**ARTICLE ON CUSTOMER CHURN ANALYSIS**

**BY: ATISH P. KALANGUTKAR**

## **PROBLEM DEFINITION**

**Customer Churn also called as Customer Attrition is the percentage of customers that stopped using your company products or service during a certain time frame or in other words percentage of customers your business lost in a set period of time.**

**The more customer you retain , the lower your churn rate.**

**The higher your churn rate, the more difficult is to grow because you will be constantly chasing the new customers to replace the ones that are churning.**

**There are two types of churn, they are**

* ****Voluntary Churn**: It happens when the customer wants to stop using your service or product and cancels their account.**
* ****Involuntary Churn:** It happens when customers payment fails and they never update their billing information.**

**So customer churn is arguably one of the most important metrics for the health of your business. In simple words if you cant retain customers then it’s nearly impossible to grow a sustainable business.**

**So to reduce the customer churn, firstly I think that you should focus your attention on your best customers. Secondly, analyze the churn as it occurs meaning to check and identify the issue about why your customers are leaving and after that analyzing how and when churn occurs in a customers lifetime with your company and using that data to improve the things where you are lacking or could not give a better service to the customer. And thirdly to show your beloved customers that you care.**

**DATA ANALYSIS**

To start with Data Analysis when the data was received firstly there was a importing of required libraries such as pandas, numpy, seaborn, matplotlib and warnings.

After that dataset was loaded which was having 7043 rows and 21 columns in which 1 column was having float type of data, 2 columns were having integer type of data and rest all the columns that is 18 columns were having object type of data.Meaning it was a data of 7043 customers.

**Among 21 columns, There are three numerical features in the dataset:**  
**Tenure:** Number of months the customer has been with the company  
**Monthly Charges:** The monthly amount charged to the customer  
**Total Charges:** The total amount charged to the customer.

**16 categorical features in the dataset:**  
**Gender:** M/F  
**Senior Citizen:** Whether the customer is a senior citizen or not (1, 0)  
**Partner:** Whether customer has a partner or not (Yes, No)  
**Dependents:** Whether customer has dependents or not (Yes, No)  
**Phone Service:** Whether the customer has a phone service or not (Yes, No)  
**Multiple Lines:** Whether the customer has multiple lines or not (Yes, No, No Phone Service)  
**Internet Service:** Customer’s internet service type (DSL, Fiber Optic, None)  
**Online Security:** Whether the customer has Online Security add-on (Yes, No, No Internet Service)  
**Online Backup:** Whether the customer has Online Backup add-on (Yes, No, No Internet Service)  
**Device Protection:** Whether the customer has Device Protection add-on (Yes, No, No Internet Service)  
**Tech Support:** Whether the customer has Tech Support add-on (Yes, No, No Internet Service)  
**Streaming TV:** Whether the customer has streaming TV or not (Yes, No, No Internet Service)  
**Streaming Movies:** Whether the customer has streaming movies or not (Yes, No,

No Internet Service)  
**Contract:** Term of the customer’s contract (Monthly, 1-Year, 2-Year  
**Paperless Billing:** Whether the customer has paperless billing or not (Yes, No)  
**Payment Method:** The customer’s payment method (E-Check, Mailed Check, Bank Transfer (Auto), Credit Card (Auto)).

Further when I was going through the columns and was looking up for the unique values which were present in the columns ,in column namely ‘total charges’ I found that all the values were in integer(float type) but when checked for the data type it was showing that it possesses object data type which was not possible. So it was time to find out why it was showing data type as object. So used ‘loc’ function to check whether there are any empty spaces in that column or not. Surprisingly there were 10 empty spaces in that column which was triggering the data type. So replaced the empty spaces with the null values and than changed the data type of that particular column and then filled the null values by using imputing method.

After that I went forward to describe the data and to examine what type of unique values present in the columns, whether they are normally distributed or not, whether there is any skewness is present or not.

After examining dataset, it came into notice that column ‘customer id’ was having the unique id’s which was not relevant at the time of prediction so that column was dropped.

It was time to check if there are any duplicates present or not in the dataset, so it was noticed that 22 duplicates were present which was later on dropped from the dataset.

After that looked into the number of unique values present in the columns and than started plotting distribution plots, count plots and pie plots for the dataset.

After that I inspected and made a different dataset as per to the gender meaning how many males and females didn't churned and how many females and males churned and did analyses by plotting count plots.

Than did analysis as per to the internet service, as which customers churned more, the one who are using DSL or the fibre optic or the ones who are not using internet service.

Further I went on to check the highest top 10 monthly charges and total charges and what type services this customers are using including the churn rate.

After that I went on to check for the correlation of features with the label and used bar plot to check whether they are positively correlated or negatively correlated and plotted heatmap to know whether there is any multicollinearity problem is there or not.

**EDA CONCLUDING REMARKS**

Lets summarize the finding from the analysis which was performed.

There was a not single null value present in the dataset but their were 22 duplicates present which was dropped from the datasets. In column 10 empty spaces was detected which was later on replaced by null values and was filled by using imputing methods.

So after going through the count plots it was noticed that dataset was having similar numbers of Male and Female. Also, proportion of customers churned is almost equal in both genders.

Our dataset has less senior citizens as compared to non-senior citizens. Overall, it was seen that there was a higher proportion of senior citizens churned than non-senior citizens. Customers without partners churn slightly more than those with partners. And customers without dependents churn slightly more than those with dependents.

It was also seen that most of the customers fall in the tenure of less than 5 months or greater than 70 months with nearly uniform distribution in between these ranges in which churn rate was more when the customers having tenure of less than 5 months and churn rate was less when the customer was having tenure of greater than 70 months.

There are three categorical features which tells us about the type of services subscribed by the customer, they are Phone Service, Multiple lines and Internet Service. By going through count plots it came into notice that customer having phone service churned more than that of customer without phone service. Customers with multiple lines have a slightly higher churn rate than those with a singular line. Than customers having only phone service and no internet service have lower churn rate than those having internet service and it also it was seen that most of the customers were using fibre optic internet service. Also fibre optic internet customers churn at significantly higher proportion than that of customers using DSL or no internet service.

Customers have subscribed for various add-on services along with phone and internet services such as online security, online backup, device protection, tech support, streaming tv and streaming movies. So after going through the plots it was seen most of the customers who have not opt for any of the add on services their churn rate is more than customers who have opt for the add on services. And the customers who have the add on services such as TV streaming and/ or movie streaming services churn more than all other add-on-services.

Than coming to the contract part, more than half of the customers use a monthly payment option. Significantly more customers churn on monthly plans. The longer the plan, less is the churn rate.

When checked about paperless billing it was seen that most of the customers have paperless billing. Also , customers with paperless billing have significantly higher churn rate than the customers with the non-paperless billing .

About payment methods, it was seen that customers with Automatic payment methods have lower churn rate than other payment methods and customers with electronic check have higher churn rate than the other payment methods.

And when the analysis was done about the highest monthly and total charges it was seen that customers using fibre optic internet was having the highest monthly and total charges and most of the customers were leaving meaning the churn rate was high as compared to the customers using DSL.

After looking at the bar plot of the correlation it was seen that **contract and tenure are highly negatively correlated with the label including column paper less billing and monthly charges but they were highly positively correlated to the label and column gender and phone service are very least correlated with the label.**

**After looking at the heatmap it was seen that column tenure was multi correlated to with the column total charges and after plotting scatter plot for the same it was noticed that they were showing positive trend or relationship and total charges columns was less correlated to the label and was also cross verified by using vif method. So the column total charges was dropped from the dataset.**

**PER-PROCESSING PIPELINE**

In this section, we tackle data per-processing steps to prepare the datasets for Machine Learning algorithm implementation.

**Encoding:** There were about 16 categorical columns in the dataset, so label encoder was applied to encode the data as machine learning algorithms only understand the numbers as they don’t understand strings.

**Outliers: After looking at the box plots it was seen that outliers were not present in the dataset.**

****Skewness:** It was seen that skewness was present in the in one of the columns so power transform method was used to treat the skewness.**

**Splitting data into training and testing sets:** Splitting data into features and labels is one of the important thing as further we need to do the feature scaling,have to used method vif and feature selection.

**Feature Scaling:** As there are lot of columns with different units so it is important to do scaling to standardize the data. So standard scaler was applied to standardize the data.

**VIF:** Variance inflation factor is used to detect multicollinearity problem between the features. So column total charges value was more than 5 and it was less related to the label so it was dropped from the dataset.

**Feature Selection:** As there are lot of columns so used feature selection method to to get the best columns to build the machine learning models.

**Smote Technique:** As the label was imbalanced, so smote technique was used to balance the data.

**BUILDING MACHINE LEARNING MODELS**

A machine learning model is built by learning and generalizing from training data, then applying that acquired knowledge to new data it has never seen before to make predictions and fulfill its purpose.

After performing all the exploratory data analysis which is useful in understanding the dataset, data cleaning, and performing all per-processing pipeline steps, the dataset was ready to build machine learning model. The steps to build machine learning models are as follows:

Firstly it is important to choose which algorithms you will be using. Than importing the libraries for that specific models.

Than the data was split into train-test-split. This splits the data into 4 parts that is x\_train, y\_train, x\_test, y\_test. So x\_train and y\_train data will be used for model training purpose and x\_test and y\_test data will be used for model testing purpose.

Further the best random state was found using algorithm Random Forest. The best random state is the one in which highest training and testing score will be there.

Afterwards the first model that was trained was Random Forest classifier and after using a function training, testing, confusion matrix and f1-score was produced.

After that to check whether our model is over fitted or under fitted we use cross validation method. K-fold cross validation randomly splits the training data into K subsets called folds. So again we use a function to get the best k-fold and we use the k-fold for all the models. And after the ROC curve was plotted.

So all this steps were repeated for 3 more algorithms namely Support Vector Machine, Ada Boost and Decision Tree.

After that data frame was made for all this models and their training, testing, cross validation score and f1 score was compared with each other and the best model was chosen according to the model which was having highest training score, less difference between training score and testing score, and cross validation score as well as f1-score is same as or closer to testing score.

So further after going through each and every aspect it was considered that the Random Forest was the best model. So going forward Hyper parameter tuning was done on the best model so as to tuned the model and to get the best training and testing score. So Grid Search CV was used to tuned the model(which is type of hyper parameter tuning). Here all the parameters of your model are specified and GridSearchCV runs combination and permutations on best model considering all those parameters. After completing this process best parameters are fetched that were responsible to provide best test and training scores based on parameters specified. This is always a time consuming process.

After that using this parameters model is again trained and tested. But unfortunately training accuracy and testing accuracy which was provided after doing hyper parameter tuning was less compared to the one on which hyper parameter was not done. Hence by using pickle the Random Forest classifier model was saved on which hyper parameter was not done.

**CONCLUDING REMARKS**

With this project, we got the idea about what type of data we can work with in building a model and what type data we should avoid. We also found how balancing the target values in a classification problem play a crucial part. In this project we found the feature engineering play a crucial role in the performance of the model.

By analyzing the dataset provided in this project, we found that the churn has a high positive correlation with monthly charges, as monthly charges increases chances of churn also increases. We also found that churn has a negative correlation with the contract and tenure, which means that if the customers contract and tenure is more and long, than the chance of churn rate goes down.

.