Portfolio Milestone

Ashraf Wan

SUID: 273192117

https://github.com/ashwan01/datascience-portfolio

Table of Contents

Introduction

The Applied Data Science program at Syracuse University's School of Information Studies is designed to have students master the fundamental aspects of data science through project-based research and deliverables. Each student is to demonstrate mastery in data collection, data analysis, and implement business decisions. Through this portfolio, I will showcase my understanding of the different aspects of data science through different courses and different tools.

IST 707 – Applied Machine Learning

Project Description

The main purpose of this project was to perform a customer personality analysis based on a grocery store's dataset and find the ideal customer to target for a marketing campaign. The dataset has information on customer's demographic, purchase history, reaction to past promotions and place of purchase. My teammate and I acquired the dataset from Kaggle. We identified missing values and outliers in our dataset, so we removed those to get a more accurate result from our analysis. We utilized various model prediction, k-means clustering and Apriori Algorithm to find the ideal customers for the marketing campaign. Halfway into the project, the dataset couldn't give us the answer we needed to support our goals. However, the algorithms we ran were able to accurately tell us customers who are more likely to not accept marketing campaigns. My teammate and I realized there's a huge untapped potential of customers and decided to change the focus to those customers.

Technologies and Techniques

- Apriori Algorithm in both WEKA and R
- K-Means Clustering in R
- Decision Tree in R
- Naïve Bayes in R
- Random Forest in R
- SVM in R
- KNN in R

Reflection and Lesson Learned

One of the main lessons I learned from this project is to not be to focus on my conclusions even before I run the calculations. Sometimes the dataset just won't fit into my predetermined goals, and I should be flexible enough to change the goal or business question based on what my data is telling me.

IST 719 – Information Visualization

Project Description

For this project, students were asked to choose a dataset and create a poster to showcase the results through visualization. The dataset was on the Olympic game's performance by each country for both the summer and winter Olympics. It had information on total medal counts for each of the Olympics and total for both, country names, and IOC codes since 1896. I acquired this dataset from Kaggle and filled out any missing information using Wikipedia.

Technologies and Techniques

- R
- ggplot
- rworldmap
- Adobe Illustrator

Reflection and Lesson Learned

One of the main issues I ran into was the dataset uses IOC codes to represent the countries while rworldmap uses ISO codes to represent countries. I had to go through each country and make sure the IOC and ISO codes were matching so I could create the world map properly. It wasn't difficult to do but was time consuming. If I was in the same situation again in the future, I'd rather look for a dataset that have both IOC and ISO codes so that I could save some time.

IST 736 – Text Mining

Project Description

Sentiment analysis and topic modeling were the focus for this project. The dataset is on financial news headlines from 2014 and the purpose was mostly to find out the sentiment of the headlines or topics to help investors make the best decision while investing. The dataset was simple and only had 2 columns which are the headlines and sentiment. There was an imbalance of sentiments total so the same number of rows from each sentiment were randomly used. Then multiple modeling algorithms were used for classification.

Technologies and Techniques

- Bernoulli Naïve Bayes in Python
- Multinomial Naïve Bayes in Python
- SVM in Python
- Topic Modelling in Python

Reflection and Lesson Learned

The topic and dataset for this project was straightforward so I didn’t run into any problems. I had to run multiple tests for each of the modeling algorithms with different parameters to get interesting results for reporting.

MBC 638 – Data Analysis & Decision Making

Project Description

The goal of this project was to showcase our understanding of various analytics techniques and excel usage abilities using dataset we collect ourselves. The only condition for the dataset is that the data can be used to show improvement in ourselves. The dataset I decided to use is my own personal bank statements because I wanted to show the improvement in my monthly savings.

Technologies and Techniques

- Descriptive Statistics in Excel
- Chi-squared Test in Excel
- Correlation in Excel
- Exponential Smoothing in Excel

Reflection and Lesson Learned

Having to use data on myself was an interesting experience. Data collection was easy because I already have all the data and if I was missing anything, all I had to do was pull up my bank statements.

Conclusion

This portfolio has demonstrated the various implementation of the learning objectives using different tools from Python to R to Excel and to and outdated tool such as WEKA. Majority of the data were collected through Kaggle and any missing information is filled in using other sources. The datasets were analyzed using statistical methods and data mining techniques such as clustering and classification and presented using visualizations. Not only were these skills showcased for individual projects, but these were also showcased successfully in a group environment as well.