, avg_overall_rating, avg_seller_rating
Go
--
```

	rating_for_user_id	rating_value	user_firstname	user_lastname	avg_overall_rating	avg_seller_rating
1	1	4	Abby	Kuss	3	4
2	1	5	Abby	Kuss	3	4
3	3	2	Barb	Barion	3	3
4	3	5	Barb	Barion	3	3
5	4	5	Barry	DeHatchett	3	5
6	6	3	Gus	Toffwind	3	3
7	7	4	Isabelle	Gunnering	3	4

The figure above, representing a query for overall average seller rating and average user rating, is a testament to the practical application of the knowledge acquired during this course. It highlights my ability to navigate and extract meaningful insights from complex relational databases—a skill that is indispensable in the field of data science.

The depth and comprehensiveness of this class were substantiated by its culminating group project assignment, where we were tasked with applying all the knowledge we had accumulated throughout the course by crafting a database from scratch.

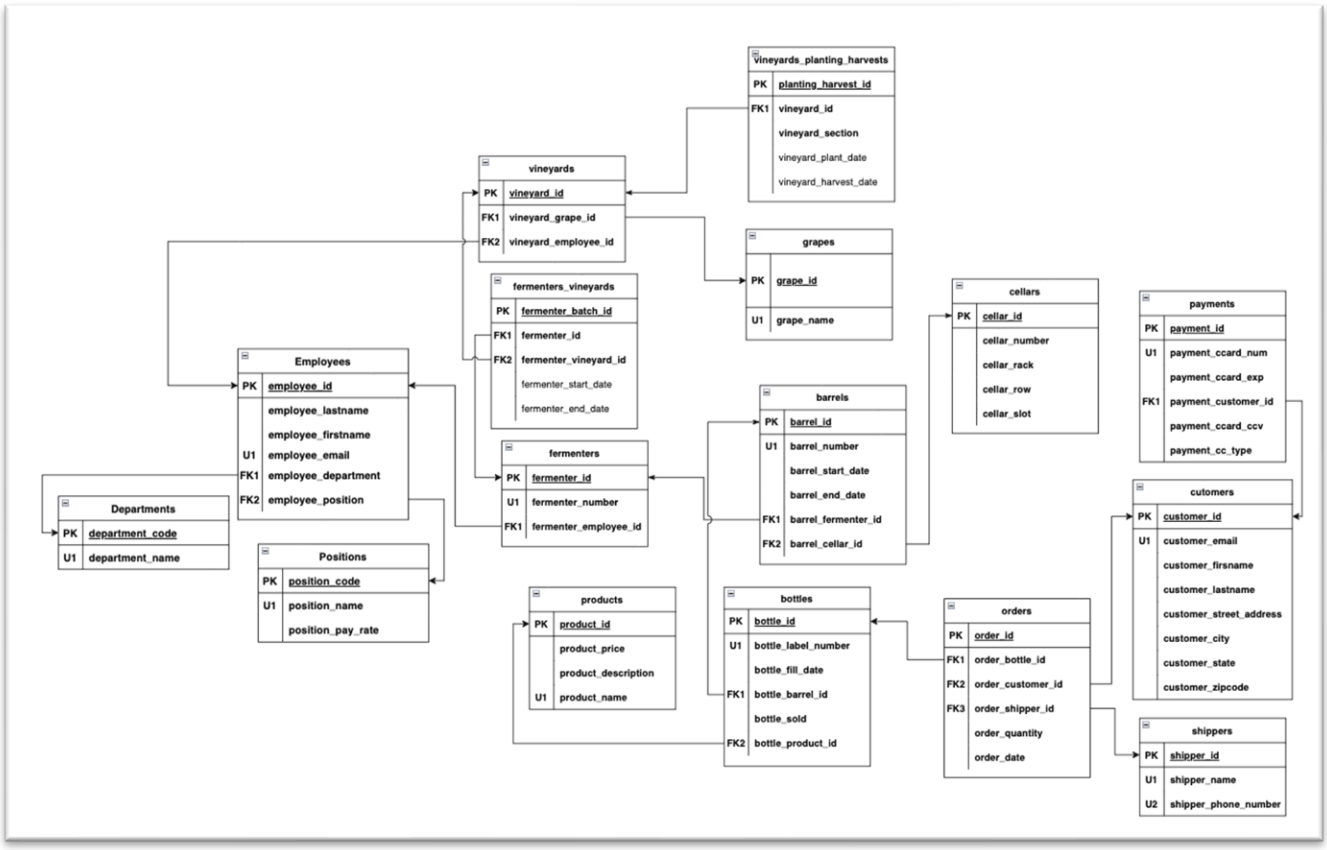
Our project entailed the creation of a fictional vineyard, aptly named "Chateau Cuse." This startup vineyard embarked on the journey of nurturing grape cultivation, wine production, and online wine sales. The database we meticulously designed revolved around three core domains of the vineyard's operations: production, sales, and human resources. Our project's scope was intricately tailored to enable tracking a grape's journey from inception through the entire production process, culminating in its transformation into a bottle of wine ready for purchase. Certain intricacies of the wine-making process were deliberately omitted to maintain project focus.

The primary objective of this database is to empower business analysts working with the vineyard to query and analyze the data comprehensively, thereby assessing the vineyard's performance and identifying avenues for enhancement. Our Database Management System (DBMS) equips these analysts with the tools to effortlessly uncover vital insights, including but not limited to:

- Identifying the top-selling products
- Quantifying orders placed within specific time frames
- Tracking real-time stock levels for each product
- Tracing each bottle's origin to its specific vineyard source
- Determining the aging duration of each product and the corresponding aging barrels
- Pinpointing bottling dates
- Calculating the duration of fermentation for each product

In essence, our database functions as a pivotal resource, enabling business analysts to access a comprehensive overview of the vineyard's operations and performance metrics. Below, you'll find the logical diagram that intricately elucidates the table relationships within our database.

Figure 8 - Logical Data model Diagram



Our sample database was developed using MySQL, featuring a total of 16 tables encompassing diverse relationship types, including one-to-one, one-to-many, and many-to-many entity relationships. The database's functionality was aptly showcased during our final presentation, where we presented statistics derived from the database through an engaging Power BI dashboard, encapsulating production, and operational insights.

Figure 9 - Sales Dashboard

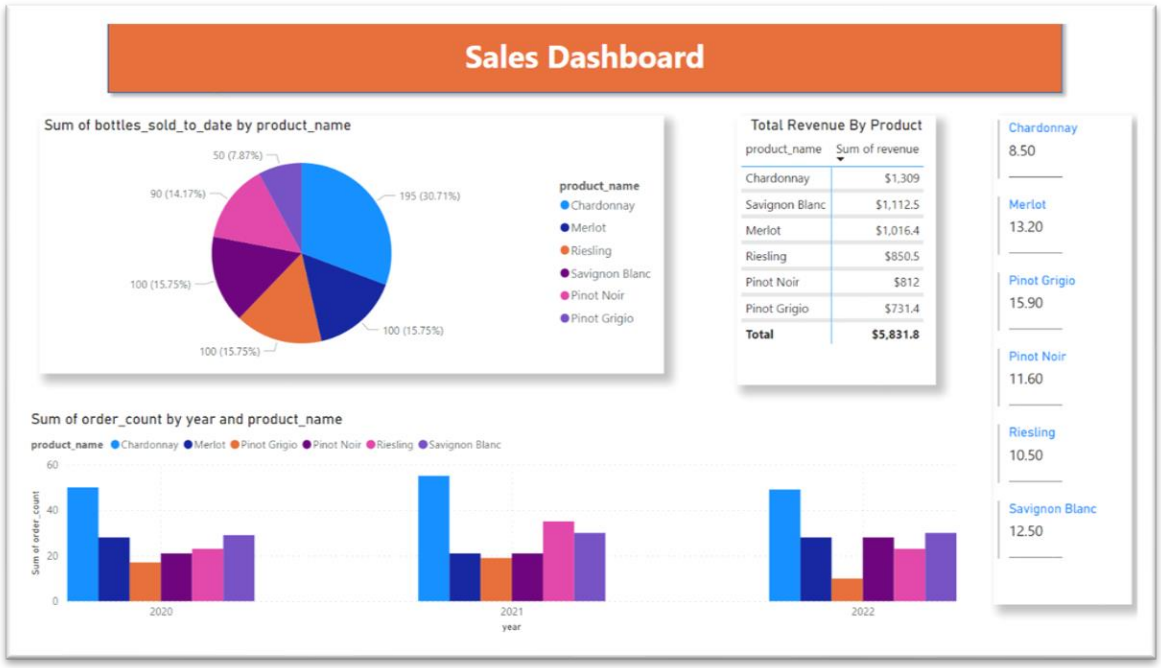
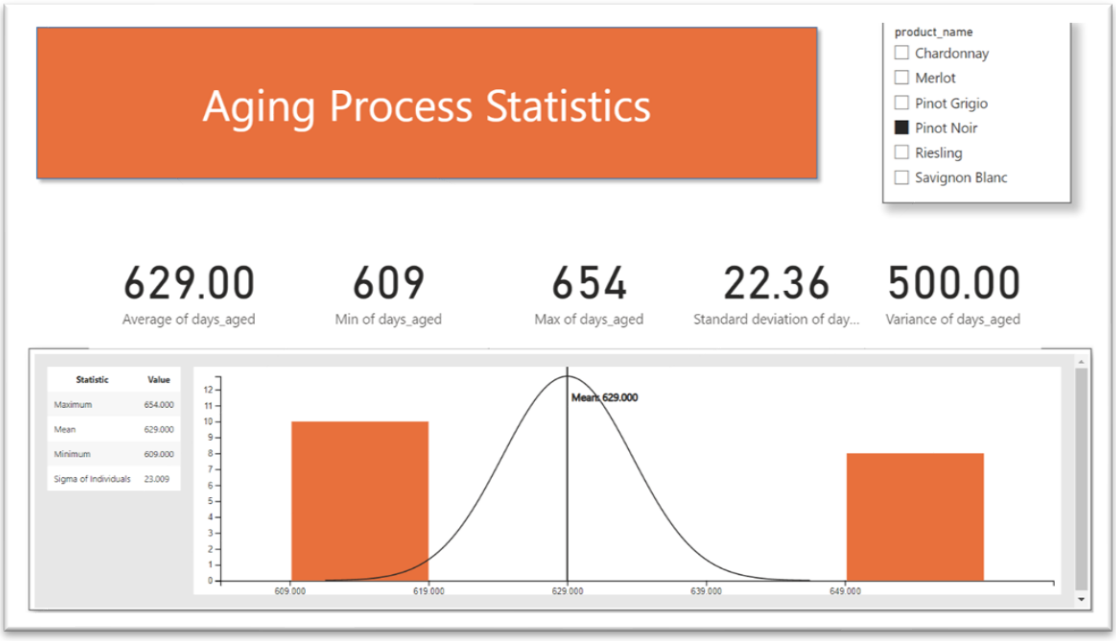


Figure 10: Aging Process Statistics



Section 2: Data Analysis and Interpretation:

IST-652 Scripting for Data Science

While each course within the applied data science master's program inherently incorporated data analysis and interpretation, a particularly illustrative instance showcasing the practical application of these proficiencies can be found in the culminating group project of the IST-652 Scripting for Data Science course. This project was meticulously designed with the aim of assisting individuals in making informed choices when it comes to selecting an optimal streaming service tailored to their individual content, genre, and quality preferences. The overarching inquiry guiding this endeavor was none other than the fundamental question: "What streaming service is best suited for me?"

To achieve this, we leveraged data from CSV files containing information on popular streaming platforms like Amazon Prime, Disney+, HBOMax, Hulu, Netflix, and Paramount+. These datasets, originally uploaded to Kaggle by Victor Soeiro in 2022, provided a snapshot of the content available on each platform in the United States as of mid-2022. The datasets, structured consistently, facilitated a comprehensive analysis. Our work highlighted not only our prowess in data manipulation and analysis but also the tangible applications of data science in helping users make informed decisions regarding their streaming preferences.

Throughout this project, we used statistical analysis and data visualization techniques to classify and recommend streaming services based on various characteristics. This endeavor underscored the practicality of data science in addressing consumer-oriented inquiries, emphasizing the role of data exploration and analysis in providing personalized recommendations for streaming service selection.

Among these informative visualizations, we presented insights into the distribution of titles across various streaming platforms, the balance between movies and TV shows available on these platforms, age certifications for each service, and the temporal distribution of release years. It's important to note that the subsequent charts represent merely a glimpse of the wide array of visualizations and interpretations generated during our thorough exploration of the dataset.

Figure 11 - Bar plot showing the number of titles belonging to each streaming service in the data frame.

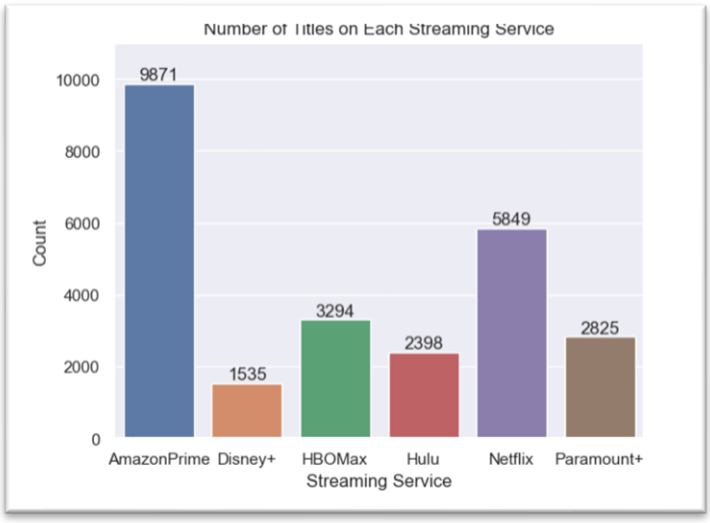


Figure 12 - Bar plot shows the number of movies and TV shows available across all the streaming services in the data frame

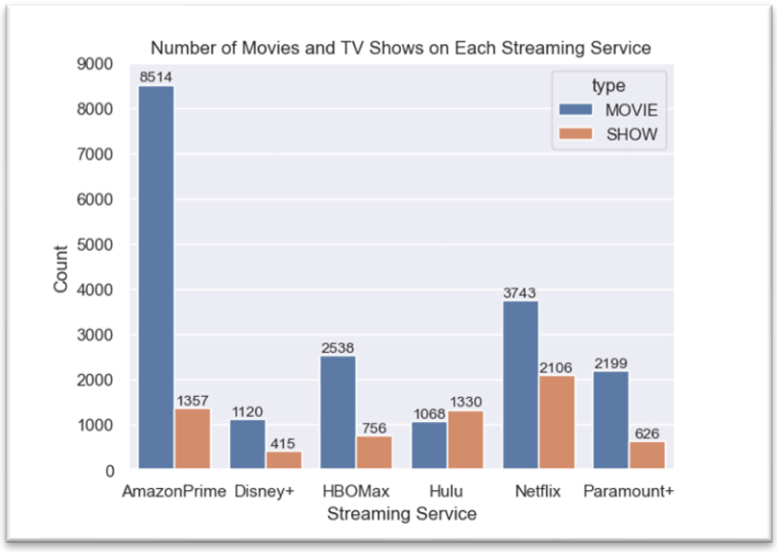


Figure 13 - Treemap showing age certification for movies

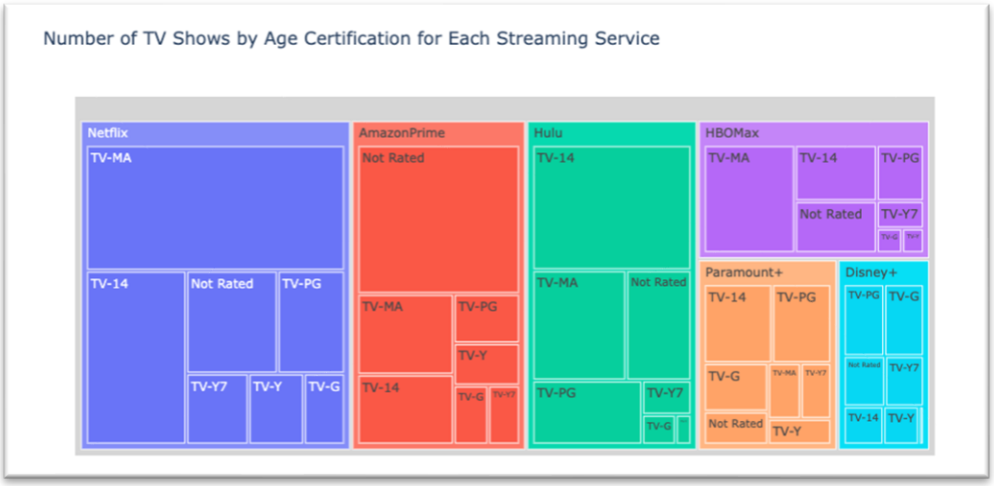
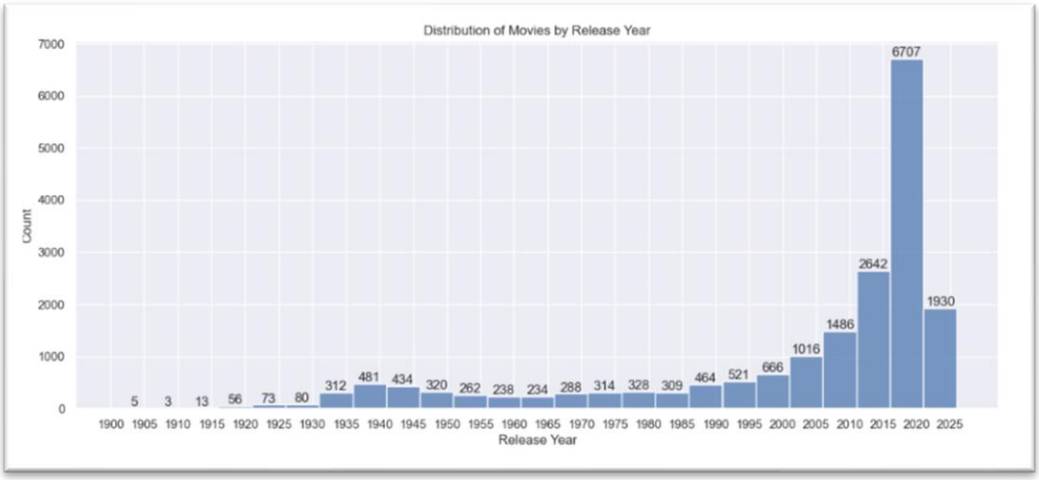


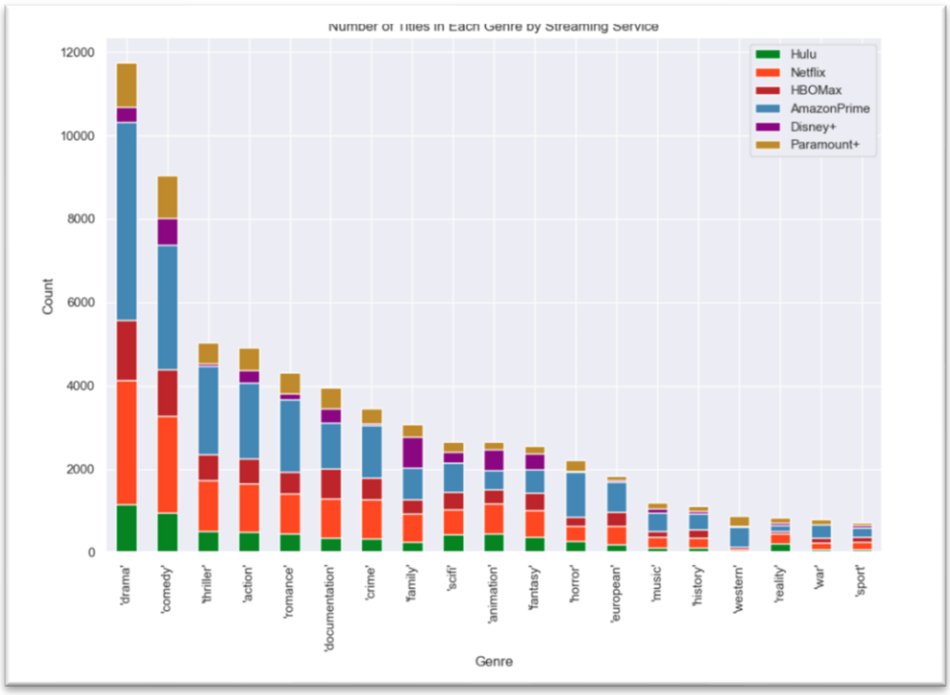
Figure 14 - Bar chart showing release year for movies



Subsequently, we set out to address a fundamental inquiry: Which streaming service offers the genres of movies and shows that align with a prospective subscriber's preferences? Recognizing that a single title could span multiple genres, we devised a custom function to tally the genres' occurrences. This function initiated an empty dictionary and an empty set, conducting an iterative sweep through each row of the input data frame. It parsed the service and genres columns, systematically appending each genre to the initialized set. For each genre and service pairing, the function scrutinized if it already existed in the dictionary. If so, it incremented the count for the pairing; if not, a new entry with a count of 1 was forged. This method assured that a genre was not counted redundantly for any given title.

The outcome was then visually represented in a stacked bar plot, where each color band in the bars depicted the distribution of titles within a specific genre available on a particular streaming platform.

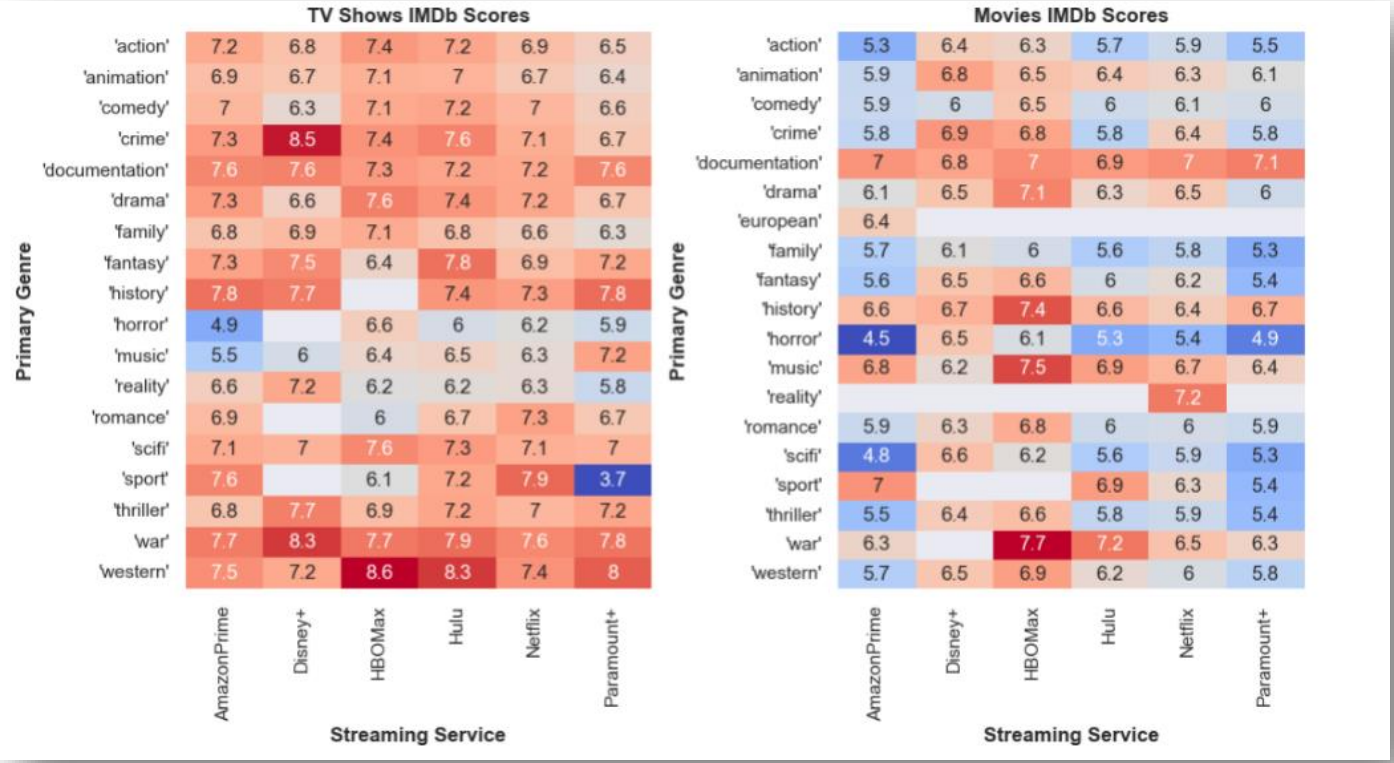
Figure 15 - Stacked bar chart showing genre distributions per service provider



Drama and comedy emerged as the two predominant genres, boasting significantly higher counts than their counterparts. Examining the color bands, they exhibited a relatively proportional distribution across each column. This equilibrium may be attributed to Amazon Prime's larger title collection compared to other streaming platforms, thus likely hosting the highest count of titles in each genre. Nevertheless, a few exceptions surfaced; notably, Disney+ held a substantial chunk of the family genre, despite being a smaller segment in other categories.

Our ensuing inquiry delved into uncovering the streaming service with the highest-rated movies and TV shows. Recognizing that viewers often gravitate toward specific genres, regardless of a title's overall rating, we segregated the data based on each title's primary genre. We designated the first genre tag in each title's genre value as its primary genre. Titles were once again categorized into movies and TV shows, acknowledging that viewer preferences may vary between the two. The ensuing heatmaps depict the mean IMDB score for each primary genre within each streaming service.

Figure 16 - Heatmaps for TV and movies by genre ad service provider.

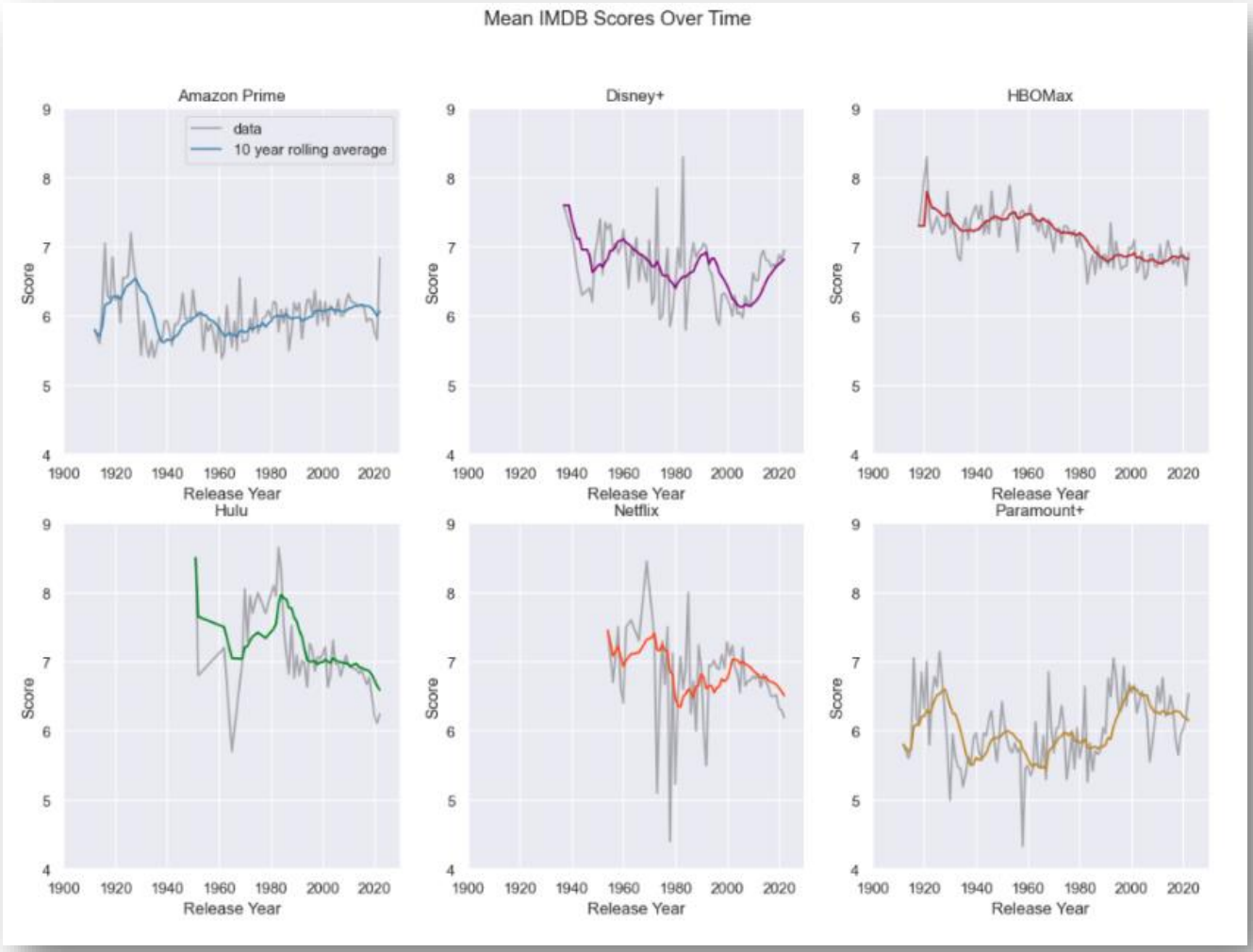


In these heatmaps, deeper shades of red signify higher average scores, while deeper blues indicate lower averages. Blank spaces without numbers signify the absence of titles in that particular primary genre on the streaming service. Perusing the heatmaps across each row permits identification of the streaming service with the highest average score in that genre. Similarly, scanning along each column unveils the highest and lowest-rated genres offered by each service. Remarkably, Western TV shows on HBOMax garnered the highest average IMDB score overall, while Paramount+ presented the lowest in sports TV shows. In the realm of movies, war films on HBOMax claimed the highest average ratings, while horror movies on Amazon Prime registered the lowest. Notably, this visualization reaffirmed that, on the whole, TV shows held higher average ratings compared to movies across these streaming services.

Analyzing the average overall IMDB scores allowed us to gauge the performance of each service provider over time. Plotting these scores against release years provided valuable insights into the evolving quality of their content, shedding light on trends of improvement or decline.

To better discern patterns of score fluctuation over time, we employed a 10-year rolling average of mean IMDB scores for all titles on each platform, superimposing it onto the line plot depicting mean IMDB scores by release year. This visualization offers a clear representation of content quality trends within the most recent decade.

Figure 17 - 10-year rolling mean IMDB scores across all titles per service, juxtaposed with the mean IMDB scores by release year.



Disney+ stands out as the sole service showing an upward trajectory in its 10-year rolling average, following a period of decline. Amazon Prime and HBOMax maintain relatively stable mean IMDB scores, hovering around 7 and 6, respectively. In contrast, Hulu and Netflix are currently experiencing a decline in scores for their most recently released content, with rolling averages dropping to the 6 to 7 range. Lastly, while Paramount+ exhibits a slight dip in ratings for its latest releases, its recent entry into the market warrants further observation to assess the performance of upcoming titles with audiences.

The project culminated in the creation of user profiles, reflecting the practical application of our analysis in guiding personalized recommendations. Here are two user profiles exemplifying tailored streaming service suggestions based on individual preferences:

User Profile 1 - Mark:

Mark is an avid viewer who seeks the best and most popular shows across various genres, always wanting to stay current with the latest buzz in the entertainment community. With ample free time each night, Mark is open to exploring a wide range of content. For Mark, the optimal streaming services would likely be Amazon Prime or HBOMax. Amazon Prime boasts an extensive content library spanning numerous genres, with the exception of family and reality. While it may not offer the highest-rated movies and shows, it provides a vast selection. On the other hand, HBOMax

offers a balanced variety of genres and consistently holds the highest ratings for both TV shows and movies. Mark can enjoy popular titles like Chernobyl, Game of Thrones, the Wire, and the Lord of the Rings movies on HBOMax.

User Profile 2 - Jennifer:

Jennifer is a dedicated movie enthusiast who revels in thrilling horror films and classics that offer insights into cinematic history. Her demanding schedule as a surgeon leaves limited time for TV series, making her focus primarily on movies. Jennifer values the uniqueness of every film, regardless of critical acclaim. In this context, Amazon Prime might not be the best fit for Jennifer, as its horror movie offerings tend to have lower ratings. Hulu, with its emphasis on TV shows, may not align with her preferences either. HBOMax, however, emerges as a promising choice for Jennifer. Despite not having the largest collection of horror movies, HBOMax hosts the highest-rated titles in this genre, some dating back to 1920. Jennifer can indulge in her love for movies and explore cinematic history through HBOMax's diverse selection.

Section 3: Strategy and Decision-Making:

IST-707 Applied Machine Learning

In the context of strategic thinking and decision-making, the IST-707 Applied Machine Learning course exemplifies a practical and strategic learning experience. This course focuses on popular data mining methods that enable the extraction of valuable insights from datasets, with a strong emphasis on applying these methods to real-world challenges.

Throughout the course, students engage in a comprehensive exploration of various data mining tasks, including data preparation, concept description, association rule mining, classification, clustering, evaluation, and analysis. These tasks serve as the fundamental building blocks for crafting data-driven solutions to complex scientific and business problems. Strategic decision-making becomes integral as students learn to strategically formulate data mining tasks, aligning them with specific problem-solving objectives.

One key facet of strategic decision-making within this course is the ability to effectively document, analyze, and translate data mining needs into technical designs and practical solutions. Students gain valuable experience in assessing real-world problems, identifying the most appropriate data mining approaches, and devising strategies to extract meaningful insights from data. Additionally, they acquire proficiency in evaluating the performance and reliability of chosen models, a pivotal step in the decision-making process.

The pinnacle of strategic application within this course is exemplified by the final group project. This project serves as a comprehensive case study where students are tasked with acquiring, cleaning, preparing, exploring, and strategically deriving actionable insights from data, ultimately leading to decisions with a tangible impact on the chosen use case. In this project, the effectiveness of various machine learning models in predicting future stock prices was explored, with the objective of providing traders and investment enthusiasts with a competitive edge.

For this project, diverse stock price time series data were sourced from Yahoo Finance and Dow Jones Industrial Average, spanning several years of historical daily prices. These datasets were meticulously processed and enriched with additional features, including mathematical indicators such as RSI, Exponential Moving Average, and MACD. Other calculations, such as percent change, average return, and variance, were also incorporated as features. Furthermore, rigorous data cleaning procedures were implemented to eliminate any missing values, ensuring the dataset's integrity and reliability.

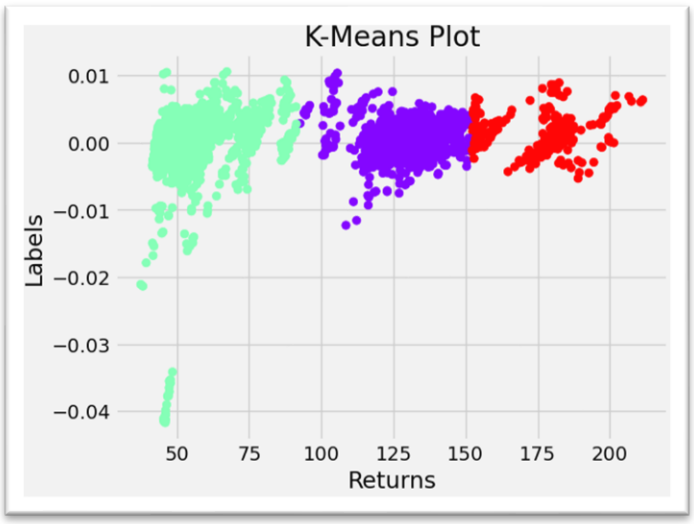
A well-diversified portfolio plays a pivotal role in the success of stock trading, serving as a safeguard against market volatility by distributing investments across various stocks, industries, and sectors. In our pursuit of achieving effective stock price diversification, we harnessed the power of K-means clustering, a widely recognized unsupervised machine

learning algorithm. This algorithm proves invaluable in crafting a diversified portfolio that mitigates risk by categorizing stocks with similar price trends into distinct clusters.

The K-means algorithm accomplishes this by assigning each stock to the cluster centroid, or central point, which is closest to it in terms of Euclidean distance. Euclidean distance serves as a geometric distance metric, effectively measuring the spatial separation between two points within an n-dimensional space. By employing this algorithm, we efficiently delineate stocks into different clusters, each characterized by its own unique set of price trends and patterns.

This strategic approach enables investors to make informed decisions about portfolio diversification. By strategically allocating investments across stocks from various clusters, investors can achieve a well-balanced and diversified portfolio that helps minimize risk. Figure 19 below vividly illustrates how K-means effectively segregates stocks based on their similarities, underscoring its prowess as an indispensable tool for crafting a diversified investment portfolio.

Figure 18 - K-Means showing 3 distinct clusters of groups of stocks



Subsequently, we embarked on a comprehensive evaluation of diverse machine learning models aimed at predicting the closing price of Google's stock, spanning both daily and 60-minute intervals. In order to maintain consistency for meaningful comparisons, we chose Google as our focal point throughout this analysis. The overarching objective was to systematically assess these models to discern their consistent performance across varying time periods and time frames.

Our repertoire of models encompassed a wide spectrum, including both traditional supervised classic classification methods and deep learning approaches. Specifically, we delved into the realms of K-Nearest Neighbors, Long Short-Term Memory (LSTM), Supervised Vector Machine (SVM), and Random Forest. While the intricacies of each model were comprehensively addressed in our final project paper and presentation, we present a succinct summary of our findings below for reference.

Figure 19 - Predicted vs actual closing price using LSTM model applied to 60 minutes Google stock

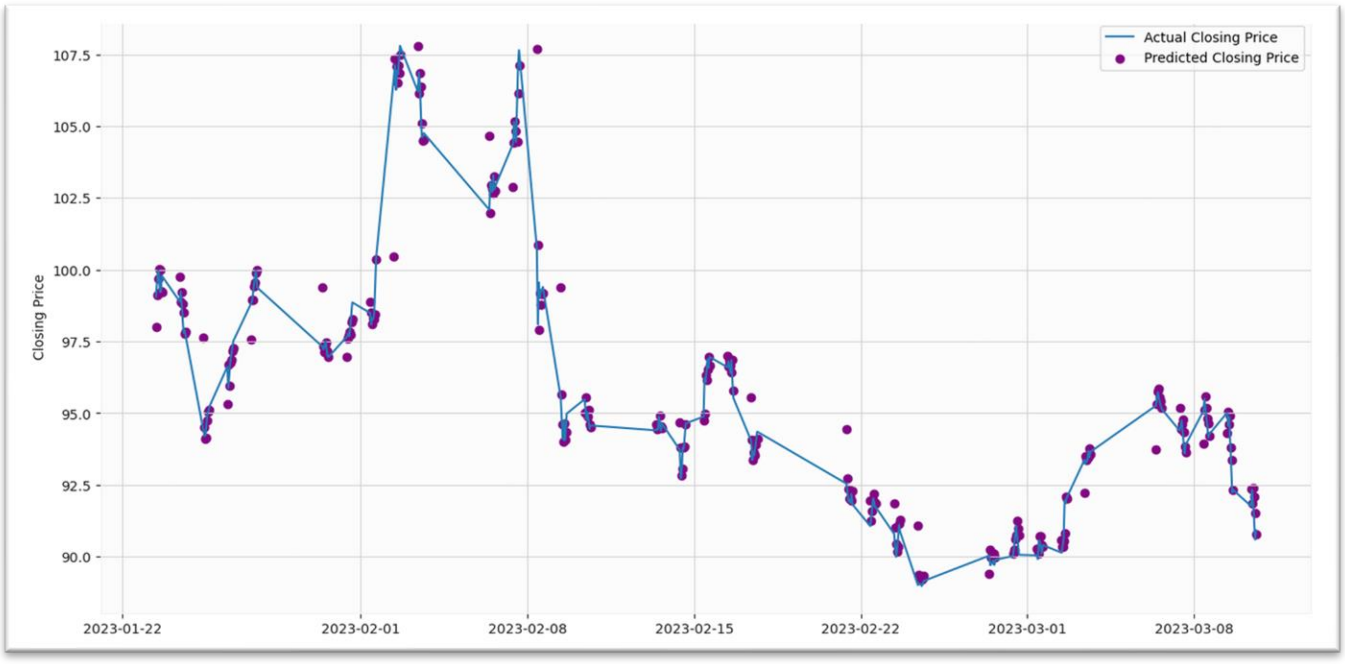


Figure 20 - Table showing sample Target Price (Actual Price) versus Predicted closing price from LSTM model

Datetime	Close	EMA	RSI	MACD	Target	Predicted Closing Price
2023-03-10 10:30:00	92.324997	93.571429	39.562627	-0.354378	91.959999	91.847229
2023-03-10 11:30:00	91.959999	93.417960	37.339704	-0.431341	91.349998	92.385834
2023-03-10 12:30:00	91.349998	93.221011	33.910466	-0.535385	90.589996	92.079391
2023-03-10 13:30:00	90.589996	92.970438	30.190291	-0.671427	90.916801	91.515457
2023-03-10 14:30:00	90.916801	92.774854	33.565362	-0.744290	90.629997	90.781067

Figure 21 - SVM model applied to 60 Min data showing signs of being overfit to the training data

Figure 22 - Predicted vs actual closing price using Random Forest model applied to Daily Google stock

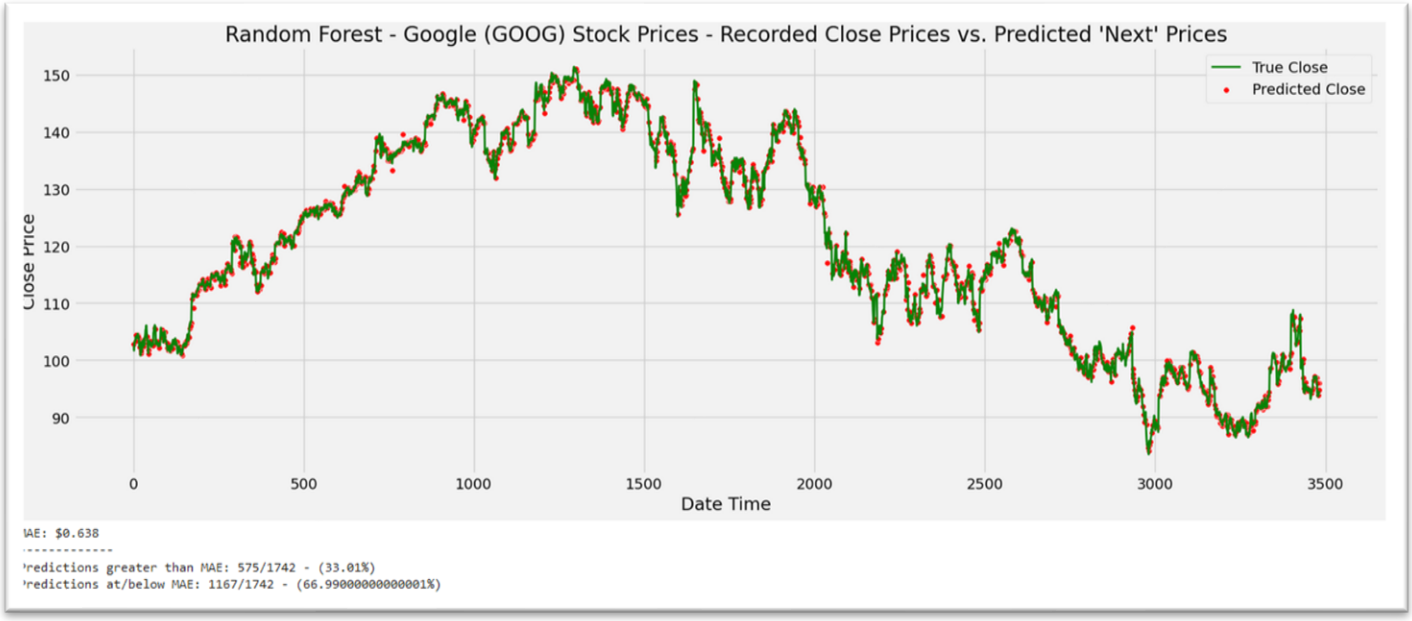


Figure 23 - Close up of Predicted vs actual closing price using Random Forest model applied to Daily Google stock

In concluding this journey through the IST-707 Applied Machine Learning course, it becomes evident that strategic thinking and decision-making play a pivotal role in the world of data science and machine learning. The application of diverse machine learning models to predict stock prices has provided a rich tapestry of insights and outcomes. From the robust K-Nearest Neighbors algorithm, which demonstrated its potential as a reliable tool for stock price prediction, to the formidable Long Short-Term Memory (LSTM) model, which showcased its prowess in capturing complex temporal dependencies, each model offers a unique perspective on the intricate nature of financial forecasting.

Furthermore, while the Support Vector Machine (SVM) model faced challenges in this specific analysis, it serves as a reminder that real-world applications demand continuous refinement and adaptation. It underscores the importance of considering not only model selection but also other critical factors such as trade costs and slippage to navigate the complexities of stock trading effectively.

Among the models, the Random Forest model emerged as a standout performer, excelling in capturing non-linear relationships within stock data. Its feature importance scoring system further highlights its capability to offer valuable insights into the significance of different features driving predictions.

Overall, this course has provided a comprehensive understanding of how machine learning models can be strategically leveraged to address real-world challenges. It emphasizes the need for data scientists and analysts to possess not only technical proficiency but also a strategic mindset to make informed decisions based on model outcomes. As we step out of this course, armed with a diverse set of models and the strategic acumen to apply them effectively, we are better equipped to navigate the dynamic and competitive landscape of data-driven decision-making.

Section 4: Implementation:

IST-691 Deep Learning in Practice (in progress)

IST-691 Deep Learning in Practice serves as a practical foray into the utilization of deep learning frameworks for solving concrete problems. The course curriculum is meticulously designed, navigating through the foundational elements of deep learning to its more sophisticated algorithms and evaluation methods. We explored a breadth of topics including Deep Learning Fundamentals, the intricacies of Backpropagation, Optimization techniques, the architecture of Convolutional Neural Networks, and advanced topics such as Transfer Learning and Autoencoding. The course also introduced us to Recurrent Neural Networks, advanced Language Processing, Embedding techniques, and the basics of Reinforcement Learning.

As the course progresses, a notable highlight awaits—my team's capstone presentation. This will feature the deployment of various deep learning models, each meticulously tailored and fine-tuned, to architect an innovative stock portfolio system. This system is designed to not only compile a strategic selection of stocks but also to incorporate a sophisticated trading algorithm capable of autonomously rebalancing the portfolio in response to market dynamics. This practical implementation underscores the program's commitment to fostering a hands-on understanding of how deep learning can be leveraged to navigate and potentially capitalize on the complexities of financial markets.

Section 5: Application of Data Science Tools:

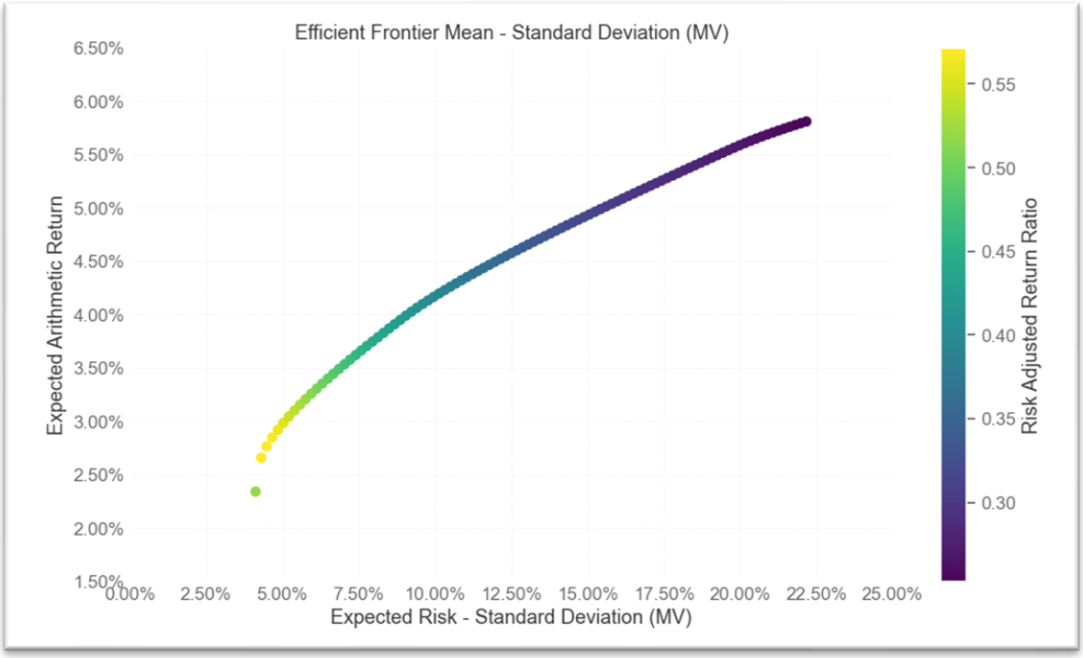
FIN-654 Financial Analytics

In the realm of application of data science tools, FIN - 654 Financial Analytics course served as a comprehensive platform for harnessing the power of programming languages, notably R and Python, to drive practical and data-driven decision-making within the financial domain. Throughout the course, an array of fundamental financial concepts, including the Capital Asset Pricing Model and portfolio sorting and optimization, were seamlessly integrated with the application of cutting-edge machine learning techniques to solve complex financial challenges. Upon completing this course, I successfully achieved the following milestones:

- Proficiently utilized Python to retrieve, analyze, and derive meaningful insights from time series data pertaining to financial securities.
- Skillfully employed Python to optimize investment portfolios based on historical data, thereby enhancing the efficacy of financial decision-making.
- Applied machine learning methodologies, both in R and Python, to address intricate problems such as predicting portfolio returns, unlocking valuable predictive insights.
- Demonstrated the ability to craft and fine-tune artificial neural network models for quantitative prediction tasks, including real-estate price forecasting and revenue prediction.

Through hands-on experiences and practical applications like the Efficient Frontier Portfolio Optimization depicted in Figure 24, this course has equipped me with the essential tools and knowledge to navigate the intricate landscape of financial analytics, making informed decisions and driving value within the financial industry.

Figure 24 - Efficient Frontier Portfolio Optimization



IST-772 Quantitative Reasoning for Data Science

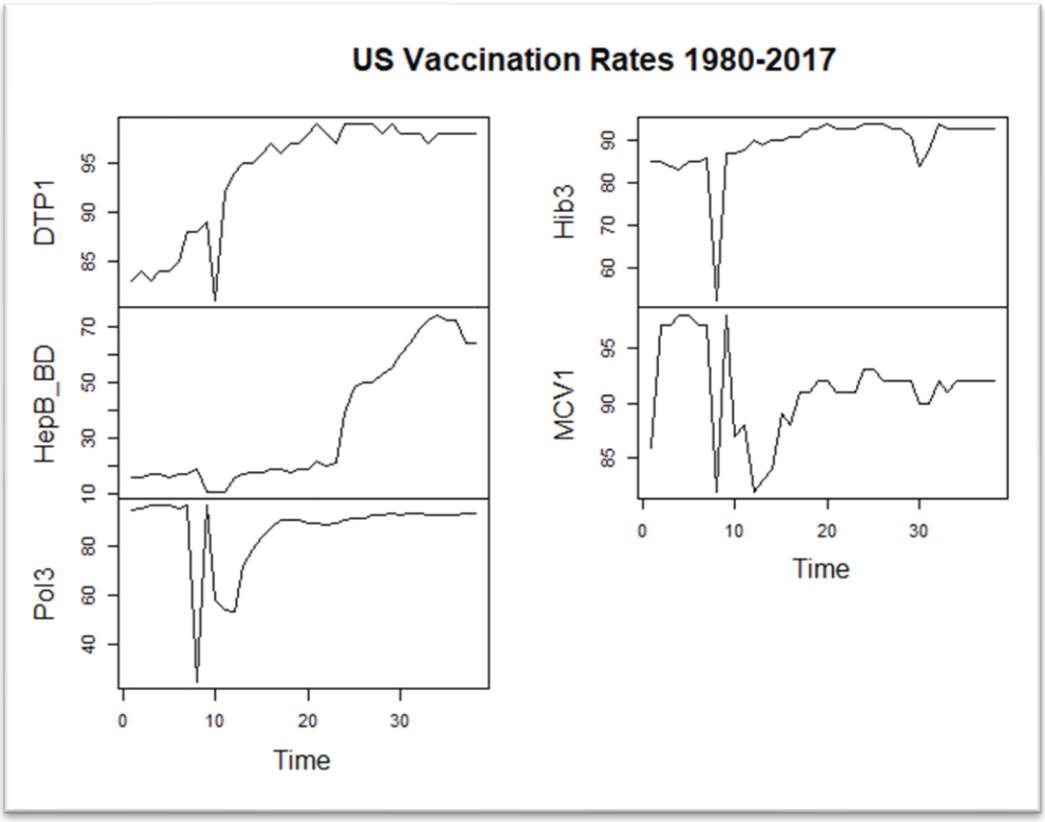
In the domain of IST-772 Quantitative Reasoning for Data Science, I delved into the core of data science, where quantitative data analytics takes center stage. This course's primary focus was to impart a deep understanding of contemporary methods for statistical inference, irrespective of the specific analytical context. The core learning objectives revolved around mastering techniques to extract meaningful insights from data samples, encompassing fundamental topics such as basic probability, sampling distributions, Bayesian theorem, and ANOVA experimentation.

The apex of this academic journey was marked by the final exam, a challenging endeavor that demanded the crafting of a written technical report. This report served as an in-depth evaluation of findings derived from the analysis of three datasets, each brimming with information regarding vaccination rates within a Californian school district and the broader U.S. vaccination landscape. The overarching goal was to conduct comprehensive analyses and present the results in a manner that could be readily understood by a scientifically astute staff member in a state legislator's office. The report was meticulously structured to include an abundance of numeric data and graphical representations, empowering the staff member to compile a comprehensive briefing tailored for legislators.

Leveraging the capabilities of R Studio, I embarked on a journey of extensive data analysis using these three vaccination-related datasets. Through these datasets, I embarked on a quest to unravel various research questions and unearth valuable insights to benefit a state legislator's office. Here, I present the salient findings that emerged from this analytical odyssey:

Time Series Analysis: I analyzed U.S. vaccination rates over time (1980-2017) for five common vaccines. HepB_DB and DTP1 rates showed an upward trend during the second third of the time series. Pol3, Hib3, and MCV rates remained relatively stable. DTP1 had the highest rate at the end, while HepB_BD had the lowest. Hib3 exhibited the highest volatility.

Figure 25 - % US Vaccination Rates time series from 1980 through 2017

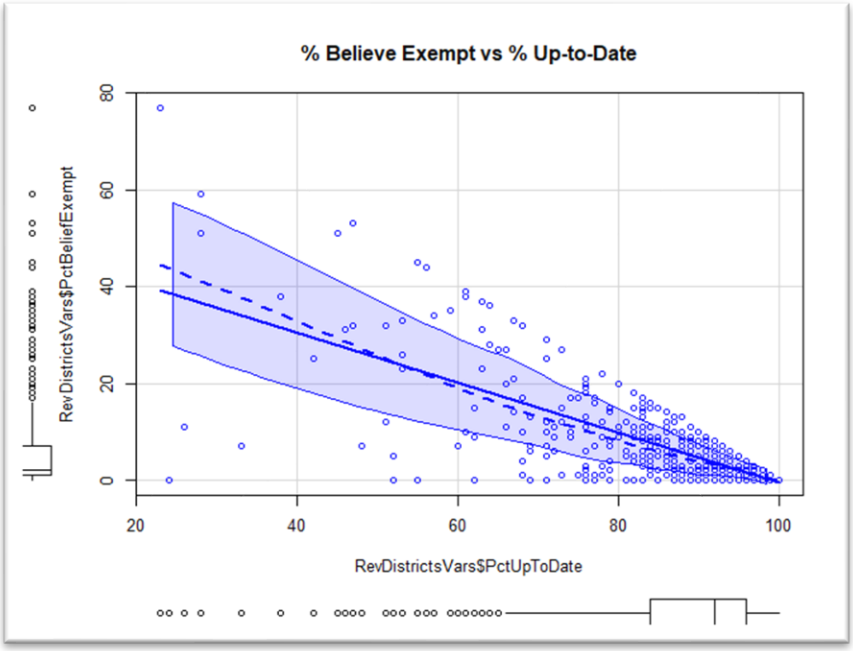


School Reporting: About 80% of public schools reported vaccination data, while approximately 20% of private schools reported. A t-test revealed a credible difference in overall reporting proportions between public and private schools.

California vs. U.S. Vaccination Rates: California's vaccination rates in 2013 were generally lower than U.S. rates reported by the World Health Organization. However, HepB rates in California were significantly higher.

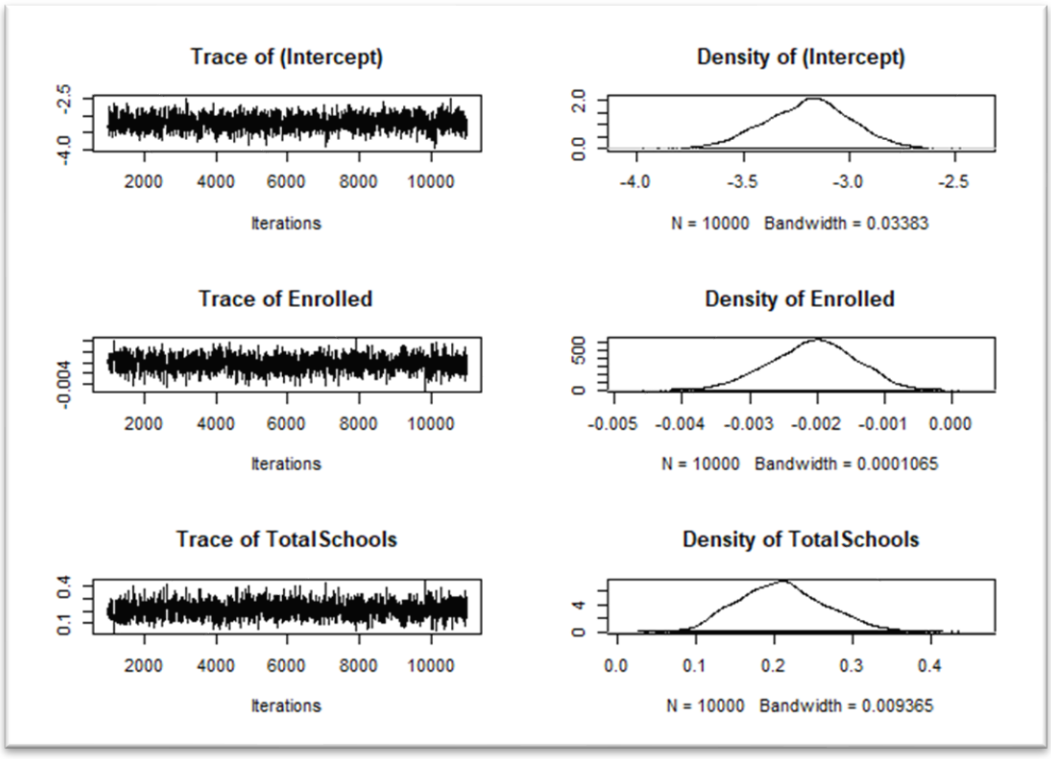
Correlation Analysis: Strong inverse correlations were observed between belief exemptions and up-to-date vaccination rates, indicating that students not up-to-date tend to have belief exemptions. Low-income indicators (PctChildPoverty, FamilyPoverty, PctFreeMeal) also correlated with low vaccination rates.

Figure 26 - Scatter plot of percent Believe vs percent up-to-date



Predictive Analysis - Reporting Completeness: Logistic regression showed that the completeness of reporting was related to enrollment and total schools. Bayesian analysis confirmed the significance of these factors.

Figure 27 - Plot of Logistic Regression showing density plots of the for completeness



Predictive Analysis - Up-to-Date Vaccination Rates: Multivariable linear regression identified PctFreeMeal and PctFamilyPoverty as significant predictors for up-to-date vaccination rates.

Predictive Analysis - Belief Exemptions: Linear regression indicated that PctFreeMeal, PctChildPoverty, and PctFamilyPoverty were significant predictors of belief exemptions.

Figure 28 - Scatter Plot showing linear relationship between % Child Poverty and % Free Meal

Conclusion: The analysis revealed that low-income demographics were associated with lower vaccination rates and belief exemptions. I recommend allocating resources to study these relationships further, especially focusing on districts with high rates of low-income demographics and belief exemptions. Tailored interventions and educational campaigns may be necessary to improve vaccination rates among these populations. These findings can inform strategic decision-making for allocating financial assistance to school districts to enhance vaccination rates and reporting compliance, ultimately contributing to public health efforts.

Section 6: Communication of Insights:

- **IST-736 Text Mining**

Data science, at its core, serves as a crucial bridge connecting a wide array of domains, ranging from business management to academic discovery. The wealth of information extracted from data holds immense potential, but its true value cannot be harnessed unless it is effectively communicated to stakeholders who stand to benefit from these revelations.

During the course of IST-736 Text Mining, we immersed ourselves in the world of text mining, absorbing essential concepts and techniques. Our journey encompassed diverse areas, including information extraction, text classification and clustering, and topic modeling. We harnessed benchmark corpora, as well as a range of commercial and open-source text analysis and visualization tools, to unearth intriguing patterns within textual data. Our exploration extended to comprehending the underlying mechanisms of advanced text mining algorithms, which find application in information extraction, text classification and clustering, and even opinion mining.

Throughout the course, we underscored the significance of communicating our findings in a manner that transcends technical jargon. We emphasized the pivotal role of effective communication in data science—a process that extends beyond language and delves into the realm of data visualization. Our assignments were structured to encapsulate this ethos, where we consistently framed our insights in a non-technical language, ensuring accessibility to diverse audiences.

For our culminating project, we embarked on a fascinating journey into the world of NBA drafts. Our objective was to leverage a plethora of text mining techniques to gain a unique perspective—essentially, to "out scout the scouts." Our strategy involved parsing the language used by draft analysts in their reports to uncover additional information that could serve as predictive indicators of a player's success or failure. To achieve this, we obtained permission to scrape pre-draft commentary dating back to the year 2000 from the NBADRAFT.NET website. Furthermore, we harnessed the NBA API to gather player statistics, including average playtime, value over replacement, and player efficiency rating, directly from the NBA website. These statistics enabled us to classify draft players as either "successes" or "busts."

Armed with our target variable, we embarked on an exploratory journey that included a series of techniques designed to unravel the intricacies of the data and tease out meaningful patterns. Among them we harnessed the power of various machine learning models and techniques including Random Forest, LSTM, Naïve Bayes and SVMs to predict if an NBA draft would be classified as a “bust” or “no bust”. We also applied LDA for topic allocation in order to identify underlying themes within the text, as well as sentiment analysis to unveil if players with more negative sentiment scores would turn out to be successful or failures. Next, are samples of the insights we uncovered, and the visualizations that helped us tell the story to our audience.

Figure 29 showcases the most frequently occurring words based on their associated sentiment scores. These WordClouds offer a visual representation of the prominent terms within the dataset, shedding light on the prevalent themes and sentiments.

Figure 29 - WordClouds based on Positive and Negative Sentiment Scores using Weaknesses Preprocessed

The objective of the sentiment analysis is to determine if negative reviews vs positive reviews by the analysts can be used as predictors to determine if a player will be successful or a "bust". Figure 30 shows that most of the players which can be considered "busts" based on their career statistics, have sentiment scores that are positive and above 0.75. Figure 31 shows detailed scores distributions for each scenario, where "bust" is a yes, no, or to be determined.

Figure 30: Sentiment Scores Distributions for Bust2 Criteria – Weakenesses_Compound Scores

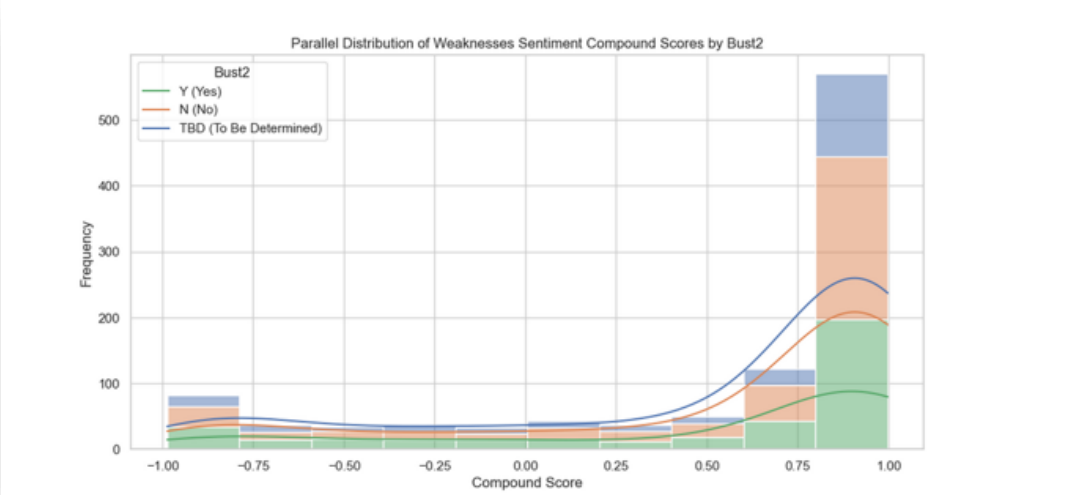
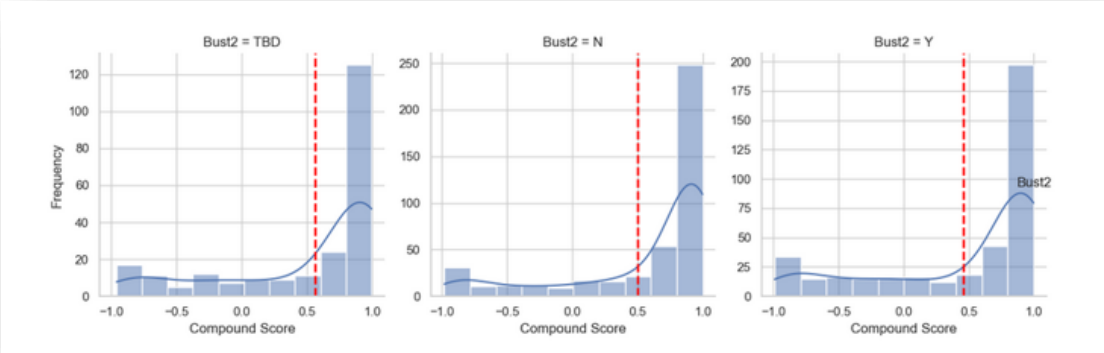


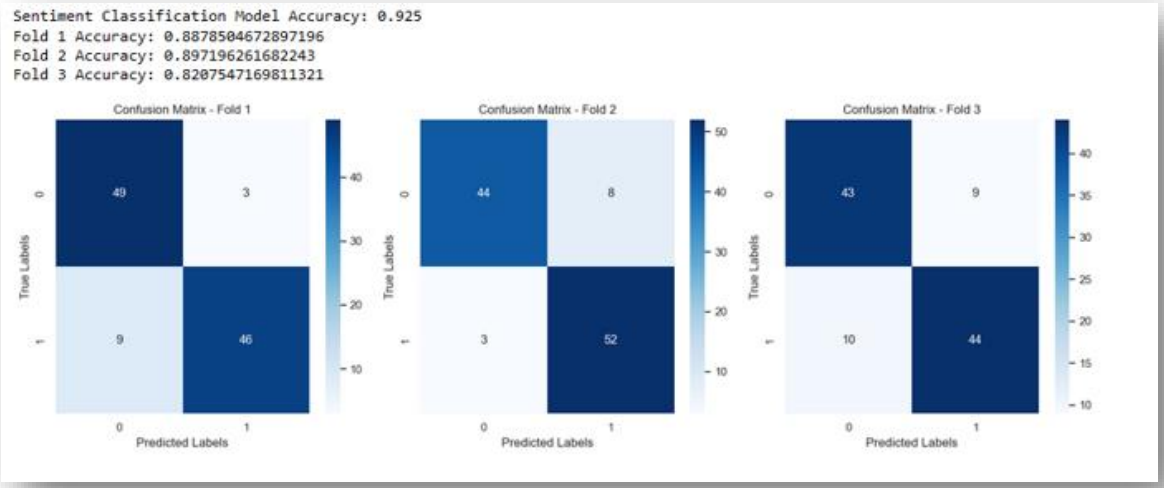
Figure 31: Weakness Compound Scores Distribution of Bust = Yes, Bust2 = N, and Bust2 = Y



The above visualizations send a clear message that our data was significantly presenting skewness towards positive sentiments, which meant we had to apply techniques to try to medicate the unbalanced nature of our sample dataset. The issue was addressed by resampling the dataset using the highest and lowest sentiment scores to compile a training dataset that would be balanced and thus able to capture both negative and positive sentiments when applied to unseen data.

Subsequently we moved to techniques that were designed to classify sentiment using supervised machine learning models. One such example is shown below where we applied an SVM model with a linear kernel and 3 cross validations for predicting positive versus negative sentiment. The model produced an overall accuracy of 92.5%. The accuracy for each fold shows some variation but remains relatively high across all folds. The heatmaps of the confusion matrices below show the individual accuracy breakdown of each of the 3 folds.

Figure 32: Results and Confusion Matrix Model 2 SVM 3CV Scenario 3



While using SVM modeling to predict “bust” vs “no bust”, we uncovered that accuracy could vary greatly as players are fed into the model. The highest SVM accuracy comes from Bust3 when fed 250 players ~75%. For RF, Bust4 does impressively well at 200 players ~74% while Bust3 does well at all 545 ~71%. This can be a balance of underfitting and overfitting while dealing with noisy text data. Also, training balance or a skew of labels should be analyzed as well.

Figure 33 - SVM plot of accuracies across the different bust criteria variables

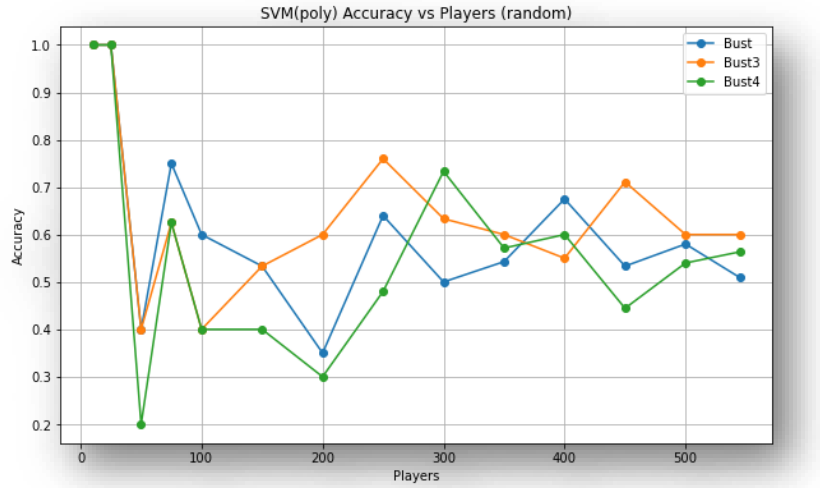
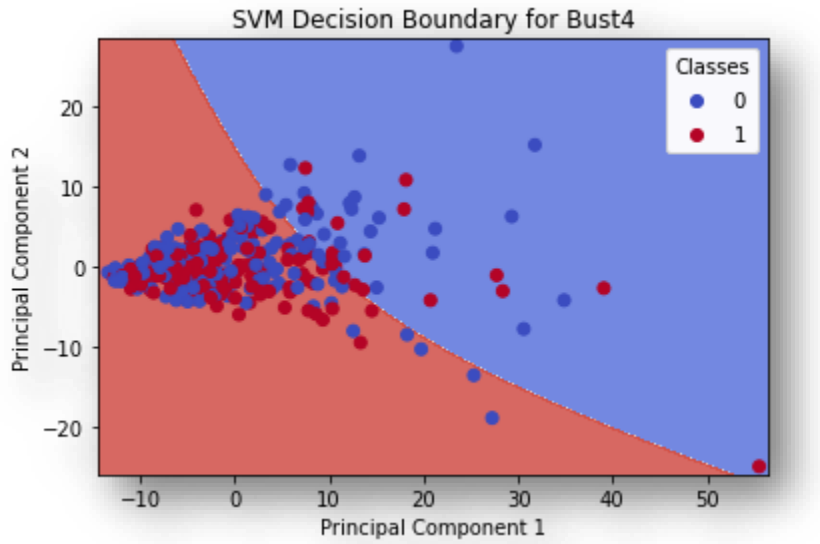
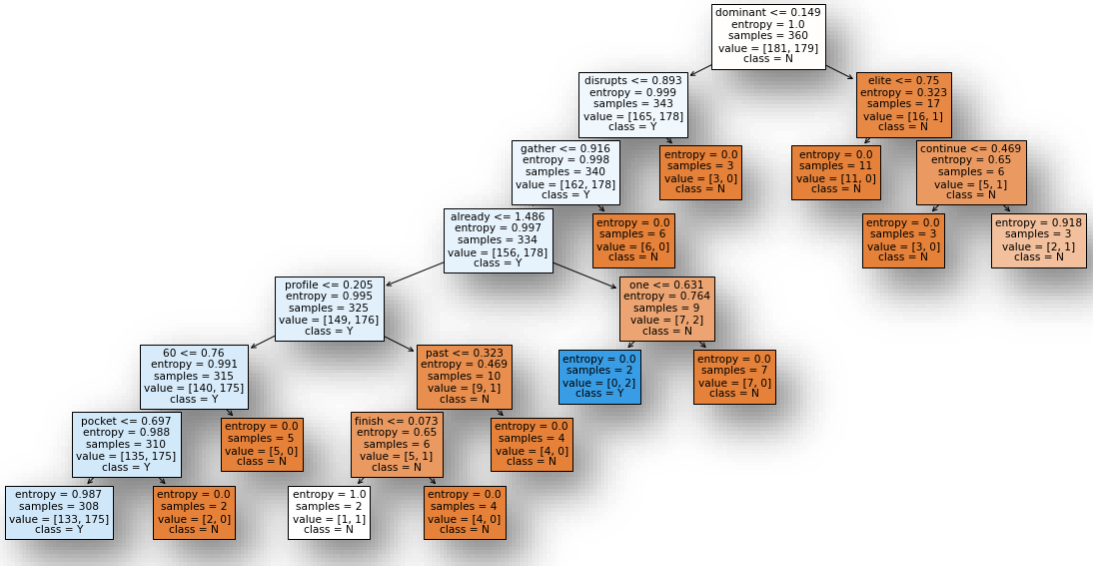


Figure 34: SVM Model Decision boundary showing disparity between 1 and 0 labels representing “bust” and “no-bust”



In the results shown above, the SVM model achieved a higher recall (it correctly identified all actual positives) but had lower precision compared to the Random Forest model. The Random Forest model had a slightly higher precision but a lower recall. The F1 score, which considers both precision and recall, was similar for both models and suggests that they have a similar overall performance. To visualize how the Random Forest models may be making decisions, a visual representation of a pruned version showing a max depth of 7 is shown:

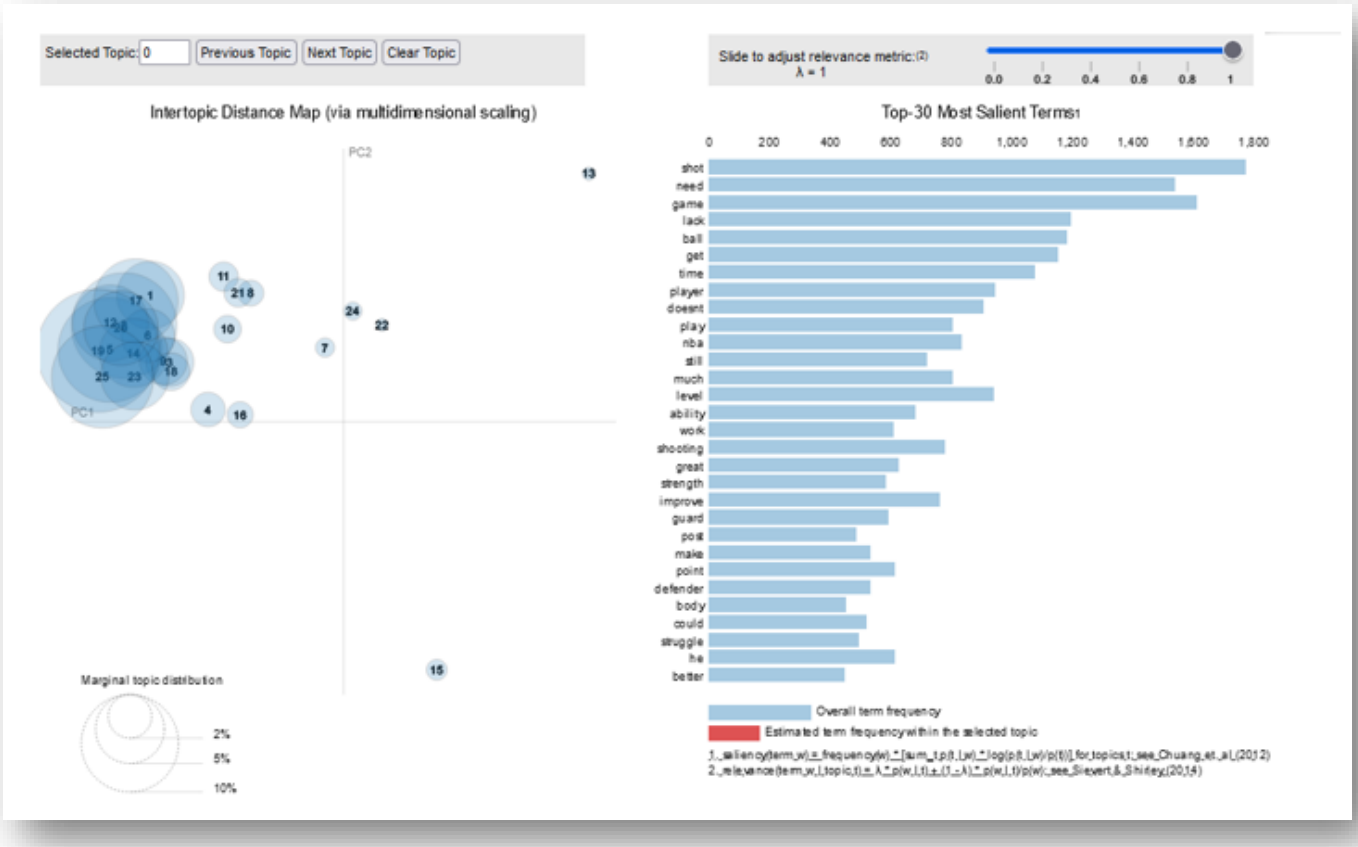
Figure 35 - Random Forest Visualiaztion (max_depth=7, min_samples_split=3, min_samples_leaf=2, max_features='sqrt', criterion='entropy', class_weight=None, splitter='random'):



Another fantastic example of deriving insights for visualizations came from PyLDAvis library. The best version of the LDA model identified 25 distinct topics based on the “Weaknesses_Preprocessed” text data. Several topics revolved around common themes like the need to improve various aspects of the players' game, such as shooting, ball-handling, and overall skills. These topics also touch upon the player's physical attributes, including strength and athleticism.

The PyLDAvis visualization below highlights the 25 topics found, where the distance between the bubbles indicates the degree of similarity or dissimilarity between topics. Specifically, it represents the inter-topic distance, also known as the topic distance map. When two bubbles are positioned close to each other on the PyLDAvis visualization, it suggests that the topics they represent share similar terms and are more related to each other. Conversely, if two bubbles are farther apart, it indicates that the topics they represent are less similar and have fewer overlapping terms. The bars represent the most frequent words present in each of the topics

Figure 36: LDA visualization (PyLDAvis) with 25 Topics from Concatenated Preprocessed Results Using Coherence Scores and Distribution of Most Frequent Words



To conclude, the world of professional basketball scouting has always been a blend of art and science. In recent years, we've witnessed a significant shift towards a more analytical approach, emphasizing data-driven insights. This study, while comprehensive in its scope, serves as an introductory exploration into the vast potential that modern scouting methods offer. The insights gleaned underscore the dynamic balance required between traditional expertise, with its nuanced understanding of the game, and the precision of cutting-edge analytical tools. What emerges is a clearer picture of the future of player evaluation, especially within the critical juncture of the NBA Draft.

Section 7: Ethical Considerations:

IST-718 Big Data

In the ever-evolving world of data analytics, ethical considerations play a pivotal role in shaping responsible and trustworthy practices. As we delve into the importance of ethics in data science, it becomes evident that ethical principles guide our actions when we create applications and interpret data.

Taking the context of our final project in IST-718 Big Data into account, our team embarked on a comprehensive assessment of credit risk, a domain where ethical considerations are of paramount importance. Our project, titled "Optimizing Credit Risk Assessment: Leveraging Customer Segmentation, Fraud Detection, and Default Risk Analysis," provided a unique opportunity to apply ethical considerations across various facets:

Data Privacy and Security: Handling sensitive financial data, including customer demographics, credit scores, and transaction history, requires strict adherence to data privacy and security regulations. Ethical considerations involve ensuring that the data is anonymized and protected to prevent potential breaches that could harm individuals' financial well-being.

Fairness in Credit Decisions: As the project involves credit risk assessment, it's crucial to ensure fairness in the models and algorithms used. Ethical concerns arise if the models discriminate against certain demographic groups or exhibit bias in credit approval or denial. Efforts should be made to eliminate bias and ensure that credit decisions are fair and unbiased.

Transparency and Explainability: Ethical considerations also include making the credit risk assessment models transparent and explainable. Customers have the right to know why their credit application was approved or denied. Ensuring that the models provide clear explanations for the outcomes is essential for ethical decision-making.

Fraud Detection and Prevention: Fraud detection is an important aspect of the project. Ethical considerations involve developing and implementing fraud detection algorithms that protect both the lender and the customers. False positives in fraud detection can negatively impact customers, so efforts should be made to minimize these instances.

Customer Consent and Data Usage: Before using customer data for analysis, obtaining informed consent is an ethical requirement. Customers should be aware of how their data will be used and for what purposes. Additionally, they should have the option to opt out if they do not wish to participate.

Responsible Use of Customer Segmentation: Customer segmentation should be used responsibly and not for discriminatory purposes. Ethical considerations involve ensuring that segmentation is done to better understand customer behavior and improve risk assessment, rather than to unfairly target or discriminate against specific groups.

Monitoring and Auditing: Regular monitoring and auditing of the credit risk assessment models are essential ethical practices. This helps identify and rectify any biases or issues that may arise over time, ensuring that the models continue to provide fair and accurate results.

Compliance with Regulatory Frameworks: Adhering to financial regulations and industry standards is a fundamental ethical consideration. Lenders must operate within the legal framework and follow industry best practices to protect both themselves and their customers.

Following are examples of descriptive and predictive analysis results obtained during the course of our exploration, which demonstrate that in every case we followed the aforementioned ethical guidelines:

Figure 37: Distribution of total income and years employed for credit approval analysis

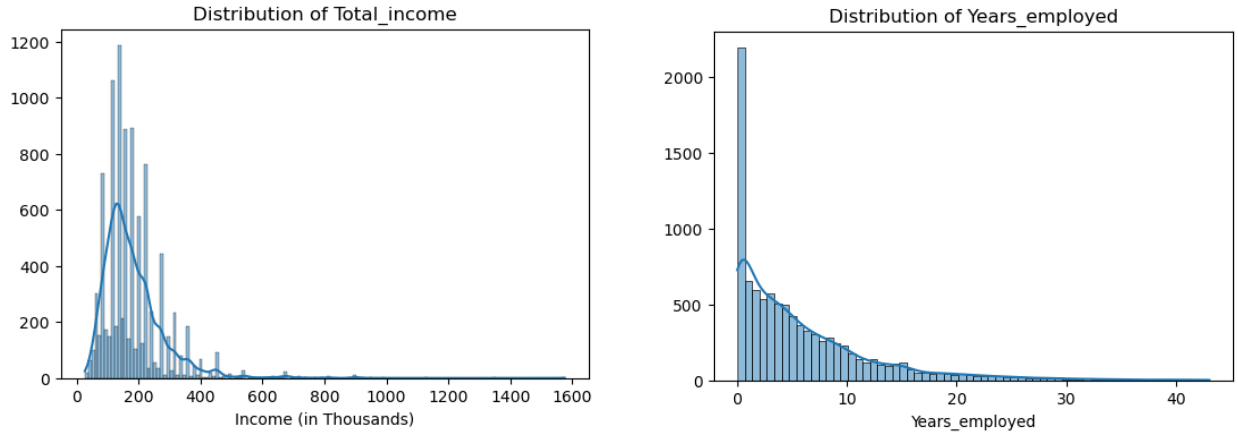


Figure38: Feature importance for credit approval using a Random Forest model

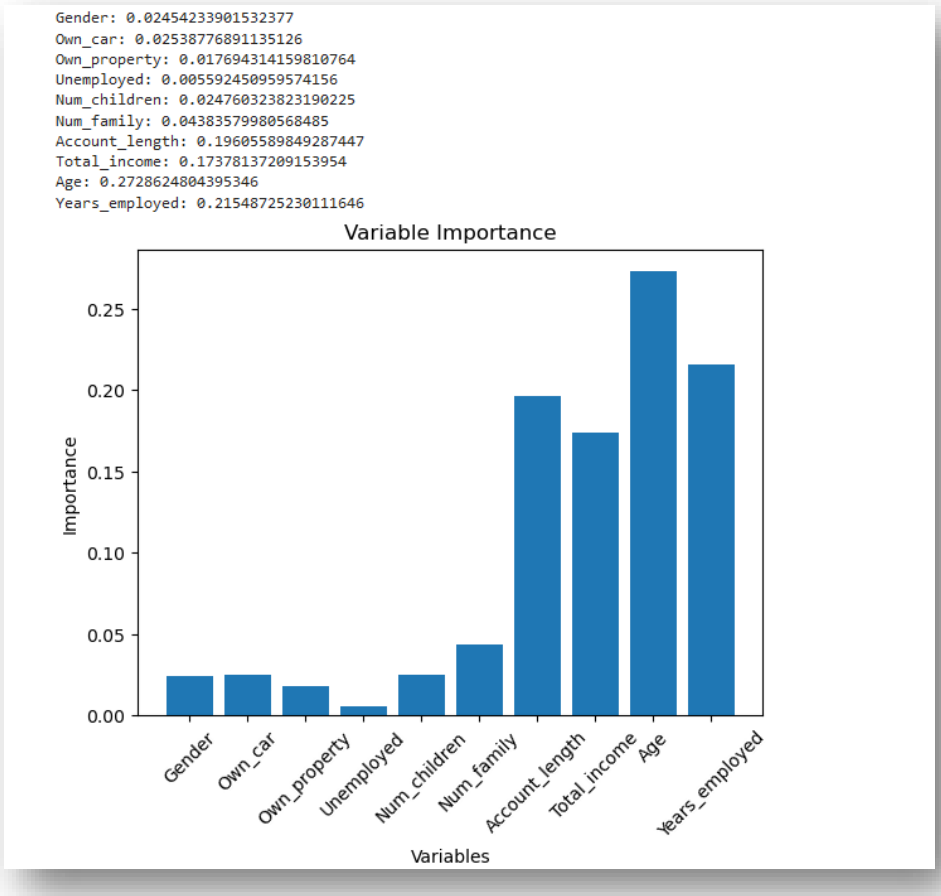
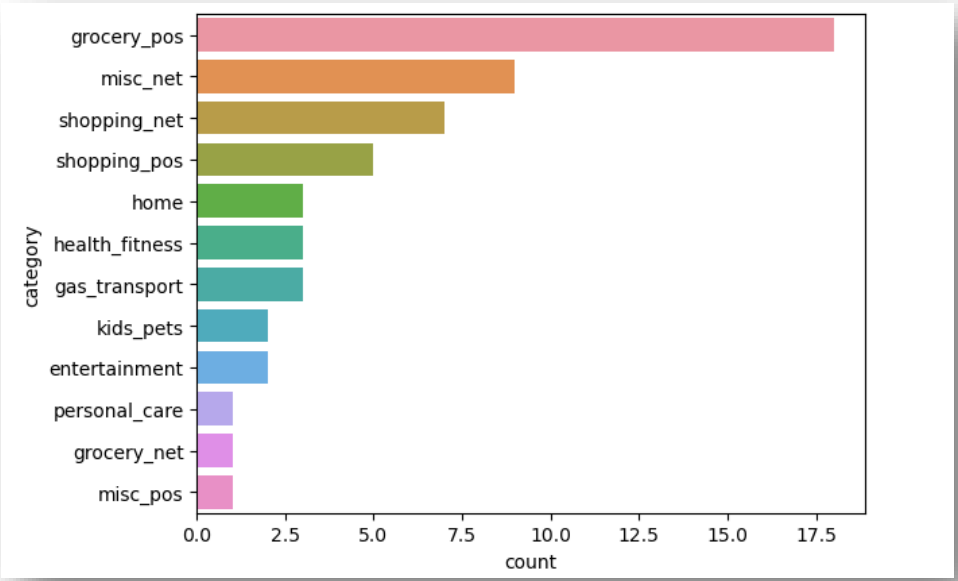


Figure 39: Fraudulent Transactions by category



In conclusion, the project offered several opportunities to apply ethical considerations, primarily in ensuring fairness, transparency, and data protection in credit risk assessment, fraud detection, and customer segmentation.

Adhering to ethical principles in these areas is crucial for maintaining trust and integrity in the financial industry and protecting the interests of both lenders and customers.

Conclusion

In synthesizing the comprehensive journey through Syracuse University's Applied Data Science program, the capstone portfolio concludes as a testament to a deeply immersive educational experience that has honed the capability to harness the vast potential of data science. This portfolio serves not only as a showcase of the technical proficiencies and strategic acumen acquired across a spectrum of real-world scenarios but also as a commitment to ethical practice and continual growth within this dynamic field.

From the meticulous gathering and management of data in a variety of formats to the application of advanced analytical methods such as machine learning and deep learning, the projects presented reflect a deep understanding of the transformative power of data science. The exploration of sentiment analysis in social media data, the strategic creation of diversified investment portfolios using clustering algorithms, and the deployment of neural networks for predictive modeling are illustrative of the multifaceted expertise developed.

Communicating insights effectively, whether through the visualization of complex data or the distillation of intricate analyses into accessible findings, has been central to the educational journey. This skill has been repeatedly demonstrated through projects that not only predict outcomes, such as the potential success of NBA draft picks but also provide actionable recommendations, such as in optimizing financial analytics for portfolio management.

Ethical considerations have underpinned all endeavors, ensuring that the pursuit of data-driven insights is balanced with the responsibility to uphold privacy, fairness, and transparency. This ethical compass will guide the continued application of data science techniques in professional practice, emphasizing the role of data scientists as stewards of data integrity and advocates for the responsible use of AI.

As this chapter of academic exploration draws to a close, the portfolio stands as a bridge to the next, where the principles and practices of data science will be applied to meet the challenges of an ever-evolving digital landscape. It is with a profound sense of accomplishment and an eagerness for lifelong learning that this journey through Syracuse's Applied Data Science program is concluded, with the anticipation of contributing to the field's future advancements and societal impact.