e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

PREDICTION OF CUSTOMER CHURN USING MACHINE LEARNING

Yash Singh*1, Yash Pandit*2, Neil Joshi*3, Prof. Vedika Avhad*4

*1,2,3BE Student, Department Of Information Technology, Mumbai, Maharashtra, India, Vasantdada Patil Pratishthan's College Of Engineering & Visual Arts, India.

*4Assistant Professor, Department Of Information Technology, Mumbai, Maharashtra, India, Vasantdada Patil Pratishthan's College Of Engineering & Visual Arts, India.

ABSTRACT

Churn prediction is a common use case in machine learning domain. If you are not familiar with the term, churn means "leaving the company". It is very important for telco businesses to know about why and when their customers are about to churn. Having a accurate and robust churn prediction model helps businesses to take precautions and necessary actions to prevent customers from churn. Churn prediction consists of predicting which customers are likely to stop a subscription to a service based on how they use the service. This paper will help various companies to understand the factors behind customer churn and the actual customer churn rate using machine learning which the company can use to reduce the customer churn and also make strategies to retain back the churned customer.

Keywords: Machine Learning, Customer Churn, Churn Rate, Prediction, Growth.

I. INTRODUCTION

Customer churn is the percentage of customers that stopped using a particular company's product or service during a certain time frame. One can calculate churn rate by dividing the number of customers that company lost during that time period by the number of customers that company had earlier before that time period. Predicting customer churn is a challenging task but extremely important too because business problem especially in industries where the cost of customer acquisition is high and difficult to manage such as in technology, telecom, finance, etc. sectors. The ability to predict that a particular customer is about to churn, while there is still time to do something about it, might increase huge additional revenue source for companies. Churn prediction consists of detecting which customers are about to cancel a subscription to a service or a product. Predicting customer churn is also a binary classification problem. Customers either churn or retain in a given period. The reasons for customer churn are divided into two types: "accidental and intentional". Accidental churn happens when the conditions are changing so as to keep the clients from utilizing the services later on, for example financial conditions that make benefits unreasonably costly for the client.

Intentional churn happens when the clients switch to another organization that gives same alike services, with better ideas from rivalry, further developed services and better cost for a similar service. In recent years, churn prediction has become an important method for telecommunication industry. To tackle customer churn rate.

In order to deal with this problem, the telecom operators must find out these customers before they churn. Therefore, developing a unique and accurate classifier that will predict future churns is vital. This classifier must be able to recognize users who might churn in the near future, so the operator can take any effective measures to stop these customers from stop using their services maybe by using discount and promotions or another strategy.

When it comes to useful business applications of machine learning, it doesn't get much better than customer churn prediction. It's a problem where you usually have a lot of high-quality, fresh data to work with, it's relatively straightforward, and solving it can be a great way to increase profits.

Churn rate is a critical metric of customer satisfaction. Low churn rates mean happy customers; high churn rates mean customers are leaving you. A small rate of monthly/quarterly churn compounds over time. 1% monthly churn quickly translates to almost 12% yearly churn.

In this project, we're using "Telecom Customer Churn" dataset which is available on Kaggle.

There are 22 features or independent variables and 1 dependent variable for 7043 clients. Dependent or labeled variable indicates if a customer has left the company (churn=yes) within the last month. Since the dependent variable has two states (yes/no or 1/0), this is a binary classification problem.

The variables are: 'customer ID', 'gender', 'Senior Citizen', 'Partner', 'Dependents', 'tenure', 'Phone Service', 'Multiple Lines', 'Internet Service', 'Online Security', 'Online Backup', 'Device Protection', 'Tech Support', 'Streaming TV', 'Streaming Movies', 'Contract', 'Paperless Billing', 'Payment Method', 'Monthly Charges', 'Total Charges', 'Churn'.

Customers are very important assets in any industry since they are considered as the main profit source. Nowadays, companies have become very active that they put much effort not only to convince or attract the customers, but also to retain their existing customers. Churned customers are persons who move to other company for various factors. To decrease the customer churn rate, the company should be able to predict the behavior of customers correctly and establish connections between customer attrition and keep elements in check. Churn prediction is a binary classification task, which differentiates churners from non-churners.

Machine learning is a data analytical model which automates model analytical building. Using algorithms that repeatedly learn from data, machine learning allows systems to explore hidden patterns without being explicitly think where to look for the problems. There are three types of machine learning: unsupervised machine learning, semi-supervised machine learning, and supervised machine learning. Supervised learning is the machine learning task of finding the patterns from dataset which already has the output within it. Unsupervised learning is the machine learning task of finding the patterns from dataset which have no output. Semi-supervised learning is a class of supervised learning tasks and techniques which also make use of unlabeled data for training – typically a small set of labelled data with a large set of unlabeled data. Semi supervised learning falls between unsupervised learning and supervised learning.

Customer Churn prediction involves these three types of analysis:

A. Prediction

Prediction can be used to specify whether customers might churn or not based on multiple factors and possibilities. There are many algorithms to do predictions but logistic regression gives the best results when the problem has two possible outputs or binary classification is needed.

B. Classification

Classification When the data are being used to predict a category, supervised learning is also called classification.

Classification predictive modeling involves assigning a class label to input examples.

- Binary classification refers to predicting one of two classes.

C. Regression

Regression analysis is a fundamental concept in the field of machine learning. It is a category of supervised learning wherein the algorithm is trained with both input and output from dataset. Its helps in identifying the relationship between the factors or variables which may help to find customer churn accurately.

II. AIM AND OBJECTIVES

A] Aim

Our aim is to build a project that will help the companies to predict whether their customers will churn or not. These projects will help the companies for customer churn rate prediction, accurately and quickly. These predictions will be based on some important factors like gender, online security, streaming services, tech support, etc.

B] Objectives

- The primary objective of the customer churn predictive model is to retain back the customers who are at the highest risk of churn by engaging with them via mails or other medium of communication.

For example: Offer a gift voucher or promotional discount to lock in the customer for more years to come.

- Prediction of customer churn will take less time.
- Various important factors affecting customer churn can be determined so that retention rate of customers can be reduced effectively.

C] Motivation for the work Issues faced by companies after customer(s) churn:

1. Time taken and investment/promotions needed is huge to bring in new customers in place of those loyal customers who stopped using services and products. 2. Credibility, market share and revenue decreases. 3. To predict the actual reason(s) behind why a customer or a no. of customers stop using their services or products.

D] Issues faced in searching for actual factor(s) behind the customer(s) churn:

1. Customer's data is stored in huge numbers, so predicting the important factors behind customer attrition is a difficult and time-consuming process.

2. The results after prediction might be not accurate enough to make a decision to reduce customer retention rate

3. The existing system to do the same task is difficult to understand or explain to others.

E] Scope

Since our project uses a telecom dataset hence this project can be used by telecom companies to predict customer attrition or churn rate and the factors behind customer churn. But it is not limited only to telecom industry only, this project can be used by other service and product-based companies too like streaming, e-commerce, automobile industries, etc. to predict customer attrition or churn and the factors affecting customer churn.

III. LITERATURE SURVEY

A] Introduction:

In Today's world the competition between different companies providing the same kind of products or services has increased over the years since the digital revolution of 4G technology which has given a significant boost to the no. of users using services and products over the internet. Along with this tracking customer retention rate, customer requirements and fulfil them accordingly is of great importance.

B] Existing System:

As the industries have grown, the need for prediction of customer attrition or churn rate has also grown because it's very difficult to bring in new customers in place of the existing, loyal customers when they stop using your company services or products. By Knowing the factors behind the customer churn may help the companies to reduce the customer retention rate and retain back the customers using various strategies and offers. In existing system customer churn is predicted using either feedback forms or commonly used ml algorithms such as Naïve Bayes, Extra tree, Decision tree, etc. which is significantly slow and less accurate as compare to xg-boost & logistic regression.

Kriti [1] in her paper Customer churn: A study of factors affecting customer churn using machine learning has been too successful to find out various factors affecting customer churn (price sensitivity, technology, customer service, tenure, security). Also comparing various algorithms to analyze customer churn and prescribe solutions to avoid this churn. She has given the future work as the predictions from the ML model can help in understanding the customers who might leave their services. Also suggested various solutions based on those predictions.

Essam Abou El Kassem and Shereen Ali Hussein [2] in their paper has explained that Customer churn is a problem for most companies because it affects the revenues of the company when a customer switches from a service provider company to another. They've used social media sentiment analysis to predict the factors behind customer churn.

Praveen Lalwani and Manas Kumar Mishra [3] in their paper has Compared the time taken to train the model and accuracy of various ML algorithms. Their paper concludes that ensemble learning techniques such as XG-Boost classifier gives maximum accuracy when compared to other models.

Saran Kumar A. [4] In his paper had conducted a survey on various ML algorithms and techniques to predict customer attrition or churn rate. Proposed to use various boosting classification techniques for better accuracy.

Pradeep B and Sushmitha Vishwanath Rao [5] in their paper have explained how to use various ML algorithms to analyze customer attrition or churn rate in the logistics industry.

This paper concludes that the purpose analysis of customer churn rate is to identify valuable customers that potentially contribute to the profitability of the company.

C] Problem Definition

➢ To build an effective customer attrition or churn rate prediction system using ML to predict whether the customer(s) may churn or not, also to find out important factors affecting customer churn.

D] Logistic Regression algorithm Advantages:

1. Logistic regression is easy to implement, very efficient to train, and interpret.
2. It can easily extend to multiple attributes (multinomial regression).
3. It is very fast and accurate for classifying unknown records.
4. Very Good accuracy for many small and large data sets and it performs well when the dataset can be separated linearly.

E] Logistic Regression algorithm Disadvantages:

1. If the number of observations is less than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.
2. The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.
3. Non-linear problems can’t be solved with logistic regression because it has a linear decision surface. Linearly separable data is hardly or rarely found in real-world scenarios.
4. Logistic Regression requires average or no multicollinearity between independent variables.

F] Need of New System:

Limitations of Existing Systems:

1. Time-consuming and not accurate.
2. Cannot predict customer attrition or churn rate based on huge dataset.
3. Feedback Forms are not sufficient.
4. Competitive Market.

IV. DESIGN AND IMPLEMENTATION

A] Proposed System

Step1->Loading Data

Step2->Data Cleaning, Feature engineering, Data Visualization

Step3->Encoding Categorical Data.

Step4->Then use the train-test split procedure to estimate the performance of machine learning algorithms when they are used to make predictions on data.

Step5->Displaying the results using Bar plot.

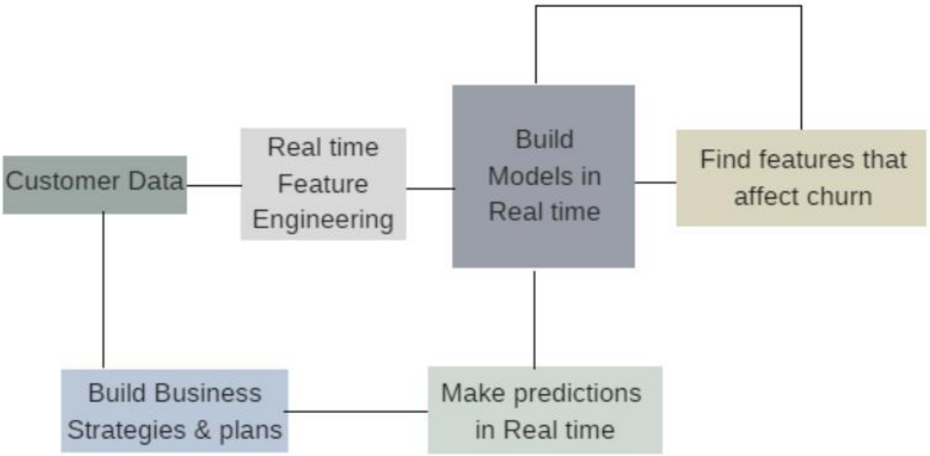

Customer Data
Real time Feature Engineering
Build Models in Real time
Find features that affect churn
Build Business Strategies & plans
Make predictions in Real time

Fig 1: Steps to predict customer attrition or churn rate using ML

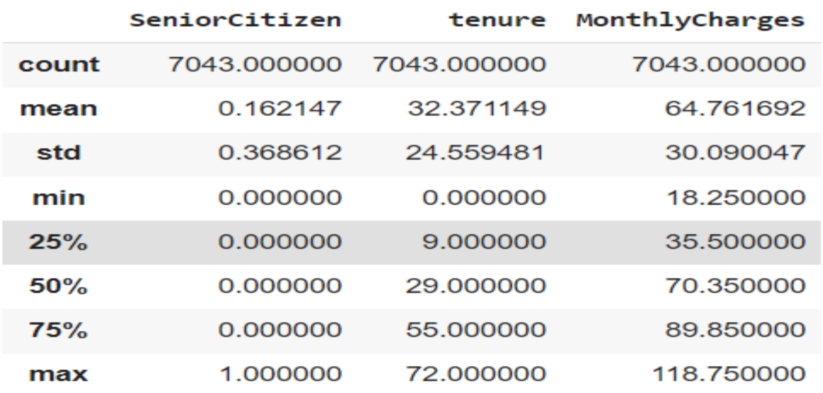
Firstly, Imported the libraries like cat boost, NumPy, pandas, sklearn to use all the ml algorithms to perform various operations on the dataset and to predict customer attrition or churn rate and the factors affecting the customer churn.

To visualize and compare the information in the dataset we've used matplotlib and seaborn libraries. Also imported the modules like Label Encoder, One Hot Encoder which will be used to rectify the data discrepancies and accuracy score to check the accuracy of various algorithms to predict customer attrition or churn rate.

Also, we've loaded the telco dataset which we got from Kaggle using google drive. Using the head function and pandas library we've checked the top 5 rows and attributes or the services on the basis of which we'll be able to calculate customer churn and find out the important factors behind customer attrition or churn rate in telecom company.

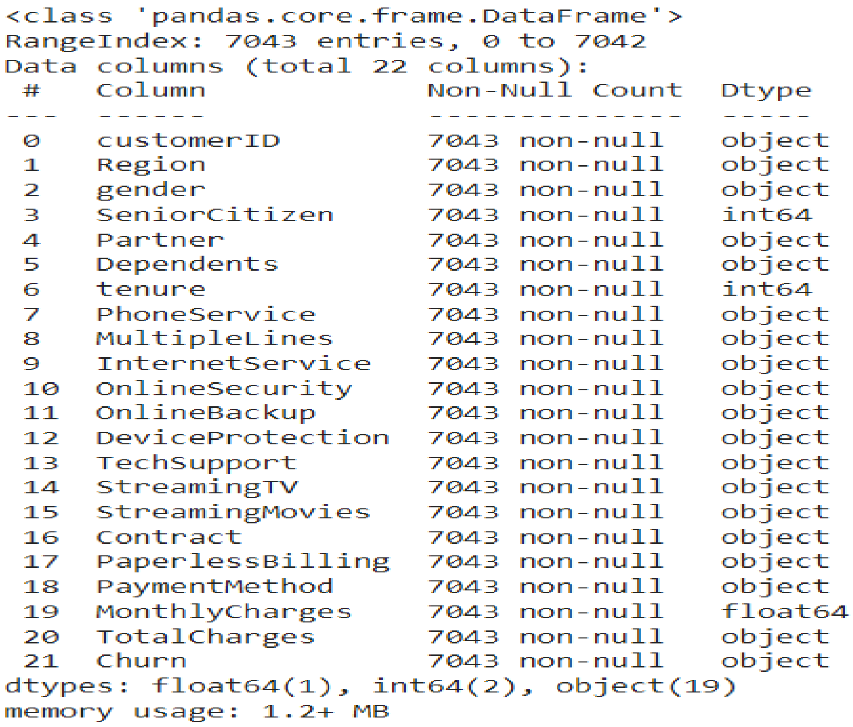
Dependent variable has imbalanced class distribution. Positive class (Churn=Yes) is much less than negative class (churn=No). Imbalanced class distributions influence the performance of a machine learning model negatively.

Then we use describe function to figure out the count of the rows, mean value, standard deviation, minimum and maximum value,25%,50%,75% values for numerical attributes (like Senior citizen, Tenure, Monthly Charges)



	SeniorCitizen	tenure	MonthlyCharges
count	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692
std	0.368612	24.559481	30.090047
min	0.000000	0.000000	18.250000
25%	0.000000	9.000000	35.500000
50%	0.000000	29.000000	70.350000
75%	0.000000	55.000000	89.850000
max	1.000000	72.000000	118.750000

Fig 2: Descriptive statistics summary of a given data frame using pandas describe function.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 22 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   Region            7043 non-null   object
 2   gender            7043 non-null   object
 3   SeniorCitizen     7043 non-null   int64
 4   Partner           7043 non-null   object
 5   Dependents        7043 non-null   object
 6   tenure            7043 non-null   int64
 7   PhoneService      7043 non-null   object
 8   MultipleLines     7043 non-null   object
 9   InternetService   7043 non-null   object
 10  OnlineSecurity    7043 non-null   object
 11  OnlineBackup      7043 non-null   object
 12  DeviceProtection  7043 non-null   object
 13  TechSupport       7043 non-null   object
 14  StreamingTV       7043 non-null   object
 15  StreamingMovies   7043 non-null   object
 16  Contract          7043 non-null   object
 17  PaperlessBilling  7043 non-null   object
 18  PaymentMethod     7043 non-null   object
 19  MonthlyCharges    7043 non-null   float64
 20  TotalCharges      7043 non-null   object
 21  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(19)
memory usage: 1.2+ MB
```

Fig 3: Pandas data frame.info () function is used to get a informative summary of the data frame.

Using the data frame(df) and info function we check for the null value in our dataset and also the datatype of the attributes /services of telco company provided in the dataset.

E] Data Cleaning.

Data Cleaning is a very important step in machine learning or data mining since it helps to clean the data or rectify the mistakes which is there in the dataset. After this we have visualize or display a bar chart to compare the customer attrition or churn rate precentage as “Yes” and not churned as “No”. Since our project is based on supervised learning it already has the output to train and test the model.

V. ALGORITHMS USED

A. Logistic Regression

-> Logistic regression is a classification model rather than a regression model. Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

->It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry. The logistic regression model is a method for binary classification that can be generalized to multiclass or multi-attributes classification.

B. XGBOOST

->With a regular machine learning model, like a decision tree, we’d simply train a single model on our dataset and use that for prediction.

->We might play around with the parameters for a bit or augment the data, but in the end, we are still using a single model. Even if we build an ensemble, all of the trained models and applied to our data separately.

->Boosting, takes a more iterative approach. It’s still technically an ensemble technique in that many models are combined together to perform the final one, but takes a cleverer approach.

C. Light GBM

Light GBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with the following advantages:

->Faster training speed and higher efficiency.

->Lower memory usage.

->Better accuracy.

D. Neural Network

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. Neural networks can adapt to changing input; so, the network generates the best possible result without needing to redesign the output criteria. The concept of neural networks, which has its roots in artificial intelligence, is swiftly gaining popularity in the development of trading systems.

E. Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.Most important features of the Random Forest Algorithm are that it can handle the data set containing continuous variables as in the regression case and categorical variables as in the case of classification. It performs better results for classification problems.

F. K-nearest neighbors (KNN)

The k-nearest neighbors (KNN) algorithm is a very simple, easy-to-implement machine learning supervised type algorithm that can be used to solve both regression problems and classification problems.

As the name (K Nearest Neighbor) suggests it considers K Nearest Neighbors (Data points) to predict the class or continuous value for the new Datapoint.

G. Extra Tree

Extra Trees, or Extremely Randomized Trees, is a type of ensemble machine learning algorithm.

Specifically, it is a kind of decision trees (ensemble) and it is related to other ensembles of decision trees algorithms such as bootstrap aggregation (bagging) and random forest.

The Extra Trees algorithm works by creating a large number of unpruned decision trees from the training dataset. Predictions are made by averaging the prediction of the decision trees in the case of regression or using majority voting in the case of classification.

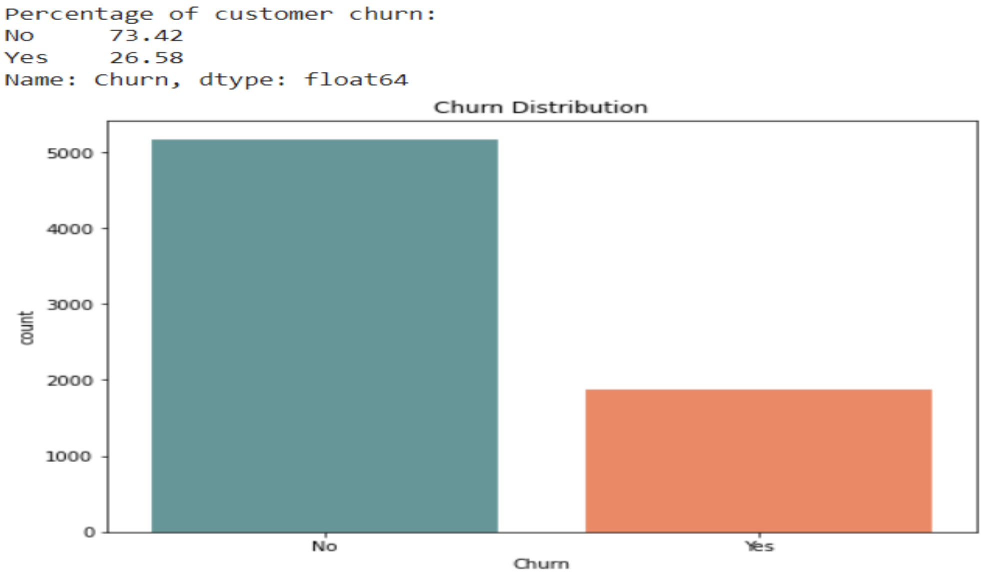
H. Gaussian Naive Bayes

A Gaussian Naive Bayes algorithm is a special type of NB algorithm. It's specifically used when the features have continuous values. It's also assumed that all the features are following a gaussian distribution i.e., normal distribution.

An approach to create a simple model is to assume that the data is described by a Gaussian distribution with no co-variance (independent dimensions) between dimensions.

VI. CHURN DISTRIBUTION

In Churn Distribution we check the customer attrition or churn rate based on the attributes given in the dataset (Gender, Dependents, Partner, Phone Services, Streaming TV, etc.



```
Percentage of customer churn:
No      73.42
Yes     26.58
Name: Churn, dtype: float64
```

Fig 4: Churn Distribution Bar Plot

VII. DISTRIBUTION OF CUSTOMER CHURN BASED ON SEVERAL VARIABLES

Gender and partner are evenly distributed with approximate percentage values. The difference in churn is slightly higher in females, but the small difference can be ignored. There's a higher proportion of churn in younger customers (Senior Citizen = No), customers with no partners, and customers with no dependents. The demographic section of data highlights on-senior citizens with no partners and dependents as a particular segment of customers likely to churn. Internet service variable is definitely important in predicting churn rate. As you can see, customers with fiber optic internet service are much likely to churn than other customers although there is not a big difference in the number of customers with DSL and fiber optic. This company may have some problems with fiber optic connection.

However, it is not a good way to make assumptions based on only one variable. Let's also check the monthly charges. Fiber optic service is much more expensive than DSL which may be one of the reasons why customers churn.

Key Observations:

- There are 7043 entries with 27 columns in the original data.

- The average tenure is about 32 months with a minimum entry of 0 months and a maximum entry of 72 months.
- The average monthly charge is about 65 with a minimum of about 18 and a maximum of about 119.
- Most customers are on a month-to-month contract.
- Most customers pay using Electronic Check (manual)
- Most customers have fiber optic cable internet service.
- Total charges were loaded in as a string and will need to be transformed in the prepare stage.
- There are multiple 'id' columns that can be dropped in the prepare stage.

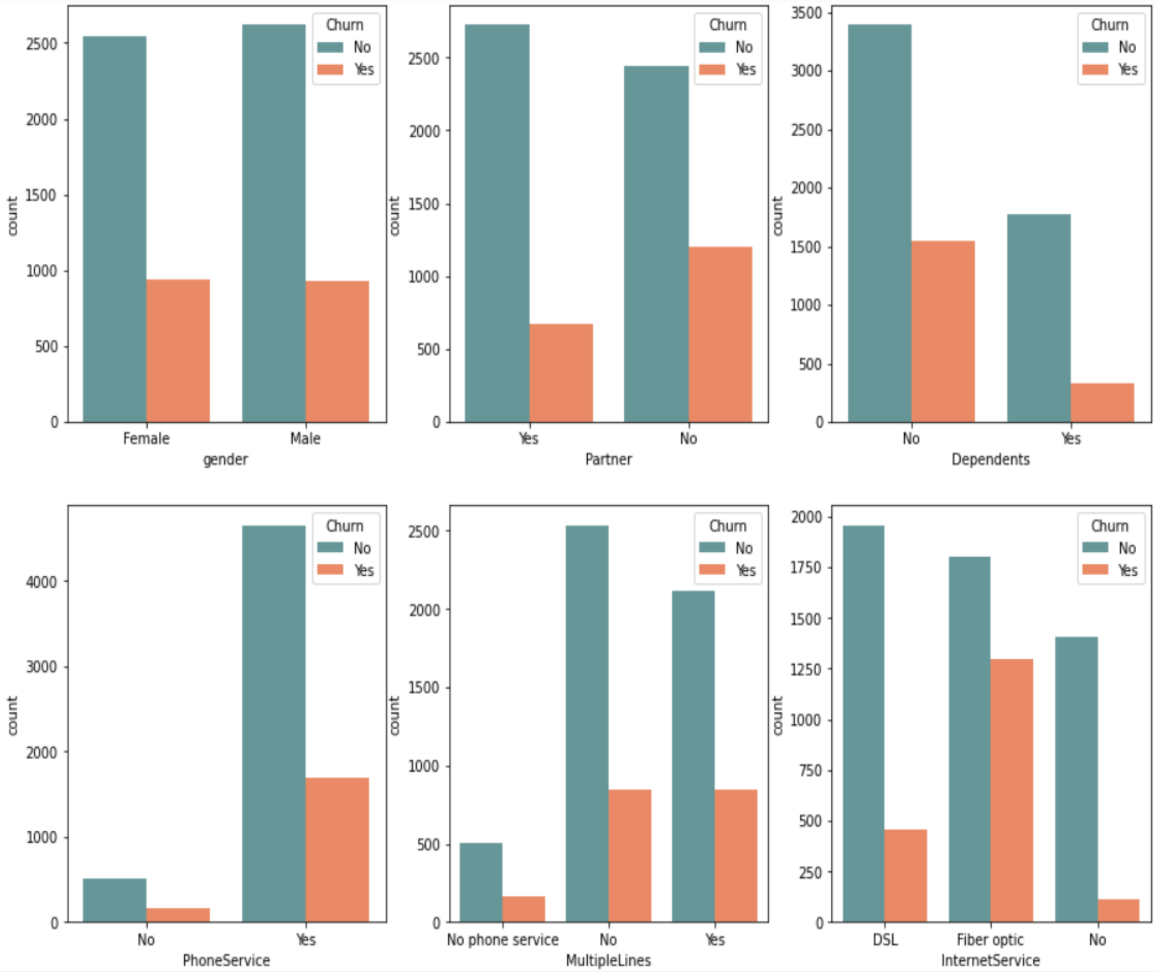


Fig 5: Customer Churn Distribution Bar Plot based on several variables

VIII. CUSTOMER DISTRIBUTION BASED ON NUMERICAL ATTRIBUTES (TENURE, TOTAL CHARGES AND MONTHLY CHARGES)

In this we check the customer churn based on some numerical attributes or continuous features like tenure, total charges and monthly charges. And we also check the skewness in the data which falls under these attributes.

Skewness is the measure of the asymmetry of an ideally symmetric probability distribution and is given by the third standardized moment.

In simple words, skewness is the measure of how much the probability distribution of a random variable deviates from the normal distribution.

Well, the normal distribution is the probability distribution without any skewness. In Fig. below which shows symmetrical distribution that's basically a normal distribution and you can see that it is symmetrical on both sides of the dashed line. Apart from this, there are two types of skewness:

- Positive Skewness
- Negative Skewness

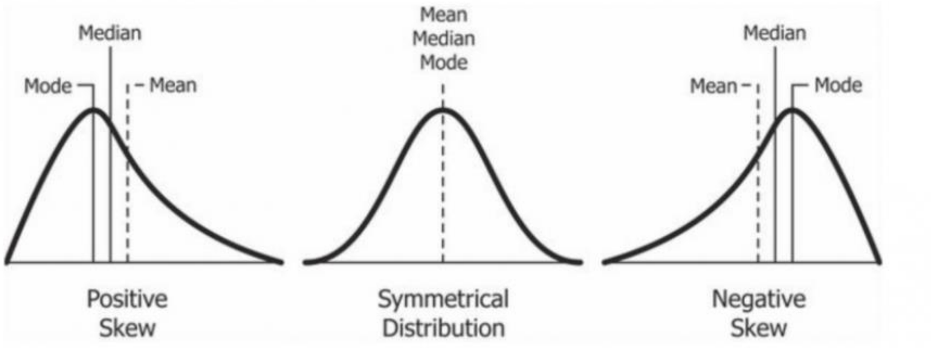


Fig 6: Types of Skewness

The probability distribution with its tail on the right side is a positively skewed distribution and the one with its tail on the left side is a negatively skewed distribution. Now, we know that the skewness is the measure of asymmetry and its types are distinguished by the side on which the tail of probability distribution lies. But why is knowing the skewness of the data important? First, linear models work on the assumption that the distribution of the independent variable and the target variable are similar. Therefore, knowing about the skewness of data helps us in creating better linear models. The continuous features are tenure, monthly charges and total charges. The amount in total charges column is proportional to tenure (months) multiplied by monthly charges. So, it is unnecessary to include total charges in the model. Adding unnecessary features will increase the model complexity. It is better to have a simpler model when possible. Complex models tend to overfit and not generalize well to new, previously unseen observations. Since the goal of a machine learning model is to predict or explain new observations, overfitting is a crucial issue.

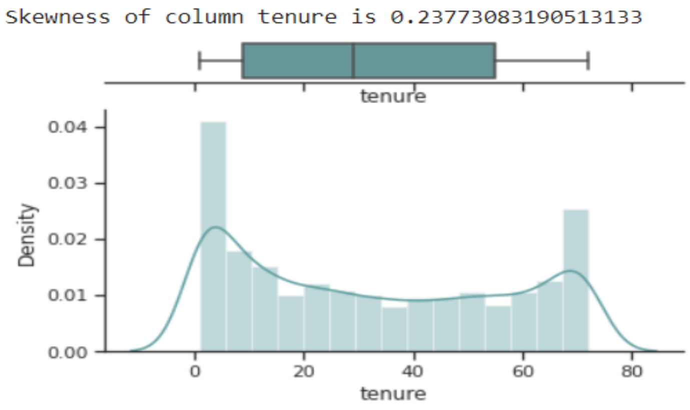

```
Skewness of column tenure is 0.23773083190513133
```

Fig 7: Skewness of Tenure Cloumn

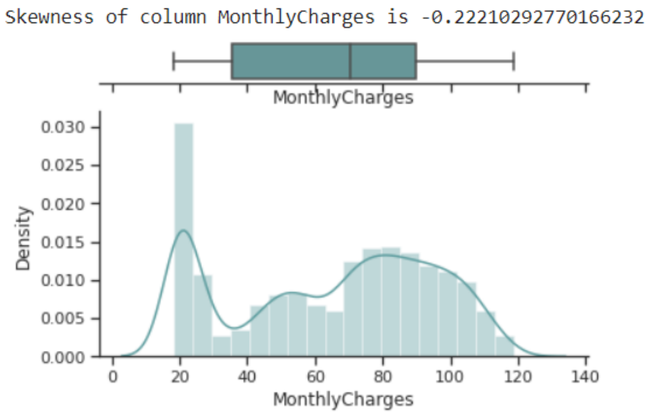
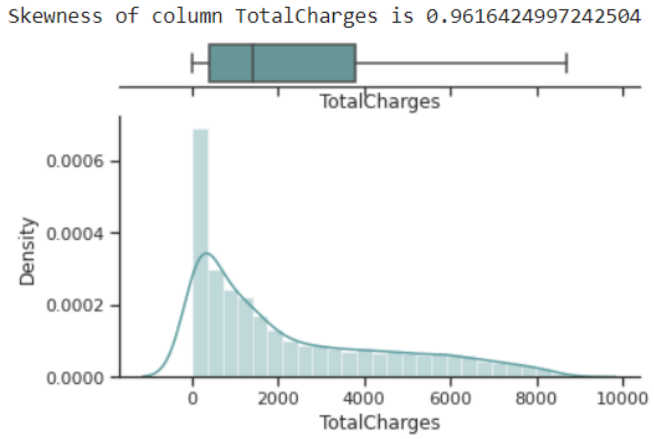

```
Skewness of column MonthlyCharges is -0.22210292770166232
```

Fig 8: Skewness of Monthly Charges Column



```
Skewness of column TotalCharges is 0.9616424997242504
```

Fig 9: Skewness of Total Charges

According to the distribution of tenure variable, most of the customers are either pretty new or have stayed for a long time with the company.

Our goal should be finding a way to keep those customers with a tenure of up to a few months.

A similar trend is seen on Monthly Charges.

There seems to be a gap between low rates and high rates. As we see, contract and tenure are highly correlated. Customers with long contracts have been a customer for longer time than customers with short-term contracts.

As contract will add little to no value to tenure feature so we will not use contract feature in the model.

After exploring the variables, we have decided not to use following variable because they add little or no informative power to the model:

- Customer ID
- Gender
- Phone Service
- Contract
- Total Charges

IX. ENCODE CATEGORICAL DATA

Categorical features need to be converted to numbers so that they can be included in calculations done by a machine learning model.

The categorical variables in our data set are not ordinal (i.e., there is no order in them).

For example, “DSL” internet service is not superior to “Fiber optic” internet service.

An example for an ordinal categorical variable would be ratings from 1 to 5 or a variable with categories “bad”, “average” and “good”.

When we encode the categorical variables, a number will be assigned to each category.

The category with higher numbers will be considered more important or effect the model more.

Therefore, we need to do encode the variables in a way that each category will be represented by a column and the value in that column will be 0 or 1.

We also need to scale continuous variables. Otherwise, variables with higher values will be given more importance which effects the accuracy of the model.

As we briefly discussed in the beginning, target variables with imbalanced class distribution are not desired for machine learning models.

We use up sampling which means increasing the number of samples of the class with less samples by randomly selecting rows from it.

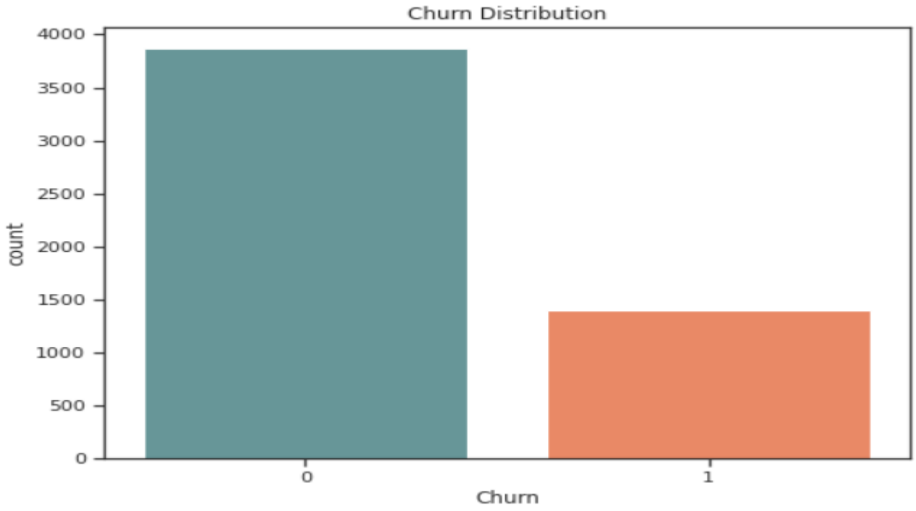


Fig 10: Churn Distribution vs Churn bar plot after encoding categorical data

X. CORRELATION

Correlation measures the linear relationship between two variables. Features with high correlation are more linearly dependent and have almost the same effect on the dependent variable. So, when two features have a high correlation, we can drop one of them. In our case, we can drop highly correlated features like Multiple Lines, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, and Streaming Movies. Churn prediction is a binary classification problem, as customers either churn or are retained in a given period. Two questions need answering to guide model building:

1. Which features make customers churn or retain?
2. What are the most important features to train a model with high performance?

Monthly Charges and Phone Service columns will not be used to reduce multicollinearity in the data.

XI. SPLIT DATA

The information in the dataset will be converted to a scale of 0-1 before splitting the data.

Using the Normalization formula.

x = (x-np.min(x))/(np.max(x)-np.min(x)). values

Normalization in machine learning is the process of converting the data into the range [0, 1] (or any other range) or simply transforming data.

We need to divide the dataset into training and test subsets so that we are able to measure the performance of our model on new, previously unseen examples.

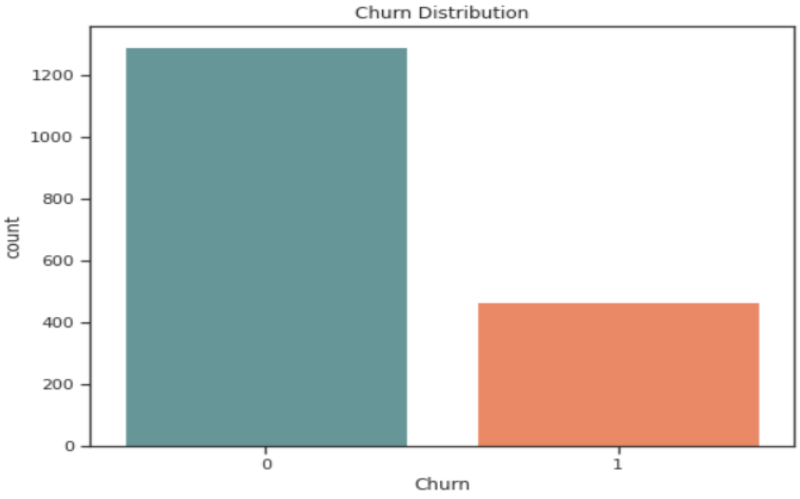
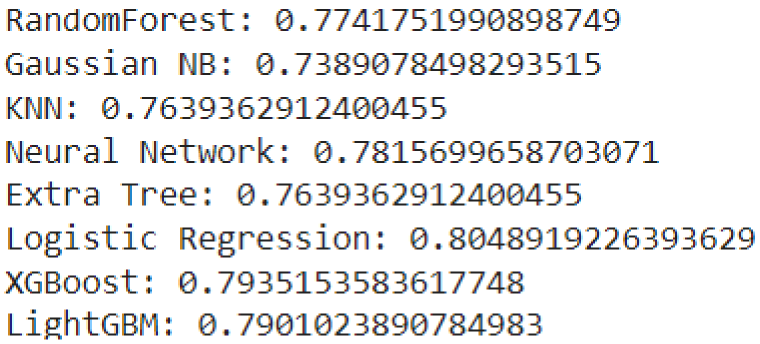


Fig 11: Churn Distribution vs Churn bar plot after Normalization

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

XII. RESULT AND EVALUATION



```
RandomForest: 0.7741751990898749
Gaussian NB: 0.7389078498293515
KNN: 0.7639362912400455
Neural Network: 0.7815699658703071
Extra Tree: 0.7639362912400455
Logistic Regression: 0.8048919226393629
XGBoost: 0.7935153583617748
LightGBM: 0.7901023890784983
```

Fig 12: Accuracy Comparison between Machine Learning algorithms

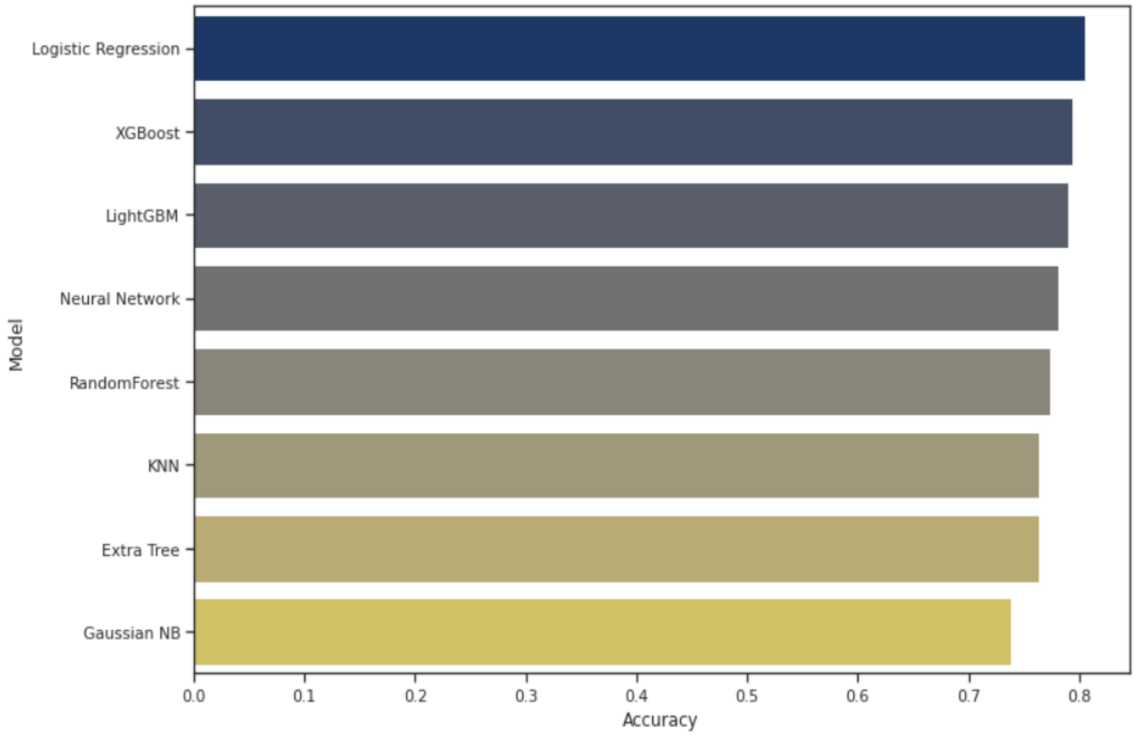


Fig 13: Horizontal Bar Plot Chart Shows the accuracy of models based on ML Algorithms to predict customer attrition or churn rate

Looking at the evaluation results from data understanding and customer distribution and churn distribution, the hypotheses can be directionally supported or refused:

- **Contract duration:** Contract duration month-to-month is the second biggest driver of churn → supported
- **Number of additional services:** This feature does not rank among the top features → refused
- **Tenure:** High tenure ranks as the strongest factor for not churning and the strongest feature overall. This is also supported by the boxplot in the EDA step. → supported
- **Monthly payment:** Total payments, which is the product of tenure and monthly payment ranks as the biggest factor for churn rate prediction. Indirectly, high monthly payments lead to churn. However, tenure is the highest driver of not churning → refused
- **Senior citizens:** Senior citizens does not have high feature weights. Also, the ratio of senior citizens who churn is much higher than that of non-churners → refused.

As we can see Top 4 base models based on different Machine Learning algorithms provides greater accuracy even if the dataset is huge, it can easily predict the customer percentage to be churned. Also, we can understand the importance of features or attributes using data understanding and data visualization.

XIII. CONCLUSION

This Customer Churn Project is certainly used for analysis of major reasons behind the customer stop using services of a company.

- Also, they can get to know the probability of various customers who might churn or leave using their services.
- After the analysis that company can plan on how to decrease the customer churn and also retain the existing customer back.

In this project, we have predicted customer churn using the Telco customer Churn dataset. We started by cleaning the data and analyzing it with visualization.

Then, to be able to build a machine learning model, we transformed the categorical data into numeric variables (feature engineering). After transforming the data, we tried different machine learning algorithms using default parameters.

Telcos typically have much more data available that could be included in the analysis, like extended customer and transaction data from CRM systems and operational data around network services provided. Also, they typically have much larger amounts of churn/non-churn events at their disposal more than the around 7000 rows data used in this project. A very high accuracy is needed to be able to identify promising customer cases where churn can be avoided as, eventually, the customer returns protected need to outweigh the costs of related retention campaigns.

XIV. FUTURE WORK

Based on the factors affecting customer churn and prediction of customer attrition or churn rate based on ml algorithms one can produce various strategies to retain back the churned customers. Also, a mobile or desktop application can be made which can further simplify and help organizations, startups to predict customer attrition or churn rate and focus on important factors affecting customer churn.

XV. REFERENCES

[1] Kriti, "Customer churn: A study of factors affecting customer churn using machine learning" Iowa State University Capstones, Theses and Dissertations, March 2019

[2] Essam Abou el Kassem, Shereen Ali Hussein, Alaa Mostafa Abdelrahman, Fahad Kamal Alsheref. "Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content" IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, November 5, 2020

[3] I Praveen Lalwani, Manas Kumar Mishra, Jasroop Singh Chadha, Pratyush Sethi. "Customer churn prediction system: a machine learning approach" in Springer

[4] Saran Kumar A., Chandrakala D. "A Survey on Customer Churn Prediction using Machine Learning Techniques" International Journal of Computer Applications (0975 – 8887) Volume 154 – No.10, November 2016.

[5] Pradeep B, Sushmitha Vishwanath Rao, Swati M Puranik, Akshay Hegde "Analysis of Customer Churn prediction in Logistic Industry using Machine Learning" International Journal of Scientific and Research Publications, Volume 7, Issue 11, November 2017 ISSN 2250-3153.