School of Information Studies, Syracuse University
M.S. Applied Data Science

Project Portfolio Milestone

Yeswanth Reddy Velapalem
SUID: 204882653
yvelapal@syr.edu
https://github.com/yeswanthrs9/Portfolio-Milestone

[Portfolio Presentation Link]

Table of Contents

Introduction

Data Science is a trending field in the business and tech industry. In this field, you can either stop at a certain point or can go on exploring further, yet it still is not enough to learn all the concepts. What makes it interesting is the un-ending curiosity to analyze the data in multiple angles and improve the decision making by discovering those concealed trends and patterns which can hugely impact the growth of an organization. I became familiar with the concept during my internship and immediately knew I had to upscale my skill in the field of data science. This curiosity led me to pursue my master's degree in data science. The M.S. Applied Data Science course curriculum is designed to meet all the necessary requirements covering all the tools and technologies which help the learner succeed in this ever-growing industry.

While pursuing my master's, I was able to work on various assignments and projects which transformed me into a better learner. I listed some of the most important projects that showcases my best work in this course.

IST 687: Introduction to Data Science

I took Introduction to Data Science in my first semester to start my master's degree. This course introduced me to fundamentals about data and methods for organizing, managing and using data. It also helped me explore some key concepts like information visualization, text mining and machine learning.

Project Description

For the project, we were assigned into a group and were given Customer data of Airline Industry. The goal of the project is to reduce the customer churn by getting ahead and identifying some leading indicators, or metrics and recommending ways to improve the customer satisfaction. The data consisted of 10282 observations and 28 attributes. Before cleaning the data, we decided to focus on flights that are not cancelled, since customers might not be able to rate their experiences. The first process of cleaning the data involved formatting the datatypes of some attributes and then scaling to bring the data under a single range of values. Then came dealing with missing values where we replaced the missing values with mean of that respective attribute.

We then split the data into three major groups based on the attribute 'net promoter score'. We visualized the cleaned data to identify the trends and patterns to find the factors influencing customer satisfaction. We also plotted a correlation matrix to find how the attributes change w.r.t to each other. Four data models were deployed using the data to find the factors which highly influence the satisfaction of the customer. The four models were Linear Regression, Association Rule Mining, Support Vector Machine, Logistic Regression. From the visualizations and data modeling we discovered some actionable insights which were communicated to the clients to help them improve their business and customer satisfaction.

I used statistical programming language R and RStudio to achieve all the objectives of the course and the project.

Reflection & Learning Goals

Through this project, I was able to learn the essential concepts of data science and scripting for data curation and management using R and R studio. I also learnt the common practices in data wrangling and data visualization along with some Machine Leaning techniques and how important it is to communicate the discovered actionable insights to the clients/decision making bodies.

Github link: https://github.com/yeswanthrs9/customer-churn-in-airline-industry

IST 718: Big Data Analytics

This course helped me by introducing analytical processing tools and techniques. It also talks about the high-level math concepts behind data science techniques. I had the opportunity to build data pipelines involving various machine learning algorithms which has taught me how to work with big data and its aspects.

Project Description

For our project, we used online news articles collected by Mashable to predict whether an article is going to be a hit or not before publishing and also predicting the number of shares it will receive on social media. This project is a case of Content Optimization technique. The news article can be indicated by number of reads, likes and shares. We had approximately 39644 articles and 61 variables. It is treated as a binary classification problem and the objective is to find the best classification learning algorithm to accurately predict if a news article will become popular or not before publication. Also, data is treated as a regression problem to predict the number of shares an article can get.

The data is preprocessed for missing values and outliers are dealt using the IQR method. Then a threshold was set for number of shares to decide if a news article is popular or not. Then exploratory data analysis is performed to find out the variation of number of shares w.r.t other attributes and find the factors which affect them the most. After this we did feature engineering to select the best attributes that contribute to the popularity of the news article. We built three pipelines where learning algorithms like Logistic Regression, Random Forest Classifier and Gradient Boosting Classifier are implemented. We also tuned the hyper-parameters to find the best possible parameters of the algorithm which improves the AUC score. Since this is also a regression problem, we built a Random Forest Regressor pipeline to predict the number of times an article will be shared. The hyper-parameters are tuned for regressor to improve the RMSE. I used Python and Apache Spark to leverage advanced analytics and gain actionable insights which helps create an advantage.

Reflections & Learning Goals

This course taught me how to use various tools and techniques to build efficient data pipelines to implement learning algorithms on data. The various results obtained from the models is an example of the importance of testing different techniques to develop the most simple and accurate prediction models. Testing alternative strategies and weighing the benefits of each technique can reduce computation costs and provide the greatest precision in data analytics tasks. This project contributed to the successful application of the learning goals through the development of alternative strategies based on the data, and the communication of observations which translate to actionable insights. Extensive Data Analysis was also used in conjunction with visualization to identify patterns in the data for use in classification tasks.

Github link: https://github.com/yeswanthrs9/Online-News-Popularity

IST 736: Text Mining

Text mining introduced me various concepts and methods to tackle large amount of text data and the application of text mining techniques for business intelligence and social behavior analysis.

Project Description

For this course project, we gathered data from online review sites which contain a lot of information regarding user preferences and experiences over multiple product domains which can be used to obtain valuable insights. Here, we examined online user reviews within the pharmaceutical field which contain information related to multiple aspects such as effectiveness of drugs and side effects, which made analysis very interesting but also challenging. Our main goal was to check the effectiveness of using Sentiment Analysis which could detect the sentiment of the review and hence be in agreement with the rating classification. There were multiple reviews for the drugs that belong to a similar condition and we investigated how the reviews for different conditions used different words and how they impacted the ratings of the drugs.

After dealing with missing values and incorrect data we filtered the reviews using regular expressions which helped to remove the non-alphanumeric characters. Then the exploratory data analysis performed on the data gave us the trend of drug rating over the years for various conditions that helped us understand the data in a clear way. We performed topic modeling to identify the important words/topics to understand why they garner huge number of shares. Since we were dealing with a multiclass classification problem, we selected required evaluation metric to validate the performance of the learning algorithms. Also, we used a Snowball stemmer to stem the reviews and stop words were removed from the reviews.

For data modeling, we implemented Naïve Bayes and Support Vector Machine algorithms by varying the type of vectorizer in the vectorization process. The models were also experimented by changing the n-grams and compared how they fared with each other. Along with sentiment analysis, Error analysis was also performed on the reviews to detect why the reviews were wrongly

predicted. This helped in understanding the necessary improvements that we needed to do to our learning algorithm to improve the evaluation metric.

Reflection & Learning Goals

This course gave a clear knowledge about text analysis and how raw text reviews can be modified and used to obtain data-driven insights. I learned text classification, clustering, topic modeling and how to use text mining concepts to develop and evaluate effective solutions. Analyzing text data is as important as numerical data since data like reviews can help a lot of companies to grow by actively leveraging the inputs of customers.

This project provided the opportunity for the collection and structuring of externally sourced data, identification of patterns within the text data. All the data we collected is public, but considerations must be made to ensure that only the relevant information is used to both balance user limitations and privacy. Text data is incredibly important to marketing analytics teams as more unstructured sources are introduced. With more content being created by customers such as reviews, social media posts, and transcriptions, the value of extracting actionable insights from text is growing in significance. As organizations project a greater social media presence, the ability to organize and analyze large collections of text allows for automation using conversational assistants, as well as predictive analytics with text mining.

Github link: https://github.com/yeswanthrs9/Sentiment-Analysis-on-drug-reviews

CIS 731: Artificial Neural Networks

This course introduced me to advanced concepts of learning algorithms and the concept of deep learning along with the mathematical concepts that drive those algorithms. I learned the functions and limitations of neural networks and custom learning algorithms in a detailed manner.

Project Description

The project for this course involved Multivariate time series weather data where we worked to forecast the temperature of a city by taking 8 years of historical data. We had a total of 420,000 observations and 14 variables related to weather forecasting. Along with univariate predictions we also made multivariate predictions using the data. The purpose of the project is to incorporate and compare different forms of Neural Networks. We compared the architectures of these networks on the basis of a loss function (mean squared error) and the computational effort required to train the models.

Initially we did a statistical analysis of data to explore the range of different variables. Adding to that we also made visualizations which explained the shifting trends and correlation between variables. The statistical analysis conducted was very helpful to detect the incorrect data and presence of outliers. We cleaned the data appropriately and did feature engineering to select

desired variables which solved the issue of multicollinearity between the variables .Then we proceeded with the sequence generation process. We generated three input sequences of data to experiment multiple scenarios. First two were only to predict the temperature at a single point in time. The third sequence is generated so that we can predict the data over 24 hours. To achieve this, we added an offset in the sequence generation process.

Once we had the input sequences, we experimented with a large number of Neural Network architectures such as Feed Forward Networks, Recurrent Neural Networks, LSTM's, Gated Recurrent Networks and Sequence to Sequence Models. We did both univariate and multivariate forecasting with each of these architectures. We optimized all the architectures by tuning the hyper-parameters such as number of layers, number of epochs and hidden layers.
Convolution Neural Networks were also built for forecasting the weather. Usually, CNN's are used for 2D data, but they can be modified to be used with 1D data. We made changes to successfully fit the data and forecasted the temperature using CNN's for both univariate and multivariate cases. All the results were compared and then evaluated on the basis of Mean Squared Error and steps were taken to make sure the models don't overfit. I used Python, keras and PyTorch to achieve all the objectives of this course.

Reflections & Learning Goals

This course taught me how to build Neural Networks for new problems and to modify the architectures to achieve desired results. I also learned how to evaluate the capabilities and limitations of different learning algorithms.

Github link: https://github.com/yeswanthrs9/CIS-731

Conclusion

During my Master's, all the courses that I have taken and the assignments and projects for those courses, taught me all the fundamental concepts along with the advanced techniques of data science. I have learned to collect, organize, curate and transform data. I have learned to visualize the data to bring out the hidden trends and patterns which help us to understand the data. I am able to perform statistical analysis which reveals the trends and abnormalities in numeric data. I have learned to leverage these results obtained and develop strategies and propose an action plan to implement these strategies. I developed good communication skills when presenting my projects and explain the trends in the data along with the ideas I developed to solve the desired problem. The ethical values of data science were reinforced when I was trying to collect data from sources by scraping only relevant data and also considering the privacy of the user when using personal data. This program gave me confidence and equipped me with skills which I can use to tackle a wide range of problem and help in the growth of the company by providing growth-driven solutions.