Customer Churn Prediction Using Machine Learning

M.S N Lakshmi[1],G. Sridatta Anirudha Nivas[2], Ch. Teja Venkat[3], J. Hemanth Purna Kumar[4], G. Prudhvi[5]

[1]Asst. Professor, Department of Computer Science and Engineering

[2,3,4] Student, Department of Computer Science and Engineering

[1,2,3,4] Vignan's Lara Institute of Technology & Science

Abstract-In any Industry customer plays a pivotal role, whether it is telecom, finance, banking, IT or any other industries. Nowadays, customer churn is a major problem in every industry. In the telecom industry customers churn i.e. they may change from one network to another network, this might be due to various reasons in their existing service such as high charges, network issues, security services etc. The main aim of this research is to solve this churn problem in the telecom industry. Due to the advancement in the machine learning field, we decided to use the machine learning techniques to overcome this issue. The churn prediction model mainly works on customer features, support features, Usage features and contextual features. Unlike most research that uses regression techniques as a method to boost the accuracy, this paper uses efficient algorithm i.e. Light GBM classifier to achieve better results compared to previous works. An accuracy of 90% was achieved. Thus , this classifier(Light GBM) is suggested for customer churn prediction.

Keywords—*Machine learning, churn, LightGBM*

I. INTRODUCTION

Nowadays, customer churn is the major problem in many industries. By considering different Industries, the churning issue is more in the telecom industry compared to others, because churning in the telecom industry is a very simple process. In telecom, the term "churn" refers to the loss of subscribers who switch from one service provider to another service provider during a given period. The globalization and advancements of the telecommunication industry, exponentially raises the number of operators in the market that escalates the competition. In telecom Industry, the communication between different customers should be fast and reliable. Nowadays customers are expecting better services with reasonable charges in the mobile network.

Based on a previous study, the estimated average churn rate for the mobile telecom industry is about 2.25% per month. This means that every month 1 in 50 subscribers of a company discontinues their services. In order to retain the new customer is better than providing better services to the existing customers to make them happy and not to churn to other service. Telecom industry have the responsibility to predict the customers who are ready to churned. With the advancements in Machine Learning, different algorithms like gradient Tree boosting, Random Forest, naïve bayes, Light GBM etc…are used to predict the customer churn.

II. LITERATURE SURVEY

Praveen et al. [1], provided comparative analysis of machine learning models for customer churn prediction, where they adopted support vector machine, decision tree, naive bayes, & logistic regression. Thereafter, they also observed how boosting algorithms performed on the classification accuracy. In the obtained results, SVM-POLY 123 P. Lalwani et al. using AdaBoost performed better than others. However, the classification accuracy may be further improved by incorporating feature selection strategies like uni-variate selection.

Horia Beleiu et al. [2], they adopted three machine learning approaches, namely, neural network, support vector machine and bayesian networks for customer churn prediction. In the feature selection process, principal component analysis (PCA) is taken into consideration to reduce the dimensions of the data. But the feature selection process can be improved using optimization algorithms which increases the classification accuracy. within the performance evaluation, gain measure and ROC curve was used.

J. Burez et al. [3], authors tried to capture the class imbalance problem. They applied logistic regression and random

forest by using re-sampling technique. In addition, boosting algorithms were also applied. within the performance analysis, AUC and Lift are taken into consideration. They also observed the effect of advanced sampling techniques such as CUBE, but the obtained outcome did not improve the performance. However, still the category imbalance problem may be solved in an exceedingly better way by using the optimization-based sampling techniques.

J. Hadden et al. [4], analyse the variables that impact churn in reverence. They also provided the comparative study of three machine learning models like neural network, regression trees and regression. The obtained results confirm that decision tree is superior to others due to its rule-based architecture. The obtained accuracy may be further improved using the prevailing feature selection techniques.

III. PROPOSED SYSTEM

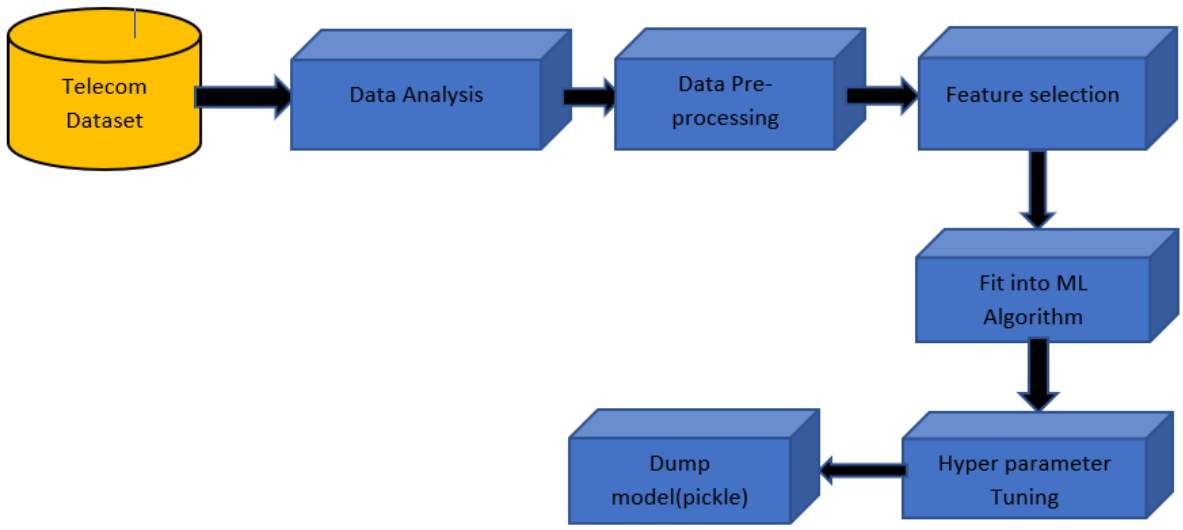
The proposed system consists of 8 phases

About Dataset

The data set contains 7043 rows (customers) and 21 columns (features) and is taken from https://www.kaggle.com/datasets/blastchar/telco-customer-churn

The data set includes information about:

- Customers who left within the last month – the column is named Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method,
- paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they need partners and dependents
- Churn - dependent feature ('Yes' denotes customers left, 'No' denotes customer stay here)



Architecture of proposed system

Data Analysis (EDA)

It is a way of exploring the hidden features that are present in the rows and columns of data by visualizing, summarizing and interpreting data.

Once EDA is done, meaningful insights are drawn that can be used for supervised and unsupervised machine learning modeling. Some different techniques can even be wont to gather more information and insights about customers by following innovative solutions. In our telecommunication data-set we divided the data-set into two parts: 1st Categorical features and 2nd Numerical features. From 21 features, 16 features were categorical and 5 were numerical.

Data Pre-processing

In the dataset, there are 11 missing values in the total charges column. So we replace these values with mean values. In this dataset there are no null values, so we don't want to perform Handling missing values**.**

Feature Engineering

The main aim in this step is to find the correlation and splitting dataset between the independent and dependent feature.

Feature Selection (SelectKBest)

In our proposed system, we mainly select 10 features which have higher correlation by using feature selection methods. From sklearn using feature selection modules importing the SelectKBest to select the important features.

According to the feature selection, we select 10 out of 21 features. These are the 10 features selected [Dependents, tenure, Online Security, Online Backup, Device Protection, Tech Support, Contract, Paperless Billing, Monthly Charges, Total charges] and we split the data into training and testing in 80-20 ratio as follows

X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.2)

Fit into Algorithm (ML Algorithm)

Since our data is divided into training and testing, we apply machine learning algorithms on training data. Apart from the existing algorithms we used algorithms such as random forest, Gradient boosting classifier and Light GBM.Due to imbalance dataset the results were not upto mark, so we implemented SMOTEENN methods to balance the dataset.

Hyper Parameter Tuning (RandomSearchCV)

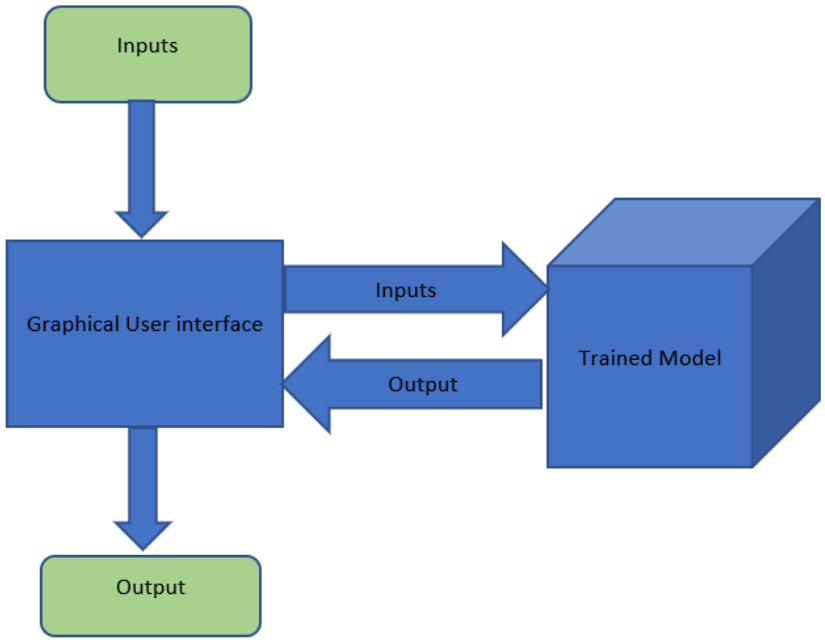
From these results we get better accuracy and TP FP ratio also increases in Light GBM, so we performed Hyperparameter Tuning for this model only. In our proposed model we perform hyperParameter tuning using RandomizedSearchCV method.

Dump model (Pickle)

We saved the model using pickle module and this model is able to predict the output based on the user inputs.

Creating GUI using Python

This user Interface will serve as a frontend to our model and it takes inputs from the user and predicts our desired output



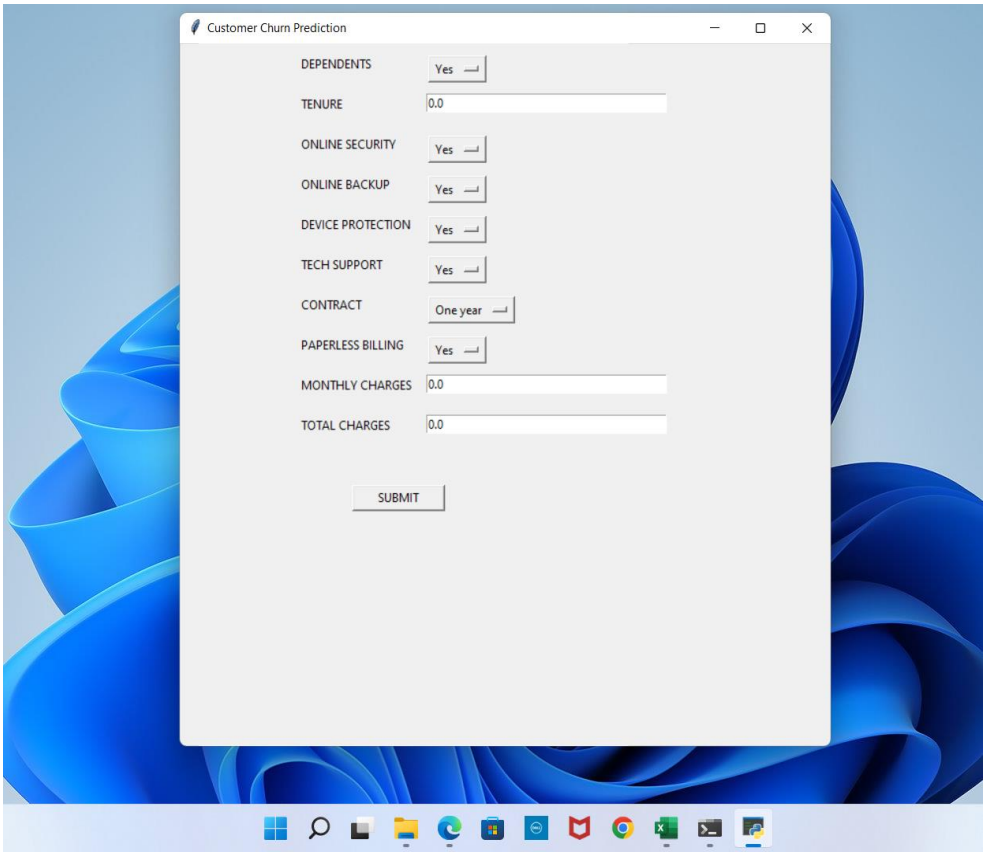
ADVANTAGES OF PROPOSED TECHNIQUE OVER THE EXISTING

The merits of the proposed algorithm have listed as follows:

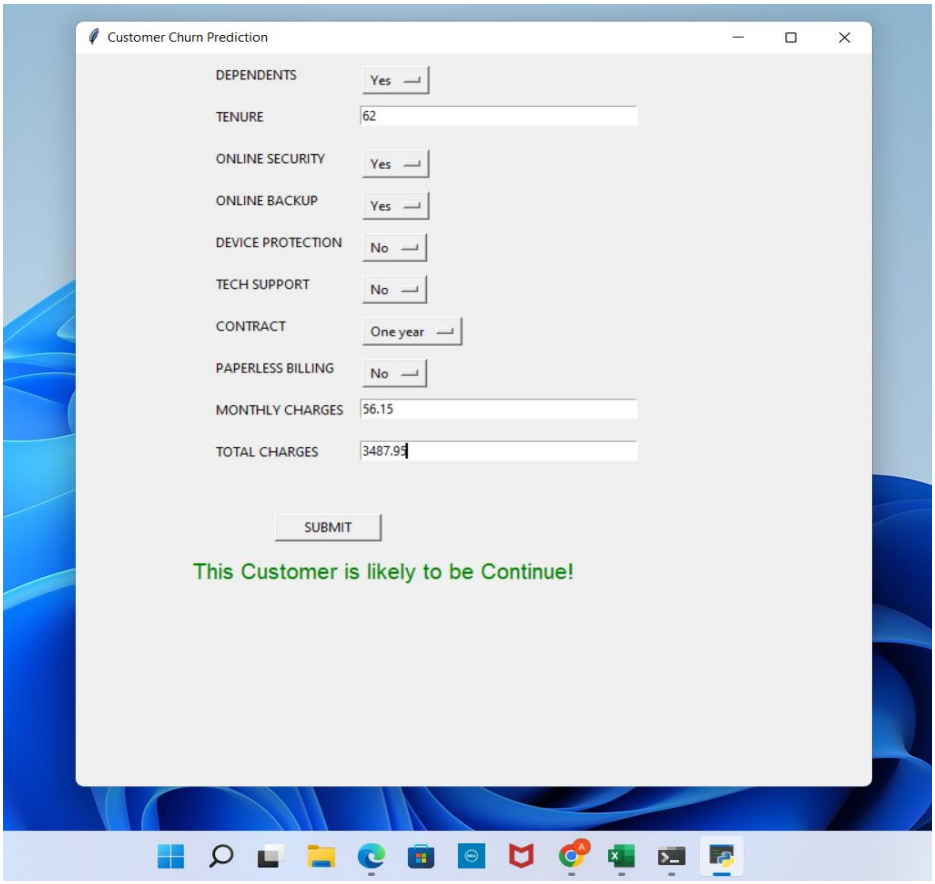
- We have applied SelectKBest to perform feature selection and to reduce the dimensions of the data-set, in contrast to existing approaches where prediction accuracy is low due to improper feature selection [5,6].
- The obtained accuracy was low in existing methods due to an imbalanced dataset, to overcome this problem we implemented SMOTEENN methods to balance the dataset.
- We have implemented User interface using python, so that the model is well able to work with new values apart from the dataset.
- Then we have evaluated the algorithms on the test set using confusion matrix and AUC curve, which have been mentioned in the form of graphs and tables in order to compare which algorithm performs best for this particular data-set, in contrast to the existing techniques where obtained results are not properly evaluated [7,8].

IV. RESULTS

The output screens of our work are shown below



The above screen shows the initial GUI with some default values and it is created using the tkinter module. This UI serves as a front end for our model at the back end. In the next step ,we enter some random values as an input to the model.



In the above screen for the given inputs the output is “This customer is likely to be continue!” which tells us the customer will not churn. The model was tested with both known and unknown inputs and an accuracy of 90% was achieved for this model.

V. FUTURE SCOPE AND CONCLUSION

Telecommunication industry has suffered from high churn rates and immense churning loss. Although the business loss is unavoidable, churn can still be managed and kept at an acceptable level. Good methods need to be developed and existing methods have to be enhanced to prevent the telecommunication industry from facing challenges. In this paper we discussed the various prediction models and also compared the quality measures of prediction models like random forest ,Gradient boosting classifier and Light GBM. We found that the accuracy achieved with lightGBM is higher than the other techniques which clearly states that it is an efficient technique.

The future scope of this paper will use hybrid classification techniques to point out the existing association between churn prediction and customer lifetime value. The retention policies need to be considered by selecting appropriate variables from the dataset. The passive and the dynamic nature of the industry ensure that data mining has become an increasingly significant aspect in the telecommunication industry prospect.

VI. REFERENCES

[1] Asthana P (2018) A comparison of machine learning techniques for customer churn prediction. International Journal of Pure and Applied Mathematics 119(10):1149–1169

[2] Brându¸soiu, I., Toderean, G., Beleiu, H.: Methods for churn prediction in the pre-paid mobile telecommunications industry. In: 2016 International conference on communications (COMM), pp. 97–100. IEEE (2016)

[3] Burez J, Van den Poel D (2009) Handling class imbalance in customer churn prediction. Expert Systems with Applications 36(3):4626–4636

[4] Hadden J, Tiwari A, Roy R, Ruta D (2007) Computer assisted customer churn management: State-of the-art and future trends. Computers & Operations Research 34(10):2902–2917

[5] Gürsoy U ¸S (2010) Customer churn analysis in the telecommunication sector. ˙İstanbul Üniversitesi ˙İşletme Fakültesi Dergisi 39(1):35–49

[6] Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., Dai, W., Yang, Q., Zeng, J.: Telco churn prediction with big data. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data, pp. 607–618 (2015)

[7] Idris, A., Khan, A., Lee, Y.S.: Genetic programming and adaboosting based churn prediction for telecom. In: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1328–1332. IEEE (2012)

[8] Kirui, C., Hong, L., Cheruiyot, W., Kirui, H.: Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining. International Journal of Computer Science Issues (IJCSI) 10(2 Part 1), 165 (2013)