RESEARCH ARTICLE OPEN ACCESS

Customer Churn prediction in Software as a Service (SaaS) Industry using Machine Learning Algorithms

Mohammad Hassan Ansari[1], Rahul Gupta[2]
[1](Amity School of Engineering & Technology, Amity University, Noida
Email: mohdhassanansari737@gmail.com)
[2](Amity Business School, Amity University, Noida
Email: rgupta10@amity.edu)

--************************----------------------------------

Abstract:

The sole reason for an organization to exist in this era of competition is to remain profitable and competitive within the industry. There is no doubt in the notion that customers directly influence the profitability of a firm, the customer can be categorized as existing and new customers. Although primary motive of a firm is to cater more percentage of target customers, the ability of a firm to retain their existing customers is equally important. Churn rate can be defined as the rate at which the existing customers that were using a service or product had stopped using it. This study is focused on predicting the churn rate of customers using various machine learning algorithms including Voting Classifier, Gradient Boost Classifier, AdaBoost, Logistic Regression and Gaussian Naïve Bayes Algorithm in the Software as a Service (SaaS) industry on the basis of features selection from the collected data. The algorithms will be implemented to identify the best algorithm with the most accuracy as compared to other algorithms. This paper contains one extra algorithm of Voting Classifier which shows highest accuracy with one more evaluation matrix feature of F2-score for every algorithm implemented in the paper. This also implemented two special classifiers for feature prediction giving different results using Gradient Boost and Adaboost classifier.

Keywords **—Churn, Machine Learning, Software as a Service (SaaS), Voting Classifier, Gradient Boost Classifier, AdaBoost, Logistic Regression and Gaussian Naïve Bayes**

--************************----------------------------------

I. INTRODUCTION

In every industry, one of the main sources of income is customers making them the resource with the highest value [1]. Plethora of academic research has been done to identify various ways to target customers and their various segments. For the same purpose, companies have their ultimate goal to cater their customers for which sometimes they succeed but have the almost equal chance of failing. Targeting new customers is a tedious task but making the existing customers retain seems to be no easier making customers churn. Therefore, it is quite important to retain the existing customers in this situation which is the problem faced by all businesses. Churn can be defined as loss of customers or clients which can also be termed as customer attrition. To retain the customers that are using product or a service of a company, churn rate is a matrix which is vital for an organization. Churning clients is vital problem that can lead to revenue loss, waste of expenses and wastage of time spent on customer acquisition. It has emerged as a requirement to identify churn for a company as it can affect the profits if left unchecked [2].

In recent years, Software as a Service (SaaS) is growing with more successful adoptions. It is a model of delivery which caters the clients of a company to a business function remotely [4]. The very nature and functions of services offered through SaaS may not be same. Packages and full-fledge software solutions can be expensive as well as a tedious task to understand its function. Normal services are offered with a specific operation desired, but there is a requirement to combine different services together to achieve a desired operations or function of an organization [5]. The user does not actually buy a license of software. The infrastructure charges, usage rights of software and other services such as hosting services, repair and maintenance are combined to offer an individual monthly basis or according to per-usage. Since SaaS allows us to bring down the Total Cost of Ownership (TCO), and higher Return on Investment (ROI), the services has gained growth in development and include some of the well know areas such as Customer Relationship Management (CRM) services offered by salesforce.com, Human Resource Management (HRM) from employease.com [6, 7].

This paper will make an attempt to bring the emerging software industry that has a business model of Software as a Service (Saas) under the implementation of Machine learning algorithms to predict churn rate for which there is very limited literature is available. This paper will gather, model and produce results for the firm in software industry as against the telecom industry for which plethora of academic research has been done. It will help discover new patterns and trends for the industry. This paper will also help researchers to investigate further in the same area.

II. OBJECTIVES

First Objective: To elaborate how machine learning algorithms can be implemented in the Industry of Software as a Service (SaaS) to determine churn of the customers.

Second Objective:

1. To elaborate the impact of churn of the customers on the Industry of Software as a Service (SaaS).

2. To determine the customer churn model's importance in the Industry of Software as a Service (SaaS).

3. To compare distinct churn prediction Machine Learning Algorithms.

III. LITERATURE REVIEW

B. Prabadevi et al. proposes using machine learning algorithms like stochastic gradient booster, random forest, logistic regression, and k-nearest neighbors to predict customer churn in the telecom industry. The data includes customer information like services used, account details, and demographics for 7044 customers from a fictional company. The algorithms are compared based on accuracy, with stochastic gradient booster achieving the highest at 83.9%, followed by random forest (82.6%), logistic regression (82.9%), and k-NN (78.1%). The paper concludes that organizations can use these predictive models, especially stochastic gradient booster, to identify customers at high risk of churning and take proactive retention measures. Further research is suggested on refining data preprocessing and hyper parameter tuning to improve model performance [8].

M. A. S. Thorat et al. employed machine learning algorithms like random forest, XGBoost, and logistic regression to predict customer churn in the telecom industry. The data includes customer usage, demographics, etc. for prediction modeling. Models are evaluated on accuracy, with random forest achieving 88%. Key factors are monthly service usage, customer tenure, contract type. Deep learning models like convolutional and recurrent neural networks are also suggested for enhanced accuracy. Results demonstrate potential of deep learning in churn prediction to inform customer retention strategies. Further data and tuning could improve model performance. Overall it shows machine and deep learning are effective for churn prediction in telecom industry [9].

S. Gowd et al. examined customer churn in the telecom industry using the Orange telecom dataset from Kaggle. Machine learning algorithms like decision trees, random forest, K-NN, Naive Bayes, and XGBoost are implemented. Models are evaluated on accuracy, precision-recall curves, and F1-scores. Results show XGBoost and random forest perform best with 95.65% and 95.20% accuracy after hyper parameter tuning. Key factors identified include number of customer service calls, international plan use, and charges correlated to call volumes. The paper concludes XGBoost is the best model for predicting churn due to higher accuracy in less execution time versus random forest. Further work could explore more advanced algorithms and ensemble techniques. Overall, it demonstrates machine learning, especially boosted models, can effectively predict churn in the telecom sector. [10].

A. M. Almana carried out survey of common data mining techniques like neural networks, statistical methods, decision trees, and covering algorithms applied to predict and understand customer churn in the telecom sector. Research shows neural networks can effectively model churn but have limitations in interpretability. Statistical techniques like regression and Naive Bayes also demonstrate utility. Decision tree methods like CART and C5.0 frequently outperform other approaches in accuracy. Covering algorithms are promising but less studied for churn modeling. The paper concludes more research is needed on newer algorithms like covering methods for churn analysis to find the most accurate solutions. Overall, it provides a useful overview of using data mining, especially decision trees, to leverage industry data and address the key business issue of reducing customer churn [11].

A.M. Aldalan et al. performed customer churn in the telecommunications industry using machine learning algorithms and investigates the factors contributing to customer churn. The authors apply four different machine learning models, namely Logistic Regression, Random Forest, Decision Tree, and Gradient Boosting, on the Orange dataset from an American telecom company. The dataset contains 22 features and 12,892 rows, with the last column indicating whether the customer has churned or not. The authors begin by performing Exploratory Data Analysis (EDA) to understand the key characteristics of the variables and their relationships with customer churn. They visualize the data and formulate hypotheses about the factors influencing churn, such as the number of voicemail messages, the presence of a voicemail plan, total daytime minutes and charges, and the number of customer service calls. The authors then preprocess the data by applying feature engineering techniques, including label encoding for categorical variables and min-max normalization for continuous variables. They also employ feature selection methods, such as the Pearson Correlation Coefficient (PCC), Chi-Squared test, and T-Test, to identify and remove irrelevant or highly correlated features. To address the class imbalance issue, where only 14.5% of customers have churned, the authors use grid search with hyper parameters and class weighting to prevent over fitting and bias toward the majority class. The authors evaluate the performance of the models using the F1-score and Area under the Receiver Operating Characteristic (ROC-AUC) curve. They conduct two experiments: one without feature selection and another with feature selection. The results show that the Gradient Boosting model with feature selection outperforms the other models, achieving an F1-score of 99% and an AUC of 99%. Random Forest and Decision Tree also perform well, with F1-scores and AUCs above 95%. Logistic Regression yields relatively poor results, with an F1-score of 49% and an AUC of 77%. The authors conclude that feature engineering and selection techniques improve the performance of the models, and ensemble models like Gradient Boosting and Random Forest are robust against high dimensionality and imbalanced datasets. They suggest further exploration of deep learning models and continuous data generation for future improvements [2].

A. A. Khan et al. applied Knowledge discovery in Database (KDD) techniques by choosing an ISP operator operating in Iran. The data was divided into ratio of 7:3 for training and testing of the dataset. Microsoft Business Intelligence Development Studio was used in which for training three models which were used which are Microsoft Decision Trees, Microsoft Neural Network and Microsoft Logistic regression and tested further for accuracy. Decision tree has 87.74% accuracy, Logistic regression has 89.01% accuracy and Neural Network has 89.08% accuracy. It was hard to come on conclusion for usage and billing features while the demographic feature has the least contribution to the prediction of the churn [12].

D. Singh et al. collected data from IBM to perform multiple Machine Learning Algorithms and determined their accuracies respectively for implementation of each. The algorithms used were Gradient Booster Classifier (78.05%), Light Gradient Boosting Machine (77.83%), Random Forest Classifier (77.38%), Ada Boost Classifier (76.88%), Extra Trees Classifier (76.53%), Logistic Regression (74.58%), Linear Discriminant Analysis (74.26%), Naïve Bayes (73%), Decision Tree Classifier (72.54%) and K Neighbors Classifier (68.97%). Although the research paper is not recent as well as accuracy is low but the purpose was that some extra Machine Learning Algorithms were used [13].

A portal was created in which there is a model which is created by adopting Random Forest and has performed and proved that 95% accuracy of the most predictions is based on the F-1 Score. Three algorithms used were Decision Tree, Random Forest and XGBoost. It also has a page for Admin login with username and password to sign in. In the portal details of plans such as Account length, International Plan, number of voice mail messages, total day call, total eve calls are given as input to predict of churn or no churn [13].

L. ÇALLI et al. focused on analyzing customer churn in the software-as-a-service (SaaS) industry using machine learning algorithms. While customer churn research is prevalent for B2C models like telecom and retail, there is limited research on B2B models like SaaS. Previous SaaS studies have used algorithms like logistic regression, random forests, and support vector machines to predict churn, with features related to customer usage and transactions. Key predictors found are number of logins, transactions, users, call quality, etc. Random forest tended to perform best with accuracy around 75-92%. However, some studies had issues with insufficient data. This paper aims to further analyzeSaaS customer churn to address the literature gap [14].

N. Phumchusri et al. focuses on predicting customer churn for a Software-as-a-Service (SaaS) inventory management software company in Thailand, which is facing a high churn rate. The authors aim to develop a customer churn prediction model using machine learning algorithms to identify customers likely to churn and provide insights into factors influencing churn behavior. The paper begins with a literature review, highlighting the importance of customer churn prediction in various industries, particularly the SaaS industry, given its subscription-based business model. The authors note that while customer churn prediction has been extensively studied in industries like telecommunications and banking, there are fewer studies focused on the SaaS industry. The methodology involves collecting data from the case-study company's database, covering customer usage behavior and business metrics from October 2015 to October 2019. The authors define churn as a customer who has been inactive for more than 14 consecutive days, based on the company's marketing team's requirements. After data cleaning and preprocessing, the authors apply feature selection techniques, including Chi-squared score and ANOVA F-values, to identify the most relevant features for predicting churn. The authors then employ four machine learning algorithms: logistic regression, support vector machine (SVM), decision tree, and random forest. These algorithms are chosen based on their successful applications in previous customer churn prediction studies, particularly in the SaaS industry. The performance of the models is evaluated using 10-fold cross-validation and various metrics, including accuracy, precision, recall, and F1-score. The authors prioritize recall and F1-score as the most important metrics, as correctly identifying churn customers is crucial for the case-study company to mitigate the high cost of acquiring new customers. The results show that the random forest model, combined with the Chi-squared feature selection method, outperforms the other algorithms, achieving a recall score of 91.6% and an F1-score of 92.6%. The authors further validate the model's performance on a holdout testing dataset, confirming its effectiveness in predicting churn with high accuracy. Additionally, the authors analyze the feature importance scores from the random forest model, highlighting that business metrics, such as the number of transactions in the current and previous months, customer spending amount, and number of actions per customer, are the most significant factors influencing churn behavior. The paper concludes that the developed random forest model can effectively predict customer churn and provide valuable insights into the factors contributing to churn. The authors suggest that the case-study company can use this model to identify potential churn customers and implement targeted marketing campaigns to retain them, thereby improving customer retention and profitability [19].

A. Kolomiiets et al. The literature review explores the application of modern machine learning models and methods for predicting customer churn, specifically focusing on their relevance to B2B Software as a Service (SaaS) companies in the IT sector. The review emphasizes the effectiveness of the random forest method and also conducts an analysis of the deep neural network method. Initially, the review discusses the optimization of the random forest model, particularly in relation to the number of trees (n_estimators) and the number of features for splitting (max_features). It notes that as the number of trees increases, the quality of the model on the training sample improves, reaching an asymptote on the test sample. Around 40 trees were chosen for the model, and the trees were built at full depth due to the low noise level in the data. Additionally, it observes that increasing max_features leads to longer training times and "more monotonous" trees, with the quality on the test data showing a unimodal trend. The review then delves into customer churn prediction, highlighting that customers with lower monthly payments are more likely to remain, while higher total charges correlate with a higher probability of retention. It emphasizes the importance for IT companies to understand the relationship between customer inflow and projected outflow in terms of revenue. A key insight provided is the effectiveness of the random forest model, attributed to the combination of a large number of relatively uncorrelated trees. The review explains that the weak correlation between the trees allows them to collectively outperform individual components, akin to diversified financial investments in a portfolio. This effect is attributed to the trees mitigating each other's mistakes until consistent errors in the same direction emerge. In conclusion, the review underscores the applicability of machine learning methods, particularly random forest and deep neural networks, in predicting customer churn for B2B SaaS companies in the IT sector. It highlights the stability and accuracy of a deep neural network with 32 hidden layers in identifying early signals of customer churn, enabling companies to implement proactive customer retention strategies through marketing activities and additional services. Overall, the review provides valuable insights into the optimization of machine learning models for churn prediction and underscores the importance of understanding customer behaviour in the context of subscription-based services. It advocates for the adoption of predictive analytics as a tool for enhancing customer retention efforts in the IT industry [20].

IV. METHODLOGY

Data description: The sample dataset was collected from Software as a Service startup with limited scope of use with privacy and security concern of the data. The data contains 21 columns. The dataset which is in .CSV format was imported into python notebook using Visual Studio code as .ipynb extension. The analysis and implementation of dataset was done using Python. The customer is found to be of the last year and customer that has not been using service more than 6 months are assumed to be churned.

Data Preprocessing: This process is an integral part of the whole process. Before the actual implementation and analysis of the dataset, it is a vital step to be followed so that all the missing as well as errors can be handled. This can help with getting desired results with more accuracy. The TotalCharges column of the dataset contains 11 missing values which are shown in Figure 3. The missing values should be handles such that it can help the analysis of data better without any difficulty in reading and generating output for better visualization of data.

Data Post processing: After applying the method, the missing values are then replaced by the mean value of the column. For the missing values, the TotalCharges column’s mean is calculated. After calculating the mean, all the null values of the column are replaced by the calculated mean value. After checking the null value, the dataset is not having any null value in the tenure column.

Algorithms Used: There are plethora of Machine Learning algorithms that are available that can be implemented to the dataset. Some of the algorithms that had been implemented to the dataset are as follows:

a. **Adaboost**: Ada – boost like Random Forest Classifier is another ensemble classifier (Ensemble classifiers consist of many classifier algorithms, the combined output of which is the outcome of those classifier methods). An individual algorithm might not classify the items very well. By using techniques such as Ada-boost, choosing the training set for a particular iteration and allocating the appropriate weight during the final voting, we may get a high accuracy score for the classifier as a whole. To put it briefly, Ada-boost retrains the algorithm repeatedly by selecting the training set according to the precision of the prior training. After combining with the Random Forest, Decision Tree, and Extra Tree classifiers in the prediction of the Churn of the telecom data-set, the Ada Boost classifier improved performance and accuracy. Similar to this, a variety of boosting methods or algorithms can be improved for greater results [15].

b. **Voting Classifier**: A machine learning model known as a voting classifier is trained on a large ensemble of models and forecasts an output (class) based on the models' best likelihood of producing the desired class. It merely compiles the results of every classifier that is fed into the voting classifier, which then predicts the output class according to the voting's largest majority. The goal is to train a single model that predicts output based on the cumulative majority of votes from each output class, rather than building individual specialized models and determining the accuracy for each one.

Two voting formats are supported by Voting Classifier.

i. **Hard Voting:** A class with the largest majority of votes, or the class with the best likelihood of being predicted by each classifier, is the predicted output class in a hard vote. Assume that three classifiers (A, A, and B) correctly predicted the output class; in this case, the majority correctly predicted A. Therefore, the ultimate forecast will be A.

ii. **Soft Voting:** In soft voting, the forecast made for a class is based on its average likelihood. Assume that the prediction probabilities for classes A and B are, respectively, (0.30, 0.47, 0.53), given a set of inputs to three models. Class A is obviously the victor because it had the highest probability summed by each classifier, with an average of 0.4333 for A and 0.3067 for B. [16].

c. **Gradient Boost**: In particular for classification challenges, gradient boosting generates robust, competitive, and comprehensible approaches. This method of creating ensemble models is based on fitting a first model—such as a linear regression or tree model—to the dataset and then creating a second model to precisely forecast the scenarios in which the first model fails. It is preferable to combine two models than to repeatedly enhance one model alone. Each new model attempts to make up for the inadequacies of the boost of the ensemble of all the models that came before it [2].

d. **Logistic Regression**: A probabilistic model called logistic regression is used to classify categorical values that depend on one or more factors in binary. Determining the relationship between variables is a statistical procedure. Data must be changed from its original form in order to solve the customer churn prediction issue, sometimes outperforming decision trees in the process. Logistic regression may be used to model categorical responses or response variables that have undergone some sort of transformation; much like linear regression can be used to represent numerical replies [2].

e. **Naïve Bayes**: A probabilistic strategy is thought to be one that uses a Naive Bayes classifier. According to Naïve Bayes, every vector characteristic is thought to be independent of the others. Class conditional independence refers to the classifier's presumption that each feature's value influences a particular class independently. We refer to it as Naïve since it is utilized to simplify computations. The Bayes theorem is a prerequisite for the Naïve Bayes concept. [10].

Evaluation Matrices: A useful tool for characterizing the difference between prediction and actual result in a binary classification is the confusion matrix, which may be used to assess and make the comparison effectiveness as well as accuracy of prediction models. The accuracy, precision, recall, and F1-score are the four dimensions in which calculations can be made using the confusion matrix [19].

1. Accuracy: The ratio of all right predictions to all observations, or accuracy, provides a rough idea of how well forecasts perform in terms of frequency. The computation is displayed in Formula (1):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (1)$$

2. Precision: The ratio of accurately anticipated churns to the overall churns predicted is known as precision. It is typically applied to problems that revolve around erroneously formulated predictions or misleading positive findings. The computation is displayed in Formula (2):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (2)$$

3. Recall: The recall measures how well real churns are anticipated in relation to all churns, or the ratio of churns predicted that are correct to the overall churns. The computation is displayed in

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

(3)the Formula (3).

4. F1-score: The average weight of recall and precision, or F1-score, shows the overall accuracy of the prediction. The computation is displayed in Formula (4).

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (4)$$

5. F2-scrore: The F2 score is the weighted harmonic mean of the precision and recall (given a threshold value). Unlike the F1 score, which gives same weight to precision and recall, the F2 score gives a little more weight to recall than that to precision. More weight should be given to recall for cases where False Negatives are considered worse than False Positives [21].

$$F2 = 5\left(\frac{(precision)\,(recall)}{((4)\,(precision)) + recall}\right) \qquad (5)$$

V. EXPERIMENTAL RESULTS

The data is visualized and interpreted using various packages available in Python to explore patterns of the data. The churn of the customers with respect to gender has been generated which shows amongst the churn customers; 939 females and 930 males while amongst the non-churn customers comprised of 2544 females and 2619 males. Out of the total customers, 73.4% did not churn while 26.6% churned. Both percentages and both genders combined in the figure. The data is visualized as in the Figure 1.

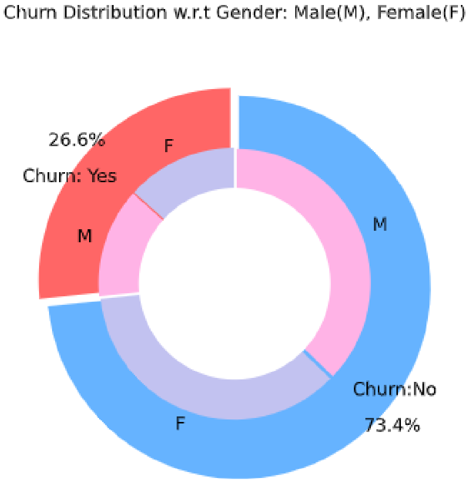

Figure 1: Churn rate with respect to gender

Customers with MonthlyCharges< 40 got churn is less and those that paid charges between 60 and 120 has higher churn rate as shown in Figure 2.

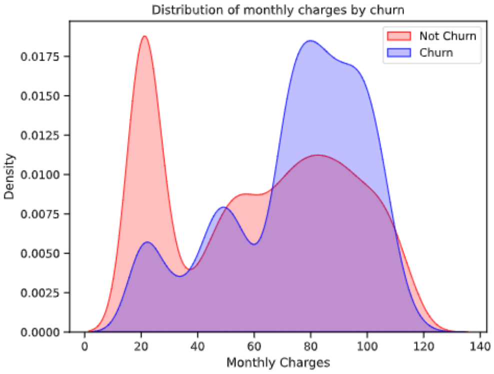

Figure 2: Distribution of monthly charges with churn

A customer with less TotalCharges has more churn rate than those customers who paid more. It depicts that the customer which uses the software less and use it less frequently has more churn rate while customers that are using the service for longer duration has less churn rate as shown in Figure 3.

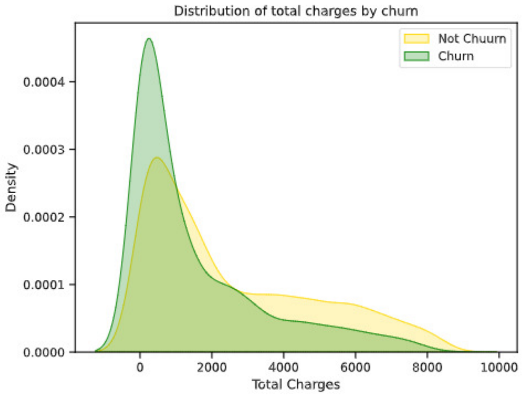

Figure 3: Distribution of total charges with churn

It is evident from the Figure 4 that customer being SeniorCitizen, using PaperlessBilling form of bill and MonthlyCharges shows positive correlation while customer having a partner, dependents, long tenure and high TotalCharges shows negative correlation with tenure showing the highest negative correlation with the churn rate of customers using the service.

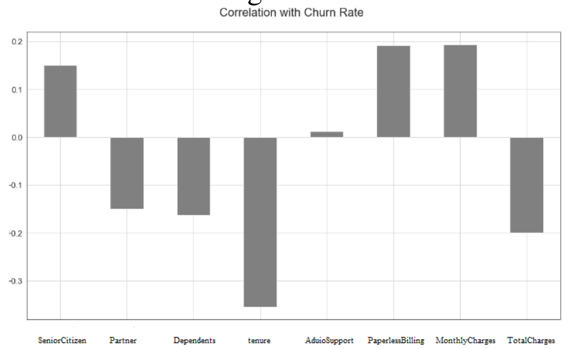

Figure 4: Correlation of attribute with churn rate

The Figure 5 shows correlation matrix of all the attributes with each attribute.

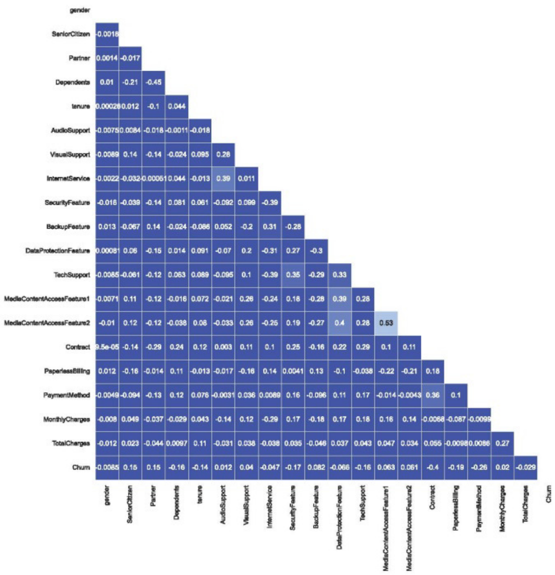

Figure 5: Confusion matrix of attributes

The Figure 6, 7 and 8 shows data distribution of feature tenure, MonthlyCharges and TotalCharges.

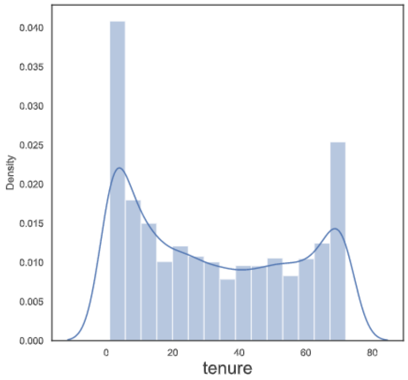

Figure 6: Tenure distribution

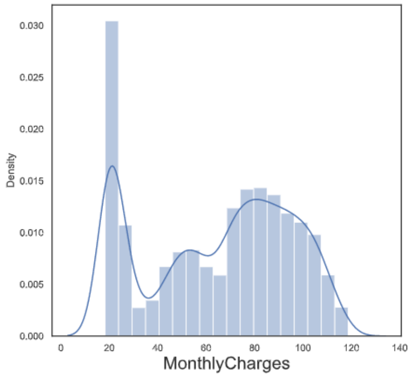

Figure 7: Monthly Charges Distribution

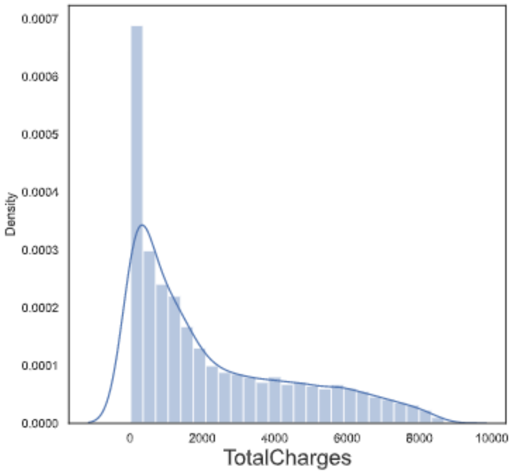

Figure 8: Total Charges Distribution

The Table 1 shows Receiver Operating Characteristics Curve's Area under the curve and Accuracy of each algorithm implemented with showing mean and standard deviation of each for every algorithm.

Algorithm	**ROC AUC Mean**	**ROC AUC STD**	**Accuracy Mean**	**Accuracy STD**
Voting Classifier	84.93	1.39	80.23	1.89
Gradient boost classifier	84.72	1.42	79.72	1.95
Adaboost	84.55	1.25	80.09	1.77
Logistic Regression	84.39	1.47	74.38	1.94
Gaussian NB	82.32	1.28	75.38	1.23

Table 1: ROC AUC and Accuracy mean and standard deviation

Every model is now evaluated in the form of confusion matrix for each algorithm. Figure 9 shows matrix for Adaboost, Figure 10 for Voting Classifier, Figure 11 for Gradient Boost, Figure 12 for logistic regression and Figure 13 for Naïve Bayes.

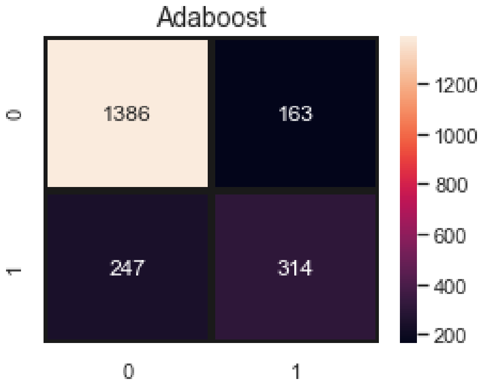

Figure 9: Adaboost Confusion Matrix

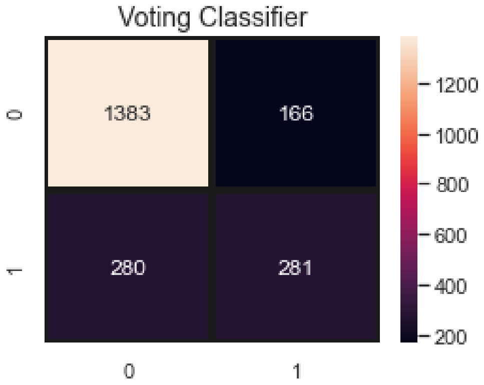

Figure 10: Voting Classifier Confusion Matrix

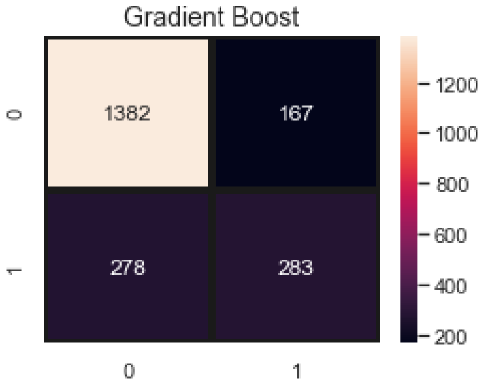

Figure 11: Gradient Boost Confusion Matrix

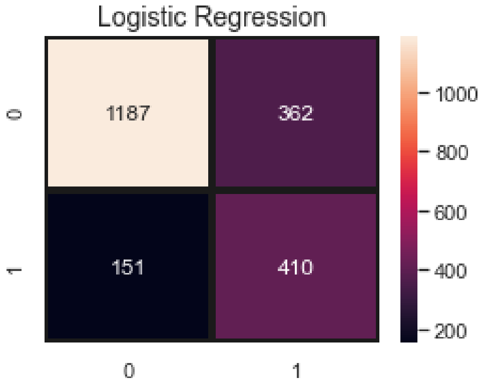

Figure 12: Logistic Regression Confusion Matrix

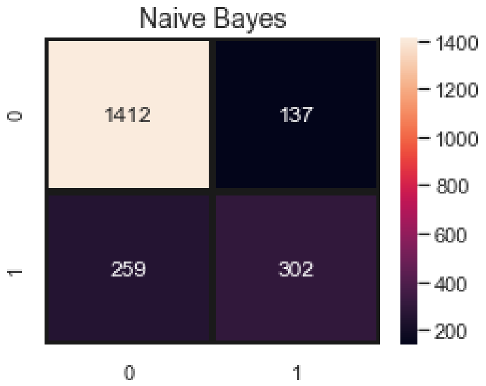

Figure 13: Naïve Bayes Confusion Matrix

Comparing True Positive Rate with False Positive Rate using ROC Curve of each algorithm.

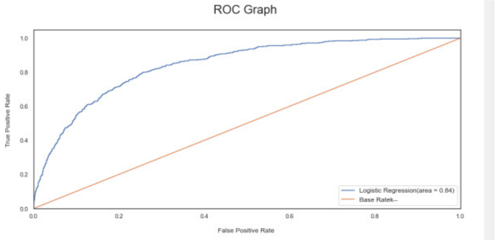

Figure 14: ROC Graph of Logistic Regression

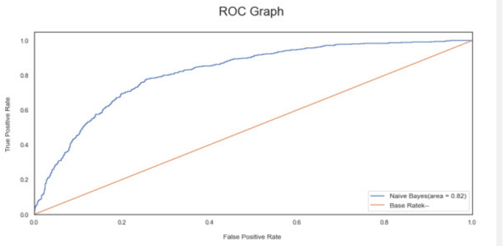

Figure 15: ROC Graph of Naïve Bayes

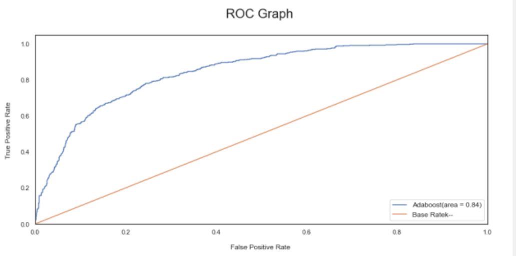

Figure 16: ROC Graph of AdaBoost

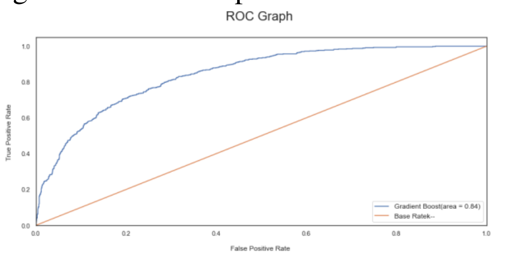

Figure 17: ROC Graph of Gradient Boost

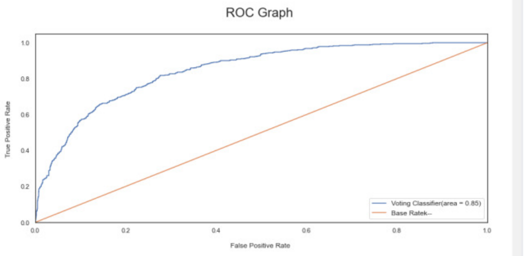

Figure 18: ROC Graph of Voting Classifier

Predicting feature importance using Gradient Boost for each variable as shown in Table 2.

	Features	Coefficient	Percentage (%)
14	Contract	0.403837	40.3837
4	tenure	0.142425	14.2425
17	MonthlyCharges	0.134165	13.4165
18	TotalCharges	0.108308	10.8308
8	SecurityFeature	0.064292	06.4292
11	TechSupport	0.055295	05.5295
7	InternetService	0.024634	02.4634
16	PaymentMethod	0.012317	01.2317
1	SeniorCitizen	0.011853	01.1853
15	PaperlessBilling	0.009874	0.9874
9	BackupFeature	0.008643	0.8643
6	Visual Support	0.006893	0.006893
10	DataProtectionFeature	0.004704	0.4704
13	MediaContentAccess Feature2	0.003343	0.3343
2	Partner	0.002785	00.2785
0	gender	0.002249	00.2249
5	AudioSupport	0.001479	00.1479
3	Dependents	0.0014712	00.147612
12	MediaContentAcessFeature1	0.001428	00.1428
	TOTAL	1.00	100%

Table 2: Coefficients using Gradient Boost

Predicting feature importance using AdaBoost Classifier as shown in Table 3.

	Features	Coefficient	Percentage (%)
18	TotalCharges	0.34	34
17	MonthlyCharge	0.20	20

	s		
4	tenure	0.14	14
14	Contract	0.12	12
16	PaymentMethod	0.04	04
8	SecurityFeature	0.04	04
10	DataProtection Feature	0.02	02
15	PaperlessBilling	0.02	02
11	TechSupport	0.02	02
9	BackupFeature	0.02	02
1	SeniorCitizen	0.02	02
6	VisualSupport	0.02	02
7	InternetService	0.00	00
12	MediaContentAcessFeature1	0.00	00
13	MediaContentAcessFeature2	0.00	00
5	AudioSupport	0.00	00
3	Dependents	0.00	00
2	Partner	0.00	00
0	gender	0.00	00
	TOTAL	1.00	100

Table 3: Coefficients using Ada Boost Classifier

Accuracy, Precision, Recall, F1 Score and F2 Score for each Model as evaluation as shown in Table 4.

Model	Accuracy	Precision	Recall	F1Score	F2Score
Adaboost	0.812322	0.687927	0.538324	0.604000	0.562803
Voting Classifier	0.808531	0.675615	0.538324	0.599206	0.561130
Gradient Boost	0.805213	0.672018	0.522282	0.587763	0.546642
Logistc Regression	0.805687	0.658281	0.559715	0.605010	0.576994
Gaussian NB	0.756872	0.531088	0.730838	0.615154	0.679708

Table 4: Evaluation Matrix

VI. CONCLUSION

Churn rate has been a serious concern for various industries and areas especially Software as a Service (SaaS) companies. It helps companies to determine the rate at which the users have stopped using the services provided by the company. Patterns revealed in the data may help companies to determine the customers and their segments which are leaving the services of an organization. Such data analysis helps companies to build strategy to target a particular segment with different strategy implemented for each segment according to the behavior of the customers of that segment. This study has helped in understanding of the data. Customers paying less monthly charges to the company are less likely to churn which shows positive correlation of monthly charges with churn rate. Customers paying more monthly charges are more likely to churn especially between 60 and 120. Customers paying less total charges to the industry are more likely to get churn which shows negative correlation with churn rate. Although negatively correlated, tenure which shows the duration the customer has been with the organization shows highest correlation with churn rate which is more than 0.3. Audio support shows the least correlation with the churn rate of the customers. Voting Classifier showed highest ROC AUC mean which is 84.93 while Gaussian NB shows the least which is 82.32. Logistic Regression has highest ROC AUC standard deviation which is 1.47 while AdaBoost shows least standard deviation which is 1.25. Voting Classifier has also shown highest accuracy mean which is 80.23 while Logistic Regression shows the least accuracy mean which is 74.38.Gradient Boost Classifier showed highest accuracy standard deviation which 1.95 and Gaussian NB shows least which is 1.23. Using GardientBoost, the feature of contract shows highest importance with coefficient .4038 while MediaContentFeatureAccess1 has .001428 value of coefficient. Using Adaboost Classifier, Total charges has highest importance with .34 value of coefficient while gender has the least importance of attribute.

REFERENCES

[1] Singh, S., Wankhede, A., &Patil, P. S. TELECOM CHURN PREDICTION USING MACHINE LEARNING ALGORITHM.

[2] Aldalan, A. M., &Almaleh, A. (2023). Customer Churn Prediction Using Four Machine Learning Algorithms Integrating Feature Selection and Normalization in the Telecom Sector. *International Journal of Electronics and Communication Engineering*, *17*(3), 76-83.

[3] Y. Bharambe, P. Deshmukh, P. Karanjawane, D. Chaudhari and N. M. Ranjan, "“Churn Prediction in Telecommunication Industry”," *2023 International Conference for Advancement in Technology (ICONAT)*, Goa, India, 2023, pp. 1-5, doi: 10.1109/ICONAT57137.2023.10080425.

[4] Sun, W., Zhang, K., Chen, SK., Zhang, X., Liang, H. (2007). Software as a Service: An Integration Perspective. In: Krämer, B.J., Lin, KJ.,Narasimhan, P. (eds) Service-Oriented Computing – ICSOC 2007. ICSOC 2007. Lecture Notes in Computer Science, vol 4749. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74974-5_52

[5] Web Site: Mashups and the Web as Platform, http://www.programmableweb.com/

[6] Web Site, http://www.employease.com

[7] Web Site: Salesforce.com AppExchange, [Online]: http://www.salesforce.com

[8] Prabadevi, B., Shalini, R., &Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*.

[9] Thorat, M. A. S., &Sonawane, V. R. (2023). CUSTOMER CHURN PREDICTION IN TELECOMMUNICATION INDUSTRY USING DEEP LEARNING. *Journal of Data Acquisition and Processing*, *38*(3), 1417.

[10] Gowd, S., Mohite, A., Chakravarty, D., &Nalbalwar, S. (2023, August). Customer Churn Analysis and Prediction in Telecommunication Sector Implementing Different Machine Learning Techniques. In *First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022)* (pp. 686-700). Atlantis Press.

[11] Almana, A. M., Aksoy, M. S., &Alzahrani, R. (2014). A survey on data mining techniques in customer churn analysis for telecom industry. *International Journal of Engineering Research and Applications*, *4*(5), 165-171.

[12] Khan, A. A., Jamwal, S., &Sepehri, M. M. (2010). Applying data mining to customer churn prediction in an internet service provider. *International Journal of Computer Applications*, *9*(7), 8-14.

[13] Singh, D., Jatana, V., &Kanchana, M. (2021). Survey paper on churn prediction on telecom. *Available at SSRN 3849664*.

[14] ÇALLI, L., & KASIM, S. (2022). Using Machine Learning Algorithms to Analyze Customer Churn in the Software as a Service (SaaS) Industry. *Academic Platform Journal of Engineering and Smart Systems*, *10*(3), 115-123.

[15] Lalwani, P., Mishra, M. K., Chadha, J. S., &Sethi, P. (2022). Customer churn prediction system: a machine learning approach. *Computing*, *104*(2), 271-294.

[16] https://www.geeksforgeeks.org/ml-voting-classifier-using-sklearn/

[17] https://www.javatpoint.com/major-kernel-functions-in-support-vector-machine#:~:text=Kernel%20Method%20in%20SVMs&text=The%20kernel%20function%20in%20SVMs,function%20computes%20their%20dot%20product.

[18] https://www.spiceworks.com/tech/big-data/articles/what-is-support-vector-machine/

[19] Amornvetchayakul, P., &Phumchusri, N. (2020, April). Customer churn prediction for a software-as-a-service inventory management software company: A case study in Thailand. In *2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA)* (pp. 514-518). IEEE.

[20] Kolomiiets, A., Mezentseva, O., &Kolesnikova, K. (2021). Customer churn prediction in the software by subscription models it business using machine learning methods. In *CEUR Workshop Proc* (Vol. 3039, pp. 119-128).

[21] https://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/scorers.html#:~:text=The%20F2%20score%20is%20the,to%20recall%20than%20to%20precision