A major Project report

on

A Churn Prediction Model in Telecom Sector Using Machine Learning

Project submitted to the Jawaharlal Nehru Technological University in Partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering

Submitted By

ADDLA MADHURI

16891A0562

SAI VAMSHI CHINTAKUNTA

16891A0574

YANALA UDAY KIRAN REDDY

16891A05B5

Under the Guidance of

Mrs. M. Swathi

Assistant Professor

Department of Computer Science and Engineering

VIGNAN INSTITUTE OF TECHNOLOGY AND SCIENCE, Deshmukhi

Affiliated to Jawaharlal Nehru Technological University. Hyderabad

2020

A Project report

on

A Churn Prediction Model in Telecom Sector Using Machine Learning

Project submitted to the Jawaharlal Nehru Technological University in Partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering

Submitted By

ADDLA MADHURI

16891A0562

SAI VAMSHI CHINTAKUNTA

16891A0574

YANALA UDAY KIRAN REDDY

16891A05B5

Under the Guidance of

Mrs. M. Swathi

Assistant Professor

Department of Computer Science and Engineering

VIGNAN INSTITUTE OF TECHNOLOGY AND SCIENCE, Deshmukhi

Affiliated to Jawaharlal Nehru Technological University. Hyderabad

2020

VIGNAN INSTITUTE OF TECHNOLOGY AND SCIENCE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the seminar report titled A CHURN PREDICTION MODEL IN TELECOM SECTOR USING MACHINE LEARNING is being submitted by ADDLA MADHURI, bearing 16891A0562, SAI VAMSHI CHNTAKUNTA, bearing 16891A0574**,** YANALA UDAY KIRAN REDDY, bearing 16891A05B5 in IV B. Tech II semester *Computer Science and Engineering* is a record bonafide work carried out by them. The results embodied in this report have not been submitted to any other University for the award of any degree.

Internal Guide **Head of the Department**

Mrs. M. Swathi **Mr.G.Raja Vikram**

PROJECT EVALUATION CERTIFICATE

This is to certify that the Project work entitled “A CHURN PREDICTION MODEL IN TELECOM SECTOR USING MACHINE LEARNING” submitted by ADDLA MADHURI 16891A0562, SAI VAMSHI CHNTAKUNTA 16891A0574**,** YANALA UDAY KIRAN REDDY 16891A05B5, has been examined and adjudged as sufficient for the partial fulfillment of the requirement of the degree of Bachelor of Technology in Computer Science and Engineering of Jawaharlal Nehru Technological University, Hyderabad.

External Examiner : ____________________________

(Signature with Date)

Internal Examiner : ____________________________

(Signature with Date)

Head of the Department : ____________________________

(Signature with Date)

ACKNOWLEDGEMENT

Success will be crowned to people who made it reality but the people whose constant guidance and encouragement made it possible will be crowned first on the eve of success.

This acknowledgement transcends the reality of formality when we would like to express deep gratitude and respect to all those people behind the screen who guided, inspired and helped us for the completion of our project work.

We would like to express our thankfulness towards **Dr.G.Durga Sukumar**, Principal, **Mr. G. Raja Vikram**, Head of the Department and **Mrs. M. Swathi**, Assistant Professor our mini project guide for giving us all the facilities, ways and means by which we were able to complete the mini project. We express our sincere gratitude to him for his constant support and valuable suggestions without which the project would not have been possible.

We thank for **Mrs. M. Swathi**, Assistant Professor for her constant supervision, guidance and boundless co-operation throughout the project. We also extend our thanks to all the staff of the Department of Computer Science and Engineering, VITS for their co-operation and support during our course work.

Finally, we would like to thank our friends and batch-mates for their co-operation to complete this project successfully.

ADDLA MADHURI (16-562)

SAI VAMSHI CHINTAKUNTA (16-574)

YANALA UDAY KIRAN REDDY (16-5B5)

DECLARATION

We declare that this written submission represents our ideas in our own words and where others ideas or words have been included; we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

ADDLA MADHURI (16-562)

SAI VAMSHI CHINTAKUNTA (16-574)

YANALA UDAY KIRAN REDDY (16-5B5)

CONTENTS

ABSTRACT

In the telecom sector, a huge volume of data is being generated on a daily basis due to a vast client base. Decision makers and business analysts emphasized that attaining new customers is costlier than retaining the existing ones. Business analysts and customer relationship management (CRM) analyzers need to know the reasons for churn customers, as well as, behavior patterns from the existing churn customers' data. This paper proposes a churn prediction model that uses classification, as well as, clustering techniques to identify the churn customers and provides the factors behind the churning of customers in the telecom sector. Feature selection is performed by using information gain and correlation attribute ranking filter. The proposed model first classifies churn customers data using classification algorithms, in which the Random Forest (RF) algorithm performed well with 88.63% correctly classified instances. Creating effective retention policies is an essential task of the CRM to prevent churners. After classification, the proposed model segments the churning customer's data by categorizing the churn customers in groups using cosine similarity to provide group-based retention offers. This paper also identified churn factors that are essential in determining the root causes of churn. By knowing the significant churn factors from customers' data, CRM can improve productivity, recommend relevant promotions to the group of likely churn customers based on similar behavior patterns, and excessively improve marketing campaigns of the company. The proposed churn prediction model is evaluated using metrics, such as accuracy, precision, recall, f-measure, and receiving operating characteristics (ROC) area. The results reveal that our proposed churn prediction model produced better churn classification using the RF algorithm and customer profiling using k-means clustering. Furthermore, it also provides factors behind the churning of churn customers through the rules generated by using the attribute-selected classifier algorithm.

LIST OF FIGURES

LIST OF ABBREVATIONS

CRM	customer relationship management
RF	Random Forest
ROC	receiving operating characteristics
2G	2nd Generation
3G	3rd Generation
4G(LTE)	4th Generation (Long Term Evolution)
WEKA	Waikato Environment for Knowledge Analysis
LR	Logistic Regression
DT	Decision Tree
SVM	Support Vector Machine
NV	Naïve Bayesian
KNN	K-nearest neighbor
ANN	Artificial Neural Networks
NIC	National Informatics Centre
IDLE	Integrated Development Environment
OS	Operating System
RAM	Random Access Memory
UML	Unified Modeling Language
OO Tools	Object Oriented Tools
SQL	Structured Query Language
CAD	Computer Aided Design
2,3,4D	2,3,4 Dimensions
GIMP	GNU Image Manipulation Program
GNU	General Public License
GUI	Graphical User Interface
CSV	Comma Separated Values
JSON	JavaScript Object Notation
API	Application programming interface

CHAPTER 1

INTRODUCTION

In the present world, a huge volume of data is being generated by telecom companies at an exceedingly fast rate. There is a range of telecom service providers competing in the market to increase their client share. Customers have multiple options in the form of better and less expensive services. The ultimate goal of telecom companies is to maximize their profit and stay alive in a competitive market place A customer churn happens when a vast percentage of clients are not satisfied with the services of any telecom company. It results in service migration of customers who start switching to other service providers. There are many reasons for churning. Unlike postpaid customers, prepaid customers are not bound to a service provider and may churn at any time. Churning also impacts the overall reputation of a company which results in its brand loss. A loyal customer, who generates high revenue for the company, gets rarely affected by the competitor companies. Such customers maximize the profit of a company by referring it to their friends, family members and colleagues. Telecom companies consider policy shift when the number of customers drops below a certain level which may result in a huge loss of revenue.

Churn prediction is vital in the telecom sector as telecom operators have to retain their valuable customers and enhance their Customer Relationship Management (CRM) administration. The most challenging job for CRM is to retain existing customers. Due to the saturated and competitive market, customers have the option to switch to other service providers. Telecom companies have developed procedures to identify and retain their customers as it is less expensive than attracting the new ones. This is due to the cost involved in advertisements, workforce, and concessions which can scale up to almost five to six times than retaining existing customers. Small attention is needed for identifying the existing churn customers, which can help in overturning the situation. The requirement of retaining customers needs to develop an accurate and high-performance model for identifying churn customers. The proposed model should have the capability to identify churn customers and then find the reasons behind churn to avoid loss of customers and provide measures to retain them. In addition, it should employ techniques to predict when such a situation is going to arise in the future.

1.1 MOTIVATION

In order to continue life-sustaining competitive advantage, many organizations focus on maximizing the marketing relationship with their customer lifetime value and customer churn management. In fact, more organizations are realizing that their most valuable resource is their current customer base. In the present study are to go through a database collected from 300 customers, including an insurance company in Iran has been used. In order to check the model presented with a desire to review a decision tree classification methods. Bayesian networks and neural networks will be paid with respect to sample. Survey results can help managers, marketers in this arena is in various industries. Reduction strategies appropriate to offer in this field.

1.2 PROBLEM DEFINATION

In a business environment, the term, customer attrition simply refers to the customers leaving one business service to another. Customer churn or subscriber churn is also similar to attrition, which is the process of customers switching from one service provider to another anonymously. From a machine learning perspective, churn prediction is a supervised problem defined as follows: Given a predefined forecast horizon, the goal is to predict the future churners over that horizon, given the data associated with each subscriber in the network. Churn Prediction is a phenomenon which is used to identify the possible churners in advance before they leave the network. This helps the CRM department to prevent subscribers who are likely to churn in future by taking the required retention policies to attract the likely churners and to retain them. Thereby, the potential loss of the company could be avoided. The input for this problem includes the data on past calls for each mobile subscriber, together with all personal and business information that is maintained by the service provider. In addition, for the training phase, labels are provided in the form of a list of churners. After the model is trained with highest accuracy, the model must be able to predict the list of churners from the real dataset which does not include any churn label. In the perspective of knowledge discovery process, this problem is categorized as predictive mining or predictive modeling.

1.3 OBJECTIVES

The objective of this contest is to predict customer churn. We are providing you a public dataset that has customer usage pattern and if the customer has churned or not. We expect you to develop an algorithm to predict the churn score based on usage pattern. The predictors provided are as follows:

1.4 LIMITATIONS

The first limitation we would like to address is the lack of coverage data per customer. We were only able to calculate the coverage at home for each customer. Loss of coverage for each customer is impossible to measure from the network side. Having adequate coverage information could have improved our model. However, the Ratio between 2G and 3G data events does imply the influence of loss of 3G coverage or insufficient 3G capacity in certain areas onto churn. Other limitations of this research are of legal nature. Namely, in most European countries stringent Data Privacy Acts or Net Neutrality Laws exist. This makes it impossible to look into individual consumption of different types of Internet use (e.g. browsing, streaming, messaging, etc), which could provide even better insights into what type of service degradation leads to churn.

Next, as usage patterns change, so do the expectations from the service quality that the network provides. Therefore, in time we expect a change in the influence on churn of the various factors that we discussed which makes the model outdated. This will especially be the case after the introduction of 4G (LTE) networks, which allow much faster Internet speed. However, these issues can be addressed by remodeling.

CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

Customer churn has become a major problem within a customer centered banking industry and banks have always tried to track customer interaction with the company, in order to detect early warning signs in customer's behavior such as reduced transactions, account status dormancy and take steps to prevent churn. This paper presents a data mining model that can be used to predict which customers are most likely to churn. The study used real-life customer records provided by a major Nigerian bank. The raw data was cleaned, pre-processed and then analyzed using WEKA, a data mining software tool for knowledge analysis. Simple K-Means was used for the clustering phase while a rule-based algorithm, JRip was used for the rule generation phase. The results obtained showed that the methods used can determine patterns in customer behaviors and help banks to identify likely churners and hence develop customer retention modalities.

Machine-learning techniques have been widely used for evaluating the probability of customer to churn. Based on a survey of the literature in churn prediction, the techniques used in the bulk of literatures fall into one of the following categories

1. Regression analysis.
2. Tree – based.
3. Support Vector Machine.
4. Bayesian algorithm.
5. Ensemble learning.
6. Sample – based learning.
7. Artificial neural network.

1) Regression analysis:

Regression analysis techniques aim mainly to investigate and estimate the relationships among a set of features. Regression includes many models for analyzing the relation between one target/response variable and a set of independent variables. Logistic Regression (LR) is the appropriate regression analysis model to use when the dependent variable is binary. LR is a predictive analysis used to explain the relationship between a dependent binary variable and a set of independent variables. For customer churn, LR has been widely used to evaluate the churn probability as a function of a set of variables or customers' features.

2) Decision Tree:

Decision Tree (DT) is a model that generates a tree-like structure that represents set of decisions. DT returns the probability scores of class membership. DT is composed of: internal Nodes: each node refers to a single variable/feature and represents a test point at feature level; branches, which represent the outcome of the test and are represented by lines that finally lead to leaf Nodes which represent the class labels. That is how decision rules are established and used to classify new instances. DT is a flexible model that supports both categorical and continuous data. Due to their flexibility they gained popularity and became one of the most commonly used models for churn prediction for new instances based on the analysis of their ancestors.

3) Support Vector Machine:

Support Vector Machine (SVM) is a supervised learning technique that performs data analysis in order to identify patterns. Given a set of labeled training data, SVM represents observations as points in a high dimensional space and tries to identify the best separating hyper planes between instances of different classes. New instances are represented in the same space and are classified to a specific class based on their proximity to the separating gap. For churn prediction, SVM techniques have been widely investigated and evaluated to be of high predictive performance.

4) Bayes Algorithm:

Bayes algorithm estimates the probability that an event will happen based on previous knowledge of variables associated with it. Naïve Bayesian (NB) is a classification technique that is based on Bayes' theorem. It adopts the idea of complete variables independence, as the presence/absence of one feature is unrelated to the presence/absence of any other feature. It considers that all variables independently contribute to the probability that the instance belongs to a certain class. NB is a supervised learning technique that bases its the proximity of his features to the customers in each classes.

5) Instance – based learning:

Also known as memory based learning, new instances are labeled based on previous instances stored in memory. The most widely used instance based learning techniques for classification is K-nearest neighbor (KNN). KNN does not try to construct an internal model and computations are not performed until the classification time. KNN only stores instances of the training data in the features space and the class of an instance is determined based on the majority votes from its neighbors. Instance is labeled with the class most common among its neighbors. KNN determine neighbors based on distance using Euclidian, Manhattan or Murkowski distance measures for continuous variables and hamming for categorical variables. Calculated distances are used to identify a set of training instances (k) that are the closest to the new point, and assign label from these. Despite its simplicity, KNN have been applied to various types of applications.

6) Ensemble – based Learning:

Ensemble based learning techniques produce their predictions based on a combination of the outputs of multiple classifiers. Ensemble learners include bagging methods (i.e. Random Forest) and boosting methods (i.e. Ada Boost, stochastic gradient boosting). Random Forest (RF) are an ensemble learning technique that can support classification and regression. It extends the basic idea of single classification tree by growing many classification trees in the training phase. To classify an instance, each tree in the forest generates its response, the model chooses the class that has receive the most votes over all the trees in the forest. One major advantage of RF over traditional decision trees is the protection against over fitting which makes the model able to deliver a high performance.

7) Artificial neural network:

Artificial Neural Networks (ANNs) are machine-learning techniques that are inspired by the biological neural network in human brain. ANNs are adaptive, can learn by example, and are fault tolerant. An ANN is composed of a set of connected nodes (neurons) organized in layers. The input layer communicates with one or more hidden layers, which in turn communicates with the output layer. Layers are connected by weighted links. Those links carry signals between neurons usually in the form of a real number. The output of each neuron is a function of the weighted sum of all its inputs. The weights on connection are adjusted during the learning phase to represent the strengths of connections between nodes. ANN can address complex problems, such as the churn prediction problem.

2.2 EXISTING SYSTEM

Earlier it was predicted only for banking sector customers' dataset also there was one algorithm used for prediction purpose at a time. Accuracy wise it wasn't accurate. However, it is not possible to extract the useful information hidden in these data sets, unless they are processed properly. In order to find out this hidden information, various analyses should be performed using data mining, which consists of numerous methods.

2.3 DISADVANTAGES OF EXISTING SYSTEM

1. We predicted only banking sector customers' dataset.
2. In today's technological conditions, new data are being produced by different sources in many sectors.
3. However, it is not possible to extract the useful information hidden in these data sets, unless they are processed properly.
4. In order to find out this hidden information, various analyses should be performed using data mining, which consists of numerous methods.

2.4 PROPOSED SYSTEM

The proposed churn prediction model is evaluated using metrics, such as accuracy, precision, recall, f-measure, and receiving operating characteristics (ROC) area. The results reveal that our proposed churn prediction model produced better churn classification. For classification purpose we are using multiple machine learning algorithms like Decision tree. Random Forest, Linear Regression (LR), Support Vector Machine (SVM), Xgboost and Adaboost classifiers and then we compare the results with highest accuracy classifier. Hence we improvised the accuracy and performance

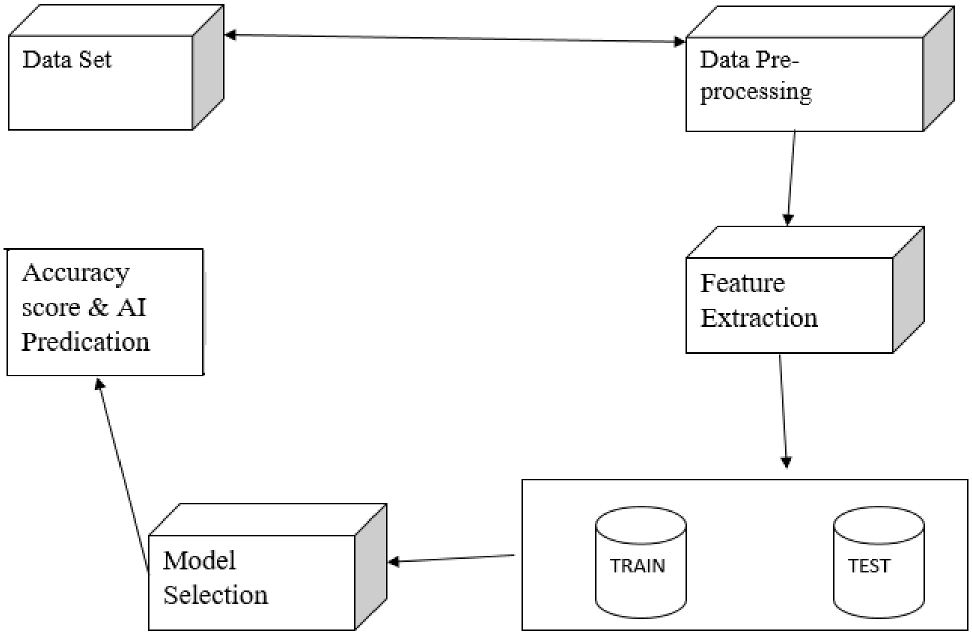


Figure 2.1 Proposed System Architecture

1. DATA COLLECTION

In Data Collection, We have some raw data related to the telecommunication Sector that can be used for analysis. This step is concerned with selecting the subset of all available data that you will be working with. So, this is one of the important steps as we need to consider all the useful subsets from the available subsets in order to get the efficient output. This step is the collection of both the trained data and the Test data that is used in the analysis.

2. DATA PRE-PROCESSING

As we know that the given data is the raw data it is important to Pre-Process the given data. So, we use some pre-processing techniques to process the data. So, we need to consider some effective methods like data cleaning, data reducing and data wrangling. So, output of this step is exactly processed data that is given to some machine learning algorithm to analyze the data. Here, the data is given as a input for the feature extraction.

3. FEATURE EXTRACTION

Next thing is to do Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally, our models are trained using Classifier algorithm. So, in the Feature Extraction we use some machine learning algorithms like Random Forest and linear regression and check the feature obtained from them. We will also use Support vector machine and Xgboost and Adaboost classifiers to analyze the complex data and to get more accuracy.

4. EVALUATION MODEL

Finally we will try to evaluate the given data and try to predict against the given model. This evaluation is the final task and we will see the output in the form of accuracy and we find the churn in the accuracy. So, in this model more is the accuracy more effective is the evaluation model and churn can be easily obtained. Here the Churn is used to obtain customer dissatisfaction, cheaper and/or better offers from the competition, more successful sales and/or marketing by the competition, or reasons having to do with the customer life cycle.

2.5 CONCLUSION

The achieved conclusions can advise the machine learning users which classifier and feature selection method to use to optimize the classification accuracy, which may be important especially in risk-sensitive applications of Machine Learning (e. g. medicine, business decisions, and control applications) as well as in the aim to reduce costs of collecting, processing and storage of unnecessary data.

CHAPTER 3

ANALYSIS

3.1 INTRODUCTION

Preliminary investigation examine project feasibility, the likelihood the system will be useful to the organization. The main objective of the feasibility study is to test the Technical, Operational and Economical feasibility for adding new modules and debugging old running system. All system is feasible if they are unlimited resources and infinite time. There are aspects in the feasibility study portion of the preliminary investigation:

1. Technical Feasibility

2. Operational Feasibility

3. Economical Feasibility

3.1.1 TECHNICAL FEASIBILITY

The technical issue usually raised during the feasibility stage of the investigation

Includes the following:

- Does the necessary technology exist to do what is suggested?
- Do the proposed equipment's have the technical capacity to hold the data required to use the new system?
- Will the proposed system provide adequate response to inquiries, regardless of the number or location of users?
- Can the system be upgraded if developed?
- Are there technical guarantees of accuracy, reliability, ease of access and data security?

Earlier no system existed to cater to the needs of 'Secure Infrastructure Implementation System'. The current system developed is technically feasible. It is a web based user interface for audit workflow. Thus it provides an easy access to the users. The database's purpose is to create, establish and maintain a workflow among various entities in order to facilitate all concerned users in their various capacities or roles. Permission to the users would be granted based on the roles specified. Therefore, it provides the technical guarantee of accuracy, reliability and security. The software and hard requirements for the development of this project are not many and are already available in-house or are available as free as open source. The work for the project is done with the current equipment and existing software technology. Necessary bandwidth exists for providing a fast feedback to the users irrespective of the number of users using the system.

3.1.2 OPERATIONAL FEASIBILITY

Proposed projects are beneficial only if they can be turned out into information system. That will meet the organization's operating requirements. Operational feasibility aspects of the project are to be taken as an important part of the project implementation. Some of the important issues raised are to test the operational feasibility of a project includes the following: -

- Is there sufficient support for the management from the users?
- Will the system be used and work properly if it is being developed and implemented?
- Will there be any resistance from the user that will undermine the possible application benefits?

This system is targeted to be in accordance with the above-mentioned issues. Beforehand, the management issues and user requirements have been taken into consideration. So there is no question of resistance from the users that can undermine the possible application benefits.

The well-planned design would ensure the optimal utilization of the computer resources and would help in the improvement of performance status.

3.1.3 ECONOMIC FEASIBILITY

A system can be developed technically and that will be used if installed must still be a good investment for the organization. In the economical feasibility, the development cost in creating the system is evaluated against the ultimate benefit derived from the new systems. Financial benefits must equal or exceed the costs.

The system is economically feasible. It does not require any addition hardware or software. Since the interface for this system is developed using the existing resources and technologies available at NIC, There is nominal expenditure and economical feasibility for certain.

3.2 SOFTWARE SPECIFICATIONS

3.2.1 SOFTWARE REQUIREMENTS

1. Python IDLE
2. Anaconda – Jupyter Notebook

3.2.2 HARDWARE REQUIREMENTS

1. OS – Windows 7,8 or 10 (32 or 64 bit)
2. RAM – 4GB

3.3 FLOW CHART

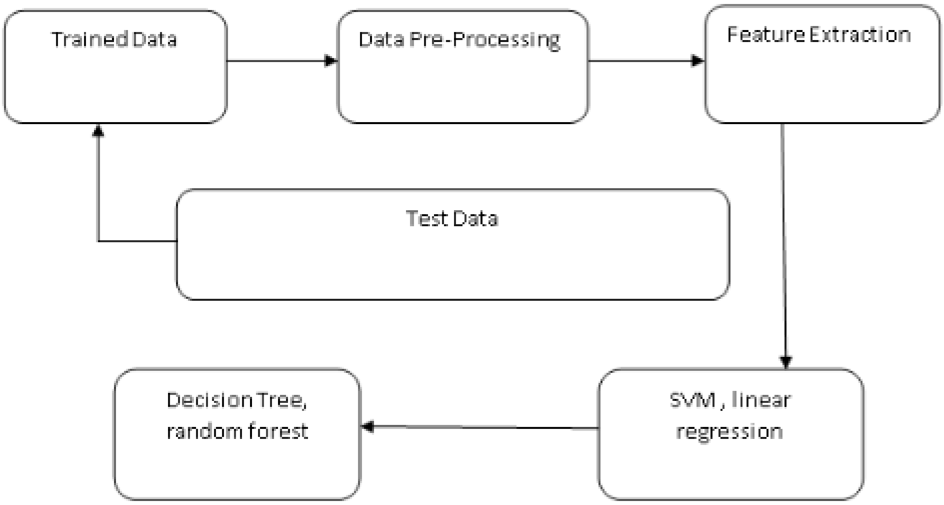


Figure 3.1 Flowchart of Churn Prediction

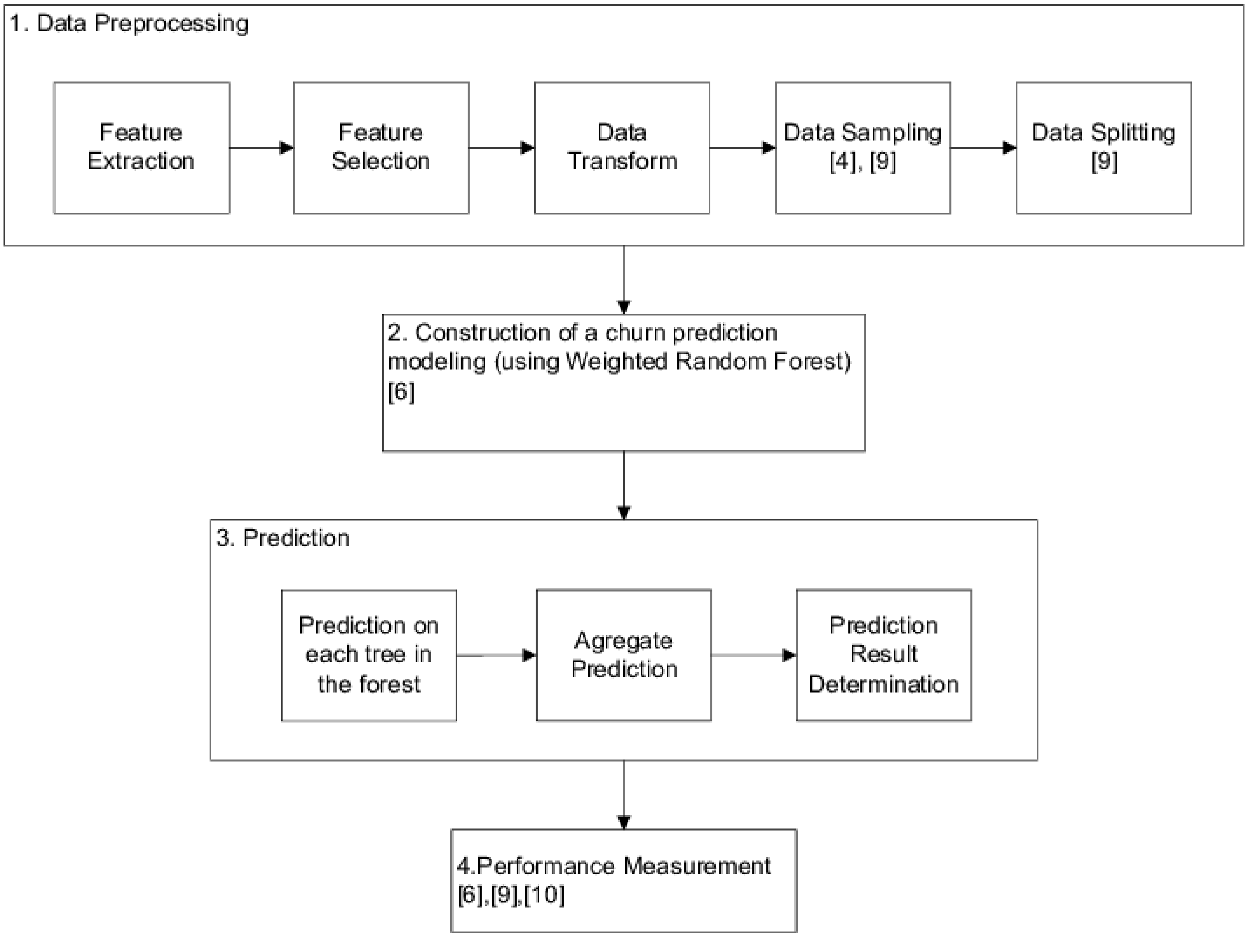

1. Data Preprocessing
Feature Extraction
Feature Selection
Data Transform
Data Sampling [4], [9]
Data Splitting [9]
2. Construction of a churn prediction modeling (using Weighted Random Forest) [6]
3. Prediction
Prediction on each tree in the forest
Agregate Prediction
Prediction Result Determination
4.Performance Measurement [6],[9],[10]

Figure 3.2 Flowchart for Data preprocessing and Prediction

CHAPTER 4

SYSTEM DESIGN

4.1 INTRODUCTION

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta -model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

4.2 UML DIAGRAMS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.

6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

4.2.1 USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use c ase diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

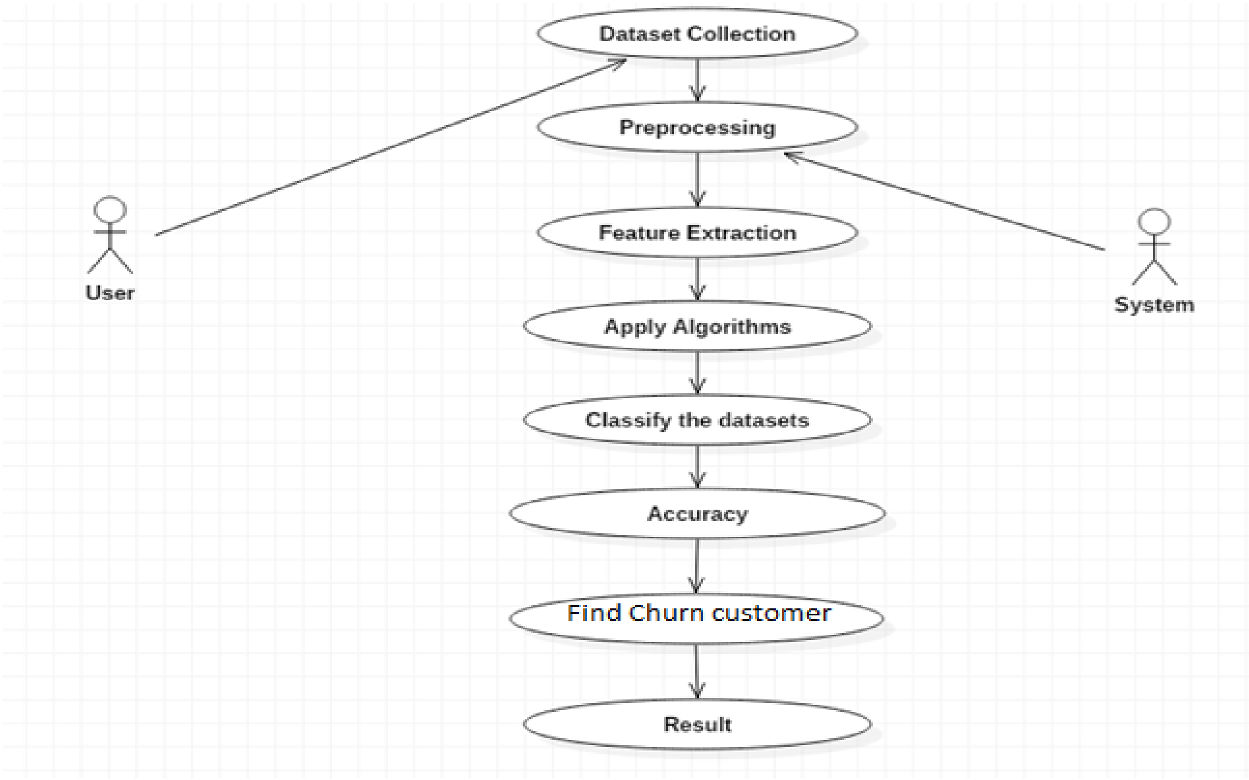


Figure 4.1 Use Case Diagram

4.2.2 SEQUENCE DIAGRAM:

Sequence Diagrams Represent the objects participating the interaction horizontally and time vertically. A Use Case is a kind of behavioral classifier that represents a declaration of an offered behavior. Each use case specifies some behavior, possibly including variants that the subject can perform in collaboration with one or more actors. Use cases define the offered behavior of the subject without reference to its internal structure. These behaviors, involving interactions between the actor and the subject, may result in changes to the state of the subject and communications with its environment. A use case can include possible variations of its basic behavior, including exceptional behavior and error handling.

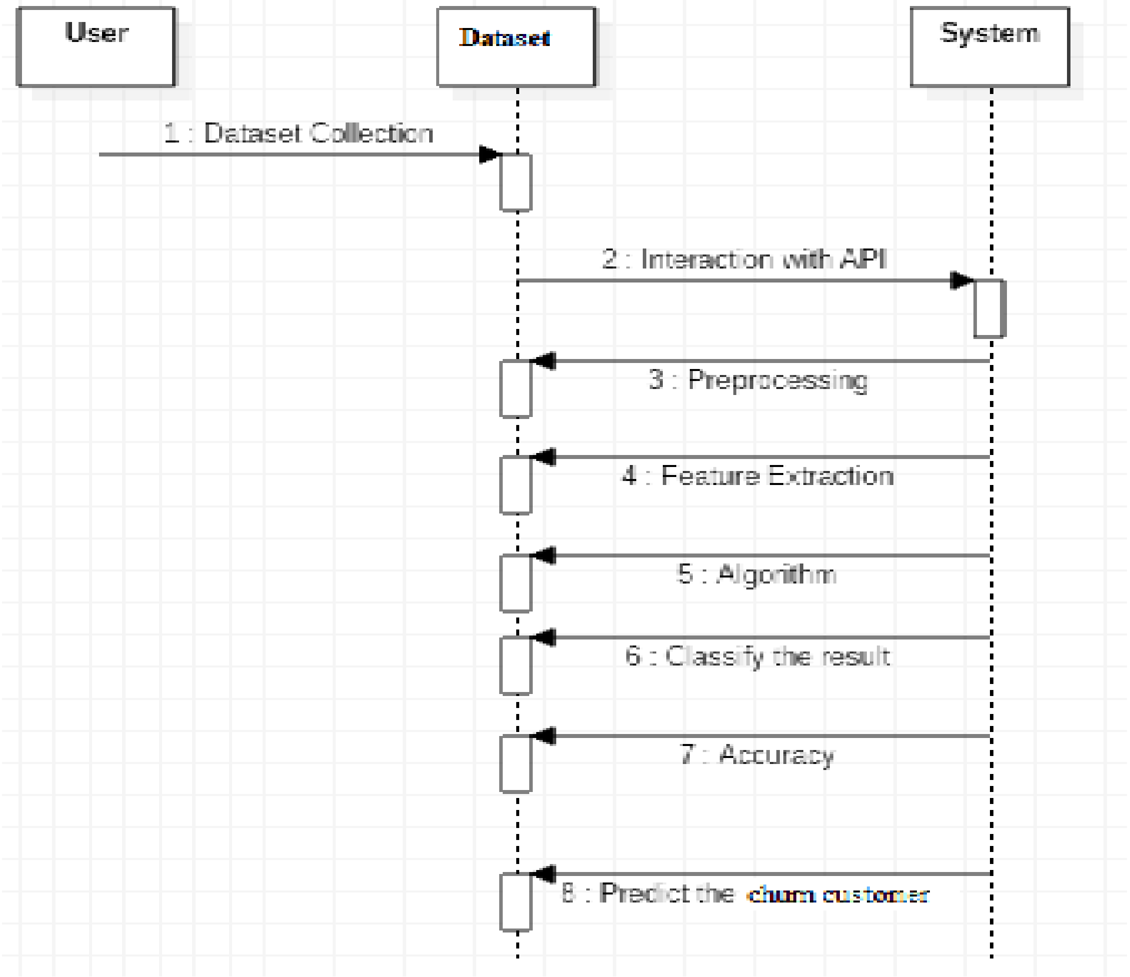


Figure 4.2 Sequence Diagram

4.2.3 ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of Workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

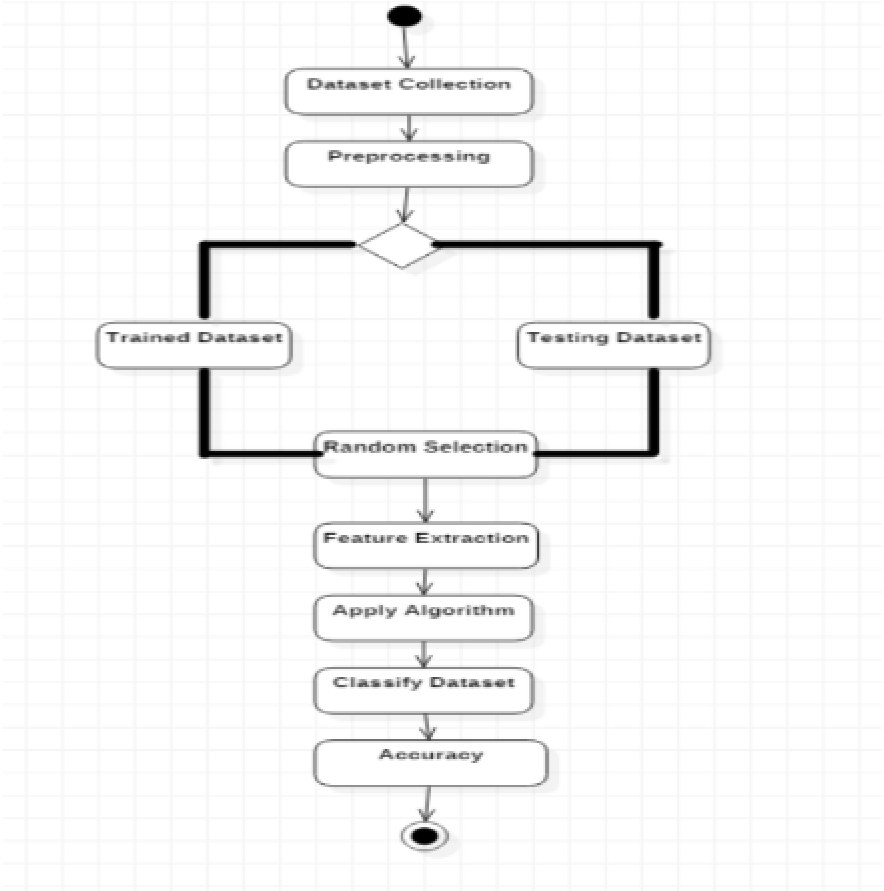


Figure 4.3 Activity Diagram

4.2.4 CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

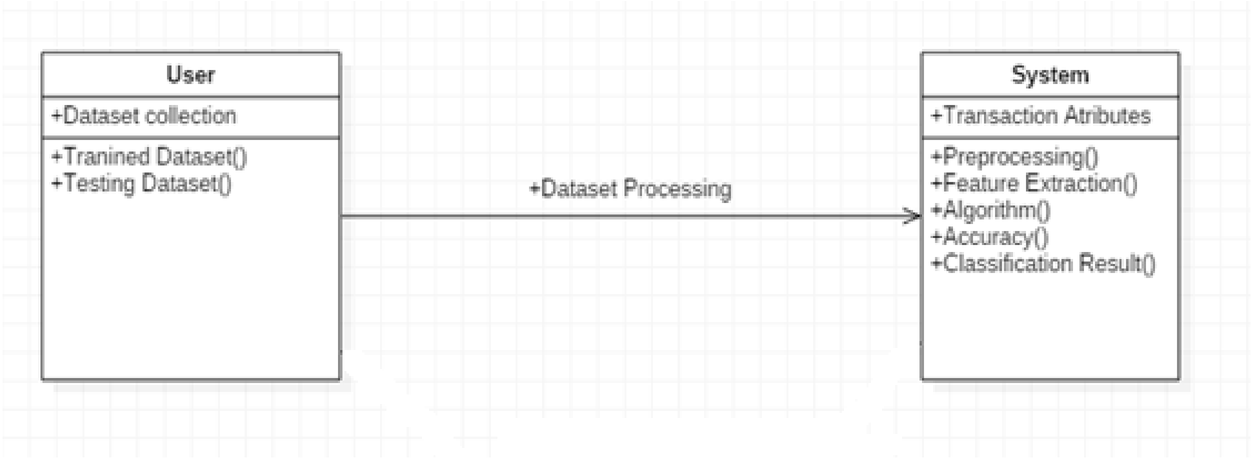


Figure 4.4 Class Diagram

CHAPTER 5

IMPLEMENTATON

5.1 INTRODUCTION

The following are the libraries used in the project.

- Python
- Anaconda Navigator
- Python environment

5.1.1 PYTHON

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python can serve as a scripting language for web applications, e.g., via mod_wsgi for the Apache web server. With Web Server Gateway Interface, a standard API has evolved to facilitate these applications. Web frameworks like Django, Pylons, Pyramid, TurboGears, web2py, Tornado, Flask, Bottle and Zope support developers in the design and maintenance of complex applications. Pyjs and Iron Python can be used to develop the client-side of Ajax-based applications. SQLAlchemy can be used as data mapper to a relational database. Twisted is a framework to program communications between computers, and is used (for example) by Dropbox. Libraries such as NumPy, SciPy and Matplotlib allow the effective use of Python in scientific computing, with specialized libraries such as Biopython and Atrophy providing domain-specific functionality.

Sage Math is a mathematical software with a notebook interface programmable in Python: its library covers many aspects of mathematics, including algebra, combinatorics, numerical mathematics, number theory, and calculus. Python has been successfully embedded in many software products as a scripting language, including in finite element method software such as

Abaqus, 3D parametric modeler like Free CAD, 3D animation packages such as 3ds Max, Blender, Cinema 4D, Lightwave, Houdini, Maya, modo, Motion Builder, Softimage, the visual effects compositor Nuke, 2D imaging programs like GIMP, Inkscape, Scribus and Paint Shop Pro, and musical notation programs like score writer and Capella. GNU Debugger uses Python as a pretty printer to show complex structures such as C++ containers.

Esri promotes Python as the best choice for writing scripts in ArcGIS. It has also been used in several video games, and has been adopted as first of the three available programming languages in Google App Engine, the other two being Java and Go. Python is commonly used in artificial intelligence projects with the help of libraries like TensorFlow, Keras, Pytorch and Scikit-learn. As a scripting language with modular architecture, simple syntax and rich text processing tools, Python is often used for natural language processing.

5.1.2 ANACONDA NAVIGATOR

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda 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, mac OS and Linux.

Why use Navigator?

In order to run, many scientific packages depend on specific versions of other packages. Data scientists often use multiple versions of many packages, and use multiple environments to separate these different versions.

The command line program conda is both a package manager and an environment manager, to help data scientists ensure that each version of each package has all the dependencies it requires and works correctly.

The following applications are available by default in Navigator:

- JupyterLab
- Jupyter Notebook
- QTConsole
- Spyder
- VSCode
- Glueviz
- Orange 3 App
- Rodeo
- RStudio

Advanced conda users can also build your own Navigator applications

How can I run code with Navigator?

The simplest way is with Spyder. From the Navigator Home tab, click Spyder, and write and execute your code. You can also use Jupyter Notebooks the same way. Jupyter Notebooks are an increasingly popular system that combine your code, descriptive text, output, images and interactive interfaces into a single notebook file that is edited, viewed and used in a web browser.

5.1.3 PYTHON ENVIRONMENT

Python is available on a wide variety of platforms including Linux and Mac OS X. Let's understand how to set up our Python environment.

Python’s standard library:

- Pandas
- NumPy
- Sklearn
- seaborn
- matplotlib
- Importing Datasets

5.1.3.1 PANDAS

Pandas is quite a game changer when it comes to analyzing data with Python and it is one of the most preferred and widely used tools in data munging/wrangling if not THE most used one. Pandas is an open source

It’s well known about Pandas is that it takes data and creates a Python object with rows and columns called data frame that looks very similar to table in a statistical software This is so much easier to work with in comparison to working with lists and/or dictionaries through for loops or list comprehension.

WORKING WITH PANDAS

Loading and Saving Data with Pandas

When you want to use Pandas for data analysis, you’ll usually use it in one of three different ways:

- Convert a Python’s list, dictionary or NumPy array to a Pandas data frame
- Open a local file using Pandas, usually a CSV file, but could also be a delimited text file (like TSV), Excel, etc.
- Open a remote file or database like a CSV or a JSONon a website through a URL or read from a SQL table/database

There are different commands to each of these options, but when you open a file, they would look like this:

➔ pd.read_filetype()

As I mentioned before, there are different filetypes Pandas can work with, so you would replace “filetype” with the actual, well, filetype (like CSV). You would give the path, filename etc inside the parenthesis. Inside the parenthesis you can also pass different arguments that relate to how to open the file. There are numerous arguments and in order to know all you them, you would have to read the documentation (for example, the documentation for pd.read_csv() would contain all the arguments you can pass in this Pandas command).

In order to convert a certain Python object (dictionary, lists etc) the basic command is:

➔ pd.DataFrame()

Inside the parenthesis you would specify the object(s) you're creating the data frame from. This command also has different arguments. You can also save a data frame you're working with/on to different kinds of files (like CSV, Excel, JSON and SQL tables). The general code for that is:

➔ df.to_filetype(filename)

VIEWING AND INSPECTING DATA

Now that you've loaded your data, it's time to take a look. How does the data frame look? Running the name of the data frame would give you the entire table, but you can also get the first n rows with df.head(n) or the last n rows with df.tail(n). df.shape would give you the number of rows and columns. df.info() would give you the index, datatype and memory information. The command s.value_counts(dropna=False) would allow you to view unique values and counts for a series (like a column or a few columns). A very useful command is df.describe() which inputs summary statistics for numerical columns. It is also possible to get statistics on the entire data frame or a series (a column etc):

- df.mean() Returns the mean of all columns
- df.corr() Returns the correlation between columns in a data frame
- df.count() Returns the number of non-null values in each data frame column

- df.max()Returns the highest value in each column
- df.min()Returns the lowest value in each column
- df.median()Returns the median of each column
- df.std()Returns the standard deviation of each column

SELECTION OF DATA

One of the things that is so much easier in Pandas is selecting the data you want in comparison to selecting a value from a list or a dictionary. You can select a column (df[col]) and return column with label col as Series or a few columns (df[[col1, col2]]) and returns columns as a new DataFrame. You can select by position (s.iloc[0]), or by index (s.loc['index_one']) . In order to select the first row you can use df.iloc[0,:] and in order to select the first element of the first column you would run df.iloc[0,0] . These can also be used in different combinations, so I hope it gives you an idea of the different selection and indexing you can perform in Pandas.

Data Cleaning

Data cleaning is a very important step in data analysis. For example, we always check for missing values in the data by running pd.isnull() which checks for null Values, and returns a boolean array (an array of true for missing values and false for non-missing values). In order to get a sum of null/missing values, run pd.isnull().sum(). pd.notnull() is the opposite of pd.isnull(). After you get a list of missing values you can get rid of them, or drop them by using df.dropna() to drop the rows or df.dropna(axis=1) to drop the columns. A different approach would be to fill the missing values with other values by using df.fillna(x) which fills the missing values with x (you can put there whatever you want) or s.fillna(s.mean()) to replace all null values with the mean (mean can be replaced with almost any function from the statistics section).

It is sometimes necessary to replace values with different values. For example, s.replace(1,'one') would replace all values equal to 1 with 'one'. It's possible to do it for multiple values: s.replace([1,3],['one','three'])would replace all 1 with 'one' and 3 with 'three'. You can also rename specific columns by running: df.rename(columns={'old_name': 'new_ name'})or use df.set_index('column_one') to change the index of the data frame.

5.1.3.2 NUMPY

NumPy is one such powerful library for array processing along with a large collection of high-level mathematical functions to operate on these arrays. These functions fall into categories like Linear Algebra, Trigonometry, Statistics, Matrix manipulation, etc.

GETTING NUMPY

NumPy's main object is a homogeneous multidimensional array. Unlike python's array class which only handles one-dimensional array, NumPy's ND array class can handle multidimensional array and provides more functionality. NumPy's dimensions are known as axes. For example, the array below has 2 dimensions or 2 axes namely rows and columns. Sometimes dimension is also known as a rank of that particular array or matrix.

IMPORTING NUMPY

NumPy is imported using the following command. Note here np is the convention followed for the alias so that we don't need to write NumPy every time.

- ➔ import numpy as np

NumPy is the basic library for scientific computations in Python and this article illustrates some of its most frequently used functions. Understanding NumPy is the first major step in the journey of machine learning and deep learning.

5.1.3.3 SKLEARN

In python, Scikit-learn library has a pre-built functionality under Sklearn. Preprocessing. Next thing is to do feature extraction Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally, our models are trained using Classifier algorithm. We use nltk. classify module on Natural Language Toolkit library on Python. We use the labelled dataset gathered. The rest of our labelled data will be used to evaluate the models.

Some machine learning algorithms were used to classify preprocessed data. The chosen classifiers were Decision tree, Support Vector Machines and Random forest. These algorithms are very popular in text classification tasks.

5.1.3.4 SEABORN

DATA VISUALIZATION IN PYTHON:

Data visualization is the discipline of trying to understand data by placing it in a visual context, so that patterns, trends and correlations that might not otherwise be detected can be exposed. Python offers multiple great graphing libraries that come packed with lots of different features. No matter if you want to create interactive, live or highly customized plots python has an excellent library.

5.1.4 TO GET A LITTLE OVERVIEW OF PLOTTING LIBRARIES

- Matplotlib: low level, provides lots of freedom
- Pandas Visualization: easy to use interface, built on Matplotlib
- Seaborn: high-level interface, great default styles
- ggplot: based on R's ggplot2, uses Grammar of Graphics
- Plotly: can create interactive plots

5.1.4.1 MATPLOTLIB

Matplotlib is the most popular python plotting library. It is a low-level library with a MATLAB like interface which offers lots of freedom at the cost of having to write more code.

- To install Matplotlib pip and conda can be used.
- pip install matplotlib
- conda install matplotlib

Matplotlib is specifically good for creating basic graphs like line charts, bar charts, histograms and many more. It can be imported by typing

➔ import matplotlib.pyplot as plt

LINE CHART

In Matplotlib we can create a line chart by calling the plot method. We can also plot multiple columns in one graph, by looping through the columns we want, and plotting each column on the same axis.

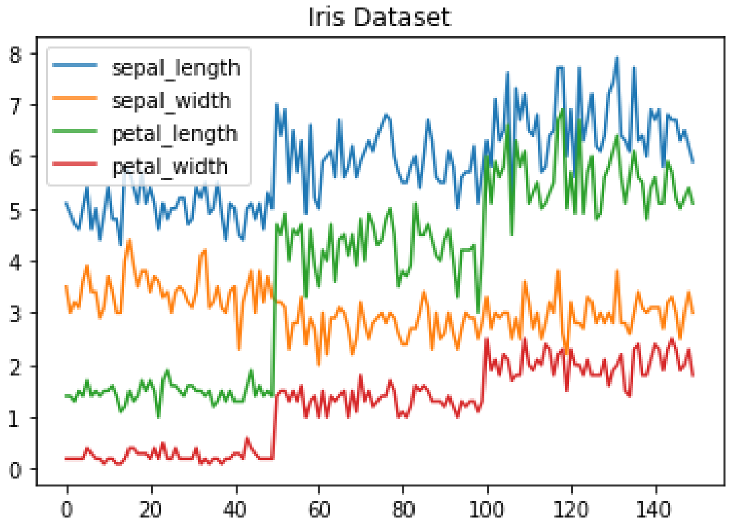


Figure 5.1 Line chart

HISTOGRAM

In Matplotlib we can create a Histogram using the hist method. If we pass it categorical data like the points column from the wine-review dataset it will automatically calculate how often each class occurs.

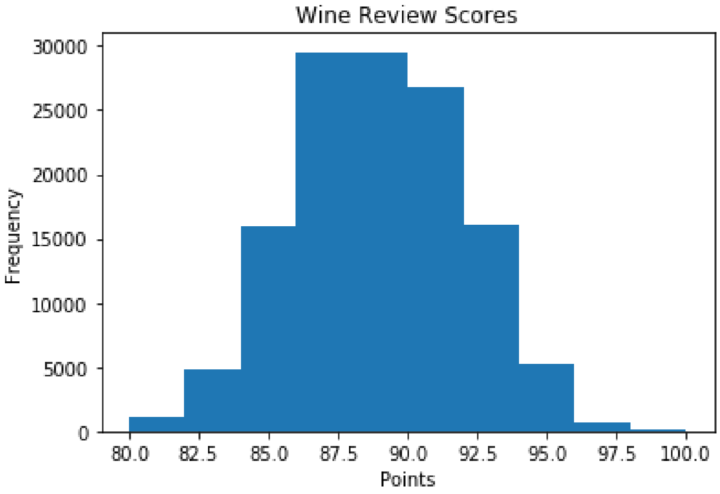


Figure 5.2 Histogram

BAR CHART

A bar-chart can be created using the bar method. The bar-chart isn't automatically calculating the frequency of a category so we are going to use pandas value counts function to do this. The bar-chart is useful for categorical data that doesn't have a lot of different categories (less than 30) because else it can get quite messy.

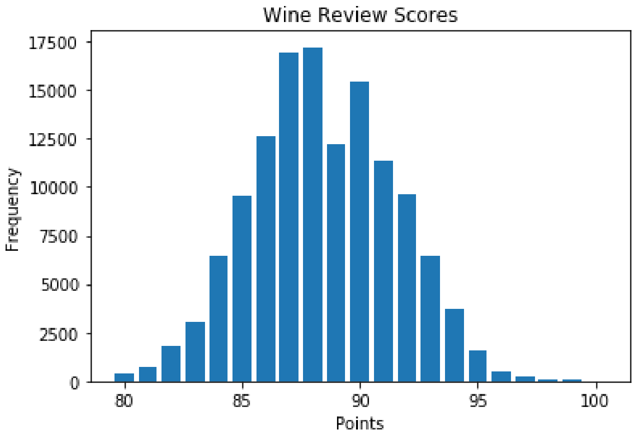


Figure 5.3 Bar Chart

HEATMAP

A Heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors. Heatmaps are perfect for exploring the correlation of features in a dataset.

To get the correlation of the features inside a dataset we can call <dataset>.corr (), which is a Pandas data frame method. This will give use the correlation matrix. We can now use either Matplotlib or Seaborn to create the heatmap.

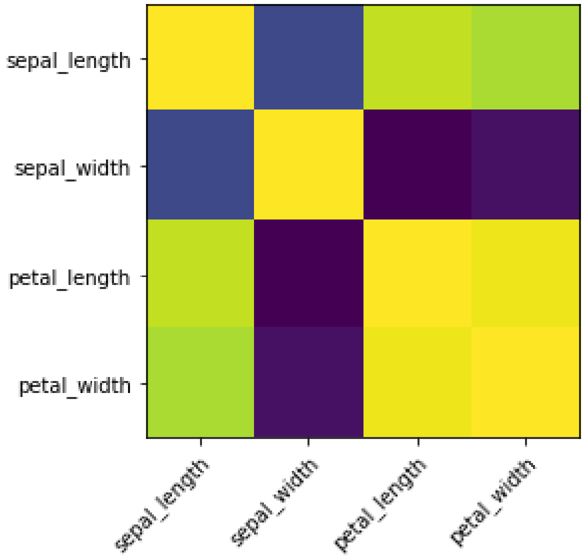


Figure 5.4 Heatmap without annotations

5.1.4.2 PANDAS VISUALIZATION

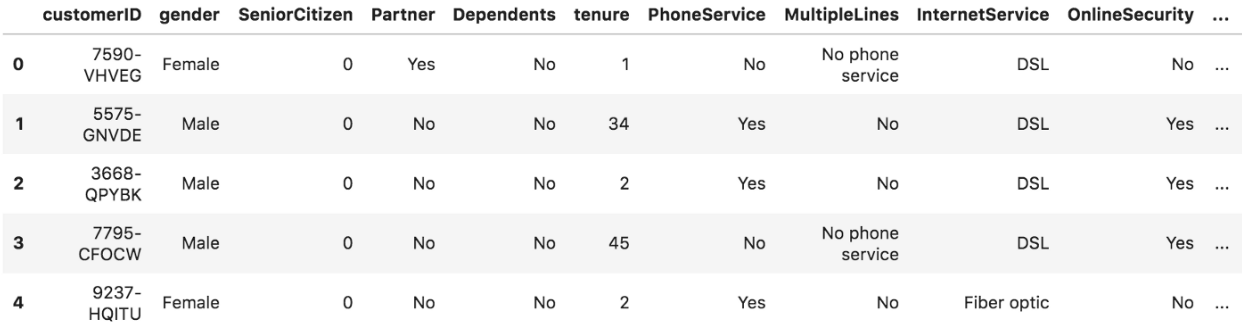
Pandas is an open source high-performance, easy-to-use library providing data structures, such as data frames, and data analysis tools like the visualization tools we will use in this article.Pandas Visualization makes it really easy to create plots out of a pandas data frame and series. It also has a higher-level API than Matplotlib and therefore we need less code for the same results. Data visualization is the discipline of trying to understand data by placing it in a visual context, so that patterns, trends and correlations that might not otherwise be detected can be exposed. Python offers multiple great graphing libraries that come packed with lots of different features. In this article we looked at Matplotlib, Pandas visualization and Seaborn.

5.2 METHOD OF IMPLEMENTATION

DATA PREPROCESSING

We use pandas to read the dataset and preprocess it. Telco dataset has one customer per line with many columns (features). There aren't any rows with all missing values or duplicates (this rarely happens with real-world datasets). There are 11 samples that have TotalCharges set to which seems like a mistake in the data. We remove those samples and set the type to numeric (float).

```
df = pd.read_csv('data/telcom.csv)
df = df.dropna(how=”all”) # remove samples with all missing values
df = df[~df.duplicated()] # remove duplicates
total_charges_filter = df.TotalCharges == " "
df = df[~total_charges_filter]
df.TotalCharges = pd.to_numeric(df.TotalCharges)
```



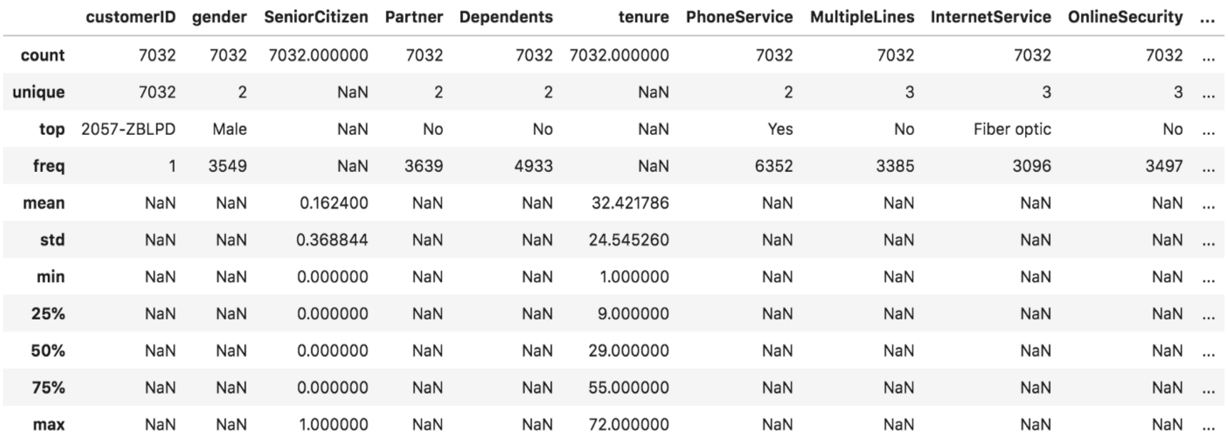
	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...
3	7795-CFOCW	Male	0	No	No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

Figure 5.5 sample telecom data model

EXPLORATORY DATA ANALYSIS:

We have 2 types of features in the dataset: categorical (two or more values and without any order) and numerical. Most of the feature names are self-explanatory, except for:

- Partner*:* whether the customer has a partner or not (Yes, No),
- Dependents: whether the customer has dependents or not (Yes, No),
- Online Backup: Whether the customer has an online backup or not (Yes, No),
- tenure: number of months the customer has stayed with the company,
- Monthly Charges: the amount charged to the customer monthly,
- Total Charges: the total amount charged to the customer.



	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...
count	7032	7032	7032.000000	7032	7032	7032.000000	7032	7032	7032	7032	...
unique	7032	2	NaN	2	2	NaN	2	3	3	3	...
top	2057-ZBLPD	Male	NaN	No	No	NaN	Yes	No	Fiber optic	No	...
freq	1	3549	NaN	3639	4933	NaN	6352	3385	3096	3497	...
mean	NaN	NaN	0.162400	NaN	NaN	32.421786	NaN	NaN	NaN	NaN	...
std	NaN	NaN	0.368844	NaN	NaN	24.545260	NaN	NaN	NaN	NaN	...
min	NaN	NaN	0.000000	NaN	NaN	1.000000	NaN	NaN	NaN	NaN	...
25%	NaN	NaN	0.000000	NaN	NaN	9.000000	NaN	NaN	NaN	NaN	...
50%	NaN	NaN	0.000000	NaN	NaN	29.000000	NaN	NaN	NaN	NaN	...
75%	NaN	NaN	0.000000	NaN	NaN	55.000000	NaN	NaN	NaN	NaN	...
max	NaN	NaN	1.000000	NaN	NaN	72.000000	NaN	NaN	NaN	NaN	...

Figure 5.6 Dataset summary

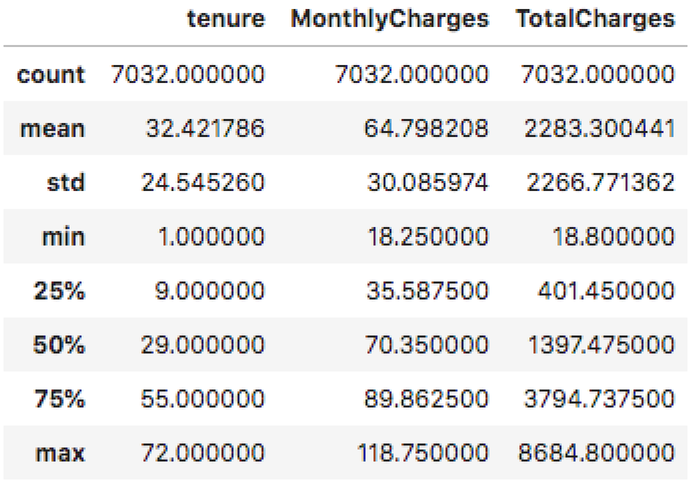
We combine features into two lists so that we can analyze them jointly.

```
categorical_features = [
 “gender”,
 “SeniorCitizen”,
 “Partner”,
 “Dependents”,
 “PhoneService”,
 “MultipleLines”,
 “InternetService”,
 “OnlineSecurity”,
 “OnlineBackup”,
 “DeviceProtection”,
 “TechSupport”,
 “StreamingTV”,
 “StreamingMovies”,
 “Contract”,
 “PaperlessBilling”,
 “PaymentMethod”,]
numerical_features = [“tenure”, “MonthlyCharges”, “TotalCharges”]
target = “Churn”
```

NUMERICAL FEATURES DISTRIBUTION

Numeric summarizing techniques (mean, standard deviation, etc.) don't show us spikes, shapes of distributions and it is hard to observe outliers with it. That is the reason we use histograms.

```
df[numerical_features].describe()
```



	tenure	MonthlyCharges	TotalCharges
count	7032.000000	7032.000000	7032.000000
mean	32.421786	64.798208	2283.300441
std	24.545260	30.085974	2266.771362
min	1.000000	18.250000	18.800000
25%	9.000000	35.587500	401.450000
50%	29.000000	70.350000	1397.475000
75%	55.000000	89.862500	3794.737500
max	72.000000	118.750000	8684.800000

Figure 5.7 Summary of numerical features

At first glance, there aren't any outliers in the data. No data point is disconnected from distribution or too far from the mean value. To confirm that we would need to calculate interquartile range and show that values of each numerical feature are within the 1.5 interquartile range from first and third quartile.

We could convert numerical features to ordinal intervals. For example, tenure is numerical, but often we don't care about small numeric differences and instead group tenure to customers with short, medium- and long-term tenure. One reason to convert it would be to reduce the noise, often small fluctuates are just noise.

```
df[numerical_features].hist(bins=30, figsize=(10, 7))
```

We look at distributions of numerical features in relation to the target variable. We can observe that the greater Total Charges and tenure are the less is the probability of churn.

```
ROWS, COLS = 4, 4
fig, ax = plt.subplots(ROWS, COLS, figsize=(18, 18))
row, col = 0, 0
for i, categorical_feature in enumerate(categorical_features):
   if col == COLS - 1:
      row += 1
   col = i % COLS
df[categorical_feature].value_counts().plot('bar',ax=ax[row,col]).set_title(categorical_feature)
```

The next step is to look at categorical features in relation to the target variable. We do this only for contract feature. Users who have a month-to-month contract are more likely to churn than users with long term contracts.

```
feature = 'Contract'
fig, ax = plt.subplots(1, 2, figsize=(14, 4))
df[df.Churn=="No"][feature].value_counts().plot('bar', ax=ax[0]).set_title('not churned')
df[df.Churn == "Yes"][feature].value_counts().plot('bar', ax=ax[1]).set_title('churned')
```

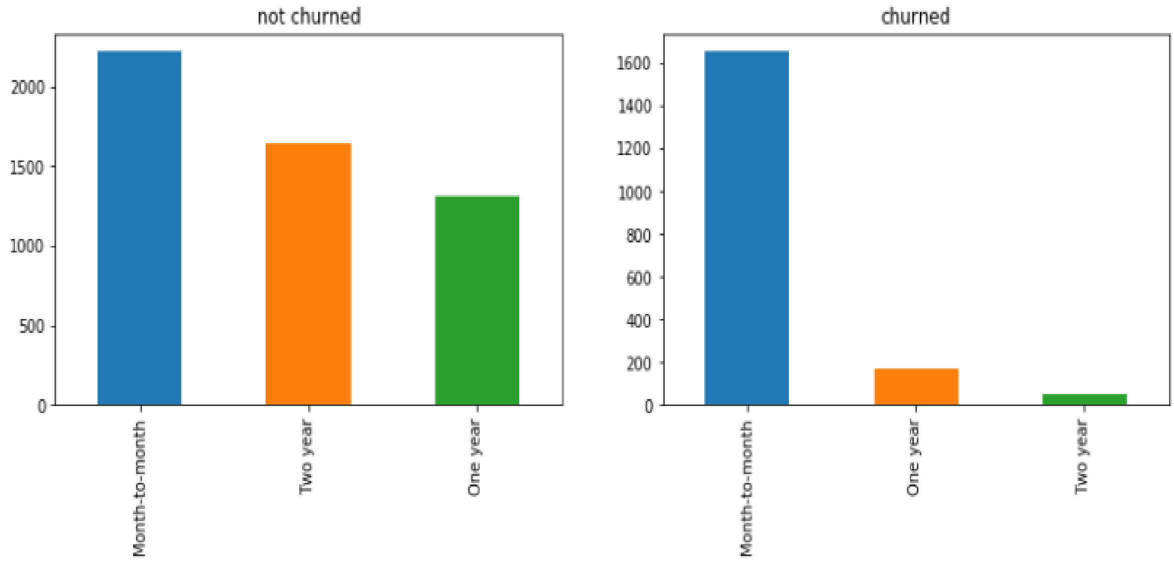


Figure 5.8 Contract feature in relation to the target variable

TARGET VARIABLE DISTRIBUTION

Target variable distribution shows that we are dealing with an imbalanced problem as there are many more non-churned as churned users. The model would achieve high accuracy as it would mostly predict majority class — users who didn't churn in our example.

```
df[target].value_counts().plot('bar').set_title('churned')
```

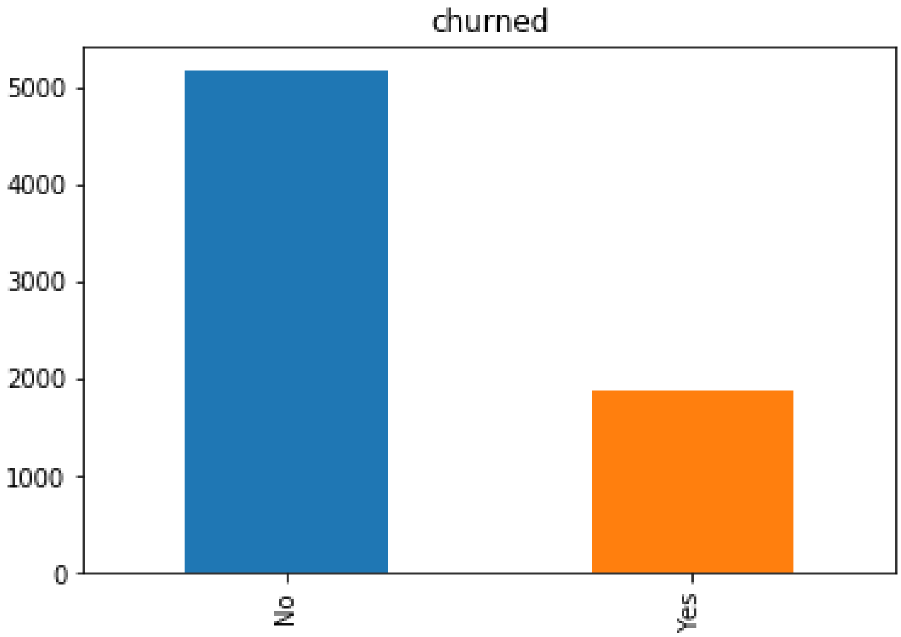


Figure 5.9 Target variable distribution

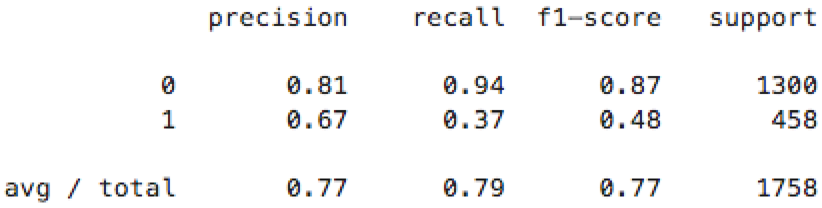
5.3 RESULT ANALYSIS:

We split the dataset to train (75% samples) and test (25% samples). We train (fit) the pipeline and make predictions. With classification_report we calculate precision and recall with actual and predicted values.

```
from sklearn.model_selection import train_test_splitdf_train, df_test = train_test_split
(df, test_size=0.25, random_state=42)pipeline.fit(df_train, df_train[target])
pred = pipeline.predict(df_test)
```

For class 1 (churned users) model achieves 0.67 precision and 0.37 recall. Precision tells us how many churned users did our classifier predicted correctly. On the other side, recall tell us how many churned users it missed. In layman terms, the classifier is not very accurate for churned users.

```
from sklearn.metrics import classification_report
print(classification_report(df_test[target], pred))
```



```
             precision    recall  f1-score   support

          0       0.81      0.94      0.87      1300
          1       0.67      0.37      0.48       458

avg / total       0.77      0.79      0.77      1758
```

Figure 5.10 Classification report

OUTPUT SCREENS

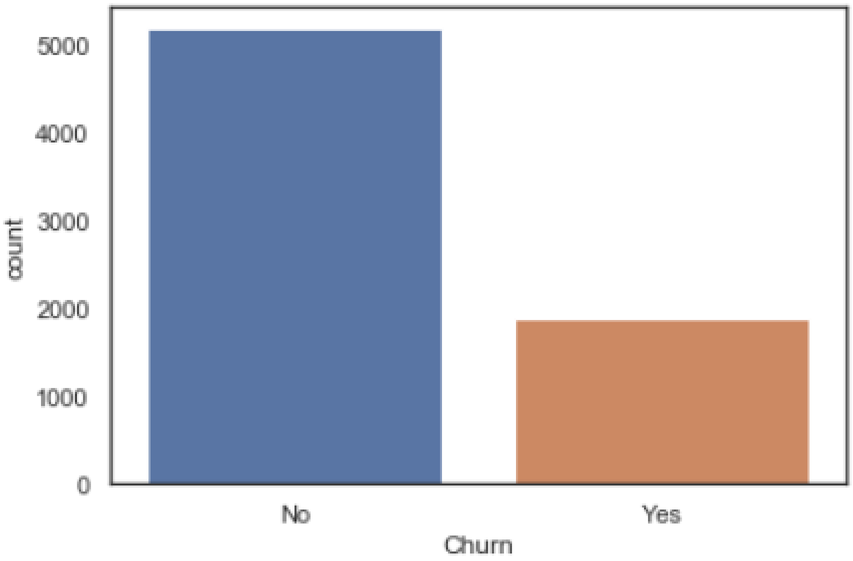


Figure 5.10 Count vs Churn graph

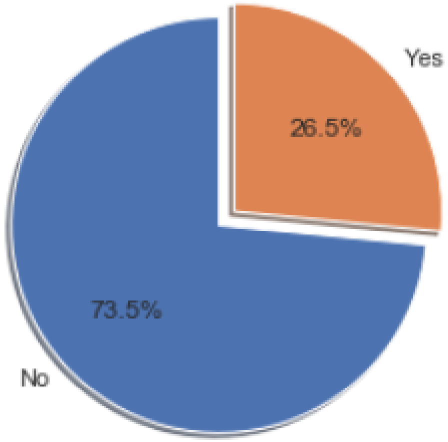


Figure 5.11 Count vs Churn pie diagram

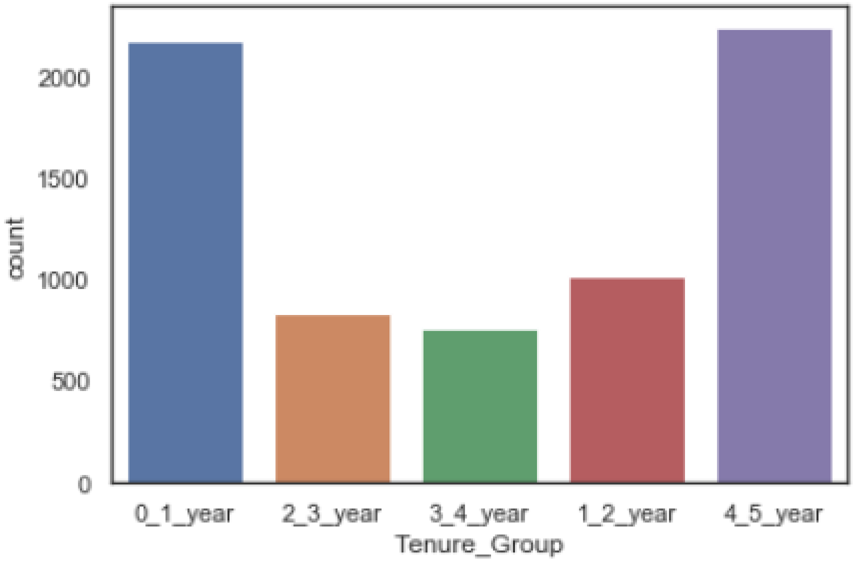


Figure 5.12 Count vs tenure bar diagram

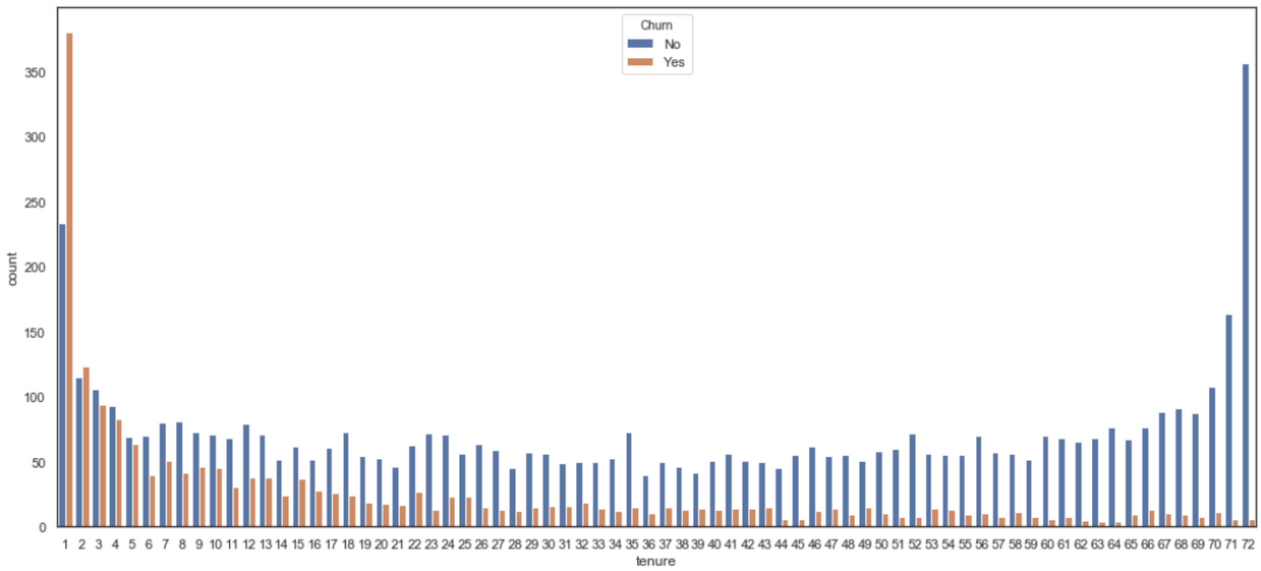


Figure 5.13 Count vs tenure(age) bar diagram

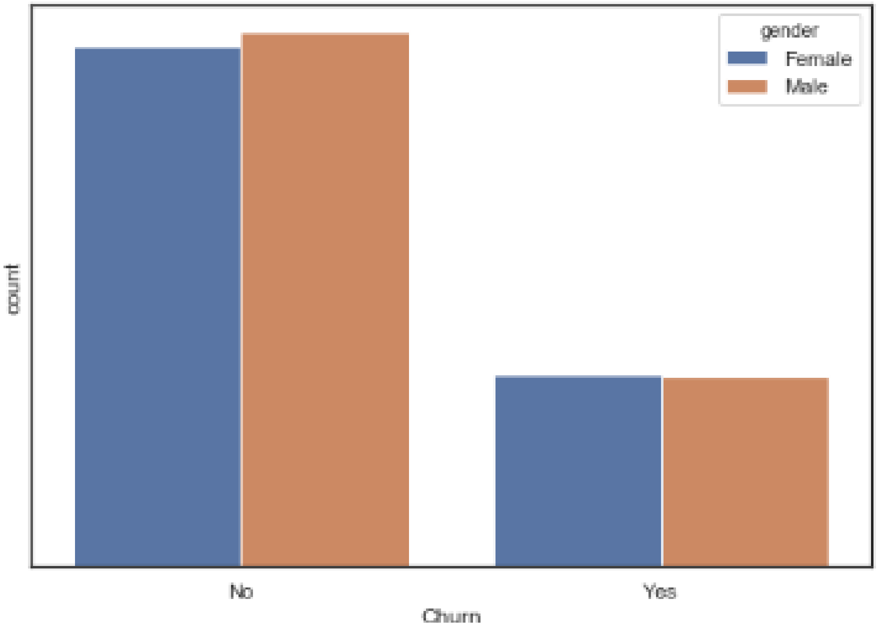


Figure 5.14 count vs churn(gender)

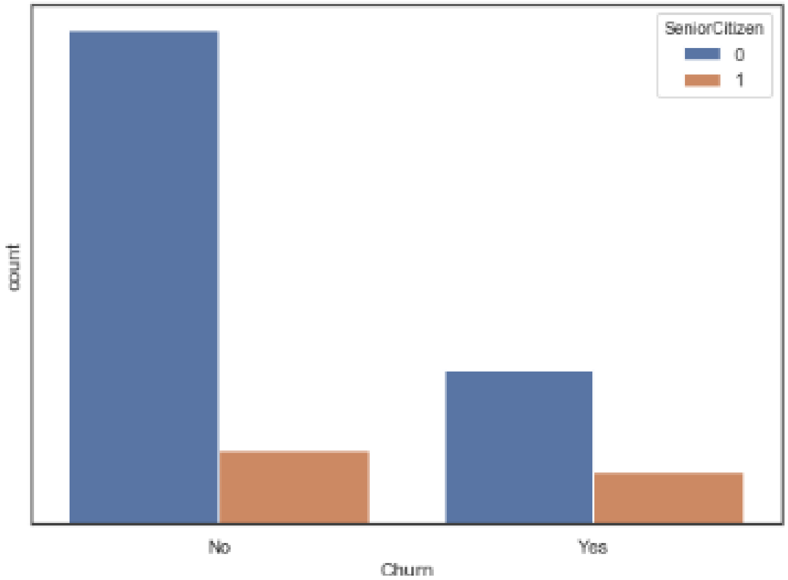


Figure 5.15 count vs churn (senior citizen)

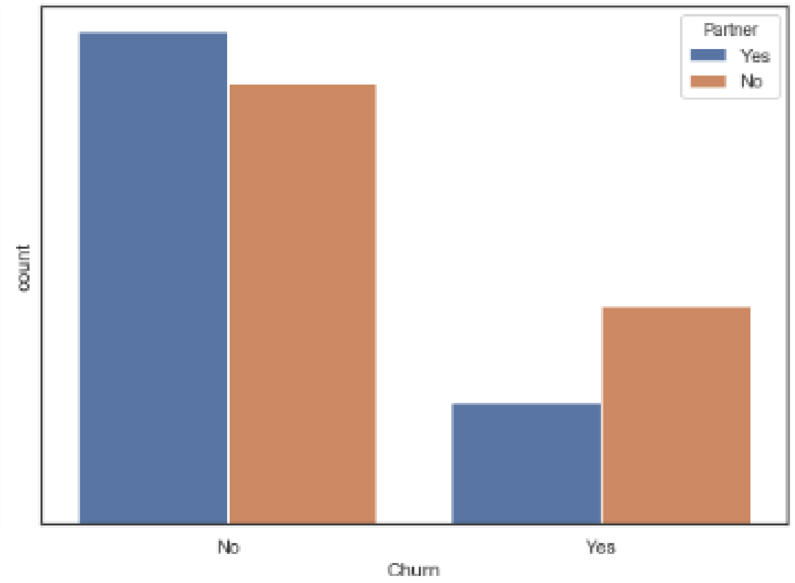


Figure 5.16 count vs churn (partner)

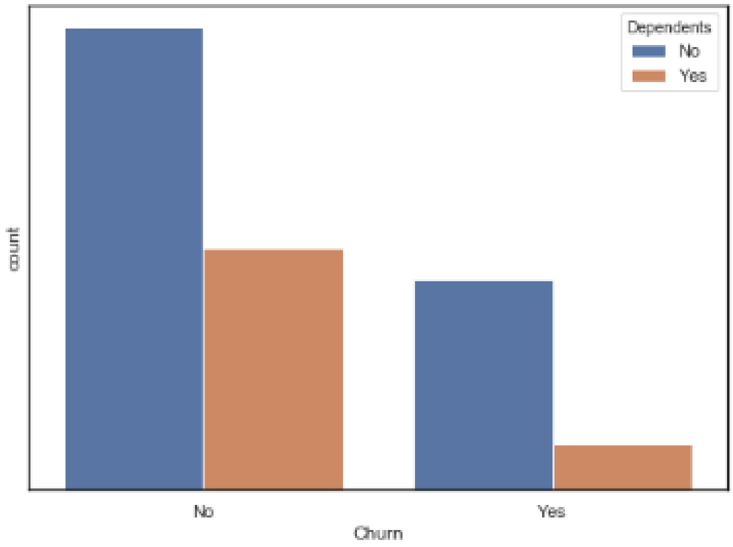
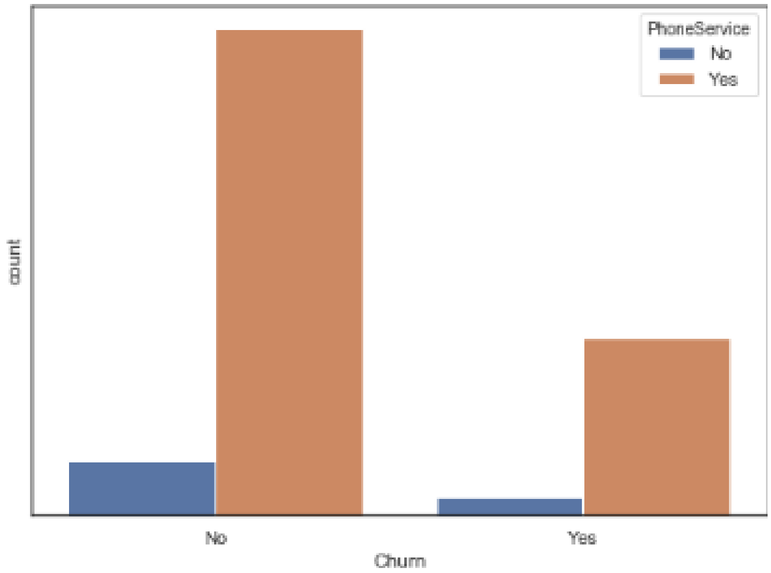
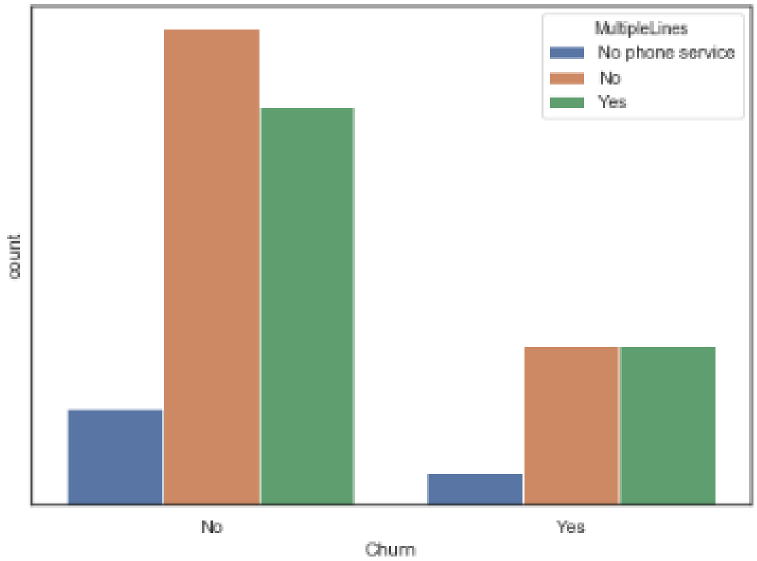


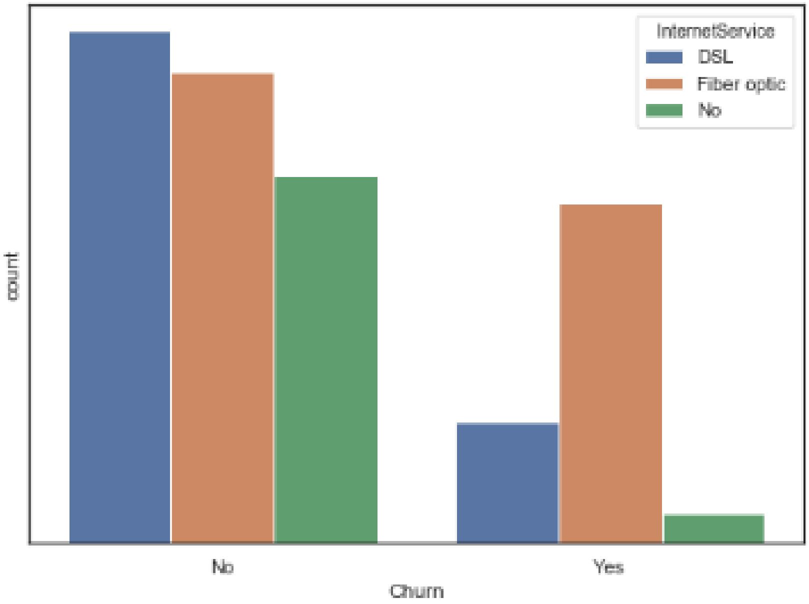
Figure 5.17 count vs churn (Dependents)



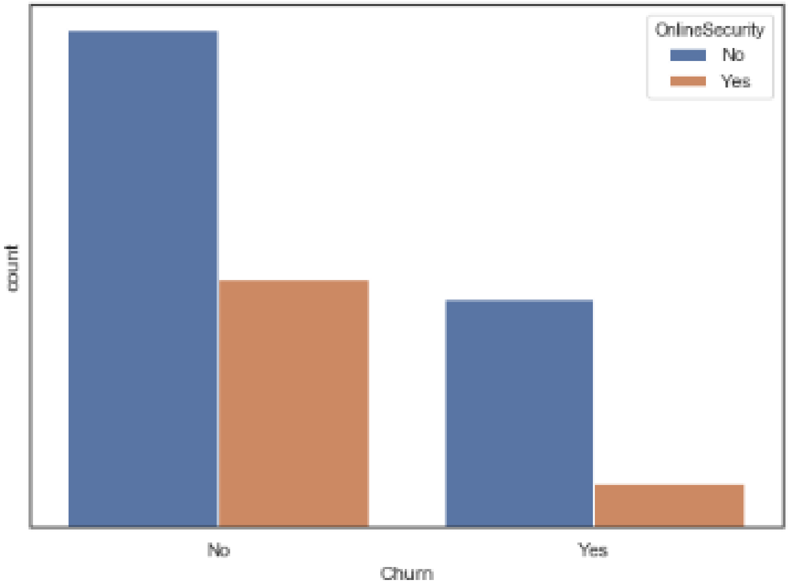
5.18 count vs churn (phone service)



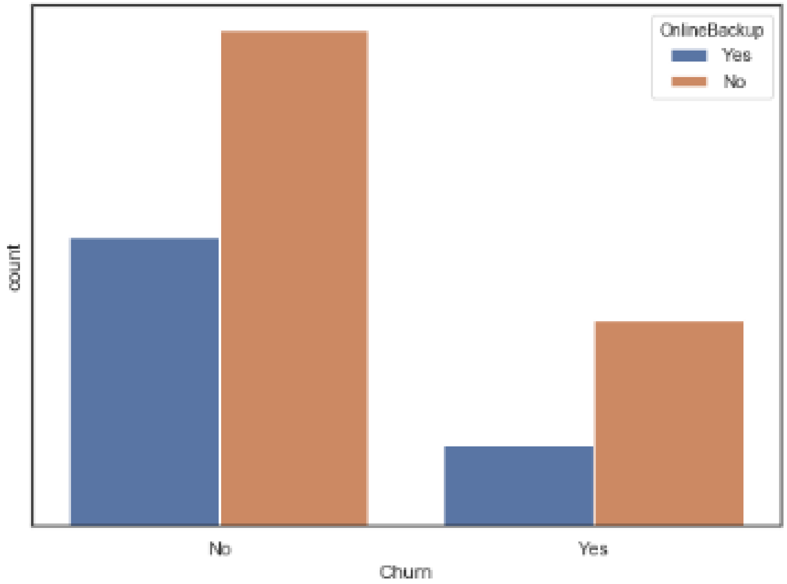
5.19 count vs churn (other networks)



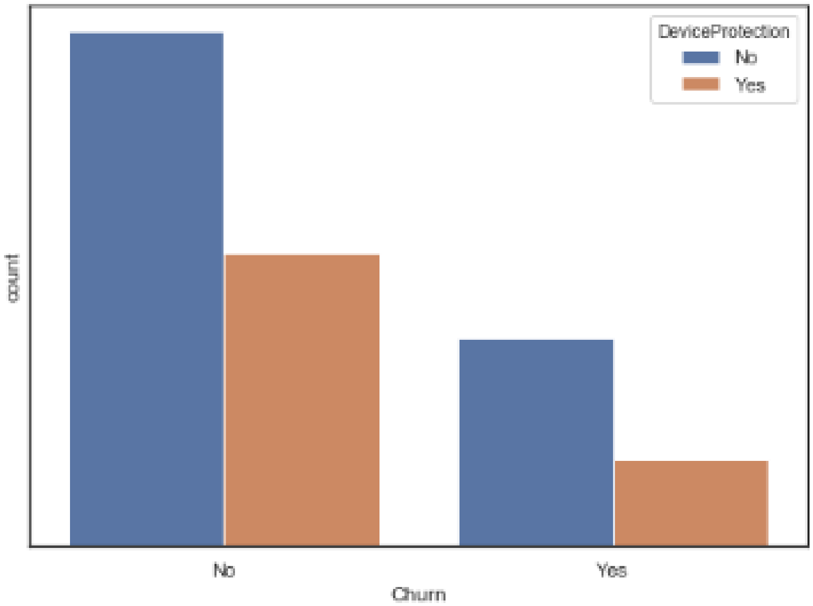
5.20 count vs churn (DSL and fiber optics)



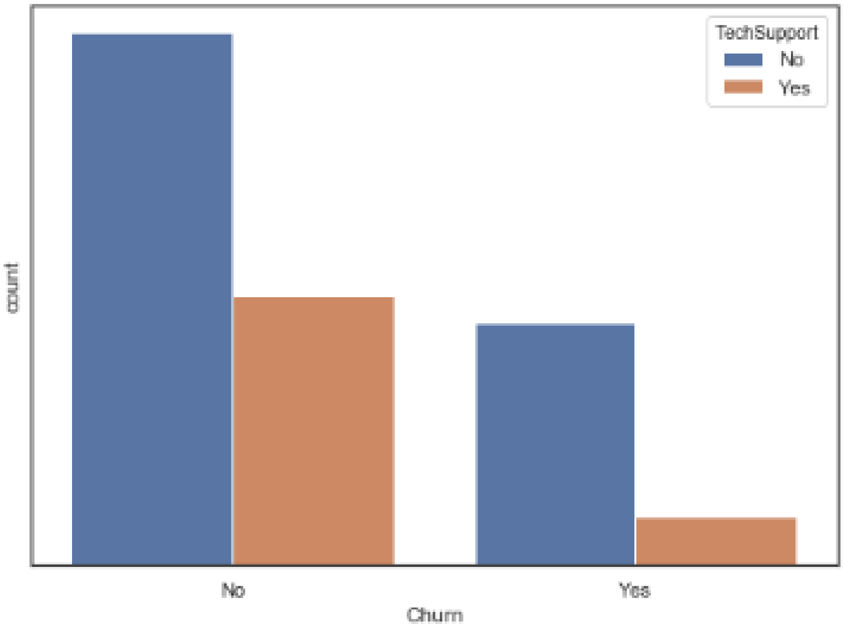
5.21 count vs churn (online security)



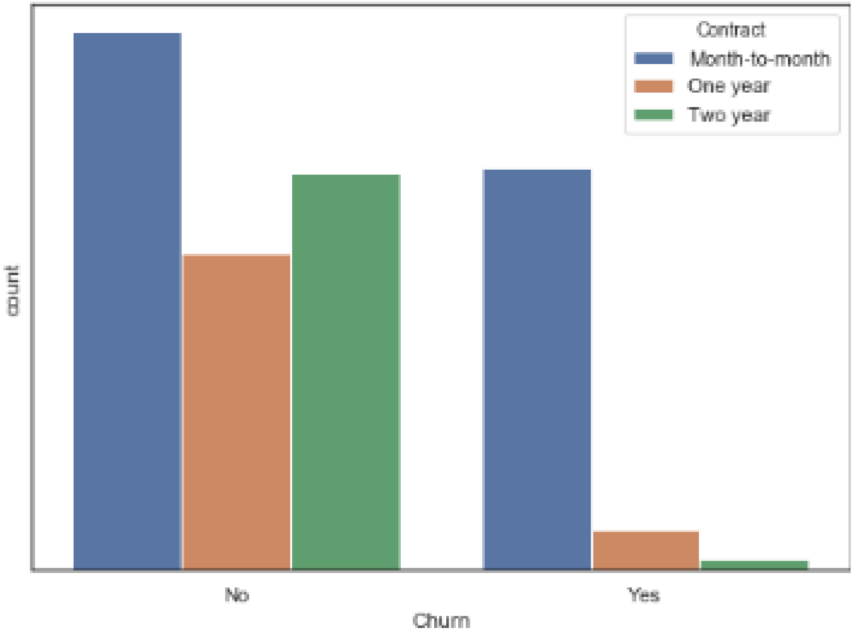
5.22 count vs churn (online backup)



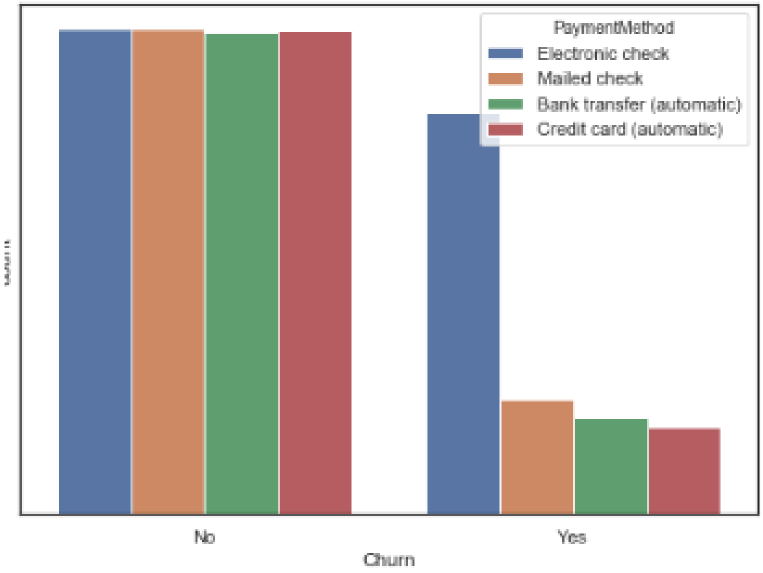
5.23 count vs churn (Device Protection)



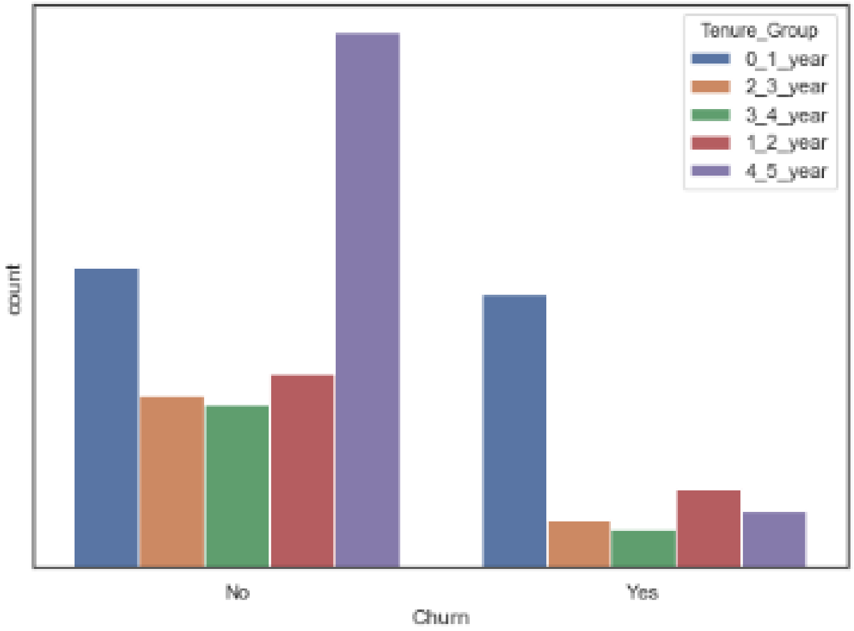
5.24 count vs churn (Tech Support)



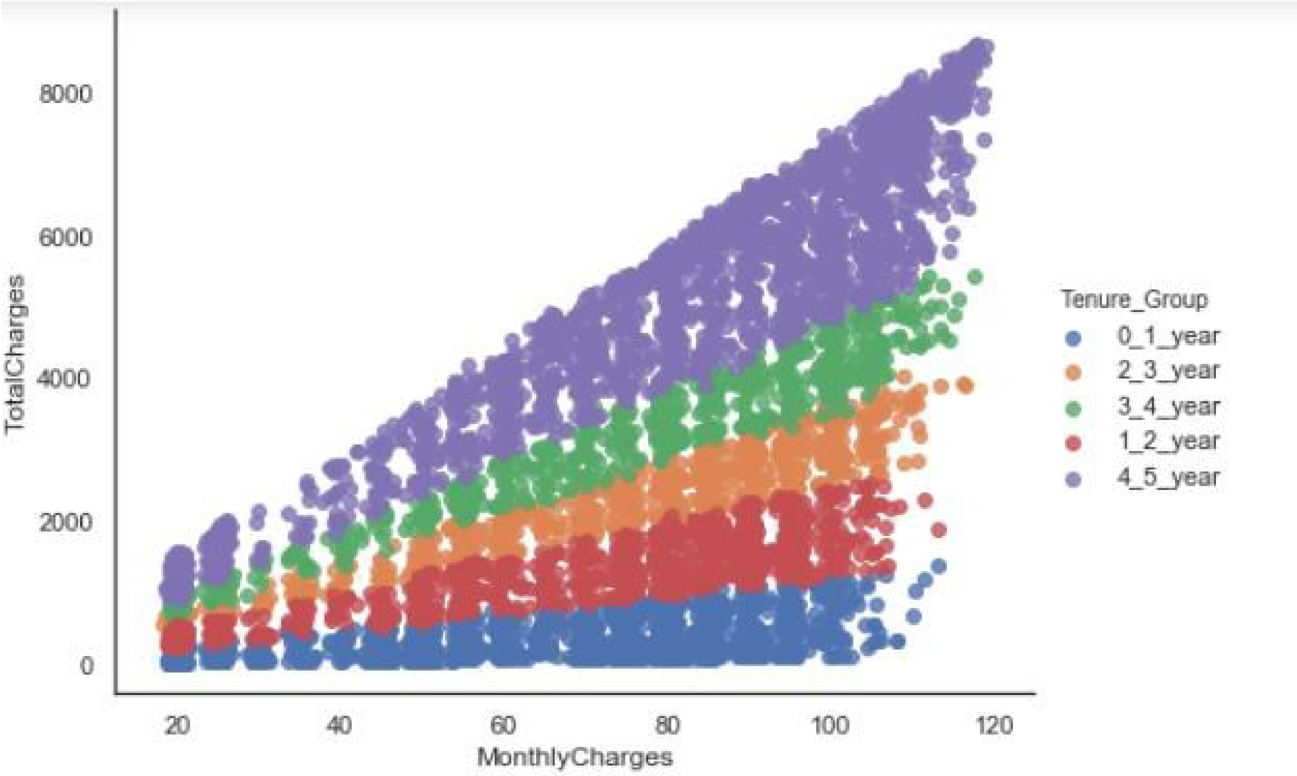
5.25 count vs churn (Type of Contract)



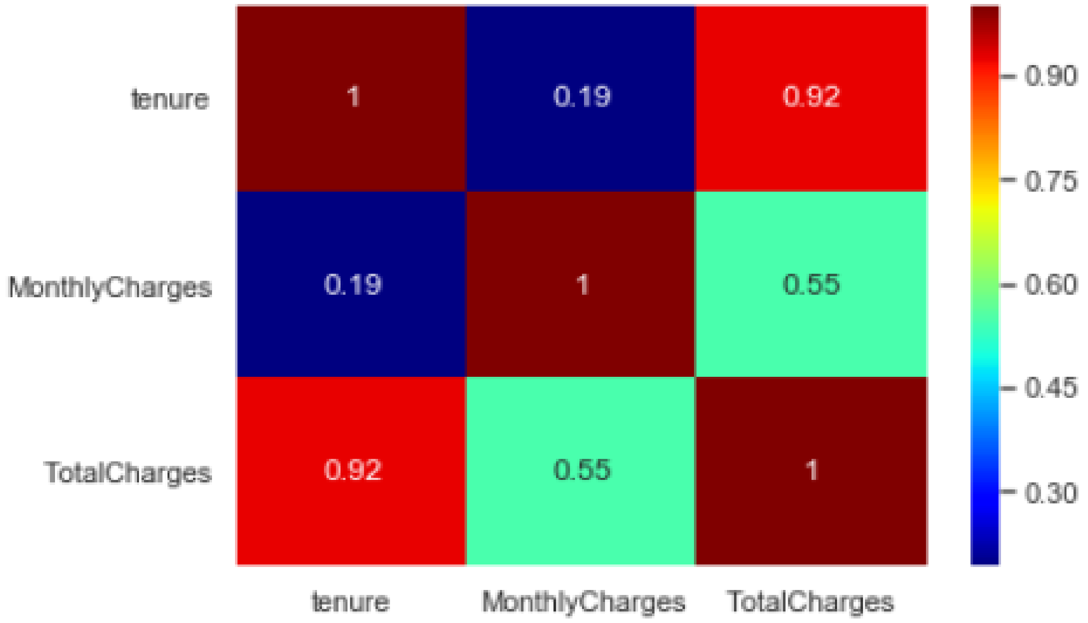
5.26 count vs churn (Billing/payment Method)



5.27 count vs churn (Tenure Group)



5.28 Total vs Monthly charges with different tenure group



5.29 heatmap for tenure, monthly and total charges

CHAPTER 6

TESTING & VALIDATION

6.1 INTRODUCTION

Software testing is a critical element of software quality assurance and represents the ultimate review of specification, design and coding. The increasing visibility of software as a system element and attendant costs associated with a software failure are motivating factors for we planned, through testing. Testing is the process of executing a program with the intent of finding an error. The design of tests for software and other engineered products can be as challenging as the initial design of the product itself. There of basically two types of testing approaches.

One is Black-Box testing – the specified function that a product has been designed to perform, tests can be conducted that demonstrate each function is fully operated.

The other is White-Box testing – knowing the internal workings of the product, tests can be conducted to ensure that the internal operation of the product performs according to specifications and all internal components have been adequately exercised.

White box and Black box testing methods have been used to test this package. The entire loop constructs have been tested for their boundary and intermediate conditions. The test data was designed with a view to check for all the conditions and logical decisions. Error handling has been taken care of by the use of exception handlers.

6.2 DESIGN OF TEST CASES

Testing is a set of activities that can be planned in advanced and conducted systematically. A strategy for software testing must accommodation low-level tests that are necessary to verify that a small source code segment has been correctly implemented as well as high-level tests that validate major system functions against customer requirements.

Software testing is one element of verification and validation. Verification refers to the set of activities that ensure that software correctly implements as specific function. Validation refers to a different set of activities that ensure that the software that has been built is traceable to customer requirements.

The main objective of software is testing to uncover errors. To fulfill this objective, a series of test steps unit, integration, validation and system tests are planned and executed. Each test step is accomplished through a series of systematic test technique that assist in the design of test cases. With each testing step, the level of abstraction with which software is considered is broadened. Testing is the only way to assure the quality of software and it is an umbrella activity rather than a separate phase. This is an activity to be performed in parallel with the software effort and one that consists of its own phases of analysis, design, implementation, execution and maintenance.

6.2.1 UNIT TESTING:

This testing method considers a module as single unit and checks the unit at interfaces and communicates with other modules rather than getting into details at statement level. Here the module will be treated as a black box, which will take some input and generate output. Outputs for a given set of input combination are pre-calculated and are generated by the module.

6.2.2 SYSTEM TESTING:

Here all the pre tested individual modules will be assembled to create the larger system and tests are carried out at system level to make sure that all modules are working in synchronous with each other. This testing methodology helps in making sure that all modules which are running perfectly when checked individually are also running in cohesion with other modules. For this testing we create test cases to check all modules once and then a generated test combination of test paths throughout the system to make sure that no path is making its way into chaos.

6.2.3 INTEGRATED TESTING:

Testing is a major quality control measure employed during software development. Its basic function is to detect errors. Sub functions when combined may not produce than it is desired. Global data structures can represent the problems. Integrated testing is a systematic technique for constructing the program structure while conducting the tests. To uncover errors that are associated with interfacing the objective is to make unit test modules and built a program structure that has been detected by design. In a non - incremental integration all the modules are combined in advance and the program is tested as a whole. Here errors will appear in an endless loop function. In incremental testing the program is constructed and tested in small segments where the errors are isolated and corrected.

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

6.2.4 REGRESSION TESTING:

Each time a new module is added as a part of integration as the software changes. Regression testing is an actually that helps to ensure changes that do not introduce unintended behavior as additional errors.

Regression testing maybe conducted manually by executing a subset of all test cases or using automated capture play back tools enables the software engineer to capture the test case and results for subsequent playback and compression. The regression suit contains different classes of test cases.

A representative sample to tests that will exercise all software functions.
Additional tests that focus on software functions that are likely to be affected by the change.

6.2.5 FUNCTIONAL TESTING

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:

- Valid messages: Identified classes of valid messages must be accepted.
- Invalid messages: Identified classes of invalid messages must be rejected.
- Functions: Identified functions must be exercised.
- Output: Identified classes of application outputs must be exercised.
- Systems/Procedures: Interfacing systems or procedures must be invoked.

6.3 VALIDATION

PERFORMANCE ANALYSIS:

Confusion matrix for different cases

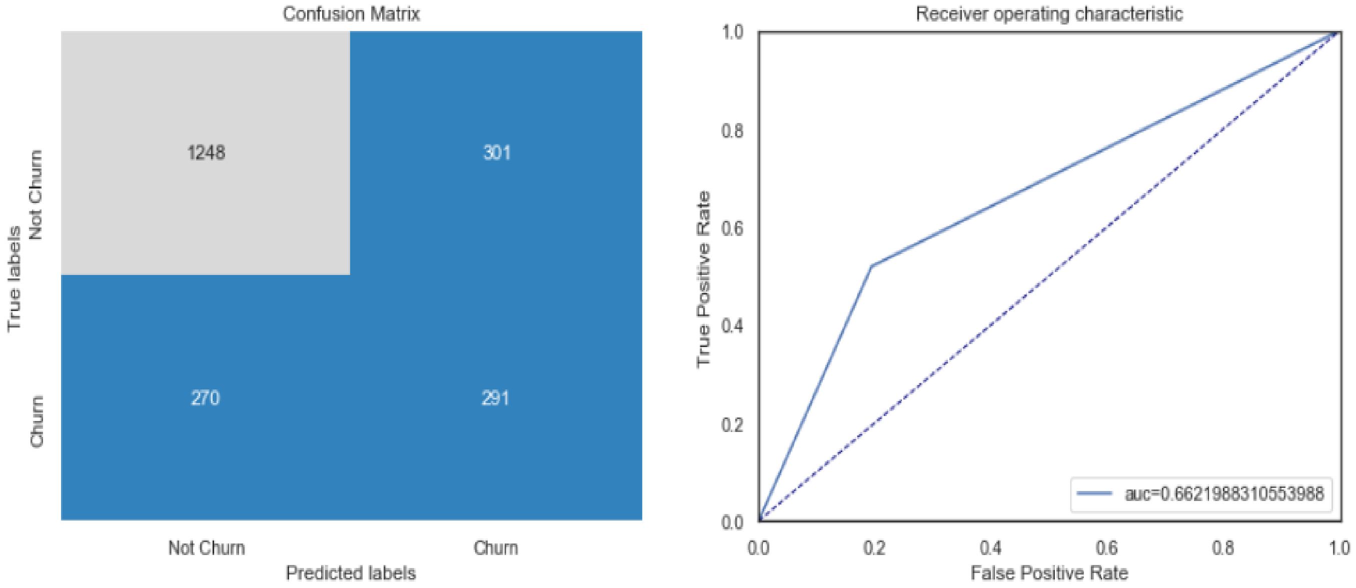


Figure 6.2

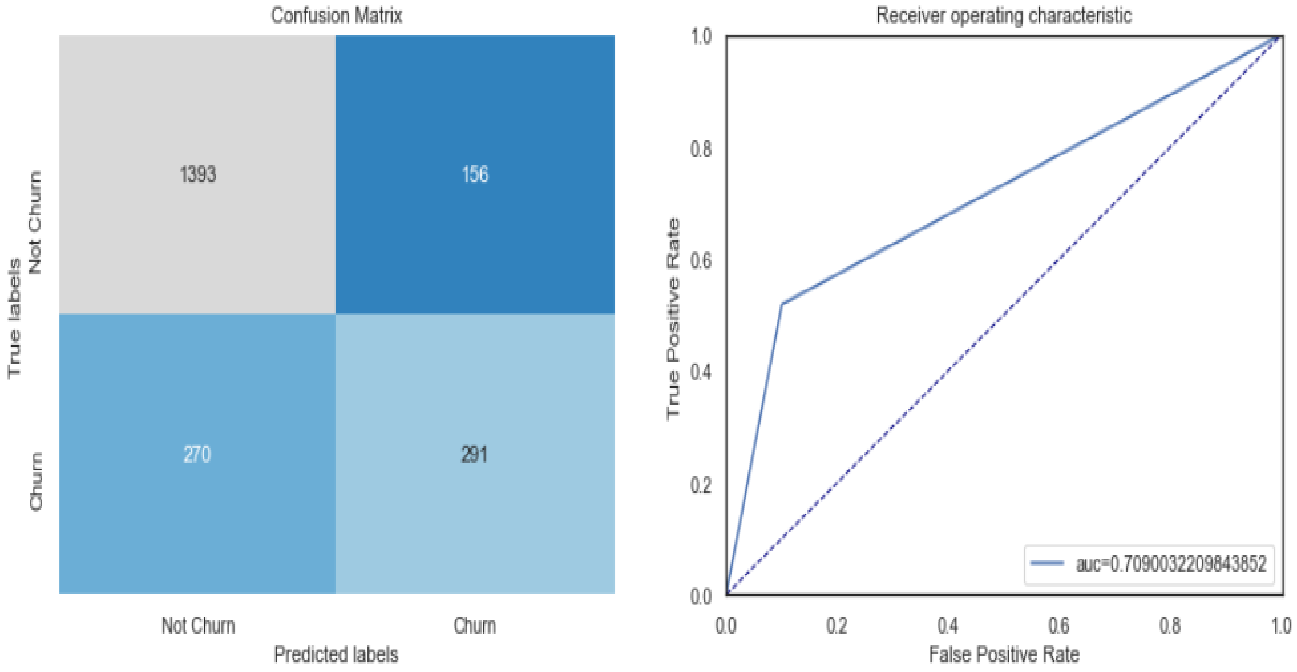


Figure 6.3

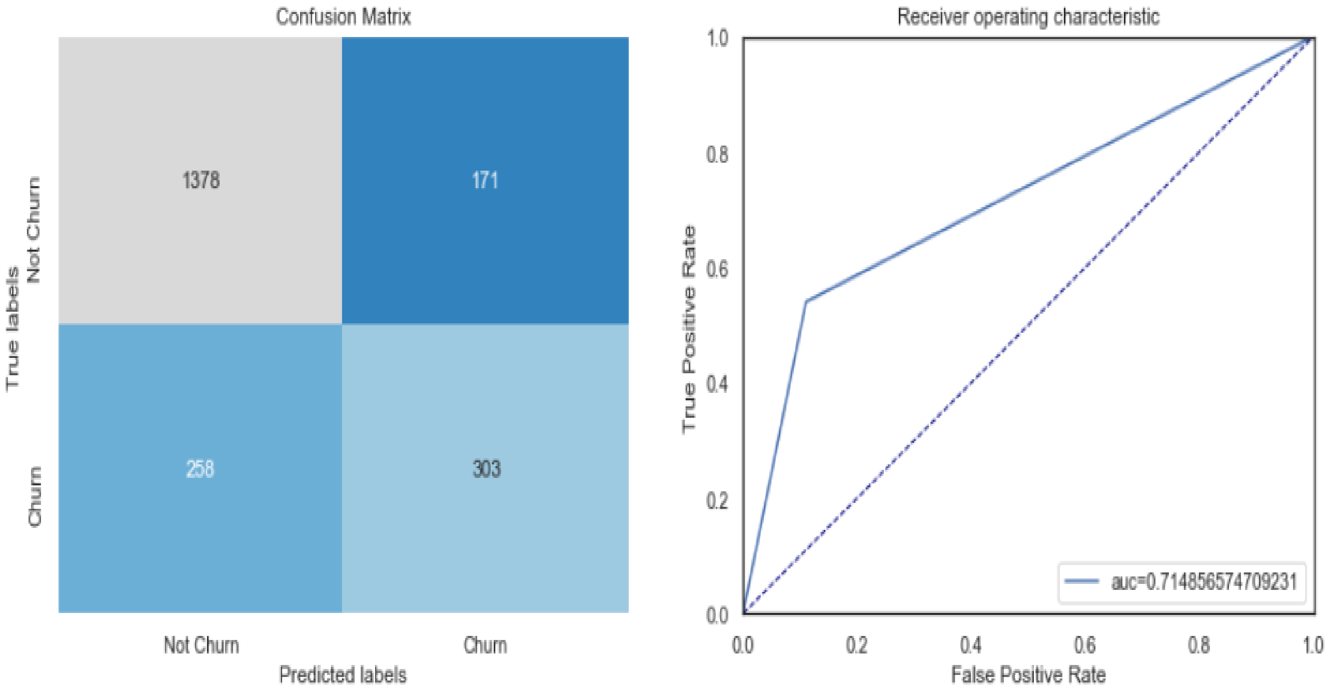


Figure 6.4

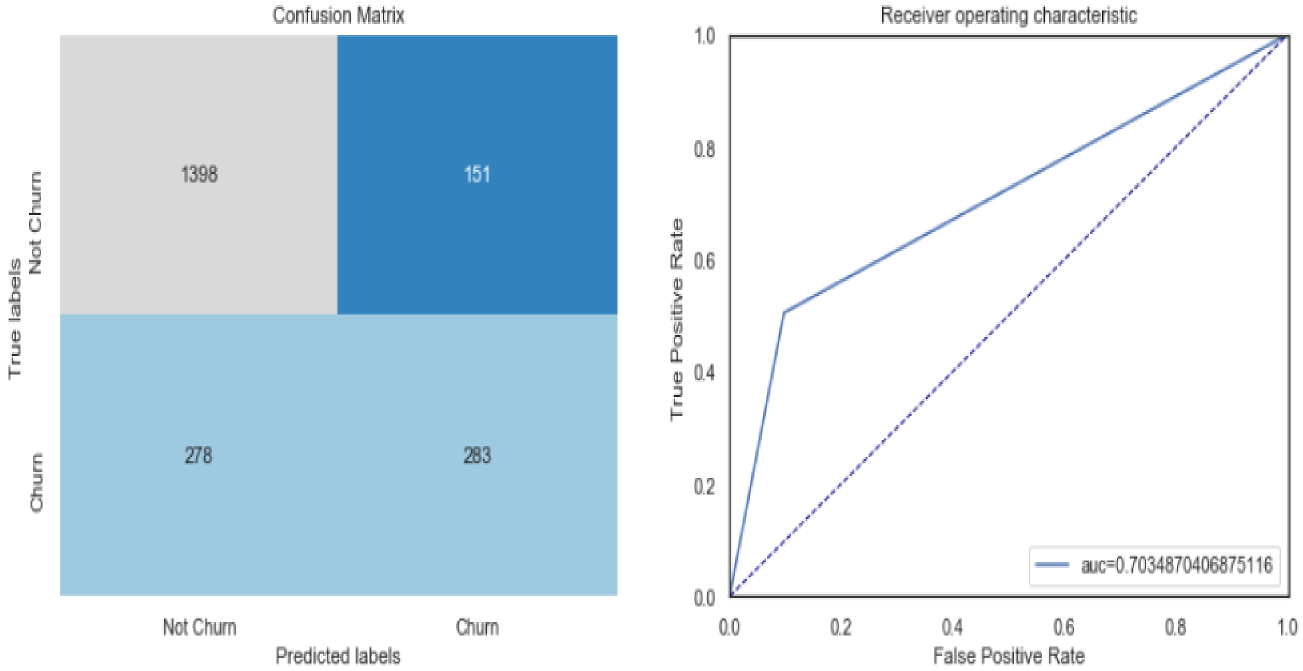


Figure 6.5

Figure descriptions:

Here in the results we are trying to show the accuracy of different algorithms by plotting them in different confusion matrix and with the false positive graph.

In the first figure (ie.,6.2) we are trying to use Decision Tree Classifier for the prediction of churn where the accuracy of the algorithm was 0.73.

In the second figure (ie.,6.3) we are trying to use Logistic Regression for the prediction of churn where the accuracy of the algorithm was 0.80

In the third figure (ie.,6.4) we are trying to use AdaBoost Classifier for the prediction of churn where the accuracy of the algorithm was 0.80

In the third figure (ie.,6.5) we are trying to use XGB Classifier for the prediction of churn where the accuracy of the algorithm was 0.80

So, from the above result we can conclude that most of the algorithms are having almost same accuracy value for the given data where Decision Tree Classifier was likely less accurate for my data where the Overall accuracy was likely equal to 0.80.

Now, we can conclude that in spite of the algorithm used sometimes the data that is used will also be considered to get the accurate values. Where as in real world the data which is being generated is huge and the data which I have considered was a part of it.

CONCLUSION

CONCLUSION

In the present competitive market of telecom domain, churn prediction is a significant issue of the CRM (Customer relation management) to retain valuable customers by identifying a similar group of customers and providing competitive offers/services to the respective groups. Therefore, in this domain, the researchers have been looking at the key factors of churn to retain customers and solve the problems of CRM. In this study, a customer churn model is provided for data analytics and validated through standard evaluation metrics. The obtained results show that our proposed churn model performed better by using several machine learning techniques. The message of this research is to show that free and open source technologies are matured enough for scientific computing domains. Python and DT are good points of start for researchers and students of computer vision

The system works satisfactorily for wide variations in illumination conditions and different types of number plates commonly found in India. It is definitely a better alternative to the existing proprietary systems, even though there are known restrictions.

FUTURE SCOPE

The future scope of this paper will use hybrid classification techniques to point out existing association between churn prediction and customer lifetime value. The retention policies need to be considered by selecting appropriate variables from the dataset. The passive and the dynamic nature of the industry ensure that data mining has become increasingly significant aspect in the telecommunication industry prospect.

An intelligent predictive churn analytics model, powered by Big Data analytics will allow businesses to process, analyze, and co-relate traditional and non-traditional metrics to achieve a holistic customer blueprint and effective insights that can trigger an alarm way before real damage is done.

REFERENCES

References

[1] K. B. Oseman, S.B.M. Shukor, N. A. Haris, F. Bakar, Data Mining in Churn Analysis Model for Telecommunication Industry, Journal of Statistical Modeling and Analytics, Vol. 1 No. 19-27, 2010.

[2] S.V. Nath, Customer Churn Analysis in the Wireless Industry: A Data Mining Approach, Technical Report, retrieved from http://download.oracle.com/owsf 2003/40332.pdf, April 14, 2014.

[3] D. V. Poel and B. Larivi. Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational Research, 157(1):196{217, 2004.

[3] V. Lazarov, M. Capota. Churn Prediction. Journal, Bus. Anal. Course. TUM Comput. Sci. 2007 publisherCiteseer.

[4] R. Baran, Christopher, M. Zerres, Customer Relationship Management. Book, 2012.

[5] B V Chowdary, A. G. Raju, B. Anuradha, R.Changala, Decision Tree Induction Approach for Data Classification Using Peano Count Trees. Volume 2, Issue 4, April 2012 ISSN: 2277 128X.

[6] A. Ntoulas, P. Zerfos, J.Cho. Downloading Textual Hidden Web Content Through Keyword Queries. In: 5th ACM/IEEE Joint Conference on Digital Libraries (Denver, USA, Jun 2005) JCDL05, pp. 100-109.

[7] Shin-Yuan Hung , David C. Yen , H. Wang, Applying data mining to telecom, Expert Systems with Applications 31 (2006) 515–524, Elseiver.

[8] V. Umayaparvathi, K. Iyakutti, Applications of Data Mining Techniques in Telecom Churn Prediction, International Journal of Computer Applications (0975 – 8887) Volume 42– No.20, March 2012.

[9] O.R. Zaiane,Introduction to Data Mining, CMPUT^() Principles of Knowledge Discovery in Databases Chapter,pp.1-15,1999.

[10] Andrew H. Karp, Using logistic regression to predict customer retention

[11] C. Wei, I. Chiu, Turning telecommunication call details to churn prediction :a data mining Approach expert System with applications (2002).

[12] Y. Wang, D. Chlang, M. Hua Hsu ,A recommender system to avoid customer churn ,"Expert System with application, 2009.

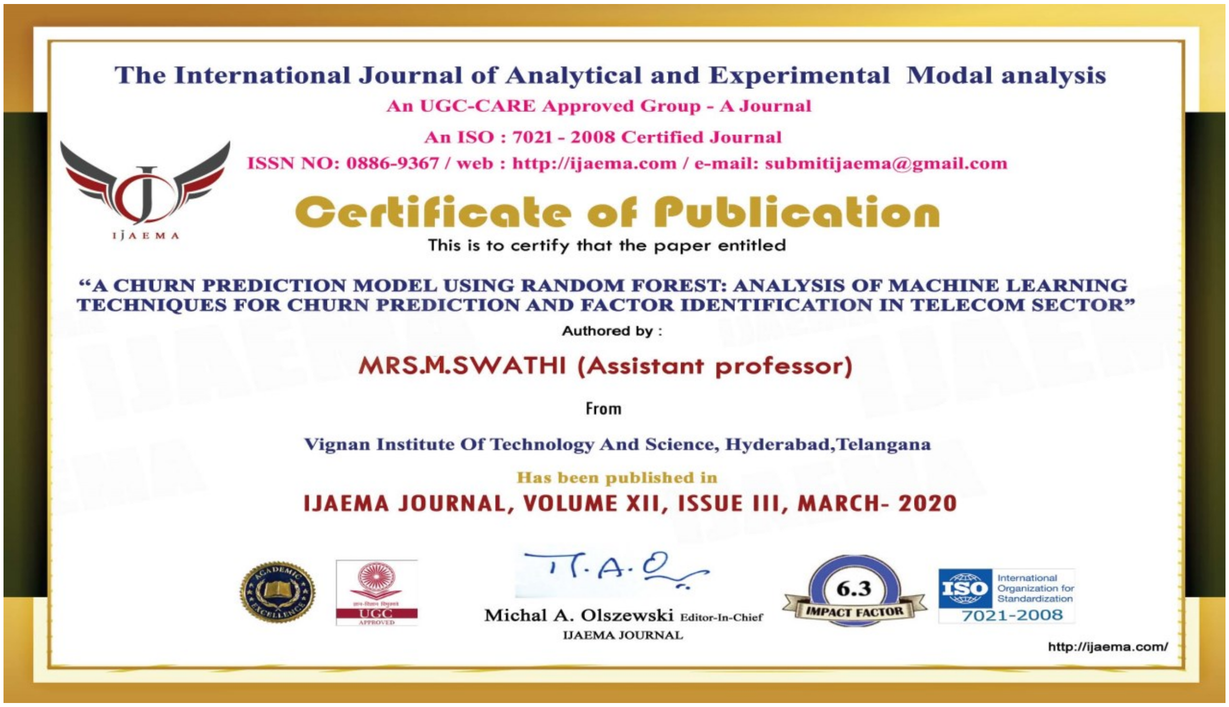
[14] R. Jadhav and U. T. Pawar, Churn Prediction in Telecommunication Using Data Mining Technology," in Proc. The (IJACSA) International Journal of Advanced Computer Scienc e and Applications, Vol. 2, No.2, February 2011.

[15] Emilía Huong Xuan Nguyen ,Customer Churn Prediction for the Icelandic Mobile Telephony Market , in proc. The Faculty of Industrial Engineering, Mechanical Engineering and Computer Science University of Iceland in September 2011.

[16] N. Kamalraj, .A.Malathi, Applying Data Mining Techniques in Telecom Churn Prediction, in proc. International Journal of Advanced Research in Computer Science and Software Engineering, 10, October 2013.

[17] L. Yangi , C. Chiu , Subscriber Churn Prediction in Telecommunications,

PUBLICATION



The International Journal of Analytical and Experimental Modal analysis

An UGC-CARE Approved Group - A Journal

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 0886-9367 / web : http://ijaema.com / e-mail: submitijaema@gmail.com

IJAEMA

Certificate of Publication

This is to certify that the paper entitled

"A CHURN PREDICTION MODEL USING RANDOM FOREST: ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR CHURN PREDICTION AND FACTOR IDENTIFICATION IN TELECOM SECTOR"

Authored by :

MRS.M.SWATHI (Assistant professor)

From

Vignan Institute Of Technology And Science, Hyderabad,Telangana

Has been published in

IJAEMA JOURNAL, VOLUME XII, ISSUE III, MARCH- 2020

ACADEMIC EXCELLENCE

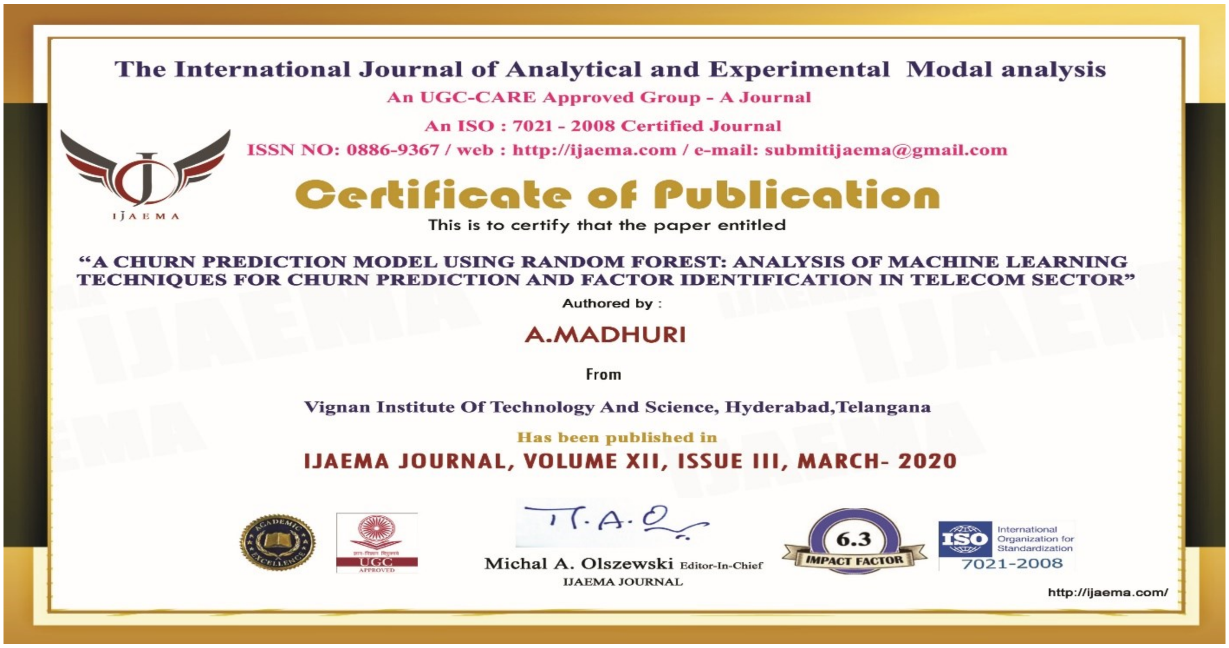
UGC APPROVED

Michal A. Olszewski Editor-In-Chief
IJAEMA JOURNAL

6.3 IMPACT FACTOR

ISO International Organization for Standardization 7021-2008

http://ijaema.com/



The International Journal of Analytical and Experimental Modal analysis

An UGC-CARE Approved Group - A Journal

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 0886-9367 / web : http://ijaema.com / e-mail: submitijaema@gmail.com

IJAEMA

Certificate of Publication

This is to certify that the paper entitled

"A CHURN PREDICTION MODEL USING RANDOM FOREST: ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR CHURN PREDICTION AND FACTOR IDENTIFICATION IN TELECOM SECTOR"

Authored by :

A.MADHURI

From

Vignan Institute Of Technology And Science, Hyderabad,Telangana

Has been published in

IJAEMA JOURNAL, VOLUME XII, ISSUE III, MARCH- 2020

ACADEMIC EXCELLENCE

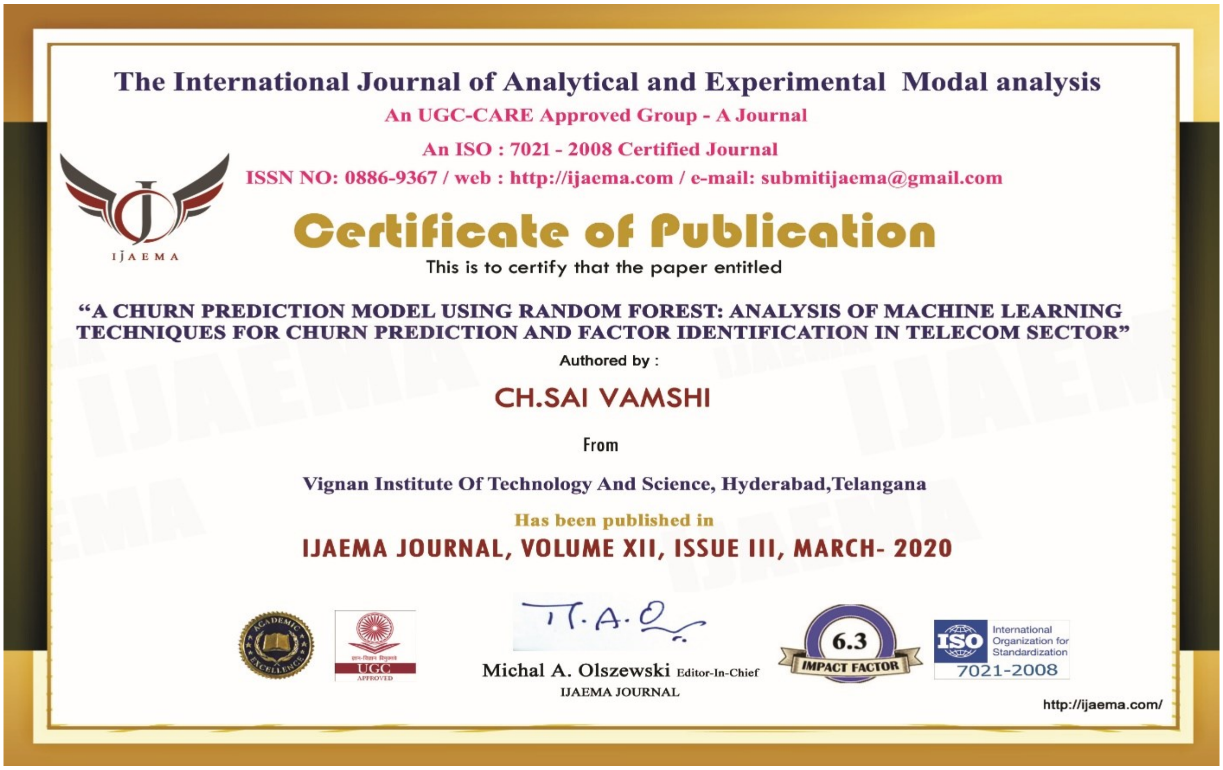
UGC APPROVED

Michal A. Olszewski Editor-In-Chief
IJAEMA JOURNAL

6.3 IMPACT FACTOR

ISO International Organization for Standardization 7021-2008

http://ijaema.com/



The International Journal of Analytical and Experimental Modal analysis

An UGC-CARE Approved Group - A Journal

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 0886-9367 / web : http://ijaema.com / e-mail: submitijaema@gmail.com

IJAEMA

Certificate of Publication

This is to certify that the paper entitled

"A CHURN PREDICTION MODEL USING RANDOM FOREST: ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR CHURN PREDICTION AND FACTOR IDENTIFICATION IN TELECOM SECTOR"

Authored by :

CH.SAI VAMSHI

From

Vignan Institute Of Technology And Science, Hyderabad,Telangana

Has been published in

IJAEMA JOURNAL, VOLUME XII, ISSUE III, MARCH- 2020

M.A.O

Michal A. Olszewski Editor-In-Chief

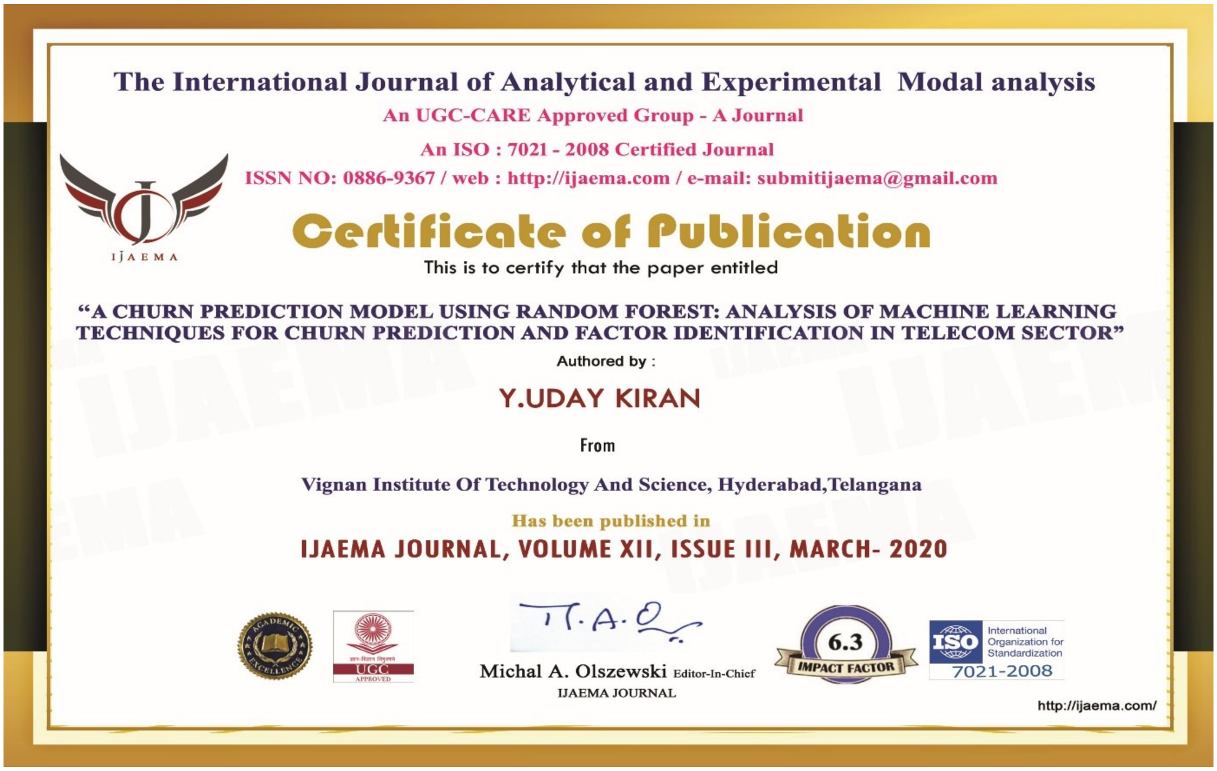
IJAEMA JOURNAL

6.3 IMPACT FACTOR

ISO International Organization for Standardization 7021-2008

UGC APPROVED

http://ijaema.com/



The International Journal of Analytical and Experimental Modal analysis

An UGC-CARE Approved Group - A Journal

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 0886-9367 / web : http://ijaema.com / e-mail: submitijaema@gmail.com

IJAEMA

Certificate of Publication

This is to certify that the paper entitled

"A CHURN PREDICTION MODEL USING RANDOM FOREST: ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR CHURN PREDICTION AND FACTOR IDENTIFICATION IN TELECOM SECTOR"

Authored by :

Y.UDAY KIRAN

From

Vignan Institute Of Technology And Science, Hyderabad,Telangana

Has been published in

IJAEMA JOURNAL, VOLUME XII, ISSUE III, MARCH- 2020

M.A.O

Michal A. Olszewski Editor-In-Chief

IJAEMA JOURNAL

6.3 IMPACT FACTOR

ISO International Organization for Standardization 7021-2008

UGC APPROVED

http://ijaema.com/