# CHURN PREDICTION OF TELCO CUSTOMERS USING VISUALIZATION TECHNIQUES AND MACHINE LEARNING ALGORITHMS

REVIEW REPORT

Submitted by

**K. SAI KOUSHIK KALAKOTA (18BCE0561)**

**ROHITH SAI G (18BCE0451)**

Prepared For

**Data Visualization (CSE3020) – PROJECT COMPONENT**

Submitted To

**Ms. NALINI N**

**Assistant Professor (Senior)**

**School of Computer Science and Engineering**


VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

**Table of Content**

## 1. Abstract

Telecom industry in the past few years has become the largest sector of Information and Communication technology, the present companies are mostly made up of telephone companies and internet service providers, these companies are playing a crucial role in development of these services. On establishment of a business, companies try to accumulate customers and based on the services if there is a higher your customer churn rate, the lower your chances of growing your business. Even if that business has some of the finest marketing campaigns in the total industry, its bottom line suffers if it is losing customers at a high rate, as the cost for acquiring new customers is always high. Customer churn is a important evaluative metric because the cost to retain existing customers is less than that it is to acquire new customers. In this project we have taken a Telecom Customer Churn dataset from Kaggle which includes many attributes which determines the churn of a customer and we have applied Visualisation techniques to visualise the data and have tried to build a predictive model using Machine learning techniques.

## 2. Keywords

Data Visualisation, Churn, Logistic Regression, Decision Tree, XGBoost, AdaBoost, Random Forest

## 3. Introduction

### 3.1 About Telecom Industry:

The telecommunications sector didn't a major contribution in the early 1950's but now it has become one of the major industries in developed countries. The gradual progress in technology and the increasing demand from numerous operators raised the level of competition in this industry. There are multiple strategies which are being used by companies to survive in this competitive market. In recent years in India, Jio has attracted many customers from other telecom companies and reduced the business of these companies, the mobile data cost in general is very high, but after a clear strategy of keeping low prices for a year, giving free services for 3 months for each Sim Card and providing high rate services enabled Jio Industries to increase the cost but keeping their churn at bare minimum. Many factors like Internet Service, Streaming Movies and many other apps developed by Jio helped them retain their customers.

### 3.2 Strategies to Reduce Churn:

In a business developing strategy, customer retention is an important factor affecting the whole business. Out of the total revenue generated, Telecom industries apply strategies like giving discounts or giving extra data packs or providing additional features like Streaming services, so that they can retain their customers. Obtaining new customers always cost more than retaining their original customers.
For a business to generate more revenues, three main strategies have been proposed
(1) acquire new customers,
(2) upsell the existing customers, and
(3) increase the retention period of customers.

• While, comparing these strategies we will be taking the value of return on investment (RoI) of each into account and it has shown that the third strategy i.e. increase in retention is the most profitable strategy. Thus, it proves that retaining an existing customer has much lower costs than acquiring a new one. To apply the third strategy, companies have to decrease the rate of customer's being churn, known as "the customer movement from one provider to another."

• Many researchers confirmed that machine learning technology is highly efficient to predict this situation. This includes ML techniques being applied for the purpose of learning from previous data. The models were also evaluated by using a testing dataset and the impact of applying models to check churn was tested. These models gave good results and the strategies

followed the factors from these models and the models were deployed to production.

### 3.3 Motivation:

Our Motivation is to find the Teleco Customers who are more likely to unsubscribe from Telco service and retaining them rather than loosing. So, the churn of Customers will be reduced and business can be grown. This is not only restricted to Telco industry but also for bank sector, entertainment sector etc.

### 3.4 Objective:

This project aims to build a predictive model to identify the Churn of the customer using the service of a Telecom company. The idea is to find out the factors in the service which is influencing the customers most. So that it would be helpful to the companies to improve their services and retain the original customers and attract the original customers. Using this we are trying to understand which factors are mainly useful to decrease the churn rate of customers.

We used various plots to visualize the data for better understanding.

After the visualizations we performed data preparation – converting data types of columns, deleting null values and creating dummy variables for categorical data. After that we perform dataset balancing using:

1.SMOTE

2.Random Oversampling

3.ADASYN

Then we perform feature engineering to check the importance of features.

Algorithms we are using for model building are:

1. Logistic Regression

2. Decision Tree

3. Random Forest

4. Bagging Classifier

5. Gradient Boosting Classifier

6. AdaBoost Classifier

7. XGBoost Classifier

After training the model we perform HyperParameter Tuning and then model testing.

### 3.5 Why to use Visualisation?

Analysing the data through visualizations and providing reasoning to them makes this complex data more accessible, understandable and usable. For a human brain, seeing the analytical results presented visually rather than text or numbers, it makes interpretation easier and makes it easier to identify patterns trends and correlations between attributes in data.

- To explore sources
- To identify areas of business that can be improved
- To tell stories
- Patterns and trends between attributes in data can be identified easily
- To understand what factors, influence customers' behaviour.

### 3.6 Significance and Applications:

- Customers' churn has been an huge concern in service sectors as they are being forced with high competitive services. On the other hand, when we are predicting the customers who are most likely to leave the company i.e. who are having most probability of churn, will represent potentially large additional revenue source if it is done in the early phase
- Having the ability to accurately predict future churn rates is necessary because it helps your business gain a better understanding of future expected revenue..
- In addition, when you're able to predict churn to forecast the potential churn rate of a particular customer, we can target the factors which the customer is having problems with, in an attempt to prevent them from discontinuing their subscription with you.
- Increased revenue
- Higher referrals
- More customer acquisition channel

### 3.7 Scope:

- To identify the Churn of the customer through visualization techniques
- To infer the factors which influencing the customers most

- Usage of Machine Learning algorithms like Linear Regression, Logistic Regression, Random Forest etc., to predict the churn of the future customers
- So that it would be helpful to the companies to improve their services and retain the original customers and attract the original customer

### 3.8 Contributions of the Project

18BCE0451 - Data pre-processing, Logistic regression, Decision Tree, Random Forest, Vis techniques

18BCE0551 – Data pre-processing, Dataset Balancing, Bagging Classifier, Gradient Boosting Classifier, AdaBoost Classifier, XGBoost Classifier

### 3.9 Dataset Used:

- The dataset we used in our project is – Telco Customer Churn, from Kaggle.com
- LINK TO THE DATASET: https://www.kaggle.com/blastchar/telco-customer-churn
- It is a single CSV file using test_train_split we divide the training and testing datasets.
- It consists of 7043 rows(instances) of 21 columns(attributes).

**Packages Used:**

- Sklearn
- Seabron
- Matplotlib
- Plotly
- Pandas
- Numpy
- Imblearn
- Xgboost

## 4. Literature Survey

### 4.1 Background Work:

Customers churn may be a considerable concern in commission sectors with highly competitive services. On the opposite hand, predicting the purchasers who are likely to go away the corporate will represent potentially large additional revenue source if it's wiped out the first phase. Having the power to accurately predict future churn rates is important because it helps your business gain a far better understanding of future expected revenue. In addition, when you are using a model for churn prediction, it can forecast the potential churn rate of a specific customer, allowing the business master to focus on that individual in an effort to prevent them from discarding their services from the business.

| Authors | Title | Publisher | Main Findings |
|---|---|---|---|
| CHAO-YING JOANNE PENG KUK LIDA LEE GARY M. INGERSOLL Indiana University-Bloomington | An Introduction to Logistic Regression Analysis and Reporting | IEEE | From the word regression analysis, we can understand that the logistic regression is a predictive analysis. This model is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.<br>It is a very powerful analytical technique for use when the outcome variable is categorical data as it shows high performance values for most data. The effectiveness of the logistic model can be throughly checked<br>(a) testing the models with null and against null models,<br>(b) the test performance off each predictive model,<br>(c) descriptive and inferential fitness of each attribute in model,<br>(d) and predictive probabilities of models. |

| | | | |
|---|---|---|---|
| Simon Bernard, Laurent Heutte and Sebastien Adam | On the Selection of Decision Trees in Random Forests | Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, June 14-19, 2009 | In this paper, for the selection of tree for Random forests algorithm is discussed. The goal was to highlight that some particular subsets of trees of a RF are able to perform better than this initial forest. The well-known selection methods used are :<br>1)SFS (Sequential Forward Selection)<br>2)SBS (Sequential Backward Selection).<br>In spite of the usage of these SFS and SBS methods, the are not perfect and this work has shown that it always exists a another subset of well selected trees which are able to produce higher performance than an ensemble grown with a RF induction algorithm . Thus an interesting aspect we found out from the experiments in paper is that when we apply some other classifier selection methods like the Branch and Bound method, they are known to be more efficient than SFS and SBS. It was a good method to allow to see till which extent a subset of trees are considered more perfect and are able to outperform the whole ensemble of trees. |
| V.Umayaparvathi1,<br>K. Iyakutti2 | Automated Feature Selection and Churn Prediction using Deep Learning Models | IRJET | There are straightforward choice standards based models and complex arrangement models for churn prediction task has been proposed in the writing. While these techniques are productive in playing out the stir expectation task, they require manual element designing cycle which tedious and error prone.This paper introduced, a procedure of utilizing machine learning models to kill the manual component engineering cycle. They made three deep neural |

| | | | |
|---|---|---|---|
| | | | organization designs for the churn prediction task. experiments were led utilizing two genuine world datasets CrowdAnalytix and Cell2Cell. Out trial results show that deep learning models performing similarly on a par with customary classifiers, for example, SVM and random forest. |
| Matthew N. O. Sadiku, Adebowale E. Shadare, Sarhan M. Musa and Cajetan M. Akujuobi | DATA VISUALIZATION | Sretech journal | This paper presents a short introduction to information image.Data visualization involves presenting information in graphical or pictorial type that makes the data straightforward to know. It helps to elucidate facts and verify courses of action. it'll profit any field of study that needs innovative ways that of presenting giant, complicated data. the arrival of tricks has formed trendy image. information image is that the method of representing information during a graphical or pictorial approach during a clear and effective manner. it's emerged as a robust and wide applicable tool for analyzing and decoding giant and sophisticated information. it's become a fast, straightforward suggests that of transfer ideas during a universal format. It should communicate complicated concepts with clarity, accuracy, and potency. These advantages have allowed information image to be helpful in several fields of study. |
| Samuel Soma Ajibade, Anthonia Adediran | An Overview of Big Data Visualization Techniques in Data Mining | IEEE | The analytics of data holds a crucial perform by the reduction of the scale and complex nature of knowledge in data processing. knowledge image could be a major methodology that aids huge knowledge to induce associate degree absolute knowledge perspective and in addition the invention of |

| | | | |
|---|---|---|---|
| | | | knowledge values.The main focus of this paper is to administer a quick survey of a number of the multi-dimensional image techniques that area unit employed in data processing, knowing absolutely well that the techniques don't seem to be restricted to those that are mentioned during this paper as there area unit way more to the present.The use of the information image techniques employed in data processing may be fascinating and every now and then difficult in addition |
| Hakki Candan Cankaya, Turker Ince | Customer Churn Prediction for Telecom Services | IEEE | Churn is the hardest problem of the three posted problems for the KDD 2009 competition. IBM has the highest 0.7651 ROC area score. In this study, we evaluate alternative machine learning methods aiming at matching or improving the best scores recorded at the KDD 2009 competition. We also focus on efficient use of computational resources. As the continuation of this study, we will include other ensemble methods and complete the comparative analysis. We will also work on intelligent ways of dealing with the dataset impurities and reducing overall complexity. We believe that this analysis should also lay a good foundation for the other churn prediction problems, where the proposed methods should be applicable with little or no modifications |
| A Proposed Churn Prediction Model | A Proposed Churn Prediction Model | International Journal of Engineering Research and Applications (IJERA) | Many churn prediction models and techniques have been presented to date. However, a simple model is required to distinguish churners from non-churners then clustering the resulted churners for providing |

| | | | |
|---|---|---|---|
| | | | retention solutions. In this paper, a simple model based on DM techniques was introduced to help a CRM department to keep track its customers and their behavior against churn. A data set of 5000 instances with 23 attributes is used to train and test the model. Using 3 different techniques which are DT, SVM, and NN for classification and simple K Means techniques for clustering results indicate that the best output for the data set in hand is SVM technique. |
| Sahand KhakAbi, Mohammad R. Gholamian, Morteza Namwar | Data Mining Applications in Customer Churn Management | IEEE | In this paper, the recent literature in the area of application of data mining techniques in customer churn management is reviewed from two different perspectives: the applied model and the statistics of publication. The primary aim of this paper is to provide a big picture for contributors to help them determine the potential research points and areas. The results also indicate that the number of publications in the area of application of data mining techniques in customer churn management has soured in the last two years and the most active publisher in this field is Expert Systems with Applications |
| Abinash Mishra, U. Srinivasulu Reddy | A Comparative Study of Customer Churn Prediction in Telecom Industry Using Ensemble Based Classifiers | IEEE | Many Telecom companies are facing difficulty to predict the customer who is likely to leave the services. In Telecom Industry, Churn Prediction is a fundamental problem which is gaining attention of many researches in the recent years. The experimental results shows that the Random Forest is the best Classifier for the Churn Prediction Problem when compared to others models, in |

| | | | |
|---|---|---|---|
| | | | terms of all the performance measures like accuracy, sensitivity, specificity and error rate. The early churn prediction can prevent the company loss by predicting the customer behaviour |
| M.BALASUBRAMANIAN, M. SELVARANI | CHURN PREDICTION IN MOBILE TELECOM SYSTEM USING DATA MINING TECHNIQUES | International Journal of Scientific and Research Publications | From this study, it's determined from the performance factors that call tree model surpasses the neural network model within the prediction of churn and it's additionally simple to construct. choosing the correct combination of attributes and fixing the correct threshold values could turn out a lot of correct results. This study limits itself with prediction of churn and no steps were analyzed to incorporate retention policies. the longer term analysis direction could also be to analyse the proper prediction policies by choosing applicable variables from the dataset. |
| S. Babu, Dr. N. R. Ananthanarayanan | A Review on Customer Churn Prediction in Telecommunication Using Data Mining Techniques | IJSER | Customer churn is the term which demonstrates the customer who is in the stage to leave the organization. Especially it is going on intermittently in the media transmission industry and the telecom enterprises are likewise in a situation to hold their client to evade the income misfortune. Prediction of such conduct is indispensable for the current market and competition. So the research is focused here and this paper reviwed around 64 research papers in the point of recognizing the data Mining methods and models used to foresee the Customer churn. Likewise, learn about the absence of the current models, there by characterizing the new model to anticipate the churn in the telecom industry. |

| | | | |
|---|---|---|---|
| Mumin Yuldiz, Songul Albayrak | Customer Churn Predictionin Telecommunication with Rotation | Research Gate | As per the outcomes, rotation forest is best to C4.5 decision tree and antminer+ churn customer rate's true prediction increasing graph is very important. The distinction of between rotation forest and antminer+ calculations is 36.31% in unique dataset for affectability rate. Balancing information is expanded all affectability rates. As indicated by this outcomes rotation forest strategy is the best calculation and 18.81% more effective than antminer+ as far as affectability |
| Maria Cristina Ferreira de Oliveira and Haim Levkowitz | From Visual Data Exploration to Visual Data Mining: A Survey | IEEE | Information Visualization is being used in some applications and in those applications we that involves data mining of large table databases to build a Web accessible resource providing information about visualization and Data Mining techniques, tools, data sets, and research projects.<br>According to the paper, a future technique is developed - “Visual DM technique” denotes more than the traditional application of a visualization technique to support non-analytic stages of a KDD algorithm steps, but in Data Mining algorithms visualization plays a major role |
| Shweta Taneja, Charu Gupta, Kratika Goyal, Dharna Gureja | An Enhanced K-Nearest Neighbor Algorithm Using Information Gain and Clustering | IEEE | In this paper, the traditional KNN algorithm is discussed along with the major factors affecting the algorithms and some improvements are suggested to overcome these factors. The techniques presented for KNN algorithm are dynamically selected, attribute weighted and distance weighted techniques. This proposed algorithm by the authors has improved the accuracy of classification and reduced the execution time of algorithm. |

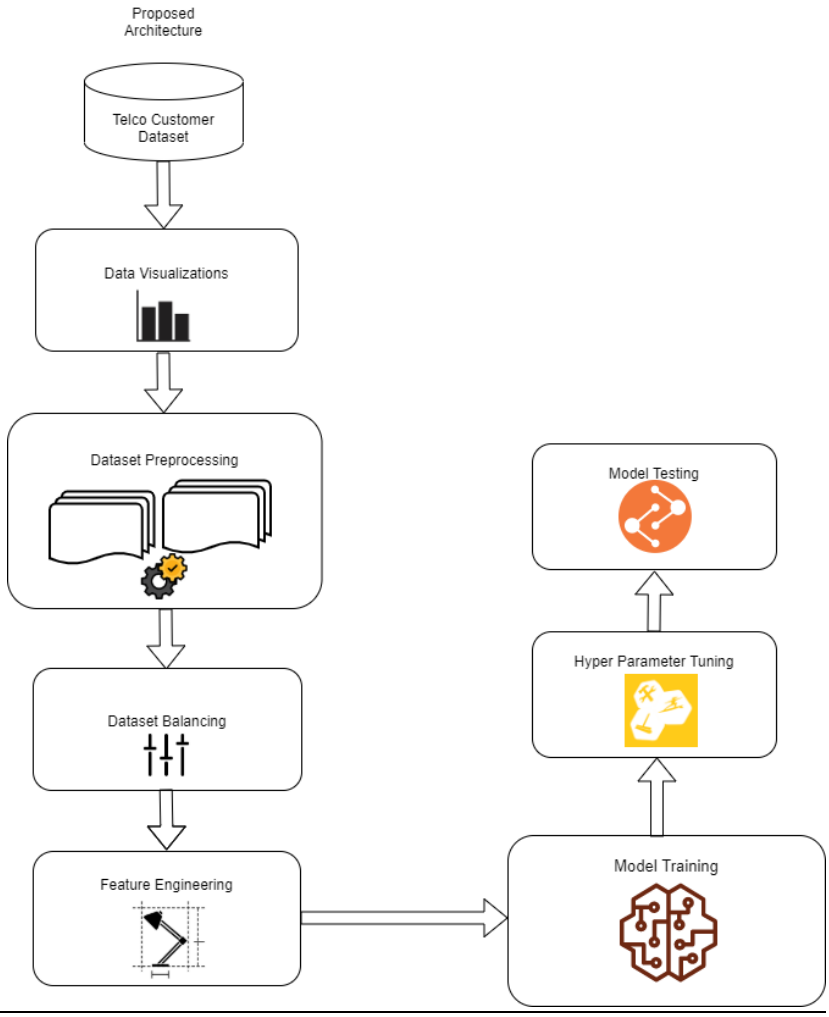
| | | | |
|---|---|---|---|
| | | | This paper has shown the combination of classification and clustering techniques. |
| Arundhati Navada, Aamir Nizam Ansari, Siddharth Patil, Balwant A.Sonkamble | Overview of Use of Decision Tree algorithms in Machine Learning | IEEE | The ID3 algorithm is a very widely used algorithm, however it has many drawbacks. The variables depend on each other, this is not taken into consideration by the ID3 algorithm; this can worsen the overall classification performance of the decision tree. Also, the variables used can only be discrete. This algorithm cannot deal with noisy data sets. But all in all, the ID3 algorithm works decently well to make simple decision trees. An improvement in the ID3 algorithm, where a global dependency between variables is taken into consideration and a ‘ looks ahead’ procedure is adopted to select good attributes |

### 4.3 Summary

All the research papers so far discussed have been giving a detailed analysis of how churn has been changing over every year but in contrast we are doing detailed analysis of each and every attribute present in the dataset on which it has effect in changing of churn and also through over analysis and through our visualisation the company can understand which factors influence the customer to change the company network. We also provide the company in which sector they need improvement through our visualization techniques.

## 5. Proposed Work:

## 5.1 Proposed Architecture




## 5.2Implementation Details

### DATASET AND PACKAGES

We have imported all the packages and libraries we will be using for the exploration of data and model building. First the data using read csv function and the path for the location of the dataset csv file is given as argument. Exploration and visualization using seaborn and plotly packages in python.

### Telco-Customer-Churn Data Overview

● Each row represents a customer, each column contains customer's attributes described on the column Metadata
● The raw data contains 7043 rows (customers) and 21 columns (features)
● After cleaning the data for later analysis, the final data set contains 7043 rows (customers) and 41 columns (features).

● Data reports:
Customers who left within the last month: Churn
Customer Services: Phone Service, Multiple Lines, Internet Service, etc.
Customer Account information: Tenure, Contract, Payment Method, Charges, etc.
Customer Demographic Information: Gender, Partner, Dependents, etc.

**Data Exploration:**

In the step of Data Exploration, we have explored the dataset available with us. We have performed some basic data exploration steps including finding the number of rows and columns, no. of unique values of each column, and changed the datatypes of some columns and handled null values by replacing with value of 0.

**Data Visualization**

The visualizations of data are performed which starts with univariate analysis, analysing the data in perspective of a single attribute then with bivariate analysis and then with multivariate which deals with more than two attributes at the same time. Here the attribute's distributions are visualized using count plots, barplots, histograms, etc. Before performing the bivariate analysis, the values of both the dimensions are scaled in order for the visual plots to appear appropriately. The bivariate analysis is done using scatter plots, box plots, violin plots and so on. Similar plots are used in multivariate analysis but the third or more dimensions are represented on two dimensions by adding colors or size to the plot attributes. Now the data is split into train and test data to perform the model building, training and testing.

**Data Pre-processing:**

In data pre-processing section we will convert Categorical variables to numeric by finding binary columns and mapping their values with 0 and 1 and after that categorical attributes with number of levels in them are identified since the categorical variables cannot be directly trained in the model. Instead we create dummy variables to represent each level in a categorical variable sort of like one hot encoding to represent the category of a particular attribute

**Dataset Balancing:**

In some cases, the dataset used for training machine learning model is not balanced. The percentage of labels that need to be predicted is biased towards one value.

In our project we have used some data balancing techniques like

1. Random Over-Sampling
2. SMOTE
3. ADASYN

Each of these techniques uses different methodology to balance the class values of predictive variable.

**Random Over-Sampling**

Random Over-Sampling simply replicates randomly the minority classes. Samples from the training dataset are selected randomly and can be replaced. It can be interpreted that the minority samples can be taken from training set an can be added repeatedly to the balanced dataset ; they are selected from the original training dataset, added to the new training dataset, and then returned or "replaced" in the original dataset, allowing them to be selected again.

**SMOTE (Synthetic Minority Over-Sampling Technique)**

SMOTE synthesis new minority instances between existing minority instances. For each minority classes smote calculates K-Nearest Neighbours from a selected point and depending upon the resampling size it chooses a number n under k and creates n samples between selected point and n choosen points.

**Adasyn (Adaptive Synthetic)**

ADASYN works almost similar to Smote algorithm the difference in this algorithm is adding some random value to the selected point between chosen point and k-neighbour point that generates synthetic data, and its greatest techniques are not copying the same minority data, and generating more data for "harder to learn" examples.

## Feature Engineering:

Feature Engineering one of the most important part of Model Building. The features in your data will directly influence the models you use and the results you can achieve. If you have better features the better results we can achieve. Here we calculated correlation between given features and drawn a heatmap for the correlation values and removed some redundant features and we have used Random Forest Classifier to know how importance each feature is and also used Recursive Feature elimination technique to find optimal number of attributes that suits our dataset on the basis of cross validation score and finally choose the best number of features so that the model performance is best.

**Model Training**

Now that we know what our data looks good, now we use some machine learning models to predict whether the customer will churn or not based on given values of the other attributes. We will use sklearn library to import various classification algorithms to train test various classification models on our data and compare the results.

Different machine algorithms used:

1. Logistic Regression

2. Decision Tree

3. Random Forest

4. Bagging Classifier

5. Gradient Boosting Classifier

6. AdaBoost Classifier

7. XGboost Classifier

The model used to predict the classification variable churn of telecom is multivariate classification model and this machine learning model is considered since we need a model capable of handling more than two attributes and therefore multiple classification is used. Multiple Classification is performed using the dummy encoded variables and then trained. The Churn attribute of the data is classified based on the remaining numerical and categorical attributes to create a classification model.

## Pseudo Code for Base models

#### Logistic Regression

Logistic Regression is primarily used for classification problems. The most common examples would be classifying whether an email is spam or not spam, and in this case, classifying whether a customer may churn or no. Logistic Regression makes use of a sigmoid function to map the data into two categories, of 0 and 1 in terms of a binary classification. Logistic Regression makes use of Maximum Likelihood Regression to estimation of regression coefficients. The coefficients obtained are that set of coefficients for which the probability of getting the data we have observed is maximum

Logistic Regression was chosen here as it is one of the most common regression algorithms for binary classification in Supervised learning.

Logistic Regression has shown an accuracy of around 78.60sss percent, which was sent as the benchmark metric to beat. Please look at Data Pre-processing section to understand how the data was encoded for this particular task.

**Random Forest pseudocode:**

i)Randomly select "k" features from total "m" features (Where k << m )
ii)Among the "k" features, calculate the node "d" using the best split point.
iii) Split the node into daughter nodes using the best split.
iv) Repeat 1 to 3 steps until "l" number of nodes has been reached.
v) Build forest by repeating steps 1 to 4 for "n" number times to create "n" number of trees.

**Pseudocode for Decision tree:**

1. Sign all training instances to the root of the tree. Set curent node to root node.
2. For each attribute a. Partition all data instances at the node by the value of the attribute.
b. Compute the information gain ratio from the partitioning.

3. Identify feature that results in the greatest information gain ratio. Set this feature to be the splitting criterion at the current node. a. If the best information gain ratio is 0, tag the current node as a leaf and return.

4. Partition all instances according to attribute value of the best feature.
5. Denote each partition as a child node of the current node.
6. For each child node: a. If the child node is "pure" (has instances from only one class) tag it as a leaf and return.
b. If not set the child node as the current node and recurse to step 2.

**Hyperparameter Tuning:**

After building different machine learning models take the best performing model and tune the hyperparameters of that machine learning model. Here in this classification problem we have selected two best performing models and tuned their parameters

1. Logistic Regression

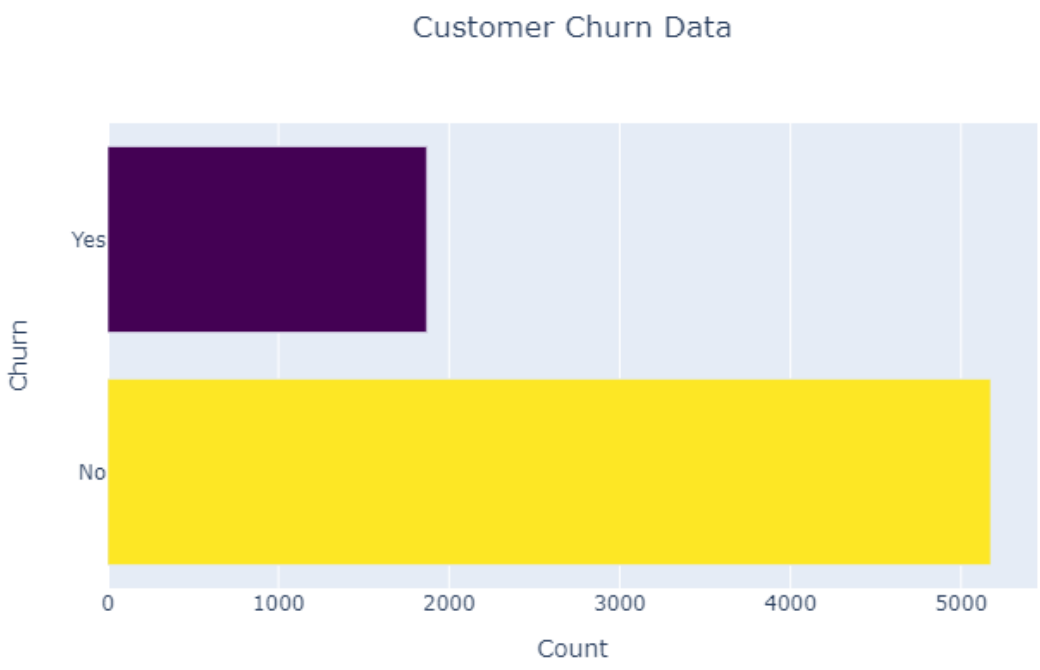
2. Adaboost Classifier

After tuning the parameters there is a slight improvement in the performance of logistic regression model and in the case of Adaboost classifier the base parameters are proven to be the best.

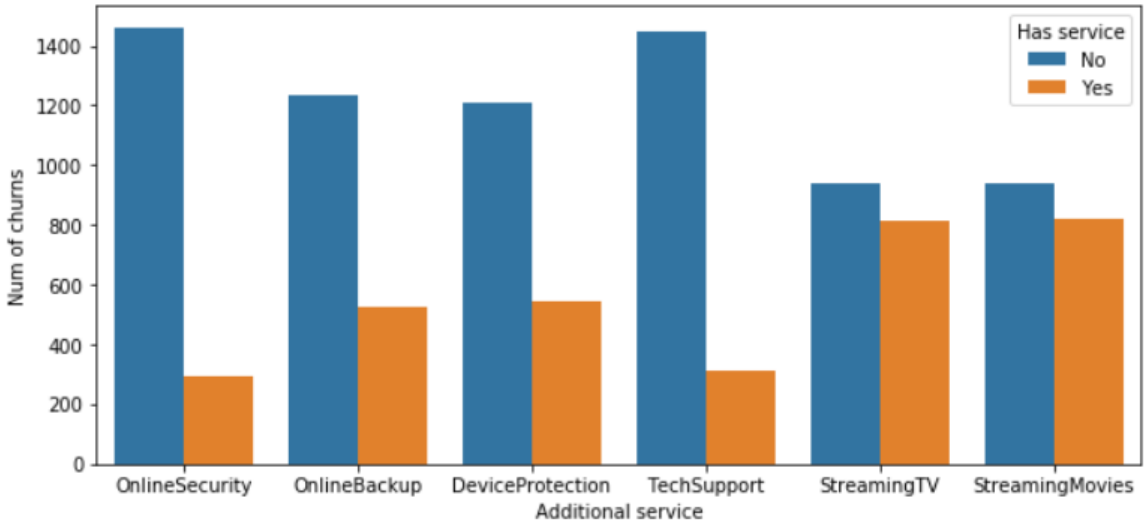
## 6. Experiment/ Results

## 6.1 Visualizations:

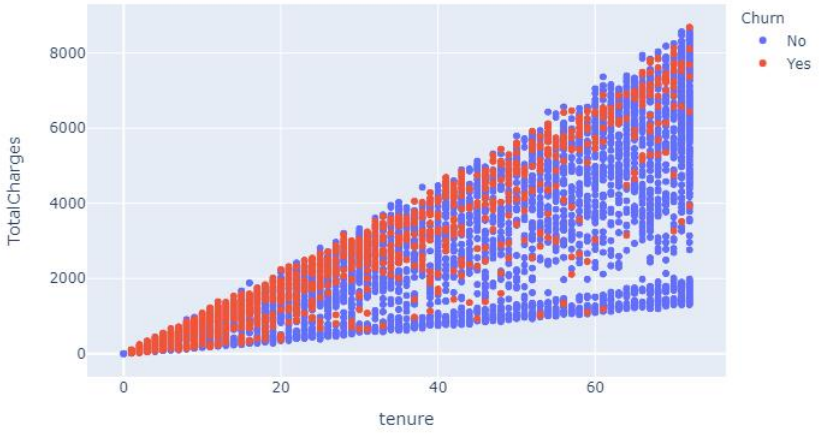
Churn classification attribute distribution

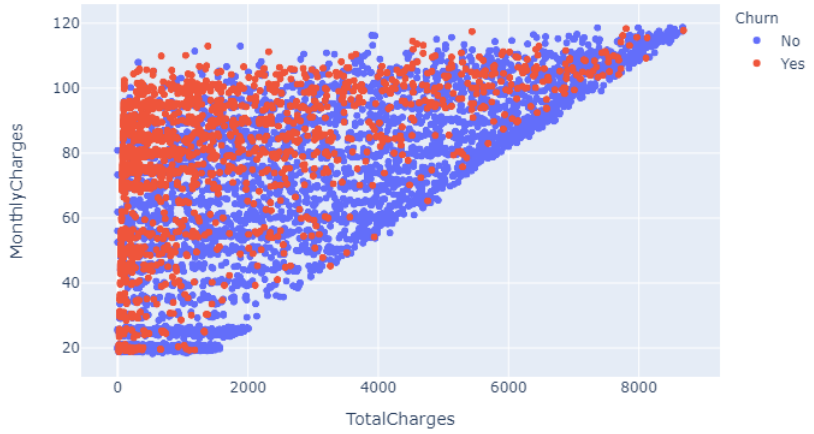


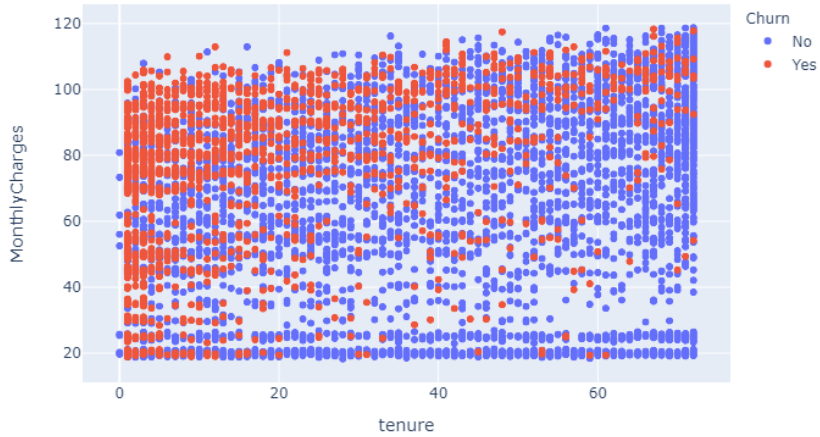

Comparing different additional services through bar graph



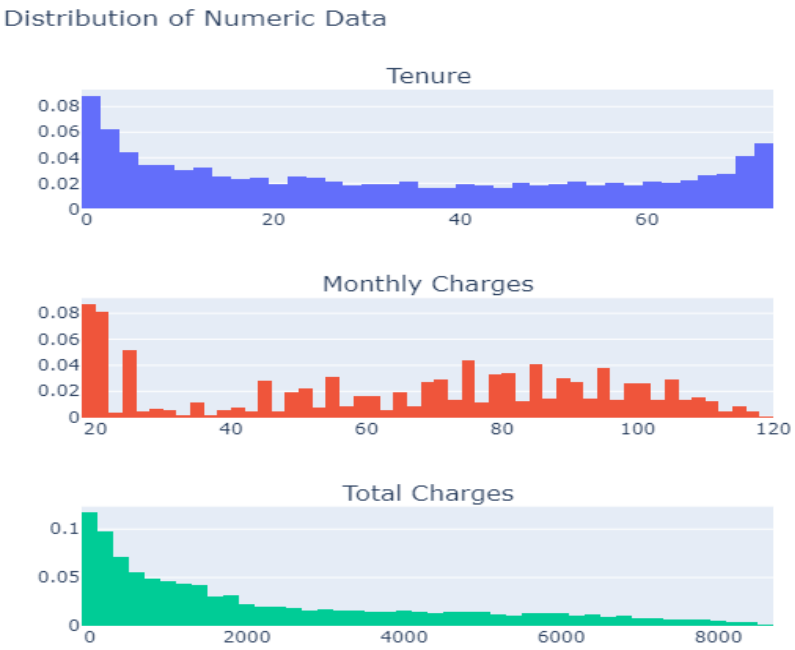

Scatter plot between various numeric variables


Churn
No
Yes
TotalCharges
8000
6000
4000
2000
0
0
20
40
60
tenure

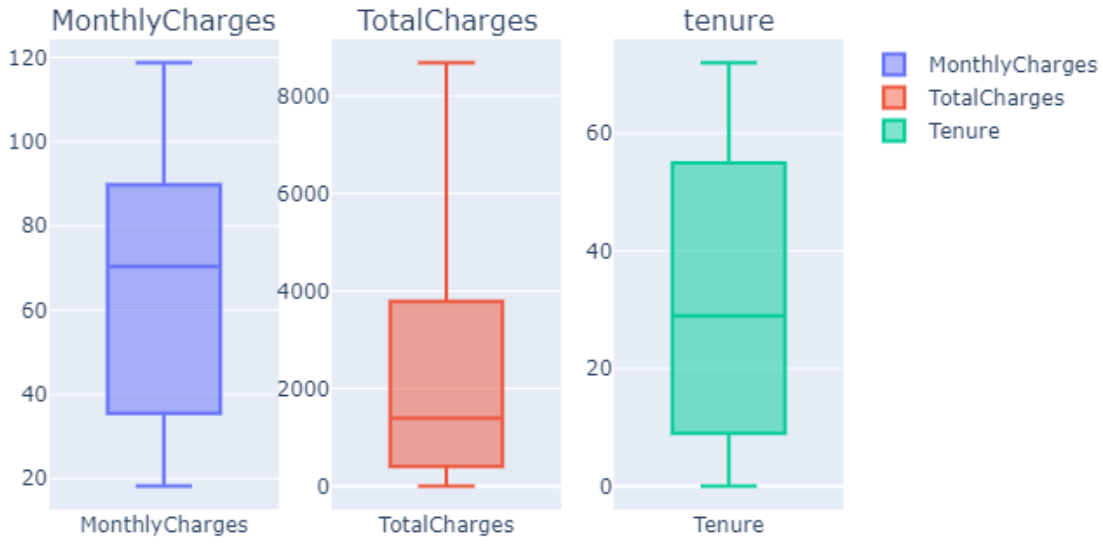

Churn
No
Yes
MonthlyCharges
120
100
80
60
40
20
0
2000
4000
6000
8000
TotalCharges


Churn
No
Yes
MonthlyCharges
120
100
80
60
40
20
0
20
40
60
tenure

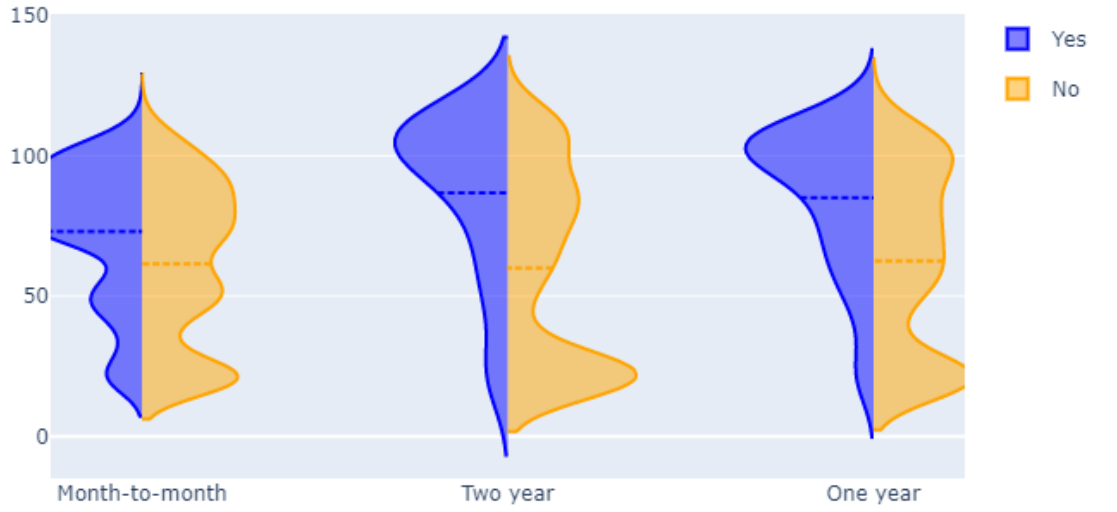
Distribution of all numeric attributes



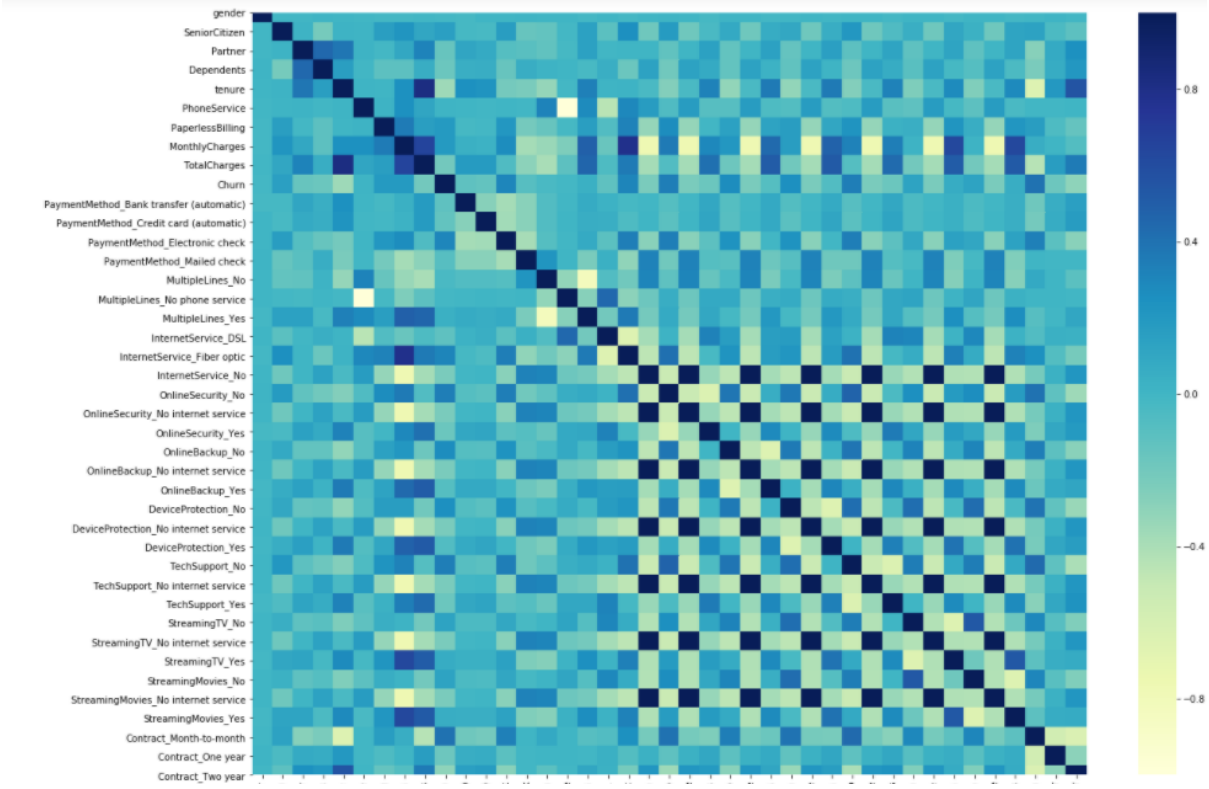

Median and Quartiles of Numeric Attributes




Violinplot between Monthly charges and Contract

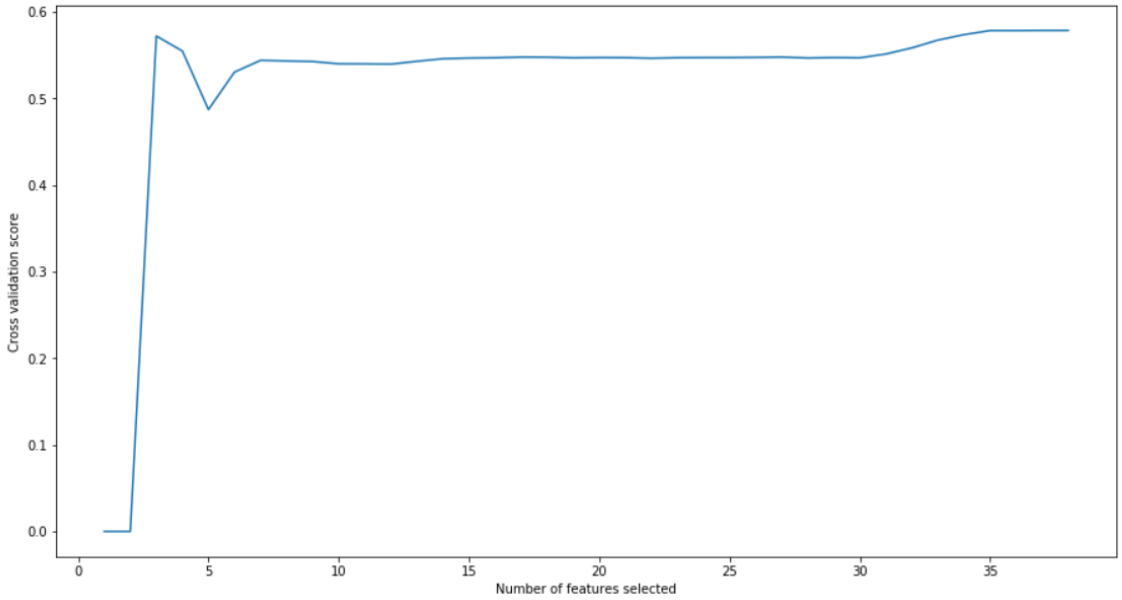



Correlation between attributes:

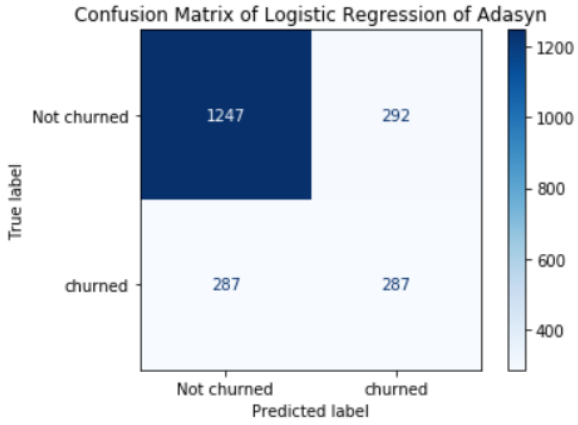


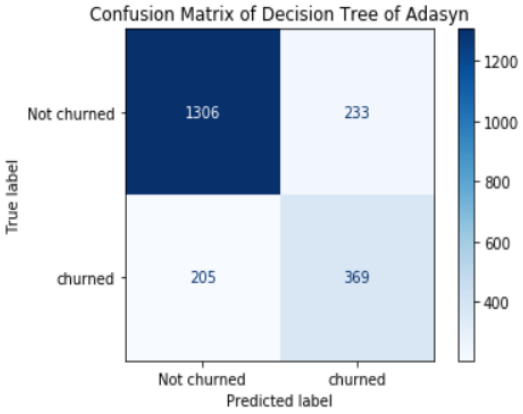

Feature Selection

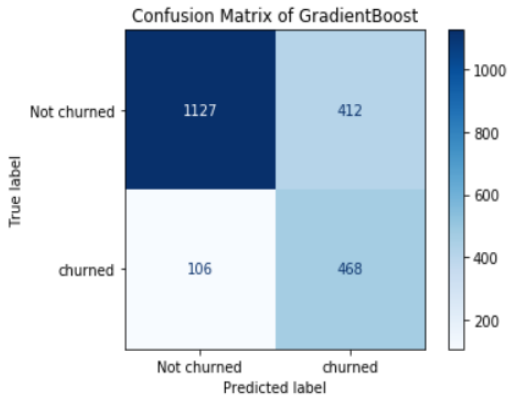
```
Optimal number of features : 38
```

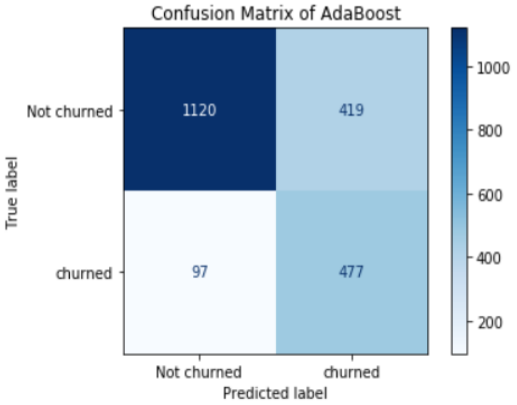



## 6.2 Machine Learning Algorithm Predictions:

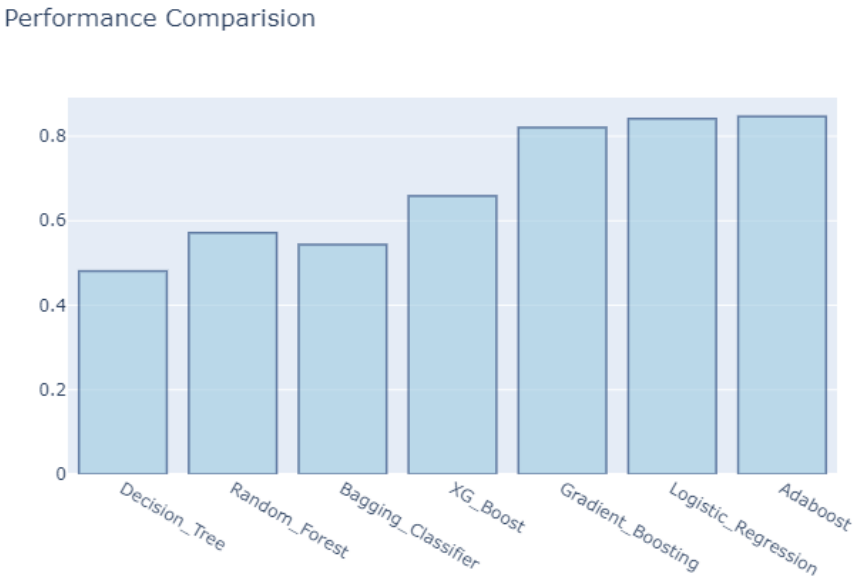


Comparison of Algorithms




## 7. Conclusion and Future Work

### 7.1 Conclusion

In order to increase our business as we have discussed in introduction there are 2 ways, one is getting new customers and retaining our old customers. The method of retaining customers always involves churn rate. Churn has become an important factor not only for telecom industries but many companies in other sectors.

In our case we tried to reduce the number of false positives i.e. recall which says the number of people who we predicted as not churned but the original outcome is churned. So, the best machine learning model according to our results are AdaBoost Classifier, after that Logistic Regression and Gradient Boost.

We also performed dataset balancing methods and the best we found is Adasyn which provided a better balance of data than other algorithms Random Oversampling and SMOTE.

Some of the insights we derived from the visualisations of data are:

- Senior citizens are more likely to churn than on-senior citizens.
- Customers with no extra services are more likely to churn.
- Customers having month-to-month contract are more likely to churn and having one year or two year are less likely to churn.
- payment method of electronic check is having most churn rate.
- Here both Two year and One Year Contract have Same trend but in all the three customers paying high monthly charges are most likely to churn

- Month to month contracts, no online security and no tech support are positively correlated with churn whereas tenure and two-year contracts are negatively correlated
- The redundant attributes in the dataset are Total Charges and No_Internet service
- Random oversampling technique is performing better than smote and adasyn techniques

### 7.2 FUTURE WORK

We will try to develop some more models for prediction of churn using Machine learning and deep learning techniques for more accuracy while predicting and also this work can be extended to also other datasets of different sectors.

## 8. References:


[1] Customer Churn Prediction for Telecom Services by Hakki Candan Cankaya , Turker Ince, IEEE 2014

[2] A Proposed Churn Prediction Model by Essam Shaaban, Yehia Helmy, Ayman Khedr, Mona Nasr.International Journal of Engineering Research and Applications (IJERA),2012
[3] Data Mining Applications in Customer Churn Management by Sahand KhakAbi, Mohammad R. Gholamian, Morteza Namwar, IEEE 2015

[4] A Comparative Study of Customer Churn Prediction in Telecom Industry Using Ensemble Based Classifiers by Abinash Mishra, U. Srinivasulu Reddy, IEEE 2017

[5] CHURN PREDICTION IN MOBILE TELECOM SYSTEM USING DATA MINING TECHNIQUES by M. BALASUBRAMANIAN, M. SELVARANI, International Journal of Scientific and Research Publications 2014

[6] A Review on Customer Churn Prediction in Telecommunication Using Data Mining Techniques by S. Babu, Dr. N. R. Ananthanarayanan, IJSER 2014

[7] Customer Churn Predictionin Telecommunication with Rotation Forest Method by Mumin Yuldiz, Songul Albayrak, IJSER 2015

[8] From Visual Data Exploration to Visual Data Mining: A Survey by Maria Cristina Ferreira de Oliveira and Haim Levkowitz, IEEE 2005

[9] An Enhanced K-Nearest Neighbor Algorithm Using Information Gain and Clustering by Shweta Taneja, Charu Gupta, Kratika Goyal, Dharna Gureja, IEEE 2014

[10] Overview of Use of Decision Tree algorithms in Machine Learning by Arundhati Navada, Aamir Nizam Ansari, Siddharth Patil, Balwant A. Sonkamble, IEEE 2014

[11] An Introduction to Logistic Regression Analysis and Reporting by CHAO-YING JOANNE PENG KUK LIDA LEE GARY M. INGERSOLL Indiana University-Bloomington IEEE 2010

[12] An Overview of Big Data Visualization Techniques in Data Mining by Samuel Soma Ajibade, Anthonia Adediran, IEEE 2011

## 9. APPENDIX A – CODING

```
#importing libraries
import warnings
warnings.filterwarnings('ignore')
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.figure_factory as ff
import plotly.offline as py
import plotly.graph_objects as go
from plotly.subplots import make_subplots

data = pd.read_csv(r'C:\Users\saiko\Desktop\Stuff\Telecom-Customer-Churn.csv')
data.head()
print("dataframe shape: ",data.shape)
null_values = data.isnull().sum().sort_values(ascending=False)
null_values = pd.DataFrame(null_values).reset_index()
null_values.columns = ['Column Name', 'No. of null values']
null_values.style.background_gradient(cmap='twilight_r')
numeric_ds.describe().T
objects_ds.describe().T
def count_plots(column):
```

```
    data_ch = churn[column].value_counts().reset_index()
    data_ch.columns = [column,'Count']
    data_nch = not_churn[column].value_counts().reset_index()
    data_nch.columns = [column,'Count']
    trace1 = go.Bar(x = data_ch[column]  , y = data_ch["Count"],
                    name = "Churn Customers",
                    marker = dict(line = dict(width = .5,color = "black")),
                    opacity = .9)

    trace2 = go.Bar(x = data_nch[column] , y = data_nch["Count"],
                    name = "Non Churn Customers",
                    marker = dict(line = dict(width = .5,color = "black")),
                    opacity = .9)
    layout = go.Layout(dict(title = "Customer Churn Across " + column,
                            title_x=0.5,
                            xaxis = dict(title = column),
                            yaxis = dict(title = "Count"),
                           )
                      )
    data = [trace1,trace2]
    fig  = go.Figure(data=data,layout=layout)
    py.iplot(fig)
df_sub = data[['tenure','MonthlyCharges','TotalCharges','Churn']]
fig1=px.scatter(df_sub,x="tenure",y="MonthlyCharges",color='Churn')
fig1.show()
fig = make_subplots(rows=1, cols=3,subplot_titles=("MonthlyCharges", "TotalCharges",
"tenure"),column_widths=[1, 1,1])
fig.add_trace(go.Box(y=data['MonthlyCharges'],name="MonthlyCharges"),row=1, col=1)
fig.add_trace(go.Box(y=data['TotalCharges'],name="TotalCharges"),row=1, col=2)
fig.add_trace(go.Box(y=data['tenure'],name="Tenure"),row=1, col=3)
fig.show()
```

```
def violin_plots(col1,col2):

    fig = go.Figure()

    fig.add_trace(go.Violin(x=data[col1][ data['Churn'] == 'Yes' ],
                            y=data[col2][ data['Churn'] == 'Yes' ],
                            legendgroup='Yes', scalegroup='Yes', name='Yes',
                            side='negative',
                            line_color='blue'))

    fig.add_trace(go.Violin(x=data[col1][ data['Churn'] == 'No' ],
                            y=data[col2][ data['Churn'] == 'No' ],
                            legendgroup='No', scalegroup='No', name='No',
                            side='positive',
                            line_color='orange'))

    fig.update_traces(meanline_visible=True)
    fig.update_layout(violingap=0, violinmode='overlay')
    fig.show()

data['gender'] = data['gender'].replace({'Female':0,'Male':1})
data['Partner'] = data['Partner'].replace({'Yes':1,'No':0})
data['Dependents'] = data['Dependents'].replace({'Yes':1,'No':0})
data['PhoneService'] = data['PhoneService'].replace({'Yes':1,'No':0})
data['PaperlessBilling'] = data['PaperlessBilling'].replace({'Yes':1,'No':0})
data['Churn'] = data['Churn'].replace({'Yes':1,'No':0})

category_cols=['PaymentMethod','MultipleLines','InternetService','OnlineSecurity','OnlineBackup','DeviceProtection','TechSupport','StreamingTV','StreamingMovies','Contract']
for cc in category_cols:
    dummies = pd.get_dummies(data[cc], drop_first=False)
```

```
    dummies = dummies.add_prefix("{}_".format(cc))
    data.drop(cc, axis=1, inplace=True)
    data = data.join(dummies)

corr=data.corr().iloc[:,:]
c1 = corr.abs().unstack()
c1.sort_values(ascending = False)

plt.subplots(figsize=(20,15))
sns.heatmap(corr,annot=False,cmap="YlGnBu")
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
lr = LogisticRegression()
model_lr = lr.fit(X_train,y_train)
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(random_state=0)
model_rf = rf.fit(X_train_new,y_train)
from sklearn.feature_selection import RFECV
from sklearn.model_selection import RepeatedStratifiedKFold
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=10, random_state=1)
rfe = RFECV(estimator = lr,cv=cv,scoring='f1')
rfe.fit(X_train_new, y_train_new)
print("Optimal number of features : %d" % rfe.n_features_)
plt.figure(figsize=(15,8))
plt.xlabel("Number of features selected")
plt.ylabel("Cross validation score")
plt.plot(range(1, len(rfe.grid_scores_) + 1), rfe.grid_scores_)
plt.show()
for i in range(X_train_new.shape[1]):
```

```
    print('Column: %d, Selected=%s, Rank: %d' % (i, rfe.support_[i], rfe.ranking_[i]))
from imblearn.over_sampling import RandomOverSampler
oversample = RandomOverSampler(sampling_strategy='minority')
X_over, y_over = oversample.fit_resample(X_train_new, y_train_new)
from imblearn.over_sampling import SMOTE
sm = SMOTE(random_state=0)
X_train_smote, y_train_smote = sm.fit_resample(X_train_new, y_train_new)
from imblearn.over_sampling import ADASYN
ad = ADASYN(random_state=0)
X_train_adasyn, y_train_adasyn = ad.fit_resample(X_train_new, y_train_new)
from sklearn.ensemble import BaggingClassifier
clf = BaggingClassifier(base_estimator = dt)
model_bagging = clf.fit(X_over,y_over)
pred_bagging_over = model_bagging.predict(X_test_new)
from sklearn.ensemble import GradientBoostingClassifier
clf_GB = GradientBoostingClassifier(random_state=36)
model_GB = clf_GB.fit(X_over,y_over)
pred_GB= model_GB.predict(X_test_new)
from sklearn.ensemble import AdaBoostClassifier
clf_AD = AdaBoostClassifier(random_state=36)
model_AD = clf_AD.fit(X_over,y_over)
pred_AD= model_AD.predict(X_test_new)
print("Accuracy Score:", accuracy_score(y_test_new,pred_AD))
print("Precision Score:", precision_score(y_test_new,pred_AD))
print("Recall Score:", recall_score(y_test_new,pred_AD))
solvers = ['newton-cg', 'lbfgs', 'liblinear']
penalty = ['l1','l2','elasticnet']
c_values = np.logspace(-4, 4, 20)
grid = dict(solver=solvers,penalty=penalty,C=c_values)
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=5, random_state=1)
grid_search = GridSearchCV(estimator=lr, param_grid=grid, n_jobs=-1, cv=cv,
scoring='f1',error_score=0,verbose=True)
grid_result_lr = grid_search.fit(X_over, y_over)
```

```
arr = []
names = ['Decision_Tree','Random_Forest','Bagging_Classifier','XG_Boost','Gradient_Boosting','Logistic_Regression','Adaboost']
dt_re = recall_score(y_test_new,pred_dt_random)
rf_re = recall_score(y_test_new,pred_rf_random)
log_re = recall_score(y_test_new,pred_lr_random)
bc_re = recall_score(y_test_new,pred_bagging_over)
xg_re = recall_score(y_test_new,pred_xgb)
gb_re = recall_score(y_test_new,pred_GB)
ad_re =  recall_score(y_test_new,pred_AD)
arr.append(dt_re)
arr.extend([rf_re,bc_re,xg_re,gb_re,log_re,ad_re])
fig = go.Figure([go.Bar(x=names, y=arr)])
fig.update_traces(marker_color='rgb(158,202,225)', marker_line_color='rgb(8,48,107)',
                  marker_line_width=1.5, opacity=0.6)
fig.update_layout(title_text='Performance Comparision')
fig.show()
```