Telecom Churn Prediction

[1]Anuja Bachhav, [2]Sharayu Patil, [3]Manisha Yadav, [4]Liju Kunjumon, [5]Mr.S.T.Datir
[1] *B. E Student, Department of Computer, Maratha Vidya Prasarak Samaj's Karmaveer Baburao Ganpatrao Thakare College of Engineering, Nashik, India*
[2] *B. E Student, Department of Computer, Maratha Vidya Prasarak Sam,aj's Karmaveer Baburao Ganpatrao Thakare College of Engineering, Nashik, India,*
[3] *B. E Student, Department of Computer, Maratha Vidya Prasarak Samaj's Karmaveer Baburao Ganpatrao Thakare College of Engineering, Nashik, India,*
4] *B. E Student, Department of Computer, Maratha Vidya Prasarak Samaj's Karmaveer Baburao Ganpatrao Thakare College of Engineering, Nashik, India,*
5] *B. E Professor, Department of Computer, Maratha Vidya Prasarak Samaj's Karmaveer Baburao Ganpatrao Thakare College of Engineering, Nashik, India*

[1]Anujabachhav26@gmail.com,[2] sharyaupatil963@gmail.com, [3]My4344974@gmail.com,
4]koshiliju8@gmail.com, 5]Datir.sunil@kbtcoe.org

Abstract— ***Telecom churn has emerged because the single largest reason behind revenue erosion for telecommunication operators. Predicting churners from the demographic and behavioral data of customers has been a subject of active analysis interest and industrial practice. In the telecom sector due to a vast client base, a huge volume of data is being generated daily. Decision-makers and business analysts highlight that attaining new customers is more expensive than retaining the existing ones. Business analysts and customer relationship management (CRM) analyzers need to know the reasons for customer churn, as well as, behavior patterns from the existing data. This project proposes a churn prediction model that uses classification, as well as, clustering techniques to identify the churn customers and help to identify the factors behind the churning of customers in the telecom sector.***
The most essential task of CRM is to create effective retention policies so as to prevent churners. After classification, the proposed model segments the churning customer's data by categorizing the churn customers into different groups to provide group-based retention offers. This project also identifies churn factors that are essential in determining the root causes of churn. By knowing the churn factors from customer data CRM can improve productivity, recommend relevant promotions based on similar behavior patterns, and improve the marketing strategies of the company. The proposed churn prediction model is evaluated using metrics, such as accuracy, precision, recall, f-measure, and receiving operating characteristics (ROC) area.

KEYWORDS: Behavioral Analysis, Customer Churn Management, SVM, XGBoost

I. INTRODUCTION

Telecom businesses generate a massive amount of data at an alarmingly quick rate in today's environment. In order to grow their client base, a variety of telecom service providers compete in the market. Customers have a variety of services to choose from, both better and less expensive.

With the reorganization of the telecom industry, the market rivalry is growing increasingly strong, and the new telecom market capacity is shrinking. Because attracting new customers is becoming increasingly expensive, reducing customer turnover has become one of the most important marketing goals for telecommunications companies.

Traditional market research approaches are very subjective and stagnant, making objective marketing judgments difficult to back.

Large volumes of client data, made available by new information technology tools, have produced opportunities as well as obstacles for firms looking to utilize the data and gain a competitive advantage. Forward-thinking businesses strive to understand each consumer personally and use that knowledge to make it easier for customers to conduct business with them rather than with competitors. The importance of data mining (DM) in boosting marketing and customer relationship management (CRM) through improving the firm's ability to create learning relationships with its customers is generally recognized. In the telephone industry, the term churn refers to any sort of customer loss, whether voluntary or involuntary.

There are three types of churn: voluntary churn, involuntary churn, and expected churn. Attrition is the result of voluntary churn. When a corporation quits a connection involuntarily, such as owing to unpaid invoices, this is known as involuntary churn. When a consumer is no longer in the target market for a product, expected churn occurs.

The focus of this paper is on voluntary churn. The set of data mining procedures used to extract and verify patterns in data is the heart of the knowledge discovery process from a methodology standpoint. These procedures include data selection, data preprocessing, data transformation, data management, and pattern interpretation and evaluation. CRM, customer value analysis, corporate strategy, and positive service mechanisms are some of the approaches used by businesses to improve customer interactions. In order to establish a good CRM, businesses must also improve marketing and sales effectiveness. CRM integrates the functions of related fields in the enterprise with customers, such as brand management, sales, facilities, and tech assistance for customer needs, and it typically uses IT to help an enterprise systematically manage customer relationships, enhancing customer experience and overall business profits. Because customers are the source of revenue, understanding how to accurately track customer turnover is critical to a company's survival and growth. Previous studies have looked at customer churn from a single perspective, such as just from the perspective of customer discontent or customer transition costs.

According to existing research, the key goal is to discover valuable churn customers using a big amount of telecom data. Existing models, on the other contrary, have inherent weaknesses that cause considerable impediments in tackling this issue in the actual world. In the telecom industry, a vast amount of data is generated, much of which comprises missing values, resulting in poor prediction model results. To address these challenges, data preparation techniques are modified to eliminate noise from data, allowing a model to correctly classify data and improve performance. Although feature selection has been employed in the literature, a number of information-rich features are often overlooked when developing models. Statistical approaches are mostly employed in a variety of domains, which leads to poor prediction model outcomes. Models have been validated in previous studies using benchmark datasets that do not reflect the true representation of the data and thus are worthless to the people who make decisions.

Multiple algorithms are applied to the same dataset to overcome this issue, and the optimal feature is chosen for retention. The intelligent framework can aid in the development of automating churn prediction as well as retention models. Feature selection is another fundamental flaw in previous models. Churn is caused by a variety of factors for each client or group of consumers. In the literature, a churning consumer is simply labeled as a churner without considering the reasons for his or her churning. Churners behave in a variety of ways, and they should not all be treated in the same way. Customers who churn are much more likely to do so than those who do not.

An average telecom customer spends $6,000 over the life of a contract. This is the equivalent of $6 per month. If you are a telecom business, you should factor this into your churn prediction. It's a good indicator of whether or not your customer retention strategy is working. Customer churn is the rate at which your customers leave your company. It's an important metric to track. You can see what your churn rate is by comparing the number of customers you have at the end of the month with the number of customers you had at the beginning of the month.

The difference between these two numbers gives you your churn rate. For example, if you have 300 customers at the end of the month and 300 customers at the beginning of the month, your churn rate would be 2%. We can calculate the churn rate using the given formula :

CCR= Total customers at end of monthTotal customers at beginning of month100

Where CCR is customer churn rate.

A model that can forecast churn customers and give retention measures such as special offers for distinct groups of churn customers depending on their churn causes is needed. Using Information Gain as well as Correlation Attributes Ranking Filter feature selection techniques, we selected the top features to get the intended outcomes, despite the limitations indicated above. As a result, as a crucial task of marketing, an effective theory should indeed be studied and implemented for customer churn forecasting.
The urgent need for businesses to keep existing consumers, along with the expensive expense of recruiting new ones, is the driving force behind this endeavor. In the telecom sector, there is a lack of competent rule-based Customer Churn Prediction (CCP) approaches, as well as the inability to predict churn with maximum accuracy, as per a review of the area. Thus, the focus of this research is to bridge these gaps by integrating a model which consists of both SVMs which use a quadratic programming problem approach, and boosting technique that is XGBoost which provides a regularizing gradient boosting framework to enhance the accuracy of determining customer churn in the telecommunication sector.

2. LITERATURE REVIEW

R. Sudharsan and Dr. E.N. Ganesh in 2020 proposed SVM based churn analysis for telecommunication. The model aim to analysis total number of subscriber in voice call routing along the period. The churn analysis can predict the churn rate and the probability for month and the year. The proposed method, classify the customer from churn and non- churn by SVM increase in accuracy for the existing system.

Atallah M. AL-Shatnwai and Mohammad Faris in 2020 they made Predicting Customer Retention using XGBoost and Balancing Methods. In this paper XGBoost classifier is experimented with different oversampling methods to improve its performance in the used imbalanced database.

Pan Tang in 2021 Telecom churn prediction model combining K-means and XGBoost Algorithm. This paper proposes a customer churn prediction model combining K-means and the XGBoost algorithm. First, K-means cluster processing is carried out on the training set, then XGBoost is used to train the clustering groups respectively, and finally, the integrated process is carried out. The results show that the model has better generalization ability.

Miss.Priyanka Parmar and Mrs. Shilpa Serasiya in 2021 made an Telecom Churn Prediction Model using XgBoost Classifier and Logistic Regression Algorithm. In this paper three tree-based algorithms were chosen because of their applicability and range in this sort of application.

Hemlata Jain, Ajay Khunteta and Sumit Private Shrivastav researched on Telecom Churn Prediction Using Seven Machine Learning Experiments integrating Features engineering and Normalization. The study states the importance of data mining techniques for a churn prediction model and proposes a very good comparison model where all machine Learning Standalone techniques, Deep Learning Technique and hybrid models with Feature Extraction tasks are being used and compared on the same dataset to evaluate the techniques performance better.

3. PROPOSED METHODOLOGY

In this system, we have proposed a Telecom Churn Prediction system.This section explains why

the suggested modified SVM & XGBoost methodology is being used. Initially, we will get the dataset, and by data filtering, we removed all the null values. Then we converted all the data into a similar form, which more natural to understand and analyze. By Using EDA analysis we Explore the Data And Find Which Mainly Features Effect to Churn By using SVM and XGBoost and having a different approach, we try to implement a predictor model for the Telecom company. Here we have a customer data set, and by preprocessing and feature selection, we divide the data set for training and testing. For this algorithm, we have made some feature engineering to have more efficient and accurate results using that algorithm. This method aids in the estimate of the customer churn rate across telecom industries.

A. Data Collection

While several other telecom service providers have contributed data to foretell churning, only a few acceptable range datasets have been made available to the research community. IBM telecommunications data, which is freely available on Kaggle, is one of the datasets which will be used in our project.

This dataset comprises one target attribute, "Churn," and also 20 independent attributes comprising 7043 entries of churners. All of the independent fields will be utilized to find target attributes as predictor attributes. The model must be trained with predictor attributes that entail more knowledge for the target attribute.

The information in the data set includes

- Customers who left within the last month – the column is called Churn
- Services to which each consumer has subscribed
- Account information of customers
- Customer demographic characteristics

B. Data Pre-Processing

Following data collection, numerous procedures are taken to investigate the data. The goal of this step is to learn about the data structure, perform preliminary preprocessing, clean the data, discover patterns and inconsistencies in the data (such as skewness, outliers, and missing values), and develop and test hypotheses. This task is responsible for doing "Data Pre Processing" and deleting unfinished, noisy, or untrustworthy data from the system. This process is considered to be the most essential phase in originating such a forecasting structure to find out the churn behavior among the customers.

Table. NAMES OF THE ATTRIBUTES & THEIR DESCRIPTIONS

No.	Attribute	Description	Representation
1	Gender	If the client is a female or a male	Female, Male
2	SeniorCitizen	If the client is a senior citizen or not	0, 1
3	Partner	If the client has a partner or not	Yes, No
4	Dependents	If the client has dependents or not	Yes, No
5	tenure	Number of months the customer has stayed with the company	Multiple different numeric values
6	Contract	Indicates the customer's current contract type	Month-to-Month, One year, Two year
7	PaperlessBilling	If the client has paperless billing or not	Yes, No
8	PaymentMethod	The customer's payment method	Electronic check, Mailed check, Bank transfer, Credit Card
9	Monthly Charges	The amount charged to the customer monthly	Multiple different numeric values
10	TotalCharges	The total amount charged to the customer	Multiple different numeric values
11	PhoneService	If the client has a phone service or not	Yes, No
12	MultipleLines	If the client has multiple lines or not	No phone service, No, Yes
13	InternetServices	If the client is subscribed to Internet service with the company	DSL, Fiber optic, No
14	OnlineSecurity	If the client has online security or not	No internet service, No, Yes
15	OnlineBackup	If the client has online backup or not	No internet service, No, Yes
16	DeviceProtection	If the client has device protection or	No internet service,

a) Storing the raw data into a Comma Separated Values format (CSV) file. Initially, the required dataset has been obtained in an excel sheet format. To Perform the rest of the processes conveniently the dataset has been converted into a CSV file format.

b) Loading the CSV with pandas.

c) Checking for the number of columns and rows

d) Removing unnecessary columns

e) Checking for null and duplicate values. The null and duplicate values make the final prediction model to be overfitted. To overcome this issue the null and duplicate values have been replaced with the mean value of the corresponding column.

f) Converting all variables to a common type. For instance, the total number of complaints and the number of negative feedbacks sent by the customers have been presented in the integer data type. Meanwhile, download bandwidth, monthly bill, and most of the other features have been presented in the double format with decimal points. The integer datatype features have been converted into the double data type.

g) Outlier detection and replacement

1. Initially, outlier values for each column have been identified. (First Quartile (Q1), Third Quartile (Q3), and InterQuartile range (IQR) values of all columns have been used in this regard.)
2. Replacement of random values within the quartile range for outlier detected cells.
3. Validation of each column value is within the quartile range limit.
4. Creation of new CSV without the outlier values.

h) Normalizing the data set between the fixed ranges.

1. Min Max Scaling technique has been adopted in this regard.
2. Normalized Value = (Actual Value – Minimum Value) / (Maximum Value – Minimum Value)

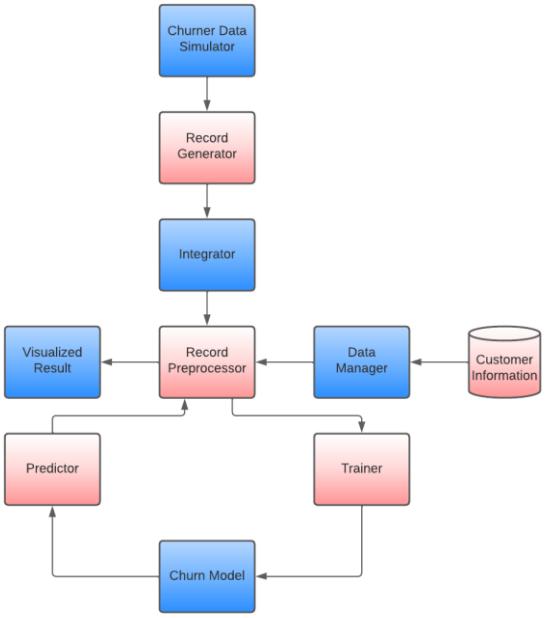

Fig 1. Block Diagram

C. Model Building

This phase includes selecting modeling techniques and architecture, as well as developing and evaluating the model. Because we intend to develop a classification model using SVM, we need to train the algorithm first and then test it, therefore we'll divide the dataset into two parts: a training set and a testing set. After multiple partitioning attempts, we concluded that the optimum performance is obtained if the training set is randomly partitioned to be approximately 80% of the original data set, with 7043 subscribers, and the testing set is randomly partitioned to be about 20% of the original data set, with subscribers.

ALGORITHMS USED

1. SVM

We employed the support vector machines methodology to develop the predictive model in an attempt to solve the subscriber churning problem under this segment.

Boser, Guyon, and Vapnik proposed the Support Vector Machines (SVM) algorithm.

SVM is an exceedingly efficient algorithm. It is the supervised learning methodology used for the two-level of operations like classification, and regression tasks, which stemmed from the statistical learning theory approach.

SVM performs via transforming training data into a higher-dimensional space. A nonlinear function is used to map the data, and then SVM implements linear regression in that dimension. SVM works on the generalization error by minimizing the upper limit rather than the training error. In n-dimensional space, SVM plots data points. SSVM plots data points in n-dimensional space, with n denoting the number of features in the dataset. Then, by establishing a HyperPlane, SVM divides data into two classes.

Even when data are not linearly separable, the kernel functions translate the data input to a high-dimensional feature space, enabling it to be categorized. A hyperplane could be created to represent the decision boundary with the maximum margin. As a result, the model only uses a small portion of the training data at the class boundaries. To demonstrate the SVM algorithm's concept, in Fig. 1. we consider the case of two-dimensional input space, i.e., x ϵ ℜ^2

The coordinates of the observations are the support vectors. These data points aid in the development of SVM. While developing SVM, the margin between the coordinates and hyperplane attempts to be maximized. The loss function is computed as follows:

$$\square(\square,\square,\square(\square)) = \{0, \square\square\ \square * \square(\square) \geq 1\ 1 - \square * \square(\square)\}$$

If the actual and anticipated values are in same range, the loss is 0; otherwise, the loss is calculated. The regularization parameter is provided to the cost function to balance margin maximization and loss. The cost function is expressed by:

$$\square\square\square_{\square}\ \square||\square||^2 + \sum_{\square=1}^{\square} (1 - \square_{\square}(\square_{\square},\square))$$

To discover gradients with regard to weights, and partial derivatives of the loss function are used. Weights are updated using a gradient.

$$\frac{\partial}{\partial \square_{\square}} \square||\square||^2 = 2\square\square_{\square}$$

$$\frac{\partial}{\partial \square_{\square}}(1 - \square_{\square}(\square_{\square},\square))_+ = \{0, \square\square\ \square_{\square}(\square_{\square},\square) \geq 1 - \square_{\square}\square_{\square\square},\square\square\square\}$$

Two key parameters are necessary to train SVM: C and Sigma. The prediction is influenced by the C parameter. It indicates the penalty cost. A significant value of C stands for high training accuracy and low testing accuracy. While a low C value suggests suboptimal accuracy. Sigma values have a greater impact on hyper-parameter partitioning in SVM. Overfitting and underfitting are induced by large and small sigma values, respectively.

2. XGBOOST

eXtreme Gradient Boosting is abbreviated as XGBoost.

The term xgboost, on the other hand, refers to the engineering aim of pushing boosted tree algorithms' computing resources to their limits. One of the reasons xgboost is so prominent is provided above. XGBoost is a high-speed, as well as high-performance gradient, boosted decision tree implementation.

Boosting is an ensemble approach for fixing flaws in old models by adding new models to them. One by one, models are introduced until no further improvements are conceivable. For example, the AdaBoost algorithm weights data points that are difficult to predict.

Gradient boosting is a method that involves creating new models that forecast the residuals or inaccuracies of previous models, which are then combined to form the final prediction. Gradient boosting gets its name from the fact that it uses a

gradient descent approach to minimize loss when adding new models.

Both regression and classification predictive modeling issues are supported by this technique.

Input: training set $\{(x_i, y_i)\}_{i=1}^{N}$, a differentiable loss function $L(y, F(x))$, a number of weak learners M and a learning rate α.

Algorithm:

1. Initialize model with a constant value:

$$\hat{f}_{(0)}(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i, \theta).$$

2. For $m = 1$ to M:

1. Compute the 'gradients' and 'hessians':

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)},$$

$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)}.$$

2. Fit a base learner (or weak learner, e.g. tree) using the training set $\left\{x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right\}_{i=1}^{N}$ by solving the optimization pro

$$\hat{\phi}_m = \arg\min_{\phi \in \Phi} \sum_{i=1}^{N} \frac{1}{2}\hat{h}_m(x_i)\left[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i)\right]^2.$$

$$\hat{f}_m(x) = \alpha\hat{\phi}_m(x).$$

3. Update the model:

$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x).$$

3. Output $\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x).$

D. Performance Measure

It's pointed out that the prediction performance for imbalance data set could not be evaluated sufficiently in terms of average accuracy. The prediction model should perform well on both positive and negative examples rather than only on one class at the cost of the other class. Thus the performance of our proposed churn prediction model is measured in terms of precision, recall, true positive rate (TPR), F-measure and G-mean. These metrics are defined as:

$$precision = \frac{TP}{TP+FP}, \; recall = \frac{TP}{TP+FN}, \; TPR = \frac{TN}{TN+FP},$$

$$F-measure = \frac{2 \cdot recall \cdot precision}{recall+precision}, \; G-Mean = \sqrt{recall \times TNR}.$$

The meanings of TP, FP, FN, TN are stated in following confusion matrix Table 2. The higher the values of all the indicators, the better the performance of the churn prediction model. Moreover churn cases are treated as positive examples and non-churn cases are treated as negative examples in this paper.

4. WORKING

Working of a system defines the flow with which the actions would be performed. Fig 2

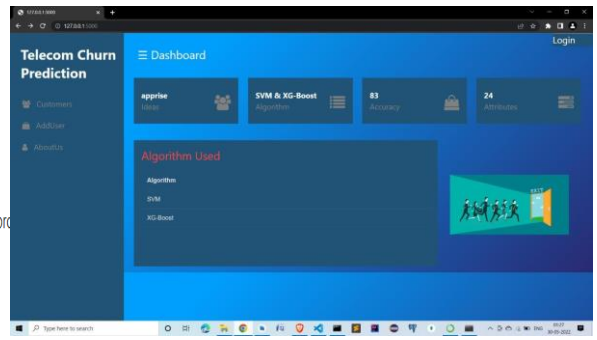

Fig 2 : Dashboard

shows the initial dashboard that is seen once the system is started. Dashboard provides with basic description about the project

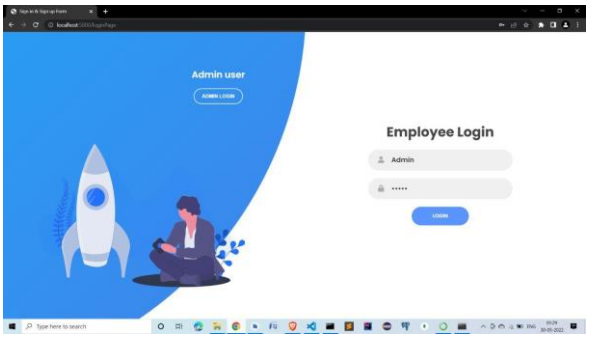

Fig 3 : Login Panel

Fig 3 shows the login panel. Here there are two login options. They are Employee Login and Admin Login.

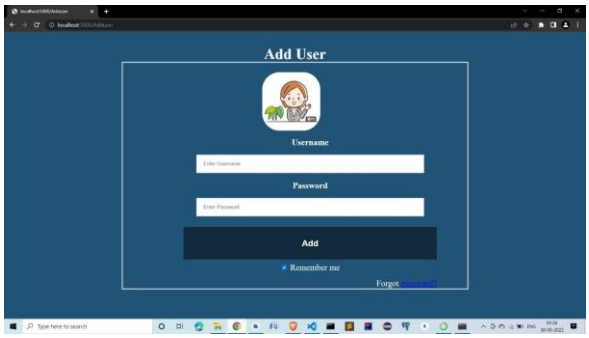

Fig 4 : Admin Panel

Fig 4 shows the Admin Login. Here once the admin logins he is able to perform certain actions like add employee, delete employee.

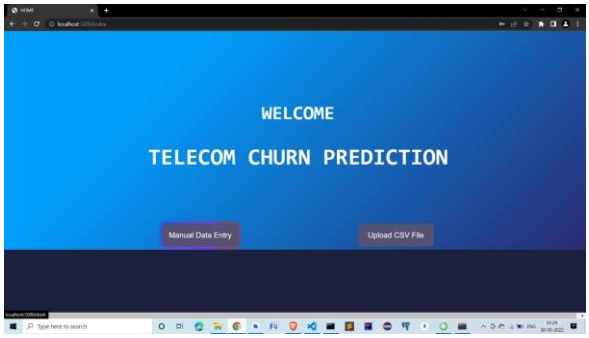

Fig 5 : Menu

Fig 5 shows the Menu. Once the employee logins he/she is able to see two options. First is Manual Entry and second is Upload csv

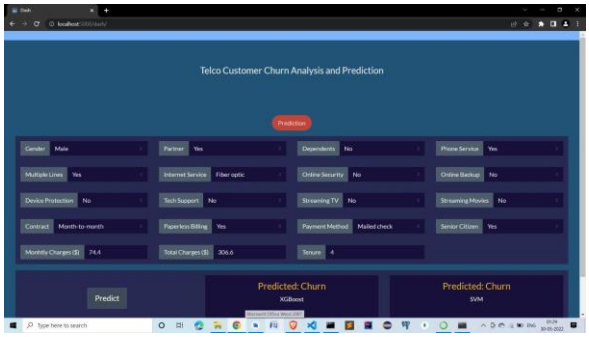

Fig 6 : Manual Entry

Once the employee clicks on Manual Entry the screen shown in Fig 6 is seen. In this the Employee inputs all the necessary data and prediction of whether the customer will churn or not is provided using SVM and XGBoost algorithm.

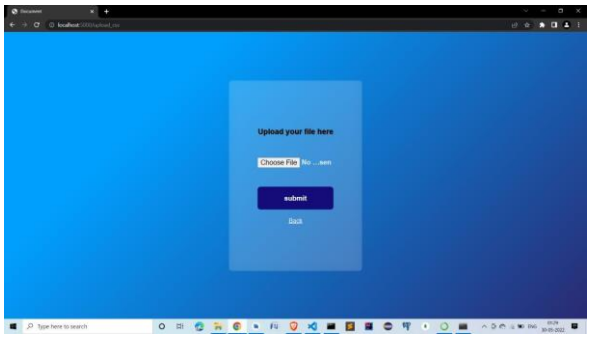

Fig 7 : Upload CSV

If the Employee clicks on Upload csv option he/she gets a screen to upload a csv file that is shown in Fig 7.

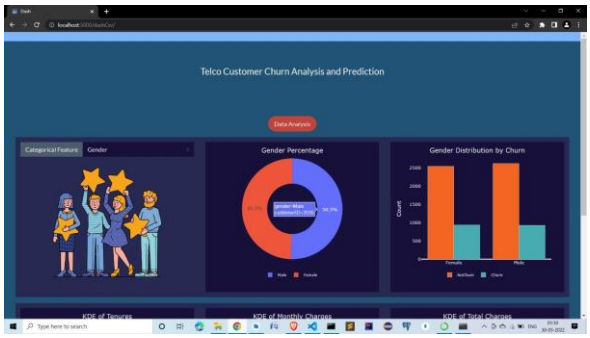

Fig 8 : Visualization

After uploading this another screen opens which is Fig 8. It shows the visualization of the dataset and helps us understand the probability of customers churning all at once

5. CONCLUSION

Customer churn is an important and complex problem that faces telecommunication companies globally. This problem can lead to serious profit loss. Therefore, companies always look for powerful prediction models which help in predicting those customers who are about leaving. However, developing models for churn prediction is not straight forward for many reasons. One of these, is the nature of the data (imbalanced data distribution) which makes applying the normal evaluation metrics is not appropriate. In this paper, we applied a Support Sector Machine (SVM)

approach for developing a churn prediction model. In order to achieve better performance, the parameters of the SVM were optimized using a grid search with a customized evaluation metric which can be adapted regarding the cost of the retention campaign strategy. The developed SVM model was evaluated and compared to literature-based classical classification techniques. XGBoost is a high-speed, as well as high-performance gradient, boosted decision tree implementation. Thus it also provides better performance. The proposed model showed promising results and high prediction power.

REFERENCES

1. R. Sudharsan, Dr. E.N. Ganesh, "SVM Based Churn Analysis For Telecommunication," IJARET_11_06_049, pp. 534-544, June 2020.
2. Atallah M. AL-Shatnwai, Mohammad Faris "Predicting Customer Retention using XGBoost and Balancing Methods" International Journal of Advanced Computer Science and Applications, Vol. 11, No. 7, 2020
3. Miss.Priyanka Parmar , Mrs. Shilpa Serasiya "Telecom Churn Prediction Model using XgBoost Classifier and Logistic Regression Algorithm" International Research Journal of Engineering and Technology (IRJET)Volume: 08 Issue: 05 | May 2021
4. P. Tang, "Telecom Customer Churn Prediction Model Combining K-means and XGBoost Algorithm," *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, 2020, pp. 1128-1131
5. Hemlata Jain, Ajay Khunteta , Sumit Private Shrivastav. "Telecom Churn Prediction Using Seven