Applied Portfolio Report

Logan Roach

# Table of Contents

[Table of Contents 2](#_Toc169081011)

[Introduction 2](#_Toc169081012)

[Demonstrating Learning Outcomes 4](#_Toc169081013)

[Business Analytics:Developing Enhanced Marketing Strategies from Previous Campaigns 4](#_Toc169081014)

[Marketing Analytics: Segmentation for Business Intelligence Insights 13](#_Toc169081015)

[Applied Machine Learning : Predicting the Winner of The Rugby World Cup 24](#_Toc169081016)

[Deep Learning: Recursive Neural Networks In Financial Markets 30](#_Toc169081017)

[Natural Language Processing: Predicting Sentiment using NLP Techniques 33](#_Toc169081018)

[Text Mining: Filtering Resumes Using Text Mining 35](#_Toc169081019)

[Conclusion 41](#_Toc169081020)

# Introduction

Integrating and applying new skills and techniques is imperative for professionals looking to cultivate success within their careers. This is especially true in the discipline of data science where the practical application of new knowledge and techniques is not only diverse but also rapidly growing across multiple business sectors and industries. The Master of Science in Applied Data Science is a program designed to equip students with the practical skills necessary to apply data science techniques across a broad spectrum of disciplines. The program goals emphasize a comprehensive approach to data science including encompassing data capture, management, analysis, and communication aimed at enhancing decision-making processes within various organizational contexts. Graduates of the program are prepared to analyze, interpret, and clearly communicate data-driven solutions to stakeholders, ensuring their recommendations are both actionable and impactful.

As a student of the program, I have been prepared to effectively utilize data science techniques in diverse fields to achieve key goals and learning outcomes. The curriculum is specifically designed to cover the entire data science lifecycle where the program goals provide me with a well-rounded skill set. I have learned to apply data science methods to real world problems, ensuring my skills are relevant and immediately applicable to professional settings. Emphasis is placed on using data effectively and ethically to inform and support decision making, allowing me to demonstrate data-driven recommendations to organizational stakeholders to enterprise operations and processes.

The MS in Applied Data Science program is structured to achieve specific learning outcomes that align with the overarching goals of the program. These outcomes ensure that graduates, including myself, are well prepared to meet the goals of the program and demands of the data science field. Specifically, I have learned to identify and leverage appropriate technologies for collecting, storing, and accessing data ensuring I can effectively manage data efficiently. I am able to create actionable insights that vary contextually demonstrating my ability to apply data science techniques to solve real-world problems. To facilitate this, I apply predictive models and incorporate visualization tools to generate these insights helping organizations and stakeholders to understand and act on the data. Additionally, the program emphasizes the use of programming languages like R and Python to support the generation of actionable insights, ensuring I am proficient in key data science tools. To convey insights from my analyses, I have gained and practiced effective communication skills to broad audiences to further support the use of actionable insights generated from the analysis. This crucial skill enables graduates such as myself to generate findings that are both accessible and understandable which otherwise would be lost or recognized comprehensively. Moreover, I have come to appreciate ethics in data science as a paramount component to the practice. I have been trained to apply ethical considerations in the development, use, and evaluation of data and predictive models where my work adheres to the highest ethical standards.

Lastly, to further demonstrate the program’s goals and learning outcomes, the program includes a diverse range of courses, assignments, and projects that provide experience to reinforce the practical application of data science techniques. The curriculum I choose to pursue focuses on data and business analytics, artificial intelligence and language analytics. Courses such as SCM 651 – Business Analytics and MAR 653 – Marketing Analytics along with their respective assignments and project are used to highlight learning outcomes from the data and business analytics focused curriculum. Courses such as IST 707 – Applied Machine Learning and IST 691 – Deep Learning in Practice along with their respective assignments and project are used to highlight learning outcomes from the artificial intelligence focused curriculum. Finally, courses such as IST 736 – Text Mining and IST 664 – Natural Language Processing along with their respective assignments and projects are used to highlight learning outcomes from the language analytics focused curriculum.

# Demonstrating Learning Outcomes

## Business Analytics:Developing Enhanced Marketing Strategies from Previous Campaigns

The goal of this assignment was to analyze previous advertising campaigns for the Whitman School of Management at Syracuse where the focus was to determine the important factors that led to effective previous advertising campaigns and apply them in future campaigns. The proposed new campaign was given a budget of $100,000 with the goal of recruiting the top students measured by GMAT scores. With the goal of recruiting the best students, the cost of advertising per student was predicted, critical aspects of the campaign were identified, and performance measurements for the new campaign were determined.

Previous campaigns used Google Advertising and the data from previous campaigns was gathered using Google Analytics. Using Google Analytics geographic information and engagement metrics gathered from previous campaigns were used to develop a new advertising campaign. This assignment introduced the concept of using Google Analytics as a data source expanding understanding of not only data sources but also how to use them appropriately and practically. The data was separated using the four previous advertisement campaigns, but the Google Analytics dashboard allowed for consolidation of key engagement metrics to use in the analysis (see below).

Table 1: Various Whitman School of Management campaigns by dates, cost, and effectiveness measurements

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Campaign*** | ***Start Date*** | ***End Date*** | ***Cost*** | ***Clicks*** | ***CPC*** | ***Sessions*** | ***Users*** | ***Bounce Rate*** | ***Pages/ Session*** |
| **Whitman.syr.**  **edu** | 26FEB11 | 26AUG11 | $37,699.45 | 9,358 | $4.03 | 7,313 | N/a | 78.20% | 1.84 |
| **MBA Marketing – iMBA** | 2FEB12 | 26OCT12 | $80,663.24 | 5,818 | $13.86 | 2,625 | 2,367 | 89.22% | 1.14 |
| **MBA Marketing – Full Time** | 26OCT12 | 01JUL13 | $71,307.56 | 4,320 | $12.03 | 4,285 | 3,774 | 82.50% | 1.27 |
| **Delta** | 1OCT13 | 31OCT13 | $10,000.00 | 22 | $454.55 | 23 | 22 | 43.48% | 2.48 |

Using these metrics from previous campaigns, a prediction analysis using Excels’s forecasting capabilities to explore actionable insights to apply to new and future campaigns. The past campaigns consisted of three internet advertising campaigns referred to as the Whitman Campaign, the iMBA Marketing Campaign and the Full-time MBA Marketing Campaign, with the last campaign being a print advertisement in magazines of Delta planes. A prediction analysis was used to calculate the cost per click (CPC) and cost per student (CPS) in a Google Ad campaign in the following year (2024). The prediction analysis of the CPC estimated an average of $88.92 CPC for 2024 and then using the forecasted CPC, the CPS was determined to be an average of $27,391.25 for the same year. The cost per student was determined using the results from the previous ad campaigns where 50 students enrolled form the Whitman Campaign, 24 students enrolled from the iMBA Marketing Campaign and, 15 students enrolled from the Full-time MBA Marketing Campaign. Visualizations of the prediction analysis come in the form of tables showing the forecasted CPC and CPS along with the lower and upper bounds of the estimate for a comprehensive understanding of the prediction, how it was calculated, and how it changes over time.

Table 2: This shows the forecasted cost per click (CPC), as shown by 2024 the CPC will be $88.92.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Year* | *CPC* | *Forecast (CPC)* | *LCB (CPC)* | *UCB (CPC)* |
| 2011 | $4.03 |  |  |  |
| 2012 | $13.86 |  |  |  |
| 2013 | $16.51 | $16.51 | $16.51 | $16.51 |
| 2014 |  | $23.68 | $19.54 | $27.82 |
| 2015 |  | $30.20 | $25.98 | $34.43 |
| 2016 |  | $36.73 | $32.33 | $41.13 |
| 2017 |  | $43.25 | $38.55 | $47.95 |
| 2018 |  | $49.78 | $44.65 | $54.90 |
| 2019 |  | $56.30 | $50.61 | $61.99 |
| 2020 |  | $62.83 | $56.45 | $69.20 |
| 2021 |  | $69.35 | $62.18 | $76.52 |
| 2022 |  | $75.88 | $67.81 | $83.94 |
| 2023 |  | $82.40 | $73.35 | $91.45 |
| 2024 |  | $88.92 | $78.81 | $99.04 |

Figure 1: Forecasting Cost Per Click (CPC) by 2024

A graph with orange lines

Description automatically generated

Table 3: Forecasted advertising cost per student (CPS) as shown by 2024 the cost of recruiting a student will be $27,391.25.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Year* | *CPS* | *Forecast (CPS)* | *LCB (CPS)* | *UCB (CPS)* |
| 2011 | $753.99 |  |  |  |
| 2012 | $3,360.97 |  |  |  |
| 2013 | $4,753.84 | $4,753.84 | $4,753.84 | $4,753.84 |
| 2014 |  | $6,910.62 | $6,210.22 | $7,611.03 |
| 2015 |  | $8,958.69 | $8,244.55 | $9,672.83 |
| 2016 |  | $11,006.75 | $10,262.74 | $11,750.76 |
| 2017 |  | $13,054.81 | $12,260.54 | $13,849.09 |
| 2018 |  | $15,102.88 | $14,235.95 | $15,969.80 |
| 2019 |  | $17,150.94 | $16,189.05 | $18,112.83 |
| 2020 |  | $19,199.00 | $18,121.27 | $20,276.74 |
| 2021 |  | $21,247.07 | $20,034.63 | $22,459.50 |
| 2022 |  | $23,295.13 | $21,931.19 | $24,659.70 |
| 2023 |  | $25,343.19 | $23,812.80 | $26,873.58 |
| 2024 |  | $27,391.25 | $25,681.01 | $29,101.50 |

Figure 2: Forecasting Cost Per Student (CPS) by 2024

A graph with orange lines

Description automatically generated

Key aspects of the previous campaigns such as geographic region, campaign key words and timing (time of day/ day of the week) were targeted as part of the assignment to optimize the performance of the future campaign. Geographic data from previous campaigns was then investigated using Google Analytics. It was found that focusing on individuals from the United States led to the most traffic and engagement in the previous ad campaigns. Drilling down in each previous campaign, the areas that produced the highest traffic within the US were identified to focus on in the future campaign. For the first campaign, named the Whitman Campaign, the top three states or regions were New York, California and then Texas. The traffic from New York state is composed of 585,394 users and 3.7 pages per session followed by California (93,416 users, 3.39 pages per session) and Texas (61,305 users, 3.42 pages per session). Further drilling down into the first campaign, looking specifically at cities within the United States, Syracuse (310,196 users), New York (98,247 users), and Washington (26,305 users) draw the highest amount of traffic in terms of the number of clicks. When looking at the metro data, the same regions are supported with similar flow of traffic, which indicates Syracuse (374,795 users), New York (198,458 users) and Washington DC (81,891 users) are the 3 most optimal cities and regions to advertise. Looking at other engagement metrics, there were no notable differences between city, metro and region data, ranging within 43-55% bounce rate along with 3.3-4.2 pages per session. Tables were created to visually summarize the top regions from previous ad campaigns found in the analysis and to further communicate the rationale of focusing on the state of New York and previously mentioned city and metro areas.

Table 4a: Whitman Campaign Advertisement Data by Region

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Region | Sessions | % New Sessions | New Users | Bounce Rate | Pages / Session | Avg. Session Duration |
| New York | 585394 | 0.450928776 | 263971 | 0.494183405 | 3.7035 | 161.4576 |
| California | 93416 | 0.675665839 | 63118 | 0.473837458 | 3.3944 | 169.8105 |
| Texas | 61305 | 0.667286518 | 40908 | 0.457271022 | 3.4153 | 171.4045 |

Table 4b: Whitman Campaign Advertisement Data by City

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| City | Sessions | % New Sessions | New Users | Bounce Rate | Pages / Session | Avg. Session Duration |
| Syracuse | 310196 | 0.333943 | 103588 | 0.5481534 | 3.3063708 | 146.723549 |
| New York | 98247 | 0.6373426 | 62617 | 0.4580801 | 3.8472625 | 162.3518174 |
| Washington | 26305 | 0.5869986 | 15441 | 0.4304504 | 3.4892606 | 184.5516442 |

Table 4c: Whitman Campaign Advertisement Data by Metro

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metro** | **Sessions** | **% New Sessions** | **New Users** | **Bounce Rate** | **Pages / Session** | **Avg. Session Duration** |
| Syracuse NY | 374795 | 0.359596 | 134775 | 0.528801 | 3.459325 | 155.408655 |
| New York, NY | 198458 | 0.613580 | 121770 | 0.444109 | 4.011352 | 166.437352 |
| Washington DC (Hagerstown MD) | 81891 | 0.600065 | 49140 | 0.45522 | 3.517724 | 185.766616 |

The same process was followed for the other previous internet marketing campaigns for a more comprehensive analysis to further optimize the geographic targeting of the new marketing campaign. The iMBA Marketing Campaign had the same top regions when comparing new users as the Whitman Campaign with New York having 421 users, California having 329 users and Texas having 200 users. Looking at cities there are again similarities across the two marketing campaigns with the top performing cities by user being New York with 160 users, with Syracuse 71 users, and Los Angeles 38 users. When viewing top metro regions by users, more similarities were seen as New York with 416 users was highest followed by Washington DC with 161 users and Los Angeles with149 users. Again, bonce rates between regions, cities and metro areas for the second ad campaign were proximal across these focused geographics ranging between 77-91% and approximately 1.1-1.5 pages per session across each geographic group.

The final campaign, the MBA Marketing – Full Time observed similar geographic trends in terms of traffic and engagement. For new users by region, New York state performed the best at 2,009 users, followed by California with 716 users and Texas with 426 users. When comparing cities, Syracuse had the most users, 607 users, followed by New York with 552 users and Washington with 156 users. Finally, for metro regions, New York once again comes first with 1,137 number of total users, followed by Syracuse with 822 users and Washington DC with 427 users. Following the same trend as previous campaigns, the bounce rates ranged from 75-87% and 1.1-1.9 pages per session showing no significant change across geographic area where there was roughly. Given the results of the top performing geographic regions from all previous campaigns, it was recommended that New York be the most focused region for the new campaign with city and metro areas such as New York City and Syracuse to be specifically targeted. Other regions, cities and metro areas would be incorporated, only after determining how to maximize converting new users to new students.

After exploring geographic areas using previous campaigns, key words from each marketing campaign were investigated using Google Analytics to find which words generated the most traffic and engagement to be used in the future campaign. Identifying the best key words to use in a campaign would improve traffic and engagement and increase the ability to recruit top students into the program. When investigating the Whitman Campaign, the keyword that generated the highest number of engagements was “MBA” generating 206 clicks and 2.54 pages per session. The worst key word to use was “Top MBA” as it only generated 3 clicks and 1.67 pages per session. A few top performing key words were identified with the iMBA Marketing Campaign. The best key word was “online MBA” with 2318 clicks and 1.14 pages per session during the time of the campaign. Other top performing key words for the campaign include “MBA” with 1058 clicks, “AACSB MBA” with 330 clicks and “AACSB MBA Programs” with 189 clicks. For the MBA Marketing – Full Time campaign, “online MBA” was found to be the top performing key word with 4,277 clicks with the remaining key words generating noticeably less traffic. It was recommended that the key words “MBA” and “online MBA” be used in the new marketing campaign to generate the most traffic and engagement, which then could lead to increasing the number of top students recruited to the program.

Going further to explore key aspects for a new campaign, the traffic and engagement from the whitman.syr.edu site was investigated again using Google Analytics with consideration to both the day of the week and the time of day. The period of time analyzed was from the beginning of 2011 to the end of 2014. This was done for a comprehensive understanding of all traffic and engagement for the Whitman School of Business website across the period of all active marketing campaigns used before. Using this information would allow the next campaign to specifically target the optimal times to show online advertisements to interested people to not only maximize traffic and engagement, but also increase the ability to recruit the best students from the traffic generated by the new campaign. It was found that traffic to the site was most prolific between Saturday and Thursday with the worst traffic towards the site on the Fridays of each week. To visualize this, the table below shows the traffic to the main page of the Whitman School of Business website during the week. Additionally, it was found 73.3% of the traffic occurred between 17:00 and 23:00 when looking at site traffic by time of day. This was also visualized in a table to further communicate the findings of analyzing the traffic by time of day. Another visualization that could be implemented is to graph the data found when analyzing the traffic of the Whitman School of Business website. This would improve the understanding of why the analysis determined Saturday – Thursday from 17:00 to 23:00 to be the optimal time during the week and during the day show advertisements in the new campaign with the goal of increasing traffic, engagement and the ability to recruit the top students to the program. The rationale behind the recommendations to limit the new campaigns to the time frames described above is that young professionals would be most likely to start an MBA program, so prospective students would be more engaged with the site after work (17:00) and not on Fridays when they are more likely to decompress from the stresses of the week. Specifically targeting the time, the campaign to appear before young professionals would be more cost effective as it would extend the length of time the campaign could run and target the audience most likely to join the program.

Table 5a: Traffic of whitman.syr.edu by day of the week from 1 January 2011 to 31 December 2014

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Day of Week | Sessions | New Users | Bounce Rate | Pages per Session |
| Monday | 1,101 | 1,035 | 78.02% | 1.80 |
| Tuesday | 1,138 | 1,035 | 76.54% | 1.90 |
| Wednesday | 993 | 936 | 79.56% | 1.73 |
| Thursday | 940 | 893 | 78.19% | 2.01 |
| Friday | 763 | 725 | 78.11% | 1.84 |
| Saturday | 1,176 | 1,133 | 77.64% | 1.89 |
| Sunday | 1,268 | 1,212 | 79.34% | 1.72 |

Table 5b: Traffic of whitman.syr.edu by time of day from 1 January 2011 to 31 December 2014

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Time of Day | Sessions | New Users | Bounce Rate | Pages per Session |
| 0000 | 248 | 235 | 74.19% | 2.33 |
| 0100 | 99 | 90 | 71.72% | 1.88 |
| 0200 | 96 | 90 | 80.21% | 1.66 |
| 0300 | 47 | 46 | 72.34% | 1.85 |
| 0400 | 45 | 43 | 86.67% | 1.42 |
| 0500 | 26 | 22 | 73.08% | 1.81 |
| 0600 | 52 | 51 | 78.85% | 1.87 |
| 0700 | 68 | 64 | 76.47% | 1.90 |
| 0800 | 107 | 100 | 78.50% | 1.73 |
| 0900 | 107 | 98 | 74.77% | 1.93 |
| 1000 | 139 | 125 | 80.58% | 1.90 |
| 1100 | 156 | 149 | 78.21% | 1.90 |
| 1200 | 137 | 127 | 83.21% | 1.67 |
| 1300 | 167 | 149 | 80.84% | 1.90 |
| 1400 | 150 | 141 | 74.67% | 1.74 |
| 1500 | 161 | 149 | 79.50% | 1.60 |
| 1600 | 167 | 160 | 74.85% | 1.90 |
| 1700 | 795 | 763 | 78.62% | 1.70 |
| 1800 | 757 | 720 | 79.13% | 1.68 |
| 1900 | 729 | 695 | 79.01% | 1.64 |
| 2000 | 806 | 770 | 79.28% | 1.79 |
| 2100 | 822 | 780 | 79.44% | 2.01 |
| 2200 | 814 | 776 | 77.64% | 2.07 |
| 2300 | 684 | 650 | 75.58% | 1.88 |

After exploring the cost of a new marketing campaign using prediction analysis and the key aspects from previous campaigns to incorporate into the new campaign, performance metrics were identified to measure the results of the new campaign. Similar metrics from previous campaigns would be used to gage the success of the new campaign including: the optimal length of time of the campaign considering the budget of $100,000, the amount of engagement or clicks on the adverts, the number of sessions, new users acquired, the bounce rate, the cost per click, the cost per student, the pages per session during the length of the campaign. Other metrics to include when evaluating the new campaign would include the behavior of users who engage with the online advertisements such as where the users would enter and exit the site from. Ideally this would help to reduce the bounce rate by optimizing the campaign to target users who would be more engaged with the site, increasing the ability to recruit users and find the best students for the program.

Keeping the goal of recruiting potential students with the highest GMAT scores, incorporating data including GMAT scores such as location within the United States with the highest GMAT scores, or previous undergraduate degrees and grades would improve the campaign optimization by targeting the audience most likely to meet the campaign goals. Additionally, if the goal is to recruit students to the programs, the number of new students gained during the time of previous marketing campaigns would be useful for measuring the success of the new campaign. Lastly, after incorporating the previous recommendations, and the additional data mentioned above, a conversion rate of users to new students in the program would be used to measure the success of the marketing campaign.

## Marketing Analytics: Segmentation for Business Intelligence Insights

The primary focus of this assignment is analysis of the data, generating insights from the analysis, and applying insights into strategies for the business at the center of the case study: Wahoo Fitness. The data was provided in several flat files in the form of excel sheets with one sheet containing survey answers from customers, a second sheet of the segmented customers surveyed clustered via k-means clustering, another excel sheet containing questions from the survey and a sheet containing the demographic information of the surveyed customers. Since the customers were already segmented, the goal was to go further to make strategy suggestions after identifying how to target the segmented customers. It’s imperative during the analysis that all ethical considerations are applied to not only reduce bias when recommending strategies, but also practice responsibly, transparently while keeping customer sensitive information private. In this case, sensitive information was removed from the data beforehand, so no identifying or compromising information can be found within the data. If this was not the case, customers would need to be anonymized by removing sensitive information and to reduce bias within the analysis.

Keeping ethical considerations in mind, the customer segments were investigated, the psychographic and demographic profiles of each segmented were determined, actionable insights applied to each identified segment, and communicate insights using visualizations created using R programming. R was also used to cluster and investigate the data for insights going further to understand what can be found within the data. For example, an elbow plot (see below) was generated to confirm that 3 clusters were optimal for the surveyed customer data. The “elbow method” is optimizing the appropriate number of clusters by determining the optimal within-cluster sum of squares (a measurement of distance between each point within the data) and the number of clusters as too many clusters would lead to overfitting the data. Other examples include a correlation analysis of the segmented customers interests using the survey questionnaire.

Figure 3: Elbow Plot used to Investigate the Optimal Number of Clusters in the Analysis

A graph of a number of clusters

Description automatically generated

After confirming that 3 clusters were optimal for the analysis, the psychographic and demographic profiles of each segmented were determined. To further ensure the distinction of three separate customer segments a correlation analysis was performed where the top 20 positive correlations within each cluster were visualized in a heat map. This allows the analysis to understand not only differences among the clusters but also provide insights into each cluster. By leveraging the segmented customers within each cluster, marketing strategies can be developed to improve the retention rate of customers using Wahoo Fitness Products. The correlation heat maps can be found below:

Table 6: Customer Segment 1 Correlation Among Survey Questions

A table with numbers and a number of variable

Description automatically generated with medium confidence

Figure 4: Customer Segment 1 Survey Question Correlation Heat Map

A diagram of a triangle

Description automatically generated

Table 6: Customer Segment 2 Correlation Among Survey Questions

A table with numbers and a few black text

Description automatically generated with medium confidence

Figure 5: Customer Segment 2 Survey Question Correlation Heat Map

A diagram of a triangle

Description automatically generated

Table 7: Customer Segment 3 Correlation Among Survey Questions

A table with numbers and a few letters

Description automatically generated with medium confidence

Figure 6: Customer Segment 3 Survey Question Correlation Heat Map

A diagram of a triangle

Description automatically generated

In the correlation analysis of the survey questions within each customer segment, several insights were gained. The first correlations are found in customer segment 1 where there was a 0.559 correlation between question Q2\_9 (Enjoy watching sports on TV) & question Q2\_10 (Enjoy going to live sporting events) showing a positive correlation of segment 1 customers enjoy viewing sports at home and live in person. The next highest positive correlation found in customer segment 1 was between questions Q2\_4 (Follow a structured training plan) & Q2\_7 (Set specific performance goals) with a correlation 0.480 suggesting these customers follow a structured training plan while also set specific performance goals. Positive correlations found in customer segment 2 were seen between questions Q2\_9 (Enjoy watching sports on TV) & Q2\_10 (Enjoy going to live sporting events) with a correlation of 0.442. Similar to segment 1 where segment 2 customers also enjoy viewing sports at home and live in person. The second highest positive correlation found in customer segment 2 was between questions Q2\_2 (Knowledge about sporting equipment) & Q2\_3 (Maintain own sporting equipment) with a correlation of 0.388 suggesting these customers in segment 2 have knowledge of sporting equipment and tend to maintain their own equipment. Lastly, positive correlations found in segment 3 were seen in questions Q2\_9 (Enjoy watching sports on TV) & Q2\_10 (Enjoy going to live sporting events) with a correlation of 0.637 was seen completing the trend of enjoying viewing sports at home and live in person across all segments. The second highest positive correlation found in customer segment 3 was between questions Q2\_5 (Track and use data to measure performance) & Q2\_7 (Set specific performance goals) with a correlation of 0.442 suggesting customers in segment 3 are similar to segment 1 regarding their interest of setting performance goals, but customers in segment 3 seem to be more interested in tracking performance suggesting a sincerity with their fitness goals.

After these initial insights among the segmented customers were found, the analysis went further into exploring the psychographics or lifestyle/ interests of the customers within each segment. Using the customers answers to survey questions in conjunction with the previous correlation analysis it was found that customers in the first segment were interested in being active, similar to the other segments but with a higher focus on enjoyment and social engagement. The third segment focused more heavily on nutrition, fitness performance and competitions. The second segment seemed to exist in between segment 1 and 2 with interests in performance but also in enjoyment of the activity. Segment 2 seemed to be less specific with their goals but still engaged in tracking their progress as well as following sporting events. The table below shows some answers to the survey questions such as the average number of sports in each segment, followed by a consolidated set of survey questions across each segment. The figures below are graphical representations of the answers of the questions shown in the table to communicate the differences between segments for most of the answers.

Table 7: Consolidated Customer Survey Responses Across Segments

A screenshot of a computer

Description automatically generated

Figure 7: Graphical Representation of Customer Responses Across Segments (Question Set 1)

A graph of colored lines

Description automatically generated

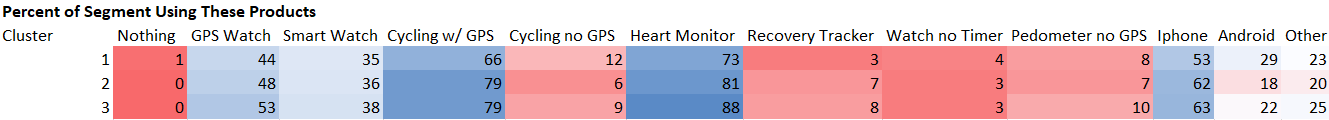
Figure 7: Graphical Representation of Customer Responses Across Segments (Question Set 2)

A graph of colored lines

Description automatically generated

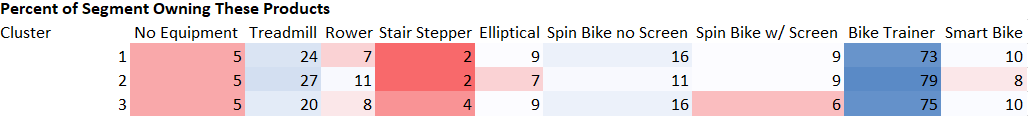
Digging down further into the psychographics of the segmented customers, the fitness preferences of each segment were analyzed for insights to build actionable strategies for Wahoo Fitness to capitalize on. Some similarities were found across all segments such as all segments owned similar fitness monitoring accessories, all segments were using tracking products such as GPS and heart monitors, all segments connected devices with their phones (mostly iPhones) and no segment significantly used recovery trackers. This can be seen visually in the table below color coded to show highly recorded percentages of the segment using the product in blue with low percentages in red.

Table 8: Percentage of Segments Using Fitness Products



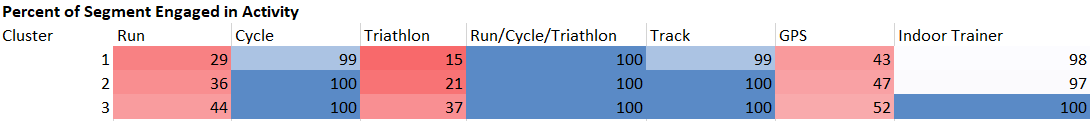
When investigating the exercise products used across segments, it was found all segments had similar products such as the bike trainer and treadmill. There were low levels of use of equipment such as the rower, spin bike with or without a screen, ellipticals, or smart bikes with the lowest exercise product used being the stair stepper. Below is a table similar to the one just above with color coding to aid in communicating the percentage of each segment using the exercise products listed above.

Table 9: Percentage of Segments Using Exercise Products



Regarding the athletic activities of the segmented population, there were some similarities seen across segments. For example, cycling, track running, all activities together, and activities involving the use of an indoor trainer were popular across samples. There was variation in activities across segments such as running, triathlons and activities involving GPS. In the activities with variations, customers in segment 1 had the least engagement with the activities while segment 3 had the highest engagement. This is consistent with prior insights gained where segment 3 seems to have a higher dedication to their fitness with an interest in competition. Another table showing the percentage of each segment engaging in each activity can be found below with a higher percentage in activity engagement in blue and a lower percentage of activity engagement in red.

Table 10: Percentage of Segments Engaged in Activity



After investigating the psychographics of the customers within each segment, the demographics of each segment were analyzed. The segments had a slight variation of age between them, but with the average age of all segments was in the mid-40s. When investigating the age distribution of each segment, there seems to be an opportunity to engage new customers in the 20-35 age range, and retaining new customers at a young age could lead to increases in customer lifetime value. Similar findings were seen when investigating the gender make up of segments where there wasn’t much variance across segments, however all segments were slightly more composed of men (1) than women (2). Below are visualizations showing the average age across segments, followed by the averaged gender make-up and lastly a histogram showing the age distribution across customer segments. These visualizations are useful for conveying the insights found within the analysis while also supporting the recommendation for opportunities in reaching out to new customers between ages 20-35.

Figure 8: Graphical Visualization of the Average Age Across Customer Segments

A graph of a graph

Description automatically generated with medium confidence

Figure 9: Graphical Visualization of the Average Age Across Customer Segments

A diagram of a person with blue lines

Description automatically generated

Figure 10: Age Distribution Across Customer Segments

A graph of different colored bars

Description automatically generated

In summary the segments were identified to develop specific actionable business strategies for Wahoo Fitness. Segment 1 is identified as customers interested in athletics for pure pleasure, referred to as the “past timers” as they have interests in fitness, but the source of their interest seems to be for pleasure rather than interests in competition or strictly improving fitness performance. Segment 2 is identified as customers that are both health conscious with an interest in sports and athletics as they have higher levels of interest in athletics than segment 1 but not segment 3. Segment 3 customers are identified as the competitors with the highest levels of interest in tracking and improving performance. From these insights and investigations into the psychographic and demographic makeup of the segments, several actionable strategies are recommended to Wahoo Fitness.

The first being to try and capture customers in the low age range (20-30) while also engaging their current customers using social media to increase the customer lifetime value across all segments. This would include shareable workouts, workout leaderboards between global users & friends with the goal of increasing customer retention (or reduce churn), engagement and a natural recruitment of new customers through leveraging the networks and friend groups of current customers. The second recommendation to increase metrics such as customer lifetime value and customer retention is to offer new products and features. There are not a lot of gaps in the customer base regarding fitness equipment, but product customization could be an avenue to incentivize those already owning equipment to purchase cosmetic add-ons, or small improvements as several segments have an interest in maintaining their own equipment. Lastly to promote growth within the current and new customers, it’s recommended that Wahoo Fitness endeavor to upcycle fee app users to a paid subscription model attached to their products. This would provide an opportunity to generate advertising revenue in a free application given to their less serious customers in segment 1 and generate income from their more dedicated customer base in segment 3 from paid subscriptions with the goal of not only increasing customer lifetime value by converting customers in segment 1 to be more like segment 2 and 3 but also retain their current customers by improving on their products and services.

## Applied Machine Learning : Predicting the Winner of The Rugby World Cup

The purpose of this assignment was to use multiple machine learning models to predict a real-world outcome where the analysis began using Association Rule Mining, Classification using Naïve Bayes, and finally Decision Trees. The outcome of interest involves predicting the winner of the 2023 Rugby World Cup. The 2023 RWC is a worldwide multinational rugby union competition, not unlike the FIFA World Cup. The winning country not only receives a monetary prize but being the third largest sporting event in the world, the winners receive international exposure only comparable to the Olympics and the FIFA World Cup. Given the prestige from winning the competition and the broader macro-economic impacts of an international competition, predicting the outcome presents not only an interesting challenge, but models could be applied or adjusted to fit other international competitions.

Association Rule Mining was used to try and see if any factors contribute to teams winning matches. Naïve Bayes was used to see if teams could be classified as winning teams using features within the data. The features used in the classification were winning margins, winning side, home team, away team, home score, and away score. Lastly, Decision Trees use tree-like models to make predictions based on input features. In this analysis, Decision Trees are used to try and predict the winner of teams within the dataset using winning team as the target variable with the deciding data features being winning side, home team, away team, home score, and away score.

The data was found on [Kaggle](https://www.kaggle.com/datasets/lylebegbie/international-rugby-union-results-from-18712022) (and limited to the previous RWC since using historical data was found to be useful in determining the top performing team historically but not useful for predicting the upcoming winner of the current RWC tournament in 2023. To further reinforce why the data was focused on data since the last RWC, the team’s winning records since 2000 and the last RWC can be compared. When looking at the winning teams since the turn of the century, New Zealand seems to be the top performing team followed by England, South Africa and Australia. When looking at the winning teams since the last RWC, the top performing teams are France followed by Ireland and then New Zealand providing a different view of the performance of each team. This can be seen in the first Association Rule Minning models because if the data was focused on all the historical data or the data since the turn of the century, one would expect New Zealand, England, South Africa and Australia to be the top performing teams, but when focusing the data since the last RWC, the recent performance of teams like Fance, New Zealand, and Ireland can be seen giving a more complete understanding of each team’s performance.

Figure 11: Association Rules Mining Plot using Data since 2000

A diagram of a network

Description automatically generated

Figure 12: Item Frequency Plot from the Association Rule Minning with Data since 2000

A graph of a number of people

Description automatically generated with medium confidence

Figure 13: Association Rules Mining Plot using Data Since Last RWC

A network of words and lines

Description automatically generated with medium confidence

Figure 14: Association Rules Mining Plot using Data Since Last RWC

A row of grey squares with black text

Description automatically generated

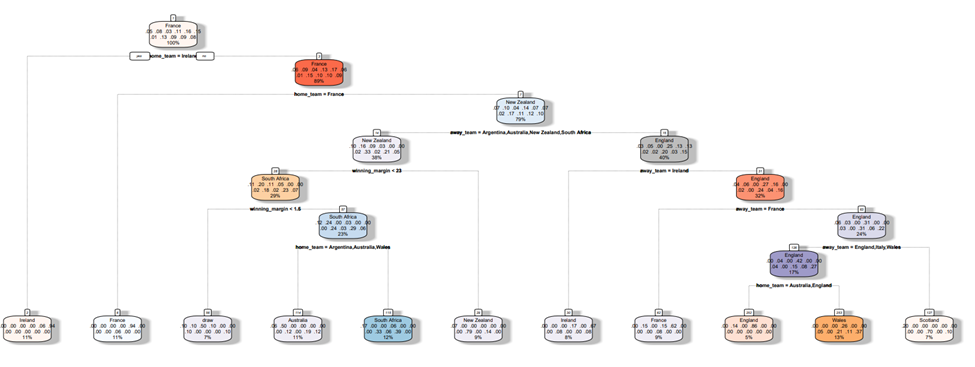
The top rule found using Association Rule Mining was the home team winning. Other rules found included France and Ireland performed the best both at home and away closely followed by Wales and New Zealand. The analysis was useful as the top feature involved with a team winning was them playing at home, and the other top performing teams participating in the RWC 2023.

The Naïve Bayes analysis endeavored to see if using features within the data could classify winning teams. The deciding data features were winning margins, winning side, home team, away team, home score, and away score while the target feature was winning team. There were 151 observations from the international matches from the previous RWC along with the 6 variables mentioned above. The data was split into training and test set where the training set contained 66% of the data and the remaining was left in the test set.

The initial results of the Naïve Bayes model had an accuracy of 53.33% in predicting the winning team, which was only slightly better than random guessing. To enhance the model's performance, Laplace smoothing was applied, resulting in an improved accuracy of 60%. Seeing improvement, 10-fold cross validation was also introduced, and the Laplace value was sequenced from 0.1 to 1.0 by 0.1. The cross-validation model using the best tunning parameters returned the same accuracy of 53.33%. These results were unsatisfactory, suggesting that further analysis using additional data and advanced techniques could potentially enhance the model's accuracy.

The final analytical model in the project utilized Decision Trees. The data was split using a training set of 66% of the data and a testing set of the remaining data. The first tree model had an overall accuracy of 48.9% where France’s winning record established France as the root node where the branches in the determine the opposition and if they were home or away. The subsequent branches represented the opposition and the location of the game (home or away). For instance, at the root node, if France played Ireland at home, the left branch indicated that Ireland would win 11% of the time. Conversely, the right branch showed that if France did not play Ireland at home, France would win 89% of the time. The remaining nodes followed a similar structure, with winners determined based on the branching criteria. The terminal nodes show the final decisions determining the winners down each branch of the tree.

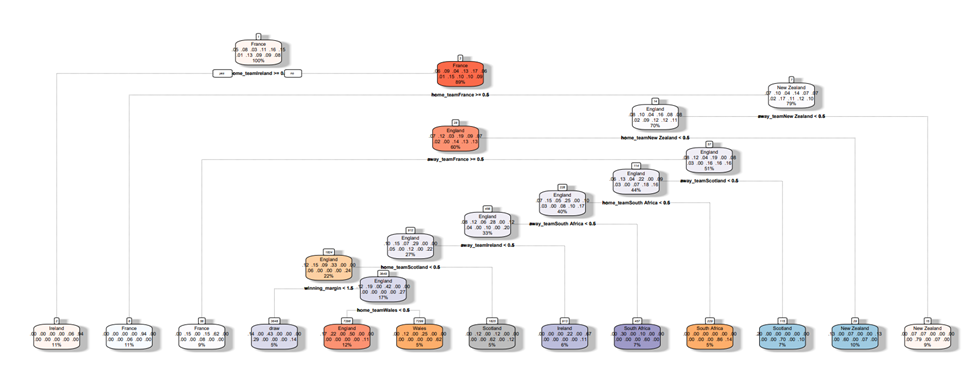
Figure 15: Initial Decision Tree Model using Data Since Last RWC



The results in the terminal nodes of the tree showed France at home winning 11% of the time and France away winning 9%, resulting in a combined winning percentage of 20%. Ireland winning 11% against France at home and 8% away from home for a combined 19% winning percentage. Wales winning 13%, South Africa winning 12%, Australia winning 11% of the time and a draw occurring 7% of the time. The model exhibited signs of overfitting, as evidenced by an illogical scenario where England was shown playing against itself when away. This clearly indicates a need for pruning to improve the accuracy and reduce the overfitting of the data.

To improve the model, a 10-fold cross validation was performed to determine the optimal complexity parameter (cp). This parameter, which balances tree complexity and accuracy during pruning, was evaluated in a sequence from 0.01 to 0.1, increasing by 0.01. Adjusting the cp of the model represents the cost associated with adding a new node to the tree. The optimal complexity parameter was found to be 0.02 increasing the model accuracy to 57.03% and modified the internal nodes and final decisions in the terminal nodes. Additionally, there was no indication of data overfitting.

Figure 16: Pruned Decision Tree Model using Data Since Last RWC (cp = 0.02)



The results in the pruned model terminal nodes showed France still held the highest combined home and away winning percentage of 20%. The first significant change was New Zealand's emergence in second place with a 19% combined winning percentage. Ireland came in third with 17%, followed by South Africa and Scotland, each with 12%. England also had a 12% winning percentage, while Wales and the occurrence of a draw both stood at 5%. In the Rugby World Cup (RWC) matches (quarter, semifinal and final matches) the outcome being a draw is unrealistic as there would be a period of sudden death where the first person to score would win the match, but it could occur in pool play which decides the quarter final match placements.

The analysis showed that France and Ireland are the top performing teams since the last Rugby World Cup. Association rules mining indicated that the home team has a significant advantage, and France, the host nation for the tournament, benefited from this. Focusing on the most recent top-performing teams, France and Ireland excelled at home, while France and New Zealand performed the best away, closely followed by Ireland. The Naïve Bayes model showed that using the criteria in the decision trees had a 53.33% accuracy in predicting the winning team, while the pruned decision tree model identified France as the top-performing team, followed closely by Ireland and New Zealand.

The analysis could be improved such as more information from match data included within the analysis. This could include the number of penalties from each participating team, the number of yellow or red cards from each team, and the number of injuries before and during the game could improve the analyses within the project. Other categorical information that could be included could be if the first- or second-string team was playing, the weather on the day of the match (raining, heat, wind speed) or the field formation the team adopts when playing in a match. This would not only aid in improving the current models but also widen the number of models one can apply to the data.

The analysis predicted France to win the RWC 2023, as they were the best-performing team since RWC 2019 and the host nation for RWC 2023. the decision tree models consistently identified France as the top performer in both initial and improved models. Consideration should also be given to teams like Ireland and New Zealand, which perform well away from home, with Ireland slightly outperforming New Zealand. South Africa had a strong incentive to defend their world title as the RWC 2019 champions. In reality South Africa ended upsetting both France and New Zealand to win the RWC 2023 showing that other data such as those mentioned above new techniques acquired could be applied to the analysis to improve the predicted outcome of the 2023 RWC.

## Deep Learning: Recursive Neural Networks In Financial Markets

The purpose of this project was to apply deep learning techniques to real world problems while comparing them to traditional machine learning techniques to not only explore this new state of the art neural networks, but also practice and compare them to classical machine learning techniques. Specifically using a recursive neural network to predict stock prices and compare them to traditional methods. Throughout the course we were encouraged to approach the problem with simple answers and models first, iterating to more complex models with more dense layers or other hyperparameters as needed. This allowed for the analysis to begin with traditional methods in the forms of a Random Forest Model and a Simple Neural Network to explore how machine learning and simpler neural networks performed compared to the more complex Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) RNN models.

The data was acquired using Y-Finance’s API where the past 10,000 days of trading data of a specific stock were queried using the API. The training data of interest was the date, the opening price of the stock, the closing price of the stock, the high and low prices of the stock during that date, the adjusted close price (close price including any modifications to the closing price such as stock splits and dividends), and the volume of trades for that stock on that date. The stocks chosen were the top performing companies from each of the stock sectors such as Apple (AAPL), Sherwin Williams (SHW), and Goldman Sachs (GS).

The neural networks were built in Python with Tensorflow an open-sourced end-to-end platform library for machine learning and Keras a high-level neural network library that runs on top of Tensorflow. The simple neural network used to compare with the Random Forest Model and the RNN’s was constructed with 16 dense layers, a ReLU (Rectified Linear Unit) activation function. The more complex RNN models were built with LSTM and GRU layers to overcome some of the traditional shortcomings of RNN models. Both LSTM and GRU networks can overcome the vanishing gradient problem allowing them to learn and retain long-range dependencies effectively. Both the LSTM and GRU models are similar with each using Sigmoid and Tanh Activation functions where the major difference between the two is the GRU model has simplified architecture making it faster to train while still being effective to capture long-term dependencies. The LSTM architecture consists of an input gate, a forget gate and an output gate where the Sigmoid activation function is used in the gates and the Tanh activation function is used in the cell state updates and the cell output. The GRU architecture is simplified by combining the forget and input gates into a single update gate in conjunction with a reset gate where the Sigmoid activation function is used in the gates and the Tanh function is used for the candidate hidden states. These functions combined with the architecture regulate the flow of information to maintain the stability of the training process. Both the LSTM and GRU models were tuned to have the same hyperparameters giving not only the best results but also uniformity for optimal comparison.

All the neural network models’ results were compared using the mean absolute error of each analysis. The mean absolute error (MAE) is a measure of the errors between paired observations expressing the same phenomenon, or how far way the prediction value is from the true value in dollar amounts. The Random Forest Model used an average accuracy for comparison where it performed the worst out of all the models tested in the analysis with an accuracy of 52.4%. The next highest performing model was the simple neural network with just dense layering and a ReLU activation function with an average MAE of 47.55. The second top performing model was the LSTM with an average MAE of 3.08 and the top performing model was the GRU model with an average MAE of 3.01, just beating the LSTM model. While the LSTM and GRU models outperformed the simple models, they are not suitable to be used as stock predictors. This is due to the complex and dynamic nature of financial markets where a combination of external factors influences the stock price. For example, the following visualizations use the best performing GRU model to show the training, validation and testing predictions from the model against the actual observations from the data where large deviations from the actual price and predicted price occur in both the validation and tests sets.

Figure 17: Training Predictions vs Observations using the GRU Model and AAPL Stock

A graph of a graph showing the growth of a training model

Description automatically generated with medium confidence

Figure 18: Validation Predictions vs Observations using the GRU Model and AAPL Stock

A graph of a graph showing the value of a model

Description automatically generated with medium confidence

Figure 19: Testing Predictions vs Observations using te GRU Model and APPL Stock

A graph of a graph showing the growth of a number of years

Description automatically generated with medium confidence

## Natural Language Processing: Predicting Sentiment using NLP Techniques

The primary purpose of the assignment was to train a Naïve Bayes Classifier Model to classify sentiment of movies reviews using natural language processing. The data was found on <https://ai.stanford.edu/~amaas/data/sentiment/> in the form of text files preorganized into folders with a train and test set of folder with both positive and negative movie reviews in both the train and test folders. The positive reviews were in one folder labeled as positive and the negative reviews found in a folder labeled negative. This allowed for data preprocessing in Python where the data was transformed from individual text files containing one review in each text file to a data frame where the folder names were used as the labels. Ethical considerations were observed throughout the preprocessing steps to remove bias from the analysis and all reviews were anonymized to protect privacy.

The text, now all in one column within a data frame, was tokenized using the Natural Language Tool Kit Python package. The tokenized data was normalized such as removing unnecessary punctuation and removing stop words. All aspects of data preprocessing, tokenization, data cleaning/ normalization, feature set building and experiment evaluation were written to be easily written and iterated throughout the experiment. This allows for easy comprehension of the code and iteration to be used or improved upon in the future.

Feature sets using the data were produced in the notation of the Natural Language Tool Kit such as a “bag-of-words” feature set where the data is represented by counting the frequency of the words in each document, and part of speech tags where words are tagged according to their grammatical categories. Additional features generated included word capitalization counts, review length, average word length, bigrams, the number of occurrences of "not" in a review, the number of negations, a VADER feature set using the Vader Sentiment Intensity Analyzer, and a TF-IDF (Term Frequency- These feature sets were used individually or in combination within experiments to identify the best-performing methods. The NLTK Naïve Bayes classifier, with 5-fold cross-validation, was employed to evaluate the experiments' performance by measuring precision, recall, F1 score, accuracy, and overall mean accuracy across experimental rounds, providing a comprehensive understanding of each experiment's results.

There were 12 experiments conducted throughout the project. The first experiment used the Naïve Bayes Classifier Model as the baseline, employing a bag-of-words feature set and all POS categories. Experiments 2 and 3 build upon Experiment 1 by removing varying amounts of text based on parts of speech tags to identify if there are any relationships between using tagged parts of speech and the performance of the classification. Experiment 2 iterated through each tagged part of speech tag, removing one tag per iteration to investigate the effect of individual POS tags on classification performance. If the performance drops after removing a part of speech, then it’s assumed that part of speech has a positive impact on classifying the sentiment of the reviews. If the performance improves then it’s assumed that the part of speech has a negative impact on the classification.

Experiment 3 uses a broader approach and removes all parts of speech tagged in the bag of words feature set where the only features left were not tagged as a part of speech to investigate if including tagged parts of speech reduces the model’s performance. Experiment 4 moves away from the bag of words feature set and creates a feature set using the frequency of the tagged parts of speech within the review to classify sentiment investigating if the volume of parts of speech give indication of positive or negative sentiment. Experiment 5 uses a feature set combining a feature set of the number of word capitalizations within a review, the average word length in each review and the average word length of each review for a comprehensive look at how these text statistics effect the performance of the classifier.

Experiment 6 uses bigrams generated from the review corpus as a feature set to see if the inclusion of bigrams results in any movement regarding the classification performance. Experiment 7 is similar to Experiment 8 where the focus is the investigation of negation within the reviews. Experiment 7 uses a feature set counting the number the word “not” appears in each review to see if the focus shares any insight into the classification performance. Experiment 8 uses a feature set of contradicted words to investigate if the negations within a review are a good barometer for classifying the sentiment of the reviews. In this case not all words that were contradicted were included on the most common words by frequency to determine if each review contained a negated or contradicted version of that word.

Experiment 9 is similar to Experiment 1-3 where a bag of words feature set is used except with TF-IDF tokenized words. The experiment should be ideal for investigating how tokenization affects the performance of classification. Experiment 10 uses the VADER third-party sentiment library to generate features with a sentiment value to investigate if using features with preassigned sentiment would improve the performance of the classification. Experiment 11 and Experiment 12 combine feature sets from the top performing experiments to see if an improved performance can be seen when classifying the sentiment. Experiment 11 combines the Vader features, TF-IDF features, and the bag of words feature set from Experiment 1 where all words were included. Experiment 12 combines Vader features, TF-IDF features, and the bag of words feature set but with nouns removed from the tagged parts of speech.

The results of each experiment were presented as a presentation to convey the results of the project. This allowed for stakeholders to digest the project from start to finish in an effective way highlighting both the successful and unsuccessful experiments. Communicating the results of the project as a presentation had the additional benefit for understanding the decisions made throughout the analysis. For example, stakeholders could understand the reason high performing feature sets and not the low performing feature sets were included in later experiments to iterate and improve the performance of the Naïve Bayes Classifier. The best result was found with an average accuracy 85% in the last experiment combing the Vader features, the TF-IDF features, and the bag of words feature set but with nouns removed from the POS tags.

## Text Mining: Filtering Resumes Using Text Mining

With the introduction of automated hiring tools, algorithmic processing of resumes has become common practice in today’s job market. The methods and models have transformed the job market and recruitment landscape. New opportunities and challenges have arisen from this transformation that affect both organizations and job seekers. The introduction of these methods is used to streamline the hiring process by reducing the time and resources spent on sifting through countless resumes to identify the most promising candidates. The automation process relies on machine learning models to optimize the recruitment process. Two models for consideration are Latent Dirichlet Allocation (LDA) and K-Nearest Neighbor (KNN) for their capacity to revolutionize resume screening, offering nuanced insights that go beyond traditional keyword matching.

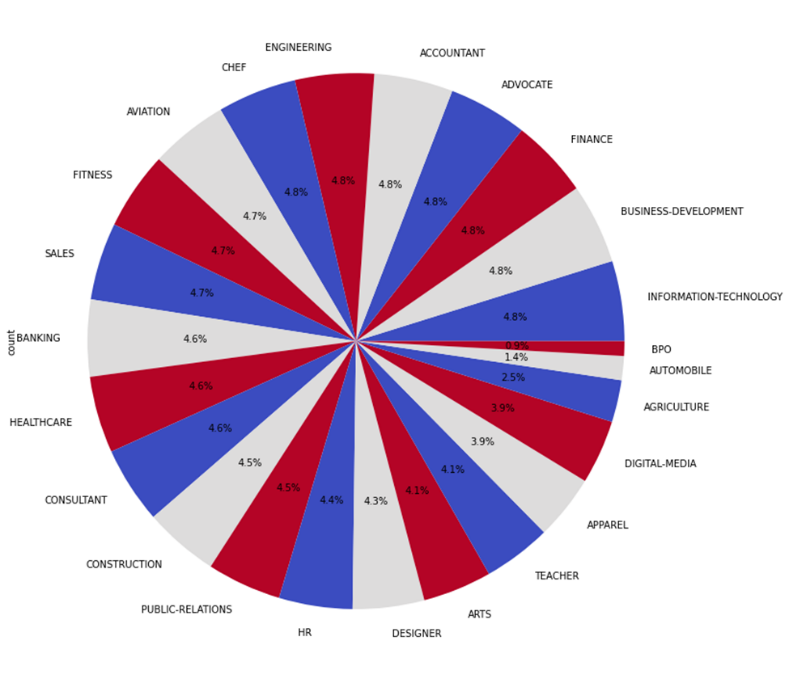
Latent Dirichlet Allocation (LDA) is a machine learning algorithm that excels at uncovering the underlying topics within a collection of documents. When applied to resumes, LDA can identify patterns and themes that highlight a candidate's skills, experiences, and professional interests. This capability provides hiring managers with a comprehensive understanding of a candidate's expertise that goes beyond simple keyword searches.

The K-Nearest Neighbor (KNN) model offers a complementary approach by focusing on the similarity between candidates. When a resume is added to the dataset, KNN evaluates it in the context of existing resumes that have been tagged as successful hires. By identifying the most similar elements within the resume pool, KNN can predict the likelihood of a candidate's success in a given role. This method is particularly useful for roles with well-defined and consistent requirements, as it allows for the comparison of candidates based on real-world outcomes. When used alongside LDA, KNN can refine its predictions by considering thematic similarities between candidates, leading to a more nuanced understanding of fit beyond just qualifications or experience.

This project aims to utilize the approaches of both models and conduct a comparative analysis using a set of resumes from various industries and categories. The data set contains a diverse set of 3,446 resumes, where both LDA and KNN models will be used in a two-step process. First, LDA will analyze the resumes to identify underlying themes, categorizing candidates by their skills and expertise. This thematic analysis provides a deeper understanding of each applicant's professional background. Then, the KNN model will compare a training set of resumes against a testing set to find the most similar profiles. By combining LDA's thematic insights with KNN's similarity-based predictions, we aim to streamline the screening process and observe how the two models perform in enhancing the resume screening process.

The text data was found on Kaggle where two different data sets were combined for a comprehensive collection of resumes. The resumes were categorized by industry, job title, and other keywords. The data cleaning began with concatenating the data samples and reading the data in through two separate steps for data uniformity. This includes ensuring the data is uniform and clean after combining the two sources, creating a pie chat to observe the categories of resumes found within the data, and word clouds were used to explore the top words of resumes with respect to each category. Functions and operations used to clean the data and apply the models mentioned above use Python and include sklearn, nltk, PyPDF2, wordcloud, numpy, pandas, seaborn and os libraries.

Figure 20: Pie Chart of the Resume Distribution by Job Title



The text from each resume was labeled by category then instantiated with a TF-IDF Vectorizer which step tokenizes, counts and computes the TF-IDF vectors for the resumes. The vectorized resumes are used as features and the resume categories as the target variables. The data is split into training and test sets where 80% of the data is used for training the model and the rest is used to test the model for performance. The training and test sets were applied to both models of interest. The KNN model is evaluated using accuracy scores and a classification report including the precision, recall, and F1 scores. The LDA model is fitted and used to assign topics to each resume using the probability distribution of topics found using the LDA model. The topic distributions were used as features for building categorization models, the first being a Support Vector Classifier and the second being a Random Forest Classifier. The performance of these models is evaluated using the accuracy of the model with and without the LDA model topics as features to determine if using the LDA model topics as features can improve the classification of the resumes when using more traditional techniques.

The KNN model provided a set of precision, recall, F1 score and support scores for each category, training over the set of resumes in each category and providing a probabilistic result. An assortment of categories, such as Testing, Web Designing, Python Developer, Database and DevOps engineer provide perfect Precision and Recall scores, implying that the samples have consistent, yet the accuracy scores only show high performance on the notably smaller samples. In the overall results, the KNN model shows a 67% precision and 65% recall, implying low false positives and moderate to high success in accurate recall over its iteration through the testing set of our data, which, when split, included a total of 690 resumes identified.

Table 8: Results of the KNN-Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Precision | Recall | F1-Score | Support |
| ACCOUNTANT | 0.5 | 0.67 | 0.57 | 24 |
| ADVOCATE | 0.5 | 0.46 | 0.48 | 24 |
| AGRICULTURE | 1 | 0.15 | 0.27 | 13 |
| APPAREL | 0.62 | 0.26 | 0.37 | 19 |
| ARTS | 0.6 | 0.29 | 0.39 | 21 |
| AUTOMOBILE | 0.67 | 0.29 | 0.4 | 7 |
| AVIATION | 0.78 | 0.3 | 0.44 | 23 |
| ADVOCATE | 1 | 0.75 | 0.86 | 4 |
| ARTS | 1 | 1 | 1 | 7 |
| AUTOMATION TESTING | 1 | 1 | 1 | 5 |
| BANKING | 0.6 | 0.39 | 0.47 | 23 |
| BPO | 0 | 0 | 0 | 4 |
| BUSINESS-DEVELOPMENT | 0.33 | 0.62 | 0.43 | 24 |
| BLOCKCHAIN | 1 | 1 | 1 | 8 |
| BUSINESS ANALYST | 1 | 0.83 | 0.91 | 6 |
| CHEF | 0.79 | 0.62 | 0.7 | 24 |
| CONSTRUCTION | 0.63 | 0.77 | 0.69 | 22 |
| CONSULTANT | 0.43 | 0.13 | 0.2 | 23 |
| CIVIL ENGINEER | 0.71 | 1 | 0.83 | 5 |
| DESIGNER | 0.65 | 0.62 | 0.63 | 21 |
| DIGITAL-MEDIA | 0.59 | 0.53 | 0.56 | 19 |
| DATA SCIENCE | 0.86 | 0.75 | 0.8 | 8 |
| DATABASE | 1 | 1 | 1 | 7 |
| DEVOPS ENGINEER | 1 | 1 | 1 | 11 |
| DOTNET DEVELOPER | 0.67 | 1 | 0.8 | 6 |
| ENGINEERING | 0.68 | 0.62 | 0.65 | 24 |
| ETL Developer | 1 | 1 | 1 | 8 |
| ELECTRICAL EGINEERING | 1 | 1 | 1 | 6 |
| FINANCE | 0.56 | 0.58 | 0.57 | 24 |
| FITNESS | 0.55 | 0.48 | 0.51 | 23 |
| HEALTHCARE | 0.42 | 0.57 | 0.48 | 23 |
| HR | 0.67 | 0.77 | 0.72 | 31 |
| HADOOP | 1 | 1 | 1 | 8 |
| HEALTH AND FITNESS | 1 | 1 | 1 | 6 |
| INFORMATION-TECHNOLOGY | 0.57 | 0.71 | 0.63 | 24 |
| JAVA DEVELOPER | 0.81 | 1 | 0.89 | 17 |
| MECHANICAL ENGINEER | 1 | 1 | 1 | 8 |
| NETWORK SECURITY ENGINEER | 0.83 | 1 | 0.91 | 5 |
| OPERATIONS MANAGER | 0.8 | 1 | 0.89 | 8 |
| PMO | 1 | 1 | 1 | 6 |
| PUBLIC-RELATIONS | 0.52 | 0.68 | 0.59 | 22 |
| PYTHON DEVELOPER | 1 | 1 | 1 | 10 |
| SALES | 0.33 | 0.52 | 0.41 | 23 |
| SAP DEVELOPER | 0.71 | 1 | 0.83 | 5 |
| SALES | 1 | 1 | 1 | 8 |
| TEACHER | 0.55 | 0.8 | 0.65 | 20 |
| TESTING | 1 | 1 | 1 | 14 |
| WEB DESGINING | 1 | 1 | 1 | 9 |
| Accuracy |  |  |  | 0.65 |
| Macro Avg | 0.75 | 0.73 | 0.72 | 690 |
| Weighted Avg | 0.67 | 0.65 | 0.64 | 690 |

The LDA model found ten topics to apply to the previously vectorized resume texts. An example of the topics generated from the LDA model can be found below visually represented in a word cloud. The topics found from the LDA model were then used for building the two categorization models. The number of topics used in the LDA model iterated across 10, 100, 500 and 1000 topics. When training the Support Vector Classifier (SVC) with the LDA determined features, the accuracy was found to be 0.1174 or 11.74%, an abysmal performance, and when using an SVC without the LDA model topics as features the accuracy increased to a more respectable 0.7391 or 73.91%. The same result was found using the Random Forest model without the LDA topics as features where the accuracy was 74.05%.

Figure 21: Example of a LDA Generated Topic

A close up of words

Description automatically generated

To try and improve the models using the LDA derived topics, the number of topics generated was set to 500 showing some improvements. The accuracy using the 500 LDA generated topics increased to 27.97% and 53.76% for the SVC and Random Forest model respectively, but nothing reaching the baseline of using the classification models traditionally without the LDA topics as features. Other iterations of trying to improve the categorization models using the LDA topics include setting the model with 100, and 1000 topics. The accuracy when using 100 LDA topics was found to be 25.07% accuracy for the SVC, 45.22% accuracy for the Random Forest and the accuracy when using the 1000 LDA topics was found to be 14.2% accuracy for the SVC, 32.75% accuracy for the Random Forest. These iterations did not show improvement when using 500 topics found using the LDA model and still performed worse than traditionally using the SVC and Random Forest Models.

Finally, applying the LDA model topics as features when building the categorization models did not turn out to be a fruitful endeavor. Using the models in a more classical analysis, simply using the vectorized data without incorporating the LDA topics, would be a more efficient use of one's time. Thus, tunning, adjusting the data processing, or parameters of the Support Vector Classifier (SVC) and Random Forest model using the vectorized data would be the most beneficial for improving the classification of the resumes. Additionally, another potential leverage method that could have been recognized was training the KNN model over the LDA model, which would provide us with a point of reference and allow for stronger accuracy for the LDA as well as KNN model. Some sources of improvement primarily lie in the stronger and further optimization of the LDA model, and it lies in finding the right stop words for the program. Increasing the number of stop words may provide us with an output that would grant us the most coherent set of words, not only for the word clouds visibly, but also for the KNN model to be accurately trained over. An increased data set can also greatly improve the model's ability to train over the set of resumes and provide us with a more effective filtration of the resumes.

# Conclusion

In conclusion, the diverse projects highlighted underscore the pivotal learning outcomes throughout the Applied Data Science Master’s Degree program. The consistent theme across these projects is the ability to effectively collect, store, and access data collection, whether through Api’s for financial data, Google Analytics for marketing insights, or text mining techniques for resume filtering. Additionally, ensuring the ethical handling of data and information was a fundamental aspect of the program further reinforcing the importance of ethics in data collection and analysis.

Another key learning outcome was the creation of actionable insights across a range of contexts. From marketing strategy optimization to predicting Rugby World Cup winners, the application of analytical techniques and predictive models was essential. Each project used various tools and methodologies to derive meaningful insights. For instance, the use of machine learning techniques and decision trees allowed for robust segmentation and prediction analyses, enhancing business intelligence and decision-making processes.

The ability to apply visualizations and predictive models to generate actionable insights was demonstrated through the effective use of programing languages, particularly R and Python. These programming skills supported the development of complex models and visualizations that facilitate the deeper understanding of the data. Visualization played a crucial role in communicating insights, with graphical representations and tables being used to convey findings to a broader audience. These skills are vital for not only ensuring data driven insights are found, but also accessible and actionable for stakeholders.

In summary, the projects collectively highlight the skills developed encompassing data collection and storage, actionable insight generation, programming proficiency, effective communication through visualizations, and ensuring ethical considerations are following best practices for maintaining trust and integrity. These competencies are critical for applying data science across disciplines including the effective application of business analytics, the development of enhanced marketing strategies, predictive models, and data-driven decision-making processes. The ability to integrate these skills across various projects demonstrates a robust understanding of data science applications in real-world scenarios. All projects discussed in this portfolio can be found [here](https://github.com/LoganRoach/Syracuse-Applied-Data-Science-MS-Portfolio).