**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.

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.