Supervised Machine Learning (Classification) Final Project

Main Objective

The main objective of the analysis is to focus on the interpretation while still maintaining above-average predictability that will generate a reasonable but accurate prediction that can be explained intuitively to the stakeholders.

Data Set Description

Datasets used in this analysis is a customer churn data from the telecom industry. It is a pre-processed dataset with 22 features and one target (columns) where each of them has 7,043 datasets (rows). As this analysis is to predict the customer churning rate, the targets is apparently the "churn_value" and features included in this analysis are "satisfaction level", "months", "contract terms", etc. The features have 5 floating values and 17 binary values, with the target is also a binary value with two outcomes of churn (0) or not churn (1). The aim is to predict will customer churn in the future according to their historical behaviour in order to prevent the customer from churning at the largest extent.

Data Exploration

In the pre-processing steps, we first separate out the features and the target data. Then, we split both the features and target data into train sets and test sets respectively. The split method we have used in this context is Stratified Shuffle Split.

Model Summary

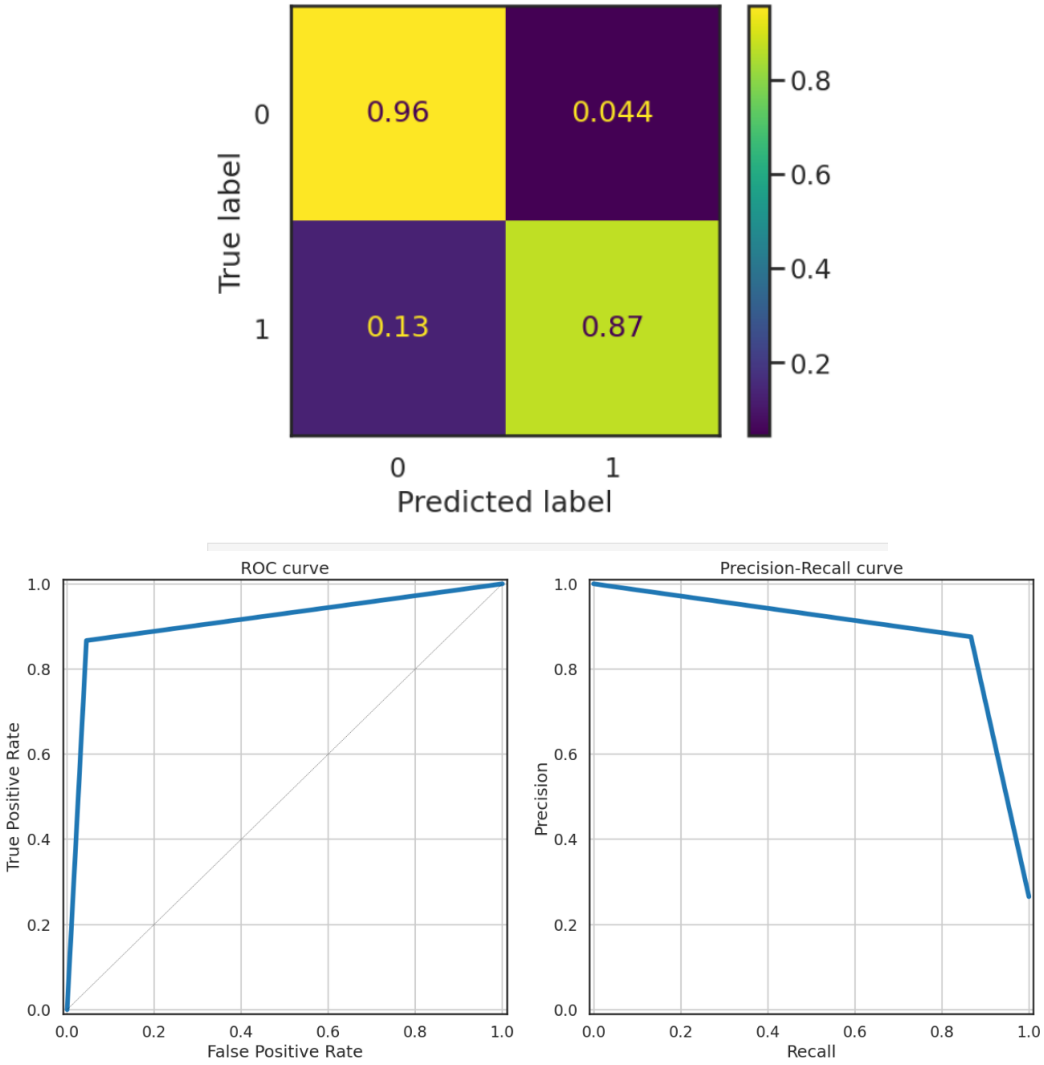
In this analysis, the three models we are using in predicting the churn value are Decision Tree Classifier, Random Forest and Extra Trees Classifier. The main reason to use these three models is to compare the outcome of these three models in answering the question of "Will the prediction be more accurate when more randomness is given to the model." The idea is that, Random Forest will provide more randomness than Decision Tree Classifier and Extra Tree Classifier will provide more randomness than Random Forest, and will more randomness lead to higher predictability?

Final Model

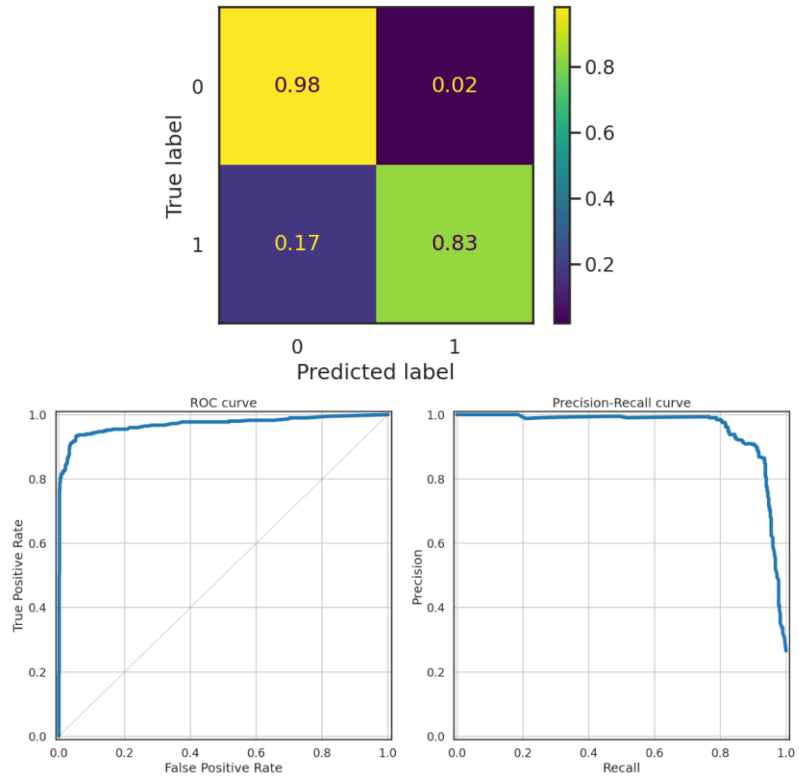
Decision Tree Classifiers is the final model recommended in this analysis that fits the needs in terms of accuracy and predictability. In this analysis, more randomness introduced does not improve much of the model. Hence, with the assurance of reasonable predictability level is matched, Decision Tree Classifier is chosen mainly because of its simplicity.

The main reason to choose Decision Tree as the final model is because Decision Tree is providing information that is accurate enough but at the same time it is the simplest model. I have conducted three matrices for the three models which are Confusion Matrix, ROC Curve and Precision-Recall Curve. Among the results of these three matrices, Decision Tree is having slightly better results in Confusion Matrix, Extra Tree is having a slightly better result in both ROC curve and Precision-Recall Curve. However, it is worth noting that the results of these three matrices do not vary much, which is around the value of 0.9 for both ROC curve and Precision-Recall Curve, as well as a true positive value of nearly 1 and true negative value of roughly 0.85 in Confusion Matrix. Hence, these three models are good models to predict the churning outcome. Therefore, since they are having roughly the same accuracy in predicting the churn value, Decision Tree Classifier is selected as the final model as it is the simplest among three of them. In this case, the interpretation process of the outcome will be much easier. (Attached below are the results of Confusion Matrix, ROC Curve and Precision-Recall Curve of Decision Tree, Random Forest and Extra Tree Classifier)

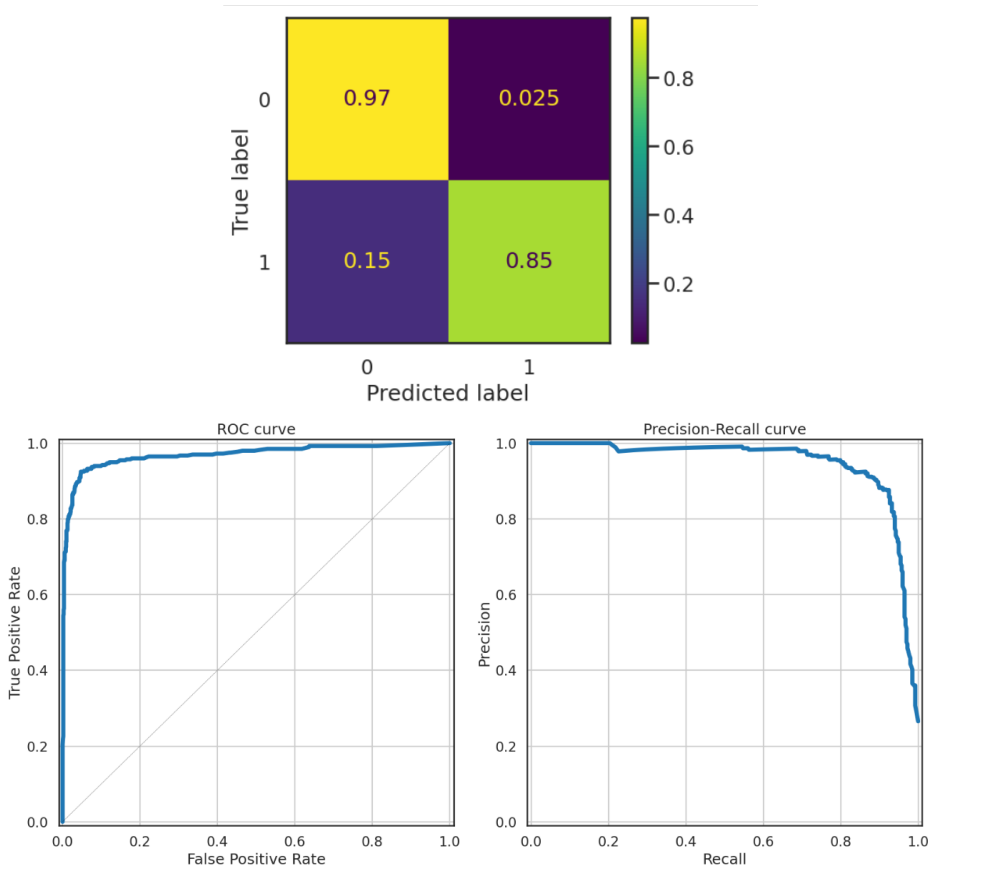
Decision Tree



Random Forest



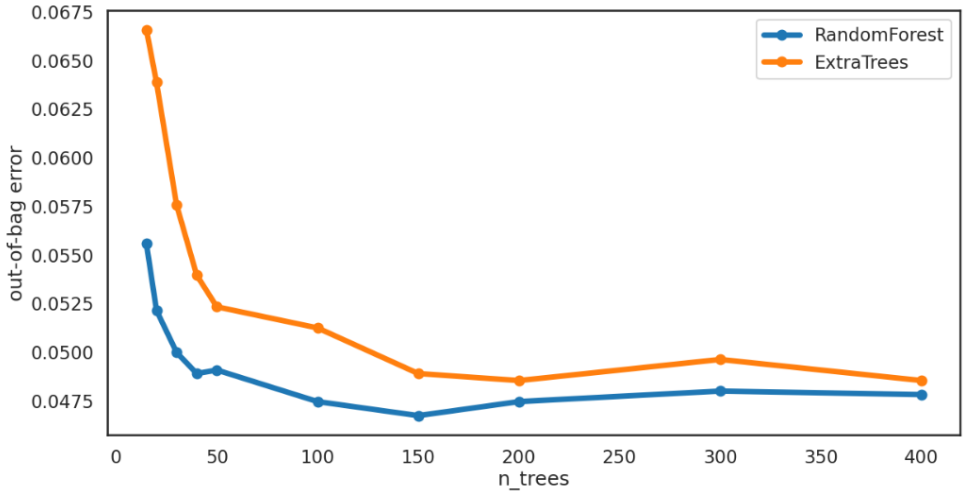
Extra Tree Classifier



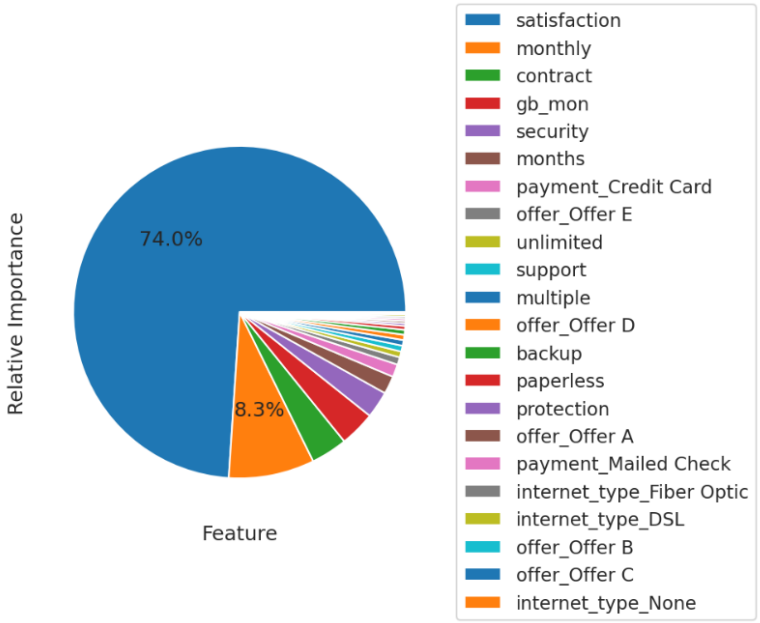
Summary Key Findings and Insights

In summary, this analysis is mainly focus on finding a model that is good in both interpretation and prediction where interpretation is the top priority so that we can explain the result in a convincing and reasonable way to the shareholder. As shown with the result about, three models have an outcome in which their predictability accuracy is roughly the same. Hence, there is a valid reason to choose the simplest Decision Tree Classifier due to its easiness in interpretation.

Also to note, both Random Tree and Extra Tree Classifier in this analysis ran is based on 150 number of trees as it has the lowest value of out-of-bag error. (As shown in the diagram below).



We can also see that, for the selected final model of Decision Tree Classifier, the top three features that account for the high importance for this model are 'satisfaction', 'monthly' and contract'. (As shown in the picture below)



Future Suggestions

In this analysis, it is obvious that one of the limitations will definitely be the number of features that we collected in the data for this analysis. In the future, when more additional features are added into the datasets, this analysis may need to be conducted again to incorporate the new feature in order to obtain a more accurate prediction. However, do also note that with more features added, there is a high possibility that model complexity may increase accordingly, and we might need to be aware of overfitting issues. In this case, we conder some model that is good in countering overfitting issues such as Lasso Regression or Ridge Regression. Additionally, we might also need to consider both upsampling and downsampling method if the collected data is imbalance. Overall, these considerations are important in increasing the predictability as well as interpretation of the model.