Project Report

On

Telecom Customer Churn Prediction

Submitted in partial fulfilment for the award of

Post Graduate Diploma in Big Data Analytics (PG-DBDA)

From Know-IT (Pune)

Guided by:

Mr. Amey Manjrekar

Submitted By:

Abhishek Gaikwad (210943025012)

Niharika Sharma (210943025031)

Rashmi Bhansali (210943025038)

Sanket Narkhede (210943025041)

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Abhishek Gaikwad (210943025012)

Niharika Sharma (210943025031)

Rashmi Bhansali (210943025038)

Sanket Narkhede (210943025041)

Have successfully completed their project on

Telecom Customer Churn Prediction

Under the guidance of Mr. Amey Manjrekar

ACKNOWLEDGEMENT

This project **"Telecom Customer Churn Prediction"** was a great learning experience for us and we are submitting this work to CDAC Know-IT (Pune).

We all are very glad to mention the name of **Mr. Amey Manjrekar** and **Mr. Sanjay Sane** for his valuable guidance to work on this project. His guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to **Mr. Vaibhav Inamdar** Manager (Know-IT), C-DAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS, Pune.

Our most heartfelt thanks goes to **Mrs. Bakul Joshi** (Course Coordinator,PG-DBDA) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC Know-IT, Pune.

From:

Abhishek Gaikwad (210943025012)

Niharika Sharma (210943025031)

Rashmi Bhansali (210943025038)

Sanket Narkhede (210943025041)

TABLE OF CONTENTS

ABSTRACT

1. INTRODUCTION

2. SYSTEM REQUIREMENTS

2.1 Software Requirements

2.2 Hardware Requirements

3. FUNCTIONAL REQUIREMENTS

4. SYSTEM ARCHITECTURE

5. METHODOLOGY

6. DATA VISUALIZATION AND REPRESENTATION

7. RESULT AND FINDING

8. CONCLUSION AND FUTURE SCOPE

References

Abstract

Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, especially in the telecom field, companies are seeking to develop means to predict potential customer to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn. The Objective is to develop a churn prediction model which assists companies to predict customers who are most likely subject to churn. We are processing ETL operation on top of that data using Apache Spark and dumping required data into Mongo Db for data visualization through PowerBI and build Machine Learning for estimating the customer churn rate. Prediction of customer churns help in planning process and strategic decision making in large companies.

1. INTRODUCTION

- Churn is the process of customers switching from one firm to another in given time. Retaining the existing customers is more profitable than fetching the new customers. The Companies concentrate to the extant customers to avert churn. A churn prediction model is needed to predict the churners.

- The fast growth of marketplace in every business is giving rise to increased subscriber base. Accordingly, companies have recognized the significance of retaining the customers who is on hand. It has become necessary for service-providers to reduce the churn rate of customers since the inattention might negatively influence profitability of the company.

- Churn prediction contributes to identify those users who are likely to switch a company over another. The dataset used for customer churn is of telecom industry. It is collected from www.github.com that contains customer and service information for telecom industry.

Datasets and features:

Data used in the project is structured in nature. It was collected from www.github.com .The Objective is to develop a churn prediction model which assists companies to predict customers who are most likely subject to churn.Logistic Regression, Decision Tree, Random forest , Stacking Classifier and Voting Classifier models were used to predict customer churn rate.

2. SYSTEM REQUIREMENTS

Hardware Requirements:

- Platform – Windows 10
- RAM – 8 GB of RAM
- Peripheral Devices – Mouse, Keyboard, Monitor
- A network connection for data recovering over network.

Software Requirements:

- Python 3
- Apache Spark
- MongoDB
- PowerBI
- **OS – Window**

3. FUNCTIONAL REQUIREMENTS

(1) Python 3:

- Python is a general purpose and high level programming language.
- It is use for developing desktop GUI applications, websites and web applications.
- Python allows to focus on core functionality of the application by taking care of common programming tasks.
- Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68,
- Small Talk, and UNIX shell and other scripting languages.

(2) Apache Spark:

- Apache Spark is an open-source cluster computing system that provides high-level API in Java, Scala, Python and R.
- Apache Spark is one of the fastest-growing big data projects in the history of the Apache Software Foundation. With its memory-oriented architecture, flexible processing libraries, and ease-of-use, Spark has emerged as a leading distributed computing framework for real-time analytics.
- Spark is used for many types of data processing – it comes packaged with support for machine learning, interactive queries (SQL), statistical queries with R, graph processing, ETL, and streaming.
- For loading and storing data, Spark integrates with a number of storage MongoDB, and more.

(3) MongoDB:

- **MongoDB**, the most popular NoSQL database, is an open-source document-oriented database.
- MongoDB allows a highly flexible and scalable document structure.
- MongoDB has built in solution for partitioning and sharing your database.
- MongoDB provides a variety of storage engines, allowing you to choose one most suited to your application.
- A real-life scenario for this kind of data manipulation is storing and querying real-time, intraday market data in MongoDB.

(4)PowerBI:

- Data visualization is the graphical representation of information and data.
- It helps create interactive elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- PowerBI is widely used for Business Intelligence but is not limited to it.
- It helps create interactive graphs and charts in the form of dashboards and worksheets to gain business insights.
- All of this is made possible with gestures as simple as drag and drop.

Data Cleaning Process:

Fig: Data Cleaning Process

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data

validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

4. SYSTEM ARCHITECTURE

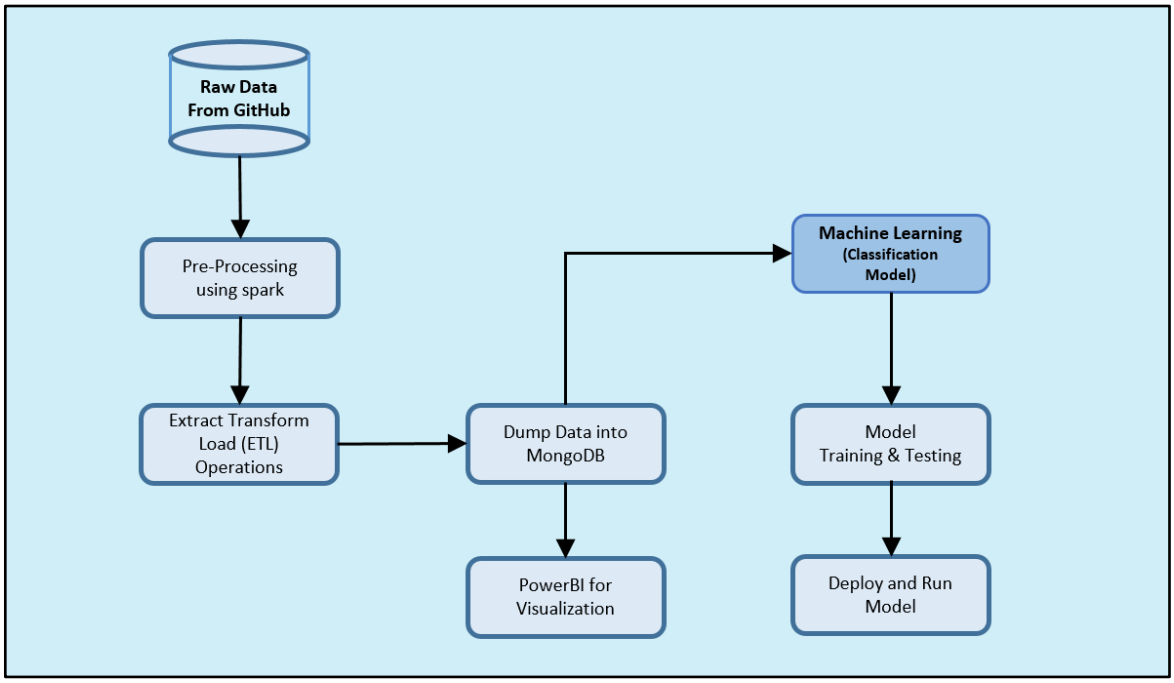


Fig: System Architecture of Telecom Customer Churn Prediction

5. METHODOLOGY

In this project we have applied various different types of Classification.

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Stacking Classifier
- Voting Classifier
- Principal component analysis (PCA)

During the implementation we analyze the accuracy of all the algorithms.

Machine Learning Algorithms

Logistic Regression

Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

Decision Tree Classifier

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

Random Forest Classifier

It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."
Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

Stacking Classifier

In stacking, a new model is trained from the predictions from various models Predicted columns act as features with response variable as the original one Usually we stack weaker models at lower level and stronger models at upper level But, there is no such rule as to which model to be used at lower level and higher level.

Voting Classifier

The voting classifier aggregates the predicted class or predicted probability on basis of hard voting or soft voting. So if we feed a variety of base models to the voting classifier it makes sure to resolve the error by any model.
Hard Voting: Voting is calculated on the predicted output class.
Soft Voting: Voting is calculated on the predicted probability of the output class.

Principal component analysis (PCA)

Principal component analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest. The importance of each component decreases when going to 1 to n, it means the 1 PC has the most importance, and n PC will have the least importance. In this technique we selected 16 PCA and performed various Machine Learning algorithms.

1.Decision Tree

2.Random forest

3.SVC(linear,rbf,poly)

6. DATA VISUALIZATION AND REPRESENTATION

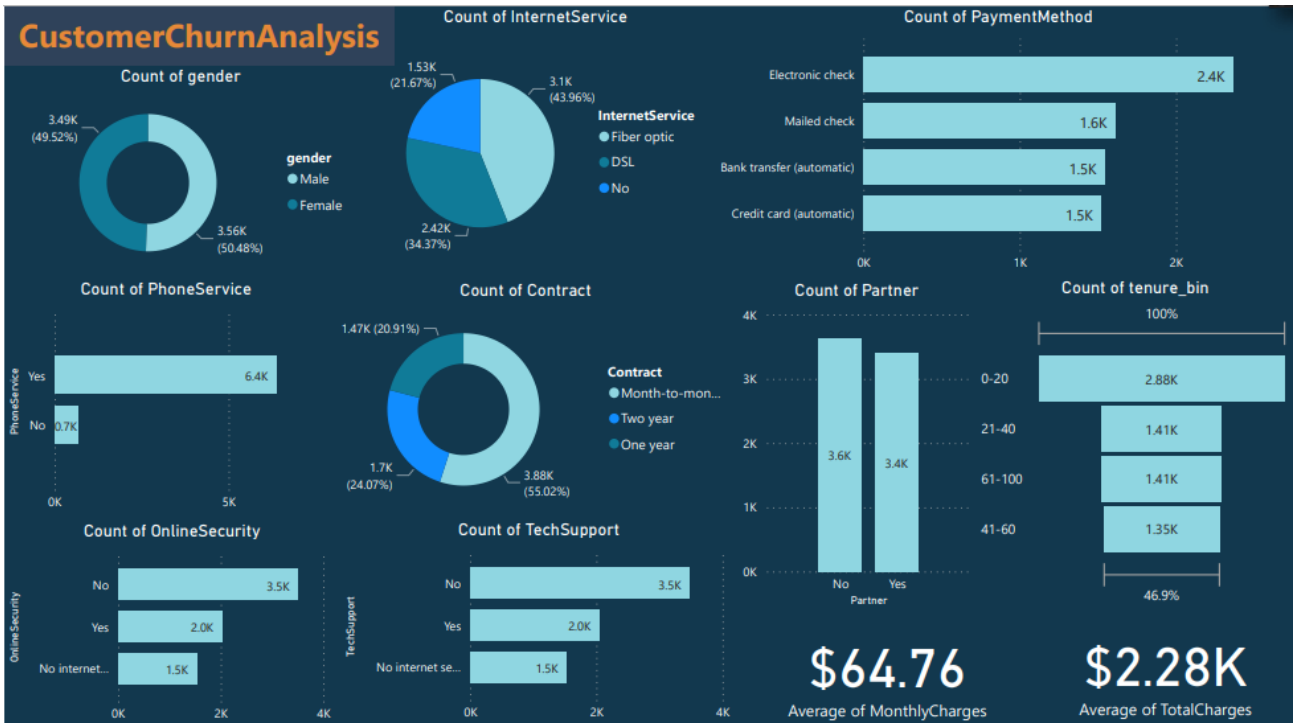


Fig.: CustomerChurn Prediction Dashboard

Fig.: Customer Category (Partner,Gender,SeniorCitizen,Dependents) Churn Percentage

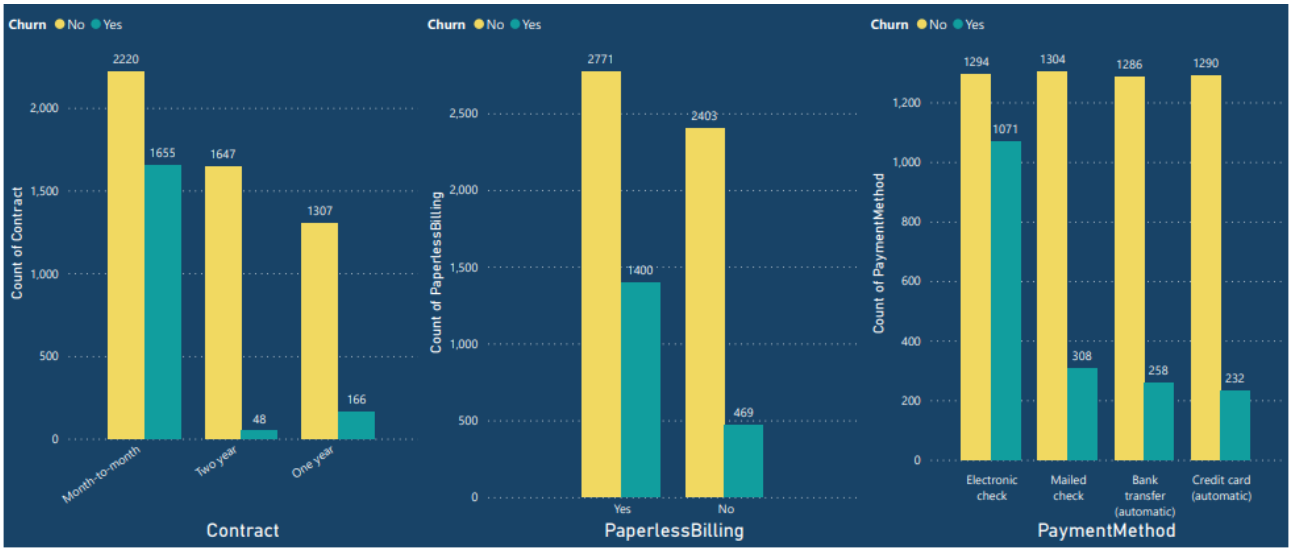


Fig.: Customer Account Information Churn Analysis

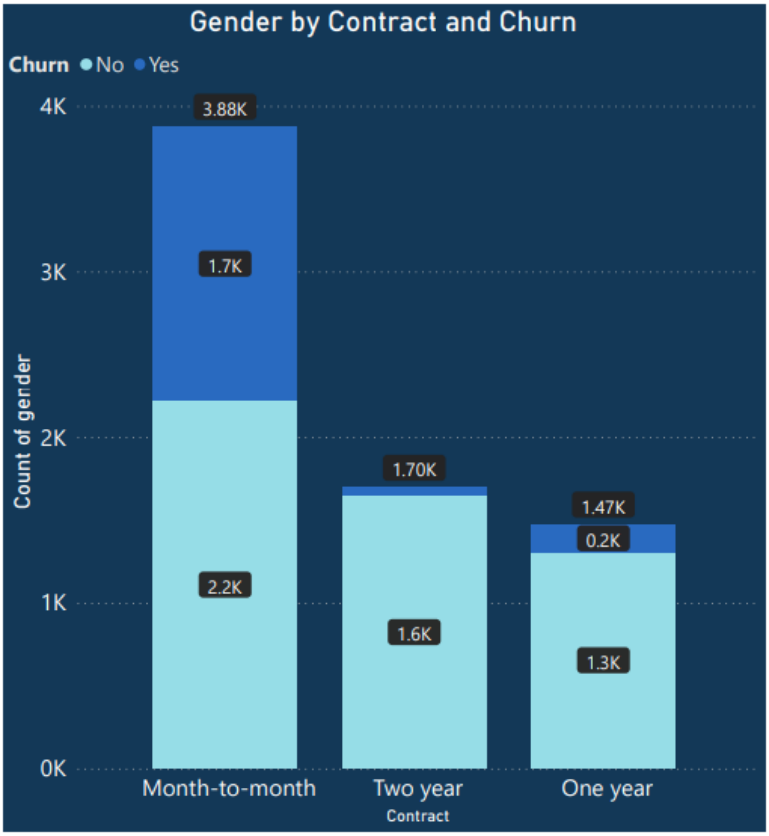


Fig.: Gender By Contract Churn Analysis

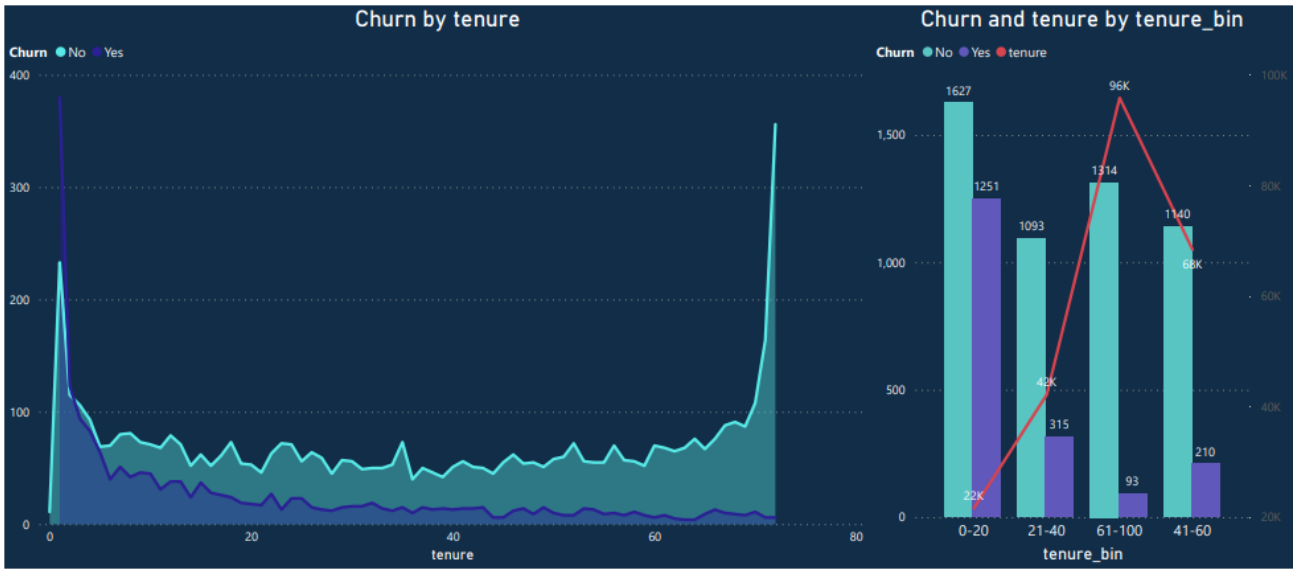


Fig.: Tenure and Churn Analysis

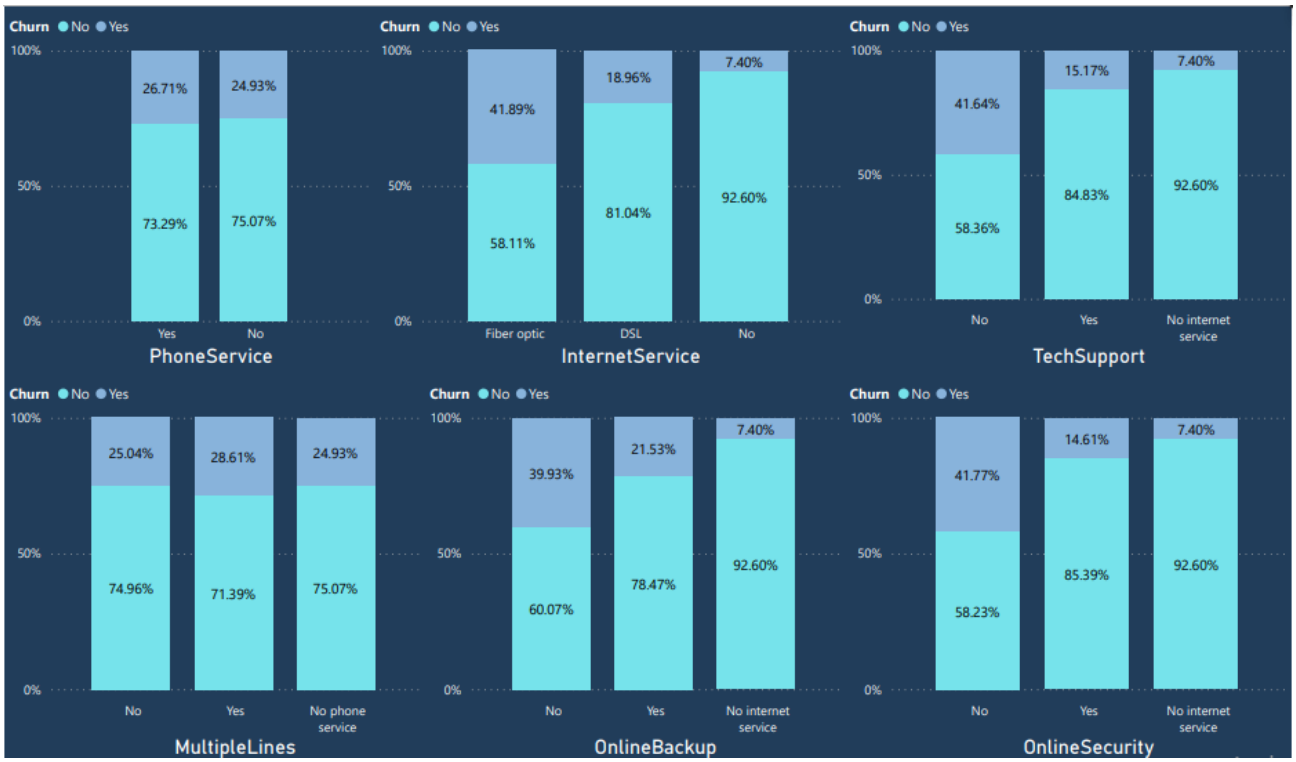
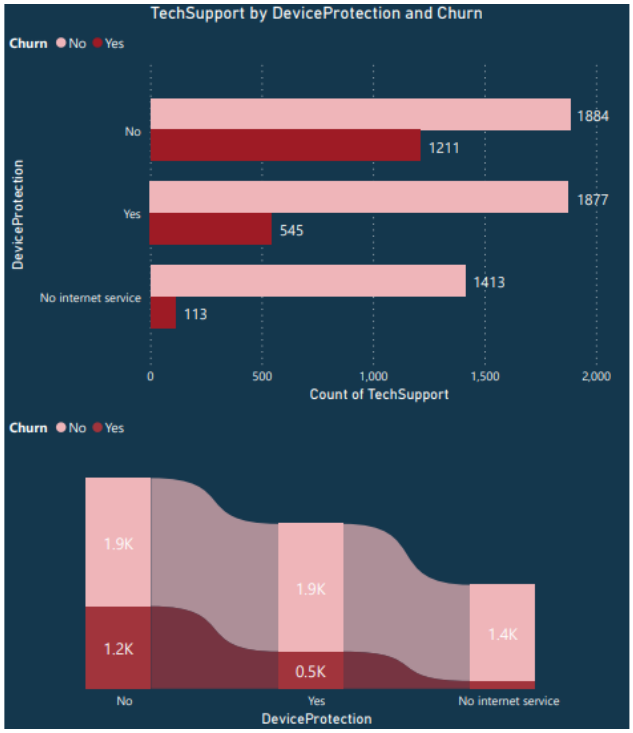


Fig.: Service Information Churn Analysis



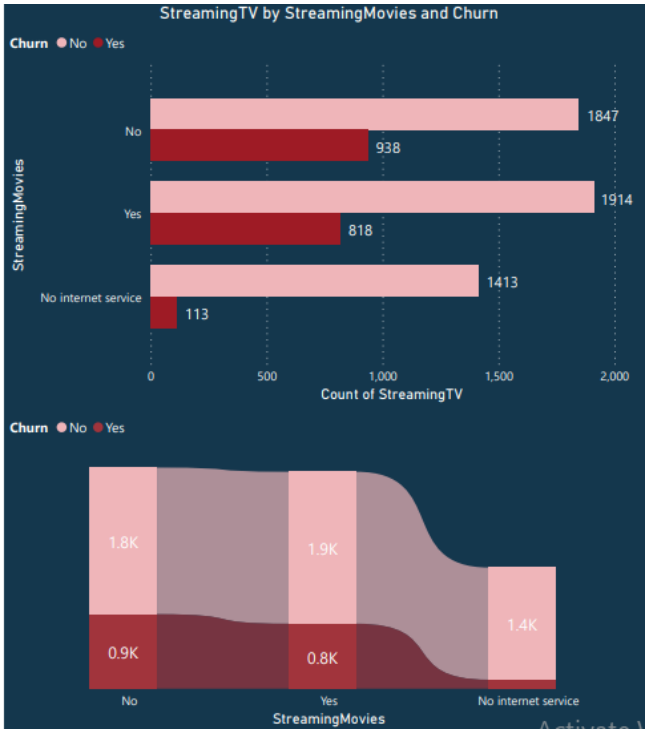


Fig.: TechSupport, Device Protection and Streaming services churn Analysis

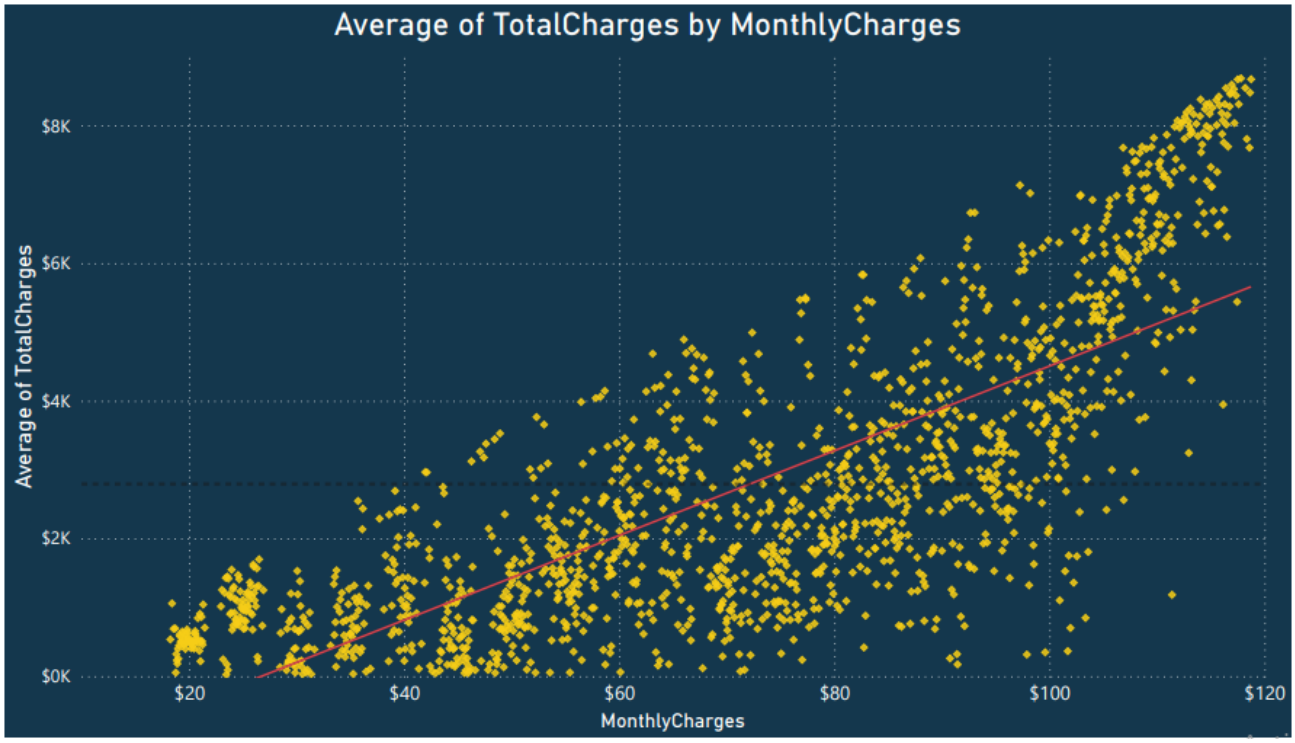


Fig.: Average of TotalCharges by Monthly Charges

7. Result and Finding

Actual Data

Algorithm	Roc_Auc	Accuracy
Logistic Regression	73.639%	79.433%
DECISION TREE	78.639%	75.177%
RandomForestClassifier	79.474%	78.156%

Algorithm	Roc_Auc	Accuracy
StackingClassifier	82.126%	80.567%
Voting Classifier	81.507%	75.177%
PCA (svc_rbf)	76.869%	73.475%
PCA(svc_linear)	80.265%	78.440%
PCA (DecisionTreeClassifier)	63.314%	71.631%

Random Over Sampled Data

Algorithm	Roc_Auc	Accuracy
Logistic Regression	80.532%	76.170%
DECISION TREE	75.766%	74.610%
RandomForestClassifier	79.899%	76.170%

Algorithm	Roc_Auc	Accuracy
StackingClassifier	80.981%	76.738%
Voting Classifier	89.570%	77.343%
PCA (svc_rbf)	89.532%	79.855%
PCA(svc_linear)	81.020%	76.170%
PCA (DecisionTreeClassifier)	67.041%	72.908%

8. CONCLUSION AND FUTURE SCOPE

Telecom customer churn is a central issue for telecom companies, since it decreases profits Furthermore, preventing customer churn is crucial As the global telecom industry is becoming more dissolving and companies are increasingly struggling to retain customers .Currently, most companies invest heavily in marketing to attract new customers. However, keeping existing customers is cheaper than acquiring new customers. Thus, it is becoming more critical and a significant concern for telecommunication companies to prevent customer churn.

We used various classification models to predict telecom churn using customer churn data of telecom industry.

Moreover, the results of this project will give us the ability to predict customer behavior and loss accurately and to optimize their strategies to improve customer retention rates.

Meanwhile, the findings will help companies reduce costs and optimize their budgets. Furthermore, for telecom companies, it will be possible to improve customer targeting through the results of this project and to increase the profits of telecom companies.

References

- https://raw.githubusercontent.com/IBM/telco-customer-churn-on-icp4d/master/data/Telco-Customer-Churn.csv
- https://towardsdatascience.com/using-principal-component-analysis-pca-for-machine-learning-b6e803f5bf1e
- https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html
- https://imbalanced-learn.org/stable/over_sampling.html
- https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
- Breiman, L., "Random forests," Machine learning, Vol. 45, No. 1, 2001, pp. 5–32.
- https://www.investopedia.com/terms/c/churnrate.asp