Customer Churn Prediction In Subscription Services

Pooja Verma
AIT-CSE
Student
Chandigarh University
India
vermapooja26200@gmail.com

Srishti Raj
AIT-CSE
Student
Chandigarh University
India
rajshrishti20@gmail.com

Harsh Kumar Deo
AIT-CSE
Student
Chandigarh University
India
harshdeo79@gmail.com

Pulkit Dwivedi
Apex Institute of Technology(CSE)
Assistance Professor
Chandigarh University
India
pulkitdwivedi90@gmail.com

***Abstract*—Customer churn analysis involves predicting customers likely to cancel their subscription services, particularly in sectors like telecommunications, finance, and insurance, and devising strategies to prevent this churn, to identify potential churners in the telecommunications industry. Machine learning techniques including Logistic Regression, K-Nearest Neighbor, Decision Trees, Random Forest, Support Vector Machines, AdaBoost, and voting classifier were applied to the datasets. Results demonstrated that the Random Forest method emerged as the most effective approach for customer churn analysis across both datasets. Once data pre-processing is complete, feature scaling is implemented to ensure equitable contribution from all features during model training. This step mitigates the risk of features with larger scales dominating the modeling process.**

***Index Terms*:Machine Learning,Logistic Regression,Support Vector Classifier,Decision Tree,KNN Classifier,Random Forest,AdaBoost Classifier,Voting Classifier**

I. Introduction

The rapid development and digitalization of the world have revolutionized business practices, compelling companies worldwide to adapt swiftly. Among the outcomes of this digital transformation, subscription-based services have emerged as a prominent model, presenting both opportunities and challenges that demand contemporary solutions. Digitalization has not only revolutionized business operations but has also inundated consumers with a plethora of subscription-based offerings,While this abundance of choices presents opportunities, it also poses challenges for companies, particularly in retaining customers amidst intense competition. Internally, digitalization within companies can yield various benefits, including reduced labor costs, heightened efficiency, and enhanced operational oversight. These advantages are crucial for maintaining competitiveness and gaining an edge in the market. To mitigate customer churn, telecom companies must accurately forecast which customers are likely to discontinue their services. This prediction is pivotal for preemptive retention strategies, ensuring the company can proactively address the needs of at-risk customers. By identifying those at high risk of churn, telecom companies can implement tailored interventions to enhance customer satisfaction and loyalty, ultimately reducing churn rates and preserving revenue streams. Through advanced data analytics and predictive modeling techniques, telecom companies can anticipate churn indicators, such as changes in usage patterns, customer complaints, or shifts in engagement levels. Armed with this predictive insight, companies can deploy targeted retention efforts, such as personalized offers, loyalty rewards, or proactive customer support, to incentivize customers to stay. By prioritizing churn prediction and implementing effective retention strategies, telecom companies can safeguard their customer base, sustain profitability, and maintain a competitive edge in the market.Telecom companies face the ongoing challenge of reducing customer churn, which necessitates the accurate prediction of customers at high risk of churn. Detecting early signs of potential churn requires a comprehensive understanding of customers and their interactions across various channels. These channels encompass physical interactions like store visits and product purchases, as well as digital interactions such as customer service calls, online transactions, and social media engagements.By developing a holistic view of customer behavior and engagement, telecom companies can identify patterns and signals indicative of potential churn. Armed with this insight, they can proactively intervene to retain at-risk customers and prevent churn.Effectively addressing churn is critical for telecom companies not only to maintain their market position but also to foster growth and sustainability. With a larger customer base, telecom companies benefit from economies of scale, reducing the cost of customer acquisition and enhancing overall profitability.In this research paper we focused on variety of machine learning algorithms to enhance predictive accuracy and identify customers at risk of discontinuing their subscriptions. These algorithms include Logistic Regression, Support Vector Classifier, Decision Tree Classifier, KNN Classifier, Random Forest, AdaBoost Classifier, Gradient Boosting Classifier, and Voting Classifier.By considering factors such as subscription contract details, usage patterns, and customer demographics, we aimed to develop robust predictive models capable of accurately forecasting customer churn. Through comprehensive analysis and model evaluation, our project aimed to provide telecom companies with actionable insights to implement targeted retention strategies and mitigate customer churn effectively.Deploying the trained churn prediction model into production systems to generate real-time predictions.

II. Related Work

Customer churn prediction is a crucial task for subscription-based businesses as it helps in retaining existing customers and maximizing revenue. In this paper, we present a comprehensive review of related work in the domain of customer churn prediction in subscription services. We discuss various approaches, methodologies, and techniques employed by researchers to address this work. D. Manzano. [1] introduce the concept of the architecture of a churn prediction system that utilizes stream mining techniques. Churn prediction refers to the task of identifying customers who are likely to stop using a service or product, commonly observed in subscription-based businesses. Stream mining deals with analyzing data streams in real-time to extract patterns and make predictions. S. Babu [2] introduces a survey focused on understanding the factors influencing churn in the telecommunications industry and how data mining techniques can be applied to analyze and predict churn behavior. A. Idris [3] introduce performance of different tree-based ensemble classifiers in conjunction with three distinct feature selection methods: maximum relevancy and minimum redundancy (mRMR) and F-score based selection schemes. Our focus is on addressing the challenging problem of churn prediction in the telecommunications industry.A.Amin [5] introduce big data analysis techniques to delve into historical churn customer data, aiming to construct a robust churn prediction model. By scrutinizing user characteristics, the study identifies customers at higher risk of churning beforehand. Subsequently, it devises targeted strategies and implements a series of retention activities tailored to retrieve these customers. L. Zhao [6]introduce A new algorithm termed KLMM (K-local maximum margin) has been introduced for feature extraction. This method delves into diversification subspace partition rules, thereby constructing a corresponding potential field structure. By scrutinizing the data source in terms of scalability dimensions, it uncovers the inherent connection between data attributes and classification outcomes. The derived features exhibit a capability to diminish the dimensionality of churn prediction in telecom data, offering potential advancements in this domain.B. Shah [7] introduce capitalizes on the notion of data certainty to refine churn prediction models. Data certainty denotes the reliability and confidence level attributed to the data at hand. Through integrating metrics of data certainty into their predictive frameworks, the authors strive to heighten the precision of identifying customers prone to churn.S. A.Qureshi [9] introduce a churn prediction model for businesses aiming to identify customers who are likely to leave and take proactive steps to retain them. However, one common challenge faced in building such models is class imbalance, where the number of churners is significantly lower than non-churners. To address this issue, various re-sampling methods can be employed. These methods involve techniques like over-sampling the minority class, under-sampling the majority class, or using more advanced techniques like SMOTE (Synthetic Minority Over-sampling Technique). By employing these methods, the model can achieve a better balance between the classes, resulting in more accurate predictions and effective retention strategies, ultimately leading to improved customer retention rates and business success.V. Lazarov [10] introduce machine learning, data mining, and hybrid approaches. These methods play a pivotal role in business decision-making and Customer Relationship Management (CRM) by enabling the identification, anticipation, and retention of churning customers. Among these techniques, decision trees stand out as a widely recognized tool for predicting issues associated with client turnover. Joao B. [14] This study introduces an innovative framework for Customer Churn Prediction (CCP) within the banking sector, particularly addressing the challenge of rare churn events that persist over time. The rarity of these events often undermines the effectiveness of traditional techniques designed for binary classification. Our objectives are outlined as follows: to present and validate a data preprocessing phase that integrates various approaches, including Feature Engineering (FE) tailored to the retail banking context, Inverse Density Tree (IDT) oversampling (IDT-over), and IDT undersampling (IDT-under).Nikita Khandelal [15] introduce Various machine learning classification models, including Random Forest, Logistic Regression, K Nearest Neighbor (KNN), AdaBoost, Decision Tree, Support Vector Machines (SVM), and Artificial Neural Networks (ANN), have been utilized for predicting customer churn. These models analyze historical customer data encompassing demographics, transaction history, customer interactions, and usage trends to forecast future churn. The selection of the most suitable model should be guided by the unique features of the industry under consideration. Factors such as data availability, client behavior, and industry-specific nuances necessitate tailored strategies for businesses operating in diverse sectors like healthcare, banking, online retail, and telecommunications.A.D.Caigny [16] introduce Decision trees and logistic regression are widely used algorithms in customer churn prediction due to their strong predictive performance and interpretability. However, decision trees may struggle with linear relationships between variables, while logistic regression may face challenges with interaction effects among variables. To address these limitations, a novel hybrid algorithm called the logit leaf model (LLM) is proposed.The LLM aims to enhance classification accuracy while retaining model comprehensibility. It operates in two stages: segmentation and prediction.This hybrid approach is compared against traditional methods such as decision trees, logistic regression, random forests, and logistic model trees in terms of predictive performance and comprehensibility. By leveraging segmentation and individualized modeling, the LLM offers a promising solution to improve churn prediction accuracy while maintaining interpretability.S. Sivakumar [17] introduce the fresh perspective on churn predictors by integrating them with an organizational competitiveness strategy. Through factor analysis, the model establishes connections between key churn predictors and the overarching competitive strategy of the organization. This innovative approach sheds light on the intricate relationship between customer churn and strategic decision-making, providing valuable insights for businesses aiming to enhance their

competitive edge.Makhtar.M [20]introduce a novel classification model grounded in Rough Set Theory for categorizing customer churn. The findings of the research demonstrate that the proposed Rough Set classification model surpasses existing models, resulting in substantial improvements in accuracy. This study thus presents a promising advancement in the realm of churn prediction, offering a more effective approach for businesses to identify and manage customer churn.Amin A. [21] introduce six prominent sampling techniques and conduct a comparative analysis of their performances. These techniques include the mega-trend diffusion function (MTDF), synthetic minority oversampling technique, adaptive synthetic sampling approach, couples top-N reverse k-nearest neighbor, majority weighted minority oversampling technique, and immune centroids oversampling technique. Additionally, we evaluate four rules-generation algorithms: the learning from example module, version 2 (LEM2), covering, exhaustive, and genetic algorithms, utilizing publicly available datasets. Our empirical findings indicate that MTDF and rules-generation based on genetic algorithms exhibit superior predictive performance compared to the other evaluated oversampling methods and rule-generation algorithms.P.T.Noi [26] introduce the performance comparison of Random Forest (RF), k-Nearest Neighbor (kNN), and Support Vector Machine (SVM) classifiers for land use/cover classification using Sentinel-2 image data. The research focuses on a 30 × 30 km2 area in the Red River Delta of Vietnam, covering six land use/cover types. Fourteen different training sample sizes, ranging from 50 to over 1250 pixels per class, were utilized, including both balanced and imbalanced datasets. Across all classification results, high overall accuracy (OA) was observed, ranging from 90% to 95%.Among the three classifiers and 14 sub-datasets, SVM exhibited the highest OA with the least sensitivity to training sample sizes, followed by RF and kNN. Regardless of sample size, all three classifiers achieved similar high OA values (over 93.85%) when the training sample size exceeded 750 pixels per class or approximately 0.25% of the total study area. Notably, this high accuracy was consistently attained across both imbalanced and balanced datasets. L. Almuqren [30]introduce a novel method leveraging social media mining to forecast customer churn within the telecommunications sector. Notably, it marks the pioneering use of Arabic Twitter mining for predicting churn specifically in Saudi Telecom companies. The efficacy of this newly proposed approach was validated through diverse standard metrics and a thorough comparison with ground-truth actual outcomes provided by a telecom company.

III. Proposed Methodology

Customer churn, the phenomenon of customers discontinuing their subscription, poses a significant challenge for subscription-based businesses. By employing machine learning algorithms, we aim to develop predictive models that can identify customers at risk of churn, enabling proactive retention strategies to be implemented.

In the problem statement of Subscription Services ,telecommunication companies face a significant challenge in retaining their customer base due to voluntary churn, where customers switch to competitors' services. This poses a considerable cost, as acquiring new customers requires substantial marketing and sales efforts compared to retaining existing ones. Thus, understanding and mitigating customer churn is crucial for maintaining profitability and market competitiveness.The challenge lies in accurately predicting which subscribers are likely to discontinue their services and implementing effective retention strategies to mitigate churn rates. Leveraging machine learning techniques, this project aims to develop a predictive model that can identify potential churners among telecommunication subscribers. By analyzing diverse datasets encompassing subscriber demographics, usage patterns, service subscriptions, and customer interactions, the goal is to build a robust churn prediction system. Key objectives include addressing data heterogeneity, feature selection, handling imbalanced datasets, and ensuring scalability and interpretability of the machine learning solution within the telecommunication infrastructure. This paper outlines the proposed methodology, which involves data feature selection as a support Vector Classifier. Various machine learning algorithms such as logistic regression, decision trees, random forests,(KNN)K Nearest Neighbor Classifier, AdaBoost Classifier, Gradient Boosting Classifier, and Voting Classifier. We will explore and compare their effectiveness in churn prediction. The proposed methodology seeks to contribute to the development of robust churn prediction systems that can assist subscription services in reducing customer churn rates and improving overall customer retention strategies.

A. *Feature Distribution*

*1) **Numerical Feature Distribution**:* By analyzing these distributions, it aids in evaluating model performance, addressing class imbalance, and enhancing model interpretability. In essence, understanding numerical feature distributions is critical for building accurate, robust, and actionable churn prediction models.

```
df[numerical_features].describe()
```

	tenure	MonthlyCharges	TotalCharges
count	7039.000000	7039.000000	7028.000000
mean	32.376332	64.762963	2284.005827
std	24.561896	30.087756	2267.193201
min	0.000000	18.250000	18.800000
25%	9.000000	35.500000	401.250000
50%	29.000000	70.350000	1397.950000
75%	55.000000	89.850000	3796.912500
max	72.000000	118.750000	8684.800000

Fig. 1: Numerical Decsription

In the figure(1) tenure, monthly charges and total charges are used to calculate mean and standard deviation,median and quartile then the total charges are computed.

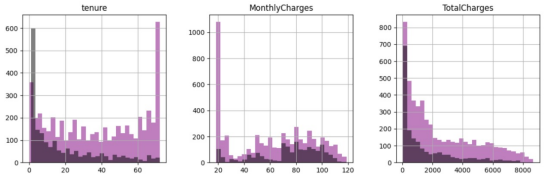

Fig. 2: comparison of active and churn customers

In Figure (2), tenure, monthly charges, churn, and no churn are depicted.In this comparison, churn predictions are depicted in black, while non-churn predictions are represented in blue colour.

*2) **Categorical Feature Distribution**:* Analyzing the distribution of categorical variables helps identify prevalent categories, detect rare or infrequent ones that may require special handling, and assess the balance within each category.

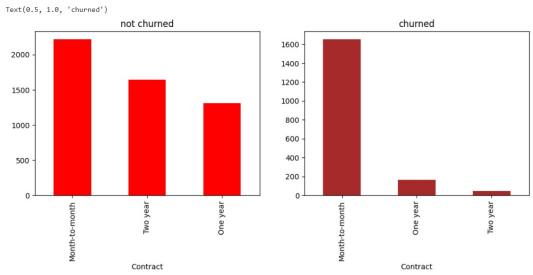

Fig. 3: categorical Distribution

In the figure(3) categorical distribution categorize variables such as gender, partner status, dependents, and senior citizen status. To assess churn or non-churn based on contract duration (1or2 years), we generate a bar graph. The brown bars represent churn predictions, while the red bars depict non-churn predictions.

*3) **Target Variable Distribution**:* These variables represent the outcome of interest—whether a customer churns or not. Target variables drive model training by providing a clear objective for the algorithm to optimize predictive performance

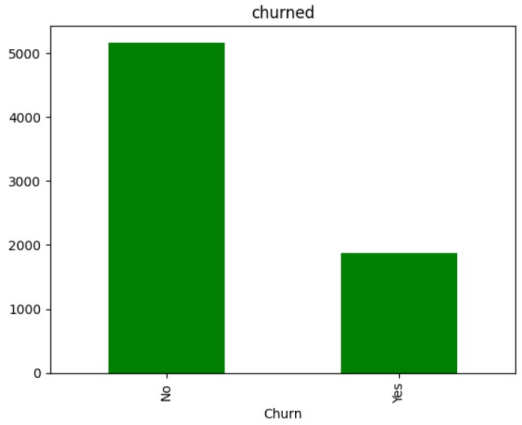

Fig. 4: Target Variable Distribution

In the figure(4) target variable represents the output or result, indicating whether a customer has churned. In visualizations, churned instances are typically depicted with green coloration.

B. Data Transformation

1) **One Hot Coding:**In churn prediction, customer attributes such as subscription plan, geographical location, or product preferences are often categorical. One-hot coding converts these categorical variables into binary vectors where each category is represented by a binary variable.

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges	gender_Female	gender_Male	Partner_No	Partner_Yes
0	0	1	29.85	29.85	1	0	0	1
1	0	34	56.95	1889.50	0	1	1	0
2	0	2	53.85	108.15	0	1	1	0
3	0	45	42.30	1840.75	0	1	1	0
4	0	2	70.70	151.65	1	0	1	0

5 rows × 47 columns

Fig. 5: One Hot Coding

In figure(5), we create multiple columns for categorical variables such as gender, partner status, contract type, internet service, online backup, etc. In this encoding scheme, we represent categorical values as binary vectors where 'Yes' is encoded as 1 and 'No' is encoded as 0.

2) **Feature Selection:** By selecting the most informative features, feature selection reduces the dimensionality of the dataset, mitigating the risk of overfitting and improving model interpretability and efficiency.
In the feature selection, imports a function is commonly used for splitting data into training and testing sets for machine learning tasks. Y = df1[Churn Yes] , It retrieves the column labeled Churn Yes from this data frame and assigns it to the variable Y. X is the data in scaled features and Y is the labels or target variables in df1(churn)yes.In test size(0.3) this argument specifies the proportion of the data to be included in the test set.30% of the data will be allocated to the test set and the remaining 70% will be used for training.random state= 46 sets the random seed for splitting the data.

C. Algorithm Analysis

This analysis typically involves assessing factors such as computational complexity, training and prediction times, scalability, model complexity, and generalization performance.

1) **Logistic Regression:**Its primary role involves estimating the likelihood of churn by fitting a logistic function to the data, which enables binary classification (churn/non-churn). Logistic regression aids in understanding the relationship between input features and the probability of churn, thereby identifying key factors influencing customer attrition.
In Figure (6), the representation of logistic regression is depicted, illustrating how it is defined.

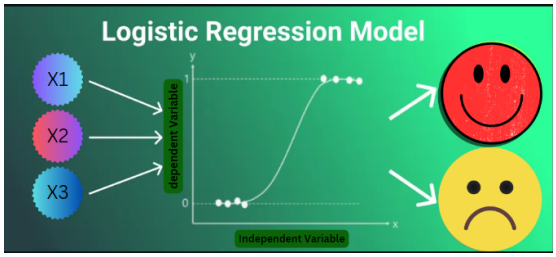

Fig. 6: Logistic Regression

2) **Support Vector Machine:**SVM aim to find the optimal hyperplane that maximizes the margin between different classes, facilitating accurate classification. SVM can handle both linear and nonlinear relationships through the use of appropriate kernel functions.
3) **K-Nearest Neighbors:** KNN classifies customers as churners or non-churners based on the majority class among their k-nearest neighbors in feature space.When a prediction is required for a new data point, KNN calculates the distance between the input data point and all other points in the training dataset.
4) **Decision Tree:** Decision trees recursively split the data based on feature values, creating a hierarchical structure that represents decision rules. This enables the model to capture nonlinear relationships and interactions between predictors and churn outcomes. Decision trees offer interpretability by visualizing decision paths and feature importance, making it easier for stakeholders to understand the factors driving churn.
5) **Random Forest:** This ensemble approach mitigates overfitting and captures complex relationships between predictors and churn outcomes. Random Forest offers improved generalization performance compared to individual decision trees, making it well-suited for handling large and diverse datasets common in churn prediction scenarios
6) **AdaBoost Classifier:** This adaptive boosting technique is particularly useful in customer churn prediction, where identifying at-risk customers accurately is paramount. Ada Boost ability to prioritize challenging cases and adapt to complex relationships between predictors and churn outcomes makes it a valuable tool for building highly accurate churn prediction models.
7) **Gradient Boosting :** By continuously refining the model based on the gradient of the loss function, gradient boosting excels at identifying subtle patterns and interactions in the data, making it particularly effective for churn prediction tasks where accurate identification of at-risk customers is paramount.
8) **Voting Classifier:** In churn prediction tasks, where the outcome is binary (churn or non-churn), the voting classifier aggregates the predictions from different models and selects the majority vote as the final prediction. This ensures that the final prediction reflects a consensus among the diverse classifiers, enhancing the reliability and accuracy of churn predictions.

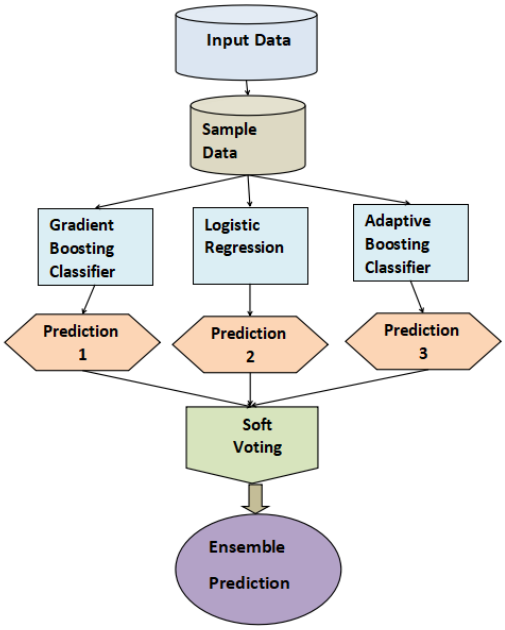

Fig. 7: Soft Voting Classifier

IV. Experimental Result

In machine learning, evaluating the classifier's performance is crucial, particularly in selecting the optimal algorithm for a given problem. Previous research on churn prediction predominantly employs metrics such as accuracy, precision, recall, and F-measure, all derived from the confusion matrix. Our study follows suit, assessing algorithm performance and efficiency using these metrics.However, when dealing with imbalanced data, accuracy may not provide an accurate reflection of the algorithm's performance. Therefore, we prioritize precision, recall, and F1-score, specifically focusing on the target class, which in our case is Churn. Among these metrics, we emphasize F1-score as it strikes a balance between precision and recall.

In our application, precision denotes the rate of correctly classified churn instances, while recall measures the model's ability to predict actual churners. Given our focus on churn prediction, recall holds greater importance than precision. Evaluation is conducted on unseen instances from the test set, ensuring that the algorithm's performance is assessed on data it has not been trained on.

A. *Performance Based Evaluation Matrix*

Confusion Matrix: In the figure(8),confusion matrix comprises True Positives (TP) and True Negatives (TN), which denote correctly classified test instances, and False Negatives (FN) and False Positives (FP), representing incorrectly classified instances.
True Positives (TP) predicts customers stay subscribed

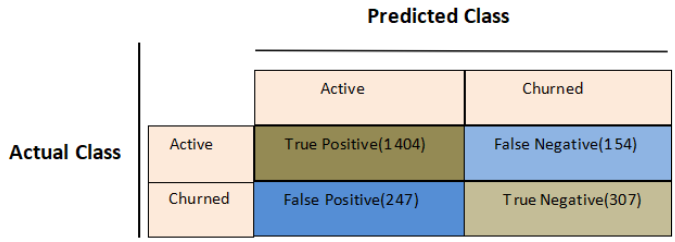

Fig. 8: confusion matrix

(correct),and customers actually stay subscribed.True Negatives (TN) predict customer does not churn (stay subscribed), unclassified outcome. False Positive (FP) predicts customer churns (incorrect), and the customer actually stays subscribed. False Negatives (FN) predicts customer stays subscribed (incorrect), customer actually churns.

Each metric provides insight into the model's performance in predicting customer behavior, crucial for refining strategies to retain valuable customers.

Recall: It specifically gauges the ratio of true positive predictions, which are the correctly identified positive instances, to all the actual positive instances in the dataset.It is the proportion of Active (or Churn) customer for the correctly identified.Calculating with the help of equation (1)

$$\text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

Precision:- It quantifies the proportion of true positive predictions (correctly identified positive instances) out of all instances predicted as positive by the model.It is the proportion of the predicted cases,which is shown in the equation (2),

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Accuracy: It is a measure that displays the percentage of cases that are correctly classified overall. Accuracy is described in equation (3),

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

F-Measure = In a prediction model, the F-measure, also known as the F1 score, predicted positive instances (precision) and the model's ability to capture all positive instances (recall).The F-measure is the harmonic mean of precision and recall and is calculated using the eqation(4),

$$F = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{4}$$

In table (1), we calculate these values by using the confusion matrix parameters precision, recall, F1-score and accuracy.

TABLE I: Performance Metrics of Various Algorithms

Algorithms Used	Precision	Recall	F1-Score	Accuracy
Decision Tree	0.82	0.8	0.81	0.72
KNN	0.84	0.88	0.86	0.79
Ada Boost	0.85	0.9	0.87	0.80
Gradient Boost	0.84	0.91	0.87	0.80
Logistic Regression	0.85	0.9	0.88	0.81
Random Forest	0.83	0.92	0.87	0.80
SVM	0.83	0.91	0.87	0.80
Voting Classifier	0.85	0.91	0.88	0.81

B. Algorithm Based Evaluation

In algorithm evaluation, We consider a variety of algorithms such as logistic regression, KNN, SVM, Voting Classifier, Gradient Boost, AdaBoost, Decision Tree, and Random Forest. We convert categorical values like Gender, Monthly Charges, and Payment Method into numerical and binary formats to assess churn and non-churn conditions in the confusion matrix.

In the figure(9), logistics regression model positive and negative rates refer to the proportion of correctly predicted positive and negative instances, respectively. With an accuracy of 0.81 this means that approximately 81.01% of the instances in the dataset are classified correctly by the logistics regression model.

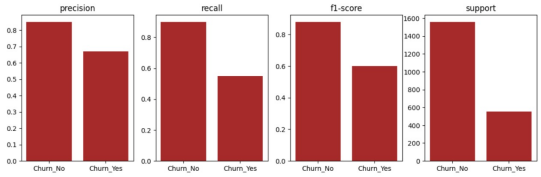

Fig. 9: Classification Report of Logistic Regression

In the support vector machine, the positive and negative rates refer to the proportion of correctly predicted positive and negative instances, respectively. With an accurracy 80.20% in the dataset are classified correctly by the support vector classifier.

In the AdaBoost Classifier, the positive and negative rates refer to the proportion of correctly predicted positive and negative instances, respectively. With an accuracy of 80.70% in the dataset are classified correctly by the AdaBoost Classifier.

In the Gradient Boosting Classifier, the positive and negative rates refer to the proportion of correctly predicted positive and negative instances, respectively. With an accuracy of 80.70% in the datasets are classified correctly by the gradient boosting Classifier.

In KNN, n neighbors is 30,the positive and negative rates refer to the proportion of correctly predicted positive and negative instances, respectively. With an accuracy 78.64% in

the dataset are classified correctly by the KNN,which is shows in fig(10)

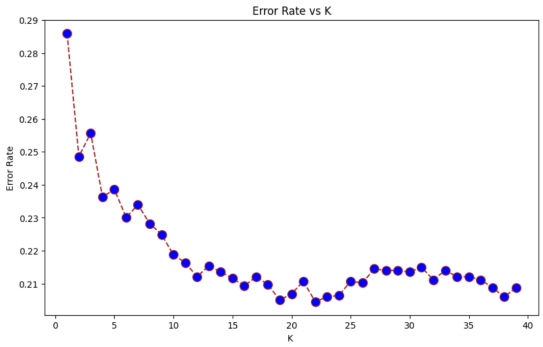

Fig. 10: Error Rate in KNN

In the random forest model, the positive and negative rates refer to the proportion of correctly predicted positive and negative instances, respectively. With an accuracy 79.78% in the dataset are classified correctly by the random forest model. In the model configuration, the parameter values are set as follows n estimators is assigned the value of 500, oob score is set to True, n jobs is set to -1 to utilize all available processors, random state is set to 50 for reproducibility, max features is specified as ”sqrt”, and max leaf nodes is set to 30. These parameter values are chosen to optimize the performance and efficiency of the model.

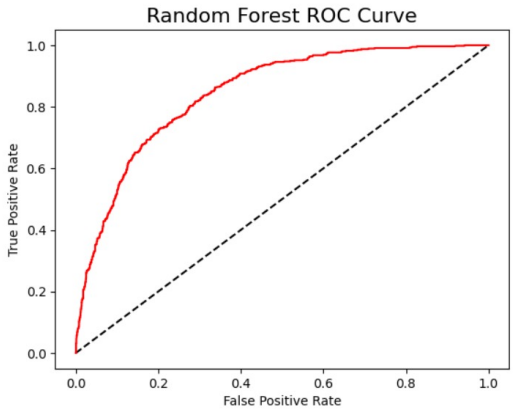

Fig. 11: ROC Curve

In the figure(11) ROC curve, the true positive rate (TPR) is plotted on the y-axis, and the false positive rate (FPR) is plotted on the x-axis. Each point on the curve represents a different threshold setting, and the curve illustrates how the trade-off between true positives and false positives changes as the threshold varies.

In the Decision tree Classifier, the positive and negative rates refer to the proportion of correctly predicted positive and negative instances, respectively. With an accuracy of 72.01% in the dataset is classified correctly by the decision tree classifier.

In the voting classifier,there is three estimators have been used, which are Gradient Boosting Classifier, Logistic Regression, and AdaBoost Classifier. Additionally, a Voting Classifier with soft voting has been employed, as it tends to yield higher accuracy by combining the predictions of multiple base estimators.These scores are calculated using the ”predicted class” and ”actual class” parameters in the confusion matrix.

In the figure (12) confusion matrix, we can observe that there are two rows and two columns named active and churned. This confusion matrix has four boxes. True positive contains the value 1253, True Negative contains the value 272, False Positive contains the value 282 and False Negative contains the value 305. We can use these values to calculate the accuracy of an algorithm.

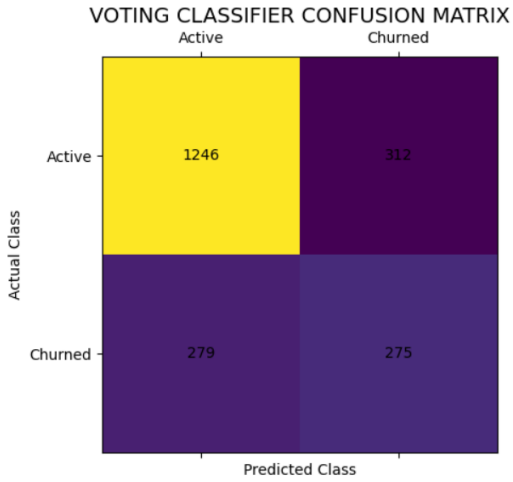

Fig. 12: Confusion Matrix Of Voting Classifier

The accuracy score in the confusion matrix of the Voting Classifier is 81.06% of the instances in the dataset.

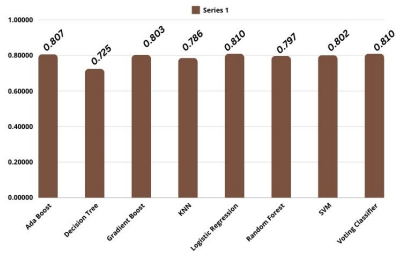

Fig. 13: Accuracy Performance Of Algorithms

The final score of the voting classifier and logistics regression algorithm demonstrating the highest accuracy, is depicted in Figure (13).

V. Conclusion

In conclusion, our analysis of customer churn prediction in subscription services reveals valuable insights when considering recall, precision, confusion matrix, F1-score. We employed various machine learning algorithms including Random Forest, Decision Tree, Support Vector Machine, Logistic Regression,AdaBoost Classifier, Voting Classifier, and Gradient Classifier. Additionally, we utilized techniques such as one-hot encoding and dummy variables for handling categorical data.

Recall and precision metrics help assess the effectiveness of our models in correctly identifying churned customers while minimizing false positives. A higher recall indicates the ability to capture more true positives, while higher precision suggests fewer false positives.The confusion matrix provides a comprehensive summary of the model's performance, depicting true positives, true negatives, false positives, and false negatives. This aids in understanding the strengths and weaknesses of the classification models.The F1-Score, which combines both recall and precision into a single metric, offers a balanced assessment of the model's overall performance. It is particularly useful when there is an imbalance between the classes. we efficiently preprocessed the data and prepared it for model training.

References

[1] D. Manzano-Machob, "The architecture of a churn prediction system based on stream mining," in Proc. Artif. Intell. Res. Develop., 16th Int. Conf. Catalan Assoc. Artif. Intell., vol. 256, Oct. 2013, p. 157.

[2] S. Babu, D. N. Ananthanarayanan, and V. Ramesh, "A survey on factors impacting churn in telecommunication using data mininig techniques," Int.J. Eng. Res. Technol., vol. 3, no. 3, pp. 1745–1748, Mar. 2014.

[3] A. Idris and A. Khan, "Customer churn prediction for telecommunication:Employing various various features selection techniques and tree based ensemble classifiers," in Proc. 15th Int. Multitopic Conf., Dec. 2012,pp. 23–27.

[4] A. Amin et al., "Customer churn prediction in the telecommunication sector using a rough set approach," Neurocomputing, vol. 237, pp. 242–254,May 2017.

[5] A. Amin et al., "Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods," Int. J. Inf. Manage., vol. 46, pp. 304–319, Jun. 2019.

[6] L. Zhao, Q. Gao, X. Dong, A. Dong, and X. Dong, "K-local maximum margin feature extraction algorithm for churn prediction in telecom,"Cluster Comput., vol. 20, no. 2, pp. 1401–1409, Jun. 2017.

[7] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar,"Customer churn prediction in telecommunication industry using data certainty," J. Bus. Res., vol. 94, pp. 290–301, Jan. 2019.

[8] H.-S. Kim and C.-H. Yoon, "Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market," Telecommun. Policy, vol. 28, nos. 9–10, pp. 751–765, Nov. 2004.

[9] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman,"Telecommunication subscribers' churn prediction model using machine learning," in Proc. 8th Int. Conf. Digit. Inf. Manage., Sep. 2013, pp. 131–136

[10] V. Lazarov and M. Capota, "Churn prediction," Bus. Anal. Course,TUM Comput. Sci, Technische Univ. München, Tech. Rep., 2007

[11] Y. Huang, B. Huang, and M.-T. Kechadi, "A rule-based method for customer churn prediction in telecommunication services," in Proc. Pacific– Asia Conf. Knowl. Discovery Data Mining. Berlin, Germany: Springer, 2011, pp. 411–422.

[12] V. L. Miguéis, D. van den Poel, A. S. Camanho, and J. F. e Cunha, "Modeling partial customer churn: On the value of first product-category purchase sequences," Expert Syst. Appl., vol. 12, no. 12, pp. 11250–11256, Sep. 2012.

[13] Nasebah Almufadi, Ali Mustafa Qamar, Rehan Ullah Khan and Mohamed Tahar Ben Othman, "Deep Learning-based Churn Prediction of Telecom Subscribers", International Journal of Engineering Research and Technology, vol. 12, no. 12, pp. 2743-2748, 2019, ISSN 0974-3154.

[14] João B. G. Brito, Guilherme B. Bucco, Rodrigo Heldt, João L. Becker, Cleo S. Silveira, Fernando B. Luce, Michel J. Anzanello, "A framework to improve churn prediction performance in retail banking", Financial Innovation, vol.10, no.1, 2024.

[15] Nikita Khandelal, Vikas Sakalle, "Customer Churn Prediction in Telecommunication, Medical Industry Using Machine Learning Classification Models", 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), vol.6, pp.1727-1734, 2023.

[16] A. D. Caigny, K. Coussement, and K. W. D. Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," Eur. J. Oper. Res., vol. 269, no. 2, pp. 760–772,Sep. 2018.

[17] S. Sivakumar,Yu W, Jutla DN.A churn-strategy alignment model for managers in mobile telecom. In: Communication networks and services research conference, vol. 3. 2005. p. 48–53.

[18] Zhan J, Guidibande V, Parsa SPK. Identification of top-k influential communities in big networks. J Big Data. 2016;3(1):16.

[19] He Y, He Z, Zhang D. A study on prediction of customer churn in fixed communication network based on data mining. In: Sixth international conference on fuzzy systems and knowledge discovery, vol. 1. 2009. p. 92–4.

[20] Makhtar M, Nafis S, Mohamed M, Awang M, Rahman M, Deris M. Churn classification model for local telecommunication company based on rough set theory. J Fundam Appl Sci. 2017;9(6):854–68.

[21] Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hawalah A, Hussain A. Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access. 2016;4:7940–57.

[22] Kumar S, Chandrakala D: A Survey on Customer Churn Prediction using Machine Learning Techniques. Int. J. Comput. Appl. 2016; 154(10): 13–16.

[23] Brandusoiu I, Toderean G, Ha B. Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. 2016. p. 97–100

[24] P. Routh, A. Roy and J. Meyer, "Estimating customer churn under competing risks", Journal of the Operational Research Society, vol. 72, pp. 1138-1155, 2021

[25] Huang F, Zhu M, Yuan K, Deng EO. Telco churn prediction with big data. In: ACM SIGMOD international conference on management of data. 2015. p .607–18.

[26] P. T. Noi and M. Kappas, "Comparison of random forest k-nearest neighbor and support vector machine classifiers for land cover classification using sentinel-2 imagery", Sensors (Basel Switzerland), vol. 18, 2018.

[27] R. I. Dzerzhinsky, M. D. Trifonov and E. V. Ledovskaya, "The support vectors and random forest methods analysis in the forecasting customer churn problem in banking services", Lecture Notes in Networks and Systems, 2021.

[28] I. Ahmad, M. Basheri, M. J. Iqbal and A. Rahim, "Performance comparison of support vector machine random forest and extreme learning machine for intrusion detection", IEEE Access, vol. 6, pp. 33789-33795, 2018.

[29] M. Pondel, M. Wuczynski, W. Gryncewicz, L. Lysik, M. Hernes, A. Rot, et al., "Deep learning for customer churn prediction in e-commerce decision support", BIS, 2021.

[30] L. Almuqren, F. Alrayes and A. I. Cristea, "An empirical study on customer churn behaviours prediction using arabic twitter mining approach", Future Internet, vol. 13, pp. 175, 2021.