School of Information Studies, Syracuse University

M.S. Applied Data Science

# Project Portfolio

Sai Praharsha Devalla

SUID: 235239596

sdevalla@syr.edu

https://github.com/saipraharsha/Portfolio

Table of Contents

## 1. Introduction:

'I have no special talent. I am only passionately curious.' - Albert Einstein. This what makes me drive towards this path. The world is one big data problem which I intend to solve few of them. My aptitude for mathematics, analytical ability, and the urge to work hard to acquire knowledge on new things paved my path to pursue my studies in the field of Statistics and Computer science i.e., Data Science. Data science is a trending now in the tech industry as it improves the decision making by identifying the hidden patterns and trends which affect the upcoming of the organization. I felt there is a need to upscale my skills after working for two years as a systems engineer at Infosys to thrive in the data industry. So, I opted for the program M.S. in Applied Data is due to the way the course structure is designed which meets the industrial requirements. The current industry uses SQL server, Apache Spark, Tableau, Power BI, Python, R and excel. So, all these tools span from storing of data to extraction, transforming and loading.

A successful student in the Applied Data Science program is expected to demonstrate the goals and outcomes of the course and I take this opportunity to demonstrate some of my projects which intends to show my best work in those courses.

## 2. IST 687: Introduction to Data Science:

**Project Description:**

In my first semester I undertook the course Introduction to Data Science, this gave me a glimpse what data science has is offering in upcoming times. This course is built in a way where all the fundamental aspects of data science are covered. It taught me to extract, transform, visualize, and clean the data. This course introduced me to analyze the patterns and trends using Machine Learning. This course has also introduced to a new scripting language 'R' which is widely used for analyzing the patterns and visualize the data.

For our final project, I was assigned to team and was given an Airline Customer Churn dataset. Our main goal is to improve the customer satisfaction of the airline by analyzing the customer data through some recommendations. The dataset for Airlines consists of 10282 rows and 32 variables. The dataset is cleaned by removing the 'NAs' and 'NULL' from data, some 'NAs' are corrected by imputing them with either mean or median of their respective attribute.

For the entire project we used statistical programming language R. We identified trends and patterns in the customer data by visualizing the variables with customer Satisfaction variable. We had identified which attributes account the most in classifying the customer satisfaction. We built 4 models' Linear model, Association Rule Mining, Support Vector Machine and Random Forest which are evaluated by accuracy and recall. We draw out interesting insights by visualizing the patterns and implementing the different models. These insights were presented to clients so that

they can improve the airline quality by improving the customer satisfaction which will enhance their business.

**Reflections & Learning Goals:**

From this course I gained valuable experience in statistical programming language 'R' which intrigued me to become a Teaching assistant for this course for undergrads. In this course I was introduced to Linear Models, SVMs and predictive modelling. I gained on hands-on experience in visualizing the patterns and trends throughout the tenure of the class. Upon performing Exploratory Data Analysis which can termed as Identifying patterns in data to obtain valuable insights, which helps in understanding of the underlying trends of the data.

GitHub Link: https://github.com/saipraharsha/Portfolio/tree/main/projects/IST%20687

## 3. IST 718: Big Data Analytics:

**Project Description:**

This course uses big data platforms which introduced me to Apache Spark, much needed skill that should be possessed by a data scientist now. This course helped me by introducing analytical processing tools and techniques. We got to learn the architecture of spark which gives in depth application of how to use map-reduce concepts with RDD's. As per the course structure we were introduced to the unsupervised and supervised learning models like Linear, logistic, decision trees, random forest, support vector machine and PCA using various case studies. I even had the opportunity to build data pipelines using various ML algorithms which were taught in the class. I did Online news popularity Classification project to fulfill the course requirements.

With the growth of the Internet in daily life, people are in a minute away to read the news, articles increasing online platforms rivalry. Due to this, every online platform is striving to publish the articles on their site which have great value and bring most shares. Project Online News Popularity intends to analyze the Mashable dataset which consists of articles data information mainly as several unique words, number of non-stop words, the postpositive polarity of words, negative polarity of words, etc. Here we intend to provide the maximum numbers of times particular article has been shared. Thus, helping platforms to realize true market potent of an article and use the information to decide which articles should be published.

For our project we used the data collected by Mashable to classify whether an article will be hit or not before publishing. This project is a classic example of content optimization. The data consist of 39644 observations and 61 attributes. Out of 61 attributes, we have 58 predictive attributes, 2 non-predictive and one goal field. The target variable 'shares' is a numeric attribute which gives us the number of times a news link is shared online. This is a regression problem, but we intend to transform it into a classification problem by setting a decision threshold on shares attribute.

Approach begins with identifying Data Quality issues and pre-processing data which includes handling missing/NA values, deleting duplicate values, filter unwanted outliers and treating special characters, categorical data, we used IQR method to remove the outliers. Data Cleaning will be followed by Exploratory Data Analysis using descriptive statistics, working on target variable, and standardizing attributes. We try to find out what attributes contribute the help classify the articles. After this we did feature engineering to select the best attribute that contribute to the popularity of news articles. Next, we will split the dataset into training and testing datasets. The training set will be used to build machine learning models and the test set will be used to validate the output of the models. We built three classifiers Random forest, Logistic and gradient boost models. These models are evaluated based on the parameters of accuracy, AUC score and recall. As this is a regression problem, we built Random forest Regressor model, which is evaluated by MSE and RMSE. The hyper parameter for both classifier and regressor is tuned to improve the scores.

**Reflections & Learning Goals:**

This course taught me to build efficient data pipelines to using the ML algorithms. This also helped me to learn a lot about Big data, Map reduce and how data is injected from big data platforms to Spark to perform analysis. Upon performing Exploratory Data Analysis which can termed as Identifying patterns in data to obtain valuable insights, which helps in understanding of the underlying trends of the data. The great thing about this course was I had to give final presentation for the course project which demonstrated my communication skills. I was able to grasp every concept taught in the class and applied it while working on my project.

GitHub Link: https://github.com/saipraharsha/Portfolio/tree/main/projects/IST%20718

## 4. CIS 731: Artificial Neural Networks:

**Project Description:**

This course introduced me to deep learning algorithms and the mathematical concepts behind those algorithms which drive them. What this course offers are the cost functions and optimization techniques which are used in implementing the different types of neural networks. This course is extensively on the various deep learning algorithms. To fulfill the course requirements, we need to implement a project to showcase the concepts which are taught in the class.

For successful store & inventory management it is crucial to anticipate the upcoming demand and stock up on the products accordingly. Understocking of products will negatively affect the store profits whereas Overstocking will lead to the store suffering from losses caused by expiration of perishable products. Hence, it is necessary to proactively procure an accurate amount of inventory to maximize revenue and profits. This can be accomplished by accurately estimating the weekly sales for every department of every store.

For our project we used Time Series Dataset contains 3 CSV files ('features.csv', 'stores.csv' & 'train.csv') that have all the predictor attributes and the target attribute ('weekly sales') data across the 3 files. This dataset contains 421K rows of weekly sales for 99 departments and 45 stores of 3 store types across 143 weeks.

Using the dataset described above, firstly, we are planning to perform data wrangling to handle data quality issues such as missing values & duplicate values. This will be followed by exploratory data analysis (correlation analysis, bi-variate plots) to identify the most important attributes required to determine the weekly sales. We will then train a traditional regression model such as Random Forest/Gradient Boosted Trees Regressor to predict the sales. Post this, we will implement the following Neural Network models:

- Feed Forward Neural Networks (FFNN),
- Recurrent Neural Network (RNN) with Backpropagation Through Time (BPTT)
- Long Short-Term Memory (LSTM) Networks
- Gated Recurrent Unit (GRU) networks

Since this is a multivariate time series forecasting problem, we will perform a couple of experiments by using the following 2 ways to vary the number of explanatory input attributes supplied to the model for each of the models listed above:

- Experiment 1: Based on what we observe through EDA & also based on the importance of the explanatory variables, we plan to experiment by dropping a couple of least important explanatory attributes from the training dataset and the training process.
- Experiment 2: We plan on experimenting with a different number of past output and past input values that will be supplied as inputs to the models for training.

As this is a sales forecasting problem with the target being a numeric continuous variable, we will use the Mean Squared Error (MSE) to evaluate the performance of multiple network models. As discussed in the above section, we will be training multiple network models on the dataset described above. These network models will be evaluated using their train MSE and test MSE. Furthermore, the models will be evaluated based on the computational effort put in training the network to obtain desirable MSE. We implemented the project using the python, Keras and Pytorch.

**Reflections & Learning Goals:**

This course has offered a lot, I got to learn how to build a neural network without using any predefined libraries and tweak the architecture to achieve the desired results. Regarding this course, it helped me develop alternative strategies and implement the business decisions by making me see the machine learning from a whole new dimension and when to choose the alternative strategies based the type of data you deal with like when to use supervised learning vs unsupervised learning, which techniques to use for performance evaluation of models and how to select the

feature engineering techniques based on type of the variables in the data. This course taught me limitations of each algorithm.

GitHub Link: https://github.com/saipraharsha/Portfolio/tree/main/projects/CIS%20731

## 5. Conclusion:

During my master's tenure through all my assignments and projects for the respective courses I gained a lot of knowledge of the fundamental concepts of data science along with the advanced topics. I have learned to extract, organize, transform data to find patterns and trends which gives us a deeper understanding of the data. We visualize the data as it provides more insights and can give better view of the data. I learnt to use the insights and patterns found in data to strategize and then implement it. I gained lot of experience in how to communicate the ideas and actionable insights through the presentation of my projects. As a future employee many times the data we deal with might be sensitive and may harm the privacy of an individual if not handled properly. So, all the courses which I undertook made me understand which data needs to be collected and which not. This program helped me gain confidence and to tackle any data driven problem, thus in turn helping the organization by providing them with data driven solution.