# Using machine learning methods to predict Customer Churn in Telecom

Candidate Number: 011039, Student ID: 700057465

## 1 Abstract

We plan to use machine learning techniques in this project as a medium to forecast the customer churn in the telecommunication industry. Based on the historical data, we will predict that the customer will stay or change the telecom service provider. We would use random forests to try to classify the loyalty of the customer to the company. We were able to forecast when a customer plans to leave the organisation with reasonable accuracy.

## 2 Introduction

### 2.1 Contextualisation: What is Churn?

One of the main concerns of telecommunications companies is customer retention[1]. Managing customer churn is of great concern to global telecommunications service companies[5]. The term churn refers to the subscribers or customers changing the service provider, triggered by better rates or services or by the benefits offered at signup by a competitor company [2]. The annual churn rate ranges from 20% to 40% in most global mobile telecommunications service companies. Figure 1 shows the data from Strategy Analytics in 2018 about the world's churn rates.

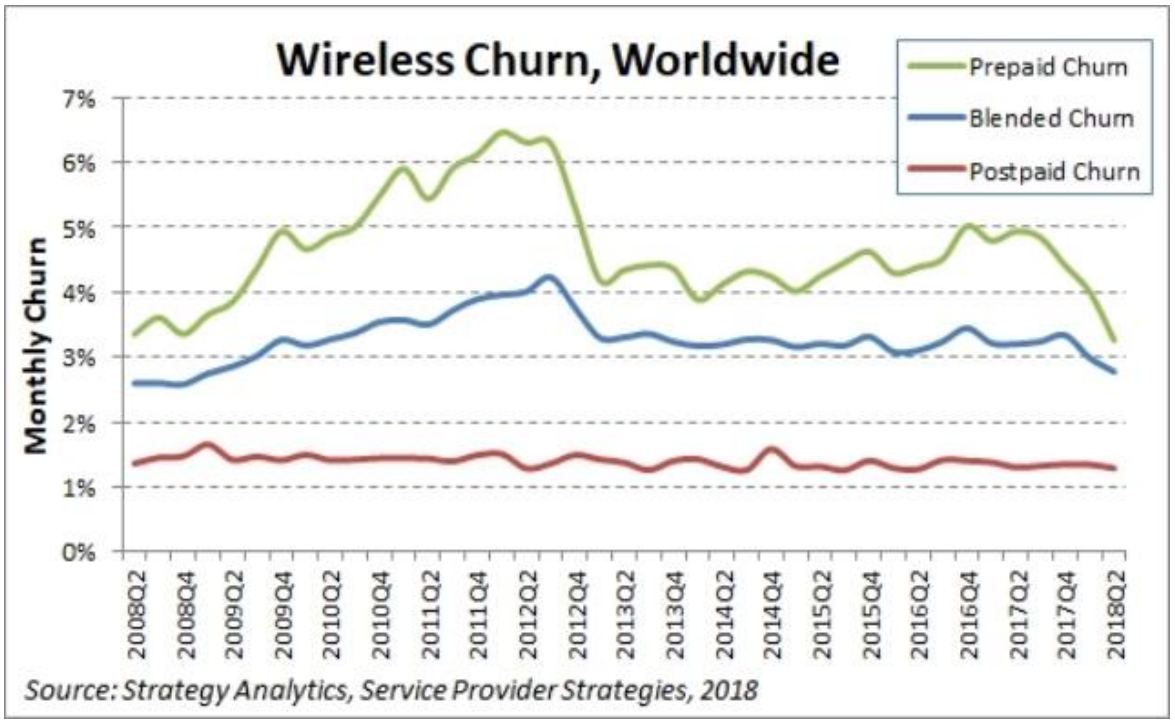



*figure 1: Worldwide monthly churn rates across different telecom operators*

### 2.2 Motivation

Due to the rapid growth in the data communication network and advancement in Information Technology, a massive amount of data is available in the telecom sector. To predict customer churn, many companies make use of data mining techniques, and it can be described as a method which helps in identifying possible churners in advance [3]. It is measured by the rate of Churn and is an essential indicator for organisations.

The introduction of the predictive model has accelerated the retention process, and the mobile telecommunications companies are achieving positive results in this competitive market. This prediction process depends strongly on the data mining techniques mainly because of the increased performance obtained by the machine learning algorithms.

### 2.3 Project Aims

The objectives of this work is to use machine learning techniques to determine whether or not the consumer is going to churn. We would use the historical data of the telephone subscriber for research. This date is classified data, and they give us information about the various situations in which the consumers churn.

We will use data learning approaches; specifically, we will use decision trees and random forests for our analysis. The aim is to see if a random forest can classify the customer will leave the existing service provider. If this is successful, we will attempt to predict which customers will churn.

## 3 Literature Review

Customer churn prediction can be regarded as a classification problem, in which each customer is classified into one of two classes, churn or non-churn. To our knowledge, little research has been conducted on using propagation models to predict customer churn. Popular churn prediction methods include logistic regression [33], decision trees [18], neural networks [41], support vector machines [8], and evolutionary algorithms [3]. Most of the studies are useful, but we will implement the random forest and compare the result with the decision tree to understand the performance with other studies.

Now there have been studies to identify the classification based on the interpersonal features of the customer. However, according to the many formal studies classification based on the service usage pattern and billing attributes, favourably contribute to the churn classification.

Service usage patterns can be described as three commonly used measures, minutes of use, frequency of use[5] and the total number of different receivers contacted by the subscriber (Wei & Chiu, 2002). In effect, the monthly charge level of service usage is one of the most prevalent behavioural predictors of defection in previous research (Buckinx & Poel, 2005). Mozer et al. (2000) conjectured that monthly charges and usage amounts are linked to churn[5]. Also, it is said that actual customer transactions or billing data may fully represent the customer's actual future decisions better than survey data or personal data[5]. So we will be using the billing information such as last payment date, recharge capabilities, loan payment ability of the customers, and our experiments.

There is evidence that access to these new reanalysis data sets allows for a great improvement in churn analysis prediction. We want to take advantage of this by applying machine learning methods to exploit the large amounts of data and computational ability now available to us.

# 4 Method

Due to Telecom business's nature, service providers have extensive customer and billing data; building the cost and computationally in-expensive dynamic models using the machine learning approaches seems to be appropriate for churn analysis and prediction. We will discuss the random forest approach and decision tree approach to predict the customer churn for the chosen dataset.

## 4.1 Data

The data we will be using is from the Kaggle data repository for the with 100000 data points. The data contains information on the customer information like mobile number, billing and payment details for different time intervals like 30days, 60day, 90 days. The data, which make up the 33 variables of the data set, have only used for the processing. All the other string variables and does not represent the importance in the processing have been dropped.

## 4.2 PCA

The PSO undersampled dataset is further treated with PCA in order to reduce the high dimensionality. PCA transforms the dataset into artificial components, which cover the maximum variance present in the dataset. In the first step, the respective mean is subtracted from each of the data dimensions to produce a zero-mean dataset. Then covariance is computed as expressed in the following equation:

$$Cov(X1,X2)=\sum(X1-M1)(X2-M2) / n$$

Where X1 and X2 represent instances of features under consideration, M1 and M2 are the respective means and n is the total number of instances.

Covariance is measured between two dimensions. Therefore, if there are more than two dimensions, then it can be represented in the form of the covariance matrix as given in the following equation:(5)

$$C=v(X1)c(X1,X2)...c(X1,Xp)c(X1,X2)v(X2)...c(X2,Xp).....c(X1,Xp)c(X2,Xp)...v(Xp)$$

Since the computed covariance matrix C is square, the eigenvectors are computed, which show the direction of maximum variance along a certain dimension. Eigenvectors are ordered using eigenvalues in descending order, which identifies the insignificant components present in the dataset. The reduced feature vector is represented with significant selective components as given in the following equation:(6)

$$FeatureVector=(eig1,eig2,eig3,...,eign)$$

Finally, this feature vector is transformed into a new reduced feature vector by taking the transpose and multiplying it on the left of the original transposed dataset as given in the following equation:

$$NewDataset=[FeatureVector]T[Data]T$$

PCA reduces the feature space to most coherent features covering maximum variance of the undersampled dataset.

## 4.3 Random Forests

Random forests consist of a large number of trees that work together to form an ensemble. [6] Ensemble methods combine different classifiers to provide better generalisation. The trees work as a committee, each gives a prediction for the class label , and the most voted for label becomes the model prediction.[6] To perform well the individual trees must have low correlations with each other[6], that way even if some trees are wrong, many trees will be correct and so protect the overall model from the individual errors of some trees. We ensure the trees in our random forest have low correlations with each other in two ways:

### 4.3.1 Bagging

Bagging enables each tree to pick and substitute random samples from the training set to be trained on.[7] As we have already mentioned, decision trees are very responsive to the training data and can be easily overfitted, so training the tree on different training sets leads to uncorrelated trees and decreases the propensity to overfit the model. Also, each tree can still be trained on the same amount of data as the original training set since we sample with replacement, meaning you dont need huge quantities of more data to generate a random forest than a single tree.

### 4.3.2 Feature Randomness

We selected our nodes from all the possible characteristics when we worked with a single tree, while each tree in our forest can only choose from a random subset of these characteristics.[6] Again, this leads to more variance and lower similarity in the trees that are generated.Together with bagging, this ensures that trees are trained on various datasets with different features that contribute to a forest of trees that have a low correlation to each other,helping them to make better decisions than a single tree or several trees with similar setups.

## 4.4 Decision Tree

Decision trees can be used in both classification and regression problems and have the advantage of being easy to understand and interpret. As our data set has several missing values, they are also resilient to noise and missing data, which is useful in our situation. As we will forecast the binary variable 'Fraud' and when fraud is over, we will concentrate on classification tree.
The trees are built up of nodes and branches. The root node at the top of the tree contains an attribute A0 to be tested. We can write the possible values of the attribute as v ∈ values(A0) then there are |v| branches off of A0 each representing one of the values. At the end of each branch there is a leaf node which classifies the data or another attribute A1, A2, .... Moving down the tree testing the attribute and then following the relevant branch will eventually lead you to the estimated classification of your data.
. Our dataset has continuous variables, so it is important to understand how these enter the tree.
The first step in building a decision tree is choosing the root node. A common metric used to do this in the case of a classification problem is the information gain which uses the principle of entropy or disorder, we wish to split the data in a way that reduces the disorder the most (i.e., splits the data into subsets with high proportions of a single class). We define entropy as,

$$\text{Entropy(S)} = -\sum_{i=1}^{c} p_i log_2(p_i)$$

Where S is our collection of training examples and pi is the fraction of S with class i. The

information gain is then,

$$\text{Gain(S, A)} = \text{Entropy(S)} - \sum_{v \in values(A)} \frac{|Sv|}{|S|} \text{Entropy}(S_V)$$

# 5 Discussion about experiments

## 5.1 The Data

The data we were working with provided 36 attributes for customer churn. Usually, customer churn is associated with customers able to stay with the company. Suppose the customer is not able to repay his pending bills or pay back his loan. Label attribute in the data is a binary data, '1' means the customer will churn, and '0' means the customers will not churn.

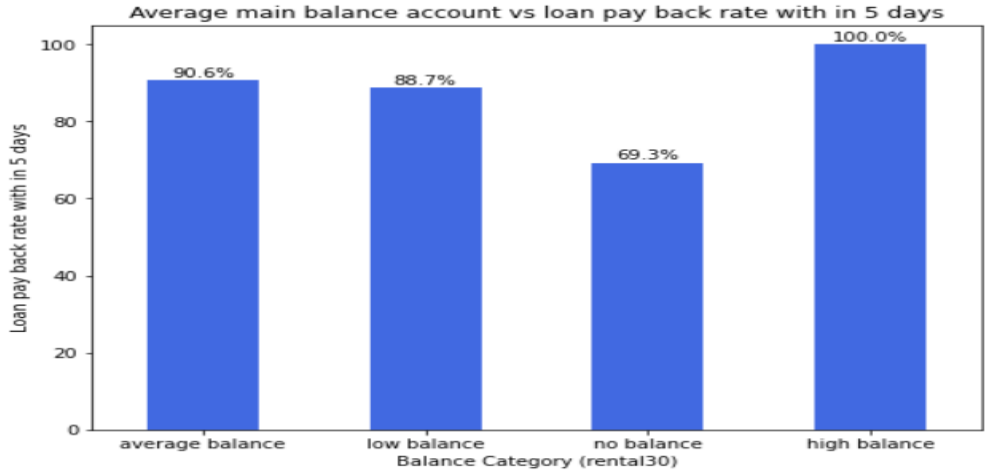
Now let us see what factors affect customer repay ability from the other vital attributes in the dataset.

## 5.2 Factors Affecting the repaying ability of the customer

### 5.2.1 Average Main Account Balance

The below plot infers us how customers with different main balance levels pay back the loan within five days. The high balance level people are with a 100% rate, i.e. they are paying loan within five days. Coming to the average and low balance people, around 10%-12% do not pay the loan within five days.

Coming to low balance level people, it is observed that around 30% of people are not paying back the loan within stipulated five days. The 30% of people with no balance or negative balance create a major loss to the company without paying back the loan within five days.

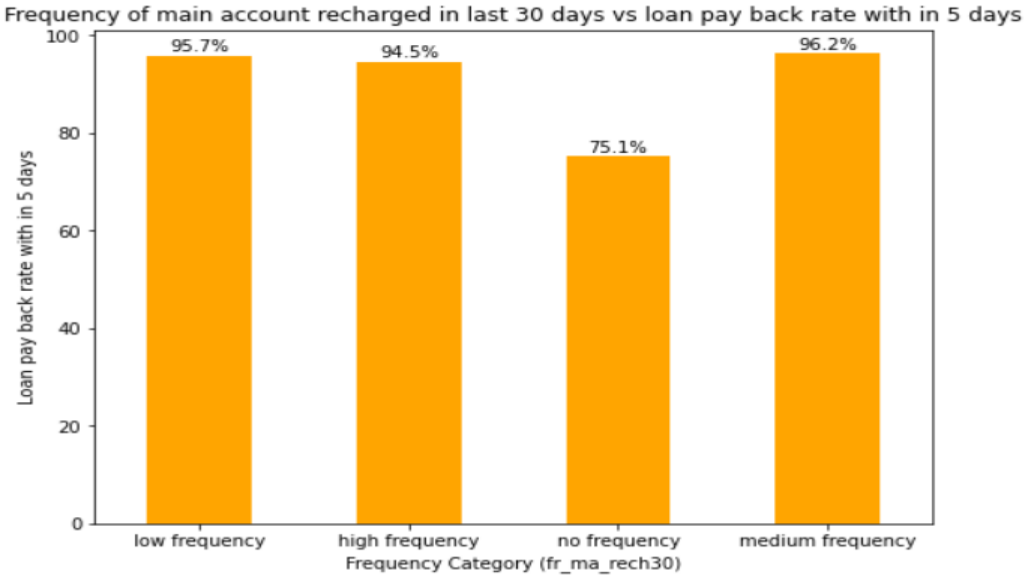



### 5.2.2 Frequence of Recharges

Below bar plot infers us how customers with different frequency levels (main account recharge) are paying back the loan within five days. The is no 100% rate in any of the frequency levels to pay back the loan within five days. Coming to the average and low & medium frequency people, it is observed that around 5%-6% of people are not paying the loan within five days.

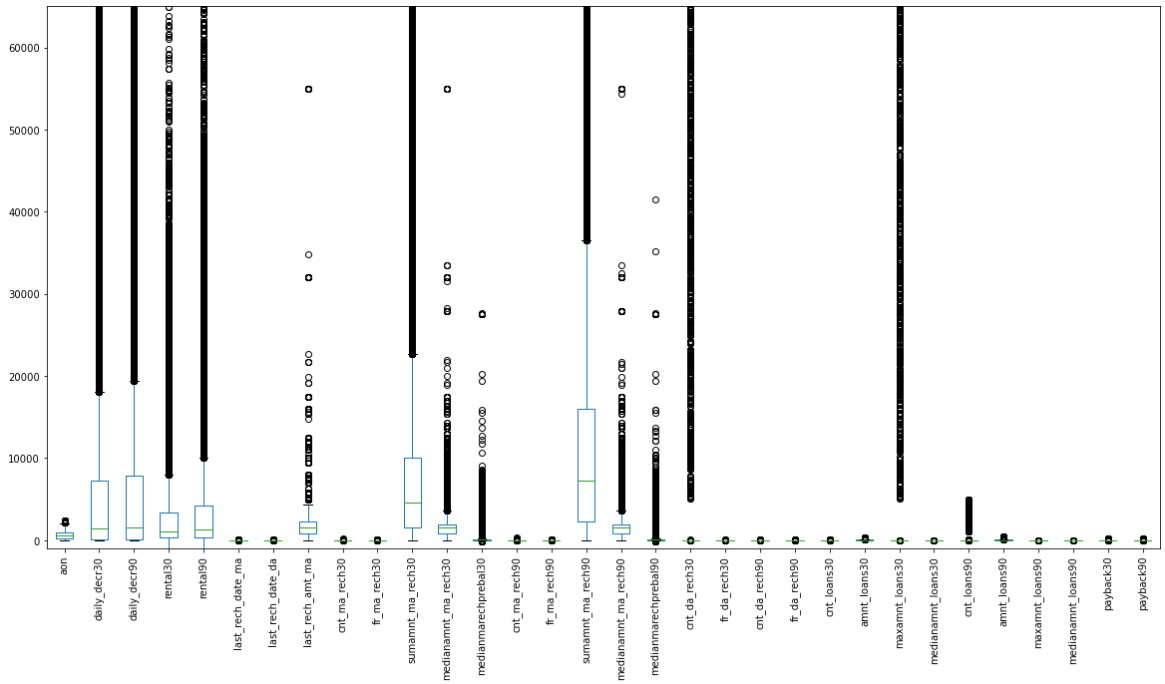
Coming to low-frequency level people, it is observed that around 25% of people are not

paying back the loan within stipulated five days of time. The 25% of people who do not get their main account recharge for 30 days create a major loss to the company without paying back the loan within five days.




## 5.3 Outliers Transformation

Another difficulty faced in this project was large amounts of outlier data. Here is the view of the outliers.

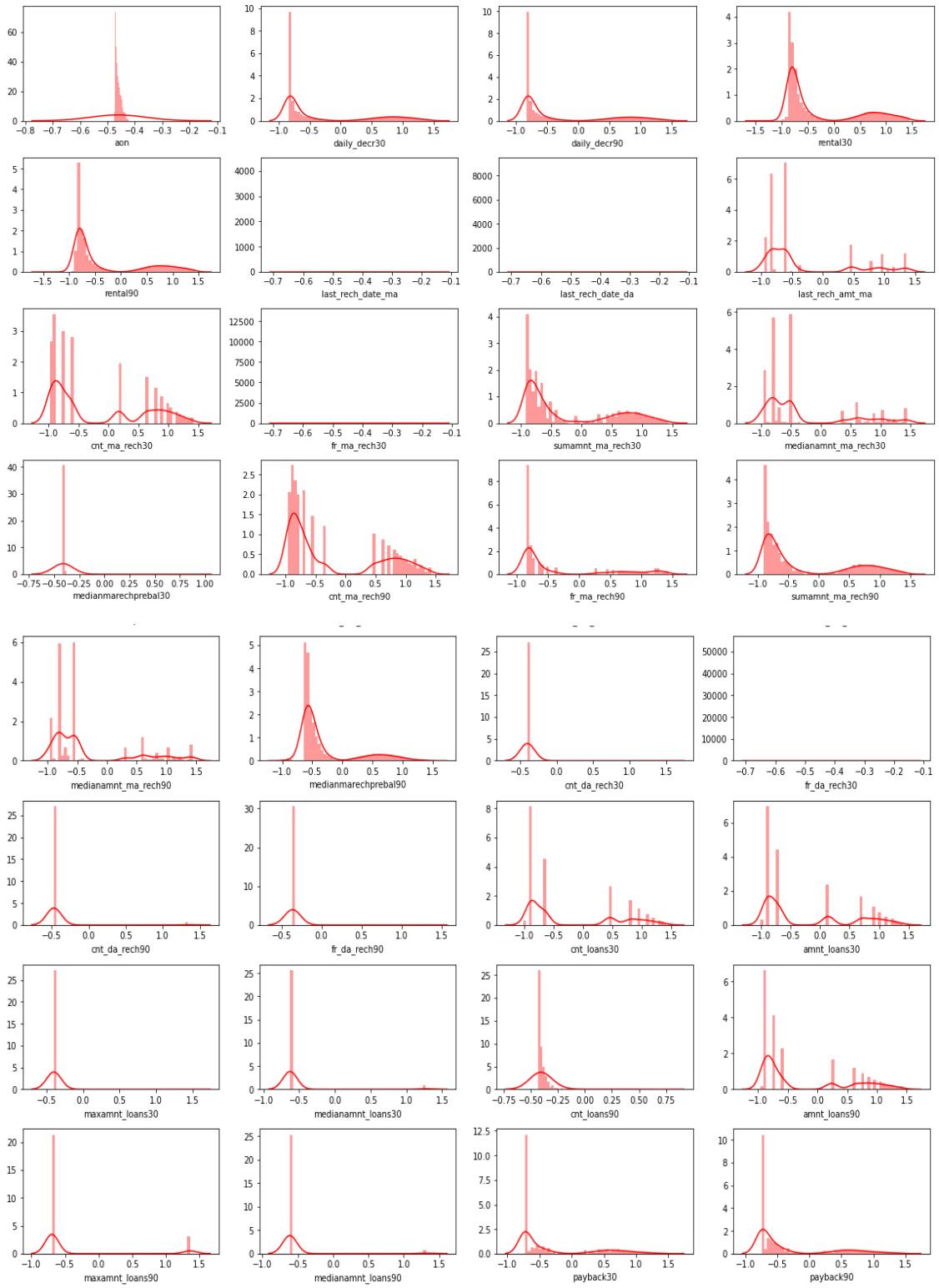



Several changes were made to the dataset to prepare it for analysis. There is no missing data in the dataset; there was no need to perform any null value imputation for the data set. There are outliers for many variables in the data set.

By observing these features, we are doing an outlier's transformation using the imputation technique for the features whose z-score > 3 and <-3. There are many ways to deal with outliers such as imputing outliers with mean, median, mode (categorical), k-NN imputation, mice imputation, or simply removing others. For this data set, we

choose to mean for imputing the outliers with the respective features. After performing mean, we also applied the cube root for the data to bring data closer to normal distribution.

After transforming the outlier using the above logic, we got the below distribution.




So, outlier imputation is far better than merely removing the outliers from the data. As the data set belongs to the loan defaulters or not the outliers are also important for us to get the unbiased results after performing machine learning algorithms.

## 5.4 Experiments

In this section, we will briefly outline the different experiments undertaken before discussing the results below. In this project, we focused on predicting the customer will churn or not from the operator company from the historical data available for some of the customers.

## 5.5 Will a Customer Churn?

To estimate whether a customer will churn or not, we used data to fit a random forest model.

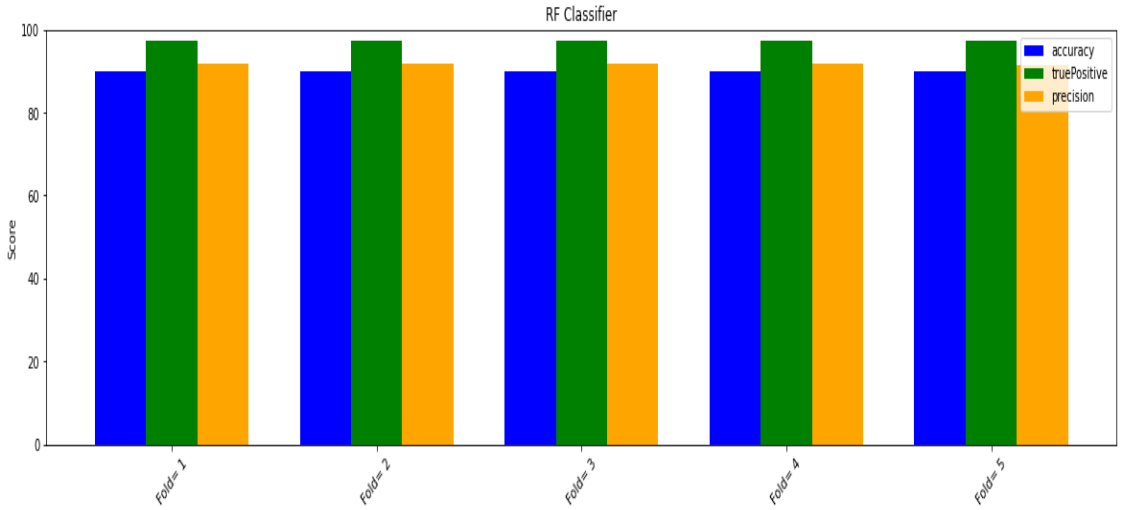
### 5.5.1 Model Selection

When selecting our model, we find the importance of each of our features for the data frames. We applied Principal Component Analysis to reduce the dimensionality of the data, eliminate the problem of multicollinearity, and improve the models' predictive power. We have chosen n_component =13 for our experiment. After we reduce the dimensionality of the data we fit the model for trees between 10 and 1200 and find the OOB score of each, the model which provides the highest OOB score is then the number of trees used to fit our final model.

Since we wanted to perform our experiment with the random forest, we applied the random forest model with different KFold values. We calculated the prediction performance indicators like Accuracy, True Positive Rate, Recall for our experiment for the different Kfold values.

### 5.5.2 Results

Below figures 2 shows the result of our experiment. Random forest model yielded around 89.97% prediction accuracy for the customer churn for the KFold value =4.



| | Accuracy | True Positive Rate | Precision |
|---|---|---|---|
| Fold= 1 | 89.930580 | 97.309613 | 91.683377 |
| Fold= 2 | 89.870942 | 97.298329 | 91.642059 |
| Fold= 3 | 89.878098 | 97.256804 | 91.689352 |
| Fold= 4 | 89.978052 | 97.152570 | 91.887234 |
| Fold= 5 | 89.825373 | 97.296707 | 91.582492 |

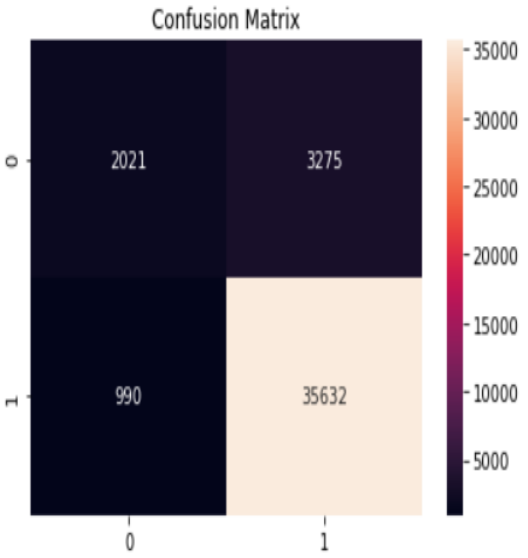



*figure 2 Random forest predictions for different KFold*

Figure 2 shows the Confusion matrix for the random forest fitted with the 41648 records. We can see that the model predicts nearly 90% of the times correctly. Around 10% of the customer churn was classified falsely. We can see that the model has performed very well in most of the cases.

#### 5.5.3 Comparison with a single decision tree

We now wish to compare the random forest with the performance of a single decision tree. We use Kfold validation to test this. We tuned the hyperparameter' max_depth' which controls the tree's complexity by setting the maximum number of nodes before a leaf node is reached. We fitted a model with all 13 features and a max depth of 14. This model's accuracy was 88.8%;

This is reasonably high but not higher than the random forest model, which we would expect. This is because the forest model contains many such trees like this one, which is often right but sometimes wrong, and when this tree makes a mistake, other trees will protect the overall model against it. Again we wish to look at the confusion matrix for this model this can be seen in figure 11. The model appears to perform very well getting most the classifications correct. The errors are evenly distributed, so the model does not show bias.

Below figure displays the result of the random forest with the decision tree. Random forest performed better than the decision tree in all of the cases.

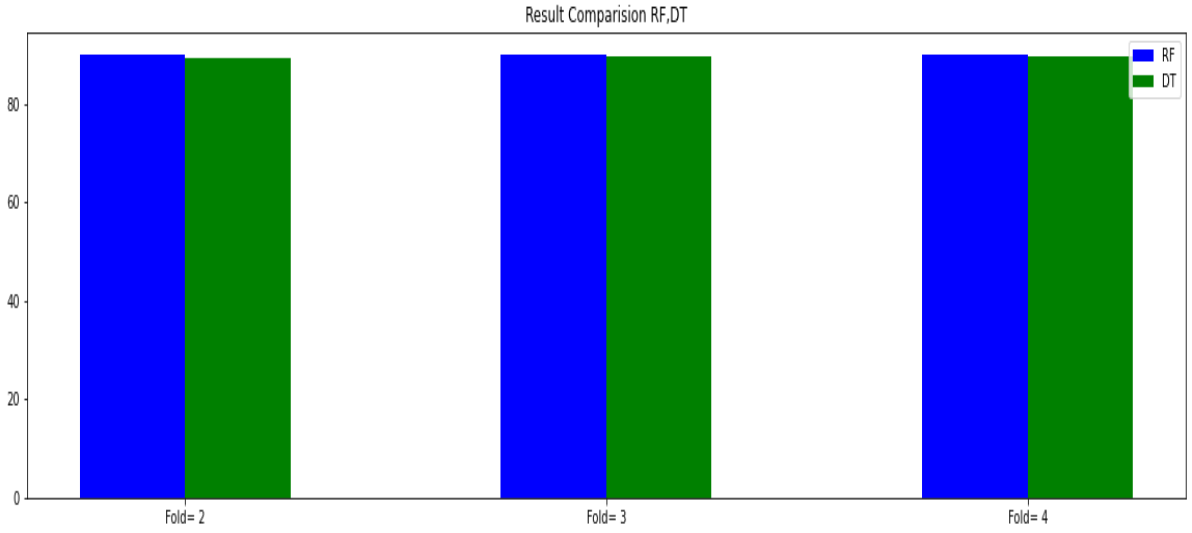



*figure 3 Random forest comparison with decision tree*

We can also see the difference in the performance using the below boxplot. We can see that random forest outperformed the decision tree in all of the cases.

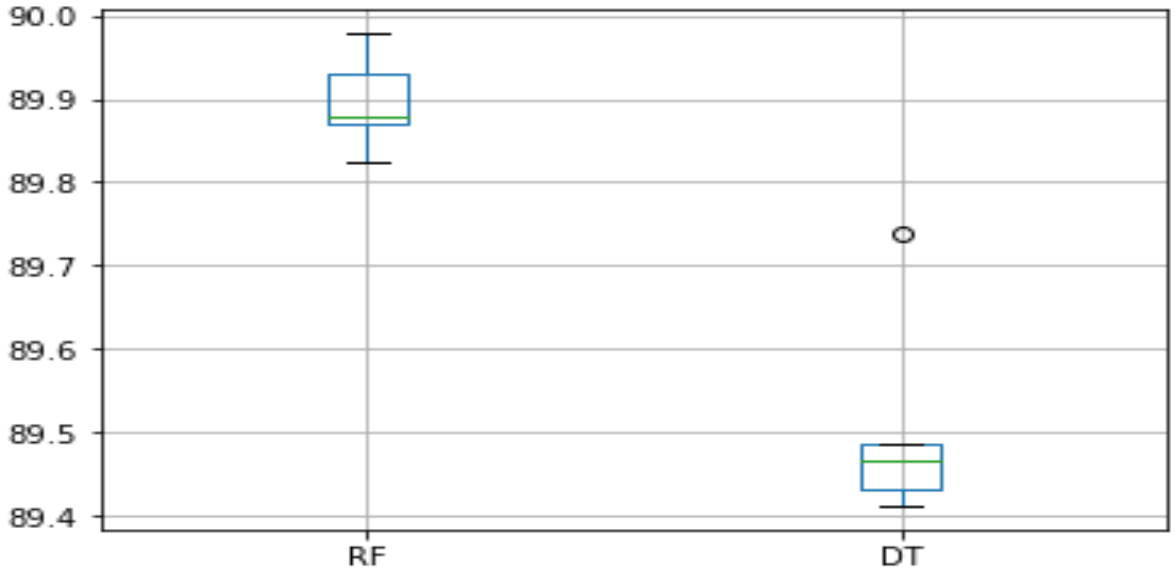



*figure 4 Random forest comparison with decision tree*

# 6 Conclusions and further work

We saw how the data science methodology to Predict Churn in the telecommunications industry. By evaluating results from a technical point of view, we observe that for predicting both churners and non-churners, the models have an overall accuracy of 90.00% for Random Forest. From a practical point of view, the models have an excellent performance of 90% in predicting churners in a telecommunications company.

Decision-making employees can build different marketing approaches to retain churners based on the predictions that have a higher importance in scoring the model performance. These churn prediction models can be used in other customer response models as well, such as cross-selling, up-selling, or customer acquisition. The predictive performance of this data mining methodology can be further improved by efficiently

including the described machine learning algorithms into an ensemble learning structure. Furthermore, text mining and social media mining techniques could be utilised in conjunction with the proposed data mining methodology to reduce the churn rate even more.

# 7 References


[1] M. Shaw, C. Subramaniam, G. W. Tan, and M. E. Welge, "Knowledge management and data mining for marketing," Decision Support Systems, Vol. 31, no. 1, pp. 127-137, 2001.

[2] C. P. Wei and I. T. Chiu, "Turning telecommunications call details to churn prediction: A data mining approach," Expert Systems with Applications, Vol. 23, pp. 103-112, 2002.

[3] V. Umayaparvathi, K. Iyakutti, Applications of Data Mining Techniques in Telecom Churn Prediction, International Journal of Computer Applications (0975 – 8887) Volume 42– No.20, March 2012.

[4]K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A.A. Nanavati, A. Joshi Social ties and their relevance to churn in mobile telecom networks Proceedings of 11th International Conference on Extending Database Technology: Advances in Database Technology, ACM, New York (2008), pp. 668-677

[5] Ahn, Jae-Hyeon & Han, Sang Pil & Lee, Yung-Seop. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications Policy. 30. 552-568. 10.1016/j.telpol.2006.09.006.

[6]T. Yiu, "https://towardsdatascience.com/understanding-random-forest-58381e0602d2."

[7] W. Koehrsen, "https://towardsdatascience.com/an-implementation-and-explanation-of-therandom-forest-in-python-77bf308a9b76."

[8] N. Bhatia, "https://towardsdatascience.com/what-is-out-of-bag-oob-score-in-random-foresta7fa23d710."