# ANALYSIS AND PREDICTION OF CUSTOMER CHURN IN TELECOM INDUSTRIES

**A PROJECT REPORT**

***Submitted by***

**SHIVAM SHARMA [RA1511003010367]**
**SRI SURYA VAMSI SARIPUDI [RA1511003010373]**

*Under the guidance of*
**Mrs.S.Ushasukhanya**
(Assistant Professor(O.G), Department of Computer Sciene and Engineering)

***in partial fulfillment for the award of the degree***

***of***

**BACHELOR OF TECHNOLOGY**

in

**COMPUTER SCIENCE AND ENGINEERING**

of

**FACULTY OF ENGINEERING AND TECHNOLOGY**

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
*Deemed to be University u/s 3 of UGC Act, 1956*

S.R.M. Nagar, Kattankulathur, Kancheepuram District

**MAY 2019**

# SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section 3 of UGC Act, 1956)

## BONAFIDE CERTIFICATE

Certified that this project report titled “**ANALYSIS AND PREDICTION OF CUSTOMER CHURN IN TELECOM INDUSTRIES** ” is the bonafide work of “**SHIVAM SHARMA [RA1511003010367], SRI SURYA VAMSI SARIPUDI [RA1511003010373]**”, who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

**SIGNATURE**

Mrs.S.Ushasukhanya
**GUIDE**
Assistant Professor(O.G)
Dept. of Computer Sciene and Engineering

**SIGNATURE**

Dr. B. Amutha.
**HEAD OF THE DEPARTMENT**
Dept. of Computer Science and engineering

Signature of the Internal Examiner

Signature of the External Examiner

# ABSTRACT

Churn Examination is one of the widespread used study on Subscription Oriented Businesses for analyzing the behavior and activities of customers in order to predict beforehand which customer is likely to exit the service agreement. Built on Machine Learning procedures and algorithms it has become very significant for companies in today ’ s market as securing of another client is more costlier than their maintenance. This paper focuses on the relevant studies on Customer Churn in Telecommunication industries to show the overall information about the frequently used data mining means, and the performance report of the methods used in our work. Initially, we obtained the telecom dataset of clients form kaggle website and which contains various customer details for analysing customer behaviour. Then, we compare our models used systems and show their performances and results. Conclusively, we review the different performance metrics that we have used for analysis . Examining all these three viewpoints is very critical for developing a more well-organized churn prediction model for telecom businesses and industries.

# ACKNOWLEDGEMENT

We express our humble gratitude to **Dr. Sandeep Sancheti**, Vice Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.We extend our sincere thanks to **Dr. C. Muthamizhchelvan**, Director, Faculty of Engineering and Technology, SRM Institute of Science and Technology, for his invaluable support.

We wish to thank **Dr. B. Amutha**, Professor  Head, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for her valuable suggestions and encouragement throughout the period of the project work.

We are extremely grateful to our Academic Advisor **Dr. K. Annapurani**, Associate Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for her great support at all the stages of project work.

We would like to convey our thanks to our Panel Head, **Dr.C.Malathy**, Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for her inputs during the project reviews.

We register our immeasurable thanks to our Faculty Advisor,**Mrs.R.Radha**,Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to our guide,**Mrs.S.Ushasukhanya**,Assistant Professor(O.G), Department of Computer Science and Engineering, SRM Institute of Science and Technology, for providing me an opportunity to pursue my project under her mentorship. She provided me the freedom and support to explore the research topics of my interest. Her passion for solving the real problems and making a difference in the world has always been inspiring.

We sincerely thank staff and students of the Computer Science and Engineering Department, SRM Institute of Science and Technology, for their help during my research. Finally, we would like to thank my parents, our family members and our friends for their unconditional love, constant support and encouragement.

**shivam sharma (RA1511003010367)**
**sri surya vamsi saripudi (RA1511003010373)**

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABBREVIATIONS

EDA–exploratory data analysis

CRM– customer relationship management

LRM– logistic regression model

SVM– support vector machines

PCA– principal component analysis

mRMR– minumum-Redundancy-Maximum-Relevance

# CHAPTER 1

# INTRODUCTION

## 1.1 Churn Prediction

customer Client information have been gathered all through the useful methodology of the cell phone organization, it is very significant for an aggressive organization to gather compelling data in over the top information assets and afterward[7] to make a joined data stage, yet it appears to be difficult to manage the helpful information utilizing the customary technique for database oversee ment,in telecom showcase rivalry coming into harsh challenge now and again, a few household media transmission company initiate to utilize a few frameworks to take care of the issue. Customer beat is a critical apparatus amid establishment and framework set up prescient model dependent on client lead. With here and there cruel challenge from the broadcast communications showcase, the media communications organization begins utilizing a few frameworks to take care of the issue. [8]Client agitate is a significant apparatus for setting up a prescient model dependent on client conduct amid the establishment and framework. In this paper we propose extrapolative models utilizing AI to foresee whether the clients in Communications/telecom firm will stir or not. We star represent the AI models with various calculations, for example, Naive Bayes , RF etc. Forecast execution of every calculation is evaluated utilizing precision grid, More testing is to set a model for Telecommunication division as there are no agreements between a client and Telecommunication concerning the length of offices/benefits .The telecom business perseveres through rising esteeming load all around. Concentrates to be done on client agitates is progressively basic for the Telecom organizations these days. Arrangement issue as characterization task goes under administered learning in AI where the fundamental objective is to build up models which are regulated by an outside specialist where the classifiers to preparing tests are realized well ahead of time. The made models distinguish the class marks of a concealed example utilizing this strategy. Feature The

Selection-Feature determination is the way toward recognizing and choosing the significant highlights . The chose highlights are extricated utilizing different element extraction strategies . a component vector is made which speaks to the arrangement of highlights with the end goal that each element vector is mapped to a class name. This aides in distinguishing class mark . Progressively over just pertinent highlights are removed as any superfluous element adds more to computational expense and irregular mistakes.

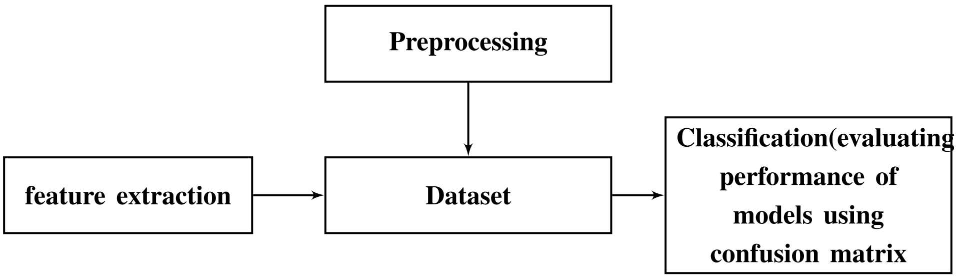

Preprocessing
feature extraction
Dataset
Classification(evaluating performance of models using confusion matrix

**Figure 1.1:** System Overview

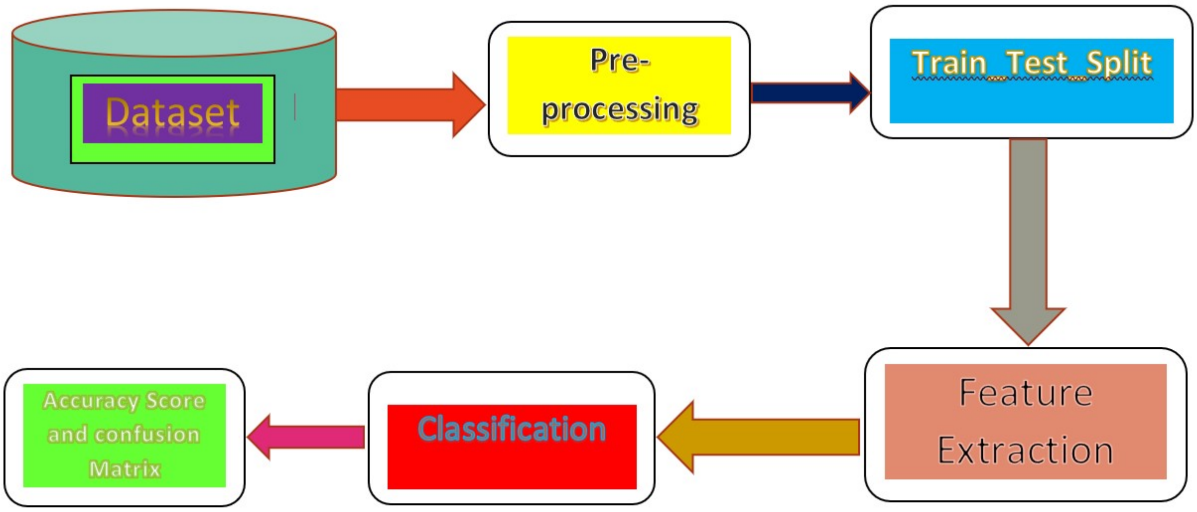


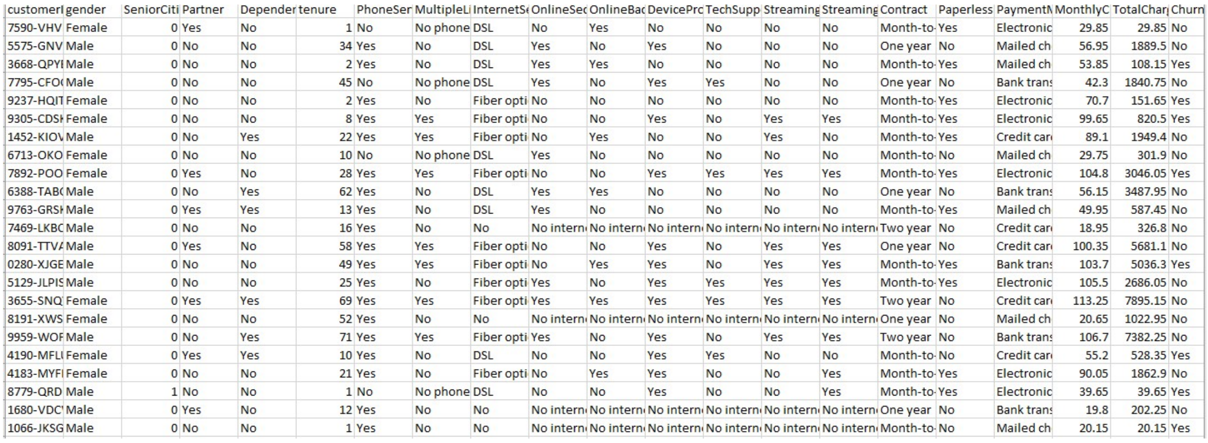
**Figure 1.2:** System Architecture

It makes a difference a great deal since associations frequently need to invest immense measure of energy and cash drawing in new clients/customers, there is a noteworthy speculation lost everytime a customer leaves/stops administration. Since Both time also work is expected to discover substitution .[9] Having the capacity to anticipate when a client is going to leave proves to be useful. The primary test here is that in a true application, a lot of information is available which is for the most part in crude structure i.e it may be excessively messy/confused to work with and consequently it is required that ought to attempt a considerably more thorough procedure for assessing our models for foreseeing precise results.the next stage would comprises of cleaning the information, highlight choice, demonstrating, and so forth.[10]Anticipating Customer Churn forecast is the procedure for distinguishing which clients are probably going to drop a membership to a specific administration or associations, for example, telecom,banking segments and so on dependent on their cooperation with the specific sort of administration .There are just two classes to which a client has a place, will agitate or not stir; henceforth it is a twofold grouping task.This paper centers around the important investigations on Customer Churn in Telecommunication businesses[12] to demonstrate the general data about the often utilized information mining implies, and the execution report of the strategies utilized in our work. At first, we got the telecom dataset of customers structure kaggle site and which contains different client subtleties for dissecting client conduct. At that point, we look at our models utilized frameworks and demonstrate their exhibitions and results. Indisputably, we survey the diverse execution measurements that we have utilized for investigation . Looking at all these three perspectives is extremely basic for building up an all the more efficient beat forecast model for telecom organizations and ventures.

### 1.1.1 Modules

#### Dataset collection

In customer churn prediction problems huge amounts of data is available thus it is important to select only the most significant one eliminating and avoiding as much noisy data as possible (outliers).as it affects model performance and can lead to incorrect predictions.Preproccessing step helps in filtering out outliers. The data set is obtained from Kaggle Website, is used in this paper for the churn analysis and prediction. This data set comprises 18 attributes and 7044 records or tuples.



| customerI | gender | SeniorCiti | Partner | Depender | tenure | PhoneSer | MultipleLi | InternetS | OnlineSec | OnlineBac | DevicePrc | TechSupp | Streaming | Streaming | Contract | Paperless | PaymentN | MonthlyC | TotalChar | Churn |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7590-VHV | Female | 0 | Yes | No | 1 | No | No phone | DSL | No | Yes | No | No | No | No | Month-to- | Yes | Electronic | 29.85 | 29.85 | No |
| 5575-GNV | Male | 0 | No | No | 34 | Yes | No | DSL | Yes | No | Yes | No | No | No | One year | No | Mailed ch | 56.95 | 1889.5 | No |
| 3668-QPYI | Male | 0 | No | No | 2 | Yes | No | DSL | Yes | Yes | No | No | No | No | Month-to- | Yes | Mailed ch | 53.85 | 108.15 | Yes |
| 7795-CFO( | Male | 0 | No | No | 45 | No | No phone | DSL | Yes | No | Yes | Yes | No | No | One year | No | Bank trans | 42.3 | 1840.75 | No |
| 9237-HQIT | Female | 0 | No | No | 2 | Yes | No | Fiber opti | No | No | No | No | No | No | Month-to- | Yes | Electronic | 70.7 | 151.65 | Yes |
| 9305-CDSI | Female | 0 | No | No | 8 | Yes | Yes | Fiber opti | No | No | Yes | No | Yes | Yes | Month-to- | Yes | Electronic | 99.65 | 820.5 | Yes |
| 1452-KIOV | Male | 0 | No | Yes | 22 | Yes | Yes | Fiber opti | No | Yes | No | No | Yes | No | Month-to- | Yes | Credit car | 89.1 | 1949.4 | No |
| 6713-OKO | Female | 0 | No | No | 10 | No | No phone | DSL | Yes | No | No | No | No | No | Month-to- | No | Mailed ch | 29.75 | 301.9 | No |
| 7892-POO | Female | 0 | Yes | No | 28 | Yes | Yes | Fiber opti | No | No | Yes | Yes | Yes | Yes | Month-to- | Yes | Electronic | 104.8 | 3046.05 | Yes |
| 6388-TAB( | Male | 0 | No | Yes | 62 | Yes | No | DSL | Yes | Yes | No | No | No | No | One year | No | Bank trans | 56.15 | 3487.95 | No |
| 9763-GRSI | Male | 0 | Yes | Yes | 13 | Yes | No | DSL | Yes | No | No | No | No | No | Month-to- | Yes | Mailed ch | 49.95 | 587.45 | No |
| 7469-LKBC | Male | 0 | No | No | 16 | Yes | No | No | No intern | No intern | No intern | No intern | No intern | No intern | Two year | No | Credit car | 18.95 | 326.8 | No |
| 8091-TTVA | Male | 0 | Yes | No | 58 | Yes | Yes | Fiber opti | No | No | Yes | No | Yes | Yes | One year | No | Credit car | 100.35 | 5681.1 | No |
| 0280-XJGE | Male | 0 | No | No | 49 | Yes | Yes | Fiber opti | No | Yes | Yes | No | Yes | Yes | Month-to- | Yes | Bank trans | 103.7 | 5036.3 | Yes |
| 5129-JLPIS | Male | 0 | No | No | 25 | Yes | No | Fiber opti | Yes | No | Yes | Yes | Yes | Yes | Month-to- | Yes | Electronic | 105.5 | 2686.05 | No |
| 3655-SNQ | Female | 0 | Yes | Yes | 69 | Yes | Yes | Fiber opti | Yes | Yes | Yes | Yes | Yes | Yes | Two year | No | Credit car | 113.25 | 7895.15 | No |
| 8191-XWS | Female | 0 | No | No | 52 | Yes | No | No | No intern | No intern | No intern | No intern | No intern | No intern | One year | No | Mailed ch | 20.65 | 1022.95 | No |
| 9959-WOF | Male | 0 | No | Yes | 71 | Yes | Yes | Fiber opti | Yes | No | Yes | No | Yes | Yes | Two year | No | Bank trans | 106.7 | 7382.25 | No |
| 4190-MFLU | Female | 0 | Yes | Yes | 10 | Yes | No | DSL | No | No | Yes | Yes | No | No | Month-to- | No | Credit car | 55.2 | 528.35 | Yes |
| 4183-MYFI | Female | 0 | No | No | 21 | Yes | No | Fiber opti | No | Yes | Yes | No | No | Yes | Month-to- | Yes | Electronic | 90.05 | 1862.9 | No |
| 8779-QRD | Male | 1 | No | No | 1 | No | No phone | DSL | No | No | Yes | No | No | Yes | Month-to- | Yes | Electronic | 39.65 | 39.65 | Yes |
| 1680-VDC' | Male | 0 | Yes | No | 12 | Yes | No | No | No intern | No intern | No intern | No intern | No intern | No intern | One year | No | Bank trans | 19.8 | 202.25 | No |
| 1066-JKSG | Male | 0 | No | No | 1 | Yes | No | No | No intern | No intern | No intern | No intern | No intern | No intern | Month-to- | No | Mailed ch | 20.15 | 20.15 | Yes |

**Figure 1.3:** Dataset for analysis

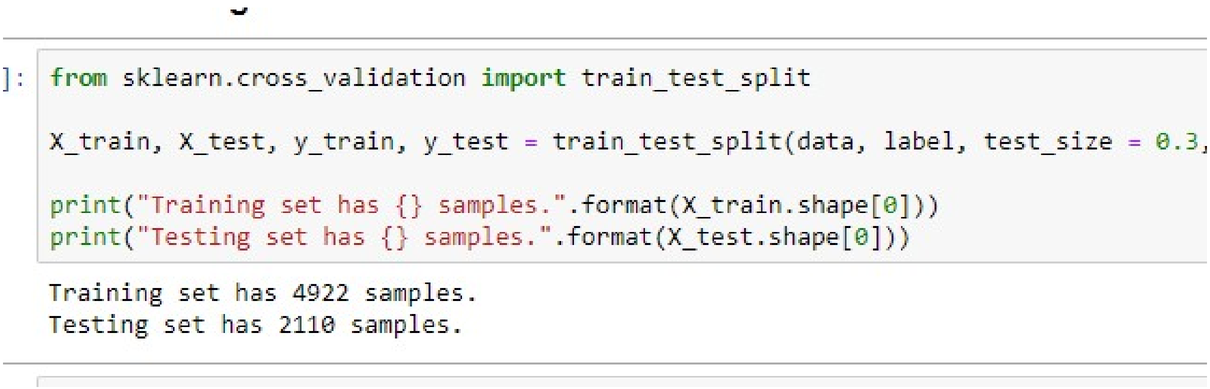
### Preprocessing phase

This phase is an very important step which prepares the dataset for the modelling part.It significantly improves performance metrics and boosts prediction accuracy.It consists of detecting outliers.there can be different kinds of outliers ,for instance,a customer might cancel his subscription to a particular service provider due to some reasons which are completely out of hand or not in control of the service provider company.[13]moving abroad or death can be some of the reasons.So it is best if these tuples are removed from the dataset as they will introduce noise in the modelling phase leading to incorrect predictions. Missing values are also treated in this phase.It is better to replace missing fields with any sort of measures of central tendency.

After detecting outliers and treatment of missing values feature selection is done.once dataset is collected it is difficult to tell which feature is relevant and which one is not.Some of the feature selection techniques most commonly used are principal component analysis(PCA) and minumum-Redundancy-Maximum-Relevance(mRMR)

### Train/Test split

The dataset is used for analysis is generally divided into two parts ,training part and test part.The training set consists of known output value which is used by the model to learn form this data inorder to classify unknown data later on. [14]The test dataset is used to test the efficiency and performance of our model.In Python this is done using the scikit-learn library which consists of traintestsplit method. This method is imported from the library to split the dataset into two parts.Before importing this method some important libraries must be imported first.Which are as follows: Panda library is use to load the dataset file into a pandas data frame for working on data, Matplotlib pyplot,seaborn can be imported for data vizualization.It is used for plotting graphs of the data, The test-size=0.3 tells that 30 percent of data is used for testing the model and the remaining 80 percent is for training the model.



```
]: from sklearn.cross_validation import train_test_split

   X_train, X_test, y_train, y_test = train_test_split(data, label, test_size = 0.3,

   print("Training set has {} samples.".format(X_train.shape[0]))
   print("Testing set has {} samples.".format(X_test.shape[0]))

   Training set has 4922 samples.
   Testing set has 2110 samples.
```

**Figure 1.4:** importing train-test-split method

# CHAPTER 2

# LITERATURE SURVEY

## 2.1 Evaluation of machine learning models for employee churn prediction

[1] This study presents a model for foreseeing worker agitate in an association. Representatives are a significant piece of association further, enlisting another worker turns out costly for any association and in this way holding current workers is the ideal arrangement. Direct help vector machine,c.5 choice tree arbitrary woods ,k-closest neighbor and naive bayes classifiers are utilized for characterization. This examination requires further investigation to limit the forecast rate.

## 2.2 Churn prediction model for effective gym customer retention

[2] This paper fabricates model to foresee client conduct in wellness businesses it is discovered that a yearly exercise center participation enables the customer to end their enrollment with barely any progressed notice.model dependent on strategic relapse, choice trees and neural systems is manufactured. This examination thinks that its elusive false positive rate for leavers at specific circumstances. Besides it expect that clients that have sporadic exchanges i.e unpredictable rec center participation are likewise named stirred clients.

## 2.3 Building comprehensible customer churn prediction model.

[3]A comprehensible customer churn prediction model is built for analyzing client behavior for determining which customer is likely to churn in the future. In order to prevent A model based on multiple kernel vector support machine approach is built .the drawback of MK-SVM is that while feature selection they can reduce some of the relevant features .This study has left the application of this framework to financial etc institutions as a future work.

## 2.4 Analysis of customer churn prediction in telecom industry

[4]This paper focuses on analyzing customer churn in telecom industry using logistic regression, neural network and decision trees, although neural networks perform very well for classification and prediction tasks it only does so for very large datasets. Moreover it takes the same amount of time for processing much smaller data sets.they left this work for future studies for making this model to handle large datasets.

## 2.5 Telecommunication subscribers' churn prediction model using machine learning

[5] This paper uses the technique of decision trees to develop a model for telecommunication subscribers churn prediction. This study canâĂŹt work on diverse data, which is considered as a drawback. .Its future work is test the approach on bigger data sets containing data over a longer period of time.

## 2.6 Customer Churn Prediction in Telecommunication Industry: With and without Counter

[6] This work made use of four different rule generation algorithms (i.e. Exhaustive, genetic ,covering and LEM2 to predict customer churn in telecom industry using the above techniques. The main focus of the problem being which classification technique could use to tactic the churn prediction in a more suitable and sufficient manner with more accuracy ,remains an open exploration problem. At the same time, the black-box model generated by SVM is also considered as one of its main drawback.

# CHAPTER 3

## SYSTEM ANALYSIS AND DESIGN

Studies shows that acquiring new customers is about 5 to 10 times expensive than retaining their existing customers and moreover keeping the customers loyal in today ' s competitive conditions has become priority for any organizations , according to reports an average business loses around 25-35 percent of their customers every year. Many companies, realizing this situation,are strongly focused in satisfying and retaining their customers in order to prevent churn. Particularly in the subscription oriented businesses, such as telecommunications, banking sectors, insurance companies, and in general in any particular field where customer relationship management(CRM) is crucial for the organization .The revenue generated and overall profits of the companies are provided by the payments/investments made by the customers periodically. Therefore the need of hour is to be able to keep customers gratified in order to be able to sustain this profits and revenue with the least expenses and minimize loss. Disadvantage:[15]In today ' s technological conditions, large volume of data is being produced from different sources in various sector .It is very important the data extractedfor large chunks of data repositories is pre-processed properly because the useful information hidden in these datasets can ' t be put into use, unless they are processed properly. In order to find out this hidden information and features, data science comes in handy for information extraction using several data mining methods and machine learning algorithms. Advantage:reviewing the relevant studies about customer churn analysis observed in the telecom industry. Predicting Customer churn is a business scenario in which a company is trying to retain a customer which is more likely to leave the services. For reducing churn rate, we have to classify which customers are most probably going to churn and which will not. Also we have some data to train our model which makes our problem as Supervised Classification problem. EDA It includes looking into the data analyzing various variables, visualization, missing value analysis, correlation analysis, chi-square test, scaling of features, Sampling. Basic Modeling Will try different machine models over pre-processed data ( Random forest,SVM,linear regression,Logistic regression). Model Evaluation

Optimization Evaluating model performances, select the best model fit for our data, optimizing hyper parameters tuning, Cost effectiveness of model. Implementation model on Final test data and to visualize the result.

## 3.1 Model Training and Performance Report

Factors considered in measuring performance are accuracy , sensitivity, specificity and precision .accuracy in the sense how good the classifier is performing sensitivity means how perfect the classifier is with respect to positive entries. There are 2 kinds tuples ,positive and negative tuples. positive tuples obey some specific rules whereas negative tuples do not . factors taken into consideration true positive(TP) tuples under consideration is perfectly positively classified i.e expected and observed tuple were positive. True negative(TN):if tuple was observed negative and expectation was same False positive(FP): data tuples were mistakenly classified as positive however expected outcome was the opposite . False negative(FN) were actually supposed to be classified as positive but were shown to be negative. We can calculate the performance measures of the created classifiers by using the following equations representing the relationship b/w the tuples wrt to the performance measuring factors.

## 3.2 Performance Metrics

### 3.2.1 Accuracy

It is the ratio of the correctly classified tuples to the entire collection of them. Accuracy answers the following question: How many customers did we correctly label out of all the customers. Accuracy = (TP+TN)/(TP+FP+FN+TN)

### 3.2.2 Precision

Precision is defined the ratio of the correctly +ve tuples classified by the model to all +ve entries. Precision answers the following: How many of those classified as churners will actually churn. Precision = TP/(TP+FP)

### 3.2.3 Recall

Recall is the ratio of the correctly +ve classified by our model to all customers that will churn . Recall answers the following question: Of all the customers who will churn, how many of those were correctly predicted. Recall = TP/(TP+FN)

### 3.2.4 F-1 Score

F1 Score takes into account both precision and recall metrics. It is calculated as the harmonic mean(average) of the precision and recall. F1 Score is best in case of imbalances between precision and recall. F1 Score = 2*(Recall * Precision) / (Recall + Precision)

### 3.2.5 Specificity

Specificity is the correctly identified -ve tuples classified by the model to all customers will not churn. Specifity answers the following question: Out of all the customers which are not going to churn, how many of those were correctly predicted. Specificity = TN/(TN+FP)

|  | positive(actual) | negative(actual) |
|---|---|---|
| positive(predicted) | true positive(TP) | false positive(FP) |
| negative(predicted) | false negative(FN) | true negative(TN) |

Table 3.1: Confusion Matrix

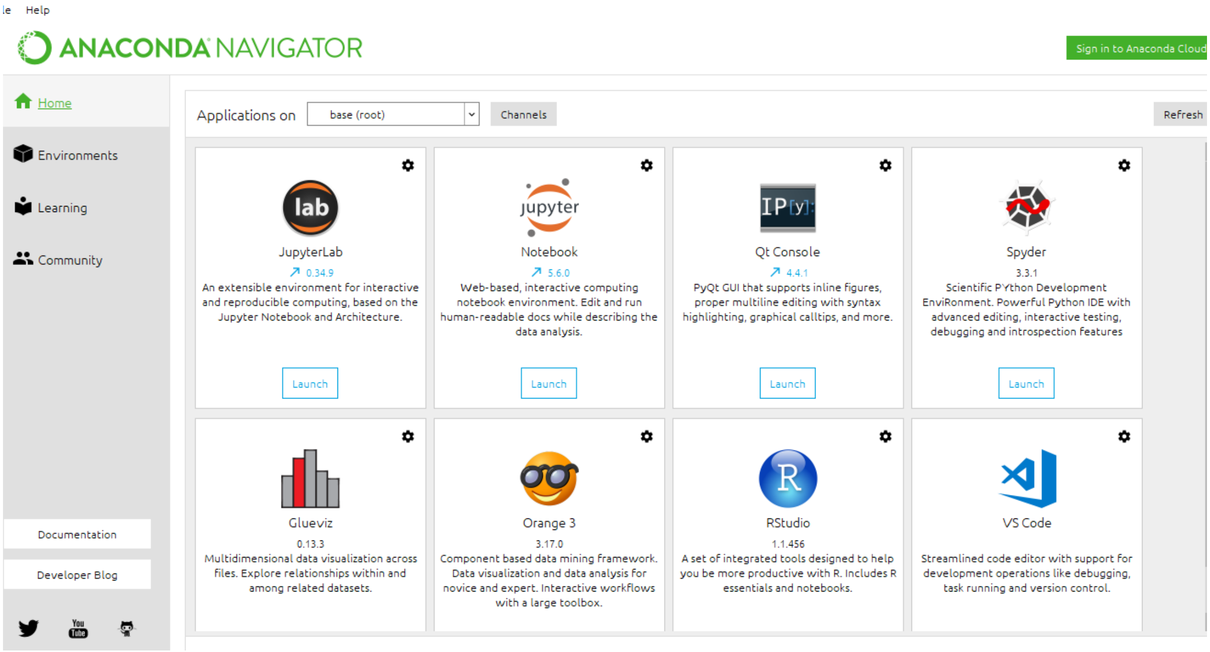
Confusion matrix is simple tool that is used to check whether tuples belonging to a class are perfectly classified.It is used to describe the perfromance of the model using a set of test data for which the true labels are known.

# CHAPTER 4

# SOFTWARE REQUIREMENTS

### 4.0.1 Anaconda Navigator

Anaconda Navigator is a desktop graphical user interface (GUI) included in AnacondaÂő distribution that allows you to launch applications and easily manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, macOS and Linux



An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Multidimensional data visualization across files. Explore relationships within and among related datasets.

Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

Streamlined code editor with support for development operations like debugging, task running and version control.

**Figure 4.1:** anaconda navigator

### 4.0.2 Jupyter notebook

The Jupyter Notebook is an open-source web application used for creating and sharing documents containing code,visualizations ,texts etc. It is used for data cleaning, transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

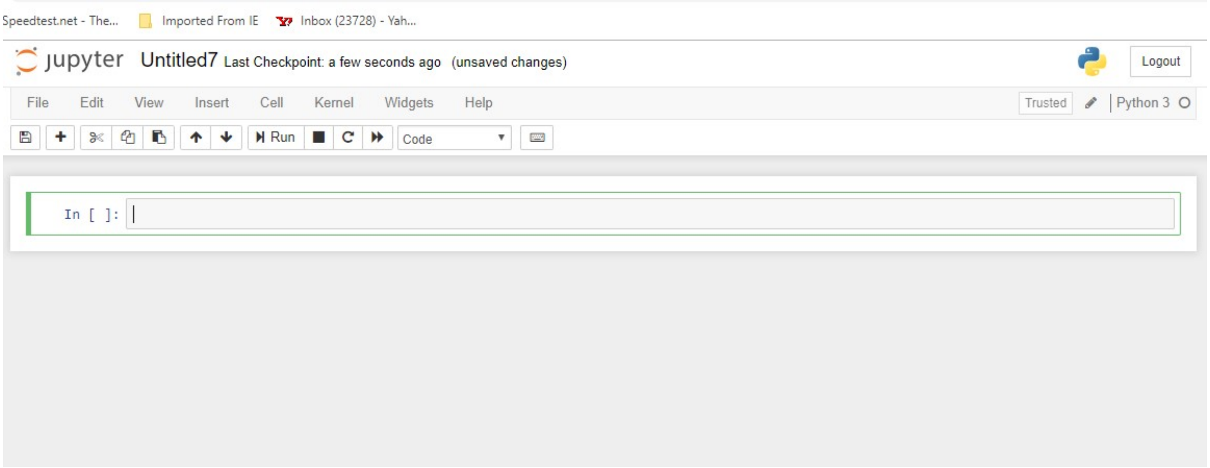
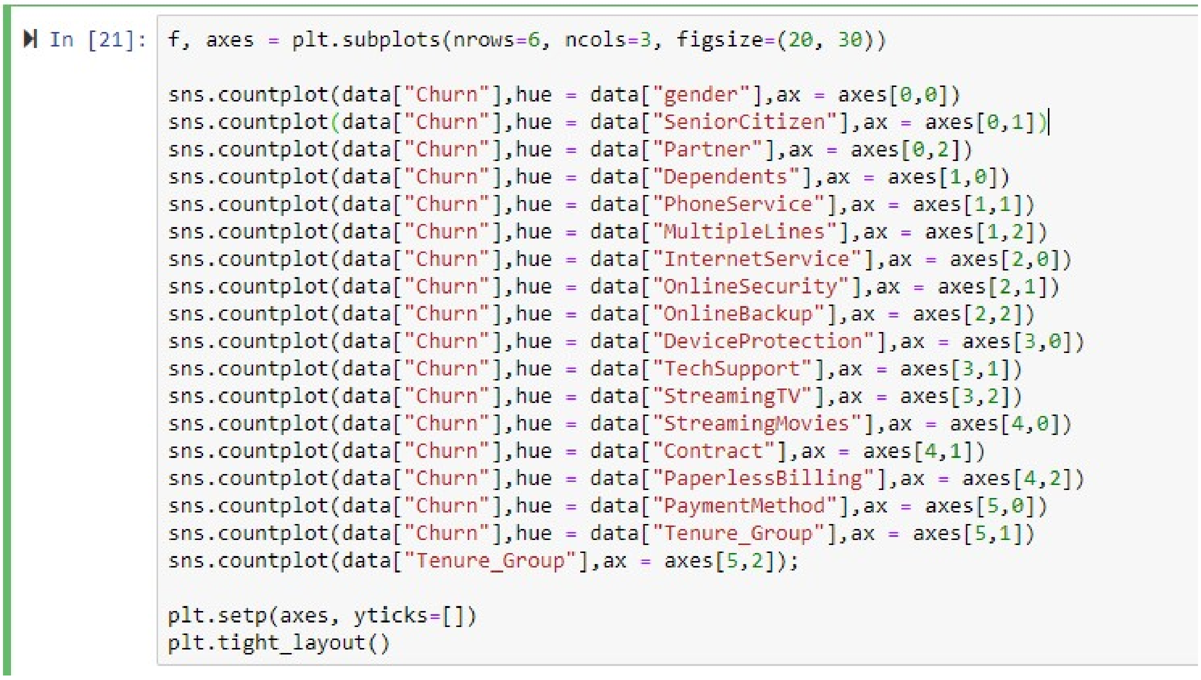


**Figure 4.2:** jupyter notebook



```
In [21]: f, axes = plt.subplots(nrows=6, ncols=3, figsize=(20, 30))

         sns.countplot(data["Churn"],hue = data["gender"],ax = axes[0,0])
         sns.countplot(data["Churn"],hue = data["SeniorCitizen"],ax = axes[0,1])
         sns.countplot(data["Churn"],hue = data["Partner"],ax = axes[0,2])
         sns.countplot(data["Churn"],hue = data["Dependents"],ax = axes[1,0])
         sns.countplot(data["Churn"],hue = data["PhoneService"],ax = axes[1,1])
         sns.countplot(data["Churn"],hue = data["MultipleLines"],ax = axes[1,2])
         sns.countplot(data["Churn"],hue = data["InternetService"],ax = axes[2,0])
         sns.countplot(data["Churn"],hue = data["OnlineSecurity"],ax = axes[2,1])
         sns.countplot(data["Churn"],hue = data["OnlineBackup"],ax = axes[2,2])
         sns.countplot(data["Churn"],hue = data["DeviceProtection"],ax = axes[3,0])
         sns.countplot(data["Churn"],hue = data["TechSupport"],ax = axes[3,1])
         sns.countplot(data["Churn"],hue = data["StreamingTV"],ax = axes[3,2])
         sns.countplot(data["Churn"],hue = data["StreamingMovies"],ax = axes[4,0])
         sns.countplot(data["Churn"],hue = data["Contract"],ax = axes[4,1])
         sns.countplot(data["Churn"],hue = data["PaperlessBilling"],ax = axes[4,2])
         sns.countplot(data["Churn"],hue = data["PaymentMethod"],ax = axes[5,0])
         sns.countplot(data["Churn"],hue = data["Tenure_Group"],ax = axes[5,1])
         sns.countplot(data["Tenure_Group"],ax = axes[5,2]);

         plt.setp(axes, yticks=[])
         plt.tight_layout()
```

**Figure 4.3:** running python 3 on jupyter notebook

# CHAPTER 5

## CONCLUSION AND FUTURE ENHANCEMENT

This study find out which machine learning model is efficient and performs well in order to predict and analyse on the customer data with the help of the predictive algorithm and is used to find whether the customer will churn or not. it got better prediction result with each algorithms for better precision,recall and f-score measures and successfully filtering out false positive errors that is classifying non-churners as churners. This study might help telecom companies what are the factors that causes churn in customers and can take the necessary steps to minimize that. We have observed that this task of predicting customers that will churn is quite complex because of the temporal nature of the problem.This makes problems that vary with time to be analyzed further for further studies.In this work we have proved that data science coupled with machine learning algorithms can help telecom industries to understand their client's behavioral characteristics which will allow them to cut investment costs in aquiring new clients by retaining current customers.

We have observed that this task of predicting customers that will churn is quite complex because of the temporal nature of the problem.This makes problems that vary with time to be analyzed further for further studies.future studies could aim at extracting new features that might allow in separating the two classes of customers better with good precision,recall and f-score measures and successfully filtering out false positive errors that is classifying non-churners as churners.Future study should gather more intel on customers to help in understanding the causes of churn more accurately .

# CHAPTER 6

## REFERENCES

[1] Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari. "Evaluation of machine learning models for employee churn prediction" , 2017 International Conference on Inventive Computing and Informatics (ICICI), 2017

[2] Jas Semrl, Alexandru Matei. "Churn prediction model for effective gym customer retention" , 2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC), 2017

[3] zhenyu chen ; zhiping fan(2011) building comprehensible customer churn prediction model.

[4] Preeti K. Dalvi; Siddhi K.Khandge ;Ashish Deomore; Aditya Bankar ;V. A. Kanade (2016) Analysis of customer churn prediction in telecom industry

[5] Saad Ahmed Qureshi ; Ammar Saleem Rehman ; Ali Mustafa Qamar; Aatif Kamal(2013) . Telecommunication subscribers' churn prediction model using machine learning

[6] Adnan Amin ;Changez Khan ;Imtiaz Ali ;Sajid Anwar(2014) Customer Churn Prediction in Telecommunication Industry: With and without Counter

[7]Alpaydin, E. (2010). Introduction to Machine Learning. London, England: The MIT Press.

[VIII] Archaux, C., Laanaya, H., Martin, A., Khenchaf, A. (2004). An SVM based Churn Detector in Prepaid Mobile Telephony. IEEE .

[8]Berson, A., Smith, S., Thearling, K. (2000). Building Data Mining Applications for CRM. New York, NY: McGraw-Hill.

[9] Brin, S., Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the 7th international conference on World Wide Web, (pp. 107-117). Brisbane, Australia.

[10] Burez, J., Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Elsevier , 4626-4636.

[11] Cetnik, B. (1990). Estimating Probabilities: A crucial task in machine learning. Ninth European Conference on Artificial Intelligence, (pp. 147-149). London.CRISP-DM - Process Model. (n.d.). Retrieved April 29, 2011, from CRISP-DM - Home:www.crisp-dm.org

[12] Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukheejea, S., Nanavati, A. A. (2008). Social ties and their relevance to churn in mobile telecom networks. EDBT '08 Proceedings of the 11th international conference on Extending database technology: Advances in database technology (pp. 668-677). Nantes, France: ACM.

[13] Dash, M., Liu, H. (1997). Feature Selection for Classification. Inrelligent Data Analysis 1 , 131-156.

[14] A Rawat and A Choubey,âĂİA survey on classification techniques in internet environment,âĂİ IJSRSET,vol.2,no.3,2016

[15] Customer prediction in mobile telecom system using data mining techniques,
M.balasubramanium,M.selvarani, international journal of scientific and research publications , volume 4,issue 4, April 2014.

# CHAPTER 7

## APPENDIX

# APPENDIX A

# SOURCE CODE

### A.0.1 importing packages

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
per matplotlib inline
import warnings
warnings.filterwarnings(’ignore’)
data = pd.readcsv("Customer-Churn.csv")
data.head()
Checking for NULL data
data.isna().sum()
Checking for duplicated data
data.duplicated().sum()
Number of Churn
sns.set(style="white", palette="deep", colorcodes=True)
sns.despine(left=True)
sns.countplot(data["Churn"]);
plt.pie(data["Churn"].valuecounts(),explode=(0,0.1),autopct=’ 1.1fperper’,
shadow=True, startangle=90,labels=data["Churn"].unique()) plt.axis(’equal’) ;
```

### A.0.2 data visualization

```
f,axes=plt.subplots(figsize=(18,8))
sns.countplot(data["tenure"],hue = data["Churn"]);
f, axes = plt.subplots(nrows=6, ncols=3, figsize=(20,30))
sns.countplot(data["Churn"],hue = data["gender"],ax = axes[0,0])
sns.countplot(data["Churn"],hue = data["SeniorCitizen"],ax = axes[0,1])
sns.countplot(data["Churn"],hue = data["Partner"],ax = axes[0,2])
sns.countplot(data["Churn"],hue = data["Dependents"],ax = axes[1,0])
sns.countplot(data["Churn"],hue = data["PhoneService"],ax = axes[1,1])
sns.countplot(data["Churn"],hue = data["MultipleLines"],ax = axes[1,2])
sns.countplot(data["Churn"],hue = data["InternetService"],ax = axes[2,0])
sns.countplot(data["Churn"],hue = data["OnlineSecurity"],ax = axes[2,1])
sns.countplot(data["Churn"],hue = data["OnlineBackup"],ax = axes[2,2]) sns.countplot(data["Churn"],hue
= data["DeviceProtection"],ax = axes[3,0])
sns.countplot(data["Churn"],hue = data["TechSupport"],ax = axes[3,1])
sns.countplot(data["Churn"],hue = data["StreamingTV"],ax = axes[3,2])
```

sns.countplot(data["Churn"],hue = data["StreamingMovies"],ax = axes[4,0])
sns.countplot(data["Churn"],hue = data["Contract"],ax = axes[4,1
sns.countplot(data["Churn"],hue = data["PaperlessBilling"],ax = axes[4,2])
sns.countplot(data["Churn"],hue = data["PaymentMethod"],ax = axes[5,0])
sns.countplot(data["Churn"],hue = data["Tenure$_{G}roup”], ax = axes[5, 1])$
$sns.countplot(data[”Tenure_{G}roup”], ax = axes[5, 2]);$
$plt.setp(axes, yticks = [])$
$plt.tightlayout()$
$f, axes = plt.subplots(ncols = 3, figsize = (20, 5))$
$sns.boxplot(x = ”Churn”, y = ”tenure”, data = data, palette =' rainbow', ax = axes[0$
$sns.boxplot(x = ”Churn”, y = ”MonthlyCharges”, data = data, palette =' rainbow', ax = axes[1])$
$sns.boxplot(x = ”Churn”, y = ”TotalCharges”, data = data, palette =' rainbow', ax = axes[2])$
$temp_{c}ols = data.drop(”SeniorCitizen”, axis = 1)$
$sns.pairplot(temp_{c}ols, hue =' Churn', palette =' rainbow')$
$f, axes = plt.subplots(ncols = 2, figsize = (20, 6))$
$sns.barplot(x =' Tenure_{G}roup', y =' MonthlyCharges', data = data, hue = ”Churn”, ax = axes[0])$
$sns.barplot(x =' Tenure_{G}roup', y =' TotalCharges', data = data, hue = ”Churn”, ax = axes[1])$

### A.0.3 Data cleaning

```
data.query("TotalCharges == ’ ’").TotalCharges.count()
data["TotalCharges"] = data["TotalCharges"].replace(" ",np.nan) data.dropna(inplace = True);
data["TotalCharges"] = data["TotalCharges"].astype("float")
data.info()
data[data["TotalCharges"]<0]["TotalCharges"].count()
tempcolumns = [col for col in data.columns if col not in
("customerID","gender","MonthlyCharges","TotalCharges","Churn")]
tempcolumns
for col in tempcolumns:
if col in
("OnlineSecurity","OnlineBackup","DeviceProtection","TechSupport"
,"StreamingTV","StreamingMovies"):
data[col] = data[col].replace(’No internet service’:’No’)
temptenure = np.array(data["tenure"].tolist())
print("min: ".format(temptenure.min()))
print("max: ".format(temptenure.max()))
def tenuretogroup(data):
if data["tenure"] <=12:
return "0-1-year"
elif (data["tenure"] > 12)  (data["tenure"] <= 24 ):
return "1-2-year"
elif (data["tenure"] > 24)  (data["tenure"] <= 36) :
```

return "$2_{3y}ear"$

$elif(data["tenure"] > 36)(data["tenure"] <= 48) :$

$return"3_{4y}ear"$

$elifdata["tenure"] > 48(data["tenure"] <= 60) :$

$return"4_{5y}ear"$

$elifdata["tenure"] > 60(data["tenure"] <= 72) :$

$return"5_{6y}ear"$

$data["Tenure_{G}roup"] = data.apply(lambdadata : tenure_{t}o_{g}roup(data), axis = 1)$

$sns.countplot(data["Tenure_{G}roup"]);$

$cat_{c}ols = [xforxindata.columnsifdata[x].nunique() < 6andx! = "Churn"]$

$num_{c}ols = [xforxindata.columnsifdata[x].nunique() > 6andx! = "customerID"]$

$id_{c}ustomer = data["customerID"]$

$label = data["Churn"]$

$label = label.apply(lambdax : 1ifx == "Yes"else0)$

$fromsklearn.preprocessingimportMinMaxScaler$

$features_{l}og_{t}ransformed = pd.DataFrame(data = data[num_{c}ols])$

$features_{l}og_{t}ransformed[num_{c}ols] = data[num_{c}ols].apply(lambdax : np.log(x + 1))$

$scaler = MinMaxScaler()$

$features_{l}og_{m}inmax_{t}ransform = pd.DataFrame(data = features_{l}og_{t}ransformed)$

$features_{l}og_{m}inmax_{t}ransform[num_{c}ols] = scaler.fit_{t}ransform(features_{l}og_{t}ransformed[num_{c}ols])$

$sns.heatmap(features_{l}og_{m}inmax_{t}ransform.corr(), annot = True, cmap =' jet');$

$features_{l}og_{m}inmax_{t}ransform.drop("tenure", inplace = True, axis = 1)$

$data.drop(["MonthlyCharges", "TotalCharges", "tenure"], axis = 1, inplace = True)$

$data = pd.concat([data, features_{l}og_{m}inmax_{t}ransform], axis = 1)$

$data.info()$

$data.duplicated().sum()$

$data.drop("Churn", inplace = True, axis = 1)$

$data.drop("customerID", inplace = True, axis = 1)$

$data.info()$

$data = pd.get_{d}ummies(data = data, columns = cat_{c}ols)$

$data.head$

$data_{o}riginal = pd.concat([data, label, id_{c}ustomer], axis = 1)$

$data_{o}riginal.info()$

$data_{o}riginal.head()$

### A.0.4 Data preprocessing

cat$_{c}ols = [xforxindata.columnsifdata[x].nunique() < 6andx! = "Churn"]$

$num_{c}ols = [xforxindata.columnsifdata[x].nunique() > 6andx! = "customerID"]$

$id_{c}ustomer = data["customerID"]$

$label = data["Churn"]$

$label = label.apply(lambdax : 1ifx == "Yes"else0)$

$fromsklearn.preprocessingimportMinMaxScaler$

$features_{l}og_{t}ransformed = pd.DataFrame(data = data[num_{c}ols])$

$features_{l}og_{t}ransformed[num_{c}ols] = data[num_{c}ols].apply(lambdax : np.log(x + 1))$

$scaler = MinMaxScaler()$

$features_{l}og_{m}inmax_{t}ransform = pd.DataFrame(data = features_{l}og_{t}ransformed)$
$features_{l}og_{m}inmax_{t}ransform[num_{c}ols] = scaler.fit_{t}ransform(features_{l}og_{t}ransformed[num_{c}ols])$
$sns.heatmap(features_{l}og_{m}inmax_{t}ransform.corr(), annot = True, cmap =' jet');$
$sns.heatmap(features_{l}og_{m}inmax_{t}ransform.corr(), annot = True, cmap =' jet');$
$data.drop(["MonthlyCharges", "TotalCharges", "tenure"], axis = 1, inplace = True)$
$data = pd.concat([data, features_{l}og_{m}inmax_{t}ransform], axis = 1)$
$data.drop("Churn", inplace = True, axis = 1)$
$data.drop("customerID", inplace = True, axis = 1)$

### A.0.5 Evaluating Algorithms

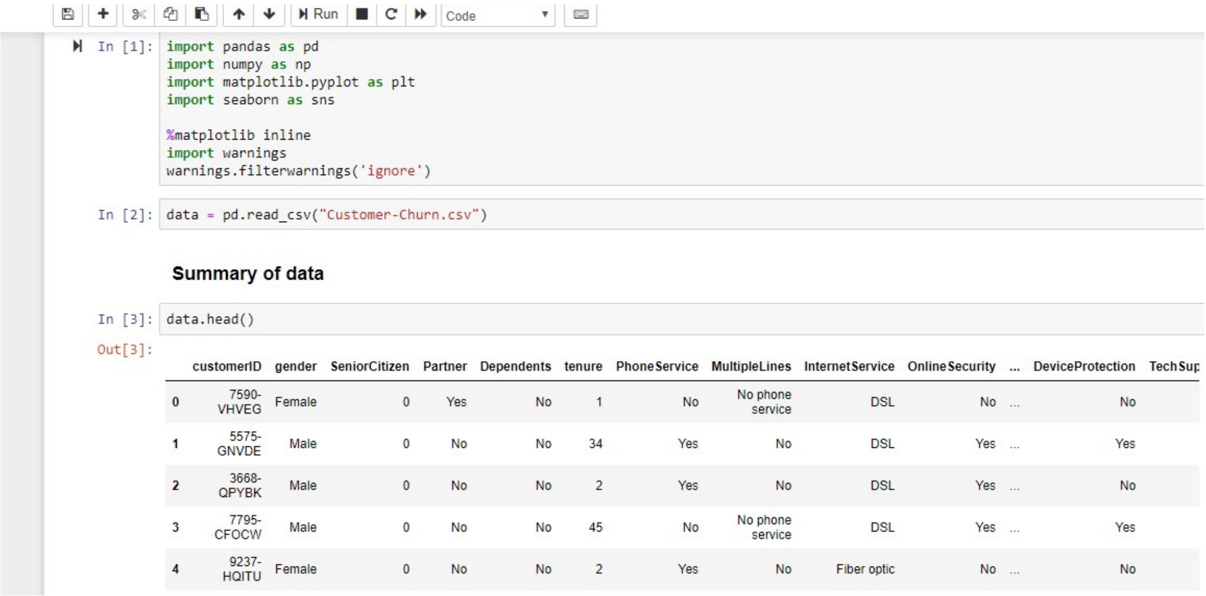
from sklearn.cross$_{v}alidationimporttrain_{t}est_{s}plit$
$X_{t}rain, X_{t}est, y_{t}rain, y_{t}est = train_{t}est_{s}plit(data, label, test_{s}ize = 0.3, random_{s}tate = 42)$
$print("Trainingsethassamples.".format(X_{t}rain.shape[0]))$
$print("Testingsethassamples.".format(X_{t}est.shape[0]))$
$fromsklearn.treeimportDecisionTreeClassifier$
$fromsklearn.linear_{m}odelimportLogisticRegression$
$fromsklearn.svmimportSVC$
$fromxgboostimportXGBClassifier$
$fromsklearn.ensembleimportRandomForestClassifier$
$fromsklearn.metricsimportconfusion_{m}atrix$
$fromsklearn.metricsimportclassification_{r}eport$
$fromsklearn.metricsimportroc_{a}uc_{s}core, roc_{c}urve$
$defapply_{c}lassifier(clf, xTrain, xTest, yTrain, yTest) :$
$clf.fit(xTrain, yTrain)$
$predictions = clf.predict(xTest)$
$conf_{m}tx = confusion_{m}atrix(yTest, predictions)$
$f, axes = plt.subplots(ncols = 2, figsize = (15, 5))$
$sns.heatmap(conf_{m}tx, annot = True, cmap =' tab20c', cbar = False, fmt = "g", ax = axes[0])$
$axes[0].set_{x}label('Predictedlabels')$
$axes[0].set_{y}label('Truelabels')$
$axes[0].set_{t}itle('ConfusionMatrix');$
$axes[0].xaxis.set_{t}icklabels(['NotChurn',' Churn']);$
$axes[0].yaxis.set_{t}icklabels(['NotChurn',' Churn']);$
$print("Classificationreport") :$
$format(classification_{r}eport(yTest, predictions)))$
$roc_{a}uc = roc_{a}uc_{s}core(yTest, predictions)$
$print("AreaunderROCcurve : ", roc_{a}uc, "newline")$
$fpr, tpr,_{=} roc_{c}urve(yTest, predictions)$
$axes[1].plot(fpr, tpr, label = "auc = " + str(roc_{a}uc));$
$axes[1].plot([0, 1], [0, 1], color =' navy', lw = 1, linestyle =' --')$
$plt.xlim([0.0, 1.0])$
$plt.ylim([0.0, 1])$
$plt.xlabel('FalsePositiveRate')$

$plt.ylabel('TruePositiveRate')$
$plt.title('Receiveroperatingcharacteristic')$
$plt.legend(loc = "lowerright")$
$decision_tree = DecisionTreeClassifier(random_state = 42);$
$apply_classifier(decision_tree, X_train, X_test, y_train, y_test)$
$logistic_reg = LogisticRegression(random_state = 42)$
$apply_classifier(logistic_reg, X_train, X_test, y_train, y_test)$
$svm_model = SVC(random_state = 42)$
$apply_classifier(svm_model, X_train, X_test, y_train, y_test)$
$random_forest = RandomForestClassifier(random_state = 42)$
$apply_classifier(random_forest, X_train, X_test, y_train, y_test)$
$Tree_parameters = "max_depth" : [3, 4, 5, 6], "min_samples_leaf" : [1, 2, 3, 4]$
$LogReg_parameters =$
$"C" : [0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 10.0], "warm_start" : ["True", "False]$
$SVM_parameters =$
$"C" : [1.0, 2.0, 3.0],$
$"cache_size" : [100, 200],$
$"decision_function_shape" : ['ovo', 'ovr'],$
$"kernel" : ['sigmoid', "linear"],$
$"tol" : [0.001, 0.0001]$
$RandomForest_parameters = "n_estimators" : [10, 15, 20, 25, 30], "criterion" : ["entropy", "gini"],$

apply$_classifier(randomForest_grid, X_train, X_test, y_train, y_test)$
$fromsklearn.ensembleimportAdaBoostClassifier$
$model = AdaBoostClassifier(base_estimator = randomForest_grid, n_estimators = 4)$
$apply_classifier(model, X_train, X_test, y_train, y_test)$
$fromsklearn.utilsimportresample$
$upsample_data = data_original$
$majority = upsample_data[upsample_data["Churn"] == 0]$
$minority = upsample_data[upsample_data["Churn"] == 1]$
$minority_upsampled = resample(minority, replace = True, n_samples = 5163, random_state =$
$42)$
$del(upsample_data)$
$upsample_data = pd.concat([majority, minority_upsampled])$
$id_customer_upsample = upsample_data["customerID"]$
$label_upsample = upsample_data["Churn"]$
$upsample_data.drop("Churn", inplace = True, axis = 1)$
$upsample_data.drop("customerID", inplace = True, axis = 1)$
$fromsklearn.cross_validationimporttrain_test_split$
$X_train_upS, X_test_upS, y_train_upS, y_test_upS = train_test_split(upsample_data, label_upsample, test_size =$
$0.3, random_state = 42)$
$print("Trainingsethassamples.".format(X_train_upS.shape[0]))$
$print("Testingsethassamples.".format(X_test_upS.shape[0]))$
$model = AdaBoostClassifier(base_estimator = random_forest, n_estimators = 4)$
$apply_classifier(model, X_train_upS, X_test_upS, y_train_upS, y_test_upS)$

# APPENDIX B

# PROJECT WORK



In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
data = pd.read_csv("Customer-Churn.csv")
```

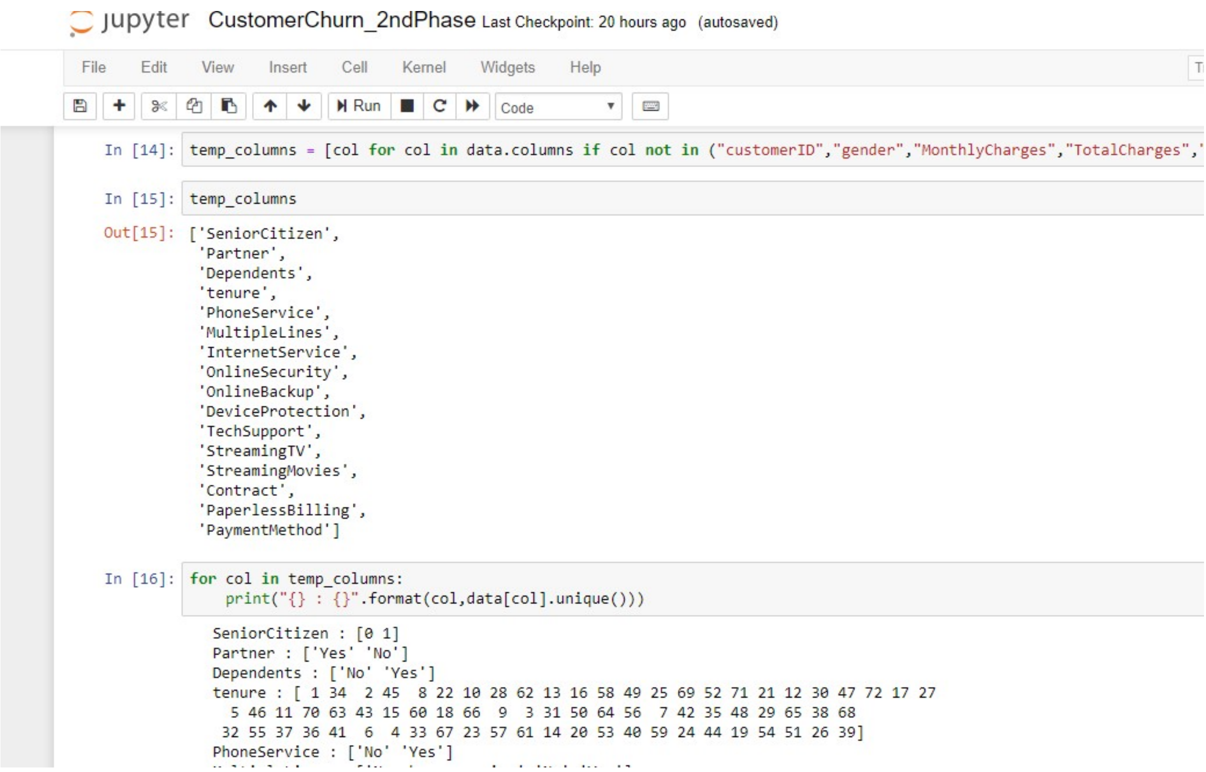
### Summary of data

In [3]:
```
data.head()
```

Out[3]:

| | customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetService | OnlineSecurity | ... | DeviceProtection | TechSup |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 7590-VHVEG | Female | 0 | Yes | No | 1 | No | No phone service | DSL | No | ... | No | |
| 1 | 5575-GNVDE | Male | 0 | No | No | 34 | Yes | No | DSL | Yes | ... | Yes | |
| 2 | 3668-QPYBK | Male | 0 | No | No | 2 | Yes | No | DSL | Yes | ... | No | |
| 3 | 7795-CFOCW | Male | 0 | No | No | 45 | No | No phone service | DSL | Yes | ... | Yes | |
| 4 | 9237-HQITU | Female | 0 | No | No | 2 | Yes | No | Fiber optic | No | ... | No | |

**Figure B.1:** Importing csv dataset


jupyter
CustomerChurn_2ndPhase Last Checkpoint: 20 hours ago (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code

```
In [14]: temp_columns = [col for col in data.columns if col not in ("customerID","gender","MonthlyCharges","TotalCharges",'

In [15]: temp_columns

Out[15]: ['SeniorCitizen',
          'Partner',
          'Dependents',
          'tenure',
          'PhoneService',
          'MultipleLines',
          'InternetService',
          'OnlineSecurity',
          'OnlineBackup',
          'DeviceProtection',
          'TechSupport',
          'StreamingTV',
          'StreamingMovies',
          'Contract',
          'PaperlessBilling',
          'PaymentMethod']

In [16]: for col in temp_columns:
             print("{} : {}".format(col,data[col].unique()))

           SeniorCitizen : [0 1]
           Partner : ['Yes' 'No']
           Dependents : ['No' 'Yes']
           tenure : [ 1 34  2 45  8 22 10 28 62 13 16 58 49 25 69 52 71 21 12 30 47 72 17 27
             5 46 11 70 63 43 15 60 18 66  9  3 31 50 64 56  7 42 35 48 29 65 38 68
            32 55 37 36 41  6  4 33 67 23 57 61 14 20 53 40 59 24 44 19 54 51 26 39]
           PhoneService : ['No' 'Yes']
```

**Figure B.2:** Customer services

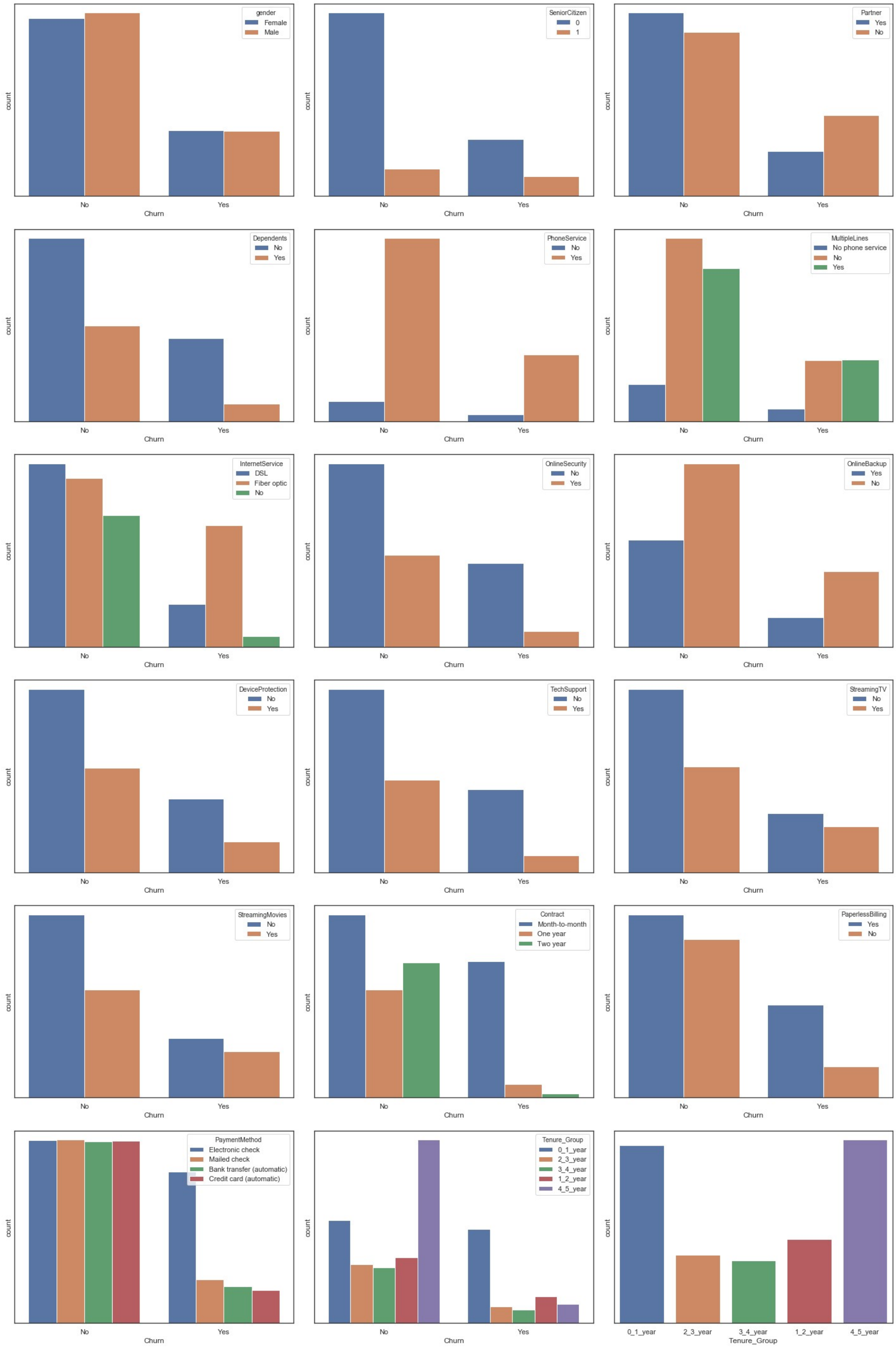
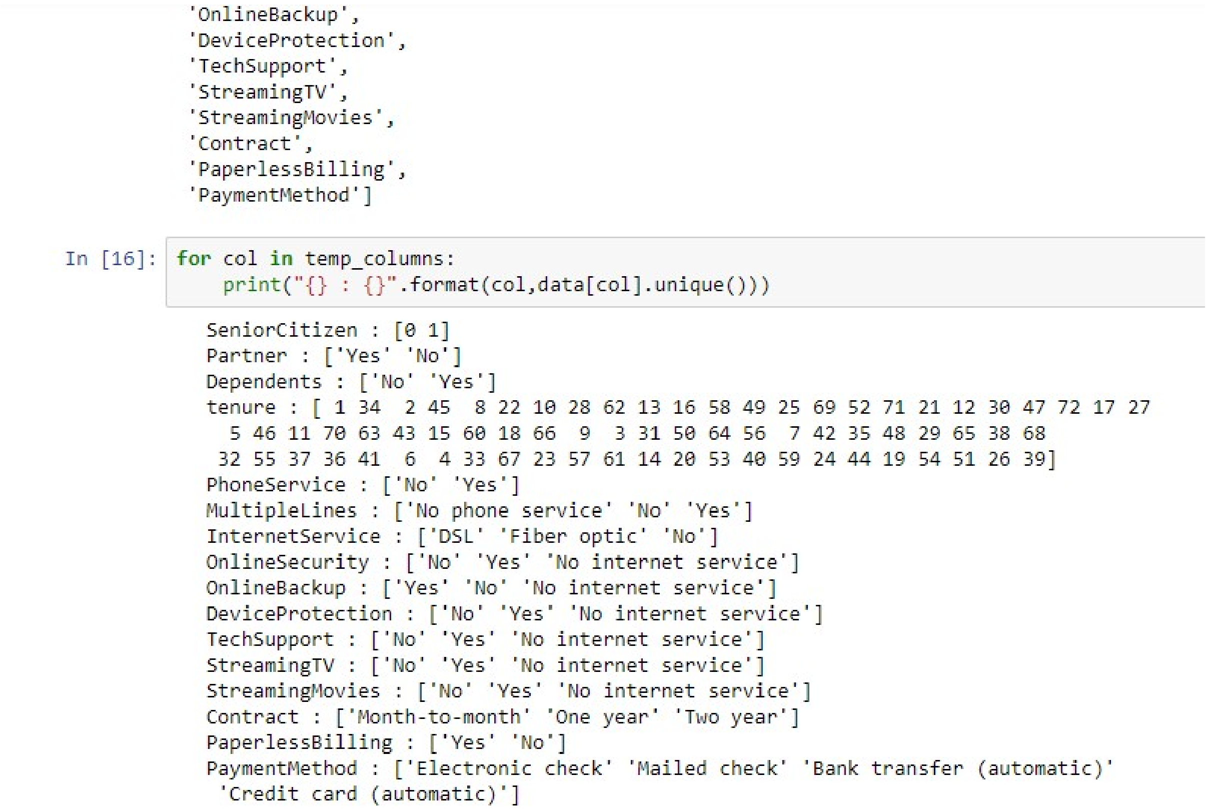


**Figure B.3:** Data visualization

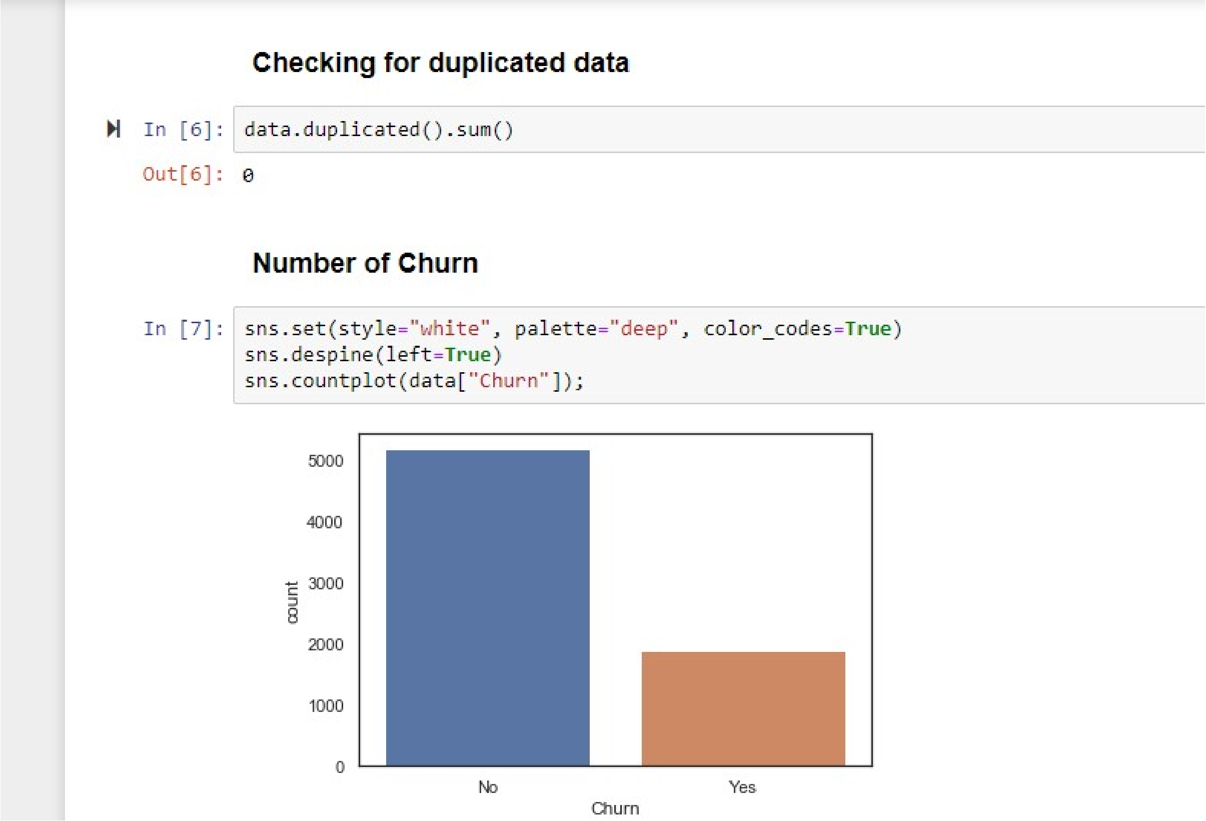


```
 'OnlineBackup',
 'DeviceProtection',
 'TechSupport',
 'StreamingTV',
 'StreamingMovies',
 'Contract',
 'PaperlessBilling',
 'PaymentMethod']
```

```
In [16]: for col in temp_columns:
             print("{} : {}".format(col,data[col].unique()))
```

```
SeniorCitizen : [0 1]
Partner : ['Yes' 'No']
Dependents : ['No' 'Yes']
tenure : [ 1 34  2 45  8 22 10 28 62 13 16 58 49 25 69 52 71 21 12 30 47 72 17 27
  5 46 11 70 63 43 15 60 18 66  9  3 31 50 64 56  7 42 35 48 29 65 38 68
 32 55 37 36 41  6  4 33 67 23 57 61 14 20 53 40 59 24 44 19 54 51 26 39]
PhoneService : ['No' 'Yes']
MultipleLines : ['No phone service' 'No' 'Yes']
InternetService : ['DSL' 'Fiber optic' 'No']
OnlineSecurity : ['No' 'Yes' 'No internet service']
OnlineBackup : ['Yes' 'No' 'No internet service']
DeviceProtection : ['No' 'Yes' 'No internet service']
TechSupport : ['No' 'Yes' 'No internet service']
StreamingTV : ['No' 'Yes' 'No internet service']
StreamingMovies : ['No' 'Yes' 'No internet service']
Contract : ['Month-to-month' 'One year' 'Two year']
PaperlessBilling : ['Yes' 'No']
PaymentMethod : ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
 'Credit card (automatic)']
```

**Figure B.4:** Feature selection



### Checking for duplicated data

```
In [6]: data.duplicated().sum()
Out[6]: 0
```

### Number of Churn

```
In [7]: sns.set(style="white", palette="deep", color_codes=True)
        sns.despine(left=True)
        sns.countplot(data["Churn"]);
```

**Figure B.5:** count vs churn

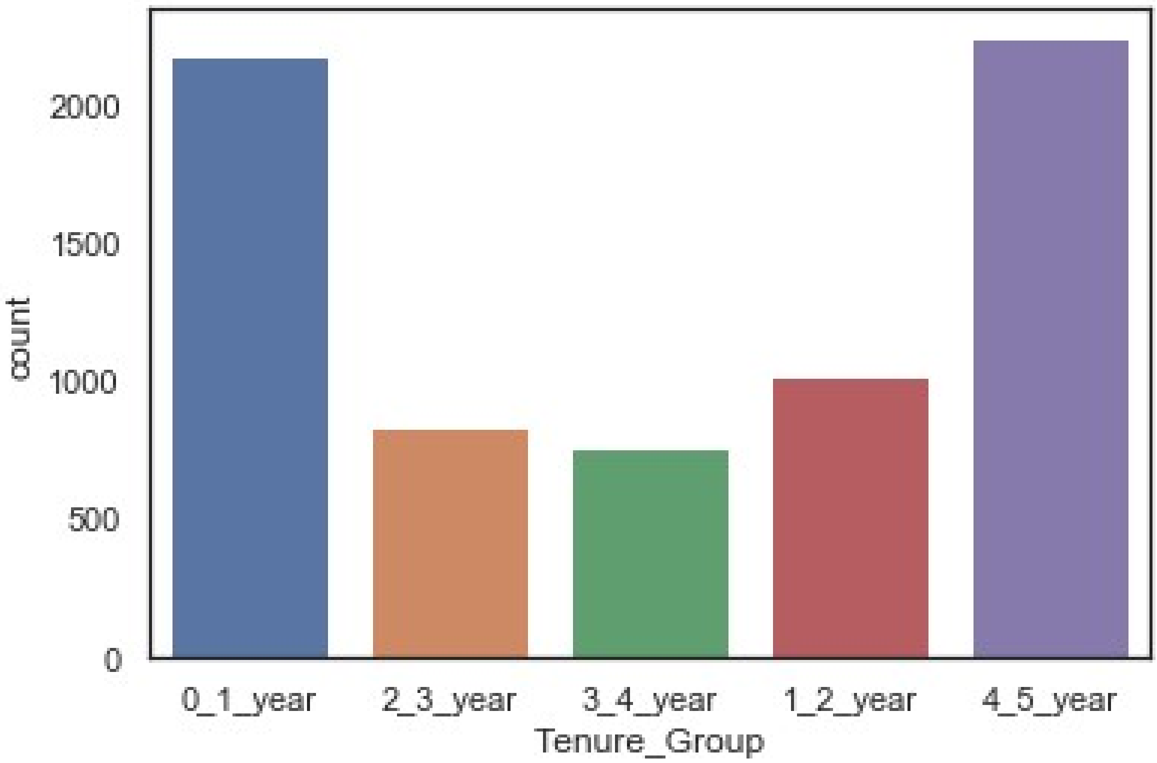


**Figure B.6:** count vs tenure

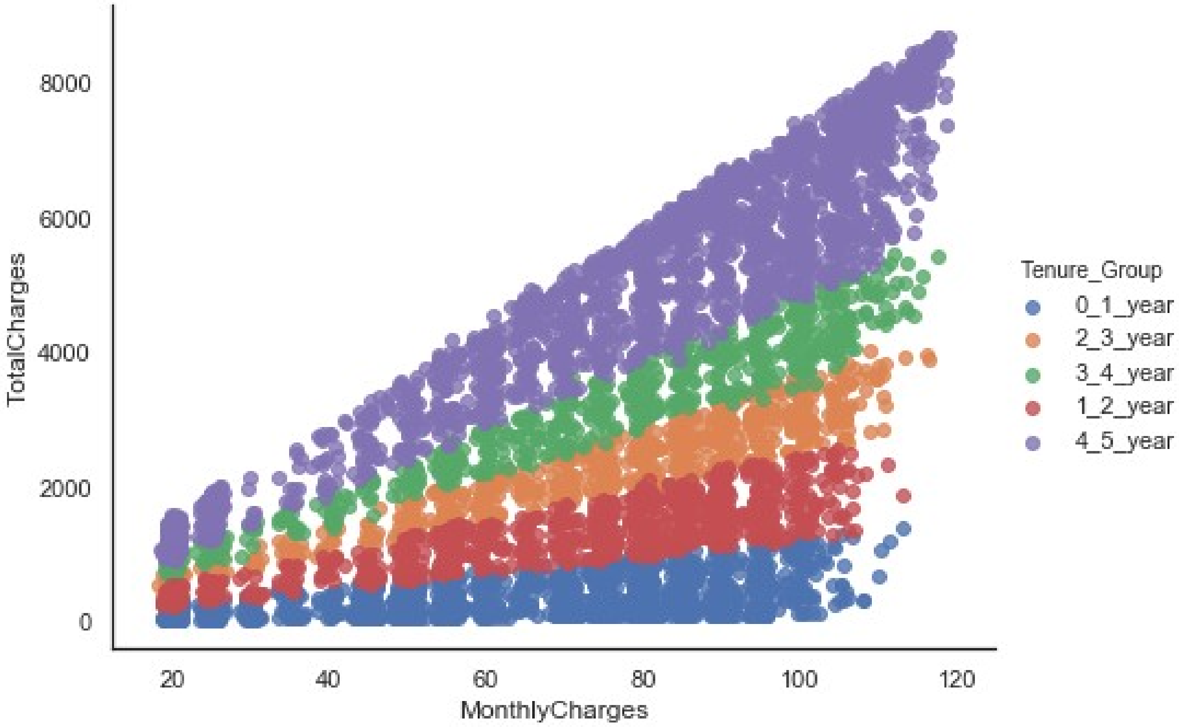


**Figure B.7:** totalcharges vs monthlycharges for different tenure groups

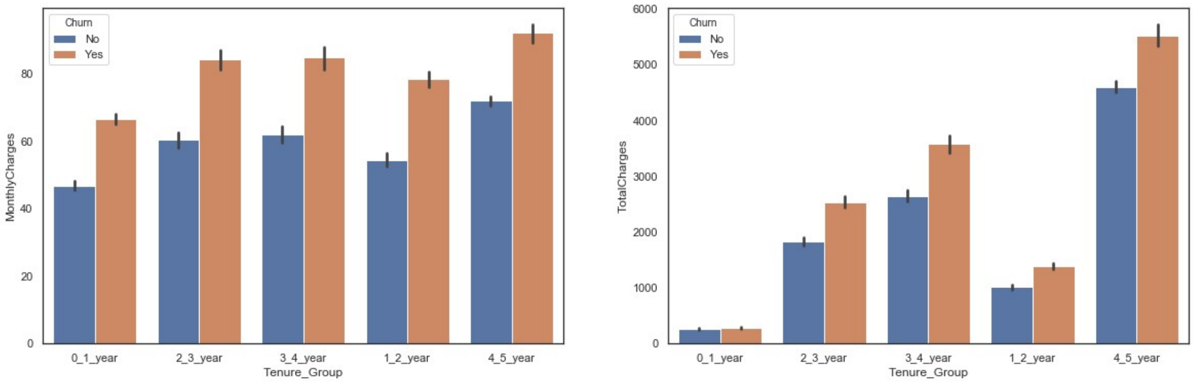
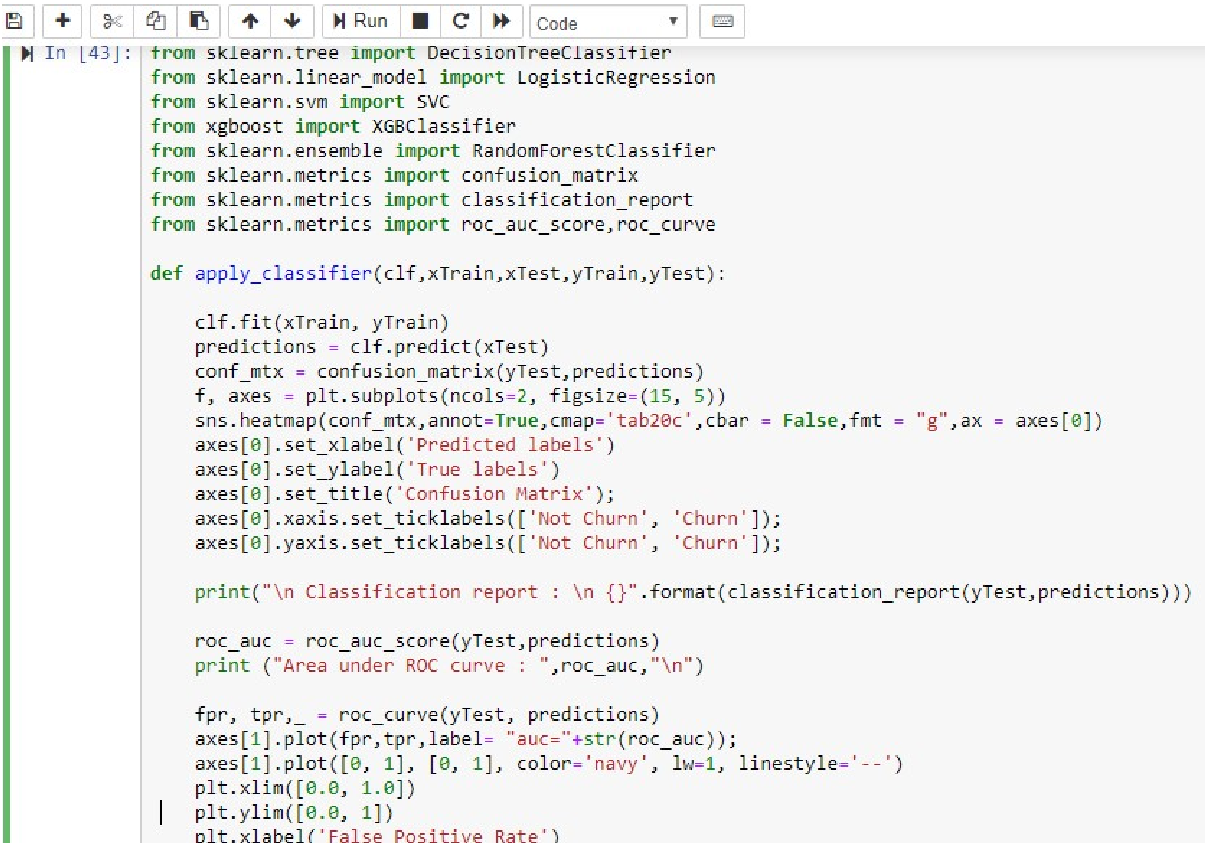
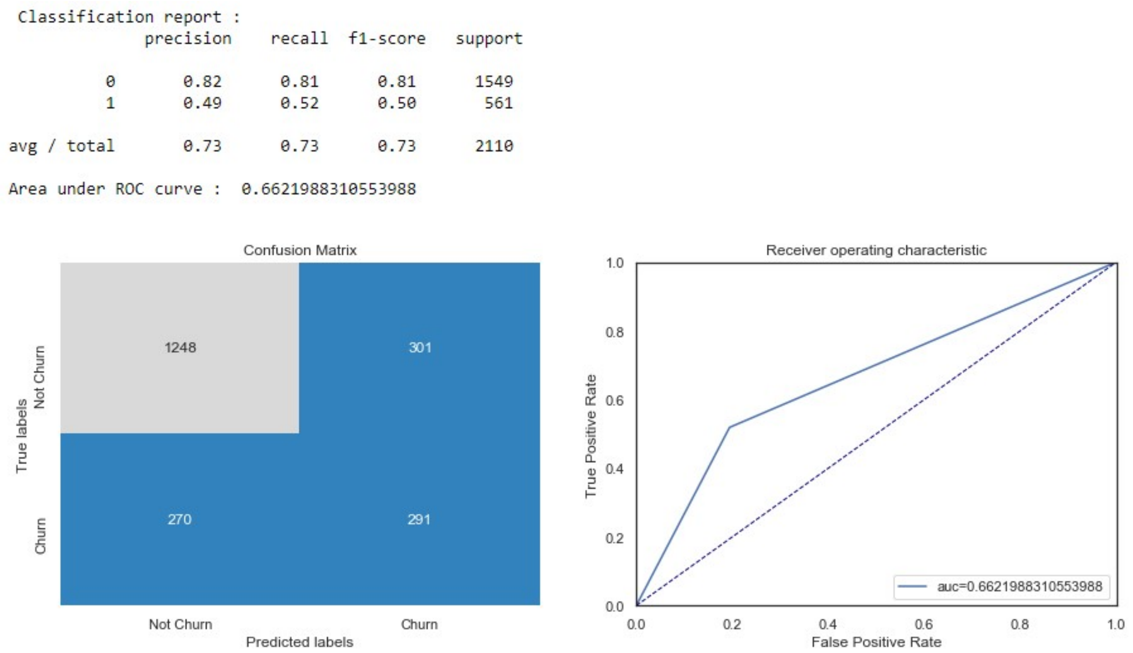


**Figure B.8:** tenure group visualization



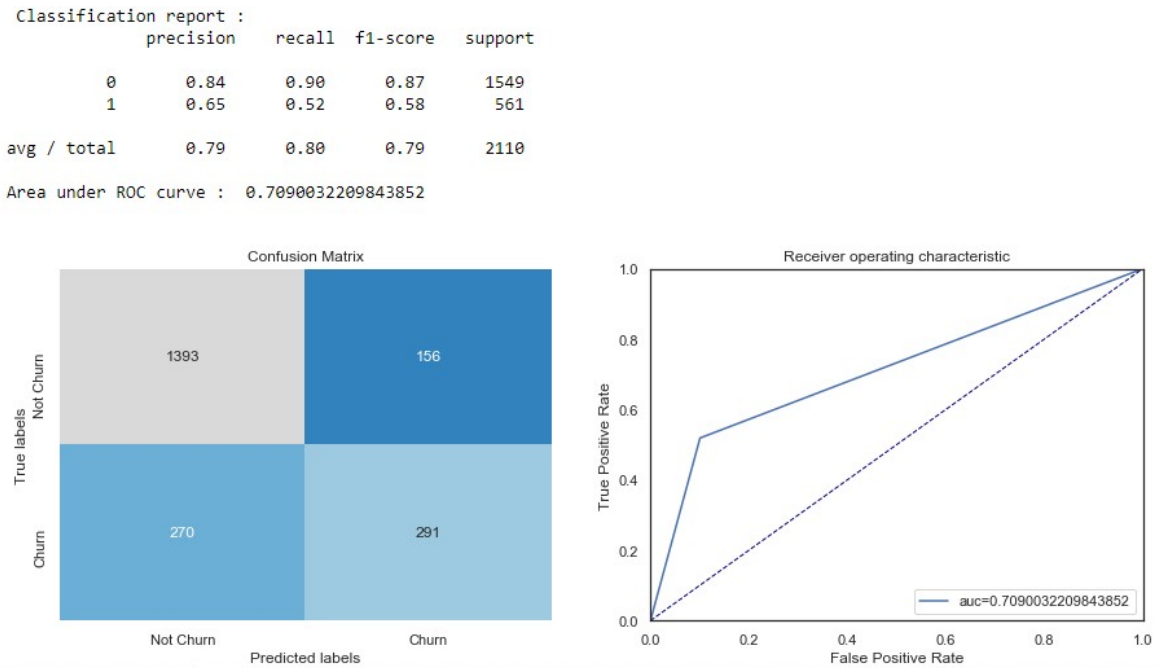
```
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from xgboost import XGBClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import roc_auc_score,roc_curve

def apply_classifier(clf,xTrain,xTest,yTrain,yTest):

    clf.fit(xTrain, yTrain)
    predictions = clf.predict(xTest)
    conf_mtx = confusion_matrix(yTest,predictions)
    f, axes = plt.subplots(ncols=2, figsize=(15, 5))
    sns.heatmap(conf_mtx,annot=True,cmap='tab20c',cbar = False,fmt = "g",ax = axes[0])
    axes[0].set_xlabel('Predicted labels')
    axes[0].set_ylabel('True labels')
    axes[0].set_title('Confusion Matrix');
    axes[0].xaxis.set_ticklabels(['Not Churn', 'Churn']);
    axes[0].yaxis.set_ticklabels(['Not Churn', 'Churn']);

    print("\n Classification report : \n {}".format(classification_report(yTest,predictions)))

    roc_auc = roc_auc_score(yTest,predictions)
    print ("Area under ROC curve : ",roc_auc,"\n")

    fpr, tpr,_ = roc_curve(yTest, predictions)
    axes[1].plot(fpr,tpr,label= "auc="+str(roc_auc));
    axes[1].plot([0, 1], [0, 1], color='navy', lw=1, linestyle='--')
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1])
    plt.xlabel('False Positive Rate')
```

**Figure B.9:** Evaluating algorithms



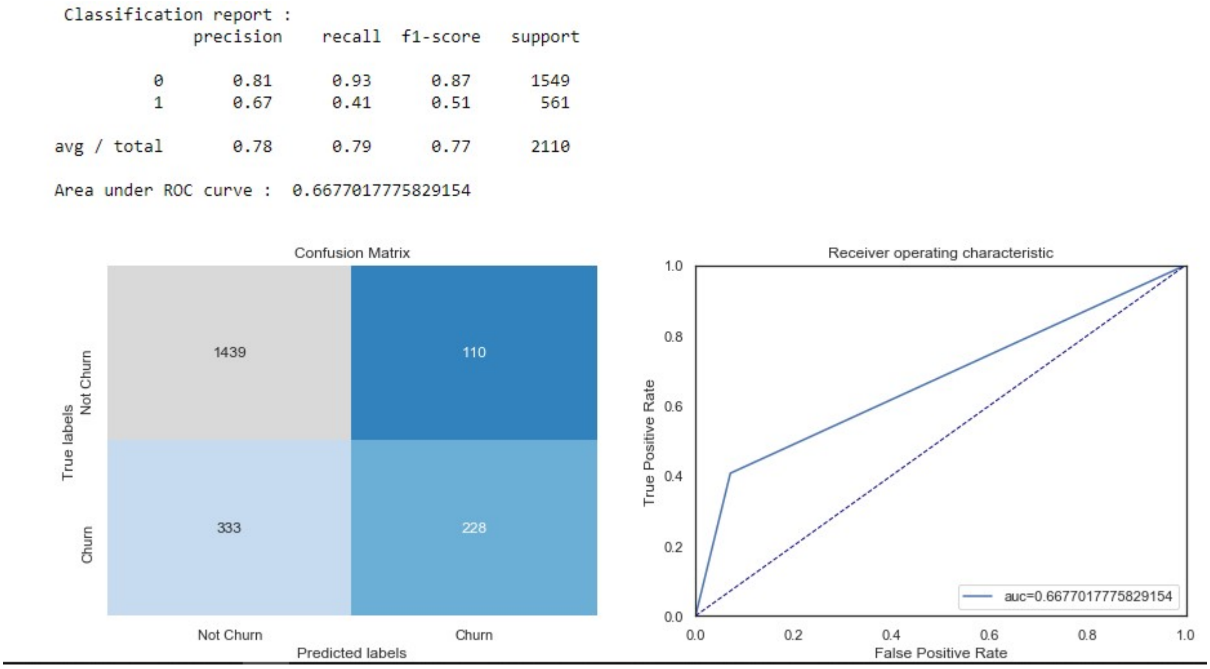
```
Classification report :
              precision    recall  f1-score   support

          0       0.82      0.81      0.81      1549
          1       0.49      0.52      0.50       561

avg / total       0.73      0.73      0.73      2110

Area under ROC curve :  0.6621988310553988
```

**Figure B.10:** decision tree performance report



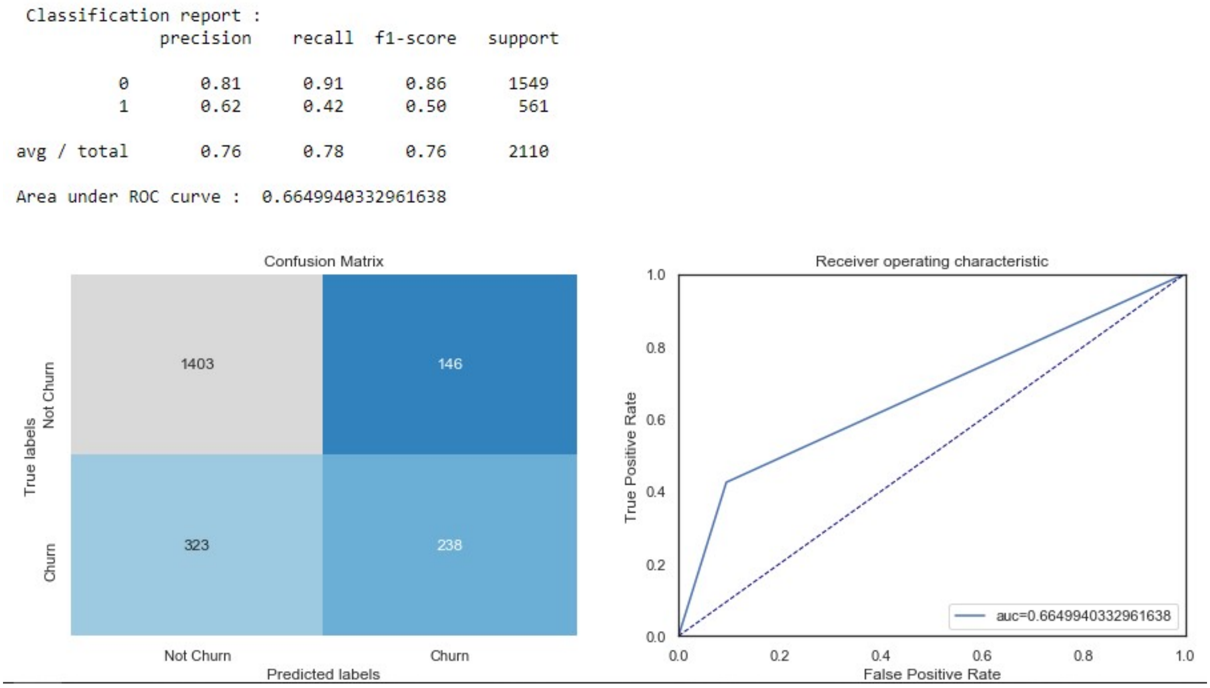
```
Classification report :
              precision    recall  f1-score   support

          0       0.84      0.90      0.87      1549
          1       0.65      0.52      0.58       561

avg / total       0.79      0.80      0.79      2110

Area under ROC curve :  0.7090032209843852
```

**Figure B.11:** logistic regression classification report



```
 Classification report :
              precision    recall  f1-score   support

           0       0.81      0.93      0.87      1549
           1       0.67      0.41      0.51       561

avg / total       0.78      0.79      0.77      2110

Area under ROC curve :  0.6677017775829154
```

**Figure B.12:** SVM performance report



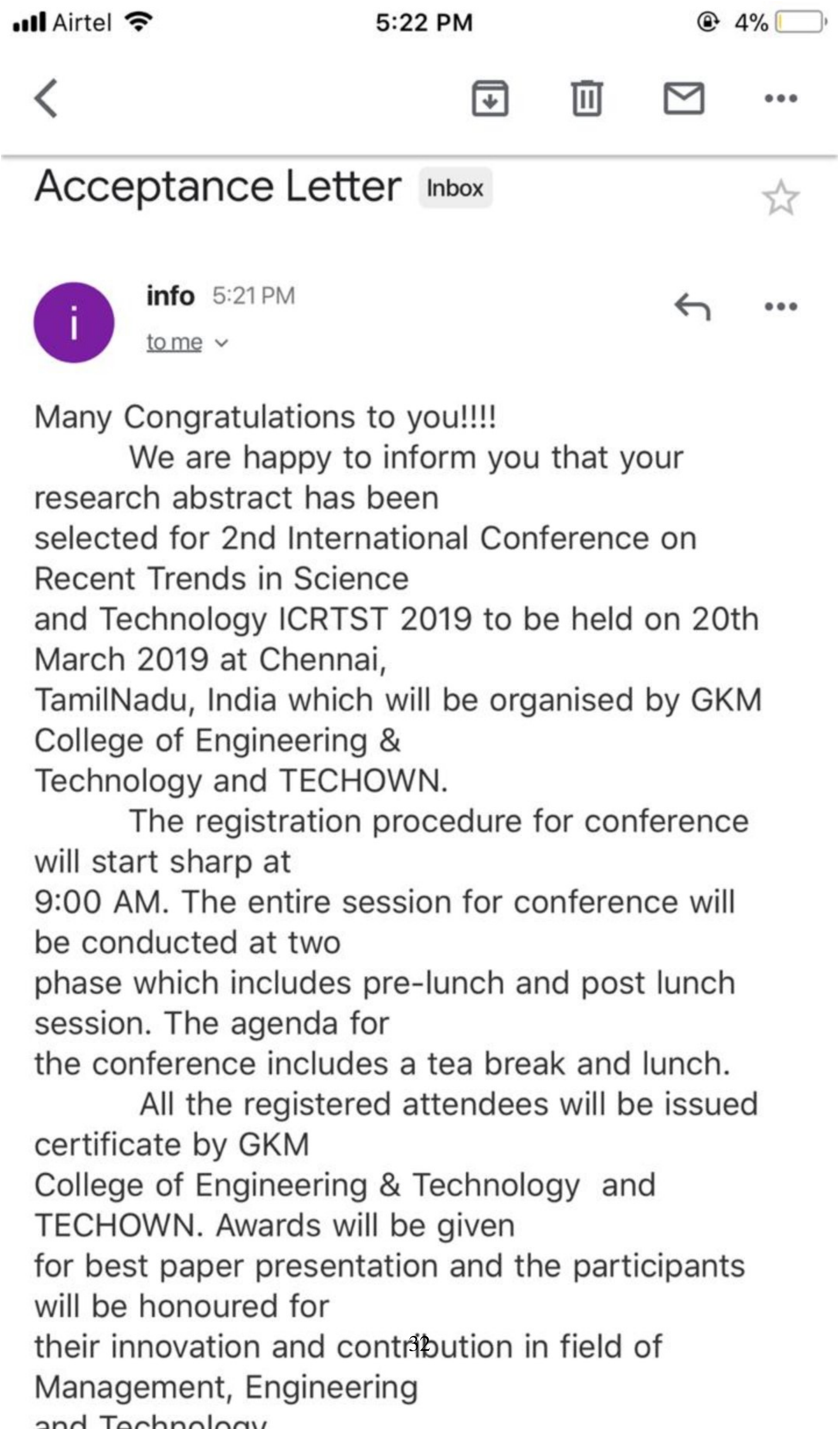
```
 Classification report :
              precision    recall  f1-score   support

           0       0.81      0.91      0.86      1549
           1       0.62      0.42      0.50       561

avg / total       0.76      0.78      0.76      2110

Area under ROC curve :  0.6649940332961638
```

**Figure B.13:** Random forest performance report

# APPENDIX C

## PAPER ACCEPTANCE

We submitted our paper for 2nd international conference on Recent Trends in Science and Technology ICRTST 2019 .We received the acceptance letter and presented our work at the ICRTST conference held on 20th march 2019 at chennai,tamilnadu, India will was organised by GKM college of engineering and technology and TECHTOWN and received certificates of participation. we have forwarded our research abstract for journal support. ICRTST received our papere and forwarded the same to publication.



Acceptance Letter

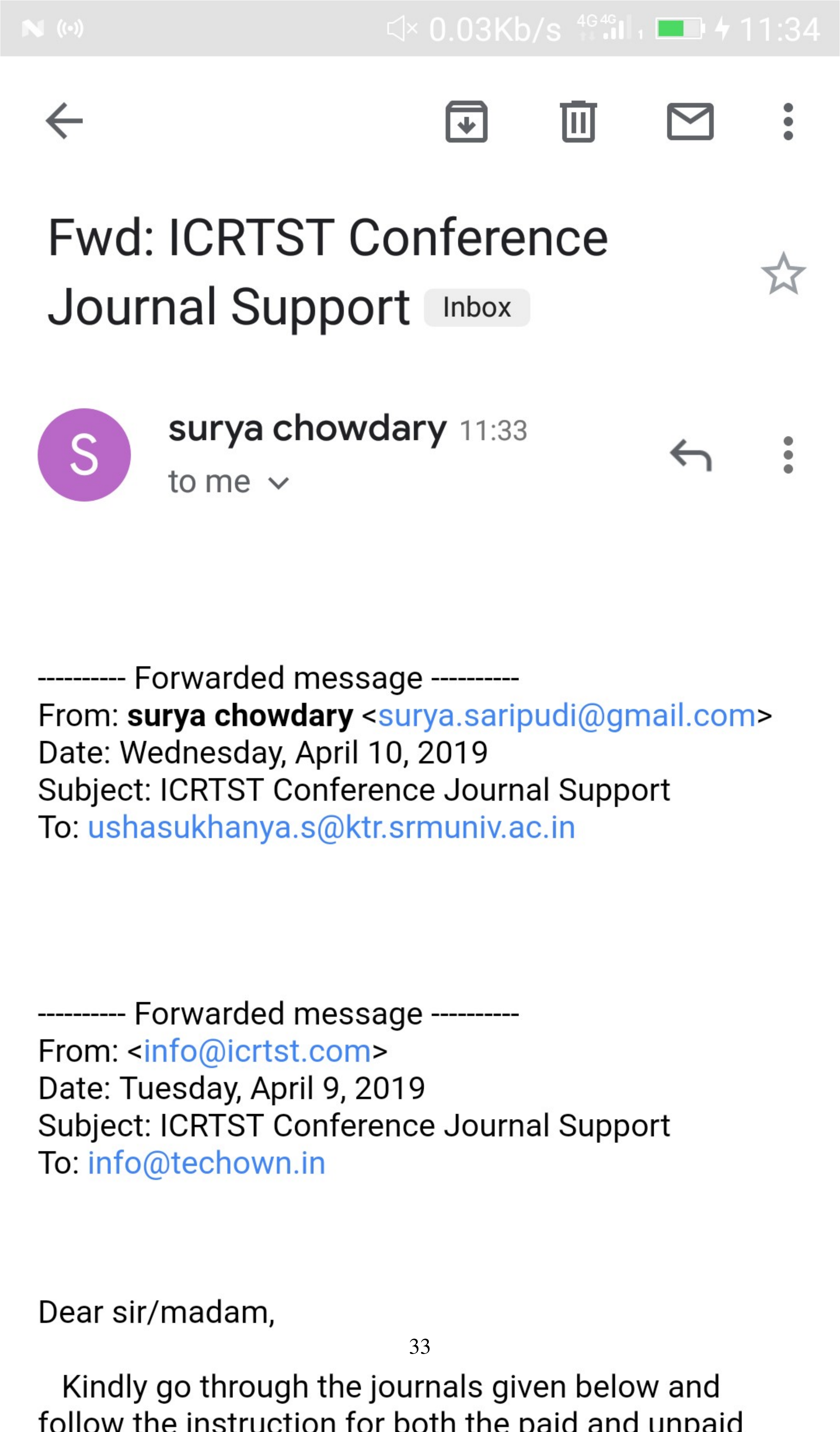
info 5:21 PM
to me

Many Congratulations to you!!!!

We are happy to inform you that your research abstract has been selected for 2nd International Conference on Recent Trends in Science and Technology ICRTST 2019 to be held on 20th March 2019 at Chennai, TamilNadu, India which will be organised by GKM College of Engineering & Technology and TECHOWN.

The registration procedure for conference will start sharp at 9:00 AM. The entire session for conference will be conducted at two phase which includes pre-lunch and post lunch session. The agenda for the conference includes a tea break and lunch.

All the registered attendees will be issued certificate by GKM College of Engineering & Technology and TECHOWN. Awards will be given for best paper presentation and the participants will be honoured for their innovation and contribution in field of Management, Engineering and Technology



Fwd: ICRTST Conference Journal Support

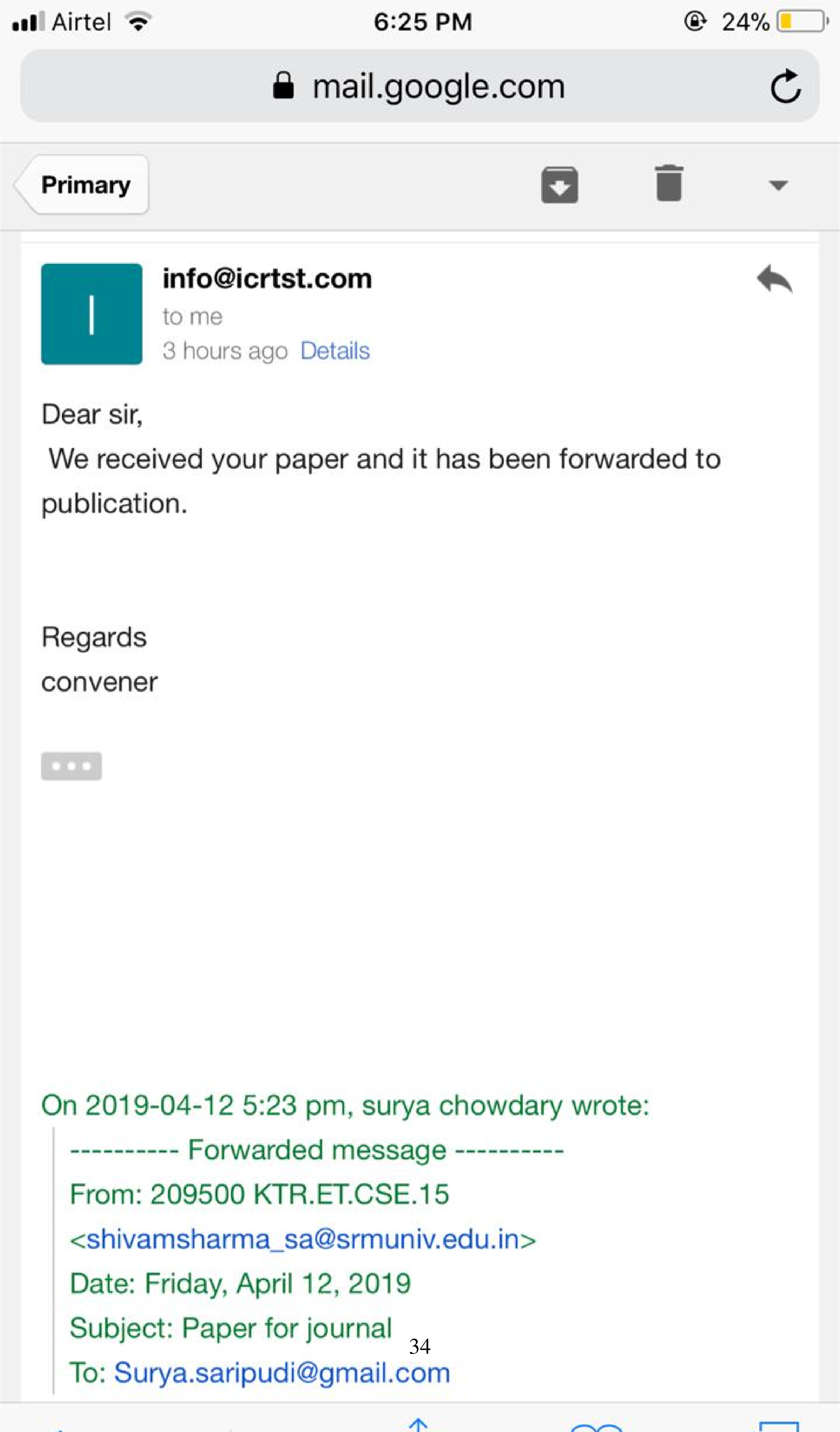
surya chowdary 11:33
to me

---------- Forwarded message ----------
From: **surya chowdary** <surya.saripudi@gmail.com>
Date: Wednesday, April 10, 2019
Subject: ICRTST Conference Journal Support
To: ushasukhanya.s@ktr.srmuniv.ac.in

---------- Forwarded message ----------
From: <info@icrtst.com>
Date: Tuesday, April 9, 2019
Subject: ICRTST Conference Journal Support
To: info@techown.in

Dear sir/madam,

Kindly go through the journals given below and follow the instruction for both the paid and unpaid



**info@icrtst.com**

to me

3 hours ago Details

Dear sir,

We received your paper and it has been forwarded to publication.

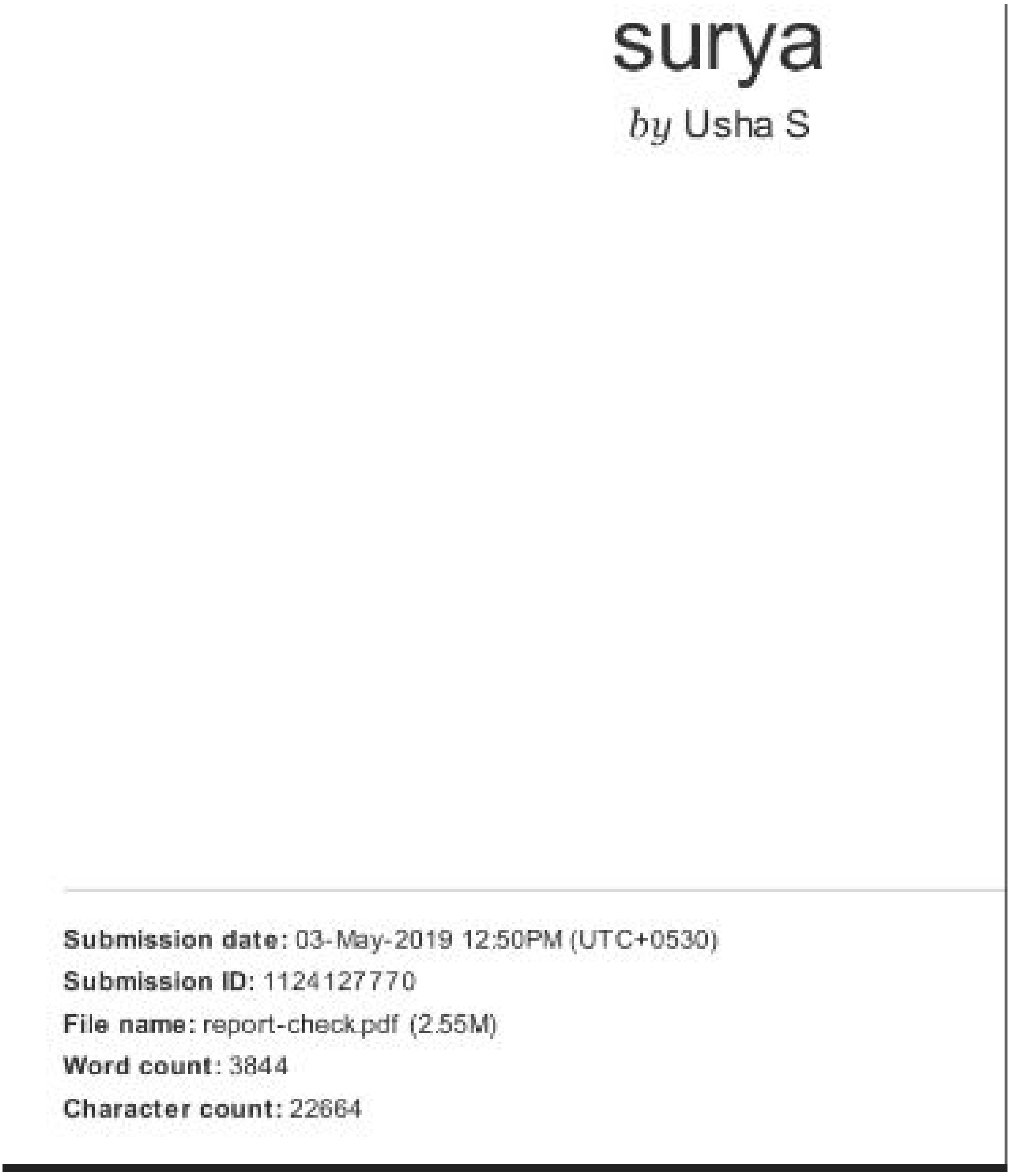
Regards
convener

On 2019-04-12 5:23 pm, surya chowdary wrote:

> ---------- Forwarded message ----------
> From: 209500 KTR.ET.CSE.15
> <shivamsharma_sa@srmuniv.edu.in>
> Date: Friday, April 12, 2019
> Subject: Paper for journal
> To: Surya.saripudi@gmail.com

# APPENDIX D

## PLAGIARISM REPORT



surya

*by* Usha S

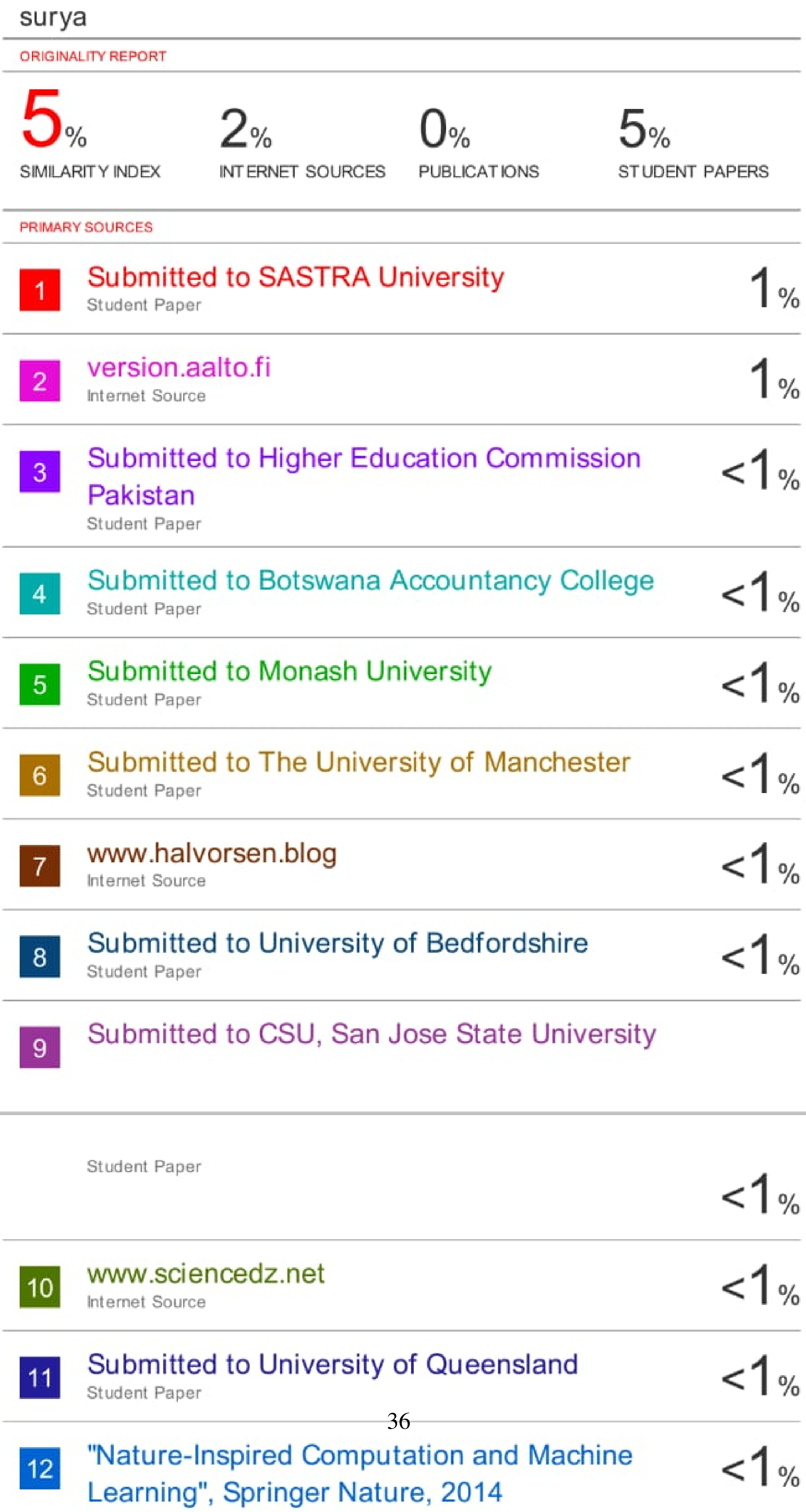
**Submission date:** 03-May-2019 12:50PM (UTC+0530)

**Submission ID:** 1124127770

**File name:** report-check.pdf (2.55M)

**Word count:** 3844

**Character count:** 22664



surya

ORIGINALITY REPORT

| 5% | 2% | 0% | 5% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to SASTRA University<br>Student Paper | 1% |
| 2 | version.aalto.fi<br>Internet Source | 1% |
| 3 | Submitted to Higher Education Commission Pakistan<br>Student Paper | <1% |
| 4 | Submitted to Botswana Accountancy College<br>Student Paper | <1% |
| 5 | Submitted to Monash University<br>Student Paper | <1% |
| 6 | Submitted to The University of Manchester<br>Student Paper | <1% |
| 7 | www.halvorsen.blog<br>Internet Source | <1% |
| 8 | Submitted to University of Bedfordshire<br>Student Paper | <1% |
| 9 | Submitted to CSU, San Jose State University<br>Student Paper | <1% |
| 10 | www.sciencedz.net<br>Internet Source | <1% |
| 11 | Submitted to University of Queensland<br>Student Paper | <1% |
| 12 | "Nature-Inspired Computation and Machine Learning", Springer Nature, 2014 | <1% |

# APPENDIX E

## CONTRIBUTION

**Shivam sharma : RA1511003010367** prepared the survey paper and PPTs for the reviews .Performed data analysis on various attributes (Data exploration),identified what all features and relevant to the target variable and extracted them from visual analysis using the data visualization packages.presented the work at the ICRTST conference held on 20th march 2019 at chennai,tamilnadu, India which was organised by GKM college of engineering and technology and TECHTOWN and received certificate of participation.Tested the performances of different algorithms.

**Surya Vamsi Saripudi : RA1511003010373** Selected the topic for project and collected dataset from kaggle website for analysis and prepared the literature survey.Performed Data cleaning by removing null values . Converted categorical into numerical.Submitted paper for conference.presented the work at the ICRTST conference held on 20th march 2019 at chennai,tamilnadu, India which was organised by GKM college of engineering and technology and TECHTOWN and received certificate of participation.Decided the algorithms to be used for making the model.