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



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 Ori-ented 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 sig-nificant 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 informa-tion 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 predic-tion 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 Engi-neering 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 en-couragement 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 in-puts during the project reviews.

We register our immeasurable thanks to our Faculty Advisor,Mrs.R.Radha,Assistant Profes-sor, Department of Computer Science and Engineering, SRM Institute of Science and Technol-ogy, 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)

iv

TABLE OF CONTENTS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| [ABSTRACT](#page3) | | |  | iii |
| [ACKNOWLEDGEMENTS](#page4) | | | | iv |
| [LIST OF TABLES](#page7) | | | | vii |
| [LIST OF FIGURES](#page8) | | | | viii |
| [ABBREVIATIONS](#page9) | | | | ix |
| [1](#page10) | [INTRODUCTION](#page10) | | | 1 |
|  | [1.1](#page10) | [Churn Prediction](#page10) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 1 |
|  |  | [1.1.1](#page12) | [Modules](#page12) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 3 |
| [2](#page15) | [LITERATURE SURVEY](#page15) | | | 6 |
|  | [2.1](#page15) | [Evaluation of machine learning models for employee churn prediction](#page15) . . . | | 6 |
|  | [2.2](#page15) | [Churn prediction model for effective gym customer retention](#page15) . . . . . . . . | | 6 |
|  | [2.3](#page15) | [Building comprehensible customer churn prediction model.](#page15) . . . . . . . . . | | 7 |
|  | [2.4](#page16) | [Analysis of customer churn prediction in telecom industry](#page16) . . . . . . . . . | | 7 |
|  | [2.5](#page16) | [Telecommunication subscribers’ churn prediction model using machine learn-](#page16) | |  |
|  |  | [ing](#page16) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 7 |
|  | [2.6](#page16) | [Customer Churn Prediction in Telecommunication Industry: With and without](#page16) | |  |
|  |  | [Counter](#page16) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 8 |
| [3 System Analysis and Design](#page18) | | | | 9 |
|  | [3.1](#page19) | [MODEL TRAINING AND PERFORMANCE REPORT](#page19) . . . . . . . . . . | | 10 |
|  | [3.2](#page19) | [PERFORMANCE METRICS](#page19) . . . . . . . . . . . . . . . . . . . . . . . . | | 10 |
|  |  | [3.2.1 ACCURACY](#page19) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 10 |
|  |  | [3.2.2](#page19) | [PRECISION](#page19) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 11 |
|  |  | [3.2.3](#page20) | [RECALL](#page20) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 11 |

v

|  |  |  |  |
| --- | --- | --- | --- |
|  | [3.2.4](#page20) | [F-1 SCORE](#page20) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 11 |
|  | [3.2.5](#page20) | [SPECIFICITY](#page20) . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 11 |
| [4](#page22) | [SOFTWARE REQUIREMENTS](#page22) | | 13 |
|  | [4.0.1](#page22) | [ANACONDA NAVIGATOR](#page22) . . . . . . . . . . . . . . . . . . . . . | 13 |
|  | [4.0.2](#page22) | [JUPYTER NOTEBOOK](#page22) . . . . . . . . . . . . . . . . . . . . . . . | 13 |
| [5](#page24) | [Conclusion and Future Enhancement](#page24) | | 15 |
| [6](#page25) | [REFERENCES](#page25) | | 16 |
| [7](#page26) | [APPENDIX](#page26) |  | 17 |
| [A](#page27) | [Source Code](#page27) |  | 18 |
|  | [A.0.1](#page27) | [importing packages](#page27) . . . . . . . . . . . . . . . . . . . . . . . . . . | 18 |
|  | [A.0.2](#page27) | [data visualization](#page27) . . . . . . . . . . . . . . . . . . . . . . . . . . . | 18 |
|  | [A.0.3](#page28) | [Data cleaning](#page28) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 19 |
|  | [A.0.4](#page29) | [Data preprocessing](#page29) . . . . . . . . . . . . . . . . . . . . . . . . . . | 20 |
|  | [A.0.5](#page30) | [Evaluating Algorithms](#page30) . . . . . . . . . . . . . . . . . . . . . . . . | 21 |
| [B](#page32) | [Project Work](#page32) | | 23 |
| [C](#page40) | [Paper acceptance](#page40) | | 31 |
| [D](#page44) | [Plagiarism Report](#page44) | | 35 |
| [E](#page46) | [Contribution](#page46) | | 37 |

LIST OF TABLES

|  |  |
| --- | --- |
| [3.1 Confusion Matrix](#page21) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 12 |

vii

LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| [1.1](#page11) | [System Overview](#page11) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 2 |
| [1.2](#page11) | [System Architecture](#page11) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 2 |
| [1.3](#page13) | [Dataset for analysis](#page13) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 4 |
| [1.4](#page14) | [importing train-test-split method](#page14) . . . . . . . . . . . . . . . . . . . . . . . | 5 |
| [4.1](#page22) | [anaconda navigator](#page22) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 13 |
| [4.2](#page23) | [jupyter notebook](#page23) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 14 |
| [4.3](#page23) | [running python 3 on jupyter notebook](#page23) . . . . . . . . . . . . . . . . . . . . | 14 |
| [B.1](#page32) | [Importing csv dataset](#page32) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 23 |
| [B.2](#page33) | [Customer services](#page33) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 24 |
| [B.3](#page34) | [Data visualization](#page34) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 25 |
| [B.4](#page35) | [Feature selection](#page35) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 26 |
| [B.5](#page35) | [count vs churn](#page35) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 26 |
| [B.6](#page36) | [count vs tenure](#page36) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 27 |
| [B.7](#page36) | [totalcharges vs monthlycharges for different tenure groups](#page36) . . . . . . . . . | 27 |
| [B.8](#page37) | [tenure group visualization](#page37) . . . . . . . . . . . . . . . . . . . . . . . . . . . | 28 |
| [B.9](#page37) | [Evaluating algorithms](#page37) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 28 |
| [B.10](#page38) | [decision tree performance report](#page38) . . . . . . . . . . . . . . . . . . . . . . . | 29 |
| [B.11](#page38) | [logistic regression classification report](#page38) . . . . . . . . . . . . . . . . . . . . | 29 |
| [B.12](#page39) | [SVM performance report](#page39) . . . . . . . . . . . . . . . . . . . . . . . . . . . | 30 |
| [B.13](#page39) | [Random forest performance report](#page39) . . . . . . . . . . . . . . . . . . . . . . | 30 |

viii

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

ix

CHAPTER 1

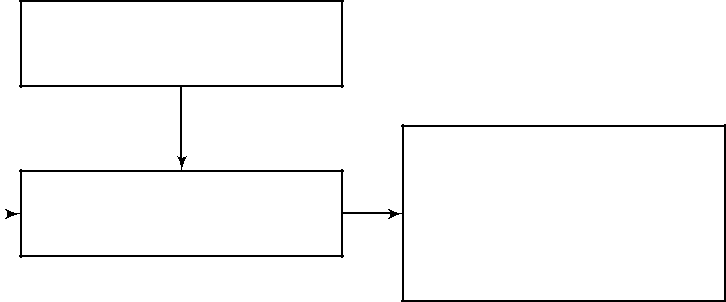
INTRODUCTION

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 set-ting 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 divi-sion 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

Classification(evaluating

performance of

|  |  |  |  |
| --- | --- | --- | --- |
| feature extraction |  | Dataset |  |
|  |  |

models using

confusion matrix

Figure 1.1: System Overview

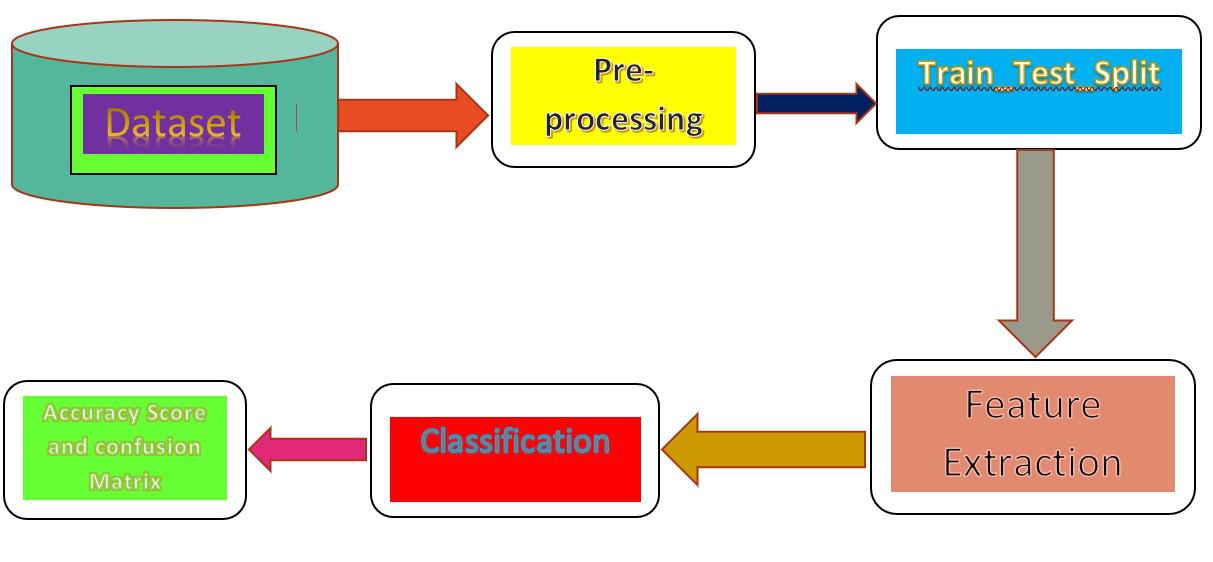


Figure 1.2: System Architecture

It makes a difference a great deal since associations frequently need to invest immense mea-sure 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 dis-cover 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

2

with and consequently it is required that ought to attempt a considerably more thorough proce-dure 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 Cus-tomer 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 Cus-tomer 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.

3

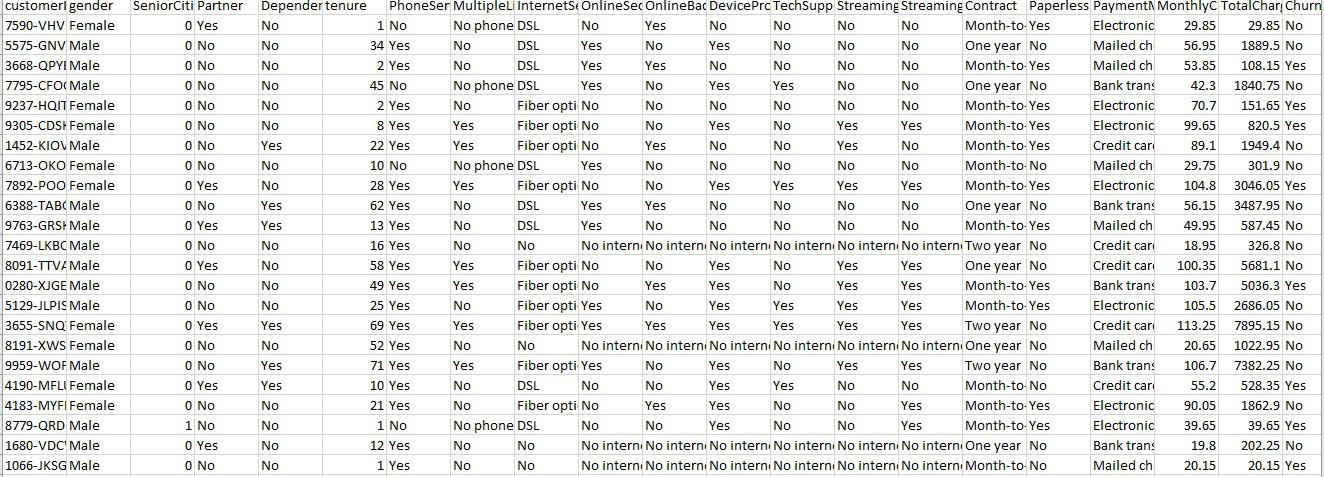


Figure 1.3: Dataset for analysis

Preprocessing phase

This phase is an very important step which prepares the dataset for the modelling part.It sig-nificantly 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 sub-scription 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

4

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.

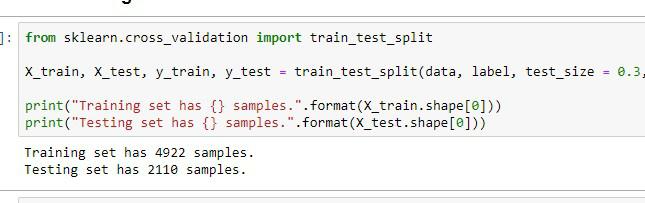


Figure 1.4: importing train-test-split method

5

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 uti-lized for characterization. This examination requires further investigation to limit the forecast rate.

2.2 Churn prediction model for effective gym customer re-

tention

1. This paper fabricates model to foresee client conduct in wellness businesses it is discov-ered 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 neu-ral 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 indus-

try

[4]This paper focuses on analyzing customer churn in telecom industry using logistic regres-sion, neural network and decision trees, although neural networks perform very well for classi-fication 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 canâAZt 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.

7

2.6 Customer Churn Prediction in Telecommunication In-

dustry: With and without Counter

1. 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.

8

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 situ-ation,are strongly focused in satisfying and retaining their customers in order to prevent churn. Particularly in the subscription oriented businesses, such as telecommunications, banking sec-tors, 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 min-ing methods and machine learning algorithms. Advantage:reviewing the relevant studies about customer churn analysis observed in the telecom industry. Predicting Customer churn is a busi-ness 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 ana-lyzing 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 preci-sion .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 classi-fied 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 per-formance 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)

10

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)

11

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | 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.

12

CHAPTER 4

SOFTWARE REQUIREMENTS

4.0.1 ANACONDA NAVIGATOR

Anaconda Navigator is a desktop graphical user interface (GUI) included in AnacondaÂo˝ distri-bution 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

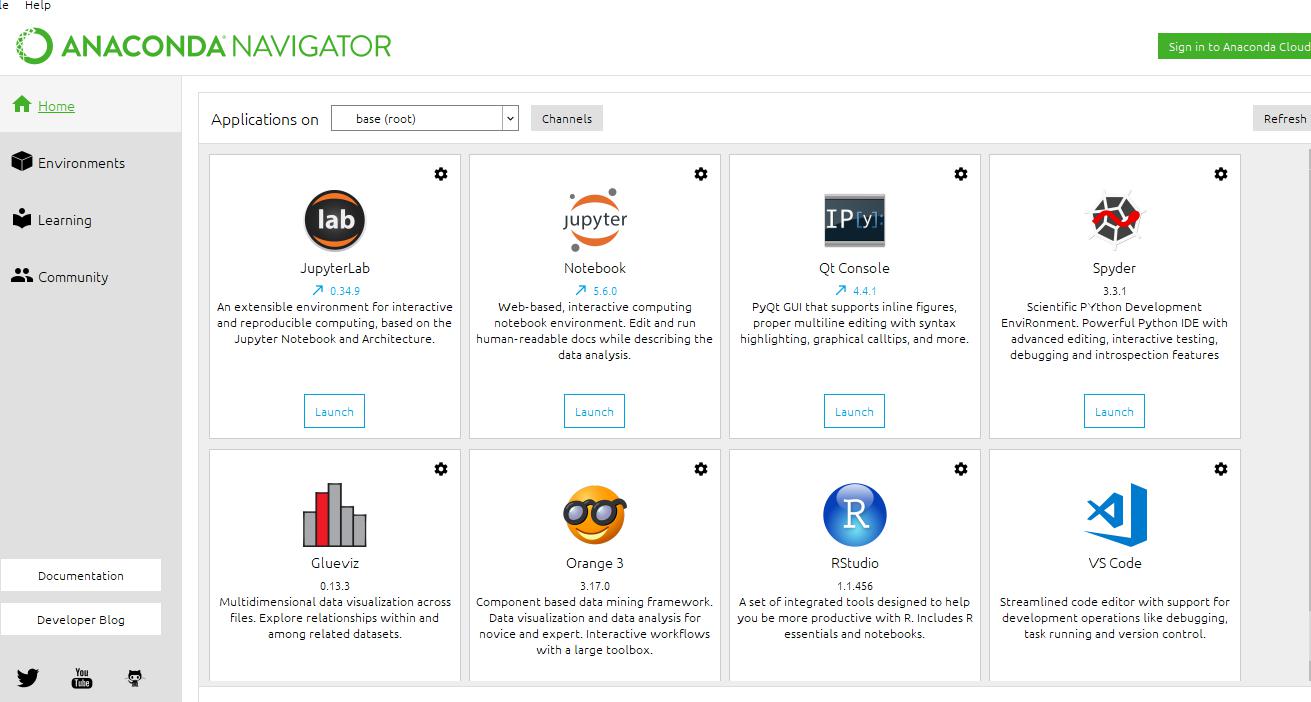


Figure 4.1: anaconda navigator

4.0.2 JUPYTER NOTEBOOK

The Jupyter Notebook is an open-source web application used for creating and sharing docu-ments containing code,visualizations ,texts etc. It is used for data cleaning, transformation, nu-

merical simulation, statistical modeling, data visualization, machine learning, and much more.

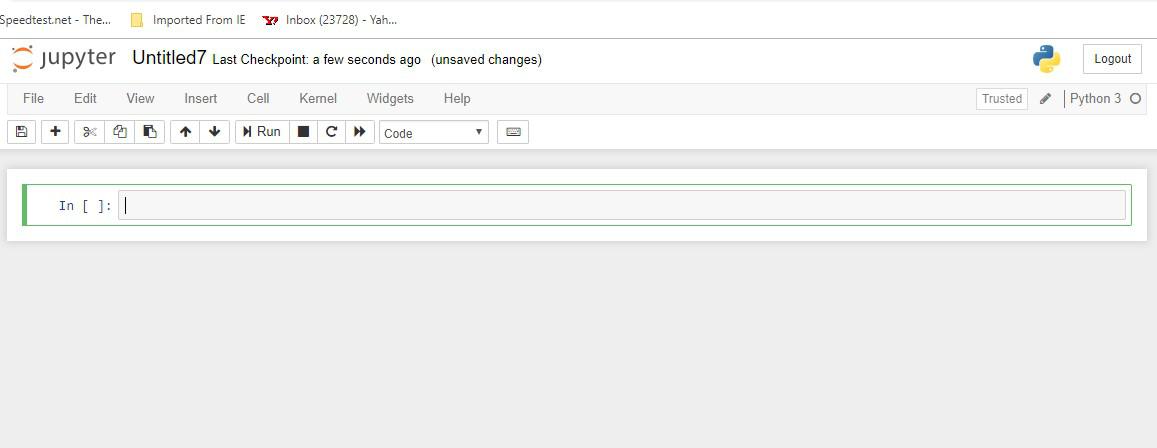


Figure 4.2: jupyter notebook



Figure 4.3: running python 3 on jupyter notebook

14

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 al-gorithms 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 sci-ence 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 churn-ers.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.

1. 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.
2. Burez, J., Van den Poel, D. (2009). Handling class imbalance in customer churn predic-tion. Elsevier , 4626-4636.
3. 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
4. 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.
5. Dash, M., Liu, H. (1997). Feature Selection for Classification. Inrelligent Data Analysis

1 , 131-156.

˘ ˙

[14] A Rawat and A Choubey,âAIA survey on classification techniques in internet environ-

˘ ˙

ment,âAI IJSRSET,vol.2,no.3,2016

1. 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

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["TenureGroup"]; ax = axes[5; 1]) sns:countplot(data["T enureGroup"]; ax = axes[5; 2]); plt:setp(axes; yticks = [])

plt:tightlayout()

f; axes = plt:subplots(ncols = 3; f igsize = (20; 5))

sns:boxplot(x = "Churn"; y = "tenure"; data = data; palette =0 rainbow0 ; ax = axes[0 sns:boxplot(x = "Churn"; y = "MonthlyCharges"; data = data; palette =0 rainbow0 ; ax = axes[1])

sns:boxplot(x = "Churn"; y = "T otalCharges"; data = data; palette =0 rainbow0 ; ax = axes[2])

tempcols = data:drop("SeniorCitizen"; axis = 1)

sns:pairplot(tempcols; hue =0 Churn0 ; palette =0 rainbow0 )

f; axes = plt:subplots(ncols = 2; f igsize = (20; 6))

sns:barplot(x =0 T enureGroup0 ; y =0 MonthlyCharges0 ; data = data; hue = "Churn"; ax = axes[0])

sns:barplot(x =0 T enureGroup0 ; y =0 T otalCharges0 ; 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) :

19

return "23year"

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

return"34year"

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

return"45year"

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

return"56year"

data["T enureGroup"] = data:apply(lambdadata : tenuretogroup(data); axis = 1)

sns:countplot(data["T enureGroup"]);

catcols = [xf orxindata:columnsif data[x]:nunique() < 6andx! = "Churn"]

numcols = [xf orxindata:columnsif data[x]:nunique() > 6andx! = "customerID"]

idcustomer = data["customerID"]

label = data["Churn"]

label = label:apply(lambdax : 1if x == "Y es"else0)

* romsklearn:preprocessingimportMinMaxScaler
* eatureslogtransf ormed = pd:DataF rame(data = data[numcols])
* eatureslogtransf ormed[numcols] = data[numcols]:apply(lambdax : np:log(x + 1)) scaler = MinMaxScaler()
* eatureslogminmaxtransf orm = pd:DataF rame(data = f eatureslogtransf ormed)
* eatureslogminmaxtransf orm[numcols] = scaler:f ittransf orm(f eatureslogtransf ormed[numcols]) sns:heatmap(f eatureslogminmaxtransf orm:corr(); annot = T rue; cmap =0 jet0 );
* eatureslogminmaxtransf orm:drop("tenure"; inplace = T rue; axis = 1)

data:drop(["MonthlyCharges"; "T otalCharges"; "tenure"]; axis = 1; inplace = T rue)

data = pd:concat([data; f eatureslogminmaxtransf orm]; axis = 1)

data:inf o()

data:duplicated():sum()

data:drop("Churn"; inplace = T rue; axis = 1)

data:drop("customerID"; inplace = T rue; axis = 1)

data:inf o()

data = pd:getdummies(data = data; columns = catcols)

data:head

dataoriginal = pd:concat([data; label; idcustomer]; axis = 1)

dataoriginal:inf o()

dataoriginal:head()

A.0.4 Data preprocessing

catcols = [xf orxindata:columnsif data[x]:nunique() < 6andx! = "Churn"]

numcols = [xf orxindata:columnsif data[x]:nunique() > 6andx! = "customerID"]

idcustomer = data["customerID"]

label = data["Churn"]

label = label:apply(lambdax : 1if x == "Y es"else0)

* romsklearn:preprocessingimportMinMaxScaler
* eatureslogtransf ormed = pd:DataF rame(data = data[numcols])
* eatureslogtransf ormed[numcols] = data[numcols]:apply(lambdax : np:log(x + 1)) scaler = MinMaxScaler()

20

f eatureslogminmaxtransf orm = pd:DataF rame(data = f eatureslogtransf ormed)

f eatureslogminmaxtransf orm[numcols] = scaler:f ittransf orm(f eatureslogtransf ormed[numcols]) sns:heatmap(f eatureslogminmaxtransf orm:corr(); annot = T rue; cmap =0 jet0 ); sns:heatmap(f eatureslogminmaxtransf orm:corr(); annot = T rue; cmap =0 jet0 ); data:drop(["MonthlyCharges"; "T otalCharges"; "tenure"]; axis = 1; inplace = T rue)

data = pd:concat([data; f eatureslogminmaxtransf orm]; axis = 1)

data:drop("Churn"; inplace = T rue; axis = 1)

data:drop("customerID"; inplace = T rue; axis = 1)

A.0.5 Evaluating Algorithms

from sklearn.crossvalidationimporttraintestsplit

Xtrain; Xtest; ytrain; ytest = traintestsplit(data; label; testsize = 0:3; randomstate =

42)

print("T rainingsethassamples:":f ormat(Xtrain:shape[0]))

print("T estingsethassamples:":f ormat(Xtest:shape[0]))

f romsklearn:treeimportDecisionT reeClassif ier

* romsklearn:linearmodelimportLogisticRegression
* romsklearn:svmimportSV C
* romxgboostimportXGBClassif ier
* romsklearn:ensembleimportRandomF orestClassif ier
* romsklearn:metricsimportconf usionmatrix
* romsklearn:metricsimportclassif icationreport
* romsklearn:metricsimportrocaucscore; roccurve

def applyclassif ier(clf; xT rain; xT est; yT rain; yT est) :

clf:f it(xT rain; yT rain)

predictions = clf:predict(xT est)

confmtx = conf usionmatrix(yT est; predictions)

f; axes = plt:subplots(ncols = 2; f igsize = (15; 5))

sns:heatmap(confmtx; annot = T rue; cmap =0 tab20c0 ; cbar = F alse; f mt = "g"; ax = axes[0])

axes[0]:setxlabel(0 P redictedlabels0 )

axes[0]:setylabel(0 T ruelabels0 )

axes[0]:settitle(0 Conf usionMatrix0 );

axes[0]:xaxis:setticklabels([0 NotChurn0 ;0 Churn0 ]);

axes[0]:yaxis:setticklabels([0 NotChurn0 ;0 Churn0 ]);

print("Classif icationreport") :

|  |  |
| --- | --- |
| f ormat(classif icationreport(yT est; predictions))) |  |
| rocauc = rocaucscore(yT est; predictions) |  |
| print("AreaunderROCcurve : "; rocauc; "newline") |  |
| f pr; tpr;= roccurve(yT est; predictions) |  |
| axes[1]:plot(f pr; tpr; label = "auc = " + str(rocauc)); |  |
| axes[1]:plot([0; 1]; [0; 1]; color =0 navy0 ; lw = 1; linestyle =0 | 0 ) |

plt:xlim([0:0; 1:0])

plt:ylim([0:0; 1])

plt:xlabel(0 F alseP ositiveRate0 )

21

plt:ylabel(0 T rueP ositiveRate0 )

plt:title(0 Receiveroperatingcharacteristic0 )

plt:legend(loc = "lowerright")

decisiontree = DecisionT reeClassif ier(randomstate = 42); applyclassif ier(decisiontree; Xtrain; Xtest; ytrain; ytest) logisticreg = LogisticRegression(randomstate = 42) applyclassif ier(logisticreg; Xtrain; Xtest; ytrain; ytest) svmmodel = SV C(randomstate = 42)

applyclassif ier(svmmodel; Xtrain; Xtest; ytrain; ytest)

randomf orest = RandomF orestClassif ier(randomstate = 42)

applyclassif ier(randomf orest; Xtrain; Xtest; ytrain; ytest)

T reeparameters = "maxdepth" : [3; 4; 5; 6]; "minsamplesleaf" : [1; 2; 3; 4] LogRegparameters =

"C" : [0:25; 0:5; 0:75; 1:0; 1:5; 2:0; 2:5; 3:0; 4:0; 10:0]; "warmstart" : ["T rue"; "F alse]

SV Mparameters =

"C" : [1:0; 2:0; 3:0];

"cachesize" : [100; 200];

"decisionf unctionshape" : [0 ovo0 ;0 ovr0 ];

"kernel" : [0 sigmoid0 ; "linear"];

"tol" : [0:001; 0:0001]

RandomF orestparameters = "nestimators" : [10; 15; 20; 25; 30]; "criterion" : ["entropy"; "gini"];

applyclassif ier(randomF orestgrid; Xtrain; Xtest; ytrain; ytest)

f romsklearn:ensembleimportAdaBoostClassif ier

model = AdaBoostClassif ier(baseestimator = randomF orestgrid; nestimators = 4)

applyclassif ier(model; Xtrain; Xtest; ytrain; ytest)

* romsklearn:utilsimportresample upsampledata = dataoriginal

majority = upsampledata[upsampledata["Churn"] == 0] minority = upsampledata[upsampledata["Churn"] == 1]

minorityupsampled = resample(minority; replace = T rue; nsamples = 5163; randomstate =

42)

del(upsampledata)

upsampledata = pd:concat([majority; minorityupsampled]) idcustomerupsample = upsampledata["customerID"] labelupsample = upsampledata["Churn"] upsampledata:drop("Churn"; inplace = T rue; axis = 1) upsampledata:drop("customerID"; inplace = T rue; axis = 1)

* romsklearn:crossvalidationimporttraintestsplit

XtrainupS; XtestupS; ytrainupS; ytestupS = traintestsplit(upsampledata; labelupsample; testsize = 0:3; randomstate = 42)

print("T rainingsethassamples:":f ormat(XtrainupS:shape[0]))

print("T estingsethassamples:":f ormat(XtestupS:shape[0]))

model = AdaBoostClassif ier(baseestimator = randomf orest; nestimators = 4)

applyclassif ier(model; XtrainupS; XtestupS; ytrainupS; ytestupS)

22

APPENDIX B

PROJECT WORK

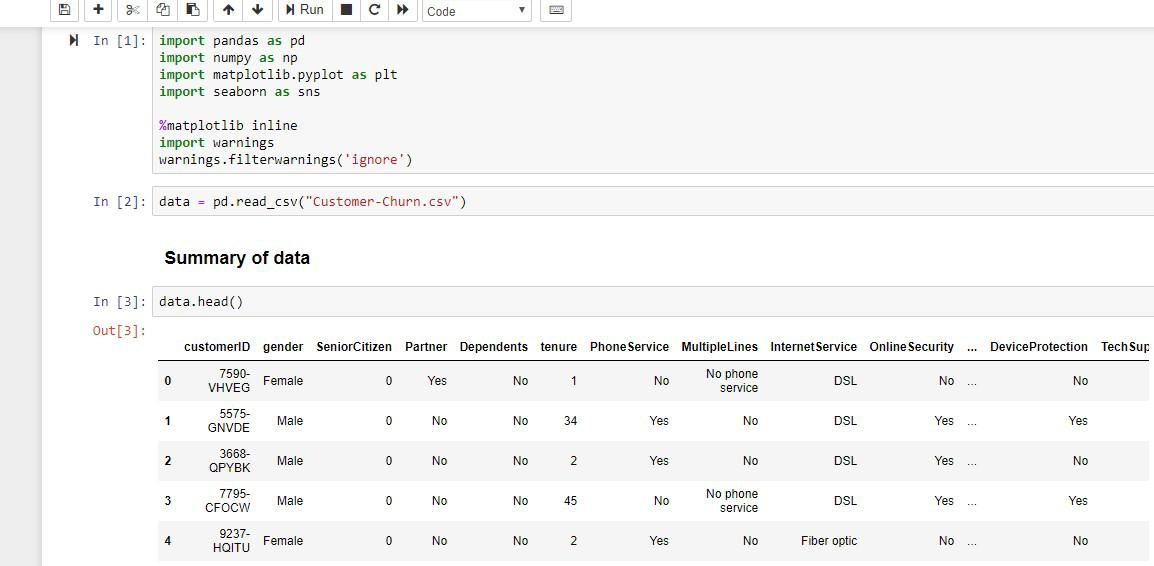


Figure B.1: Importing csv dataset

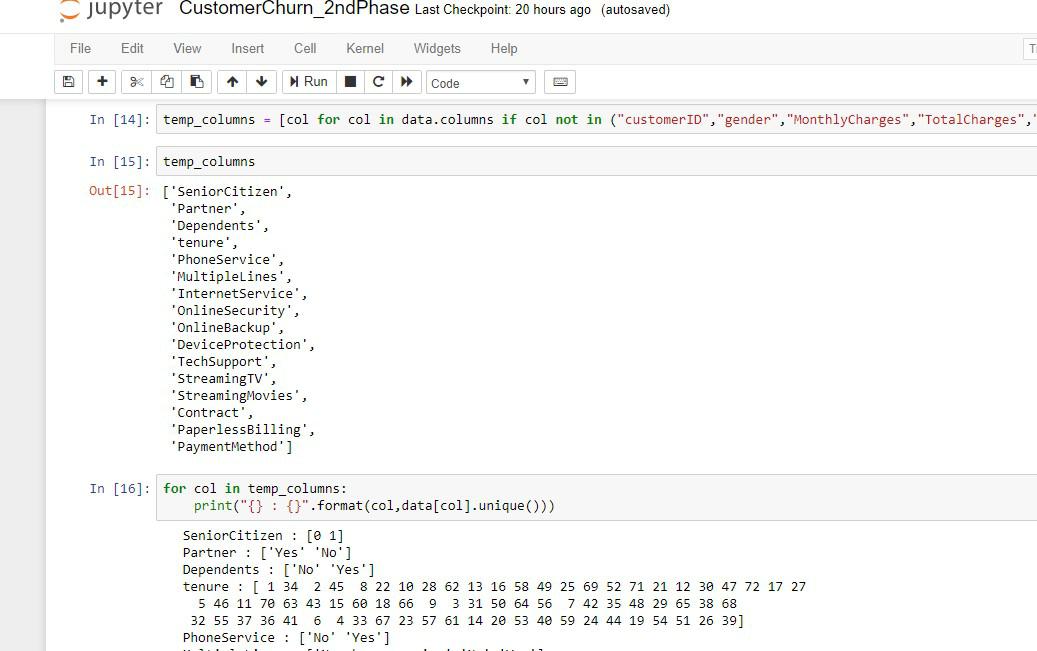


Figure B.2: Customer services

24

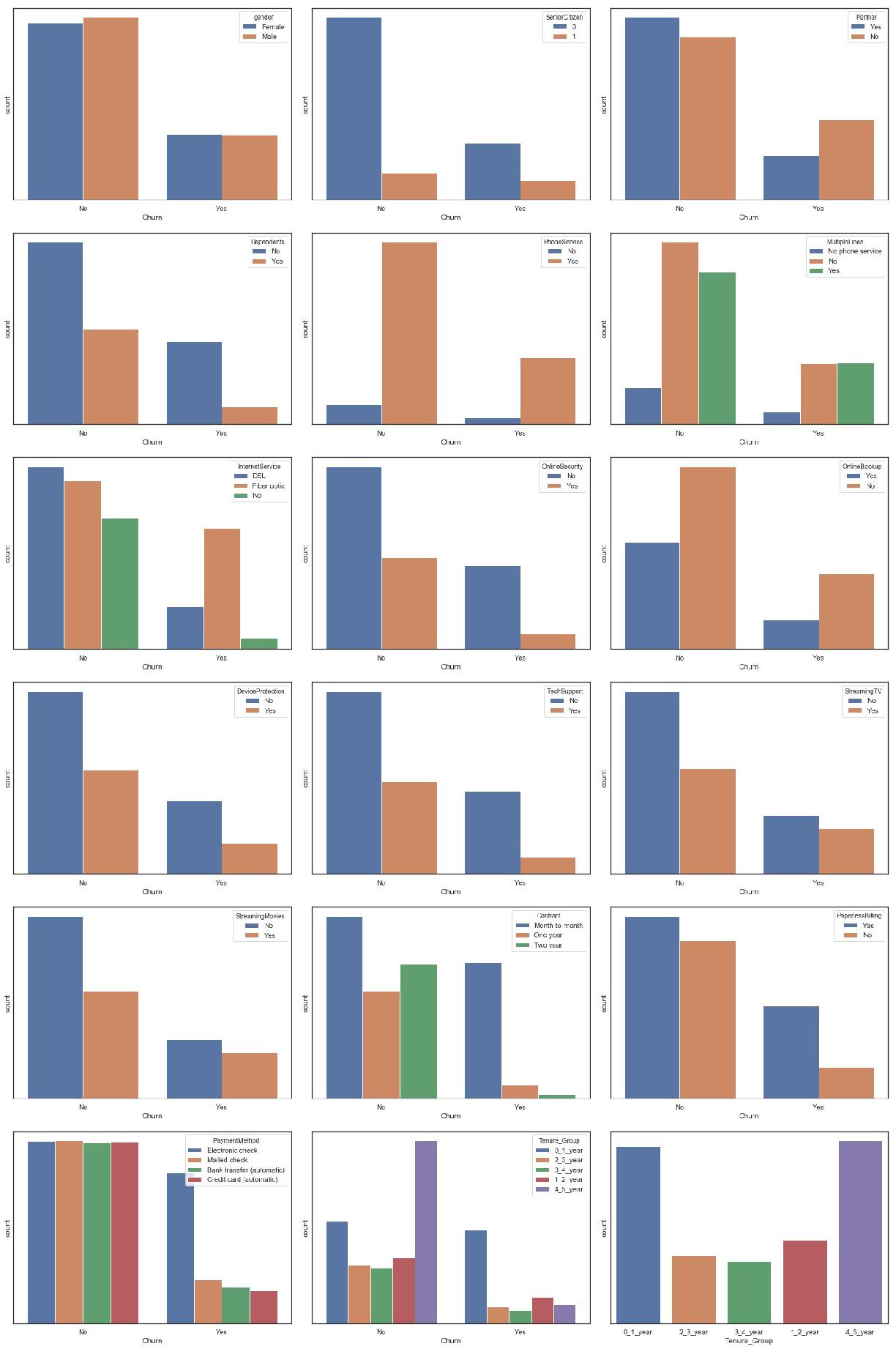


Figure B.3: Data visualization

25

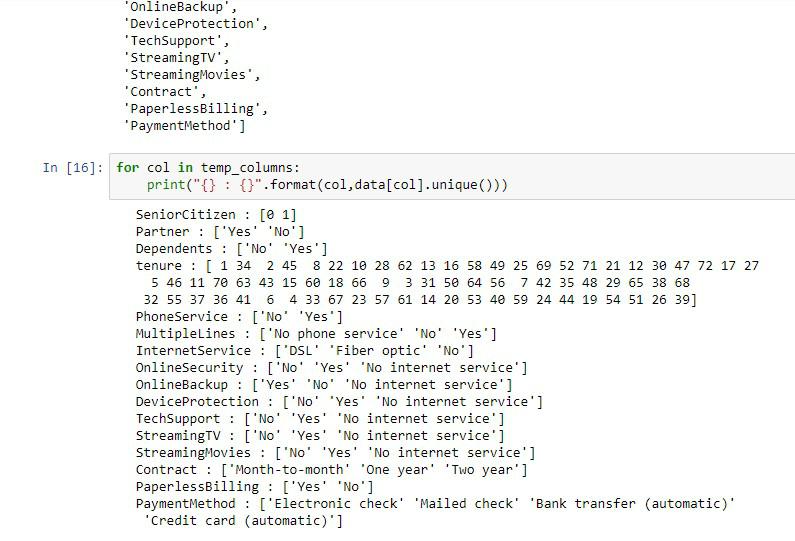


Figure B.4: Feature selection

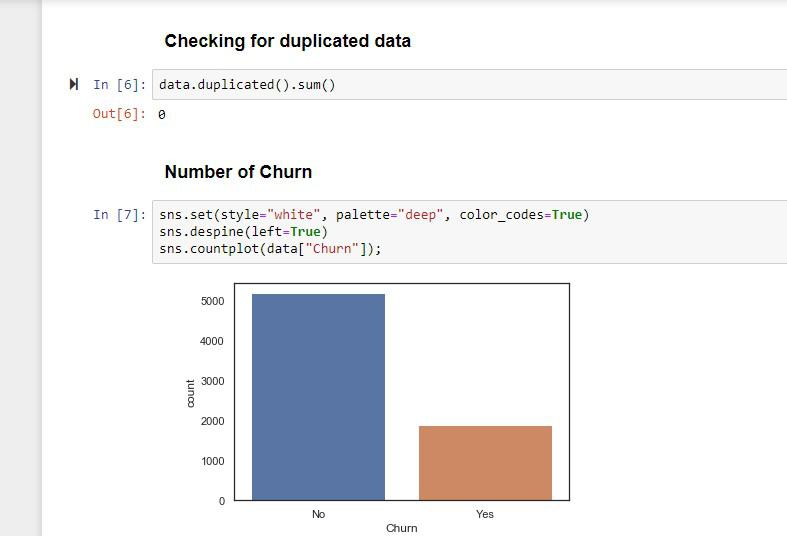


Figure B.5: count vs churn

26

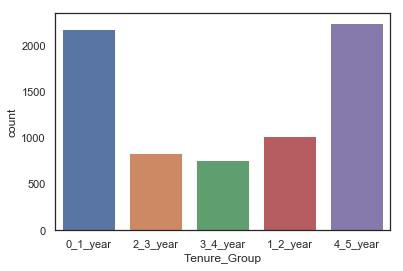


Figure B.6: count vs tenure

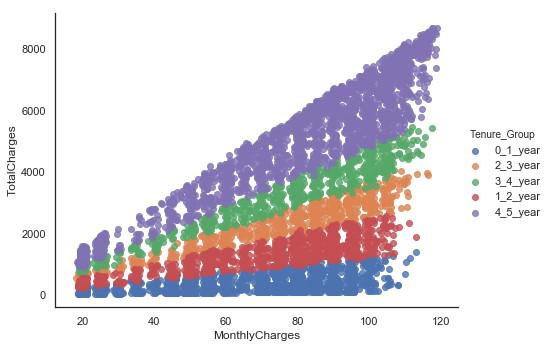


Figure B.7: totalcharges vs monthlycharges for different tenure groups

27

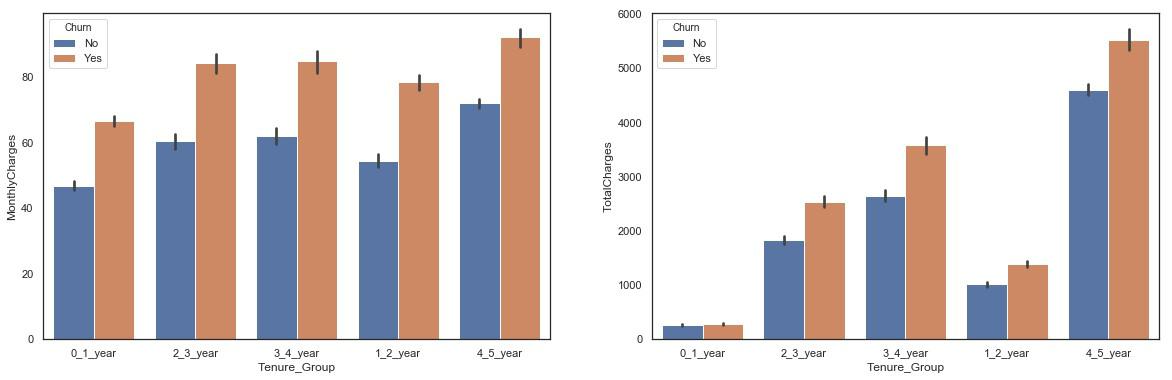


Figure B.8: tenure group visualization

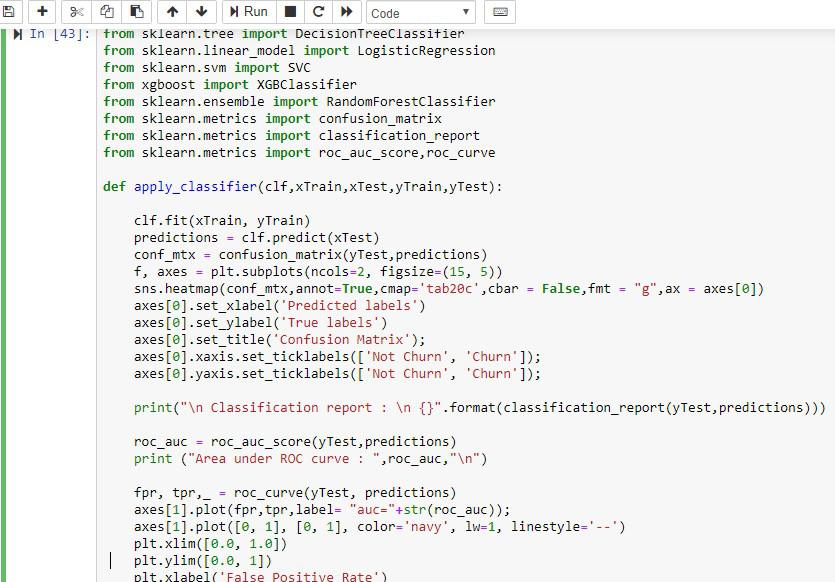


Figure B.9: Evaluating algorithms

28

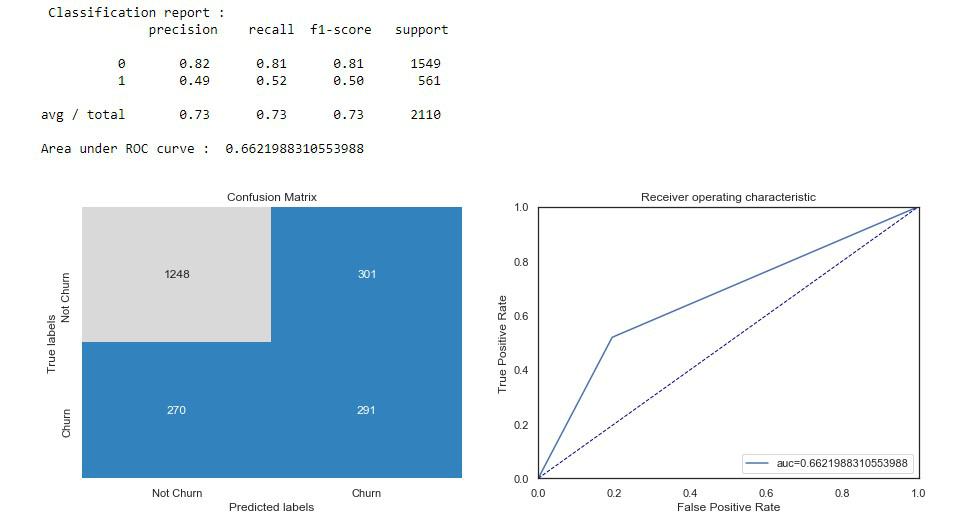


Figure B.10: decision tree performance report

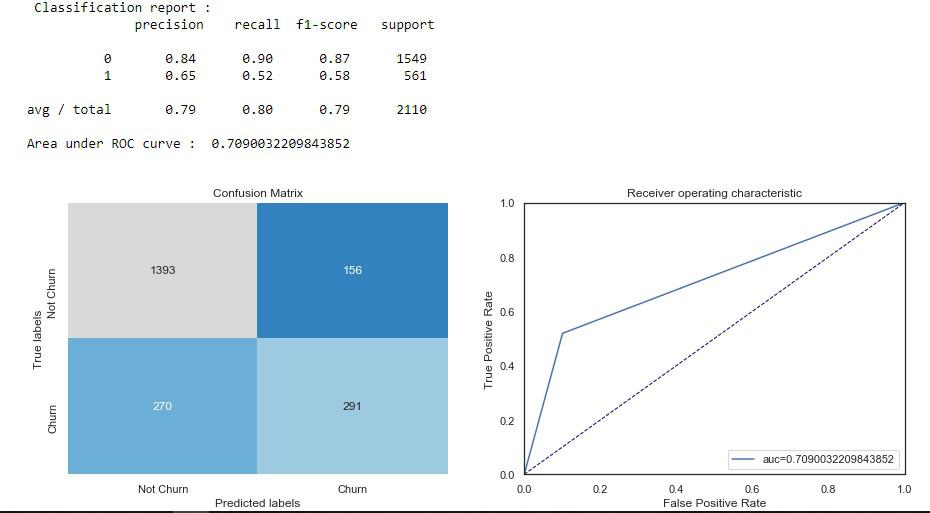


Figure B.11: logistic regression classification report

29

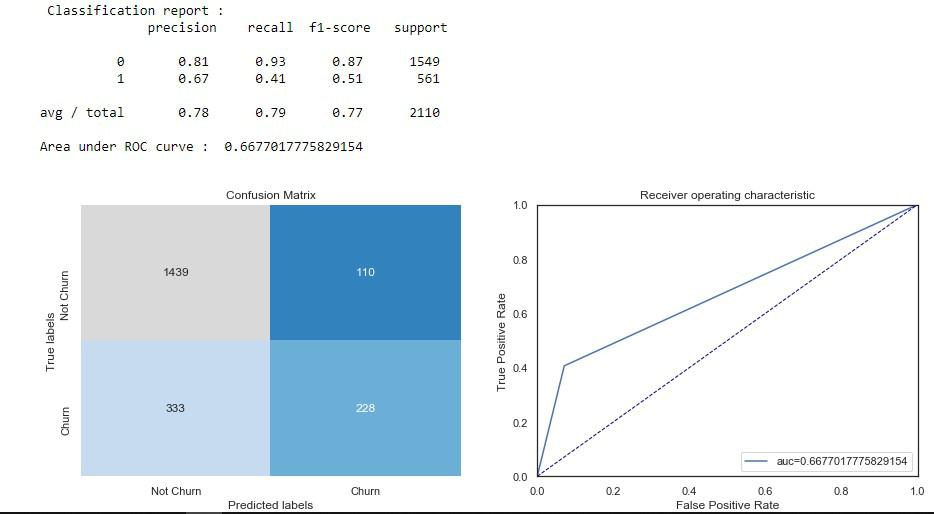


Figure B.12: SVM performance report

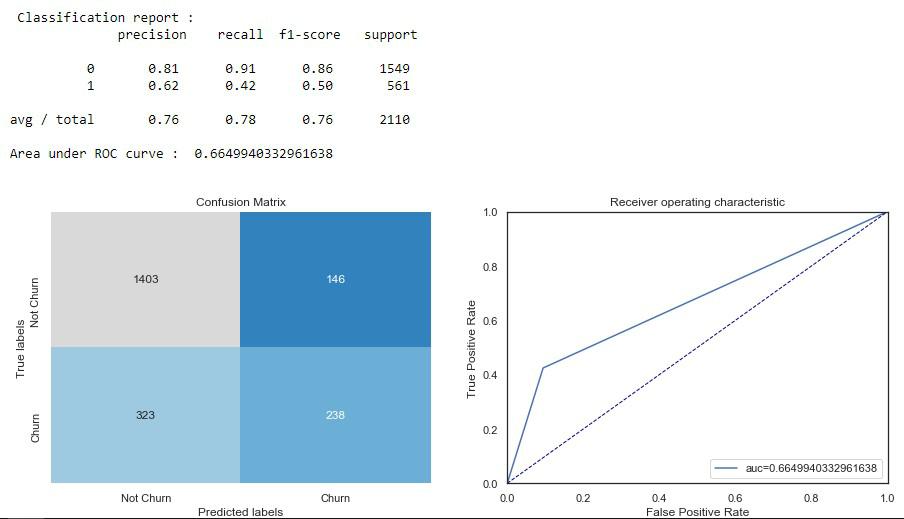


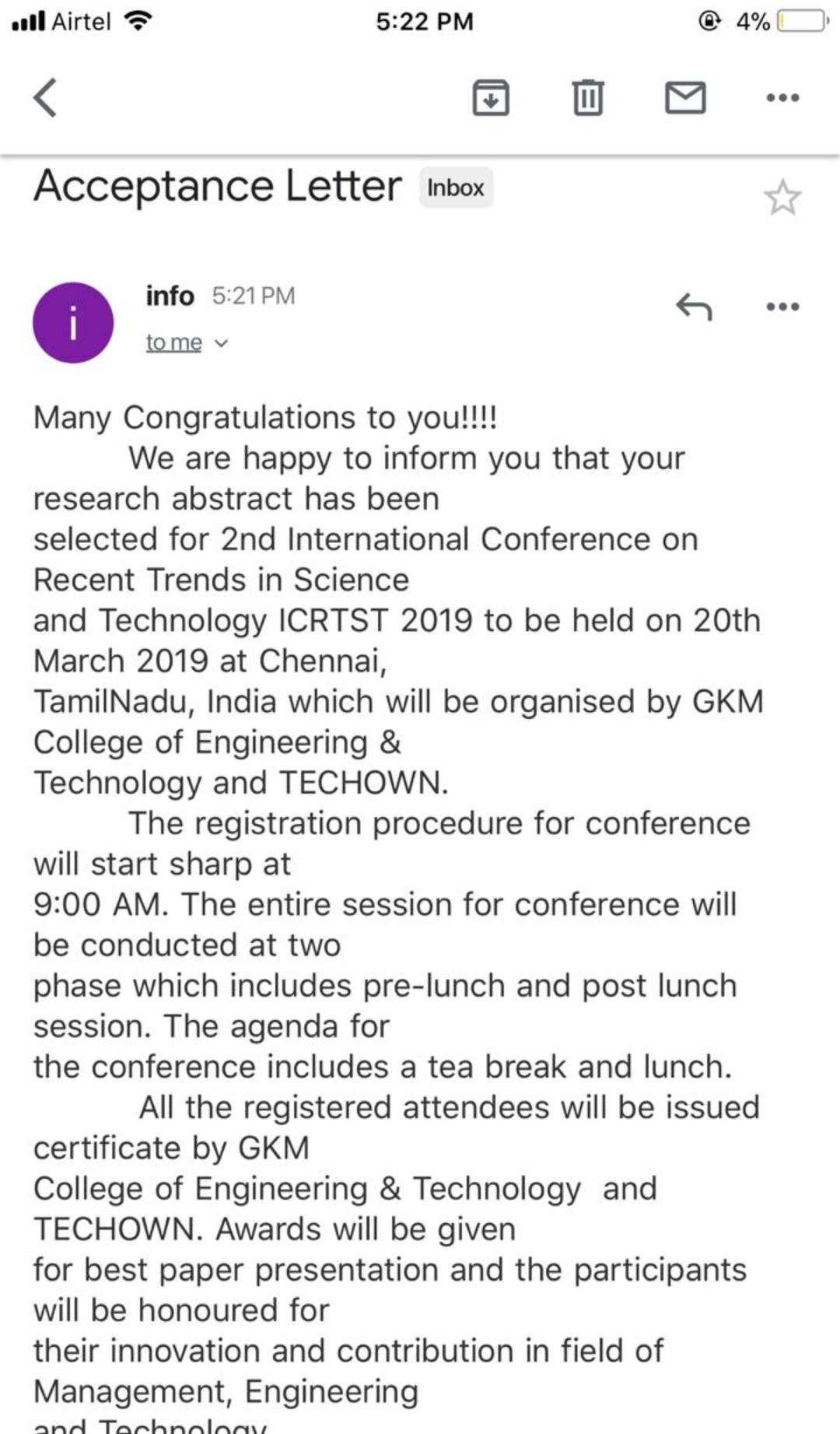
Figure B.13: Random forest performance report

30

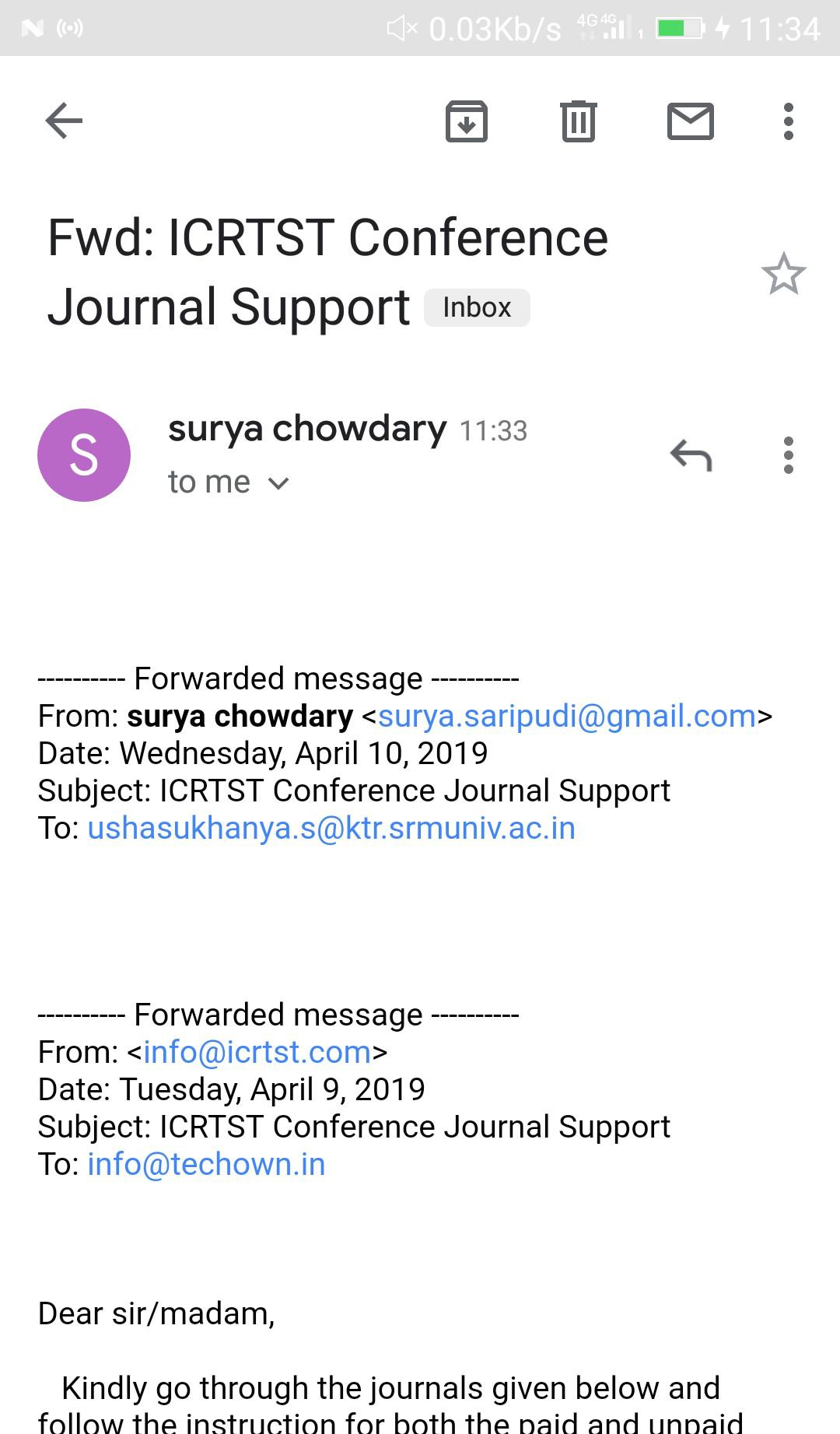
APPENDIX C

PAPER ACCEPTANCE

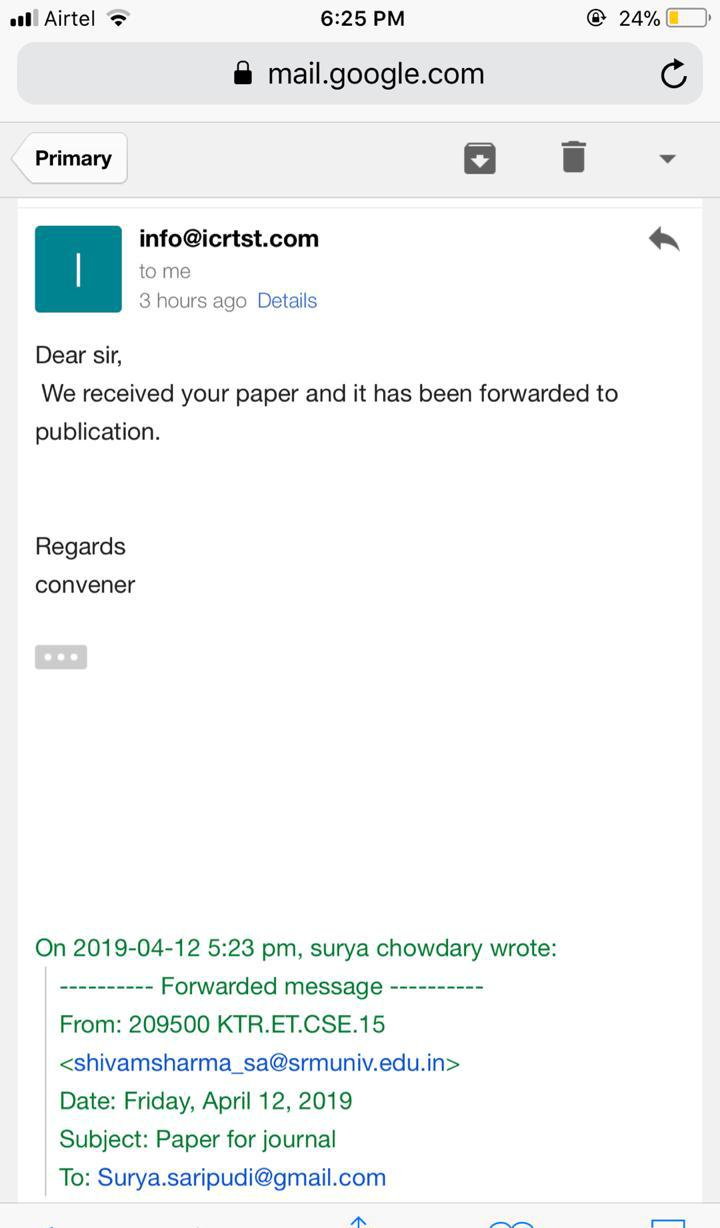
We submitted our paper for 2nd international conference on Recent Trends in Science and Technology ICRTST 2019 .We received the acceptance letter and presented our work at the ICRTST conference held on 20th march 2019 at chennai,tamilnadu, India will was organised by GKM college of engineering and technology 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.



32



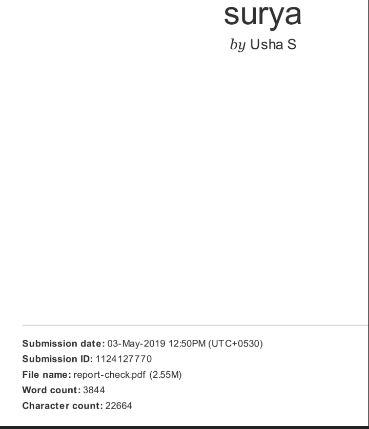
33

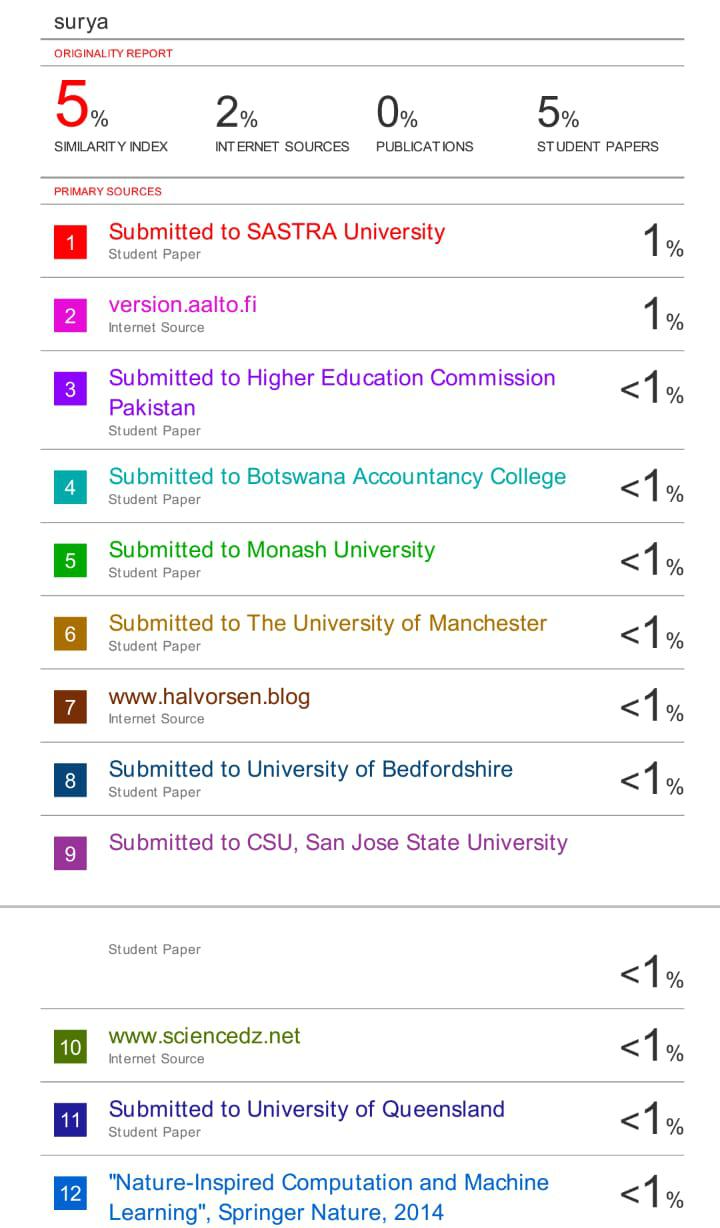


34

APPENDIX D

PLAGIARISM REPORT





36

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 visu-alization 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 chen-nai,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.