IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Churn Analysis Using ML

[1]Prof.Tushar Surwadkar, [2]Dr.Varsha Shah, [3]Prof.Nargis Shaikh, , [4]Prof.Chandramohan Konduri, [5]Prof.Sachin Charbhe,[6]Basit Jagrala,[7]Pratik Thakur,

[2]Principal, [3]HOD, [1,4,5]Assistant Professor,[6,7] Students

[1]Rizvi College of Engineering, Mumbai, India

Abstract: Customers are the foundation for any business benefit and that is why firms become conscious of the significance of acquiring satisfaction of customers. Customer churn is one of the major problems and it is regarded as one of the most essential concerns among companies because of increasing among firms, increased significance of marketing policies and customers awareness in present years. Organizations must develop different policies to solve the churn issues depending on the services they offer. Customer churn practice is essential in competitive and rapidly developing in telecom sector. The process of changing from one service provider to another telecom service provider occurs due to good services or rates or due to benefits to the customers which the competitor firm provides customers when signing up. Due to the greater cost related with acquiring new customers the prediction of customer churn has developed as an indispensable part of planning process and strategic decision making in telecom sector. The main aim of the study is to explore the customer churn prediction in telecom using in big machine learning data platform. Machine learning techniques have been used for estimating the customer probability to churn. This study makes use of logistic regression and KNN with big data for predicting consumer churn in the telecom sector. Logistic regression has been used widely to estimate the probability of churn as a function of variables set or features of customers. Similarly, for churn K-Nearest Neighbour is used to examine if a customer churns or not based on their feature's proximity to customers in every class. This study uses Kaggle website for dataset in predicting and analysing churn

Index Terms **- Churn, Telecom, Machine learning, Random Forest, logistic regression, decision tree, KNN.**

I. INTRODUCTION

Learning why customers buy more goods or services or don't is essential to a company's benefit and progress. The initial step in gaining this data is to find out how many customers are quitting your company. Then you may proceed further to identify patterns among existing customers, recognize areas for improvement, and avert losing more customers. Customer Churn Analysis is the technique of determining how quickly people turn back a product, website, or service. It permits teams to take action by answering the questions they are uncertain about. The telecom sector overall has exhibited stable long-term growth, as telecommunications has become an increasingly important basic industry, impervious to business cycles. Churn analysis is an essential practice in such industries. Managing existing customers, user experience, getting new customers and business strategy becomes easier if the company analyses the customer behavior and makes suitable changes in the system to improve the product or service quality and scale the business. In this project, we'll do telecom customer churn analysis. We have used Telecom sample dataset for this project

Literature Survey:

Survey Existing system: Churn Analysis is a common analysis technique used in industries. The rate at which customers exit from doing business with a company is known as the customer churn rate. Limitation Existing system or research gap: Churn is unavoidable. It happens in every business and every Industry. You can only work towards reducing the churn rate by identifying the reason behind it. Machine learning algorithms can be used to accomplish better outcomes. Problem Statement and Objective: To analyse the data and use various Machine Learning Algorithms to predict the churn and analyse the results.

It is recommended that telecom service providers must increase engagement of customers. In this competitive world customers are bombarded constantly by information and choices from all around. With the appropriate strategy of marketing in place and by concentrating on customer retention and satisfaction service providers must increase engagement of customers and nurture big term relations. Telecom service providers must implement tailored programs specifically to support their customers perceive the advantages of their services and products. It is recommended that telecom service providers must delight and surprise their customers. A satisfied customer is the best strategy among all solutions to reduce the churn rate. Putting a smile on the face of customer is as easy as providing the best recognition award to customer. Telecom service provider must do something outstanding to show how much they value them. Thus, the survival of any business is based on its capability to retain customers and put huge amount of efforts in reducing the churn rate of customers.

Analysis/Framework/Algorithm:
We have implemented Exploratory Data Analysis or EDA on the dataset of tele communication company. Steps include: 1. Importing the Libraries and data. 2. Exploratory Data Analysis. 3. Data Cleaning. 4. Splitting the data. 5. Building the Machine Learning Models. 6. Comparing the Accuracies.

Methodology:
DATA: The dataset used in this project is Telecom dataset which is a telecommunication company's sample dataset with many features.EDA: Exploratory Data Analysis is the most essential step before working on any dataset. Understanding the data makes it easier to model it and implement suitable machine learning algorithms so that we can achieve results with high accuracies. Here, we have used multiple techniques to understand the given dataset.

Descriptive Statistics:
• The data file used is csv file referred from Kaggle.com/datasets.
data = pd.read_csv('/content/WA_Fn-UseC_-Telco-Customer-Churn.csv')
• The data is studied and understood by using various commands and cleaned before doing further analysis.

• Churn Frequency: -The pie chart plotted suggests that Total of 26.5% of customers switched to another company.

• Gender and Churn Distributions: -We tried to find out any relationship between total churning customers to the gender ratio. But we found that, there is minor difference in customer percentage who switched the service provider. Both genders showed similar patterns when it comes to switching to another service provider.

• Customer Contract Distribution w.r.t Churn
• We found relation between total churned customers and the contract done between the company and customers
A customer with month-to-month contract has a probability of 42.71 % churn
A customer with one year contract has a probability of 11.27 % churn
A customer with two-year contract has a probability of 2.83 % churn
About 43% of customer with Month-to-Month Contract preferred to exit as compared to 11% of customers with One Year Contract and 3% with Two Year Contract. A major percent of people who left the company had Month-to-Month Contract.

• Payment Method Distribution w.r.t Churn: -Major customers who moved out had Electronic Check as Payment Method.Customers who chose Credit-Card automatic transfer or Bank Automatic Transfer and Mailed Check as Payment Method were less likely to leave.

• Churn Distribution w.r.t Internet Service: -A lot of customers choose the Fiber optic service and it's also noticeable that the customers who use Fiber optic have high churn rate, this might suggest a dissatisfaction with this type of internet service.
• Customers having DSL service are more in number and have less churn rate compared to Fibre optic service.

• Dependents Distribution:-Customers without dependents are more likely to churn

• Partner Distribution:-Customers that don't have partners are more likely to churn.

• Senior Citizen Distributions:-It can be observed that there are very few senior citizens who churns.

• Phone Service Distribution:-Very small fraction of customers doesn't have a phone service.

• Outlier Detection:-The presence of outliers in a classification or regression dataset can result in a poor fit and lower predictive modelling performance, therefore we should see there are outliers in the data and we found out that there is no outlier using box plot.

Confusion matrix:
A confusion matrix is a matrix which sum up the performance of a ML model on a set of test data. It is mostly used to measure the performance of classification models, whose objective is to predict a categorical label for each input instance. The matrix shows the number of true positives (TP), true negatives(TN), false positives(FP), false negatives(FN) produced by the models on test data.
Binary classification- matrix will be a 2x2 table
Multi class classification- the matrix shape will be equal to the number of classes.(for eg, for n classes it will be nxn).
A 2x2 confusion matrix is shown below for the image recognization having a cat image or a not cat image.

Table no.1

		Actual	
		cat	**Not cat**
Predicted	**cat**	True Positive (TP)	False Positive (FP)
	Not cat	False Negative (FN)	True Negative (TN)

True Positive (TP): it is the total counts having both predicted and actual values are cat.
True Negative (TN): it is the total counts having both predicted and actual values are not cat.
False Positive (FP): it is the total counts having prediction is cat while actually not cat.
False Negative (FN): it is the total counts having prediction is not cat while actually, it is cat.

In machine learning we use some common evaluation metrics to assess the performance of classification models. These metrics below provide insights to the models ability to correctly identify positive and negative instances in a given datast.

Precision:
Precision measures the proportion of correctly predicted positive instances (true positives) out of all instances predicted as positive.
It is calculated as-
Precision = TP / (TP + FP)
TP refers to true positives (instances correctly predicted as positive), and FP refers to false positives (instances incorrectly predicted as positive).
Precision indicates the accuracy of the model when it predicts positive instances. A higher precision value indicates fewer false positives, meaning the model makes fewer incorrect positive predictions.

Recall (Sensitivity or True Positive Rate):
Recall measures the proportion of correctly predicted positive instances (true positives) out of all actual positive instances in the dataset.
It is calculated as-
Recall = TP / (TP + FN)
TP refers to true positives (instances correctly predicted as positive), and FN refers to false negatives (instances incorrectly predicted as negative).
Recall indicates the model's ability to identify all positive instances correctly. A higher recall value indicates fewer false negatives, meaning the model captures a higher proportion of positive instances in the dataset.

F1-score:
The F1-score is the harmonic mean of precision and recall.
It is calculated as:
F1-score = 2 * (Precision * Recall) / (Precision + Recall)
The F1-score provides a balance between precision and recall. It is useful when we want to consider both false positives and false negatives in the model's performance.
The F1-score ranges from 0 to 1, where a value of 1 represents perfect precision and recall.

Support:
Support refers to the number of instances in each class in the dataset.
It indicates the number of actual instances that belong to a particular class.
Support values are often used to calculate weighted averages of precision, recall, and F1-score when dealing with multi-class classification problems.

All Models:
1) Random Forest Classifier
2) Decision Tree Classifier
3) Logistic Regression
4) K Nearest Neighbor (KNN)

Models used:

• Random Forest classifier:

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. 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.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:

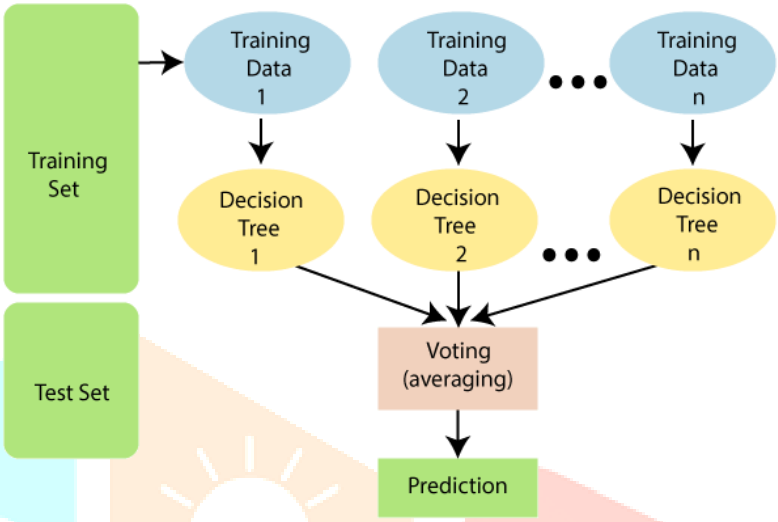

Fig no.1: working of Random Forest

Why use Random Forest?

Below are some points that explain why we should use the Random Forest algorithm

It consumes less training time as compared to other algorithms.

It predicts output with high accuracy, even for the large dataset it runs efficiently.

It can also maintain accuracy when a large proportion of data is missing.

Advantages of Random Forest

Random Forest is capable of performing both Classification and Regression tasks.

It is capable of handling large datasets with high dimensionality.

It enhances the accuracy of the model and prevents the overfitting issue.

Comparison of the Models:

1.Random Forest Classifier

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions.

It handles complex datasets and high-dimensional feature spaces.

It tend to have good accuracy and are less prone to overfitting.

The train accuracy and test accuracy should be relatively high, with the test accuracy slightly lower than the train accuracy

The test AUC score should also be high, indicating good discrimination between classes.

2.Decision Tree Classifier

Decision Trees are simple, it makes predictions based on a series of hierarchical decisions, it can handle both categorical and numerical data. Decision Trees tend to overfit the training data, resulting in high train accuracy but lower test accuracy. The test AUC score may vary depending on the dataset, but it might be lower compared to more advanced models like Random Forest or Logistic Regression.

3.Logistic Regression

Logistic Regression is a linear model. It is ly used for binary classification problems. It models the relationship between the input variables and the probability of the outcome using a logistic function. Logistic Regression performs well when the relationship between the features and the target variable is relatively linear. The train accuracy and test accuracy should be comparable, but the test accuracy might be slightly lower. The test AUC score should be moderate to high, depending on the dataset and the presence of class imbalance.

4.K Nearest Neighbor (KNN)

KNN is a non-parametric classification algorithm which classifies new instances based on their similarity to existing instances. It assigns a class label to a new data point based on the majority vote of its k nearest neighbors in the training set. KNN is computationally expensive during the prediction phase. It is simple to understand and apply. The train accuracy and test accuracy can vary depending on the choice of k and the structure of the data. The test AUC score may vary, and it could be influenced by the choice of k and the distribution of the classes.

In summary, Random Forest and Logistic Regression are widely used compared to Decision Trees and KNN. Random Forest can handle complex datasets and provide good accuracy, while Logistic Regression is effective for linear relationships between features and the target variable. Decision Trees tend to overfit, while KNN's performance depends on the choice of k and data structure.

Proposed Model: Random Forest Classifier

IV. RESULTS AND DISCUSSION

Results of Descriptive Statics of Study Variables

AUC on train data: 0.9999751451195744
AUC on test data: 0.8410015553789624
F1 score is: 0.5570400822199383
Precision is: 0.6791979949874687
Recall is: 0.4721254355400697

Figures and Tables:

Table no. 2: Accuracy of 4 models used

Model Name	Train Accuracy	Test Accuracy	Test Auc Score
Random Forest	0.998377	0.796025	0.841002
Decision Tree	0.998377	0.733554	0.654051
Logistic Regression	0.803043	0.807856	0.858507
K Nearest Neighbour	0.829412	0.769995	0.751636

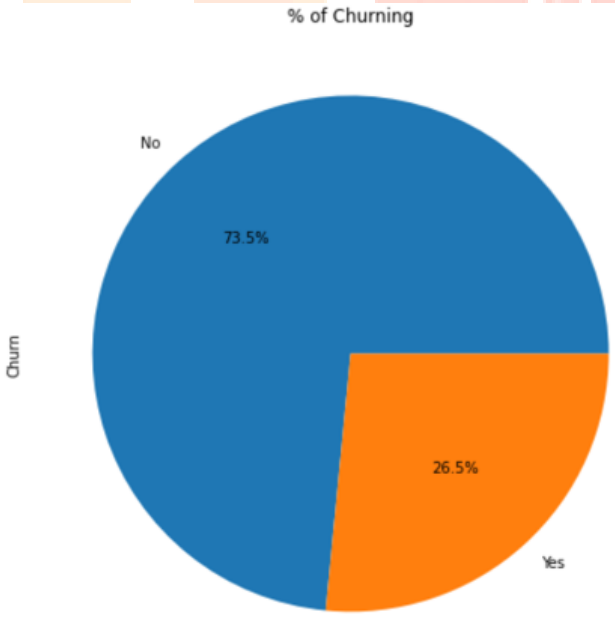

Fig.no.2: 26.5% of customers switched to another company

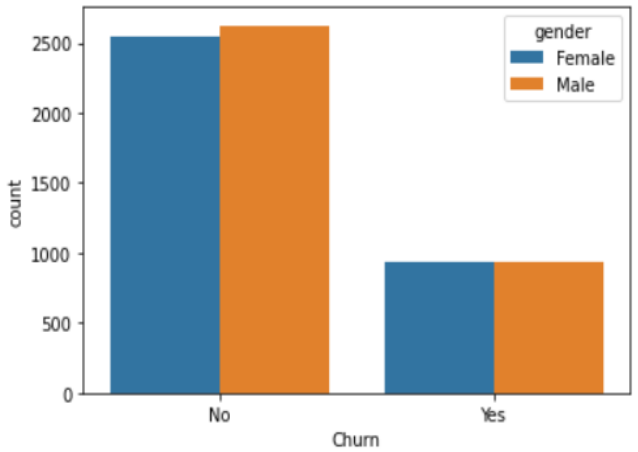

Fig.no.3: Gender and Churn Distributions

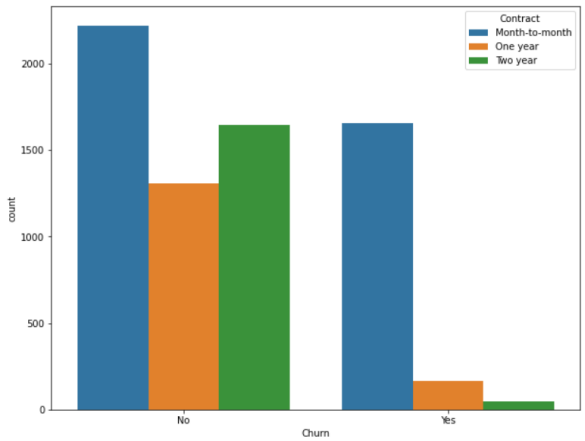

Fig no.4: Customer Contract Distribution w.r.t Churn

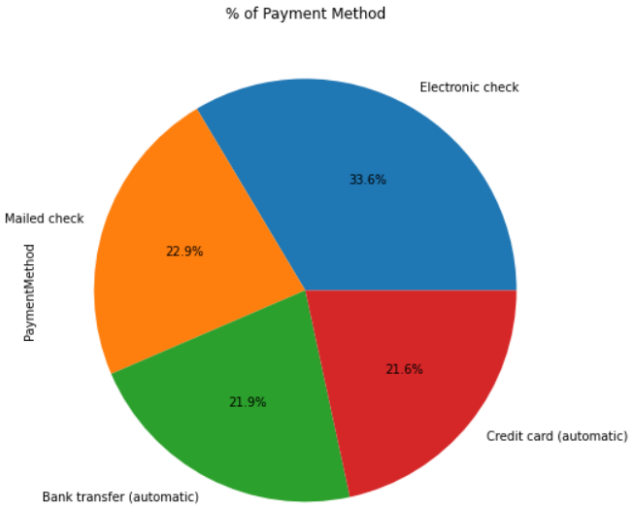

Fig.no.5: Payment Method Distribution w.r.t Churn

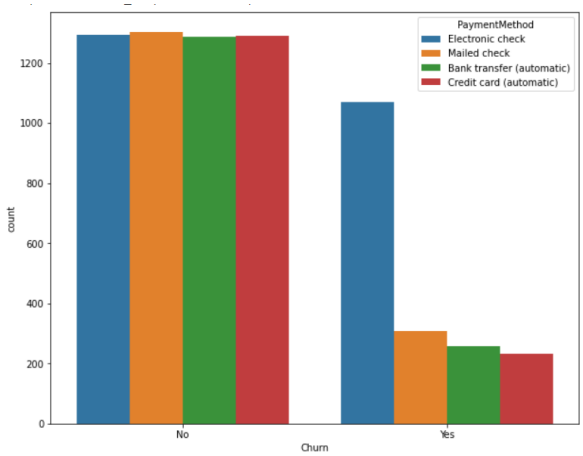

Fig.no.6: Churn Distribution w.r.t Payment Method

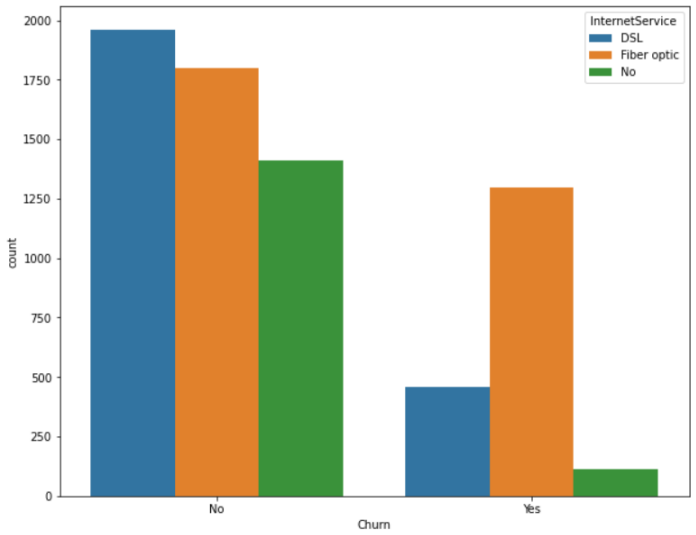

Fig.no.6: Churn Distribution w.r.t Internet Service

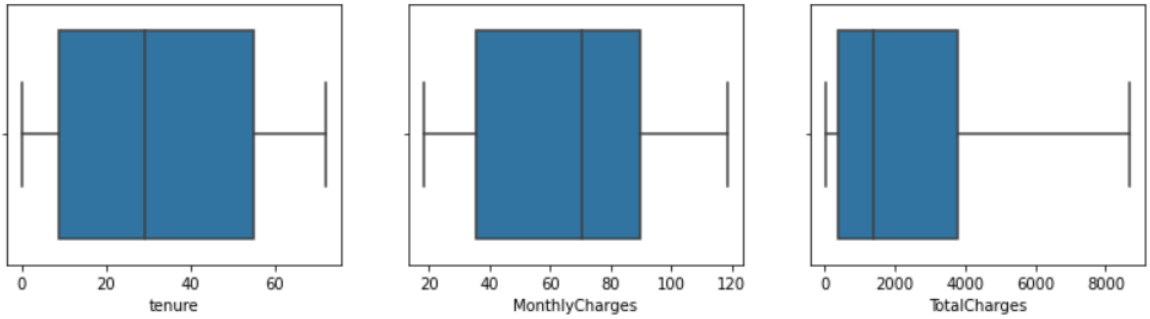

Fig no.7: - box plots

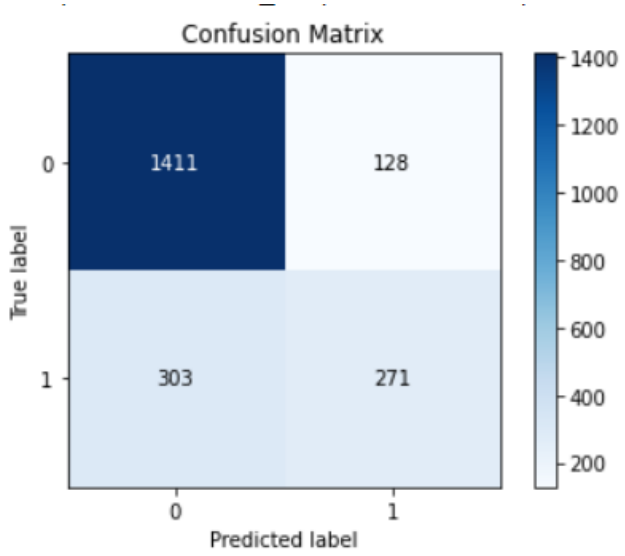

Fig.no.8: confusion matrix

CONCLUSION AND FUTURE SCOPE:

We conclude that this project 'Churn Analysis using Machine learning' was successful in achieving the results after analyzing the data, cleaning it and applying the Machine learning algorithms like Random Forest, Logistic Regression, KNN, Decision Tree. Logistic Regression algorithm showed the highest accuracy with 0.807856 test accuracy and 0.858507 Test Auc Score. With more observations and utilizing more sophisticated models, the accuracies can be improved and advanced models can be prepared for better results in any industry.

II. ACKNOWLEDGMENT

Thanks to our Principal Dr, Varsha Shah, Prof. Nargis Shaikh, HOD of Artificial Intelligence & Data Science Department & Prof. Tushar Surwadkar for their continuous support and guidance.

REFERENCES

[1]. https://scikit-learn.org/stable/index.html
[2] https://www.javatpoint.com/
[3] https://youtube.com/
[4] https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
[5] https://www.kaggle.com/