Anticipating Customer Churn in Telecommunication using Machine Learning Algorithms for Customer Retention

[1]Rajarajan.K
Department of Computing Technologies, College of Engineering and Technology,
SRM Institute of Science and Technology,
Kattankulathur - 603 203, India
rk8764@srmist.edu.in

[2]Vishal Reddy.D
Department of Computing Technologies, College of Engineering and Technology,
SRM Institute of Science and Technology,
Kattankulathur - 603 203, India
vd4465@srmist.edu.in

*[3]Priya.S
Department of Computing Technologies , College of Engineering and Technology,
SRM Institute of Science and Technology,
Kattankulathur - 603 203, India
priyas3@srmist.edu.in

***Abstract*— Customer attrition is a significant issue and a top priority for large corporations. This study addresses this pressing issue. The dataset is obtained from Kaggle which constitutes of training set and testing set, 80 percent and 20 percent of the entire dataset respectively for identifying customers who tend to unsubscribe in the telecommunications industry. Predicting customer churn with precision is a difficult endeavor, mostly due to the dependence on a singular prediction model in most existing projects. To address this, this study proposes a novel approach that compares multiple prediction models like logistic regression, KNN, random forest, SVV, Gaussian NB, Kernel SVM, Support Vector Machine, to estimate customer churn. With ROC AUC Mean score of 84% for Logistic Regression produced a better result. In the future, the study might be expanded to investigate the evolving behavioral patterns of churn consumers with the application of Artificial Intelligence approaches for predictive and trend analysis to preserve valuable client.**

Keywords—customer churn; Telecom sector; churn prediction models; Logistic Regression

I. INTRODUCTION

A. Overview

The percentage of customers who end their commercial relationship with a company over time is known as customer churn, also known as customer attrition. This can happen for a number of reasons, such as changes in the customer's needs, receiving a better offer from a competitor, or unhappiness with the product or service. Customer attrition is a significant issue and a top priority for large corporations. Companies, particularly in the telecom industry, are actively researching ways to predict client turnover in order to mitigate its impact on their revenues. Preventing customer attrition and retaining their client base has now become a crucial problem for corporate operations and growth. Identifying and resolving customer churn is crucial, as it has a substantial impact on a company's financial performance. When clients terminate their commercial association with a company, it not only diminishes revenue but also escalates the expenses incurred in attracting new customers to fill their place. Moreover, comprehending the underlying causes of customer attrition can aid the organization in improving its offerings, services, and overall customer satisfaction. The user's text is empty. By identifying these problems, the decrease in customer attrition can result in many advantages, such as enhanced client allegiance, heightened customer lifetime value, favourable oral message promotion, and a competitive border. One significant method that can help with both forecasting and reducing client attrition is machine learning. Machine learning techniques that analyse historical data and identify patterns that indicate which customers are more likely to leave, such as logistic regression, decision trees, random forests, and neural networks, can be used to forecast customer churn. After identifying these vulnerable customers, companies can employ tactics such as providing exclusive incentives or discounts to ensure their loyalty. Moreover, machine learning has the capability to identify the precise elements that lead to customer churn, such as price issues, customer service issues, or superior products, by looking at behavioural data and customer reviews. Machine learning is an effective way for businesses to increase sales, lower attrition, improve customer retention, and increase profitability.

B. Approach

Customer churn is the rate at which customers abandon their relationship with a company within a given timeframe. It is essential for businesses because it directly affects revenue and profitability. Churn can occur due to various factors such as subpar customer service, exorbitant prices, inadequate product quality, or insufficient customer participation. External factors such as market shifts, new entrants, and economic downturns can all have an impact on churn. To reduce churn, businesses must prioritize endeavors aimed at enhancing the customer experience, delivering exceptional customer service, offering competitive pricing, and consistently engaging with customers. The discernment of churn patterns also empowers businesses to pinpoint areas of vulnerability and enact strategic enhancements to retain customers.

II. RELATED WORK

The authors Nyoman et al. have designed a framework with various preprocessing techniques combined and an ensemble of two machine learning models XGBoost and random forest [1]. The dataset is obtained from a public dataset platform; the experiment uses two different sectors: telecom as well as insurance. This study achieved 0.8501 F1-score in the dataset for telecom sector and the insurance sector resulted in 0.9471 F1-score and twenty eight seconds in processing time. Compared with the latest work in the same dataset, our model achieved a greater effectiveness in F1-score performance and efficiency performance in dataset 1, but slower algorithm time in dataset 2.

The extensive analysis which encompassed a spectrum of machine learning techniques that include Support Vector Machines, Random Forests, and Decision Trees, Logistic Regression, Naive Bayes was made by Diaa Azzam et al in [2]. Furthermore, they have utilized ensemble learning approach to increase the predictive accuracy of our churn prediction models. We also used a voting classifier refined with the best features within our dataset. The voting classifier yielded an accuracy rate of 81.565%, underscoring the effectiveness of our approach in addressing the critical issue of customer churn in the Telecom industry. The authors Yunjie Lin et al. have developed and SNA concepts are applied in the Improved Churn Prediction Method to predict telecom customer churn [3]. The analysis starts by emphasizing the importance of addressing customer turnover in the current competitive environment, pointing out the role of data analytics in these efforts [4]. It proceeds to examine different machine learning techniques that can predict customer churn, and it compares their outcomes to identify the most effective algorithm across various real-life settings, such as telecommunications, finance, and online retail. The study identifies that Decision Tree, XGBoost, AdaBoost, and Random Forest Classification methods are best suited for predicting churn. Moreover, the paper discusses how to apply these insights within a churn prediction tool. The analysis in paper [5] by Eunjo Lee et al. reveals that the distribution of customer lifetime value is heavily skewed, with a small number of users accounting for the majority of sales, while most users who leave the service tend not to make purchases. As a result, it's more economically viable to target churn prediction efforts towards loyal, paying customers who bring substantial value. Additionally, optimizing the prediction model's threshold to maximize expected profit proves to be more beneficial than focusing solely on accuracy. This approach was implemented in the context of Aion, a highly popular online game in South Korea, demonstrating that targeting a specific segment of users yields greater cost-efficiency in terms of campaign expenditure and conversion rates, compared to a model that considers all users.

Drawing from the predictive outcomes and certainty levels of both decision tree and neural network models, this study constructs an integrated customer churn prediction model and performs empirical analysis to evaluate its efficacy [6]. The findings reveal that this hybrid model surpasses the accuracy and predictive performance of standalone churn prediction models. Additionally, it offers a more straightforward representation of the key attributes of customers likely to churn. The research in paper [7] aims to assess the precision of Machine Learning Models in forecasting retail customer churn by comparing their k-means clustering predictions to the customers' real RFM (Recency, Frequency, Monetary) attributes. Through this analytical comparison, the study seeks to determine the reliability of using the output from one predictive model as input for another and its potential implications for further applications. In paper [8], the authors employed a range of predictive models, XGBoost among them, to analyze the dataset. They attained an accuracy of 82.81% on the original dataset, demonstrating the efficiency and advanced technological prowess of the predictive model.

The research in [9] focuses on developing supervised models to identify customers at risk of departing from companies. It involved testing 21 different classification techniques across datasets from the telecommunications, finance, and online retail sectors. Furthermore, the study utilized RFM (Recency, Frequency, Monetary Value) segmentation, a straightforward yet powerful tool for marketing analysis that ranks customers based on their purchasing behavior, in conjunction with the Chi-Square Test for dimensionality reduction. The goal was to derive feature subsets of optimal size and to evaluate the performance of models both before and after the selection of features. The authors Venkatesh and Jeyakarthic have presented a novel approach combining an optimal genetic algorithm (OGA) with a support vector machine (SVM) model for Customer Churn Prediction (CCP) [10]. The OGA is developed through the application of a double chain quantum genetic algorithm. Subsequently, this optimized OGA is utilized to fine-tune the SVM parameters, specifically C and gamma (Y). The effectiveness of the OGA-SVM model is evaluated using a standard dataset from the telecommunications sector. The experimental results demonstrate the model's superior performance, achieving a sensitivity of 94.51, specificity of 66.059, accuracy of 90.269, F-Score of 94.31, and a kappa value of 61.167.

The methodology introduced in paper [11] integrates customer segmentation with churn prediction, offering telecom companies a detailed analysis of customer churn. This strategy assists in reducing operational costs and enhancing customer retention through tailored interventions, thereby providing a competitive advantage by utilizing data-driven insights for boosting overall profit margins and customer contentment. In the experimental phase, four machine learning (ML) classifiers are utilized to ascertain the likelihood of customer churn. To address the challenges posed by imbalanced datasets, the Synthetic Minority Oversampling Technique (SMOTE) is implemented. The experimental results identify the Gradient Boosting Classifier as the most effective technique, achieving an impressive accuracy rate of 95.13%. The work in [12] concentrates on diverse machine learning strategies for forecasting customer churn, enabling the construction of classification models including SVM, Logistic Regression, Random Forest, and Gradient Boosted Trees. Additionally, it provides a comparative analysis of the performance of these models.
A classifier-based approach for predicting customer churn based on profile data has been introduced by the authors Mohamed Galal et al [13]. The study applies and compares several supervised classification techniques, such as KNN (k-Nearest Neighbors), Logistic Regression, AdaBoost, Gradient Boosting, and Random Forest. To improve prediction accuracy, a unified voting mechanism is implemented across these classifiers. The model's performance is further optimized through hyperparameter tuning. Key features are identified and ranked using the Random Forest approach. In a case study involving a dataset of 10,000 entries, the model reached an accuracy rate of 87.00%. The authors explore and implement a method for analyzing customers who have left using data mining techniques, which not only aids in retaining customers at risk

of leaving but also supports businesses in focusing on product enhancement and managerial improvement [14]. They construct a bank credit card customer churn prediction model using a Stacking approach, aimed at forecasting the departure of credit card clients. This model demonstrates superior accuracy compared to those based on XGBoost, Random Forest, and Logistic Regression, performs efficiently, and offers improved implications for customer management. The researchers who presented their work in [15] utilize a dataset that includes customer transactions and demographic details to develop a predictive model for customer churn, applying various data mining techniques like Decision Trees (Dt), Random Forests, and Support Vector Machines (SVMs). This research provides a solid example of how data mining methods can be used to forecast customer churn in practical situations. The importance of selecting appropriate data mining techniques and feature selection methods to improve the accuracy of churn prediction models is highlighted. The findings of this study have practical applications for businesses across several sectors, such as finance, e-commerce, and telecommunications, which depend on retaining customers for their success.

The research work in [16] focuses on developing a classification model utilizing an optimized artificial neural network (ANN) algorithm through deep learning to forecast customer churn rates. The flexible nature of deep-learning ANN is leveraged as an advantage. The findings reveal that the optimized ANN model achieves an accuracy of 76.35%. The construction of the model incorporates an epoch parameter set to 30, a hidden layer count of 50, and uses tanh as the activation function. Among the variables, contract type, service type, and IPTV emerge as the top three factors influencing customer churn. According to the predictions of the optimized deep learning-ANN model, there are 2,568 customers identified as likely to churn, while 4,388 customers are predicted to remain loyal. The churn rate identified by this model stands at 36.41%, marking an 11% increase from the previous year. The study in [17] highlights the importance of early prediction of customer behavior for maintaining customer loyalty, especially in real-time marketing strategies. The research presents the challenges associated with forecasting customer turnover in the motor insurance industry, incorporating a variety of data mining approaches, including recent developments in deep learning and machine learning. Additionally, it focuses on customer churn within the management cycle and its environmental context. A systematic survey overview is provided, detailing the development of churn prediction models, the diverse predictive techniques employed, and their implementation in the business realm. It is comprehended and understood in general that making new clients are costlier than to holding existing client [18]. There is a current issue that customer leave the organization because of obscure reasons. In investigation by the authors in this research, after predicting the churn behavior of the client by utilizing diverse data mining methods, which will in the long run help in breaking down client's behavior and characterize whether it is a churning client or not. We use online available and accessible data set at Kaggle repository and for forecasting the Customer behavior we use various algorithms while having achieved 99.79% accuracy level using Bagging Algorithms.

The data mining methods including classification, clustering, and association rules are utilized in [19]. The success of any customer retention effort critically depends on the accuracy and precision of the techniques employed. Ultimately, without identifying a customer on the verge of departing from the business, a company cannot take appropriate measures to retain that customer. Authors of this research paper in [20] have analyzed various studies and collected and conducted researches revolving around some type of feature selection by comparison of different methods. Here feature selection constituted of methods like embedded method, wrapper method and filter method. The most established and accurate feature selection type found in this survey was Fisher Score, which falls under the filter method, which was then followed by Random Forest, which falls under the embedded method.

III. Proposed Work

Several elements are typically included in the system architecture for Customer Churn Prediction using Machine Learning in order to develop, train, and execute a churn prediction model as mentioned in Figure 1. The first step, data collection, is obtaining and preparing customer data from various sources, such as transaction logs, customer reviews, and demographic data. Following that, the data is modified, cleaned, and prepared for modeling. The second part is feature engineering, which involves choosing and modifying pertinent data properties to raise the prediction model's accuracy. Techniques like feature scaling, feature selection, and dimensionality reduction may be used for this. To efficiently detect and anticipate customer churn inside a firm, the customer churn prediction system architecture consists of several components and phases. Once the data collection phase is completed after that, the data is kept for later processing and analysis in a central data repository, such a data warehouse or big data platform. The data preparation or preprocessing, which involves feature engineering, cleaning, and transformation of the collected data. This step ensures that the data's format is suitable for analysis and modeling. Activities that could be included include handling missing values, data normalization, and feature creation based on domain knowledge. The model creation step of the design comes after data evaluation.

At this point, statistical or machine learning methods are used to develop prediction models. In order to find trends and pinpoint the main causes of customer attrition, these models are trained on historical customer data, which includes both non-churned and churned customers. Various techniques, based on the available data and the complexity of the task, can be used, by including logistic regression, decision trees, and neural networks. Models are evaluated using appropriate performance metrics after development, including area under the curve (AUC), recall, accuracy, and precision. This analysis helps assess the model's predictive accuracy for client attrition. If the model's performance is judged acceptable, it can be used for real-time prediction in a production setting. During the implementation stage, the predictive model needs to be incorporated into the customer relationship management (CRM) platforms or existing business systems. The model uses input data, such as customer characteristics and behavior, to generate churn forecasts for individual customers. Through resource allocation, personalization of customer interactions, and prioritization of retention initiatives, these forecasts can help lower the probability of attrition. Last but not least, a feedback loop is included in the system architecture to support the churn prediction model's continuous improvement. By collecting feedback on the forecast accuracy and monitoring the actual churn results, the model may be regularly retrained and adjusted to take into account changing consumer behavior and market dynamics. To sum up, the customer churn prediction system architecture includes data collection, preprocessing, model building, assessment, deployment, and ongoing improvement. It enables businesses to better target customer retention strategies, proactively identify potential department customer and take preventive actions so that the consumers never get any dissatisfaction.

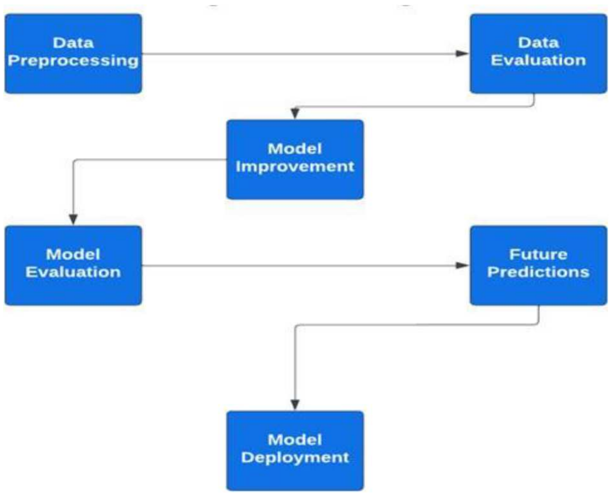

Figure 1. Proposed Workflow of customer churn prediction

Choosing the best machine learning technique for the task at hand is the third component, or model selection and improvement. Finding the best approach in terms of accuracy, precision, recall, and F1 score may necessitate comparing a number of techniques, such as logistic regression, support vector machines, decision trees, and random forests. The training of the model, which is the fourth component, entails using techniques like cross-validation and hyperparameter adjustment to optimize the model's performance by training the chosen algorithm on the prepared data. To make sure the trained model generalizes well to new data, its performance is assessed on a holdout set of data as the last step which is known as model evaluation. The final stage, known as "model deployment," entails using the learned model to create predictions by applying it to fresh data. In order to accomplish this, the model might be made available as a microservice or REST API that can be incorporated into already-existing apps. Overall, the Customer Churn Prediction using Machine Learning system architecture has a number of components that require expertise in machine learning algorithms, feature engineering, data pretreatment, and deployment infrastructure. Additionally, the system needs to be scalable, trustworthy, and maintainable in order to handle enormous data volumes and continue making accurate predictions over time. Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (where there are only two possible outcomes).

In our study, LR is used to estimate the probability of churn based on customer features. Kernel Support Vector Machine (SVM)is an extension of the support vector machine that uses a nonlinear mapping to transform the original training data into a higher dimension. In this space, a linear separator is then constructed. With kernel SVM, we tackle the non-linear relationships between customer attributes and churn, enhancing the model's prediction capability. Random Forest (RF)is an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the mode of the classes output by individual trees. RF contributes to the robustness of our predictive model by addressing the variance in predictions. K-Nearest Neighbors (KNN)is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. KNN is included in our approach to incorporate the principle of similarity, where the churn status of similar customers is used to predict the churn likelihood of a new customer. Decision Tree algorithm is a decision support tool that uses a tree-like model of decisions and their possible consequences. It is one way to display an algorithm that only contains conditional control statements. Decision trees help in breaking down a complex decision-making process into more manageable parts, making the model's predictions more interpretable. These models were selected for their diverse approaches to capture various aspects of customer behavior and attributes. By integrating and comparing these models, we aim to harness their collective predictive power for a more accurate and comprehensive churn prediction system.

IV. Results And Discussion

Significant findings from your data may support the growth of the business. Graphical analysis of data reveals patterns, linkages, and surprising discoveries that may not have been evident otherwise. Utilizing data visualization can enhance your narrative skills by bringing your data to life and revealing its underlying significance, which is shown in Figure 2.

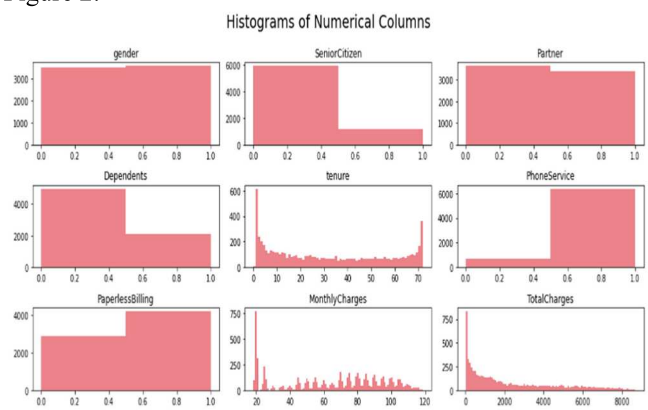

Figure 2: Exploratory Data analysis

Exploring and visualizing the data set by doing distribution of independent variables to better understand the patterns in the data and to potentially form some hypothesis is shown in Figure 2. Through dynamic reports, graphical presentations, live data visualizations, and other visuals, data visualization enables users to quickly and effectively provide appealing business insights. The gender distribution of the dataset shows that the proportion of male and female clients is about equal. There are exactly equal numbers of males and women in our dataset.

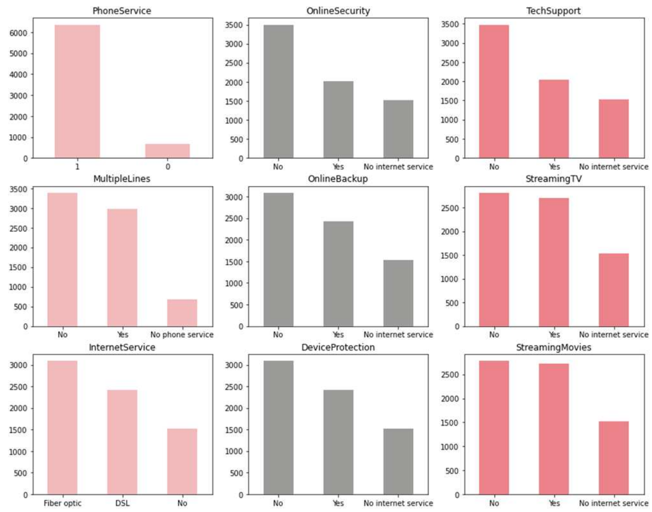

Figure 3 Distribution of label encoded variable

From the plots of Figure 3, virtually all clients are subscribed to smart phone assistance, over 50% buyers have multiple lines of service. Over half of internet users view films and TV series online, and 3/4 of users select DSL and fiber-optic Internet connection. Very few customers have utilized the safety precautions, technical assistance, and internet backup features. Among the many lessons the plots provide are the following: Churning customers tend to be older than those who are retained. Positive and negative correlations with churn rate is shown graphically in Figure 4. Correlation matrix helps us to discover the bivariate relationship between independent variables in a dataset. Which is basically a statistical technique used to evaluate the relationship between two variables in a data set. The churn rate increases with monthly charges and age. In contrast Partner, Dependents and Tenure seem to be negatively related to churn.

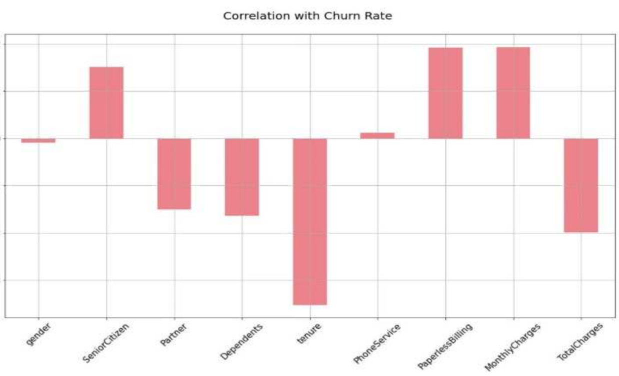

Figure 4. Correlation with churn rate

Classification Accuracy is a widely used metric for evaluating baseline algorithms. It measures the proportion of right predictions out of the total number of predictions. Nevertheless, it is not the optimal measure when we encounter a class imbalance problem. Therefore, let us arrange the outcomes according to the 'Mean AUC' metric, which represents the model's capacity to differentiate between positive and negative categories. Table 1 presents a comparative analysis of different classification techniques.

TABLE I. COMPARISON OF BASELINE CLASSIFICATION ALGORITHMS - 1ST ITERATION

Algorithm	*ROC AUC Mean*	*ROC AUC STD*	*Accuracy Mean*	*Accuracy STD*
Logistic Regression	84.12	1.65	74.6	1.26
SVC	83.64	1.68	79.98	1.08
Gaussian NB	81.82	1.79	68.99	1.46
Random Forest	81.72	2.02	78.47	1.57
Kernel SVM	79.66	2.12	79.85	1.08
KNN	77.04	2.38	75.77	1.09
Decision Tree Classifier	65.44	1.67	72.75	1.45

As for second iteration of comparison of base line classification of algorithms as show in table 2. In the subsequent phase of evaluating baseline classification methods, the optimized parameters for KNN and Random Forest models will be employed. Additionally, it is understood that false negatives incur higher costs compared to false positives in a churn scenario. Therefore, we will utilize precision, recall, and F2 scores as the optimal metrics for selecting the model.

TABLE II. COMPARISON OF BASELINE CLASSIFICATION ALGORITHMS – 2ND ITERATION

Model	*Accuracy*	*Precision*	*Recall*	*F1 Score*	*F2 Score*
Logistic Regression	0.803407	0.652038	0.55615	0.600289	0.573003
SVM (Linear)	0.803407	0.650155	0.561497	0.602582	0.57724
Kernel SVM	0.791341	0.637931	0.494652	0.557229	0.517917
Random Forest	0.779276	0.6171	0.44385	0.51633	0.470255
K-Nearest Neighbors	0.76863	0.570175	0.52139	0.544693	0.530468
Decision Tree	0.739532	0.508997	0.529412	0.519004	0.525199
Naive Byes	0.703336	0.467359	0.842246	0.601145	0.725806

In the subsequent phase of evaluating baseline classification methods, the optimized parameters for KNN and Random Forest models will be employed. Additionally, it is understood that false negatives incur higher costs compared to false positives in a churn scenario. Therefore, we will utilize precision, recall, and F2 scores as the optimal metrics for selecting the model.

A. Future Predictions

Unpredictability and risk are inherent characteristics that are closely associated with predictive algorithms. Hence, in practical scenarios, it is advisable to construct a propensity score in addition to an absolute projected outcome. Rather of solely obtaining a binary predicted goal outcome (0 or 1), it is suggested that each 'Customer ID' be assigned an additional propensity score, which would indicate the percentage of probability associated with initiating the target action. Using propensity score we divide the customers into two categories, High-risk for churning category and low-risk for churning category as shown in table 3 and table 4.

TABLE III. CATEGORY – HIGHEST RISK FOR CHURN

Index	*customerID*	*Churn*	*predictions*	*propensity_to_churn(%)*	*Ranking*
5532	9300-AGZNL I	1	1	85.99	1
5173	5567-WSELE I	1	1	84.87	1
7010	1415-YFWLT I	1	1	83.25	1
6507	2446-BEGGB I	1	1	83.22	1
5985	5277-ZLOOR I	1	1	83.08	1
5636	8884-ADFVN I	1	1	82.44	1
6667	6630-UJZMY I	1	1	82.38	1
6501	2865-TCHJW I	1	1	82.22	1
2367	4750-ZRXIU I	1	1	82.2	1
1018	2660-EMUB I	1	1	81.8	1

TABLE IV. CATEGORY – LOWEST RISK FOR CHURN

Index	*customerID*	*Churn*	*predictions*	*propensity_to_churn(%)*	*Ranking*
5532	4957 - SREEC	0	0	0.43	10
5173	1403 - GYAFC	0	0	0.44	10
7010	1052 - QJIBV	0	0	0.46	10
6507	8590 - YFFQO	0	0	0.47	10
5985	5787 - KXGIY	0	0	0.48	10
5636	0464 - WJTKO	0	0	0.49	10
6667	7876 - DNYAP	0	0	0.5	10
6501	3642 - BYHDO	0	0	0.5	10
2367	8229 - BUJHX	0	0	0.51	10

1018	3678-MNGZX	0	0	0.52	10

The ranks are divided into 10 quantiles, effectively dividing the entire dataset into 10 groups of equal size. Then labels are assigned to these quantiles ranging from 10 to 1, where 10 would be assigned to the highest propensity scores and 1 to the lowest. The ultimate goal of employing a propensity score is to mitigate selection bias by establishing equivalence between groups determined by these factors. By employing this approach, the validity of comparisons between the treatment and control groups in terms of outcomes is enhanced, thereby closely resembling the observations that would be made in a randomized controlled trial.

V. CONCLUSION

In the current highly competitive telecommunications industry, churn prediction is a crucial concern for customer relationship management (CRM). Its purpose is to retain valuable customers by identifying groups of customers with similar characteristics and offering them competitive offers and services. Classification accuracy is a quantitative measure used to assess the efficiency of classification methods. This measure is often used for assessing basic classification methods. However, when there are significant differences in social classes, this statistic becomes less useful. The algorithms and their corresponding accuracy are compared in Table 1. The first round of classification showed that the LRM is superior than the other five models.

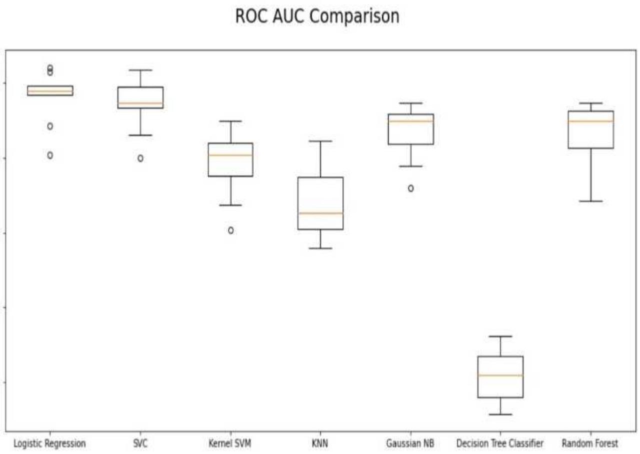

Figure 5. ROC AUC Comparison between algorithms

Upon examining graphical representation of the Accuracy ratings in Figure 5, it becomes evident that logistic regression outperforms the other approaches in terms of accuracy. Overall, the models perform effectively, and logistic regression demonstrated the highest utility in this particular case. As a result, efforts to develop this model have been concentrated, and the accuracy has increased. In this work, a customer churn-over model for data analytics is presented and then assessed using standard evaluation metrics. The results obtained show that our suggested churn model performs better thanks to the application of machine learning methods. With an ROC AUC mean of 88%, Logistic regression produced a better result. After analyzing the dataset, we determined the primary factors that contribute to customer churn. We next conducted cluster profiling based on the level of risk associated with each element. Ultimately, we furnished decision-makers of telecom corporations with explicit instructions on customer retention. In the future, the study might be expanded to investigate the evolving behavioral patterns of churn consumers with the application of Artificial Intelligence approaches for predictive and trend analysis. So, we conclude that we made use of a customer churn dataset from Kaggle to build a machine learning classifier that predicts the propensity of any customer to churn in months to come with a reasonable accuracy score of 80% to 84%.

References

[1] Nyoman Mahayasa Adiputra, Paweena Wanchai, "Customer Churn Prediction Using Weight Average Ensemble Machine Learning Model," 2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE).

[2] Diaa Azzam, Manar Hamed, Nora Kasiem, Yomna Eid, Walaa Medhat," Customer Churn Prediction Using Apriori Algorithm and Ensemble Learning," 2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES).

[3] Yunjie Lin, Mu Shengdong, Gu jijian, Nadia Nedjah, "Intelligent prediction using customer churn with a Fused attentional Deep learning model," 2022 Multidisciplinary Digital Publishing Institute(MDPI),China.

[4] Parth Pulkundwar, Krishna Rudani, Omkar Rane, Chintan Shah, Shyamal Virnodkar, "A Comparison of Machine Learning Algorithms for Customer Churn Prediction" 2023 6th International Conference on Advances in Science and Technology (ICAST).

[5] Eunjo Lee, Boram Kim, Sungwook, Kang, Byungsoo Kang, Yoonjae Jang, and Huy Kang Kim,"Profit Optimizing Churn Prediction For Long Term Loyal Customers in Online Games," 2020 IEEE Transactions on Games.

[6] Xin Hu, Yanfei Yang, Lanhua Chen , Siru 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).

[7] Harish A S, Malathy C, "Evaluative study of cluster based customer churn prediction against conventional RFM based churn model", 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT).

[8] Kiran Deep Singh, Prabh Deep Singh, Ankit Bansal, Gaganpreet Kaur, Vikas Khullar, Vikas Tripathi, "Exploratory Data Analysis and Customer Churn Prediction for the Telecommunication Industry," 2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS).

[9] Mustafa Büyükkeçeci, Mehmet Cudi Okur, "A Data Mining Application in Customer Churn Prediction," 2022 IEEE Conference.[10] S. Venkatesh, M. Jeyakarthic, "An Optimal Genetic Algorithm with Support Vector Machine for Cloud Based Customer Churn Prediction," 2020 IEEE Conference.

[11] Aishwarya H M, Bindhiya T, S Tanisha, Soundarya B, C Christlin Shanuja, "Customer Churn Prediction Using Synthetic Minority Oversampling Technique," 2023 4th International Conference on Communication, Computing and Industry 6.0 (C216).

[12] Abhishek Gaur, Ratnesh Dubey, "Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT).

[13] Mohamed Galal, Sherine Rady, Mostafa Aref, "Enhancing Customer Churn Prediction in Digital Banking using Ensemble Modeling," 2022 4th Novel Intelligent and Leading Emerging Sciences Conference (NILES).

[14] Fulu Chen, Xiaowei Wei, Saisai Yu, Pengfei Ma, Shuqing He, "Customer Churn Prediction based on Stacking Model," 2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL).

[15] 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, 2021.

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

[17] Priya Gopal, Nazri Bin MohdNawi, "A Survey on Customer Churn Prediction using Machine Learning and data mining Techniques in E-commerce," 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE).

[18] 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) Nepal.

[19] 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.

[20] Yuwadi Thein Naing, Mafas Raheem, Nowshath K Batcha," Feature Selection for Customer Churn Prediction: A Review on the Methods & Techniques applied in the Telecom Industry," 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE).