**MS Applied Data Science Project Portfolio Milestone**

Fangzhou Yuan

Student ID: 41211045

Spending 2 years here in the Applied Data Science program at Syracuse University, I build a thorough understanding of statistics and gain valuable hands-on experiences through various assignments and projects. So, I expect to include three projects I have done in this portfolio to demonstrate my proficiency in practical data analytics and prove that as a new graduate from the program, I meet the program’s expected outcomes. The three projects are Disease Prediction (with machine learning classification models), Sarcasm Detection (with text mining), and Health Insurance & Vehicle Insurance Cross Sell Prediction (machine learning with PySpark). I will explain my understanding and achievements and use the projects mentioned above as examples and representatives of my learning experience.

I started my journey in data science by learning how to organize and clean data. As the purpose of analysis and data type varies, I am capable of using different methods to prepare the dataset for further visualization and building models. In the Disease Prediction Project, I started data wrangling with detecting missing values and outliers to ensure that our analysis is built on reliable data. Since the goal of this project is to use machine learning algorithms to classify and predict if a patient has a specific type of disease based on the health statistics provided, our target variable needs to be converted to categorical. Similarly, for predicting variables such as gender, it would make our life easier to replace the strings with numeric dummy variables and treat them as categorical. Besides, sometimes collapsing or decomposing variables can be helpful too. For instance, I created a calculated field of the Body Mass Index with the height and weight variables, which is a more informative method of telling a person’s body fat conditions.

After the data is collected and cleaned, the next step is to know how the data points are distributed. Some commonly used methods include descriptive visualizations and summary statistics. Creating plots about the data not only helps us to learn the patterns of the data but also gives us insights insight on model building. I will use the Disease Prediction and Cross Sell Prediction projects collectively as an example since the goals of the projects are similar. Then I will go through the visualizations created in the Sarcasm Detection Project to demonstrate how to identify patterns in text data.

For both of the prediction projects, I created a correlation matrix before visualizing the data to show an overview of how the variables are correlated. If multi-collinearity is detected, i.e. two or more predicting variables have a strong linear correlation, measures need to be taken before building models. Then based on the data type and our focuses, I used histograms, bar plots, density plots, box plots, etc. to compare and contrast how the data points are distributed and how they can potentially influence our predicting results. If the visualizations are properly done with in-depth analysis, we can often find that the results from our trained models are consistent or complementary with our explanatory data analysis.

The importance of data visualization and pattern identification in text mining projects is nonnegligible too. Because textual data is far less intuitive at the first glance, yet we need to understand the dataset without reading every single word. Two methods that I always find helpful are word clouds and frequent word lists. For instance, in the Sarcasm Detection project, I created a word cloud for documents with each label to see how the frequent words differ and overlap in the flat and sarcastic data. Then I found some words such as “trump” are frequently mentioned in both labels. Thus, when I was vectorizing the documents, I paid attention to the max document frequency and made sure that the models do not learn from these frequent terms which can be misleading or distracting.

There are no standard data analytical procedures that work well with every dataset and task. As proficient data science students, we should be able to develop alternative strategies based on the data. For example, in my Health Insurance and Vehicle Insurance Cross Sell Prediction project, there are several alternations that I did to better achieve the project goal. To begin with, the dataset we got has around 400 thousand rows and 12 columns. Among the 400 thousand observations, more than 85% of the data belong to the negative class. In other words, we have a large dataset which could make the training process to be extremely time consuming, and we will introduce bias due to the imbalance of the data if we use the original data directly. To avoid such problems, my groupmates and I decided to implement model building and training with PySpark, which is a cluster computing framework known for its efficiency for performing data analysis at scale. Besides, we experimented with three different re-sampling methods: over-sampling, under-sampling, and SMOTEomek from the “Imbalanced-Learn” package, so that we could prepare a balanced training dataset. Furthermore, we applied principal component analysis on the re-sampled data to bring the predicting variables to a lower dimension and visualize the data in two dimensions to see how the data points are distributed.

After the dataset is well prepared and we have a thorough understanding of the dataset, we are ready to build and train our algorithms on the data. For the two major types of statistical learning and machine learning algorithms, regression or classification, I usually start with logistic regression or linear regression (depending on the target variable), which are easy to implement and run fairly fast. Then I would use the results as a benchmark to compare with other models to briefly tell which models fit or do not fit the project. Speaking of the results, there are multiple parameters that we can use to measure model performance. Again, there is no universal parameter that fits every project; instead, we as analysts need to make the call based on what we would like to prioritize for each project.

I will use my Cross Sell project here as an example to explain how I select, train, and evaluate models. As described above, my groupmates and I started with logistic regression and trained the model with grid search to find the best hyperparameters. Since the goal of the project is a supervised binary classification, we thought of either distance-based classification algorithms (e.g. KNN, KMeans, and SVM) or tree-based models (e.g. Decision Tree, Random Forest, and other Ensemble Learning algorithms). After several trials of building models from each category, we found, in general, tree-based models outperformed distance-based models in terms of accuracy, the area under curve, precision, and recall. This result is consistent with our findings from the EDA and PCA, which pointed out that there are no straightforward differences in the distribution of data points between the two target classes.

During the model training process, our strategy for performance measurement was prioritizing the AUC score, because the AUC score represents the degree of separability very well. In addition, the overall goal of the project is to help the company identify potential customers from their health insurance sales who are interested in a newly developed vehicle insurance service. Therefore, the recall (true positive rate) was another parameter that we paid attention to, since we could tolerant more about sending promotions to customers who are not interested comparing to miss-classifying a potential customer and losing revenue. So maximizing the AUC score brought a positive by-product of having high recall rates, given the fact that the AUC score is calculated as recall divided by false positive rate. Our final decision on model selection was based on the model performance on the validation dataset. In general, random forest was the best choice given the high AUC score and recall together.

Oftentimes, the final deliverables of data science projects are not predicting results or model performance comparison. The fundamental function of data science is to dig out insights from data and improve decision making in business. Therefore, different contents need to be presented as the audience changes. If the audience consists of people with no data science background, e.g. marketing and sales, then actionable plans and possible profits are more persuasive. Meanwhile, if we are presenting to IT professionals, programmers, etc., numbers such as confusion matrix are self-explanatory. Besides, a personal taste of mine is to create a flow chart of data analytical procedures to demonstrate that our analysis is logically reasonable, and more credibility is added to our analysis. Overall, being able to speak the audience’s language is crucial for data science storytellers.

Last but not least, while data science profoundly influences decision making today, ethical concerns such as privacy, implicit bias, and freedom of choices are no laughing matter. From an analyst and developer’s perspective, I am always aware of personally identifiable data, and how such data should be preserved, analyzed, and or encrypted. Because failures may cause chain reactions to damage the customers, organizations, or even the whole society.