Swiss Federal Institute of Technology (ETH)

363-1098-00L Business Analytics

Supervised Learning Project: Telecommunication Customer Churn Prediction

Syed Shahvaiz Ahmed,
Naël M. H. Prélaz,
Clive Charles Javara

Supervised by
Prof. Dr. Stefan Feuerriegel

Abstract

The primary objective of this project was to use supervised machine learning techniques to predict as accurately as possible, the customer attrition rate of a telecommunication firm. The dataset was obtained from Kaggle [1]. A series of exploratory data analysis was conducted to get a holistic idea of the dataset and after crossing a series of preprocessing milestones, the dataset was balanced using the Synthetic Minority Oversampling Technique (SMOTE). A total of six machine learning models were applied and their performance metrics were compared. Principal component analysis (PCA) was carried out to reduce the dimensions of the feature space before going forward with the Logistic Regression model. For all other models, the higher dimensional feature space was used as it led to a better prediction. An average accuracy rate of 82% was achieved across all models when tested against the test dataset while the highest accuracy score was 87% through the Random Forest classifier.

Contents

1 Introduction **3**

2 Data Overview **3**

3 Exploratory Data Analysis **3**
3.1 Demographics . . . 4
3.2 Product Subscriptions . . . 4
3.3 Financial Data . . . 4
3.4 Dataset Interpretation . . . 5

4 Data Preprocessing **5**
4.1 Overall Preprocessing . . . 5
4.2 Tackling Class Imbalance . . . 5

5 Model Building **6**
5.1 Strategy . . . 6
5.2 Training & Test Data split . . . 6
5.3 Models Evaluated in Reduced Dimensional Space . . . 6
5.3.1 Principal Component Analysis . . . 7
5.3.2 Logistic Regression (Baseline Model) . . . 7
5.4 Models Evaluated in Higher Dimensional Space . . . 7

6 Model Comparison **8**
6.1 Overall Comparison . . . 8
6.2 Confusion Matrices Comparison . . . 9
6.3 ROC and AUC Analysis . . . 10

7 Conclusion and Further Discussion **11**

1 Introduction

Companies of all major service sectors compete rigorously among themselves to either maintain or improve their market share. This has been the basic strife since the 60's from DDB Worldwide at Madison Avenue in New York to Boase Massimi Pollitt in London. These frivolous ad agencies were the ones responsible to analyze customer data of their clients and devise marketing strategies to reduce the churn rate or amplify presence. Half a millennium later, with the stark uproar of technological upheaval of machine learning and artificial intelligence, this job has been transferred to consultancies like McKinsey, BCG, Bain & Company etc. Using machine learning techniques, analysts now can easily extract the most critical features of the customer base that the firm's management should focus on when devising strategies to reduce the churn rate.

A similar approach has been carried out through the course of this project where a telecommunication dataset was analysed through supervised machine learning (ML) techniques. The idea was to find the best suitable model which accurately predicts the type of customers that are prone to churn. Such a model can be used to target these customers in isolation, in order to decrease the churn rate. These models are essential in devising focused customer retention programs to increase/maintain the generated revenue, and this methodology can be replicated for other similar datasets. In simplistic terms, this was a classification task which was handled through six ML techniques namely: Logistic Regression with Cross Validation (Baseline model), Support Vector Machines, K-Nearest Neighbours, Decision Trees, Random Forest, and Neural Networks.

2 Data Overview

The telecommunication dataset was sourced primarily from IBM Watson analytics community [2] and second-handed from Kaggle [1]. It consists of 7043 unique rows which represent a customer each and 21 columns which represent various features such as customers who left within the last month, services each customer signed up for, customer demographics, and the amount of money they spent on their telecom package. The data types in the dataset range from numerical, discrete, continuous, and categorical. The initial structure of the dataset was only interpretable enough to understand it intuitively at surface level, therefore significant preprocessing was needed for a fruitful EDA and a model building phase. As little as eleven missing values were identified and subsequently led to the ejection of eleven rows entirely.

The dataset had sixteen unique categorical features measuring varying customer data in the "yes" or "no" category. Furthermore, a categorical value, redundant across multiple categorical features which did not represent atomic and contextual labelling was discovered. Namely, the labelling of customers that had no internet service for products such as "OnlineSecurity" and "TechSupport" for which having an internet service only apply. Therefore, through maintaining the "yes" or "no" setting of the dataset's categorical variables, the normalization intended to group these outliers into the "no" category and drop the third categorical variable entirely. The following code snippet clearly explains:

```
# Replacing 'No internet service' to 'No' for the following columns

align_columns = [ 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport','
    StreamingTV', 'StreamingMovies']
for i in align_columns :
    telco_df[i] = telco_df[i].replace({'No internet service' : 'No'})
```

This initial data cleaning phase resulted in a total of 16 categorical features and 3 numerical features, making the nature of the dataset majorly categorical and therefore orientating the project objective towards a classification task.

3 Exploratory Data Analysis

A customer churn, in other words when a customer ends their relationship with a firm is the primary factor in determining the firm's profitability and market share. As mentioned before, the project's main focus was to predict a clear distinction between customers who churn and customers who don't. But, before any sort of model building could be executed, a sensible EDA phase was done which primarily consisted of probing the relationships and distinctions between Churn and Non-Churn customers in the dataset until a holistic "story" emerged. Figure 1 represents the distribution of Churn to Non-Churn customers. What's interesting to point out is that already with this basic visualization, one obtains a sense of the profitability of this telecommunication firm, even though no temporal metrics accompany this dataset to contextualize Churn over time. What the figure also points towards is the clear imbalance in the Churn and Non-Churn categories, a typical ML issue which will be tackled in the preprocessing section.

3.1 Demographics

Since the dataset is constituted of customers, demographics remained a major focal point. The data suggests that there exists no gender bias. Both male and female are approximately equally likely to Churn, revealing perhaps that the business marketing practice is unisex, thereby capturing a larger market segment.

What is particularly revealing about two specific features in the dataset, namely whether customers have a partner, and whether customers have dependents, is that for the majority, they classify as "no" in both feature settings across both Churn and Non-Churn distributions. One conjecture about these characteristics suggest that the clientele of the company comprises majorly of a young populous.

The last demographic feature worthy of exploration was seniority. Seniors make up less than a quarter of active customers and roughly one quarter of churned customers. This fact appears to confirm the assumptions which were revealed in the previous paragraph about age distribution. It is important to consider that seniors inevitably constitute a portion of customers that Churn perhaps involuntarily due to mortality.

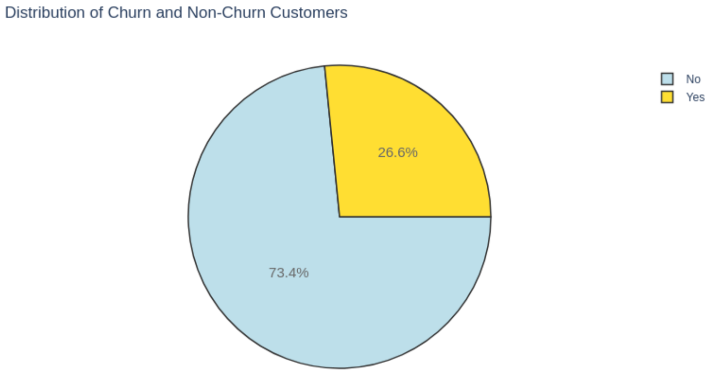


Figure 1: Overall distribution of Churn and Non-Churn customers in the dataset.

3.2 Product Subscriptions

The most popular service customers opted for was "PhoneService", constituting roughly 90% equally across both Churn and Non-Churn customer distributions. Simultaneously, almost 70% of Churn customers that opted for "InternetService" had a fiber optic connectivity medium as opposed to DSL or no internet service. What's insightful in the product feature set is that added-services such as "TechSupport", "OnlineSecurity" and "DeviceProtection" which usually complement a primary service like "InternetService" have on average been largely renounced by customers, ranging between 70% and 85% of the time.

3.3 Financial Data

Customers preferred a "month-to-month" contract. 88.6% of Churn customers held this type of contract, terminating after one month (low tenure) with exposure to high monthly charges. What is revealing about the illustration (Figure 2) is that Churn customers are often new customers, on-boarded with high monthly charges, too soon in the customer life-cycle.

Another insight worthy of mentioning is data on the "PaymentMethod" feature. Amongst the four payment method options offered by the company, Non-Churn customers are distributed approximately evenly at one quarter each. However, the Churn customer segment depicts a higher tendency towards the "Electronic check" payment method. The important remark is that almost 60% of Churn customers opted for "Electronic check", a method requiring manual settlement on behalf of the client instead of an automatic procedure such as "Bank transfer" or "Credit card".

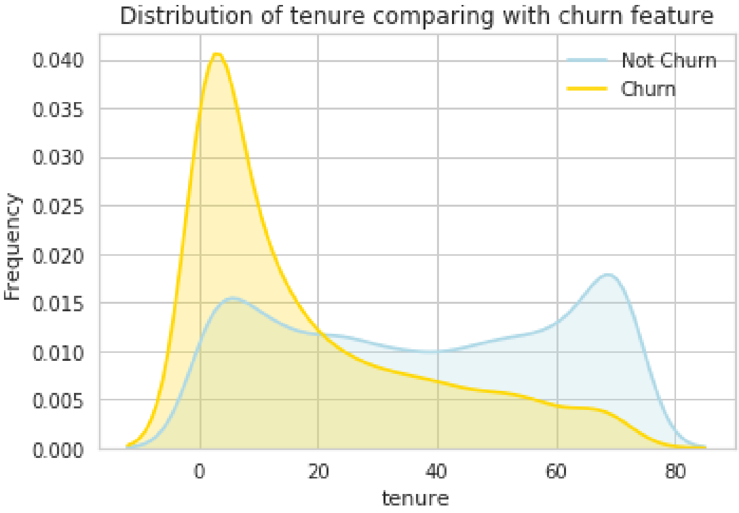


Figure 2: Tenure distribution of Churn and Non-Churn customers in the dataset.

3.4 Dataset Interpretation

An objective summary of the present dataset suggests a young user base, mostly celibate, few with dependents and a gender distribution that is almost identical. Seniors make up a small minority of the population but make up one whole quarter of Churns. Customers are tenacious for phone services and internet services but reluctant to commit to added services that complement the former. With regards to finance and payments, customers have a tendency to fall out of long term financial commitments, especially in the face of high initial monthly charges. Newcomers tend to drop out of contracts after one month.

4 Data Preprocessing

The large volume of categorical features in the dataset required significant preprocessing steps in order to prepare the data for computation. This required transforming feature representations of categorical variables from "yes" or "no" to binary values, 0 or 1. Additionally, a class imbalance problem was identified and tackled accordingly.

4.1 Overall Preprocessing

As the dataset contained a blend of categorical and numerical variables, it was important to treat such categories differently. In order to get a rich feature set and to use the information gain that was realized from the EDA, a series of methods were undertaken to enhance the feature space. Beginning with the "tenure" feature, the customers were categorized into one of five string representations (dummies), namely: 0-12 months, 12-24 months, 24-48 months, 48-60 months, and greater than 60 months. Similarly, variables such as the "PaymentMethod" and "Contract" type of a customer were also converted into dummy representations to extend the feature space. This whole procedure resulted in the culmination of 14 categorical features and 18 numerical features. Categorical features were encoded into binary values and merged with numerical features to obtain a dataframe for model building. Lastly, the numerical features "MonthlyCharges" and "TotalCharges" were normalized using `sklearn.preprocessing.StandardScaler()` [3] that had the effect of demeaning the data to zero and scaling it to unit variance.

4.2 Tackling Class Imbalance

As pointed out at the start of Section 3, an examination of the overall distribution of the dependent variable, "Churn", concluded that a class imbalance problem existed in the dataset. Specifically, the number of Non-Churn customers were predominately greater than the number of Churn customers, with a ratio of 73% to 27% respectively. Upon an initial training of multiple models with the imbalanced dataset, a lack of performance among the models was noticed. The table below summarizes the model performances. Notice the 'Precision' column averages around at only 60%.

Model	Accuracy score	Recall score	Precision score	F1-score	Area under Curve
Logistic Regression CV	0.704	0.864	0.456	0.597	0.756
Decision Trees	0.721	0.414	0.448	0.430	0.620
Random Forest	0.794	0.456	0.630	0.529	0.682
Support Vector Machine	0.796	0.447	0.641	0.527	0.681
K-Nearest Neighbour	0.797	0.480	0.632	0.546	0.693
Neural Network	0.799	0.499	0.632	0.558	0.699

Table 1: Model Performance Evaluation with Imbalanced data

In order to overcome this class imbalance problem, a statistical sampling technique called 'SMOTE' was used. It takes the entire dataset as input and increases the ratio of the minority subset in the selected dependent feature class by creating new instances from the existing samples. Working with a balanced dataset had positively impacted our models in terms of performance and predicting churning customers (Section 6). The visualization below (Figure 3) demonstrates how imbalanced the dataset was before and after applying the oversampling technique.

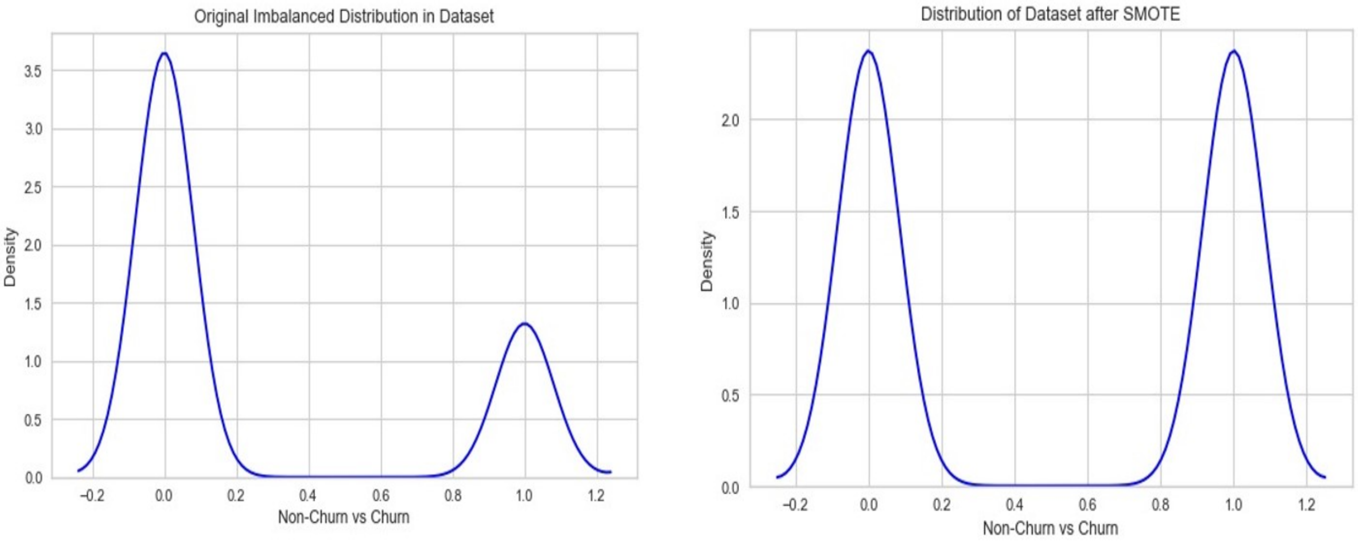


Figure 3: Balancing Churn and Non-Churn distribution using SMOTE. Left: Original Distribution of the dataset. Right: Balanced distribution of the dataset using SMOTE oversampling technique.

5 Model Building

This section explains the strategy involved in building and evaluating the supervised learning models to classify a customer as either Churn or Non-Churn. In total, six different classification algorithms were applied and compared for this task, namely: Logistic Regression with CV, K-Nearest Neighbors, Support Vector Machines, Decision Trees, Random Forest, and lastly a sequential Neural Network.

5.1 Strategy

Machine Learning (ML) models are optimized through tuning their hyperparamters. What was found in our two month long exploration with the aforementioned models is that some models worked better in a higher dimensional feature space (the original feature space), while some models worked better after a reduction of higher dimensional feature space (after PCA). Thus, we specified two separate sections: 1) Models evaluated in Reduced Dimensional space 2) Models evaluated in Higher Dimensional space. The evaluation parameters of each model was the overall Accuracy, Area under the Receiver Operating Characteristic curve, Precision score, Recall score, and the F1 score of class 1 (Churn).

5.2 Training & Test Data split

Like any traditional ML model, it was a necessary requirement that prediction models do not only recall samples given during the training procedure but at the same time classify unseen instances. To tackle this issue, the models were trained only on a subset of the total dataset. This method is well known as the train and test split, where we split the dataset into a training and a testing class. Once the models have been trained using the training set, the accuracy of the models can be measured on the data we have put aside for testing. In order to preserve homogeneity across all the classification models used for this task, the dataset was split into a conventional 75% training and 25% testing split.

5.3 Models Evaluated in Reduced Dimensional Space

In any ML classification task, all relevant models are compared with a baseline model. This baseline model is supposed to be the most simplified version of an ML algorithm, but industry conventions dictate us to optimize the baseline model to its maximum and then move forward with a definitive model comparison. In our case, the baseline model was Logistic Regression. There were two ways through which a decent optimization was achieved: firstly through incorporating cross validation and secondly through reducing the original feature space through principal component analysis (PCA).

5.3.1 Principal Component Analysis

In order to avoid redundant information (avoiding noise modelling) of features, we generate a new set of variables (Principal Components). Each Principal Component is a linear combination of the original features. We know that Principal Components explain a part (or full) of the variance. Typically, we want the explained variance to be between 95–99%. After setting n_components equal to two, we observed that the PCA plot showed two separate clusters of samples and almost clearly distinguish between Churn and Non-Churn customers (Figure 4). Also, the explained variance of the two components was ≈100% so there was no need to add another component.

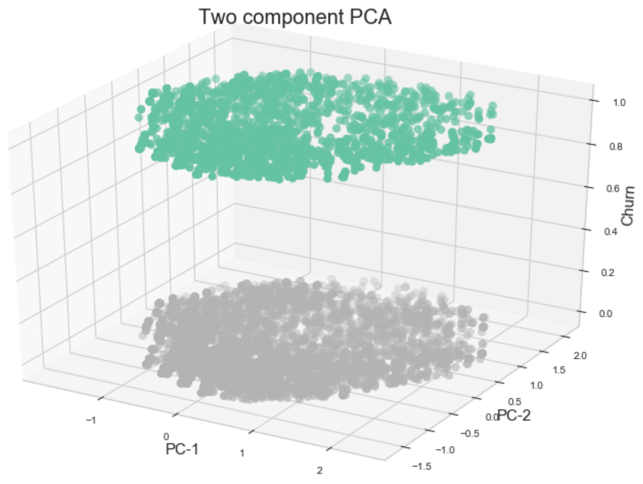


Figure 4: Reduction of feature space using PCA showing clear distinction between the two classes.

5.3.2 Logistic Regression (Baseline Model)

In the setup of the baseline Logistic Regression, Grid Search with cross validation was implemented to find the optimal hyper-parameters. The base of any logit model relies under a group wise Binomial distribution on which any goodness of fit statistics like the deviance statistics or the Pearson chi-square is valid. Because of this validation on a group-wise distribution, we found out that feature reduction actually helped in getting a better fit on the dataset. In order to further enhance the predictive power of the baseline model, we used the MinMax scaler rather than the StandardScaler in the train and test split. Experience with ML models teaches us that MinMax Scaler works better for Logistic models. In general, the speed of updating each feature depends on its scale, and in the end some coefficients are closer to the optimum than the others. Scaling softens this, because coefficients are now at the same scale and update roughly at the same speed. Through such parameterization, an accuracy rate of 78% was achieved as compared to 70% (Table 1) without any such method.

5.4 Models Evaluated in Higher Dimensional Space

In this category of model building, we used the original feature space (which essentially was a higher dimensional space). A set of different techniques were tried to further extend this feature set of thirty by the use of monomials, and as well as one-hot encoding of all features, but both of these techniques resulted in overfitting of the models. Thus, no quest for accuracy was initiated primarily to avoid over-fitting. The following binary classifiers namely: K-Nearest Neighbors, Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF) and a straightforward Neural Network were evaluated. In other words, each of these models were trained without applying any dimensionality reduction technique like the one done for Logistic Regression in the section above. The primary reason for segregating this class of models separate was due to the loss of information during PCA. While there are other techniques such as the Linear Discriminant Analysis (LDA) and Local Linear Embedding (LLE) for dimensionality reduction, the course of evaluation and model building was restricted to PCA in this project.

All models went through a Grid Search with cross validation to find the best parameters for training. Once identified, each respective model was trained with those parameters and the model evaluation indicators were recorded. The models that are evaluated under this pretext are conventional ML models which are often used for binary classification tasks. The Decision Tree Classifier has a tree shaped structure which is used to discover rules in a classification task. The aim was to predict the target variable by selecting features with the highest information gain at each node of the tree. On the other hand, SVM tries to efficiently discriminate between Churn and Non-Churn customers by constructing hyper-planes while simultaneously maximizing the space between the hyper-planes. Using a kernel function the SVM maps the space where the data points can't be linearly separated into a space where they will separated. We exclusively report on SVM with 'RBF' kernel and a penalty parameter of 10^3.

An extenion of DT is the Random Forest classifier which is an ensemble learning classification model. It randomly selects feature and creates a cluster of Decision Trees based on the training data provided. This helps us solve any over fitting problem which could be caused by a sole Decision Tree. The K-Nearest Neighbor model however, is used to classify the sample depending on the K-most similar or nearest instance in the training data. Specifically, we compute the euclidean distance between a single data point against all points in the training set. In this classification task, we used KNN with a k value of one, to achieve the the most optimal results. And lastly, a Neural Net with a depth of two layers with an adam optimizer, sigmoid activation, and binary cross entropy loss was also used for this classification task. The next section deals with the performance of each model in detail.

6 Model Comparison

As mentioned in the previous section, evaluation parameters of each model were the Overall accuracy, Recall score, Precision score, F1-score, and Area under the Receiver Operating Characteristic curve (AUROC). The positive predictive value of a model is the ratio between actual churners and total number of cases classified as churners (the sum of True and False positives), this is also called Precision. Recall is the ratio between True Positives and actual number of Positives, i.e. True Positives plus False Negatives. The F1-score gives the skill of the model for a specific probability threshold and is the harmonic mean of Precision and Recall. The AUROC shows the performance of the classification model at all thresholds. It plots the True Positive Rate, i.e. Recall, against the False Positive Rate.

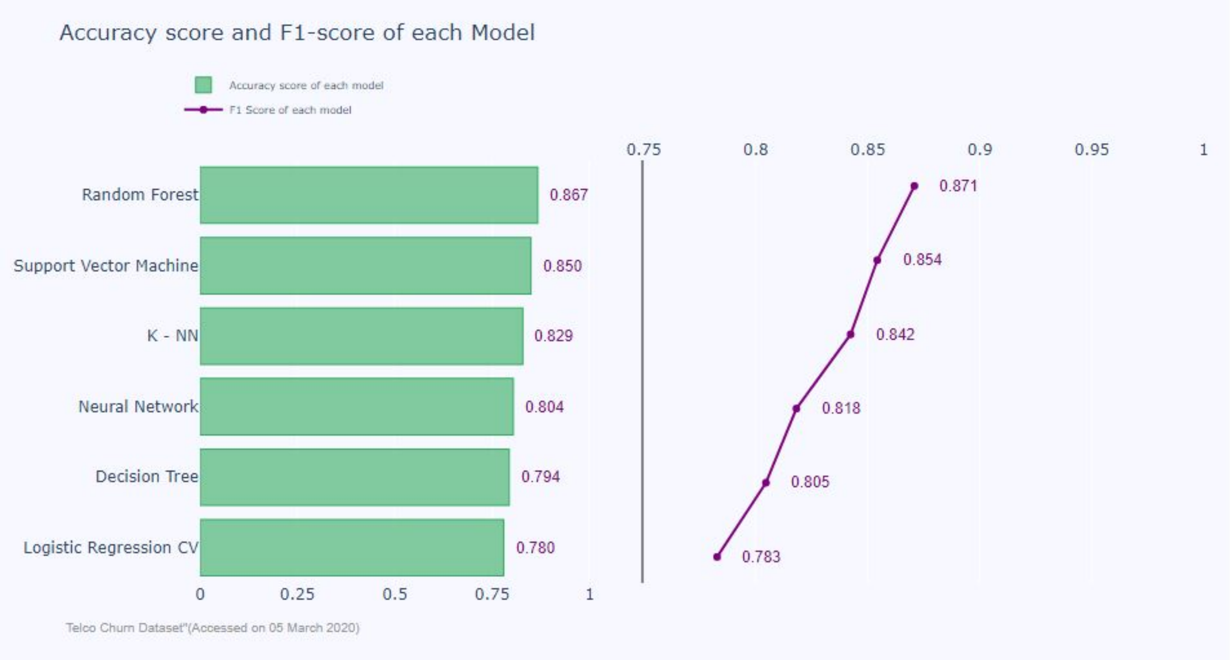
6.1 Overall Comparison

Table 2 below reports the combined model metrics for all the above mentioned models when classifying Churn and Non-Churn customers. Inspecting the figures from the table below we can see that all models perform above the baseline model. Moreover, we notice that the Random Forest classifier outperformed all other models in terms of Accuracy (87%), Precision (85%), F1-score (87%) and AUC (87%). The inferior performance of the Logistic Regression (baseline model) can be elaborated by the mere fact that it tries to fit a linear decision boundary between the binary classes which clearly aren't linearly separable. On the other hand, with a more flexible decision boundary we gain significantly higher classification accuracy in all other models. The results where accomplished after using the oversampling technique SMOTE which greatly improved the performance of each model.

Model	Accuracy score	Recall score	Precision score	F1-score	Area under Curve
Logistic Regression CV	0.780	0.785	0.780	0.783	0.779
Decision Trees	0.794	0.839	0.773	0.805	0.794
Neural Network	0.804	0.870	0.767	0.818	0.799
K-Nearest Neighbour	0.829	0.905	0.788	0.842	0.828
Support Vector Machine	0.850	0.871	0.838	0.854	0.849
Random Forest	0.867	0.897	0.851	0.871	0.868

Table 2: Model Performance Evaluation

Figure 5 below provides an even clearer visualization comparing the Accuracy scores and F1-scores for each model. The line graph shows an overall F1-score which represents the correct classification of True Positives and False Positives (further elaborated in the next subsection) while the green bars on the left represent the respective accuracy score.



Telco Churn Dataset"(Accessed on 05 March 2020)

Figure 5: Accuracy score and F1-score of each model.

6.2 Confusion Matrices Comparison

The obtained results were further visualised using confusion matrices, which measure the performance for a given classification algorithm. The confusion matrices are illustrated as an array with 4 different combination of Churn and Non-Churn. The numeric values represent the actual and predicted classifications resolved by the model. The context of an individual confusion matrix is explained below:

[Actual (Non-Churn), Predicted (Non-Churn)] - # of instances correctly classified as Non-Churn. (TN)
[Actual (Non-Churn), Predicted (Churn)] - # of instances predicted as Churn but Non-Churn in real. (FP)
[Actual (Churn), Predicted (Non-Churn)] - # of instances predicted Non-Churn but Churn in real. (FN)
[Actual(Churn) , Predicted(Churn)] - # of instances correctly classified as Churn. (TP)

The confusion matrices in Figure 6 below provides an overview of the TP, FP, TN, FN for all the models evaluated. For instance, the accuracy rate for a single classifier can be deliberated from the proportions of instances correctly classified as Churn and Non-Churn by the amount of samples which where correctly and wrongly predicted as Churn and Non-Churn. Furthermore, since we are mainly interested in churning customers (customers who leave the firm) the Precision score is calculated by taking the TP ([Churn , Churn]) and dividing by the sum of TP and FP. One interesting thing that can be seen from the Confusion Matrices below is that the K-NN model predicts the customers who will Churn, better than the Random Forest model, but falls in overall accuracy due to the severe misclassification of the Non-Churn customers.

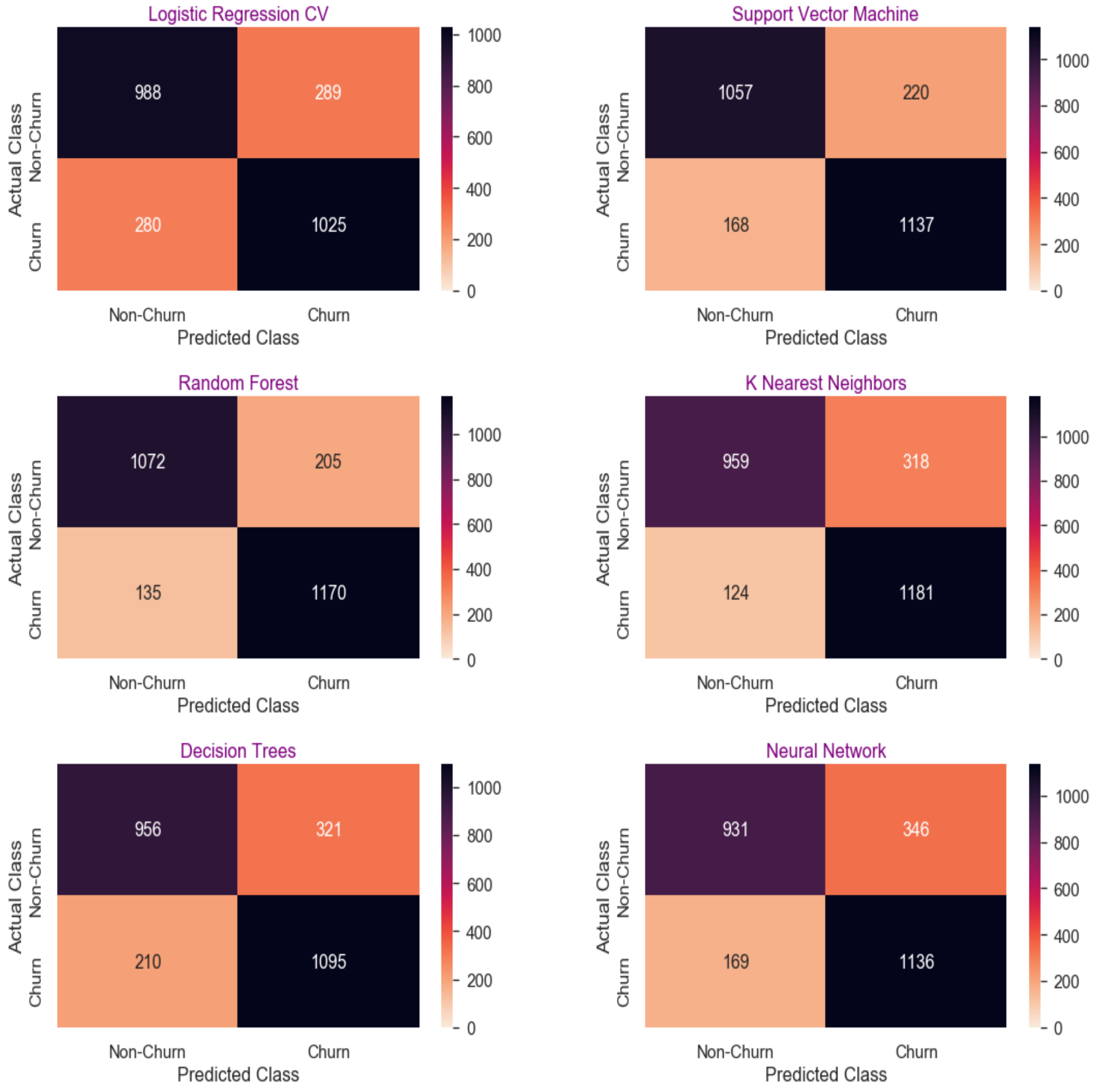


Figure 6: Confusion Matrices for all Models.

6.3 ROC and AUC Analysis

A graphical representation of the Receiver Operating Characteristics (ROC) curves of all evaluated models can be seen from Figure 7 below. The ROC is plotted as a two dimensional graph. The x-axis represents the False Positive Rate (FPR) that is the proportion of samples who did not Churn but where classified as Churn. Similarly, the y-axis takes the True Positive Rate (TPR) that is the proportion of samples who churned and where actually classified as Churn. The performance of a certain classifier can be evaluated by examining how close the curve can stretch out to the coordinate region at (1,0) conveying a False Positive Rate of zero and a True Positive Rate of one indicating a perfect classifier. Additionally, the ROC analysis delivers us an additional measure when drawn over a series of threshold points that is the Area Under the Curve (AUC). This measure tells us how efficient our model can discern True classes from False instances.

In the figure below, it can be clearly seen how well the the Random Forest classifier performs with an AUC value of ≈ 87%. Moreover, we can note a clear performance improvement when comparing the baseline classifier (Logistic Regression) with any other supervised learning model. In essence, the top three classifiers that have the highest AUC across all compared models are: K-NN (with 1 neighbour), SVM (with rbf kernel), and Random Forest (with max depth 15).

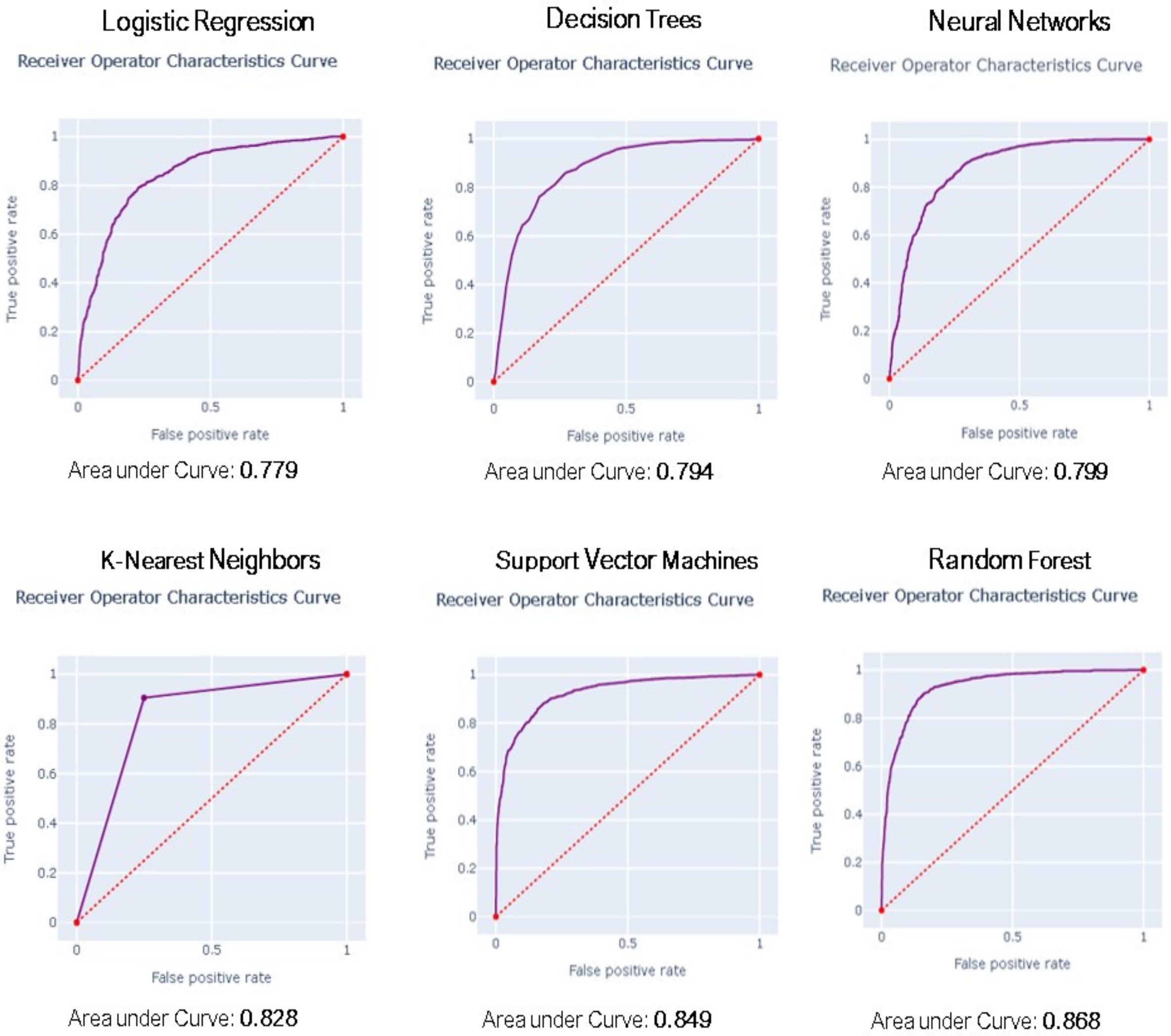


Figure 7: ROC Curves for all Models.

7 Conclusion and Further Discussion

Churning customers are always going to be an inescapable factor in the telecommunication industry. However, it is in the sole hands of the company to reduce the attrition rate through customer retention strategies. In this project, six supervised machine learning models were evaluated for correctly predicting the churned customers. The original dataset was split into 75% training and 25% as testing class. In order to find the best hyper-parameters across all models, Grid Search technique along with cross validation was used. The dataset was heavily imbalanced towards the Non-Churn customers which initially resulted in poor performance across all models. In order to overcome such a barrier, an oversampling technique SMOTE was used to induce balance in the binary classes. The result of such a process was a significantly improved accuracy score across all classification models. Through our obtained results it could be seen that the Random Forest classifier outmatched all other tested machine learning models with an accuracy score of $\approx 87\%$.

While the accuracy score on this particular classification task was at a decent percentage, there are other techniques that can be evaluated to further enhance this accuracy level. For example, there are multiple techniques for dimensionality reduction like Linear Discriminant Analysis (LDA) and Local Linear Embedding (LLE) which are non-linear transformation techniques that can be used instead of the conventional PCA, to check if it is a better fit for the dataset. More so, optimization techniques such as XGBoost for the Random Forest classifier can also be used to further enhance the achieved accuracy. Similarly, a deep neural network can also be constructed to exploit the feature space and learn more from it. However, in any case of further improvement of the current accuracy score, it shouldn't be forgotten that an exclusive quest of accuracy often leads to noise modelling and potential overfitting. Keeping such limitations in our mind, we as a group felt that their was no further need of improvement in accuracy.

References

[1] Telco Customer Churn, Kaggle, February 2018. https://www.kaggle.com/blastchar/telco-customer-churn, last visit May 29, 2020.

[2] Archived | Build a customer churn predictor using Watson Studio and Jupyter Notebooks, IBM Watson Studio, Heba El-Shimy, Scott Dangelo, November 2018. https://developer.ibm.com/technologies/data-science/patterns/predict-customer-churn-using-watson-studio-and-jupyter-notebooks/#, last visit May 29, 2020.

[3] StandardScaler, scikit-learn 0.23.1, February 2018. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html, last visit June 02, 2020.