# *Customer Churn Prediction using Machine Learning*

Samiksha Upadhyay
*Dept. of Computing Technologies*
*School of Computing*
*SRM Institute of Science and Technology*
Kattankulathur,
India
su8458@srmist.edu.in

Rajalakshmi. M
Assistant Professor
*Dept. of Computing Technologies*
*School of Computing*
*SRM Institute of Science and Technology*
Kattankulathur,
India
rajalakm2@srmist.edu.in

***Abstract*—**One significant problem that businesses face is customer attrition. It has become crucial for corporate operations and growth to prevent customer churn and work to keep clients. It is challenging to effectively estimate customer turnover because the majority of the existing projections use a single prediction model. Concentrating on the results of predictions of the models of machine learning, this study proposes a combination of estimating model for customer turnover and performs practical research on the model's efficacy. The combined prediction model outperforms the single customer churn prediction model in terms of accuracy and predictive impact, according to the findings of the predictions. It can also more naturally express the fundamental traits of the churn consumers.

## I. INTRODUCTION

The percentage of customers who discontinue doing business with a company over time is referred to as customer churn. This may be the result of a number of factors, including unhappiness with the good or service, a better deal from a rival, or a change in the client's requirements.

Businesses should monitor and lower customer churn because it can significantly affect their revenue and profitability. A high rate of client turnover may be a sign that a business is underperforming and losing ground to rivals.

Companies can take a variety of actions to lower customer turnover, including enhancing the quality of their goods and services, offering better customer support, establishing loyalty reward schemes, and assessing customer input to pinpoint improvement areas. Finding customer turnover is crucial because it significantly affects a company's revenue and profitability. When customers leave doing business with a firm, it affects both the revenue of the business and the expense of attracting new clients to take their place.

Also, figuring out why customers leave a company might help that company develop better goods, services, or customer experiences. Companies may take corrective action to fix those issues, boost customer happiness, and lessen the possibility of future churn by identifying why customers are leaving[6]. Reducing customer turnover can also result in more loyal customers, which has a number of advantages like a higher lifetime value for customers, positive word-of-mouth marketing, and a market edge. Customer turnover may be predicted and reduced with the help of machine learning, which is a potent instrument. Machine learning models can help organizations identify customers who are at danger of leaving and assist

them in taking preventative actions to keep them. Many machine learning algorithms, such as logistic regression, decision trees, random forests, and neural networks, can be used to forecast customer turnover[10]. These algorithms can be taught using past data to spot trends and indicate which clients are most likely to leave. A company can take a number of actions to keep its customers after determining which ones are likely to leave, like providing them with special offers or discounts. The elements that are causing client churn, such as quality of product, customer service, or pricing, can also be found using machine learning[4]. Businesses can learn important insights into the reasons why customers are departing and take corrective action to address those issues by evaluating customer feedback and behavior data. Overall, machine learning can be an effective tool for firms to boost client retention and decrease churn, which will increase sales and profitability. Figure 1 depicts the architecture model of the system described in this paper.

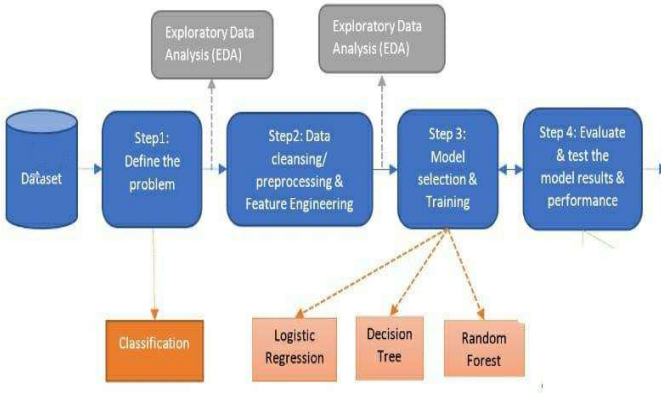


*Figure 1. Architecture model*

## II. LITERATURE SURVEY

At the moment, classical statistics-based forecasts and predictions based on integrated classifiers are both used in domiciliary and global users turnover prediction algorithms. In order to estimate customer attrition using machine learning techniques and statistical theory. employed consumer visual insights to find relations between indicators[1]. MGUIIS and CO. created a predictive model using logistic regression depending on how long retail customers spend on average per transaction. The augmented decision tree model was used by Du Gang and Huang Zhenyu to anticipate the buying habits of consumers[2]. To confirm the usefulness of the new technique in predicting consumer buying behavior, they compared the effects of their analysis before and after optimization.

The customer retention analysis was conducted by the authors using a logistic regression model. Age, gender, the kind of registration and length of service use, the type of phone, and the monthly price are used to train and evaluate the model. This initial step's test results show 74% correctness. The accuracy level of this model increased to 79% after researchers supplemented the prior data set with subscriber internet usage data[8]. Lastly, at the end of this investigation, researchers demonstrated that combining the two distinct data sets indicated above can significantly raise the level of accuracy[9]. My opinion is that when looking at the churn prediction model, they neglected to consider several crucial elements that could affect subscriber decision-making processes, such as the most recent packages utilized by customers, their happiness with customer support, etc[11]. So, this model cannot be used in the future to identify client turnover causes. In any case, this research is beneficial.

In the Improved Churn Prediction Method, SNA ideas are used to forecast the churn of telecom consumers[3]. The method entails three steps: quantifying tie strength, applying machine learning techniques to combine traditional and social variables, and influence propagation model[13]. Before clients cancel the service, a pattern analysis framework is suggested to give strategic planners advice. The chat graph approach of churn prediction concentrates on forecasting the churn in the conversation activity[5]. This methodology does not take into account the social aspects derived from graph theory. According to their online actions, users are divided into groups for

churn prediction using the clustering method, which then applies retention rules to keep them from leaving[15]. The Churn Prediction by Local Community Detection approach uses a greedy scheme to divide the network into communities. This approach disregards the network's structure and content. Identification of potential users inside an operator's network is the main emphasis of Churn Prediction by Utilizing the Diffusion Process. The network's diffusion process can be directed using graph theory.

## III. PROPOSED WORK

Predictive churning model is a tool for classifying, a system that examines the traits of potential consumers to determine what traits are essential in forecasting turnover rates. Let's imagine we have a dataset with information on ten thousand clients who are taking money out of a bank[1]. These clients' characteristics, including their country of residence, credit score, age, and balance, among others, are described in the data.

The outcome of the user's turnover should be predicted by our model. Hence, the target variable will be terminated. The data should be examined with an emphasis as to how various aspects connect to the customer churn status[14].

We are prepared to construct many models in search of the optimum fit. Forecasting customer turnover is a problem of binary classification since clients can leave or stay for a predetermined amount of time.
We'll test:

- Logistic regression classifier
- Naive Bayesian
- Decision Tree
- Kernel SVM
- KNN
- Random Forest
- Support Vector Machine with Radial basis function kernel

These models need to be worked on and we'll do so using the the given steps:

- Search for Parameters: We'll choose the parameters and values we want to look for in each of our models. The best parameters found in our model will be set when we run the GridSearchCV.
- Best Models Fit: We train the system using the train dataset after determining the best estimator.
- Performance Evaluation: Using our test set, we will evaluate the models that performed the best after being trained on our training dataset.

## IV. IMPLEMENTATION

### *1. Data Visualisation*

Substantial discoveries from your data can aid the company's development. However, the problem is that one cannot necessarily draw conclusions from simply looking at the data. Patterns, links, and other astounding revelations that might not otherwise be obvious become evident when you visually analyze your information[11].You develop storytelling skills by using data visualization to bring your data to life and reveal the hidden meanings. Through live data visualizations, dynamic reports, graphical displays, and various other illustrations, data visualization allows consumers to swiftly and efficiently develop compelling

business insights.

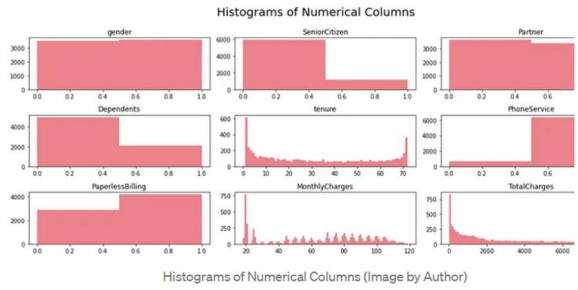


*Figure 2. Histogram of numerical data*

We can draw a bunch of conclusions based on the histograms represented in Figure 2.

According to the dataset's gender distribution, there are roughly equal numbers of male and female clients. In our dataset, we have an equal number of men and women.

Younger clients make up the majority of the dataset's customers.

While roughly 50 percent of the users are sharing the plan with their partner, it seems that few customers have dependents.

The company has a large number of recent clients, along with a devoted consumer division which are constant for, on average, more than 70 months.

Most customers seem to require access to a cellphone, and 75 percent of them prefer transactions without paper. Invoice fees per user each month vary between eighteen dollars to one hundred and eighteen dollars, with a large majority of consumers falling into the twenty dollars bracket.

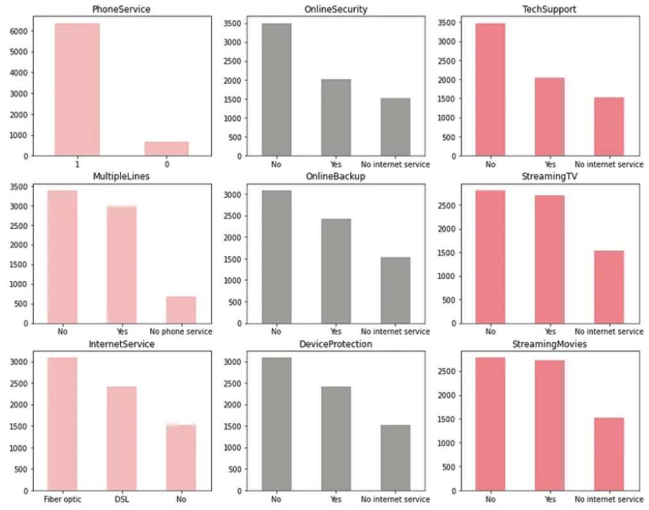


*Figure 3. Distribution of Label Encoded Categorical Variables*

From the plots of Figure 3, we learn several things:

A. Nearly half of consumers have several lines of service, while almost all of consumers have a connection to cell phone service.
B. More than half of internet users watch TV shows and movies online, and 3/4 of customers choose fiber-optic and DSL lines for their internet access.
C. Only a tiny number of consumers have used the safety measures, technological help, and internet backup features.

The plots teach us various things, including:

A. Churning customers are older than those who are kept.

b. There is no distinction between lost and maintained clients in terms of the median credit score or tenure.

d. The majority of the clients that leave the business appear to still have a sizable sum in their bank accounts.

d. Customer churn appears to be unaffected by expected wage or the number of items.

We need to undertake some feature engineering before we search for a model to forecast customer turnover.
Several of the characteristics in the dataset are clubbed together to make the latest characteristics that more accurately characterize our clients. While credit scores usually rise over time (and subsequently with age), as we previously observed, they have no impact on churning, we will develop a new feature to take this into consideration.

## 2. *Exploratory Data Analysis*

Key findings of EDA performed on this model

- There are no incorrect or missing data values in the dataset.
- Monthly Charges and Age have the strongest positive correlations with the goal qualities, whereas Partner, Dependents, and Tenure have the strongest negative correlations.
- Due to a lot of consumers having motion, the dataset is not balanced.
- Relationship between Monthly-Charges and Total-Charges is multicollinear. The VIF values have significantly lowered as a result of dropping Total Charges.
- Younger clients make up the majority of the dataset's customers.
- A substantial portion of recent customers (those under a year old) make up the majority of the company's clientele, which is accompanied by a loyal clientele that is older than 70 months.
- The majority of users appear to have phone service, with monthly costs per user ranging from $18 to $118.
- If they have chosen to making payments with online checks, a significant number of consumers with an every month subscription have an excellent chance of doing so as well.

## 3. *Classification Models*

Classification precision is one of among the most well-liked classification assessment indicators used to assess baseline techniques due to the quantity of precise forecasts made as a fraction of all predictions[5]. Nevertheless, when there are issues with disparities in class, it is not the most beneficial statistic. The "Mean AUC" score, which gauges the extent to which the model's predictions are able to differentiate between both favorable and adverse classes, will thus be used to categorize the data[4].

| | Algorithm | ROC AUC Mean | ROC AUC STD | Accuracy Mean | Accuracy STD |
|---|---|---|---|---|---|
| 0 | Logistic Regression | 84.12 | 1.65 | 74.60 | 1.26 |
| 1 | SVC | 83.64 | 1.68 | 79.98 | 1.08 |
| 4 | Gaussian NB | 81.82 | 1.79 | 68.99 | 1.46 |
| 6 | Random Forest | 81.72 | 2.02 | 78.47 | 1.57 |
| 2 | Kernel SVM | 79.66 | 2.12 | 79.85 | 1.08 |
| 3 | KNN | 77.04 | 2.38 | 75.77 | 1.09 |
| 5 | Decision Tree Classifier | 65.44 | 1.67 | 72.75 | 1.45 |

*Figure 4.Compare Baseline Classification Algorithms 1st Iteration*

Figure 4 depicts the comparison of the algorithms used and their accuracy compared. The first cycle of foundation algorithms for classification revealed that the logistic regression model and SVC scored better than the remaining five models, according to the dataset's greatest mean AUC Scores. Figure 5 compares the Accuracy scores in graphical form and we can see that logistic regression has a good accuracy compared to the rest.

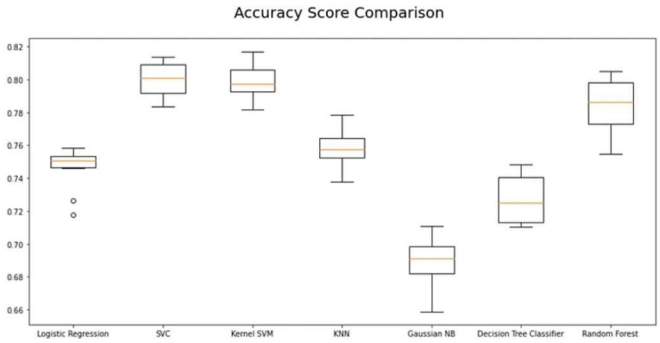


*Figure 5. Accuracy Score Comparison*

## V. RESULT

Overall, the models run successfully and we found logistic regression to be most useful in this case. Hence, the improvement of this model has been focused on and we have got better accuracy. The final result is depicted in the form of a confusion matrix as shown below in Figure 6.

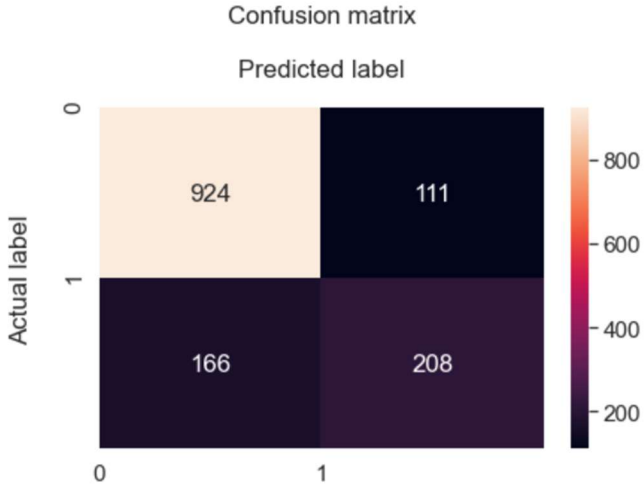


*Figure 6. Confusion Matrix*

We have 208+924 correct predictions, according to the Confusion matrix, and 166+111 wrong ones. With an accuracy of 80%, our model demonstrates the qualities of a respectable model.

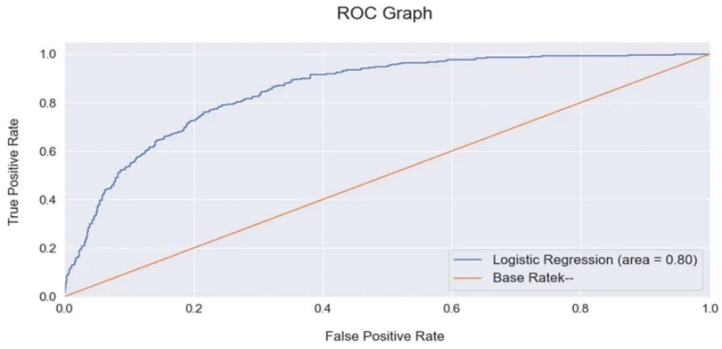


*Figure 7. ROC Curve*

It makes sense to reevaluate the system by the Receiver operating characteristic graph. Depending on the AUC Mean score The ROC graph in Figure 7 depicts a model's ability to differentiate among categories. The orange line depicts the Base Rate which is the ROC curve of a random classifier, is something that a machine learning model tries to avoid the best it can. The graph above shows that the enhanced Logistic Regression model had a greater area under the curve score.

## VI. CONCLUSION

The logistic regression model forecasts an increase in the turnover rate because of factors including a monthly contract, optical fiber internet connection, online payments, no guarantee of secure payment, and technical help.

Whereas, if any customer has a one-year contract, online security subscription, or has chosen to use postal checks as their payment method, the model predicts a negative link with churn.

## VII. REFERENCES

[1]C. Zhenhai, Liu. Wei, “Logistic Regression Model and Its Application,” Journal of Yanbian University(Natural Science Edition), vol. 38(01), pp 28-32, 2012.

[2]Z. Qian, Meng. Deyu, Xu. Zongben, “L_(1/2) Regularized Logistic Regression,” Pattern Recognition and Artificial Intelligence, vol. (05), pp. 721-728, 2012.

[3]X. Zhang, G. Feng and H. Hui, "Customer-Churn Research Based on Customer Segmentation," 2009

International Conference on Electronic Commerce and Business Intelligence, Beijing, China, 2009, pp. 443-446, doi: 10.1109/ECBI.2009.

[4]Peng Li, Siben Li, Tingting Bi and Yang Liu, "Telecom customer churn prediction method based on cluster stratified sampling logistic regression," International Conference on Software Intelligence Technologies and Applications & International Conference on Frontiers of Internet of Things 2014, Hsinchu, 2014, pp. 282-287, doi: 10.1049/cp.201

[5]O. Rezaeian, S. S. Haghighi and J. Shahrabi, "Customer Churn Prediction Using Data Mining Techniques for an Iranian Payment Application," 2021 12th International Conference on Information and Knowledge Technology (IKT), Babol, Iran, Islamic Republic of, 2021

[6]A. Larasati, D. Ramadhanti, Y. W. Chen and A. Muid, "Optimizing Deep Learning ANN Model to Predict Customer Churn," 2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), Malang, Indonesia, 2021

[7]P. Bhuse, A. Gandhi, P. Meswani, R. Muni and N. Katre, "Machine Learning Based Telecom-Customer Churn Prediction," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India, 2020

[8]Y. Xiao, C. He and J. Xiao, "Study on Customer Churn Prediction Methods Based on Multiple Classifiers Combination," *2009 Third International Symposium on Intelligent Information Technology Application*, Nanchang, China, 2009

[9]I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," in *IEEE Access*, vol. 7

[10]I. Kaur and J. Kaur, "Customer Churn Analysis and Prediction in Banking Industry using Machine Learning," *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Waknaghat, India, 2020

[11]X. Hu, Y. Yang, L. Chen and S. Zhu, "Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network," *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, Chengdu, China, 2020

[12]Chen Mingliang. Discussion on the framework of the basic theoretical
system of customer relationship management[J]. Journal of
Management Engineering, 2006,20 (4): 36- 41

[13]A. SaranKumar and D. Chandrakala, “A survey on customer churn prediction
using machine learning technique,” Int. J. Comput. Appl., vol. 154,
no. 10, 2016

[14]M. Ali, A. U. Rehman and S. Hafeez, "Prediction of Churning Behavior of Customers in Telecom Sector Using Supervised Learning Techniques," *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*, Kathmandu, Nepal, 2018

[15]I. M. M. Mitkees, S. M. Badr and A. I. B. ElSeddawy, "Customer churn prediction model using data mining techniques," *2017 13th International Computer Engineering Conference (ICENCO)*, Cairo, Egypt, 2017