A Mini-Project Report

on

"Prediction of Customer Churn"

Submitted to the

Pune Institute of Computer Technology, Pune

In partial fulfillment for the award of the Degree of

Bachelor of Engineering

in

Information Technology

by

AYUSHI PATANI 71829052G

VARUN GAWANDE 71828724L

JASH GUJARATHI 71828800K

YASH GUPTA 71828752F

Under the guidance of

Dr Emmanuel M.

Prof. R. Murumkar

Department Of Information Technology

Pune Institute of Computer Technology College of Engineering

Sr. No 27, Pune-Satara Road, Dhankawadi, Pune - 411 043.

2019-2020

CERTIFICATE

This is to certify that the project report entitled

Predicting customer churn

Submitted by:

AYUSHI PATANI 71829052G

VARUN GAWANDE 71828724L

JASH GUJARATHI 71828800K

YASH GUPTA 71828752F

is a bonafide work carried out by them under the supervision of Dr. Emmanuel and it is approved for the partial fulfillment of the requirement of Software Laboratory Course-2015 for the award of the Degree of Bachelor of Engineering (Information Technology)

Dr.Emmanuel
Internal Guide
Department of Information Technology

Dr. A.M.Bagade
Head of Department
Department of Information Technology

Dr.Emmanuel
Internal Guide
Date :

Place:
Date:

ACKNOWLEDGEMENT

We thank everyone who have helped and provided valuable suggestions for successfully creating a wonderful project.

We are very grateful to our guide, Dr. Emmanuel, Head of Department Dr. A. M. Bagade and our principal Dr. P. T. Kulkarni. They have been very supportive and have ensured that all facilities remained available for smooth progress of the project.

We would like to thank our Professor R. Murumkar Sir for providing very valuable and timely suggestions and help. We would also like the entire project staff team for providing valuable reviews and suggestions from time to time.

We would like to thank our entire department and college staff for the very valuable help and coordination throughout the duration of the project.

We would also like to thank our families and all our friends for the valuable support they provided throughout the duration of the project.

Ayushi Patani

Varun Gawande

Jash Gujarathi

Yash Gupta

Abstract

Customer attrition, also known as customer churn, customer turnover, or customer defection, is the loss of clients or customers. It is the percentage of customers that stopped using a company's product or service during a certain time frame. Customer churn is one of the most important metrics for a growing business to evaluate. While it's not the happiest measure, it's a number that can give a company the hard truth about its customer retention.

Taking into consideration the competitiveness between major giants in the TeleCom Industry, we believe Customer Churn is highly appropriate as the scope of a TeleCom company's Agenda. So we built a Customer Churn Prediction model that predicts whether a given customer has a high probability of churning, using, Random Forest Classifier, K Nearest Neighbor Classifier, Logistic Regression.

We have used a dataset containing customer-wise details that could help us understand what affects a churn. This includes location details like, area code, plan details, usage, charges, and Customer Service experiences. PreProcessing includes Label Encoding, Stripping of response values and redundant columns.

CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Motivation

Customer attrition, also known as customer churn, customer turnover, or customer defection, is the loss of clients or customers.Telephone service companies, Internet service providers, pay TV companies, insurance firms, and alarm monitoring services, often use customer attrition analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Customer loss is very closely related with customer loyalty. Today's economic trend dictates that price cuts are not the only way to build customer loyalty. Accordingly, adding new value added services to the products has become an industry norm to have loyal customers. The main goal of customer lost study is to figure out a customer who will likely be lost and is to calculate cost of obtaining those customers back again.Companies from these sectors often have customer service branches which attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients.

1.2 Purpose

Telecommunication sector provides services like rapidly growing local and intercity calls, and other communications services such as voice, fax, e-mail and other data traffic. Telecommunication market is rapidly developing and becoming competitive in many countries due to regulations, new computer and communication technologies. In this situation, data mining is required for understanding the business needs, defining the telecommunications model, using sources effectively and improving service quality. Telecommunication industry is now a mature market and aware of the importance of Customer Relationship Management. Because of having a mature market, developments in the following fields are achieved.

Cross Selling and up-selling: to maximize profits from existing customers.

Retaining and up-selling: to retain profitable customers or get rid of inappropriate customers to the company profile.

Poaching: to poach new Customers from rival companies.

Obtaining new customers is more expensive than retaining existing customers. It is for that, telecommunication companies realize that to keep existing customers is getting more important and agrees that churn analysis is one of the important data mining application areas.

1.3 Need

Mobile operators wish to retain their subscribers and satisfy their needs. Hence, they need to predict the possible churners and then utilize the limited resources to retain those customers.Churn Analysis is applied to research why customers switch service providers. In Churn analysis applications, the first thing is to access the customer data. Then, factors are classified to decide which factor or factors affect customer churn decision. After determining which customers are likely to churn, different and specific marketing and retention strategies can be applied to the target customers, in a defined time period. Churn Analysis is not just applied in marketing; it is also applied in customer service, sales and finance applications. These departments need to identify what the possible results of churn are, how much financial impact the company has, how sales and customer service are affected by churn.

Chapter 2

LITERATURE SURVEY

1.

Title: "A Churn Prediction Model Using Random Forest:Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector**"**
Author: Irfan Ullah, Basit Raza, Ahmad Kamran Malik, Muhamad Imran, Saif Ul Islam , Sung Won Kim

Methodology:

Proposed model for Customer Churn Prediction:
In the 1st step, data preprocessing is performed which includes data filtering for noise removal, removal of imbalanced data features and normalization of the data. Important features are extracted from data using information gain attributes ranking lter and correlation attributes ranking lter. In the second step, different classi- cation algorithms are applied for categorizing the customers into the churn and non-churn customers. The classification algorithms include Random Tree (RT), J48, Random Forest (RF), Decision Stump, AdaboostM1 C Decision Stump, Bagging C Random Tree, Naïve Bayes (NB), Multilayer Perceptron (MLP), Logistic Regression (LR), IBKandLWL. This step also identifies factors which are used in the next step for applying clustering algorithms. In the third step, customer profiling is performed using k-means clustering techniques. Cluster analysis is based on the patterns of customer transactional behavior captured from the data. In the nal step, the model recommends retention strategies for each category of churn customers.

Limitation and Future Scope:
The main limitation in carrying out this project is the limited dataset. The next logical step in the direction to improve the accuracy of the prediction problem at hand would be to test out the approaches and various methodologies proposed in this paper using a larger and more representative dataset. Also this study encourages to extend the candidate classifier set considered to a more exhaustive list and compare the performances among them.

2.

Title: "Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression**"**
Author: Preeti K. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, Prof. V. A. Kanade

Methodology:

Proposed model for Customer Churn Prediction:
The system will have three main options namely View performance analysis – which displays the results obtained by applying logistic regression and decision tree on the available dataset, Testing – to construct a list of customers which have a high probability to churn from the input, given that the attributes of the input data are same as the available dataset used for training, Training and testing – which builds a model along with generating a churn list if any other type of dataset is provided. In performance analysis the results after using logistic regression and decision trees on the available dataset is illustrated using confusion matrix analysis. In the next operation the user can provide data for testing the system provided the features of the data are the same as that used for training using the publicly available dataset.The system building has three main phases i.e. developing the web interface module, feature extraction module and prediction module. The web interface will provide a graphical overview of the results obtained, which can be created using the R package called Shiny. The feature extraction module will consist of estimation of parameters in logistic regression and generation of rules for decision trees. In decision tree algorithm the information gain is calculated for each feature and the maximum value feature is used to split the dataset, this is continued until all the features are used and then the features which do not provide enough information are pruned from the tree to get an optimized tree for the best possible estimate.

Conclusion and Future Scope:
The proposed system provides a statistical survival analysis tool to predict customer churn based on comparison between decision trees and logistic regression. Selecting the right combination of attributes and fixing the proper threshold values may produce more accurate results of predicting churn customers. The proposed model suggests that data mining techniques can be a promising solution for the customer churn management. Using this model the telecom companies can predict in advance which customers are at risk of leaving, and can target those customers consequently saving a lot of revenues namely the ones which are used for replacing the lost customers and also the ones that are wasted for retaining already loyal customers.

3.

Title: "Predicting Customer Churn Prediction in telecom sector using various Machine learning techniques**"**
Author: Abhishek Gaur, Ratnesh Dubey

Methodology:

Proposed model for Customer Churn Prediction:
1. **Dataset**:- A telecom dataset is taken for predicting churn which identifies trends in customer churn at a telecom company and the data which we take is in .csv format. The data given to us contains 7043 observations and 21 variables extracted from a dataset.
2. **Data Preparation**: Since the dataset acquired cannot be applied directly to the churn prediction models, we can name each attribute.
3. **Data Preprocessing**: Data preprocessing is the most important phase in prediction models as the data consists of ambiguities, errors, redundancy and transformation which
needs to be cleaned beforehand.
4. **Data Extraction**: The attributes are identified for the classifying process.
5. **Decision**: Based on data extraction and classification models we can take a decision whether the employee is churner or not.

Conclusion and Future Scope:
In order to retain existing customers, Telecom providers need to know the reasons for churn, which can be realized through the knowledge extracted from Telecom data. In this paper, we train four machine learning models which is Logistic Regression, SVM, Random Forest and Gradient boosted tree and we can say that Gradient boosting is best in among four models and the Logistic regression and Random forest is an average and SVM is underperforming between these models.

Chapter 3
Design Details and Implementation

3.1 Dataset Description

Dataset	Description
customer_chrun.csv	3333rows and 21 columns

customer_churn.csv

Column	Data Type	Description
state	int64	defines the state to which customer belongs
account length	int64	specifies the account length
area code	int 64	specifies the area code
phone number	object	specifies the phone number of the customer
international plan	int64	specifies the international plan of the customer
voicemail plan	int64	specifies the voicemail plan of the customer
number vmail messages	int64	specifies the number of vmail messages
total day minutes	float64	specifies the total day minutes
total day calls	int64	specifies the total day calls
total day charge	float64	specifies the total day charges
total eve minutes	float64	specifies the total eve minutes
total eve calls	int64	specifies the total eve calls
total eve charge	float64	specifies the total eve charges
total night minutes	float64	specifies the total night minutes

total night calls	int64	specifies the total night calls
total night charge	float64	specifies the total night charges
total intl minutes	float64	specifies the total intl minutes
total intl calls	int64	specifies the total intl calls
total intl charge	float64	specifies the total intl charges
customer service calls	int64	specifies the number of customer service calls
churn	bool	specifies if the customer will churn nor not

3.2 Dataset Preprocessing

Initially we just went through Label Encoding on the 'state', 'international plan', 'voice mail plan' column to encode the String Values of the feature. LabelEncoding is one of many techniques used for Representation of Raw Data into usable features. We know that Textual or String values cannot be used on a model. Hence we must convert it to numeric form somehow, this is where LabelEncoding helps us. LabelEncoding simply maps a unique numeric value for every distinct value in a given feature. For example, the State Feature contains many discrete String values like KH,OS,NJ,etc. LabelEncoding maps each one a unique integer, e.g KH is given 0, OS is given 1, and NJ is given 2, hence now 'state' is now a feature containing Integer Values.

We then further dropped the 'phone number' and 'churn' columns. The 'phone number' simply features unique values, while the 'churn' is already stored in Y as the output class. After building the Feature Matrix and Standardizing it's values,we know the Response values are not balanced. Hence, performed Stratified Cross Validation too.

Cross Validation is a technique used to counter overfitting of data while trying to maintain the most accurate hyperparameters. In its most basic form, we divide the dataset into K parts or folds, K-1 folds are used in training and the last one left is used for testing. All the above combinations are trained for multiple models and then a final one is evaluated. Stratified Cross Validation goes a step further, it speaks about how the folds/splits should have an even distribution so that the training is "experienced" and fair. There are ways like randomizing and even customized functions to accomplish the aforementioned distribution.

3.3 Data Model Selection

Having completed PreProcessing, we move forward with building and training a model, namely a Machine Learning Algorithm on our data. By nature of our target variable, we must construct a binary classifier.

We chose three classifiers for the problem:

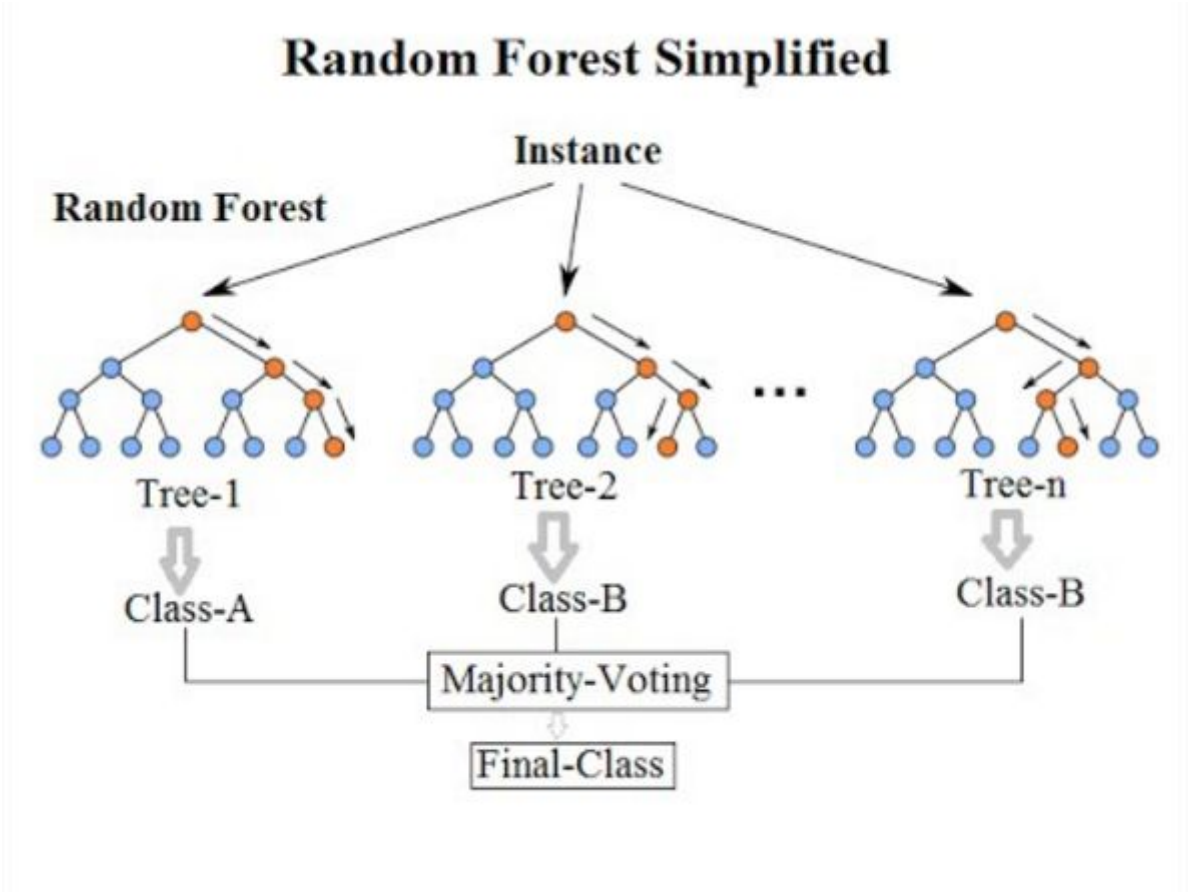
1. Random Forest Classifier
2. K Nearest Neighbour Classifier
3. Logistic Regression .

1. Random Forest Classifier:

Random Forests are an instance of using Ensemble Learning. Ensemble Learning is a way of using multiple learning algorithms to obtain better predictive results than just a single model predicting results.

Random Forests work in the following steps:

- Divide the Training Set into multiple parts
- Construct Classification Trees on all the multiple parts.
- Test our testing data on all the trees.
- Pick the mode out of the resulting classifications. That is basically, the outputs of the trees will be put to vote and the highest occuring one will be picked.



2. K Nearest Neighbour Classifier:

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

Algorithm:

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

Hamming Distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

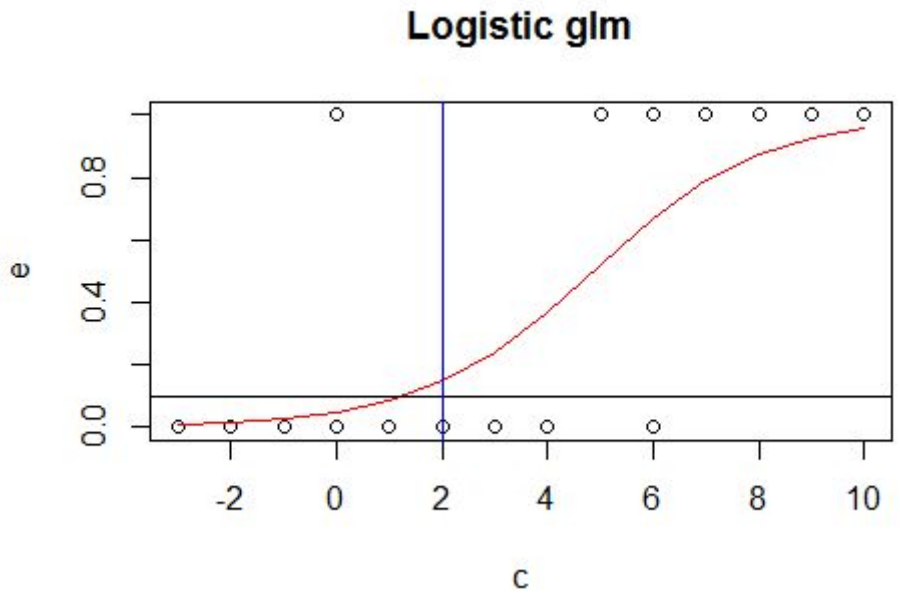
$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN.

3. Logistic Regression:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, **logistic regression** is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"), the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a *logit*, from *logistic units*, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.



Regression Analysis

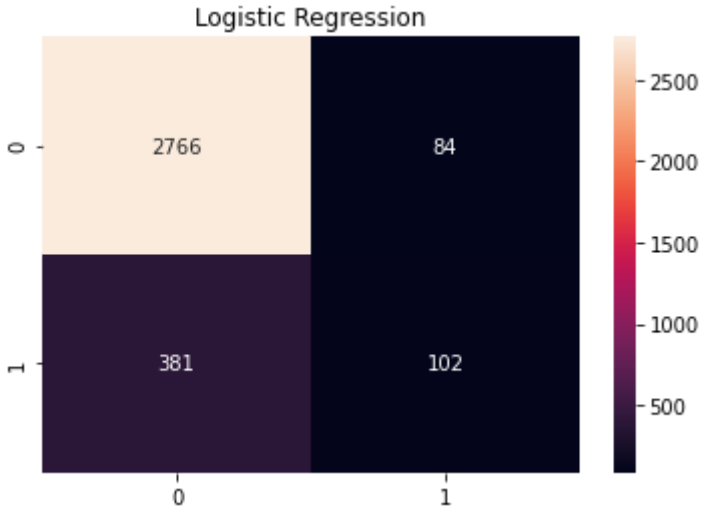
Prediction

The predictions from 3 models were found to be as follows:

1. Logistic Regression:

	Precision	Recall	F1-Score	Support
False	0.88	0.97	0.92	2850
True	0.55	0.21	0.30	483

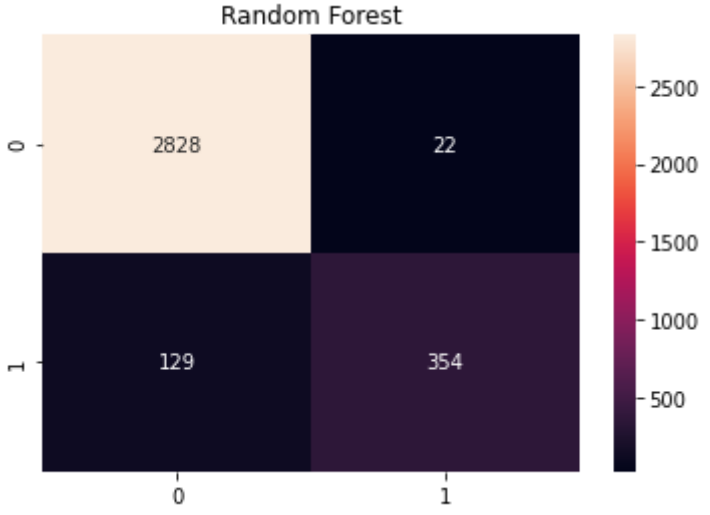
Confusion matrix


Logistic Regression
2766
84
381
102
0
1
0
1
2500
2000
1500
1000
500

2. Random Forest Classifier:

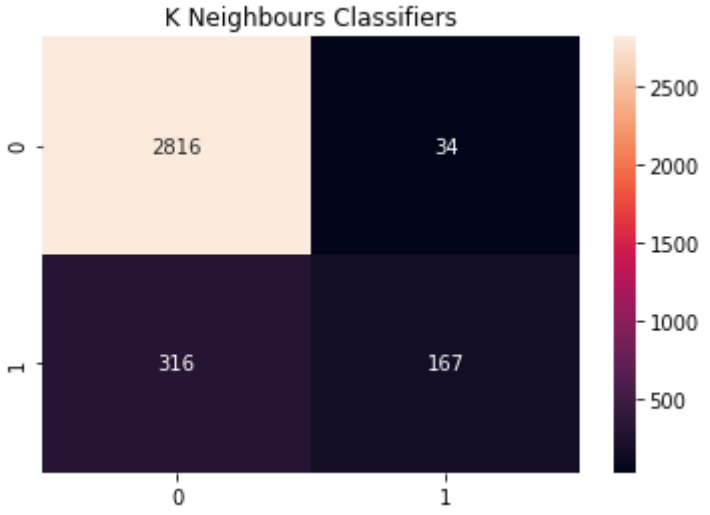
	Precision	Recall	F1-Score	Support
False	0.96	0.99	0.97	2850
True	0.94	0.73	0.83	483

Confusion matrix:



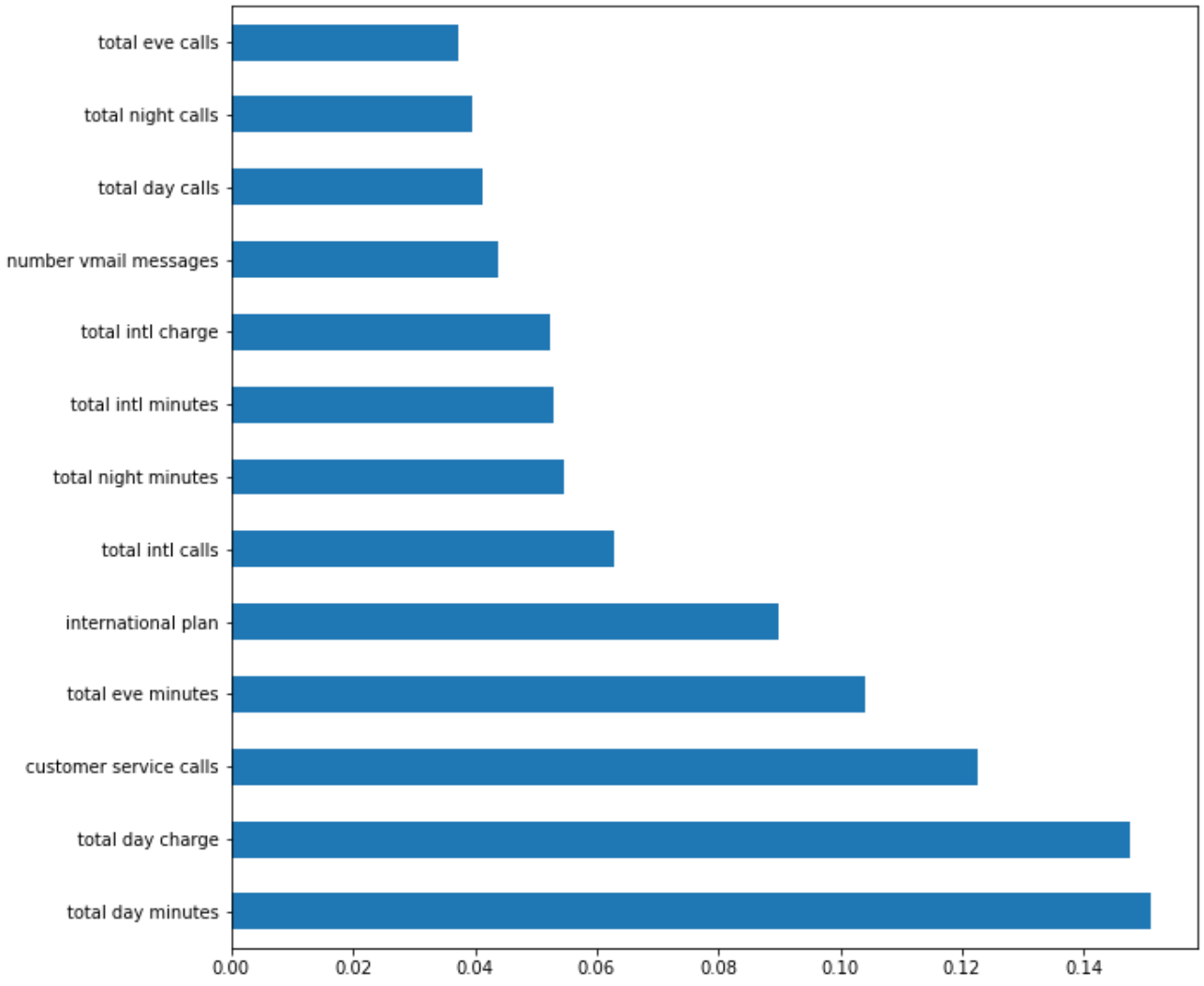
3. K Nearest Neighbor Classifier:

	Precision	Recall	F1-Score	Support
False	0.90	0.99	0.94	2850
True	0.83	0.35	0.49	483

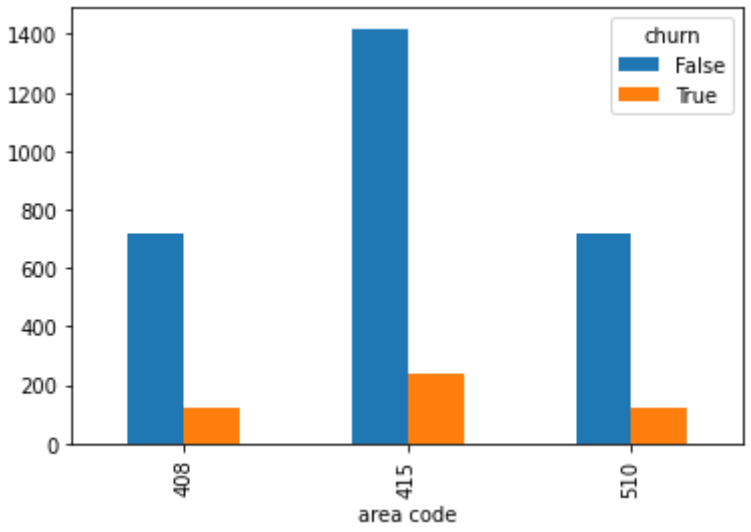


Visualization

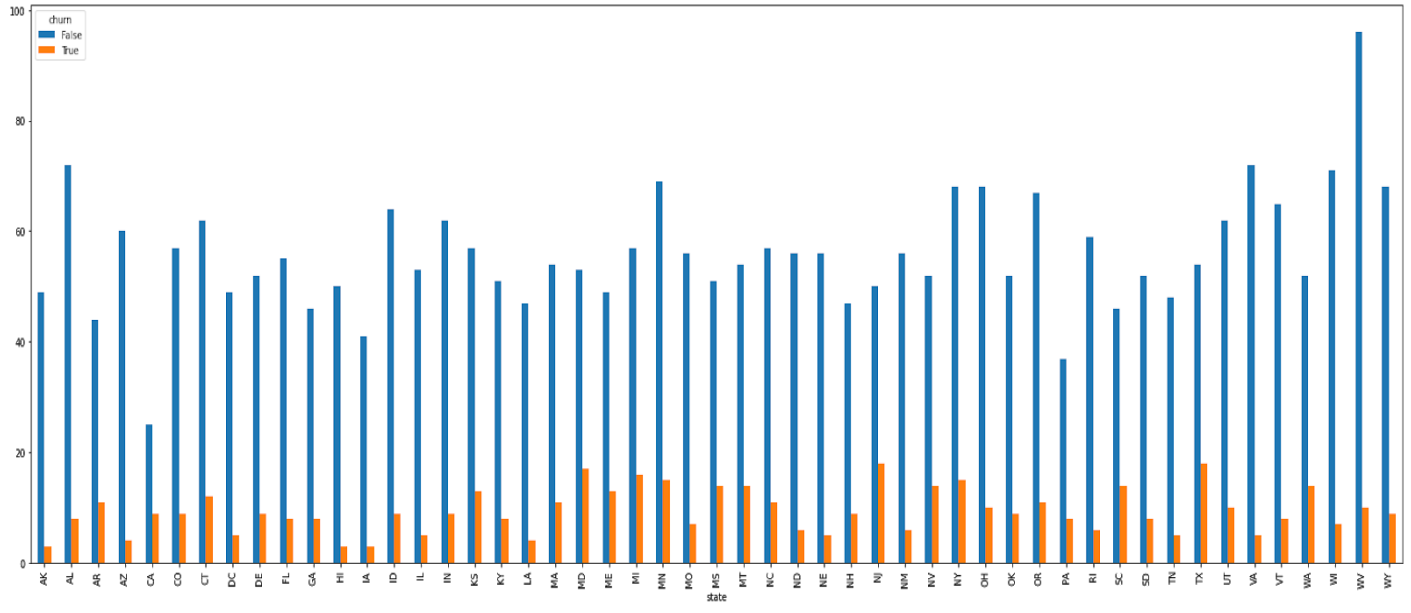
1.Feature importance graph



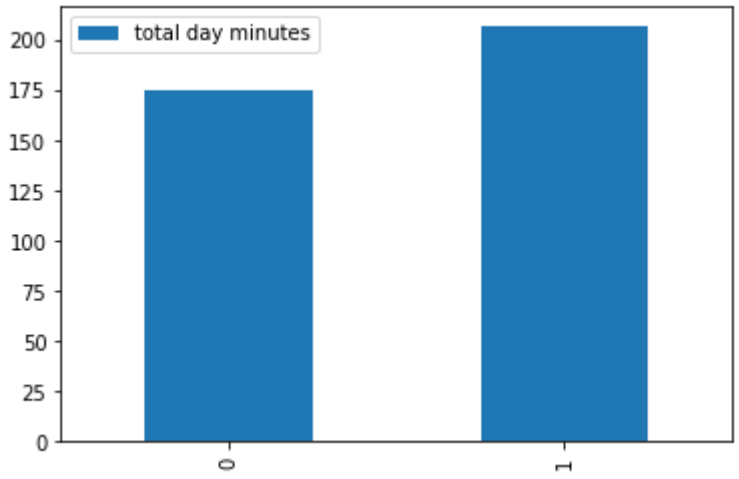
2.Area wise churn



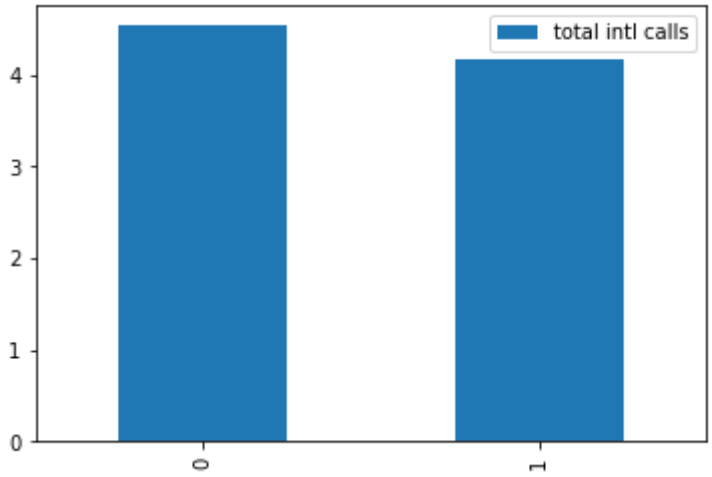
3.State wise churn



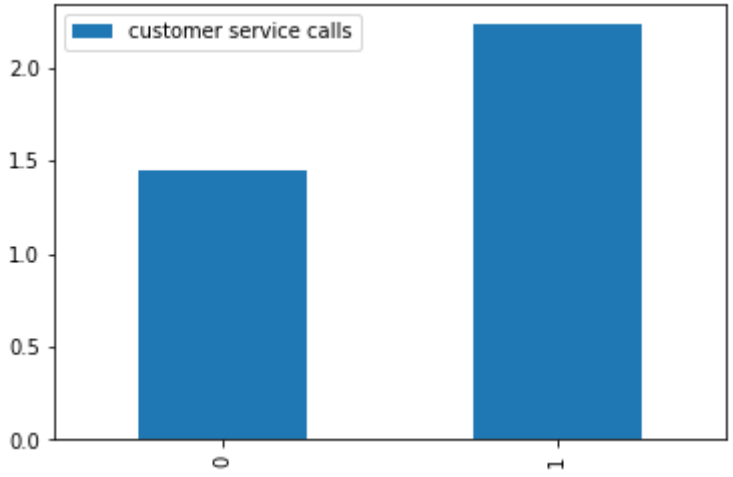
4.Total day min vs churn



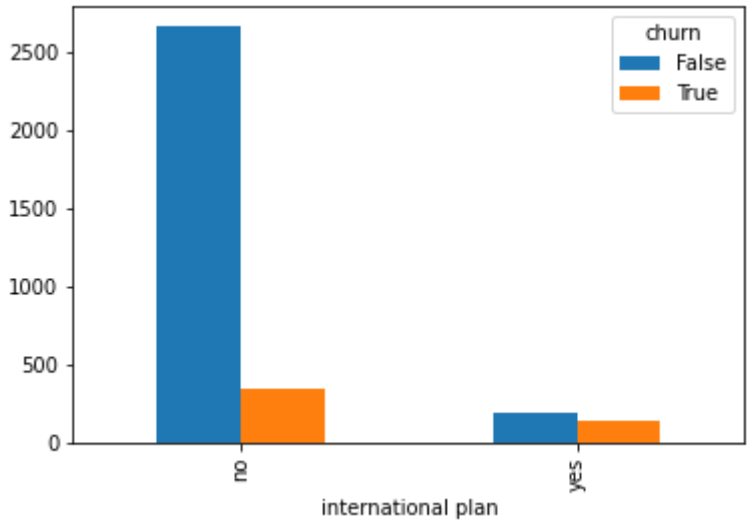
5.Total international calls vs churn



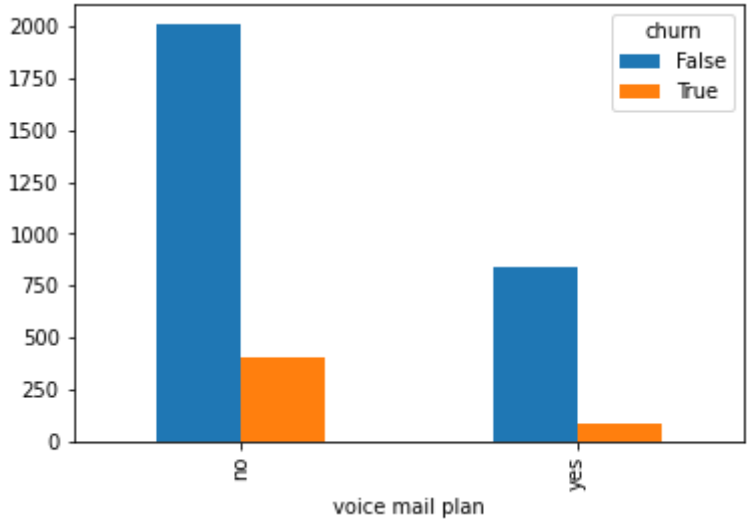
6.Total customer service calls vs churn



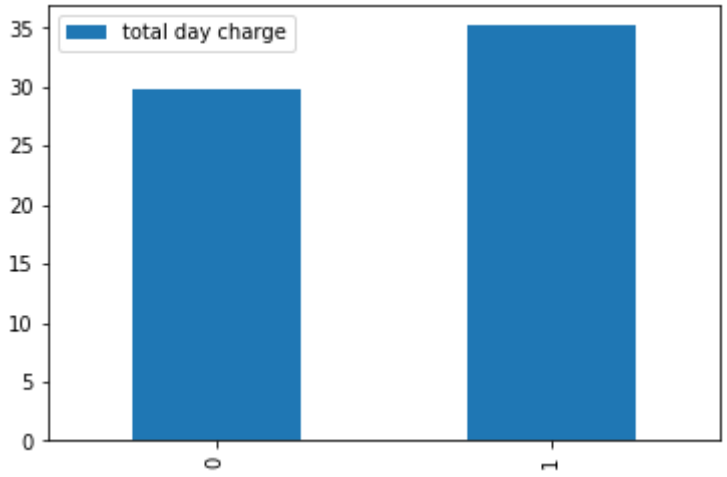
7.International plan vs churn



8.Voice mail plan vs churn



9.Total day charge vs churn



CHAPTER 4
OPTIMIZATION AND EVALUATION

We have applied optimization techniques in the model building and feature selection phase for boosting the accuracy of the model while also mitigating various challenges that plague classification models such as imbalanced training data, over-fitting , high dimensionality and lack of standardization.
The techniques employed are as follows:

1. **Stratified Cross Validation :**

   Cross-validation is a statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k − 1 folds. It is a proven technique that mitigates the over fitting problem of classification models , something that is very detrimental to imbalanced datasets.

2. **Recursive Feature Elimination :**

   Datasets used to train classification and regression algorithms are high dimensional in nature — this means that they contain many features or attributes. In textual datasets each feature is a word and as you can imagine the vocabulary used in the dataset can be very large. Not all features however, contribute to the prediction variable. Removing features of low importance can improve accuracy, and reduce both model complexity and overfitting. Training time can also be reduced for very large datasets. In this blog post performing Recursive Feature Elimination (RFE) with Scikit Learn will be covered.

   Recursive Feature Elimination (RFE) as its title suggests recursively removes features, builds a model using the remaining attributes and calculates model accuracy. RFE is able to work out the combination of attributes that contribute to the prediction on the target variable (or class). Scikit Learn does most of the heavy lifting just import RFE from sklearn.feature_selection and pass any classifier model to the RFE() method with the number of features to select. Using familiar Scikit Learn syntax, the .fit() method must then be called.

CHAPTER 5

RESULTS

1.Results from built model

Model	Precision
Random Forest classifier	0.95
K Nearest Neighbour	0.90
Logistic Regression	0.86

2. Results from visualization:

Deducing from the visualisations generated, we understand that:

- Features like ‘state’, ‘area code’ and ‘account length’ affect Target Variable the least and hence can be dropped from prediction.
- The feature of ‘Total day minutes’ is the most important feature while ‘Total evening calls’ is the least important feature.
- By considering only 13 features we can get the maximum precision.
- Area with code 415 has the greatest churn value with 1419 customers who did not churn and 236 who churned. And in comparison we see that it has a very similar probability distribution to other Area Codes, and hence does not differ much and is dropped.
- We can draw a similar conclusion behind State-wise Churn and hence drop it.
- The features with some of the highest weightages turn out to be:
    - total day minutes
    - total day charge
    - customer service calls
    - total eve minutes
    - international plan

CHAPTER 6
LIBRARIES AND SOFTWARE USED

Libraries used:

1. numpy
2. pandas
3. matplotlib.pyplot
4. sklearn.preprocessing.LabelEncoder
5. sklearn.ensemble.RandomForestClassifier
6. sklearn.tree
7. sklearn.model_selection.train_test_split
8. sklearn.metrics.accuracy_score
9. sklearn.model_selection.GridSearchCV
10. xgboost
11. seaborn
12. plotly

Software used:

1. Python 3.7.4
2. Jupyter Notebook 6.0.1
3. Kaggle Notebook
4. Browser: Mozilla Firefox Quantum 49.0

CHAPTER 7

CONCLUSION AND FUTURE WORKS

In our project, historical sample data has been collected for the Telephone service company, and useful features have been extracted after preprocessing of data. Deciding an ideal set of attributes encourages company owners to search for potential churners and then utilize the limited resources to retain those customers, which is important in this competitive environment. Detail state by state information is gathered and collected for the analysis and the problem was modeled as a classification problem. Data Visualization was carried out for Area wise churn, State wise churn, Voicemail plan vs churn, Total day charge vs churn, Total night charge vs churn, etc, which help us to predict the potential churners so that telecommunication owners can retain by providing certain services.

The proposed system provides a statistical survival analysis tool to predict customer churn based on, selecting the right combination of attributes and fixing the proper threshold values may produce more accurate results of predicting churn customers. The proposed model suggests that Random Forest Classifier techniques can be a promising solution for the customer churn management. Using this model the telecom companies can predict in advance which customers are at risk of leaving, and can target those customers consequently saving a lot of revenues namely the ones which are used for replacing the lost customers and also the ones that are wasted for retaining already loyal customers.