Milestone Report

Data science is vital to companies from all industries because it allows the companies to make decisions that can really drive it forward. These decisions come in the form of predictions, finding ways to get ahead of the competition or to improve their product. Companies also use data science for optimization problems, where companies find ways to streamline their products or services in order to work at its most optimal level.

For my capstone, the credit card default data set, the problem is figuring out how to create a model that can predict which customers will default on their next credit card payment. The ramifications of such a model are vital to any company that lends money to its customers.

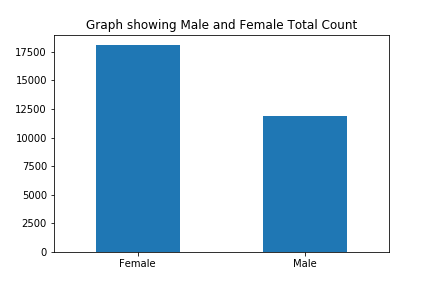
This includes banks, major and minor, credit unions, and even companies that allow their customers to buy items on credit such as car dealerships and furniture stores. New emergents in the credit industry such as many technology companies often give credit to their customers for advertising space and time would also benefit greatly from this model as it is their primary source of revenue.

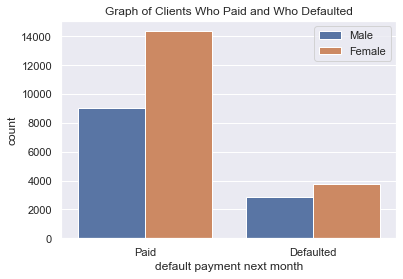
Being able to predict which customers are most likely to default on their next payment can save companies millions of dollars as their model relies on the customers paying on time. If the customer does not pay, then this creates gaps in the what the company expected to be their revenue and their actual revenue. For many companies, existing and emerging, a model such as this could really be the backbone and everything else regarding how the company does business could stem from this model or similar models.

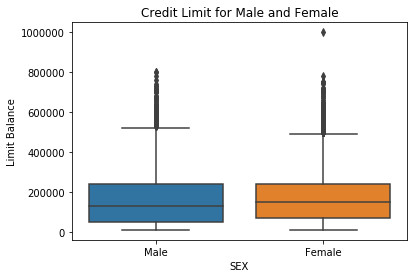
The data set was obtained through the UCI (the University of California at Irvine) repository that was posted to their website. It consisted of 30,000 unique customers and 25 unique data points for each customer. As the dataset was made to be used for students of machine learning, the data was (fortunately) already tidy enough to be used.

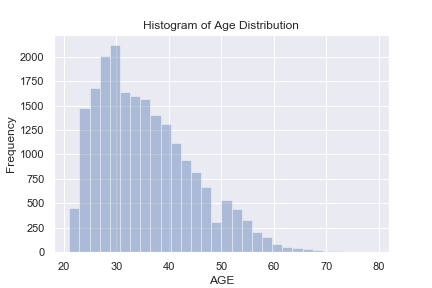
This means that there was very minimal data wrangling that was needed. As I looked through the datasets and implemented certain measures such as dropping null values or seeing if there were any inappropriate values, the majority of it had already been done. I thought that for the purposes of simplicity, naming the columns based on the month rather than having a numerical value would have been easier but that was as far as it needed to go. Each column had the appropriate data filled and there were no null values as observed in this wrangling portion.

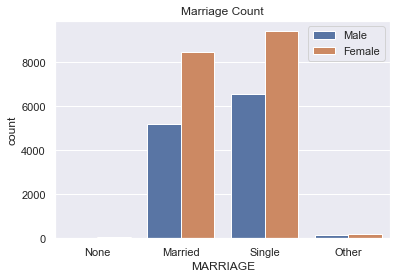
The EDA (exploratory data analysis) portion was very interesting as having this many data points made for great visualizations and analysis. I divided the data into three sets, a training set, a validation set and a test set in a ratio of 70:15:15.



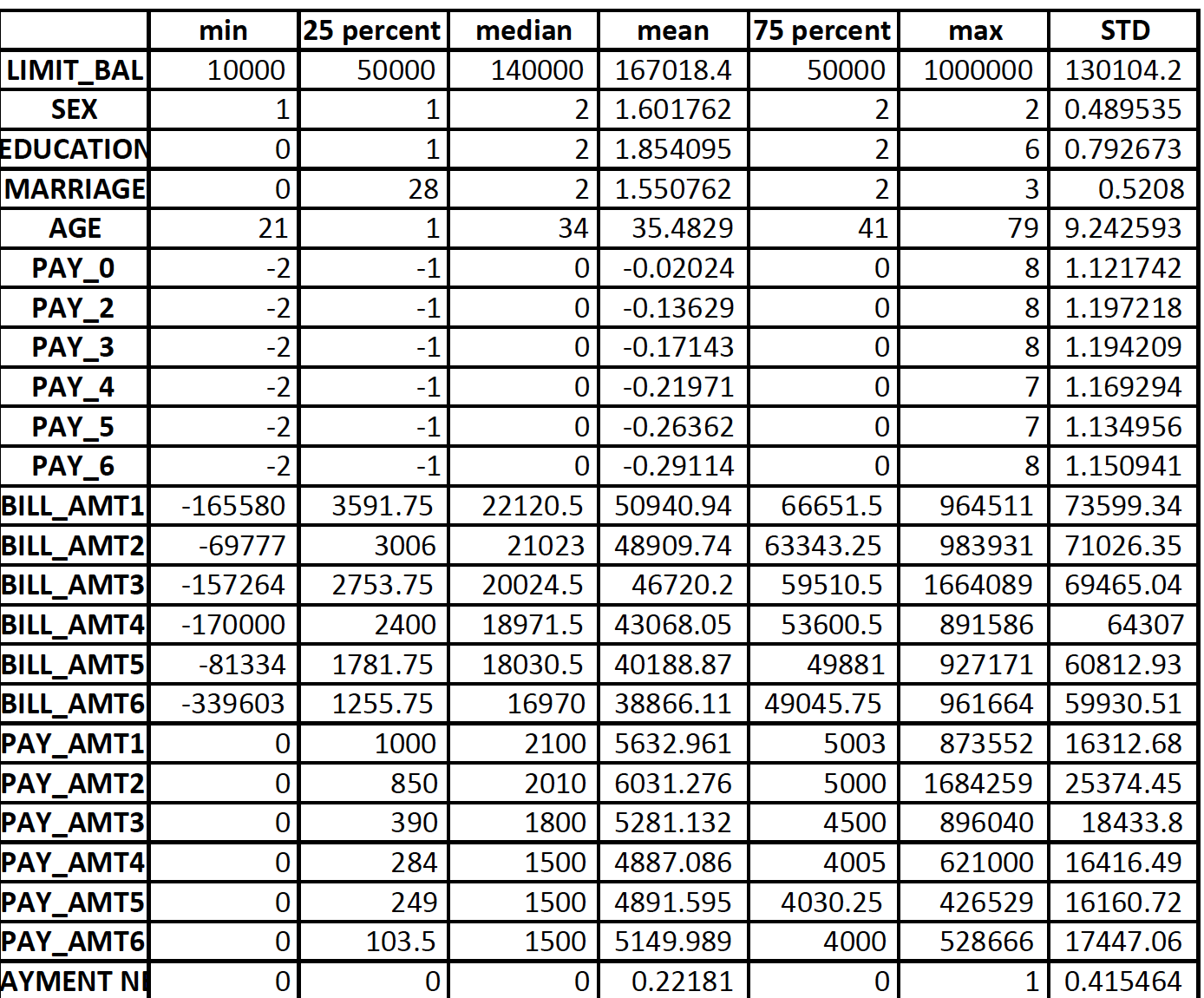


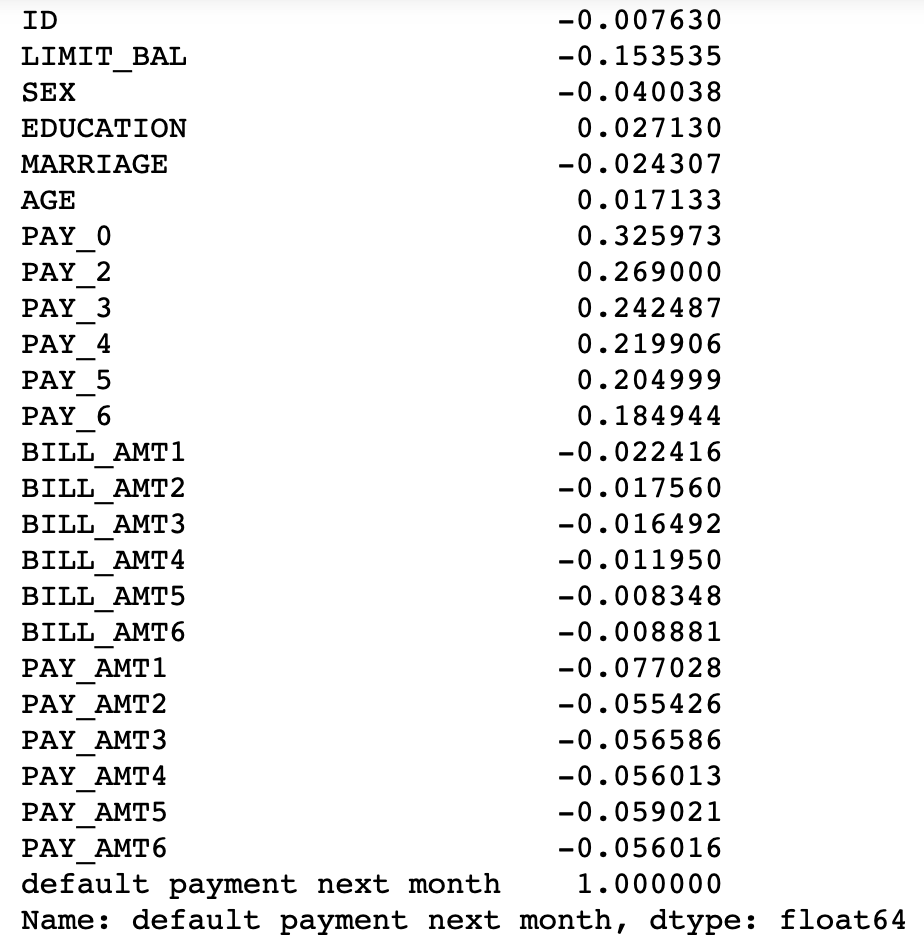






The following information was gathered exclusively on the training data set.





The graphs and statistical analyses showed that there were no significant differences in gender and education. They also showed that the Payment Status, Labelled as PAY\_0, PAY\_2, Pay\_3, and so forth, had the highest correlations to the dependent variable.

When building the model, using sklearn’s LogisticRegression function, the Classification report as well as the confusion matrix are included below.

Another model was built using

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.78 | 1.00 | 0.87 | 3497 |
| 1 | 0.00 | 0.00 | 0.00 | 1003 |
| accuracy |  |  | 0.78 | 4500 |
| Macro avg | 0.39 | 0.50 | 0.44 | 4500 |
| Weighted avg | 0.60 | 0.78 | 0.68 | 4500 |

The confusion maxtrix for y\_test,y\_pred was :

[[3497 0]

[1003 0]]

We can see that based on the recall column and the f1-score, the model performs fairly well. What is left is to use a grid search to optimize the hyperparameters in order to tune the model perfectly.

A model using K-Nearest neighbors was also attempted. The following table is a classification report of this model using n = 3 which was the optimal number.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | Recall | f1-score | support |
| 0 | 0.83 | 0.95 | 0.89 | 2428 |
| 1 | 0.69 | 0.34 | 0.45 | 722 |
| accuracy |  |  | 0.81 | 3150 |
| macro avg | 0.76 | 0.65 | 0.67 | 3150 |
| Weighted avg | 0.80 | 0.81 | 0.79 | 3150 |

The confusion matrix was as shown:

[[2317 111]

[ 478 244]]

A model using SVM was used as well. The following table is a classification report to show the results of this model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | Recall | f1-score | support |
| 0 | 0.83 | 0.95 | 0.89 | 2428 |
| 1 | 0.69 | 0.34 | 0.45 | 722 |
| accuracy |  |  | 0.81 | 3150 |
| macro avg | 0.76 | 0.65 | 0.67 | 3150 |
| Weighted avg | 0.80 | 0.81 | 0.79 | 3150 |

The confusion matrix here was as shown:

[[2316 112]

[ 477 245]]

**In Depth- Analysis**

Exploratory Data analysis is such a vital part of Data Science. Using the right analytical techniques, a person can deduce much of what is happening in the data and really get an idea how to proceed with the project at hand. If used the right way, it can really narrow down the large array of options one has when trying to narrow down the appropriate machine learning technique(s) to use.

In my exploratory data analysis, I found that the data had been normally distributed, and that certain factors, such as the final payment status prior to the predicted payment played a major role. Factors such as sex, and marriage status played less of a role but still held some weight nevertheless.

After this, I began thinking about which machine learning methods to use. I fortunate to be able to find a data set that had the appropriate labels and much of data had been formatted in a ‘tidy’ way.. This meant that I was able to choose among the array of supervised learning techniques. Because the goal of the project was to be able to predict a person as a ‘yes’ or ‘no’, I knew that I would be using a supervised, classification technique. My immediate intuition, because of the dimensionality of the data, was to use a technique such as logistic regression which is known to be able to handle such high-dimensional data. But nevertheless I knew that techniques such as KNN and SVM should never be underestimated.

Using the training data I made during my statistical Explora Data Analysis, I created and trained the three models. For each model, I was able to see the optimal F1 scores and analyze which one would be best to proceed with as my final classification model.

The logistic regression had the highest score (as had been my initial suspicion) and I ended up using the logistic model as my classification. It had a predictive score of about 82% and it had the best numbers in regards to recall and precision.

After much discussion with my mentor, who had also recommended the use of decision Trees and random forests, we ultimately agreed that the logistic model would be able to do the job competently.

Using the logistic model we were able to utilize the features that were given in the dataset to be able to make accurate predictions.

Using many optimization techniques such as cross-validating hyperparameters for the best performance, and encoding certain variables to be better interpreted by the model also helped make the model tighter.

This capstone project has made me realize the incredible importance of model selection as a Data Scientist. It has also made me realize that there is often a lot of leg work to do much before that model selection. I am aware that my data was mostly formatted in the way I needed it to be, so my data wrangling was minimal. But I know that oftentimes in the real world, it is the job of the Data Scientist to wrangle this data and format it in such a way that it is tidy and workable.

I have also realized the importance of having workable hypotheses about the data, and actions such as taking the right steps during the exploratory data analysis are all vital to aid in model selection.

The model, once selected, should also be tweaked and adjusted for the data at hand. I have learned that there is no such thing as a perfect ‘out of the box’ model if you truly want to make the best possible model.

Reference

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.