STAT-515 Final Project Report

Introduction:

This dataset Customer Churn which is also known as customer attrition occurs when a customer stops using a company's products or services. Customer Churn affects profitability, especially in industries where revenues are heavily dependent on subscriptions such as banks, telephone and internet service providers, TV companies etc. It is estimated that acquiring a new customer can cost up to five times more than retaining an existing one.

Therefore, customer churn analysis is essential as it can help a business which identify problems in its services and make correct strategic decisions that would lead to higher customer satisfaction and consequently higher customer retention. We want to predict the customer churn for a bank. Specifically, we will initially perform exploratory data analysis to identify and visualise the factors that contribute to customer churn. This analysis has helped us to build a model to predict whether a customer will churn or not. This is a typical classification problem as we have used the recall since correctly classifying elements of customers who churned are more important for the bank.

About the dataset:

The source of our dataset is from Kaggle website. Our dataset consists of 14 features and just above 10,000 customers or instances. The key feature is the 'Exited' which is the target variable and estimates whether the customer has churned or not here '0' says NO and '1' says YES. The dataset also consists of instances such as CustomerID, Surname, Creditscore, Geography, Gender, Age, Tenure, Balance, Number of products, has credit card, Is active member and estimated salary. (Adam, 2018)

Exploratory Analysis:

Analysis is a very crucial part for exploring the dataset. To get an in-depth analysis, we have performed correlation tests. We have used the correlation test to measure the linear dependence among the variables. We have performed few visualizations and generated summary statistics for some columns. So different visualization techniques apply to different types of variables, so it's important to differentiate between continuous and categorical variables and analyse them separately.

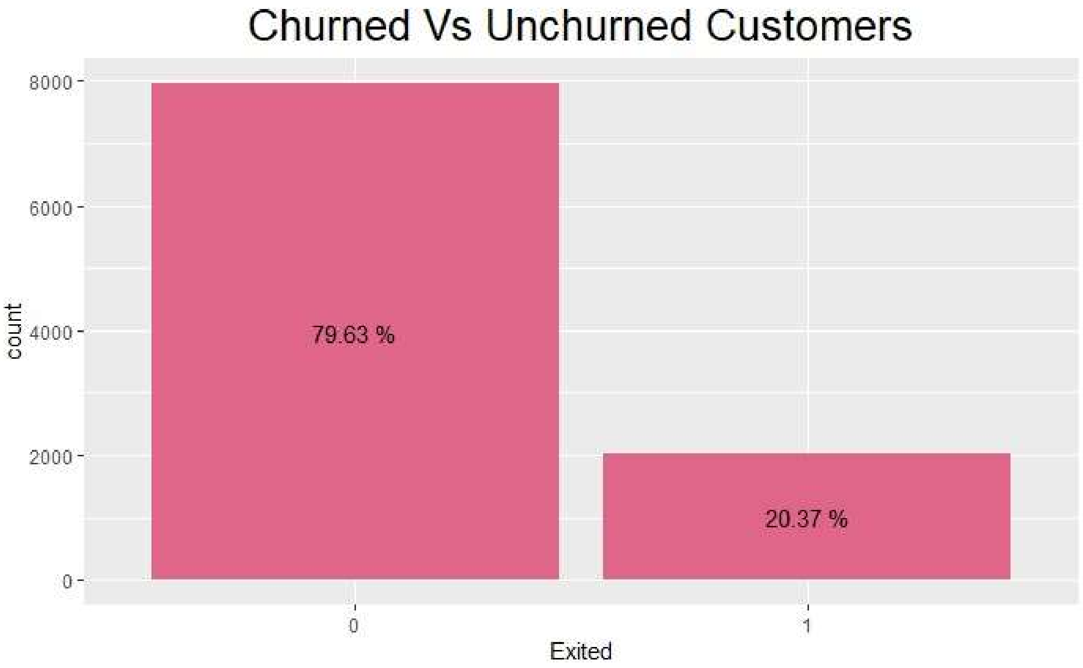


Fig.1

Firstly, we have plotted histogram for exited feature to observe the number of churned and unchurned customers.

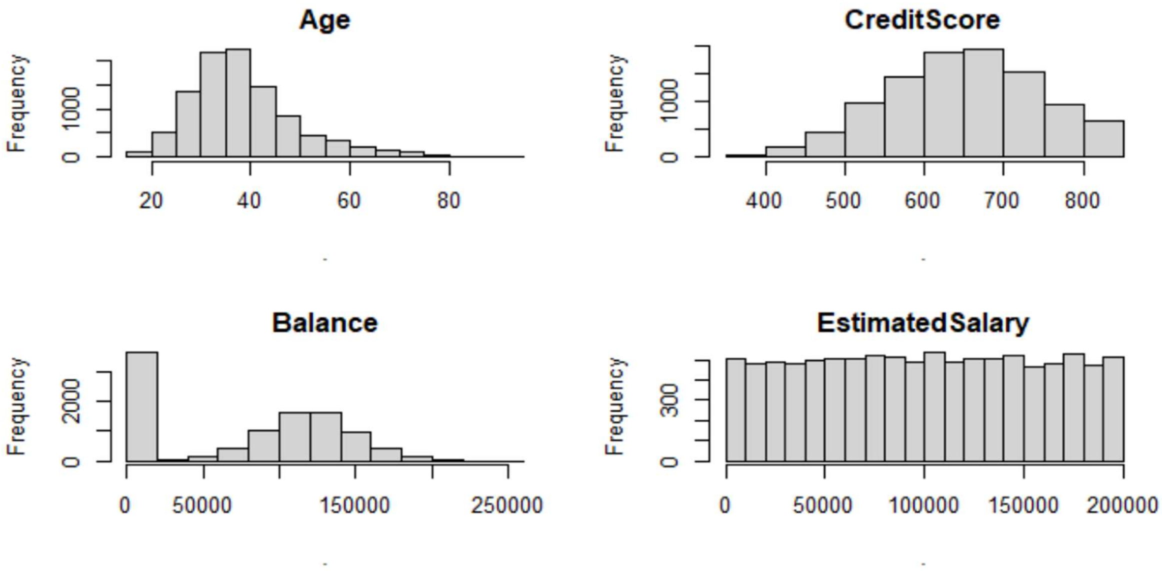
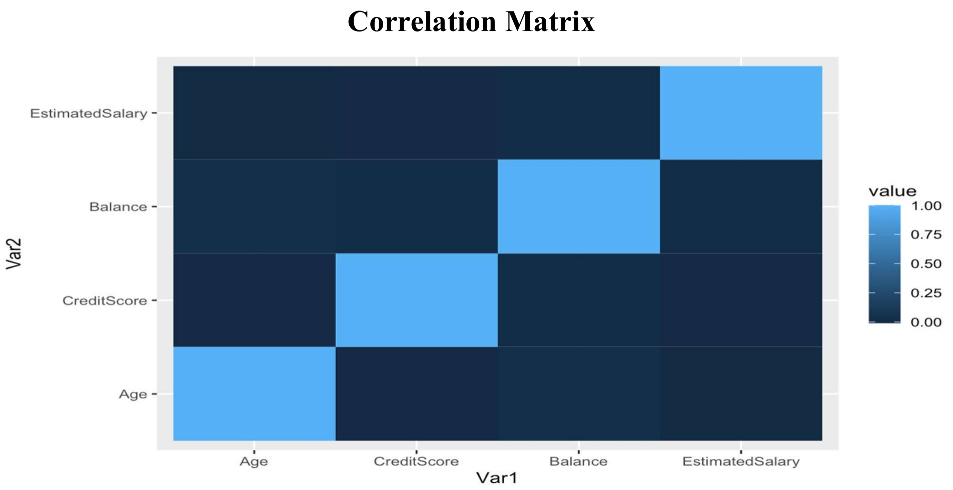


Fig.2

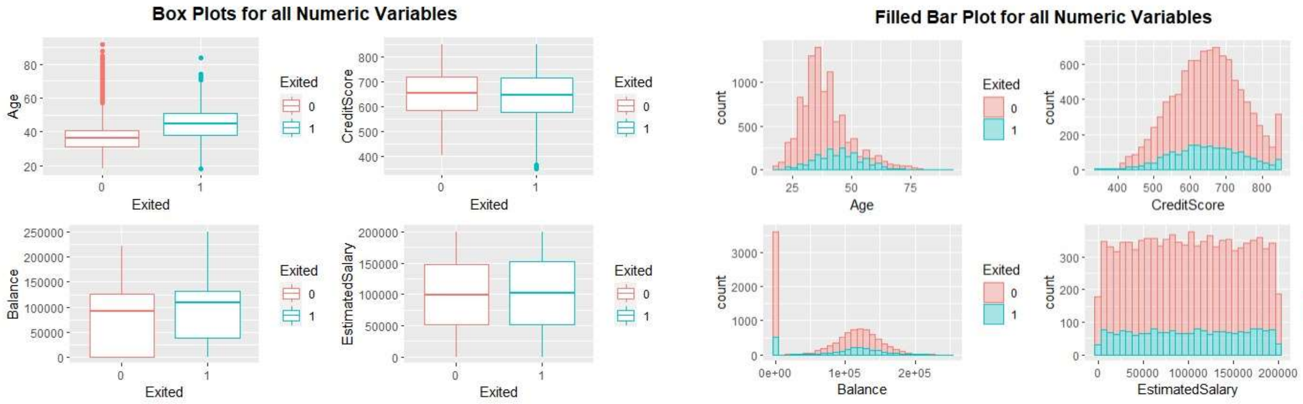
Later we plotted histograms for each of the four continuous numeric features which are Age, CreditScore, Balance and EstimatedSalary. We observed that the age extends furthermore to the left of the median than to the right of the median. Maximum of the credit score values are higher than 600. The balance follows a fairly normal distribution, and the distribution of Estimated Salary is much uniform and provides less information. (MathMastery, 2016)



Correlation Matrix

Fig.3

We tried to look for correlations between the coefficients between every pair of features. But we found that there is no significant intercorrelation between the features. So, we tried to look at these features in greater detail.



Box Plots for all Numeric Variables

Filled Bar Plot for all Numeric Variables

Fig.4

As we used the box plot to observe the relation between the features, interestingly there is a significant difference between the age groups since customers closer to the age of 40 are more likely to churn. This will potentially indicate that there is a preference change with age and the bank has not grouped their strategies to meet their requirements. Similarly, the credit score of churned customers is higher than 600, balance of churned customers is close to $100000.00 and the estimated salary of churned customers is $100000.00. Consequently, we could conclude that salary does not have a significant effect on the churn.

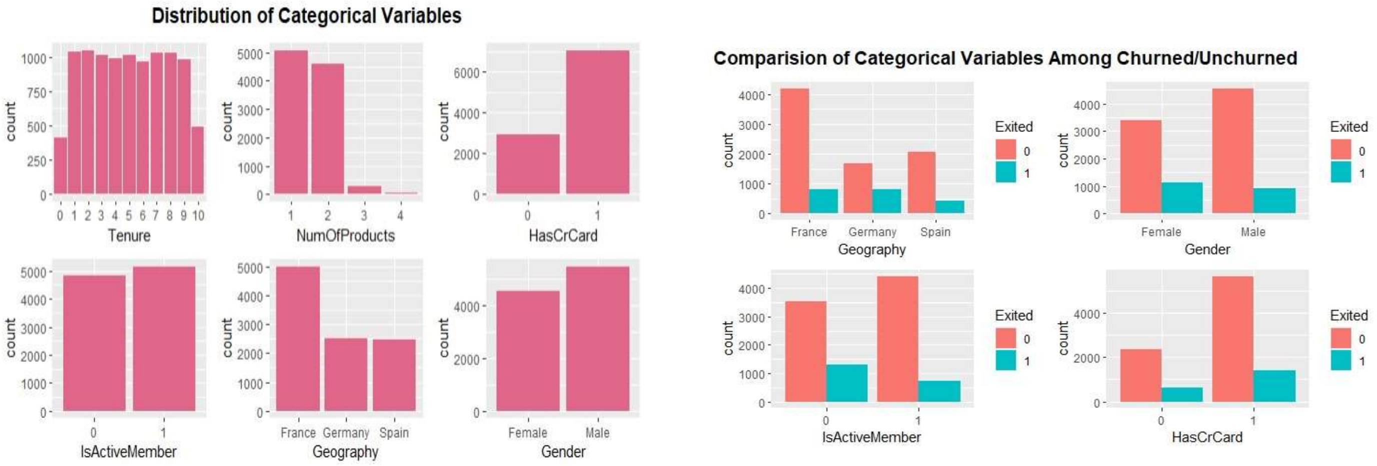


Fig.5

Now we have plotted histograms for categorical features and observed that the bank has only small percentage within the first year but the count of customers in tenure years between 1-9 years is almost the same. Most of the customers have a credit card and almost 50% of customers are not active. The bank has customers in three countries which are France, Germany and Spain and majority of the customers are in France. In the last histogram we observed that there are more male customers than female customers. From the above plot it is observed that customers in France are more likely to churn than the customers in Germany and Spain. There might be plenty of reasons that France has more customers who churn than other countries.

Data Pre-Processing:

Data pre-processing is the process of transforming unstructured data into a format that can be used to create and train Machine Learning models.

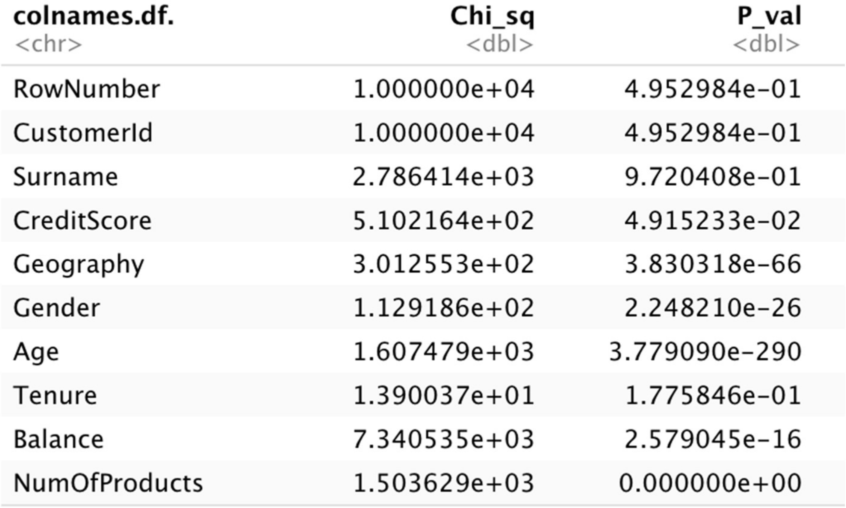
Feature Selection:

EDA revealed a few more features that can be removed because they don't help us predict our target variable:

- Dropping columns 'RowNumber', 'CustomerId', and 'Surname' as they don’t have any significance in modelling.

- 'EstimatedSalary' displays a uniform distribution for both types of customers and can be dropped.
- The categories in 'Tenure' is deemed redundant. This can be confirmed from a chi-square test

- 'Tenure' has a small chi-square and a p-value greater than 0.05 (the standard cut-off value), which confirms our initial hypothesis that these features do not convey any useful information.



colnames.df. <chr>	Chi_sq <dbl>	P_val <dbl>
RowNumber	1.000000e+04	4.952984e-01
CustomerId	1.000000e+04	4.952984e-01
Surname	2.786414e+03	9.720408e-01
CreditScore	5.102164e+02	4.915233e-02
Geography	3.012553e+02	3.830318e-66
Gender	1.129186e+02	2.248210e-26
Age	1.607479e+03	3.779090e-290
Tenure	1.390037e+01	1.775846e-01
Balance	7.340535e+03	2.579045e-16
NumOfProducts	1.503629e+03	0.000000e+00

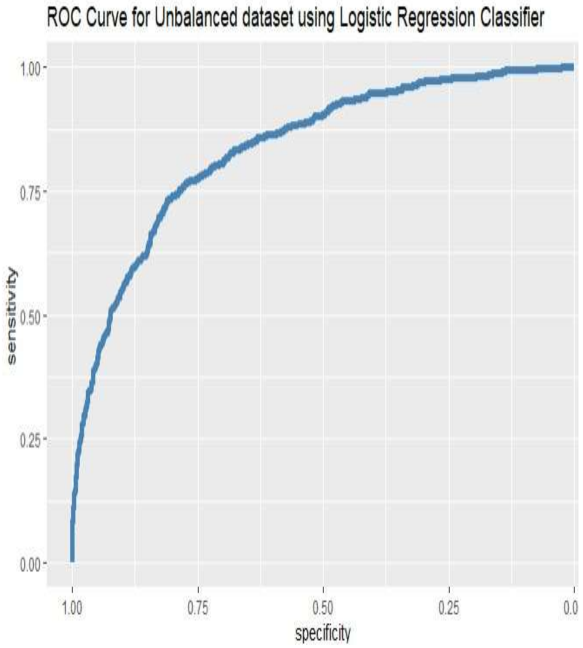
Fig.6

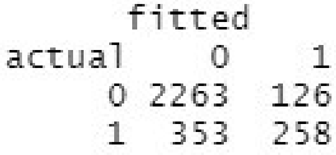
Balancing and Unbalancing data:

Our dataset predicts the occurrence or non-occurrence of an event using binary variables with two potential outcomes (1 and 0 values, respectively).When fitting a logistic regression model, we looked at the statistical effects of balancing data, with the balanced-data scenario. Balancing or not balancing the data to be used for fitting logistic regression models may have an effect on the conclusions that can be drawn from the model. We have tried modelling with both kinds of data.

Logistic Regression:

We have used logistic regression as want to identify whether the customer "will churn" or "won't churn" from the data that accumulates to a binary separation. A logistic regression model would attempt to predict the likelihood of belonging to one of the two groups. The logistic regression is basically a linear regression with the exception that the expected outcome value is in the range [0, 1]. The model will identify relationships between our target feature, Churn, and our remaining features to apply probabilistic calculations for determining which class the customer should belong to. Initially, we have chosen logistic regression because it uses sigmoid function which is optimal to predict binary target variable. So, we ran the model with 70% trained data and 30% of data as test data and observed that unchurned customers were more when compare it to the churned customers which resulted an accuracy of 84.41%. As the classes are unbalanced in the ratio of close to 1:4 ratio we have applied a technique called as SMOT for oversampling which has balanced both the classes and after rerunning the model we observed an accuracy of 73.4%. (Cole, 2020)



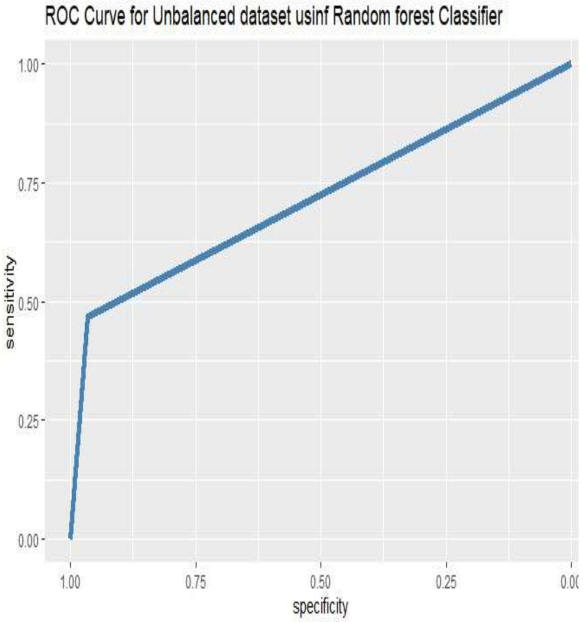


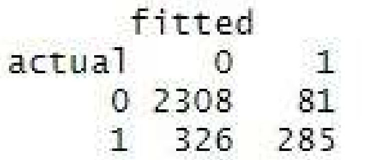
```
         fitted
actual      0    1
     0 2263  126
     1  353  258
```

Fig.7(Confusion matrix for logistic regression classifier Unbalanced dataset)

Random Forest:

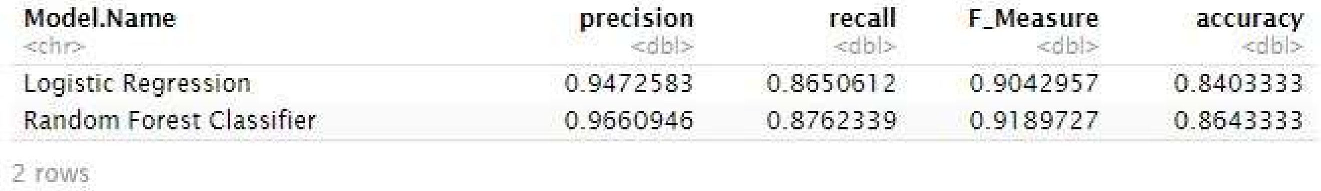
A random forest is an ensemble of decision trees for regression. Each of the regression tree models is learned on a different set of rows or records and a different set of columns describing the attributes, whereby the latter can also be a bit/byte/double vector descriptor. The output model describes an ensemble of regression tree models and is applied in the corresponding predictor node using a simple mean of the individual predictions. Random forest is an average multiple deep decision tree that is trained using a training set and with the goal of overcoming the over-fitting problem of the classification or simple decision tree. Random forest is used when we have both continuous and categorical variables. We have implemented this model as our target variable is binary, we have used the random forest classifier to predict the class probabilities for all the data instances to check the instances fall under which class. Considering 0.5(probability) as base line to divide the classes. Also, as there are equal number of numerical and categorical variables in the dataset, we have used the random forest model. We ran the model with 70% of trained data and 30% of test data by which we observed an accuracy of 86.5%. But we observed the ROC curve seems to have a share elbow edge with sensitivity less than 0.5 which shows that we have scarcity of data in one of the classes. Because of which we have used a technique called as SMOTE oversampling, because of oversampling we could attain an accuracy of 79.5%. (Yui, 2019)





```
       fitted
actual    0    1
     0 2308   81
     1  326  285
```

Fig.8(Confusion matrix for random forest classifier Unbalanced dataset)



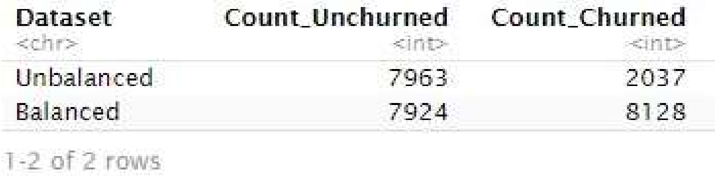
Model.Name <chr>	precision <dbl>	recall <dbl>	F_Measure <dbl>	accuracy <dbl>
Logistic Regression	0.9472583	0.8650612	0.9042957	0.8403333
Random Forest Classifier	0.9660946	0.8762339	0.9189727	0.8643333

2 rows

Fig.9

SMOTE (Synthetic Minority Oversampling Technique):

In the dataset we have 10000 samples out of which around 8000 samples have a negative response and around 2000 samples have a positive response. This is clearly an unbalanced dataset. So, to balance the rare event occurred in the dataset, we are using SMOTE function. SMOTE is nothing but Synthetic Minority Oversampling Technique which is used to over sample the rare event in the dataset. This function internally uses Bootstrapping and K-Nearest Neighbors for synthetically creating more events. By applying the SMOTE function, we can achieve a balanced response for each positive and negative event with around 8000 samples in each event. (Brownlee, 2020)



Dataset <chr>	Count_Unchurned <int>	Count_Churned <int>
Unbalanced	7963	2037
Balanced	7924	8128

1-2 of 2 rows

Fig.10

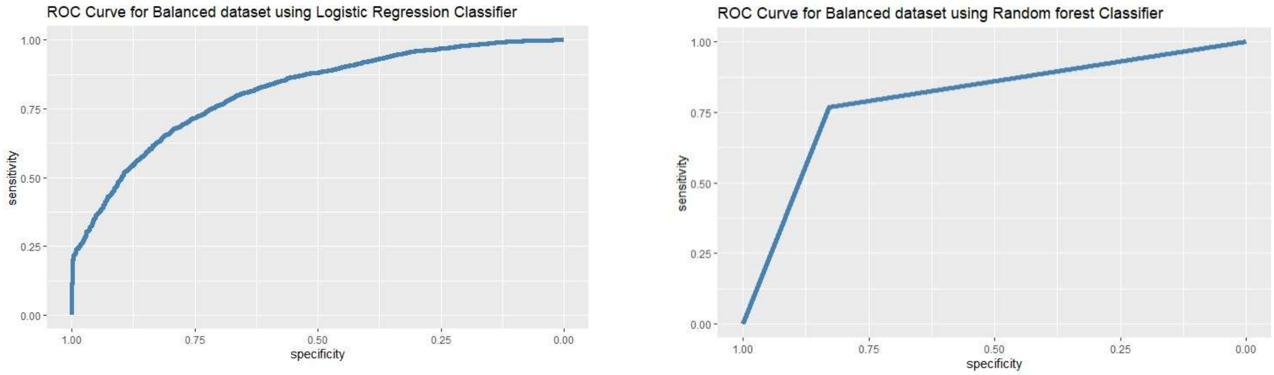
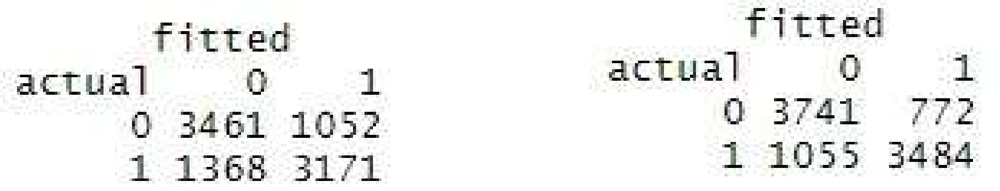


Fig.11

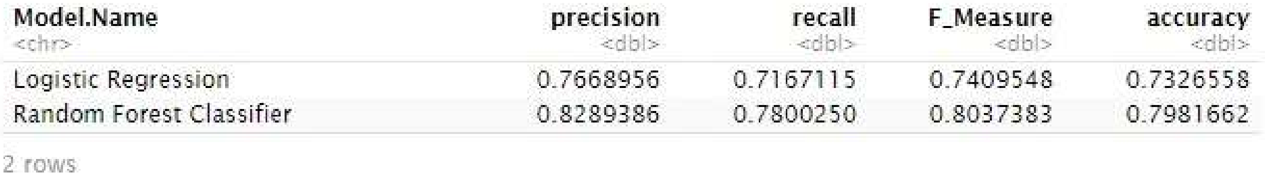
The area under the logistic ROC curve decreased for the balanced dataset compared to the imbalanced. The Random Forest ROC for balanced data also has a sharp bend elbow which indicates that there is scarcity of points, but the area under the curve increased now. This indicates that prediction accuracy with balanced dataset has increased. Random Forest has a better ROC area compared to the Logistic model for balanced dataset.



```
      fitted
actual    0    1
     0 3461 1052
     1 1368 3171
```

```
      fitted
actual    0    1
     0 3741  772
     1 1055 3484
```

Fig.12(Confusion matrix for logistic regression and random forest Classification balanced)



Model.Name	precision	recall	F_Measure	accuracy
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
Logistic Regression	0.7668956	0.7167115	0.7409548	0.7326558
Random Forest Classifier	0.8289386	0.7800250	0.8037383	0.7981662

2 rows

Fig.12

Conclusion:

We built Logistic Regression and Random forest classification models to check which model gives best prediction of churning. As the churned data is imbalanced (1:4), we applied oversampling technique called SMOTE to get a balanced dataset. So, we compared both the models before and after balancing the data. We observed that accuracy of the Logistic Model is 83.7% and that of the Random Forest is 85.9% with imbalanced dataset. We observe that there is an increase in accuracy when Random Forest classifier is used on imbalanced data. When the data is balanced we observed that the accuracy of both the models decrease with 73.2% for logistic regression and 79.8% for random forest classifier. So are the Precision, Recall and F – Measure measures. So, now we tried to compare both the models using ROC (Receiver Operating Characteristics) curve plot, which evaluates the prediction Accuracy with True Positive Rate (Sensitivity) on Y- axis and False Positive Rate (Specificity) on X – axis. We observe that Logistic ROC for imbalanced dataset is curved because prediction values are in range [0,1]. But the Random Forest ROC has a sharp bend elbow which indicates that there is scarcity of points. Logistic has a better ROC area compared to the Random Forest for imbalanced dataset. The area under the logistic ROC curve decreased for the balanced dataset compared to the imbalanced. The Random Forest ROC for balanced data also has a sharp bend elbow which indicates that there is scarcity of points, but the area under the curve increased now. This indicates that prediction accuracy with balanced dataset has increased. Random Forest has a better ROC area compared to the Logistic model for balanced dataset. As the sensitivity increases in the Random Forest Model, generalizations with new data can be done more effectively. So, Random Forest Model with balanced dataset has the best prediction accuracy compared to the Logistic.

References

Adam. (2018, Mar). *kaggle*. Retrieved from https://www.kaggle.com/adammaus/predicting-churn-for-bank-customers

Brownlee, J. (2020, Jan). *machine learning mastery*. Retrieved from https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Cole, A. (2020, May). *towadsdatascience*. Retrieved from https://towardsdatascience.com/predicting-customer-churn-using-logistic-regression-c6076f37eaca

MathMastery. (2016). *math is fun*. Retrieved from https://www.mathsisfun.com/data/standard-normal-distribution.html

Yui, T. (2019, June). *towardsdatascience*. Retrieved from https://towardsdatascience.com/understanding-random-forest-58381e0602d2