Customer Churn Prediction -Telecom industry

MD Manisha

***CCE** - Online Course on AI and Machine Learning with Python*
IISC
Bangalore, India
manisha_md@yahoo.com

***Abstract*—**Customer churn is one of the most important metrics to evaluate for the growing business. It is the number that gives the company the hard truth about its customer relationship and business. The objective of this project is to handle churn problem by analyzing telecom industry dataset. In this project we have found causes of the churn for a telecom industry by taking into consideration their past records & then recommending them new services to retain the customers to avoid churns in future. We used pie charts to check churning percentage later analyze whether there are any outliers [using box plot] then dropped some features which were of less importance then converted all categorical data into numerical by using [Label Encoding for multiple category data & map function for two category data] then spitted the data using train test split.

Keywords— customer churn, analysis, prediction, categorical data

I. INTRODUCTION

The telecom industry is growing day by day hence user as well as operators are investing into this industry, such a customer driven industry faces a huge financial issue if customer tend to leave their services. By using machine learning we can analyze, predict the way customer respond to these services. Churn is a key driver of EDITDA margin and an industry-wide challenge.

In this Customer Churn prediction & retention we are analyzing the past behavior of customers and accordingly finding the real cause of the churn, then predicting whether churn will happen in future by customers. By considering details like Monthly charges, services they have subscribed for tenures, contract they will contribute into end results i.e., prediction. Here aim is to use machine learning concepts to not only predict & retain customers but also to avoid further churns which would be beneficial to industry.

II. DATA PROCESS FLOW

A. Dataset Overview:

The Dataset in this project is "Telco Customer Churn Data" from Kaggle. It has 33 independent variables columns that indicate the characteristics of clients of a Telecommunication company including Churn Label variable (target variable) as YES or NO, other variables are Features, which are data points that describe the customers.

From the dataset – Telco Customer Churn, containing all features is used to analysis churn. Demographic Information: Gender (male, female), Senior Citizen (yes, no), Partner (yes, no), Dependents (yes, no).

Customer Account Information: Contract (Month-to-Month, One year, Two year), Payment Method (Electronic check, Mailed check, Bank transfer, Credit Card), Monthly Charges (numeric values), Total Charges (numeric values), Lat Long, Latitude(float values), Longitude(float values).

Services Information: Phone Service (yes, no), Multiple Lines (No phone service, No, Yes), Internet Services (DSL, Fibre optic, No), Online Security (No internet service, No, Yes), Device Protection (No internet service, No, Yes).

B. Data Pre-Processing

Data pre-processing is important task in machine learning. It converts raw data into clean data. Following are technique, applied on data:

Missing Values – Here we had missing values in Total charges feature which we then eliminated and adjusted them with mean values. These are the missing row values within data if not handled would later lead to errors for converting data type as it takes string value for empty spaces.

Label Encoder – For categorical variables this is perfect method to convert them into numeric values, best used when having multiple categories. We had various categorical values converted them into numeric for further use in algorithms.

Drop Columns – As we took insights from the data we came to know some of the features were of less importance so we dropped them to reduce number of features.

C. Features Selection:

Telco Customer Churn dataset having all features is used to analyse churn using Python code. Different parameters are used in finding more insights. we will study the distribution, which is very important in any machine-learning problem. It helps us to decide which algorithm we are going to use ahead.

Fig. 1. Distribution plot for in understanding dataset better for scaling

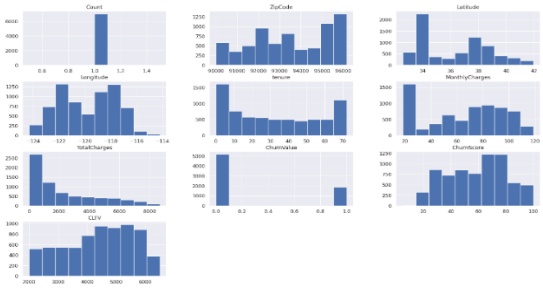

Fig. 1. Distribution plot of various columns

Data distribution of Target variable results as highly imbalanced data, ratio = 73.46%.

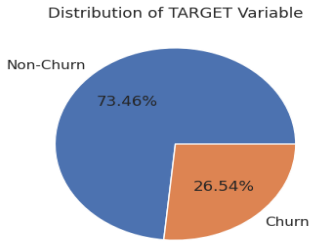

Fig. 2. Pie chart showing Distribution of Target Variables

To further understand the correlation between variables coefficient heat map is plotted for variables to conduct the thermodynamic chart analysis. The degree of correlation between variables can be judged according to the magnitude of correlation coefficients corresponding to the colours of different blocks in the correlation coefficient diagram. The highly correlated variables are close to 1 score. For example, the 'Total Charges' is 0.93, indicating that if the service charges are pocket-friendly, then customers will remain for a more extended period.

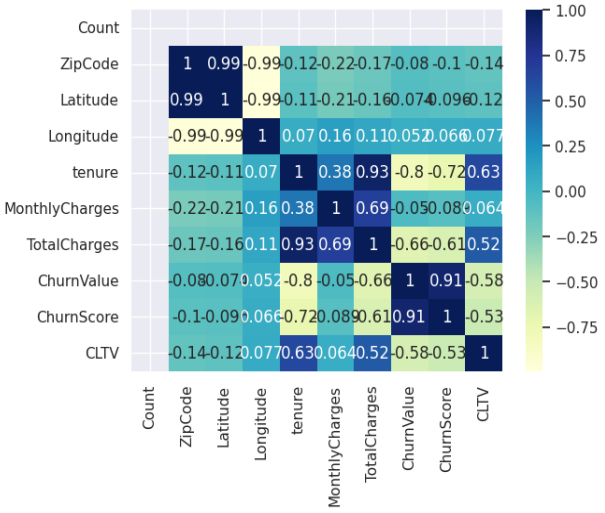

Fig. 3. Correlation Heat Map for variables

D. *Exploratory Data Analysis:*

In this phase we will look towards those features which we did not consider in feature selection but are contributing factor for prediction. EDA is nothing but a Data exploration technique to understand various aspects of the data. EDA is the first step towards data analyzing process and best way to manipulate the dataset in visual form.

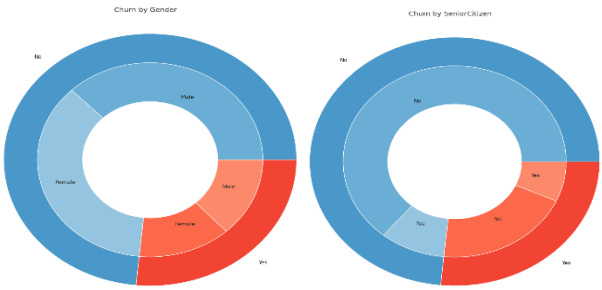

Fig. 4. Gender and Senior citizen pie plot

From Fig. 4&5, its evident that Senior citizens are less likely to churn

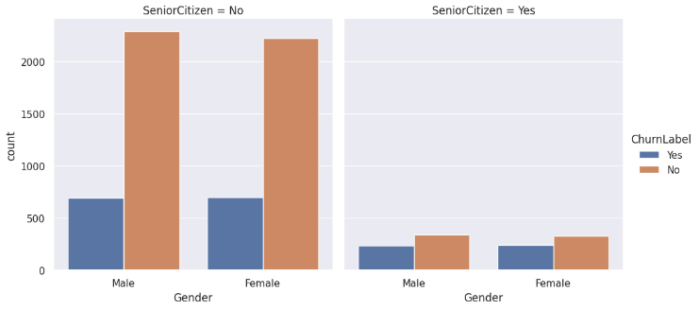

Fig. 5. Churn plot on Senior citizen and non-senior citizen

As tenure increases customer retention is more, churn is high for short tenure , from fig, 6.

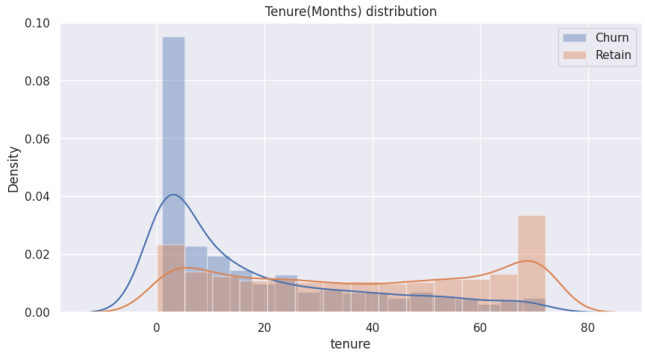

Fig.6. Tenure distribution for churn ratio

Contract type with Long tenure are less likely to Churn (fig. 7.)

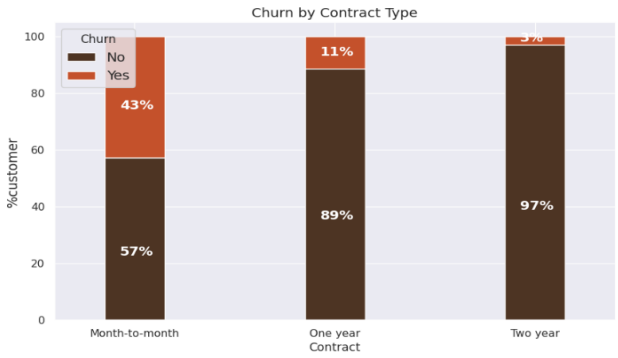

Fig. 7 . Churn by contract type

Scatter map plot to shows High ChurnScore customers location and City

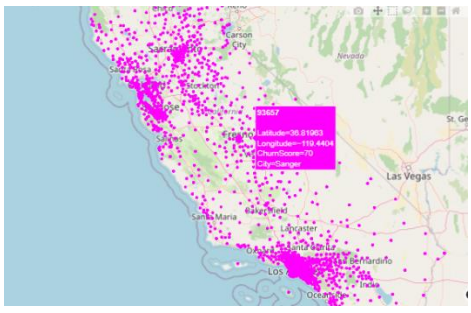

Fig.8. High churners score and their location

E. Result and Discussion

Now after all the cleaning up & pre-processing of the data now we separate our data for further applying algorithms on it.

By using:1. Train-Test Split
2. Model
3. Tuning Model

Results of two types:

1) Class imbalanced data not handled

a) XGBoost

Here we tune the model to increase model performance without overfitting the model.

XGBoost stands for extreme Gradient Boosting. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

Models	Accurarcy
XGBoost	65.80%

b) Logistic Regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

	precision	recall	f1-score	support
0	0.83	0.84	0.83	1048
1	0.51	0.49	0.50	361
accuracy			0.75	1409
macro avg	0.67	0.67	0.67	1409
weighted avg	0.75	0.75	0.75	1409

Models	Accurarcy
Logistic Regression	75.01%

c) Computing confusion matrix:

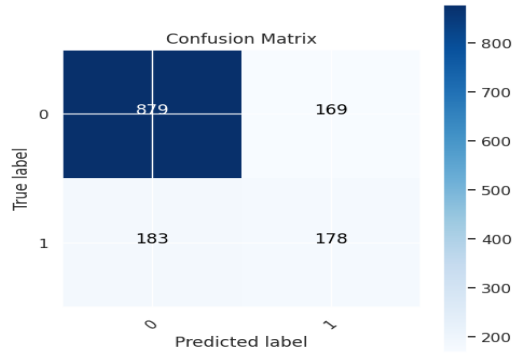

Fig. 8. Confusion matrix

2) Class imbalanced data handled

Applying SMOTE technique to handle class imbalance problem.

Applying PCA - Dimensionality Reduction on training and test data.

Logistic Regression result accuracy is 75.01%.

III. CONCLUSION

Here we had past records of customers who had churned and using that data we predicted whether new customer would tend to churn or not, this will help the companies to get to know the behaviour of customer & how to maintain their interests into the services of company. Further the company can also use recommender system to retain customers and avoid the further churns. We used various algorithms wherein Logistic regression as compared to XGBoost.

With PCA - Dimensionality Reduction, we could not see any better results, hence finalizing the model by Logistic Regression.

REFERENCES

[1] https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset

[2] https://www.kaggle.com/code/hely333/customer-churn-eda-prediction-f1-score-87

[3] http://theprofessionalspoint.blogspot.com/2019/03/implement-pca-in-python-using-scikit.html

[4] https://scikit-learn.org/0.18/auto_examples/model_selection/plot_confusion_matrix.html