# A Non-Technical Report of Machine Learning Models for Predicting Churning

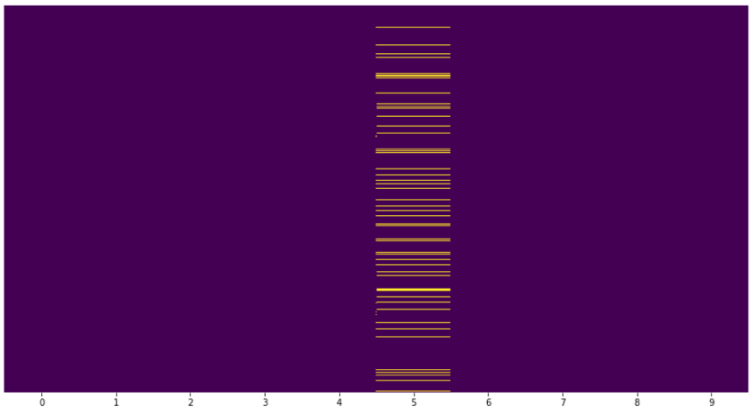
By

Sans Basnet
Dated: May 2, 2021

## DATASET

The dataset provided was based out on advertisement campaign online. It has 10 columns with data on the cost per lead, client's state, duration of business, number of products, calls, change in client's cost per lead, churn, average monthly budget, business category and clicks. Out of all, the change in client's cost per lead has missing entries. The figure on the left visually shows the missing rows.

Missing data are a common occurrence with significant effects on analysis. The missing entries were imputed with mean values using Pandas library. There were two categorical columns i.e. variables on those columns had limited number of possible values. Those columns were transformed before building out the machine learning model.




I performed other exploratory data analysis to better understand the dataset and form a plan. It is important to look at the datatype say if there are numbers or words/sentences and determine to dropped them or transform in a way machine can understand. Once the pre-processing was completed to satisfaction, I moved ahead to model building. I generated a correlation heatmap and visualized the distributions of the continuous variables and others.

## MODEL

The cleaned data is now divided between the testing and training set of 70:30 proportions. We want to know everything about the model so we train the model on the 70% of the dataset and test it on the unseen 30% of the test set to gather model performance metrics. There is nothing special about 70:30 ratio it could be any. I chose two well-known machine learning models known as Decision Tree and Random Forest. These were supervised learning meaning the dataset I used had labels to predict with.

A Decision Tree or DT model is a series of sequential decision nodes designed to reach a particular result. Random Forest or RF algorithms is an ensemble method that combines many Decision Trees trained with different sample of the datasets. Literature reads that Random Forest outperforms Decision Tree replicated in my analysis where the average accuracy was about 82% and 75% respectively.

## ANALYSIS

The models seem to generate reliable outcomes for with anticipated results for the dataset with about 10,000 rows and 10 inputs. It had 1092 imputed entries that were replaced by synthetic data. In addition, class imbalance had some effect in final outcome. The accuracy for RF was 70% and 62% for DT i.e. the fraction of predictions our model got it correct. A 70% accuracy implies 70 correct predictions out of 100 total examples. This is one of many metrics to evaluate our machine learning model. The modeler needs to be aware of over or underfitting the dataset to achieve a stable good prediction performance. DT and RF are generally good at optimizing these trade-offs provided the dataset quality and evaluation was done thoroughly.

## CONCLUSION

In conclusion, the two machine learning models performed a good threshold of roughly 70% accuracy of correct churn prediction. In a real world, this information could be reliably handed down to the client's decision board. The accuracy, precision, recall metrics are used to defend our position and a technical report to justify certain positions. Machine Learning enjoys bigger data and more sorted ones, so one could recommend the client's company to enter data with less biases and correctly.