3rd National Level Students' Research Conference on "Innovative Ideas and Inventions in Computer Science & IT with Its Sustainability"
In association with International Journal of Scientific Research in Computer Science, Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

Customer Churn Analysis in Telecom Industry using Machine Learning Algorithms

Vinit Gawali, Vatsal Tikiwala, Dr. Sachin Bhoite
School of Computer Science, MIT World Peace University, Pune, Maharashtra, India

ABSTRACT

Customers play a vital role in the telecom industry. Churn prediction is having significant importance according to the telecommunication industry. The Churn analysis is helpful for the company to discover the customers who are probably discontinuing a subscription to a service. Recently, the mobile telecommunication market has been modified from a boom market to a shape of overcrowded. The focus of telecommunication companies is to move from large customer growth to keep customers reliable. For that reason, it is crucial to know which customers are likely to leave the services of the company in future. Our proposed solution is for customer churn prediction for telecommunication companies by applying various machine learning techniques like Logistic Regression. An available dataset on Kaggle is used for model building.

Keywords—Churn analysis, churn prediction, machine learning, data mining, customer relationship management.

I. INTRODUCTION

The term customer churn is usually defined as the percentage (%) of consumers who will stop using the company's outcome or assistance after a certain period of time. Eventually, the studies on customer churn analysis started when the concept of Customer Relation Management (CRM) came into existence. As and when the market started to impregnate due to the internationalism of businesses and furious competition, the hiring cost of customers increased. As and when Loyal customers also increased. Due to such loyal customers there is a healthy competition between the companies in the market and help the companies in their own aspects.
Different kinds of various algorithms have been applied on this dataset such as decision tree, linear regression, logistic regression etc. By referring various research papers, we have concluded that logistic regression should be used as it is a classification task.

A. Types of Churners

Churners are divided into two classes: 1) voluntary and 2) involuntary. The Voluntary Churners are again divided into two categories i.e., deliberate churners and incidental churners. Involuntary churn happens due to unavoidable reasons such as payment failure or server failure of our website while transferring the funds etc. The

voluntary churning is an act where the customers are about to exit the services due to factors such as they are not satisfied with they are getting or they getting better services from other companies. Such Churners are difficult to find.

The incidental churners are those churners which have no intention to leave the service but due to other factors such as change in location or change in Social Structure etc. The deliberate churners happen due to customers which demand change in the technology of current service or want the products on great deals / offers.

II. LITERATURE SURVEY

The existing research, which mainly focuses on regression, classification algorithms or decision trees, and such other machine learning techniques. Customer churn makes mention of circumstances that customers no longer buy services or products of a company due to different facts [1].

While dealing with the customer churn issue in telecommunication business, researchers have largely conducted similar research in many important aspects like churn reasons, win – back approaches, and models creation [2].

The network scale has a major bond with the churn of telecom customers, scrutinized under the influence of client comfort and shifting price of customer churn on other telecommunications services and supposed that when the customer happiness stays constant, the higher the switching cost is, the less likely it is to churn [3].

This paper discussed the tie up among client's point of view, diverting hurdles, customer satisfaction, and customer retention. It is understood that customer's happiness totally correlates with the customer's retention. [4].

In the existing studies, it has shown us that churn analysis plays an important role in customer relationship management. However, in management practice, this analysis will bring mighty losses to the profits and evolution of the enterprises. [5].

III. DATASET

Data/Source: Kaggle

Dataset Description:

Each row in the dataset represents a customer. Whereas each column contains the customer's various attributes like gender, tenure, internet service, contract, mode of payment, monthly charge and total charge. This dataset contains 7043 records and 21 customer's attributes.

Our Data set has 12 String columns, 5 Boolean columns, 5 Integer columns and 2 Others

The "Churn" column is the target attribute of our dataset.

The dataset is explained:

- Services for each are customers signed on for – phones, multiple lines, internet, online security and streaming TV and movies.
- Customer Bank Account Information – how long they have been the client for the enterprise, payment method, paperless billing, monthly and overall billing.
- Demographic information about customers – gender, age range, and if they are married or bachelors and in which area they stay.

IV. METHODOLOGY

In Methodology, we have explained about some of the machine learning algorithms that have been used on the data.

A. EDA

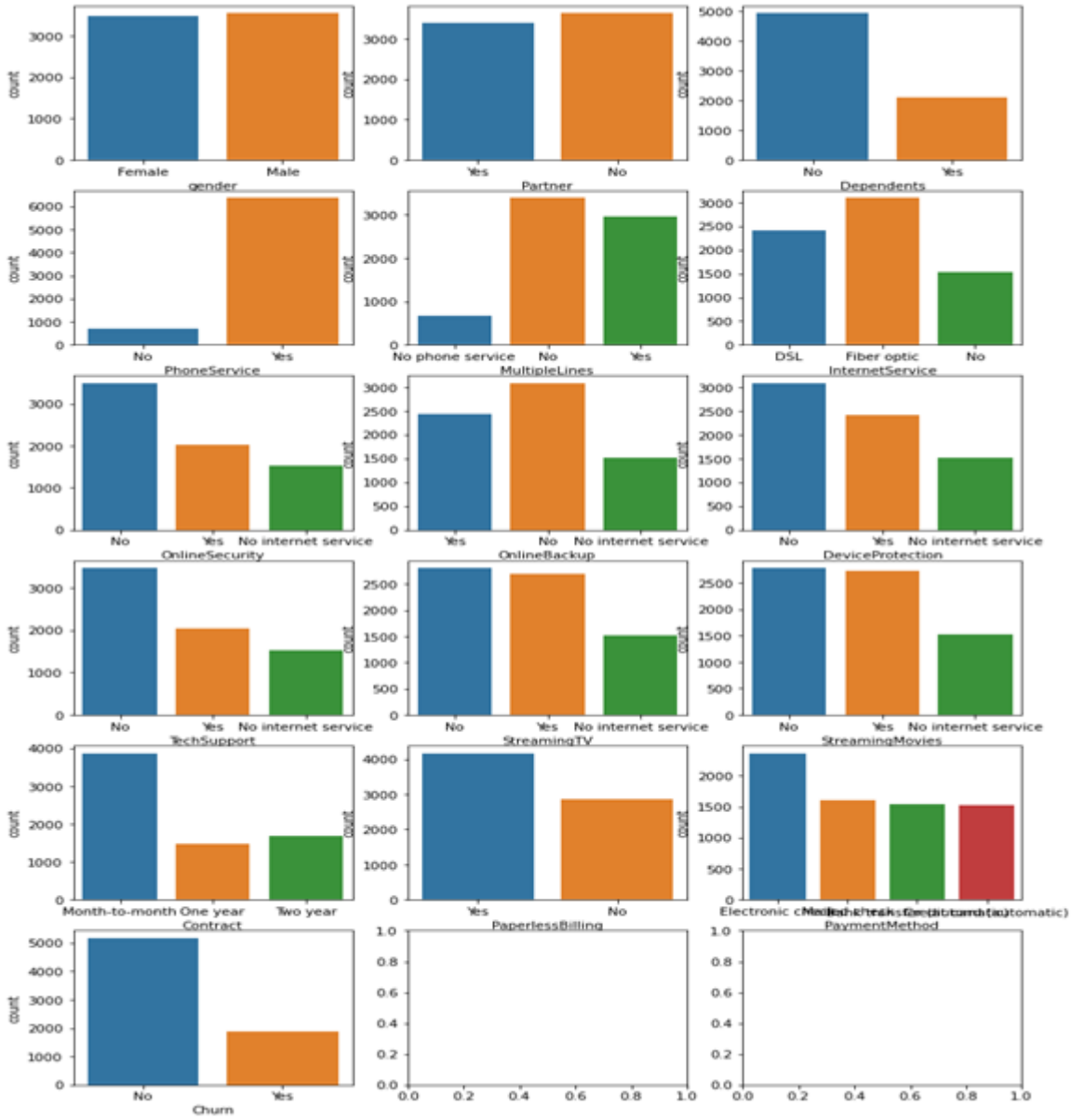

Fig. 1. Countplot

From this count plot, we can see no. of customers in every attributes distributed in three different classes i.e., “Yes”, “No”, “No phone service”.

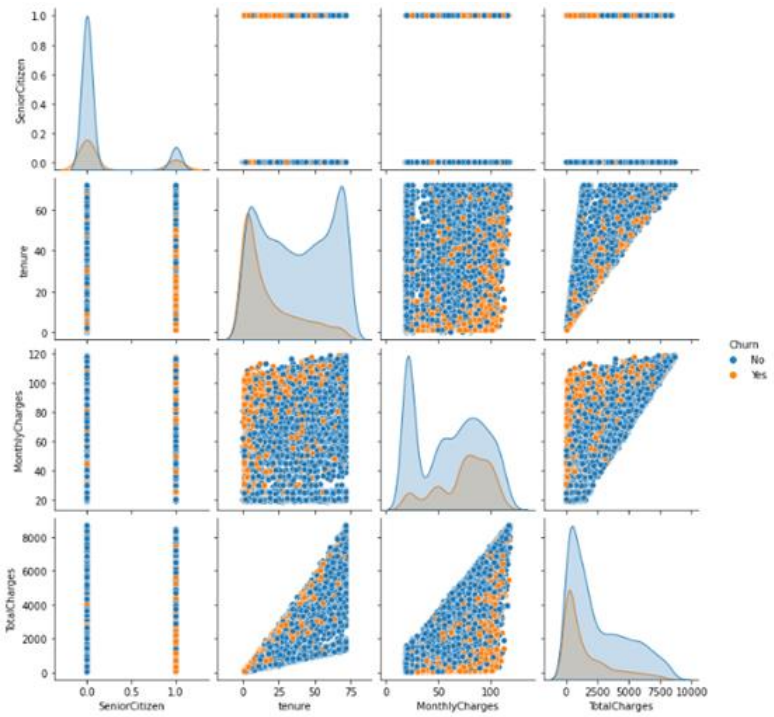

Fig. 2. Pairplot

Target attribute's correlation with other attributes:

Churn	1.000000
PaperlessBilling	0.191825
MonthlyCharges	0.183523
SeniorCitizen	0.150889
PaymentMethod	0.107062
MultipleLines	0.038037
PhoneService	0.011942
gender	-0.008612
customerID	-0.017447
StreamingTV	-0.036581
StreamingMovies	-0.038492
InternetService	-0.047291
Partner	-0.150448
Dependents	-0.164221
DeviceProtection	-0.178134
OnlineBackup	-0.195525
TotalCharges	-0.231873
TechSupport	-0.282492
OnlineSecurity	-0.289309
tenure	-0.352229
Contract	-0.396713

Name: Churn, dtype: float64

Fig. 3. Target's Correlation

Through the above Exploratory Data Analysis, we can understand the correlation of attributes with each other as well as with respect to target attribute.

B. Logistic Regression

Logistic Regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for classification problems as this customer churn analysis is a classification problem. Logistic regression is used to describe data and to explain relationship between one.

Dependent variable and one or more nominal, ordinal independent variables. Therefore, the outcome must be a categorical or discrete value. It must be either Yes or No, 0 or 1, True or False, etc. Logistic Regression can be used to classify the observations using different data types and can easily determine the most effectual variables.

Model accuracy achieved after using Logistic Regression was around 79.61%.

V. RESULTS

Algorithm	True Positive	False Positive	True Negative	False Negative	Accuracy (in %)
Logistic Regression	235	140	1158	228	79.61

In the result we can seen how the confusion matrix tells how the performance of our algorithm is on the dataset. It has given us insights about all the mistakes made by our algorithm and how the confusion matrix helps to interpret the results.

We can say that the variables `Total Charges` and `Senior Citizen` are highly positively correlated. In the correlation matrix we can see the direct correlation between these two variables. We can also spot that there are no outliers present in them.

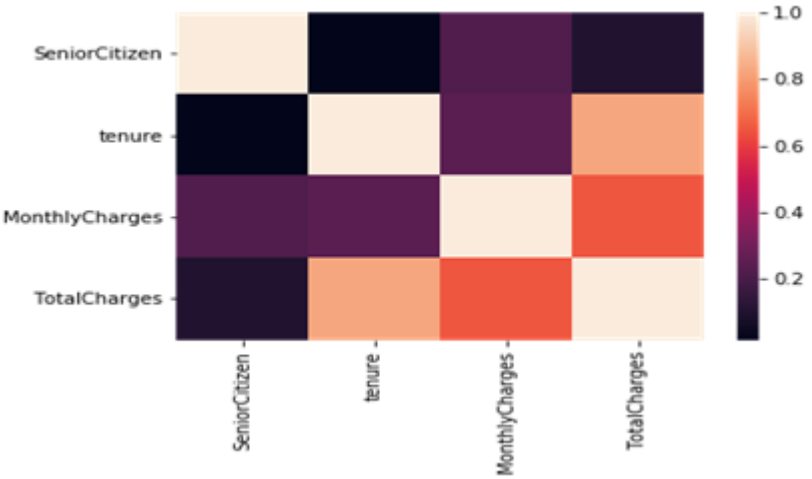

Fig. 4. Correlation Matrix

```
# fit the model with data
lr_model_single.fit(X_train,y_train)
y_pred=lr_model_single.predict(X_test)

lr_acc = metrics.accuracy_score(y_test, y_pred)
print("Accuracy: ",lr_acc)
```

```
Accuracy:  0.7961385576377058
```

Fig. 5. Accuracy

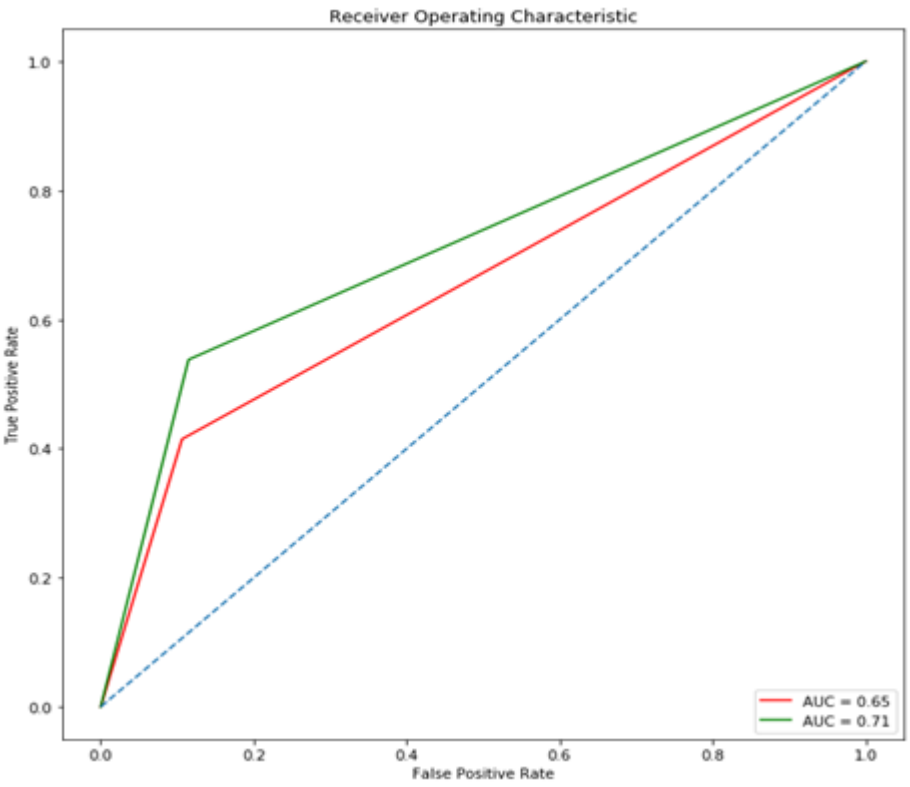

Fig. 6. ROC Curve

Through above ROC (Receiver Operating Characteristic) curve, we can understand the true positive (TP) rate of three classes. We can say that, our proposed model is accurate enough which is providing less false positive (FP) rate than TP (true positive) rate. Because of this behaviour of our model, we are able to achieve 79.61% ≈ 80% accuracy.

VI. CONCLUSION

The main objective of our research is to help companies to make profit. The main purpose of our proposed solution is to predict whether the customers are likely to churn in a telecom industry or not. The analysis of the organizations (with their past data for a particular period of time) helps them to understand a particular day's circumstances of the company and helps the companies to evaluate the churning factor of the company so that the telecom companies can implement new strategies to attract new customers as well as the customers already using their product or services.

VII. REFERENCES

[1]. G. Olle, "A Hybrid Churn Prediction Model in Mobile Telecommunication Industry,", International Journal of E-Education, e-Business, e-Management and e-Learning, https://doi.org/10.7763/ijeeee.2014.v4.302, 2014.

[2]. J. H. Ahn, S. P. Han and Y. S. Lee, "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications Policy,", vol. 30, pp. 552-568, 10.1016/j.telpol.2006.09.006, 2006.

[3]. K. J. Back and B. Barrett, "Influencing factors on restaurant customers' revisit intention: The roles of emotions and switching barriers,", International Journal of Hospitality Management, vol. 28, pp. 563-572, 10.1016/j.ijhm.2009.03.005, 2009.

[4]. V. Umayaparvathi and K. Iyakutti,. "A survey on customer churn prediction in the telecom industry: Datasets, methods and metrics", International Research Journal of Engineering and Technology (IRJET), vol. 3(04), 2016.

[5]. S. Qureshi, A. Rehman, A. Qamar, A. Kamal and A. Rehman, "Telecommunication Subscribers' Churn Prediction Model Using Machine Learning,", 8th International Conference on Digital Information Management, ICDIM, 10.1109/ICDIM.2013.6693977, 2013.

[6]. Q. Bi, "Cultivating loyal customers through online customer communities: A psychological contract perspective", Journal of Business Research, vol. 103, pp. 34–44, 2019, https://doi.org/10.1016/j.jbusres.2019.06.005, 2019.

[7]. Y. Xiao, C. Li, L. Song, J. Yang and J. Su, "A Multidimensional Information Fusion-Based Matching Decision Method for Manufacturing Service Resource", IEEE Access, vol. 9, pp. 39839–39851, https://doi.org/10.1109/access.2021.3063277, 2021.

[8]. F. Reichheld and E. Sasser, "Zero Defections: Quality Comes to Services. Harvard business review", vol. 68, pp. 105-11, 1990.

[9]. T. O. Jones and W. E. Sasser, "Why satisfied customers de- fect," Harvard Business Review, 1995.

[10].H. Jain, A. Khunteta and S. Srivastava, "Churn Prediction in Telecommunication using Logistic Regression and Logit Boost", Procedia Computer Science, vol. 167, pp. 101–112, https://doi.org/10.1016/j.procs.2020.03.187, 2020.

[11].S. Chandrasekhar, "Predicting the Churn in Telecom Industry", 2015.

[12].J. Ondrus and Y. Pigneur, "Coupling mobile payments and crm in the retail industry", 2004.