IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

Telecom Customer Churn Prediction using SMLT

Reni Hena Helan R[1], Aruljothi K[2], Raguvarman K[3], Yashwanthraj J S[4], Venkatraj R[5]
Assistant Professor, Department of Computer Science[1,2]
Students, Department of Computer Science[3,4,5]
Dhanalakshmi College of Engineering, Chennai, India

Abstract: *Customer churn is a significant issue and one of the top issues for big businesses. Companies are working to create methods to predict probable customer churn because it has a direct impact on their revenues, particularly in the telecom industry. In order to reduce customer churn, it is crucial to identify the variables that contribute to this churn. Our work's key contribution is the creation of a churn prediction model that helps telecom providers identify consumers who are most likely to experience churn. This project's objective is to present a fresh method for identifying potential customers who might leave so that marketing retention tactics can be created accordingly. The historical dataset is gathered and used to create a machine learning algorithm model. The required pre-processing methods, such as univariate and bivariate analysis, are put into practice. In order to better understand the properties of the data, it is visualized. A classification model is then developed using a machine learning algorithm, and the effectiveness of the various algorithms is compared using metrics like accuracy, F1 score recall, etc.*

Keywords: Customer Churn, Machine Learning, SVM, Logisitic Regression, Decision Tree**.**

I. INTRODUCTION

The telecommunications business is one of the largest in the world, and as a result, competition has intensified due to technological advancements and an ever-growing number of operators. In order to survive in this competitive industry, telecom companies have developed a number of strategies that aim to generate sizable amounts of income. It's crucial for businesses to reduce the likelihood of customer churn, also known as "the movement of customers from one service provider to another service provider," in order to increase customer retention rates. With more competitive service industries, it has been highlighted that customer churn is a significant problem. To anticipate this circumstance, machine learning applications are becoming increasingly useful. The fundamental idea behind predicting client attrition in terms of The telecom sector must estimate the number of subscribers who are actually considering leaving a company they have previously used and offer measures to stop big churns. In the current climate of intense competition among businesses, it is now vital to estimate churners before they leave. The crucial role that the telecom sector plays makes it even more important to develop prediction techniques in conjunction with churn prediction. Only a few studies demonstrate the importance of user retention in this sector. One study shows that a 1% increase in customer retention may effectively result in a 5% gain in all of the company shares. Another study showed that the monthly ratio of customer turnover in the telecommunications sector might increase by 5%. customer churn was 27% each year, and the annual rate was 2.2%. As a result of more competitive services, keeping customers in the telecom sector has already proven impossible. proposed a sophisticated data mining technique employing the machine learning algorithms (neural networks) and svm (support vector machines) to identify customer churn. The second set of data focused on how well machine learning algorithms anticipate client attrition. He and colleagues (2009) employed a machine learning model to address the problem of customer attrition in major telecommunications companies. Huang et al. (2015) looked into the problem of customer turnover on the big data platform with the goal of highlighting how big data considerably advances the process of calculating attrition based on variety. Speed and quantity of the statistics. Ahmad, et al (2019) distinctive that the social network analysis application attributes enhance the results of estimating the churn in telecommunication zone. Amongst all other sectors, over the a long time telecom zone are going through the most yearly price of churn starting from 20 to 40 percentage this basically has financial outcomes on a firm, because it bears 5 to ten instances lots to bring in a brand new user than keeping an old user in the firm nowadays organizations intend to build strong courting with their users therefore, it has come to be a conviction that the satisfactory promotional

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

coverage is to keep the old customers or extra simply to cope with patron emphasized that system getting to know approaches paintings with high dimensional, massive, nonlinear datasets with stepped forward prediction accuracy but taken into consideration complicated close to actual-global programs.

II. EXISTING MODEL

User churn is a significant problem for internet businesses that jeopardizes their viability and profitability. The majority of earlier studies on churn prediction convert the issue into a binary classification process in which the users are classified as either churned or not. Recently, some efforts have attempted to translate the prediction of user churn into the prediction of user return time. In this strategy, which is more accurate for real-world online services, the model forecasts the user return time at each time-step rather than a churn label? The earlier studies in this field, however, lack generality and demand complicated computations. We introduce ChOracle in this work, an oracle that forecasts user churn by simulating user return.by combining Temporal Point Processes and Recurrent Neural Networks to reduce service response times. In order to simulate the latent user loyalty to the system, we also add latent factors into the proposed recurrent neural network.

2.1 Disadvantages:

- They are not implementing machine learning algorithms for prediction
- They are not getting the accuracy of their model.
- They are not implementing the deployment part

III. PROPOSED MODEL

The suggested approach involves applying machine learning to create a prediction of telecom customer churn. We intend to create an AI-based model. Data are required to train our model. Therefore, the dataset for telecom customer churn can be used to train the model. We need to be aware of the training intentions in order to make use of this dataset. An intent is the purpose for which a user interacts with a predictive model or the purpose for which each piece of data that a certain user provides to the model is provided. These intents may differ from one another depending on the domain for which you are designing an AI solution. The plan is to create training examples for each individual intent, define those intents, and instruct your AI accordingly. Model training data comprised of those training sample data, with model training categories representing intentions. The vectorization technique, which uses the vectors to comprehend the data, is used to develop the model. We can get a better AI model with the highest accuracy by using multiple algorithms. Following model construction, the model is assessed using a variety of measures, including confusion metrics, precision, recall, sensitivity, F1 score, and others.

3.1 Advantages

- We are making a predictive AI model for Telecom Customer Churn Prediction.
- We are implementing Machine Learning Algorithms for getting more accuracy.
- We can calculate performance metrics for better performance results.

IV. LIST OF MODULES

- Data Pre-processing
- Data Analysis of Visualization
- Implementing Support Vector Machine
- Implementing Logistic Regression
- Implementing Random Forest Classifier
- Implementing Decision Tree Classifier

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

4.1 Module Description

A. Data Pre-processing:

The error rate of the machine learning (ML) model is obtained using validation procedures, and is thought to be as close to the actual error rate of the dataset as possible. You might not need the validation approaches if the data volume is sizable enough to be representative of the population. However, working with data samples that could not be a realistic reflection of the population of a given dataset in real-world circumstances. Finding duplicate values, missing values, and information about the data type— whether a float variable or an integer—are all necessary. the data sample that was used to objectively assess how well a model fit the training dataset while adjusting model hyper parameters.

As skill from the validation dataset is incorporated into the model setup, the evaluation gets increasingly skewed. A given model is evaluated using the validation set, although this is done frequently. This information is used by machine learning developers to adjust the model hyper parameters. The process of gathering data, analysing it, and dealing with its structure, quality, and substance can take a lot of time. Understanding your data and its characteristics can assist you choose the algorithm to utilize to create your model during the data identification phase.

Several distinct data cleaning jobs utilizing Python's Pandas module, with a focus on missing values possibly the biggest data cleaning task and the ability to clean data more quickly. Less effort should be spent cleaning data, and more time should be spent investigating and modelling.

Some of these sources contain merely careless errors. Sometimes there may be a more significant cause for missing data. It's crucial from a statistical standpoint to comprehend these various missing data kinds. The kind of missing data will affect how it is handled in terms of filling in the blanks, identifying missing values, basic imputation, and a thorough statistical methodology. Prior to entering code, it's crucial to know where the missing data are coming from. Here are some typical reasons why data is missing

- User neglected to complete a field.
- While manually transferring data from a legacy database, data was lost.
- A programming problem occurred
- Users declined to enter information in a field because they had preconceived notions about how the results would be used or interpreted. Variable identification with univariate, bivariate, and multivariate analysis
- Import libraries for access and functionality; read the dataset
- General properties of analysis
- Display dataset as data frame
- Show columns
- Shape of data frame
- Describe dataset
- Check data type and dataset information
- Check for duplicate data
- Check missing values of data frame
- To rename and delete the provided data frame
- To describe the kind of values, to add more columns, and more

B. Data Validation/ Cleaning/Preparing Process:

Loading the specified dataset while importing the library packages. To evaluate the missing values, duplicate values, and variable identification by data type, shape, and size. A validation dataset is a sample of data withheld from model training and used to measure model competence when fine-tuning models and processes that you may employ to make the best use of validation and test datasets when assessing your models. To analyse the univariate, bivariate, and multivariate processes, data cleaning and preparation steps include renaming the provided dataset and deleting columns, among other things. Depending on the dataset, different procedures and methods will be used to clean the data. Data cleaning's main objective is to find and eliminate mistakes and abnormalities in order to maximise the value of analytics and decision-making using data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

C. Exploration data analysis of visualization:

In applied statistics and machine learning, data visualisation is a crucial ability. In fact, the main focus of statistics is on numerical estimates and descriptions of data. An essential set of tools for obtaining a qualitative understanding is provided by data visualisation. This can be useful for discovering trends, corrupt data, outliers, and much more when exploring and getting to know a dataset. Data visualisations can be used to communicate and illustrate crucial relationships in plots and charts that are more visceral and engaging to stakeholders than measurements of association or significance with a little subject knowledge. It will recommend a closer look at some of the books mentioned at the conclusion because data visualisation and exploratory data analysis are entire topics in themselves. Data may not always make sense until it can view in a visual format, like using graphs and charts. Both applied statistics and applied machine learning value the ability to easily visualise data samples and other objects. It will show you how to utilise various plot types to comprehend your own data and the many plot types you'll need to be familiar with when visualising data in Python.

- How to visualise categorical data using bar charts and time series data using line graphs.
- How to use histograms and box plots to summarise data distributions.

D. Comparing Algorithm with prediction in the form of best accuracy result:

It is crucial to systematically compare the performance of various machine learning algorithms, and it will become clear that Scikit-learn in Python may be used to build a test harness for this purpose. You can apply this test harness as a model for your own machine learning issues and include additional and various algorithms to contrast. There will be variations in the performance attributes of each model. You may unobserved data by using resampling techniques like cross validation. It must be able to select one or two of the best models from the group of models you have developed using these estimates. It is a good idea to use new datasets to visualise the data utilising diverse methods in order to view the data from various angles. The choice of models follows the same logic. In order to select the one or two that will be used for finalisation, you should examine the estimated accuracy of your machine learning algorithms in a variety of methods. Using various visualisation techniques to display the average accuracy, variance, and other characteristics of the distribution of model accuracies is one way to achieve this. You will learn precisely how to achieve that in Python using Scikit-learn in the following section. Making sure that each algorithm is evaluated uniformly on the same data is essential for conducting a fair comparison of machine learning algorithms, and this may be done by requiring that each algorithm be tested using a uniform test harness.

In the example below 4 different algorithms are compared:

- Support Vector Machine
- Logistic Regression
- Random Forest Classifier
- Decision Tree Classifier

Each algorithm is tested using the K-fold cross validation technique, which is crucially configured with the same random seed to guarantee that the splits to the training data are carried out consistently and that each algorithm is evaluated in the same manner. Prior to the comparison algorithm, installing Scikit-Learn libraries and creating a machine learning model. Pre-processing, a linear model with the logistic regression method, cross-validation using the K Fold method, an ensemble with the random forest method, and a tree with a decision tree classifier must all be completed in this library package. Separating the train set and test set is also a good idea. to compare accuracy and anticipating the outcome.

The adjustments made to our data prior to feeding it to the algorithm are referred to as pre-processing. Data Pre-processing is a method for transforming unclean data into clean data sets. In other words, anytime data is acquired from various sources, it is done so in a raw manner that makes analysis impossible. The data must be organised properly in order to get better results from the machine learning method used to apply the model. Some machine learning models have specific information requirements, such as the Random Forest method not supporting null values. Therefore, null values from the initial raw data collection must be controlled in order to run the random forest method. And that data

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

set is another facet. Should be structured to allow for the execution of many Machine Learning and Deep Learning algorithms on a given dataset.

V. ALGORITHM AND TECHNIQUES Algorithm Explanation:

Classification is a supervised learning strategy used in machine learning and statistics, where a computer programme learns from the data input provided to it and then applies this learning to classify fresh observations. This data set may be multiclass or it may just be bi-class (for example, indicating if the individual is male or female or whether the email is spam or not). Speech recognition, handwriting recognition, biometric identity, document classification, etc. are a few instances of classification issues. Algorithms are taught by supervised learning using labelled data. After gaining a grasp of the data, the algorithm decides which label new data should receive based on patterns and associations with the previously unlabeled new data.

Used Python Packages:

sklearn:

In python, sklearn is a machine learning package which include a lot of ML algorithms.

Here, we are using some of its modules like train test split, Decision Tree Classifier or Logistic Regression and accuracy score

NumPy:

It is a numerical Python module that offers quick mathematical operations for calculations.

It is used to manipulate data and read data from numpy arrays.

Pandas:

Used to read and write different files.

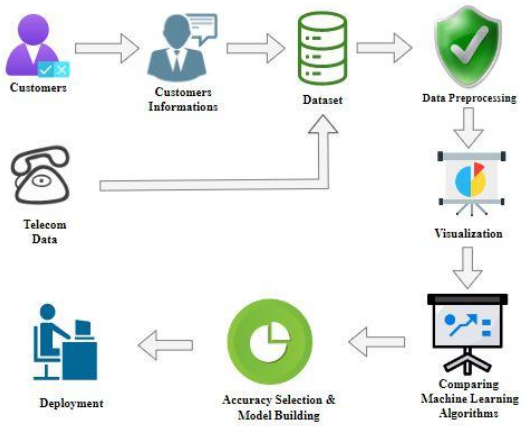
Data manipulation can be done easily with data frames.

Matplotlib:

Data visualization is a useful way to help with identify the patterns from given dataset.

Data manipulation can be done easily with data frames

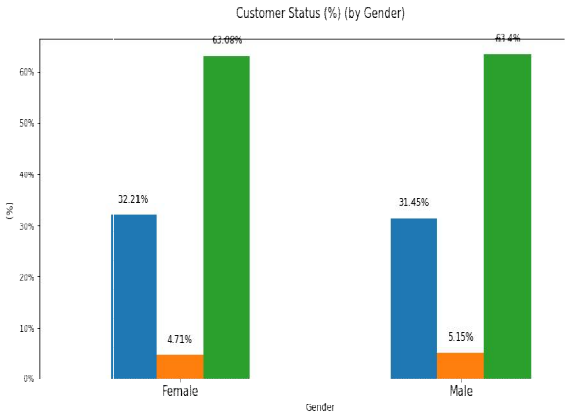
VI. SYSTEM ARCHITECTURE



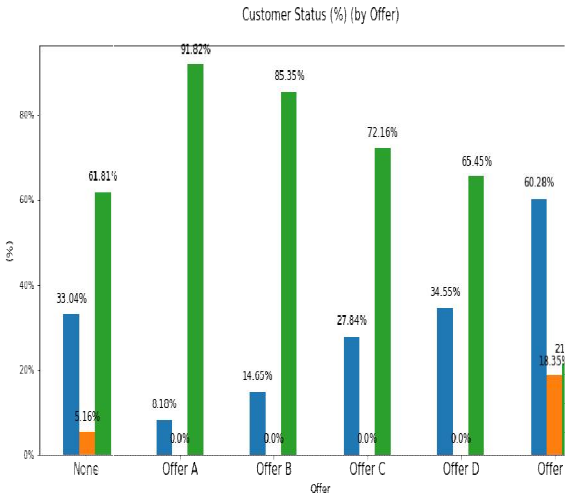
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

VII. RESULT

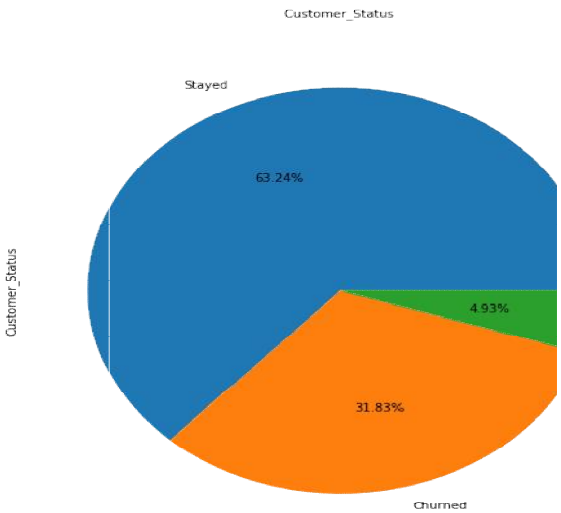
Customer status by Gender



Customer status by Offer



Customer status



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

Django:

A high-level Python web framework called Django enables the quick creation of safe and dependable websites. Django, which was created by seasoned programmers, handles a lot of the hassle associated with web development, allowing you to concentrate on developing your app without having to invent the wheel. It is free and open source, has a vibrant community, excellent documentation, and a wide range of free and paid support options.

VIII. CONCLUSION

Data preparation and processing, analysis of missing values, exploratory analysis, and model construction and evaluation came first in the analytical process. The highest accuracy score on the public test set will be discovered. This project can assist in determining the probability of telecom customers quitting.

IX. FUTURE WORK

- Telecom Customer Churn prediction to connect with Cloud.
- To optimize the work to implement in web development.

ACKNOWLEDGMENT

The Telecom Churn Prediction Project funded this research. Many thanks to all of the team members who had the opportunity to work on this and related projects; each of them made a significant contribution. We are also appreciative of the colleagues at the customer service department and the information system centre who provided knowledge that was extremely helpful to the research.

REFERENCES

[1]. Gerpott TJ, Rams W, Schindler A. Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. Telecommun Policy. 2001;25:249–69.

[2]. Wei CP, Chiu IT. Turning telecommunications call details to churn prediction: a data mining approach. Expert Syst Appl. 2002;23(2):103–12.

[3]. Qureshii SA, Rehman AS, Qamar AM, Kamal A, Rehman A. Telecommunication subscribers' churn prediction model using machine learning. In: Eighth international conference on digital information management. 2013. p. 131–6.

[4]. Ascarza E, Iyengar R, Schleicher M. The perils of proactive churn prevention using plan recommendations: evidence from a feld experiment. J Market Res. 2016;53(1):46–60.

[5]. Bott. Predicting customer churn in telecom industry using multilayer preceptron neural networks: modeling and analysis. Igarss. 2014;11(1):1–5.

[6]. Umayaparvathi V, Iyakutti K. A survey on customer churn prediction in telecom industry: datasets, methods and metric. Int Res J Eng Technol. 2016;3(4):1065–70a

[7]. Yu W, Jutla DN, Sivakumar SC. A churn-strategy alignment model for managers in mobile telecom. In: Communication networks and services research conference, vol. 3. 2005. p. 48– 53.

[8]. Burez D, den Poel V. Handling class imbalance in customer churn prediction. Expert Syst Appl. 2009;36(3):4626–36.