Adam Vajdak

[avajdak@syr.edu](mailto:avajdak@syr.edu.com)

SUID: 780108296

Milestone Paper

Over my time in the Syracuse Data Science program, I have learned a variety of applications and techniques to explore and show data such as R, SQL, Python, along with a variety of BI programs such as Tableau, Power BI, and other free tools. Throughout this paper, I will explain how I learned about managing data, creating insights, applying data visualizations and predictive models to create insights, and communicating those insights with visualizations.

For my first semester, I learned for the first time how to use R to create visualizations and to use the software for basic statistical analysis of data. (687 Project)The first data set I worked with was one of the many movie datasets from Kaggle. The data needed to be cleaned and fixed for inconsistencies. We used the basic Tidyverse package to create a couple of visualizations with ggplot to visually explore and show the data in a presentation. The strongest correlations were with votes and revenue, revenue and budget, and votes and budget. We then attempted to model those values with regression analysis using linear models to achieve an R squared of 72%of revenue explained by budget popularity and vote counts.

Alongside this, I was taking my Business analytics class which showed me snippets of deeper learning algorithms through R commander using basic applications of Linear regression, machine learning, and neural networks. Although this class did not have a project I learned a lot from this along with reaffirming business ethics to clean the data correctly to not only protect those in the data from malicious but to remove biases from the data to ensure integrity. We used neural networks with the Titanic data set to create models and evaluate them alongside Excel, for applications such as sentiment analysis, and power BI.

My first interaction with Python was in class 652 scripting for data analysis. For the project, we had to combine two data sets to answer questions. My group decided to combine the global happiness dataset on Kaggle alongside the internet usage dataset to see if there was anything we can learn from the internet usage to the happiness statistic. While the project is not highly technical it showed me that there is not a significant difference in coding as the concepts are very similar to each other and it's best to understand the concepts than trying to remember the exact code needed for processes. We joined multiple data sets before our final merged dataset that was regions, internet usage, and the global happiness sets. There was a higher correlation between wealth and internet usage affecting happiness. I expected there to be some negative correlations between the internet and happiness with how it has been used for negatives the last couple of years but it still shows a positive correlation. Our strongest factors for happiness were a country's social support followed by GDP, Income, and then internet user rate.

For the use of models and predictive analytics to find answers my larger project working on a data set was to predict which segment of the existing market will new customers be in based on older data of current customers. (707 projects) We were tasked with creating 4 segments and lumping new people into these segments. The professor recommended we just try as many possible ways of figuring out the data as we can which prompted us clustering, association rule mining, decision trees such as K nearest neighbors and random forest, SVM, and deep learning models. This project taught me a lot about different ways to apply data to models and evaluate models on their performance parameters and gave me a better understanding of when to apply which models for certain datasets as I watched my classmates present their projects.

To help demonstrate my ability to communicate my findings through visuals would be my time in 719 Information Visualization. ( 719 folder) the project shows that I have used R code with Adobe Illustrator to create a Poster to be able to present the finding and information of the data set at a glance. This project helped me focus more on how to present data more efficiently and avoid cluttering it up with non-relevant information such as removing excess lines and backgrounds while maintaining a consistent theme and color across charts.

In conclusion, this program has taught me a lot of valuable skills when it comes to data science. I am eager to use them and expand my skills more as I understand we mostly just glossed over subjects in class and there is a lot more depth to be had in this field and more to know. This is not the end of my learning journey but only another start.