World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:17, No:3, 2023

Customer Churn Prediction Using Four Machine Learning Algorithms Integrating Feature Selection and Normalization in the Telecom Sector

Alanoud Moraya Aldalan, Abdulaziz Almaleh

Abstract—A crucial part of maintaining a customer-oriented business in the telecommunications industry is understanding the reasons and factors that lead to customer churn. Competition between telecom companies has greatly increased in recent years, which has made it more important to understand customers' needs in this strong market. For those who are looking to turn over their service providers, understanding their needs is especially important. Predictive churn is now a mandatory requirement for retaining customers in the telecommunications industry. Machine learning can be used to accomplish this. Churn Prediction has become a very important topic in terms of machine learning classification in the telecommunications industry. Understanding the factors of customer churn and how they behave is very important to building an effective churn prediction model. This paper aims to predict churn and identify factors of customers' churn based on their past service usage history. Aiming at this objective, the study makes use of feature selection, normalization, and feature engineering. Then, this study compared the performance of four different machine learning algorithms on the Orange dataset: Logistic Regression, Random Forest, Decision Tree, and Gradient Boosting. Evaluation of the performance was conducted by using the F1 score and ROC-AUC. Comparing the results of this study with existing models has proven to produce better results. The results showed the Gradients Boosting with feature selection technique outperformed in this study by achieving a 99% F1-score and 99% AUC, and all other experiments achieved good results as well.

Keywords—Machine Learning, Gradient Boosting, Logistic Regression, Churn, Random Forest, Decision Tree, ROC, AUC, F1-score.

I. Introduction

THE availability of services to businesses and consumers in the world has increased dramatically as advances in telecommunications network technology have broadened the capabilities of the different company. The number of operators in the market has exponentially increased due to globalization and advancements in the telecommunications industry [1]. Therefore, the competitive environment increases the risk of customer churn and allows customers to select services based on their preferences and better prices. Additionally, it has become essential for telecom companies to maximize profits periodically; the cost of acquiring a new customer is five times greater than the cost of retaining an existing one, thus increasing client retention by 0.5 can increase profits by 25-95%. Furthermore, sales success rates for existing customers are between 60-70%, while those of new customers are 5-20% [2]. Churning customers is a crucial issue for an organization, as it can result in severe issues such as financial loss, expense, and time spent to acquire new customers, as well as customer dissatisfaction [3]. Considering the situation, it is vital to retain existing customers; this is something that every business must deal with. There is a new direction to save customers by detecting churn. Churn can be defined as customer attrition, turnover, or customer defection; they are all referring to the loss of clients or customers [4]. In reality, there is always a risk of churn, but in the long run, it can greatly harm your bottom line if churn is left unchecked; in other words, a higher churn rate means a lower retention rate. The main reason for churn is dissatisfaction with customer service and the system of support. Predicting which customers are likely to leave the company in the future will be the key to unlocking solutions to this problem. Once we identify those customers most likely to leave your company, it will be time to do something to keep them. You can target them with campaigns, offer them incentives to stay, or contact them to find out the reason and address the problem; all of these procedures maximize the chances that the customer will remain and increase profits. Therefore, establishing strategies to detect churn customers has become an essential requirement. The churn customer prediction models aim to detect the churn rate earlier [5], as well as factors that lead to churn. Consequently, companies have realized that their current customer data are their most valuable assets [6], as mentioned by Abasimehr [7] that churn prediction is a helpful solution to predict customers at risk and keep the company safe. Companies are centralizing their efforts on developing models for predicting churn and putting effort into retaining their existing customers and preventing churn. Researchers have contributed to customer retention by devising very effective experiments for predicting churn.

This paper aims to predict customer churn in the future and to identify the factors that lead to churn by evaluating their earlier usage history. It uses previous models and leverages some new works, such as applying feature selection tasks to improve performance and compares them with each other and with previous works, which are mentioned in the Related Work section. The idea of comparing the performance of different techniques is advantageous for future research. Section II of this paper provides information about previous related work done by researchers. Section III describes the proposed solution, including the dataset used and the methodologies applied in this study. Results are discussed and compared in

Ms. Alanoud Moraya Aldalan and Mr. Abdulaziz Almaleh are with the Department of Computer Science, King Khalid University, Saudi Arabia (e-mail: aldalanalanoud12@gmail.com, ajoyrulah@kku.edu.sa).

Section IV, and Section V is the Conclusion, highlighting what we achieved and what is expected in the future.

II. Related Work

Churn prediction analysis has been widely studied in the literature. Recently, however, customer churn prediction has become the primary focus of researchers. For instance, Jain et al. presented an experimental study for churn prediction by using two techniques of machine learning, namely Logistic Regression and Logit Boost. The results showed that both techniques outperformed one another, with not much difference in their results. Logistic Regression had an accuracy of 85.2385%, while Logit Boost had an accuracy of 85.1785%. Two similar techniques were used and compared; however, they could use more other different techniques to enhance their results [8].

Lalwani et al. applied a machine learning approach for churn prediction by using seven different algorithms. Their results showed that the ensemble learning techniques: Xgboost classifier and AdaBoost classifier gave the highest accuracy with 81.71% and 80.8%, respectively, and also had the highest AUC score with 84% over the others [9].

Karanovic et al. researched a CNN neural network to determine if it could predict customer churn, in addition to its most common use case of image classification. They achieved an accuracy of 98.85%, however, the data set was small and most features had missing data and were deleted. This may have caused the deletion of important performance features, so we cannot be sure that the CNN performs well in this problem [10].

Esteves et al. applied five machine learning algorithms - KNN, Naive Bayes, Decision Tree, Random Forest, AdaBoost, and a deep learning technique. The study yielded better results than the old models, with Random Forest outperforming the other algorithms by achieving 95% accuracy, 99% AUC, and 99% Sensitivity. All other experiments also produced very good results. This study highlights the importance of data mining techniques for a customer churn prediction model and suggests a good comparison model with standalone machine learning techniques, deep learning techniques, and hybrid models with feature extraction tasks, which are being used and compared on the same data set to evaluate the performance of the techniques better [11].

Jain et al. provided a comparative analysis of machine learning models for customer churn prediction, wherein they adopted seven algorithms: Logit Boost, Xgboost, Decision Tree, Logistic Regression, CNN-VAE, and PCA. Both hybrid and standalone algorithms were used with a feature engineering technique; the results showed that the random forest achieved the highest accuracy of 95%, while the accuracy without feature engineering was 93% [12].

In another study, Swetha et al. attempted to capture the churn prediction problem using a modified random forest. These algorithms have the ability to train on high-dimensional datasets and can express the uncovered patterns of a dataset. The authors proposed a modified random forest technique to solve the churn prediction problem. This modified random forest technique performs better than any other technique in terms of accuracy and area under the curve (AUC). In telecom services, this method is useful for estimating customer churn rates; the results showed that they proved their hypothesis by getting 92.7% accuracy with a modified random forest, while they got 74.64% accuracy for the standard RF [13].

III. Proposed Solution

The proposed solution is to consider churn customer detection as a classification problem that can predict whether a customer is churning or not with the help of machine learning algorithms. Thereafter, we will compare the performance by evaluation metrics, such as the F-score and AUC of all algorithms, to reach the best result. We aim to accomplish the following for this experiment: 1. Determine the factors that contribute to customer churn and visualize them. 2. Based on model performance, create a predictive model that can identify whether or not a customer will churn. All data analyses and classification are performed using Python.

A. Input Data Set

The dataset contains information about the customer usage history of a service from an American telecom company called Orange [14]. It includes 22 columns and 12,892 rows, the last column demonstrating whether the customer churns or not. The total number of churn customers is 1823, approximately 14.5% of the total number of customers. Table I shows the description of the Orange data set.

B. Exploratory Data Analysis

To develop an appropriate model for a problem and correctly interpret its results, a data analyst must conduct an EDA before engaging in machine learning or statistical modeling. Ensuring that the results they generate are valid, correctly interpreted, and applicable to the desired business context is valuable to a data scientist [15]. In order to understand the key characteristics of various variables in the data set for features, it is necessary and important to adjust the data set for analysis and observation by applying Pandas, NumPy, Statistical Methods, and Data Visualization libraries, and to try to find all possible hypotheses [16]. Once EDA is completed, meaningful insights are uncovered that can be used for machine learning models. Different techniques can be used to collect more information and insights about customers by implementing innovative approaches [17]. During this analysis, we will explore the relationship between the attributes and the Churn feature. Additionally, we will indicate the most significant relationship that we found in the dataset and formulate all possible hypotheses that could assist the telecom sector in understanding the causes and factors of churn.

Fig. 1 illustrates the proportion of churn customers, which is approximately 14.5% of the total number of customers, representing 1823 customers. Therefore, the proposed model should predict that 14% of customers will churn. Since 14% is a small number, we must ensure that the model selected predicts this 14% with high accuracy, as it is of interest to

World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:17, No:3, 2023

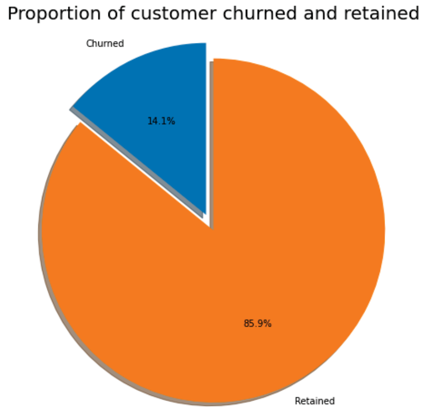

Fig. 1 Proportion of churn customers

the telecom sector to identify and retain this small group rather than just accurately estimating those customers that are retained.

TABLE I
ORANGE COMPANY CUSTOMER DATA SET DESCRIPTION

Feature	Type	Description
Record ID	Integer	Primary key of the record
State	String	51 states of the United States
Account-length	Integer	Account length
Area code	String	Area code (SF, OAK, SJ)
Intel Plan	Boolean	A customer has international plan or not
Vmail Plan	Boolean	A customer has a voice plan or not
Vmail messages	Integer	Number of voicemail messages
Minutes in day	Integer	Number of minutes in the morning
Calls in day	Integer	Number of calls in the morning
Charges in day	Decimal	Cost to customers in the morning
Evening minutes	Decimal	Total minutes spent in the evening
Evening calls	Integer	Number of calls made in the evening
Evening charge	Decimal	Cost to the customers in the morning
Night minutes	Decimal	Total of minutes spent at night
Night calls	Integer	Number of calls at night
Night charge	Decimal	Cost to the customers at night
Intel minutes	Decimal	International minutes
Intel calls	Integer	International calls
Intel charge	Decimal	International charges
Customer services	Integer	Number of calls to Customer Services
Churn	Boolean	Customer has churned or not
Customer ID	String	Enterprise ID of the customer

As shown in Table I, we have both categorical and continuous variables; the attributes of the record ID and customer ID, which are specific to each customer, are excluded to avoid profiling.

The voicemail feature displayed in Fig. 2 indicates that churn happens when there are more than 20 voicemail messages. This could be because of the poor quality of the voicemail after 20 messages. By providing service upgrade services or limiting the number of voicemails to 25, it may be possible to improve the voicemail feature and reduce churn rates.

As shown in Fig. 3, we can notice the churn in those customers who do not have a voicemail plan; they may be unsatisfied with their bills or more adversely affected than others, and as a result, have to change their service provider.

We can deduce from the boxplot in Fig. 4 that those

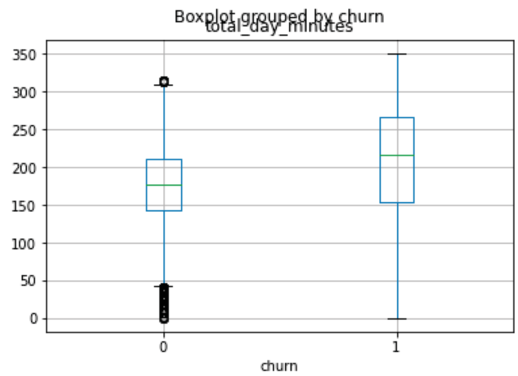

Fig. 2 Voice Mail Number vs Churn

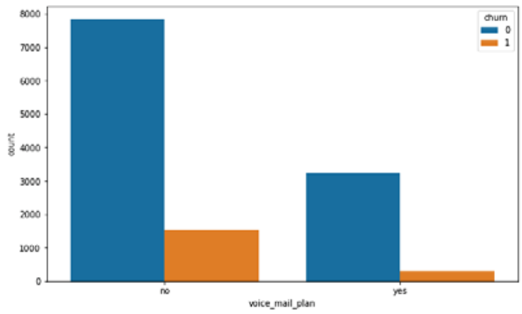

Fig. 3 Voice Mail Plan vs Churn

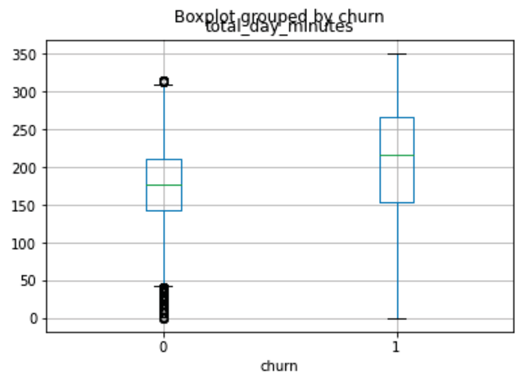

Fig. 4 Total Day Minutes vs Churn

customers who spend more than 225 minutes are churning. The reason may be due to network problems during daytime calls, causing them to change service providers. The telecom company should pay attention to these problems and examine the quality of service, to see if it needs to be upgraded using advanced technologies.

From the box plot in Fig. 5, we can notice that customers who spend more minutes on the network are more likely to switch their service provider, so the company should review its own charge prices and offer a discount to those customers.

Fig. 6 illustrates that customer feedback and confirmation in situations like these would be useful to ensure the customer receives good service, as customers who do not actively resolve their issues and do not contact customer service are most likely to leave the company.

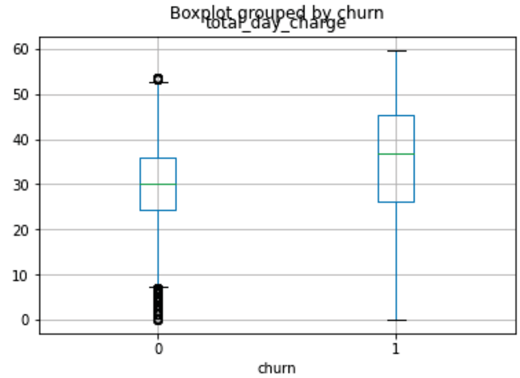

Fig. 5 Total Day Charge vs Churn

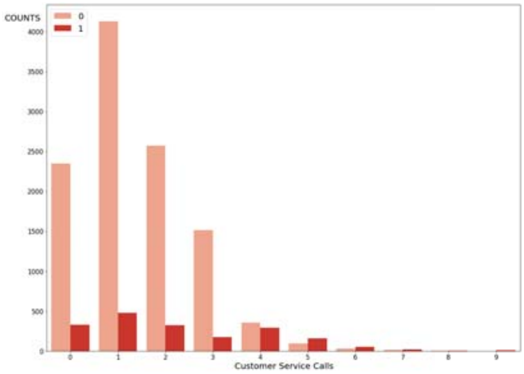

Fig. 6 Customer Service Call vs Churn

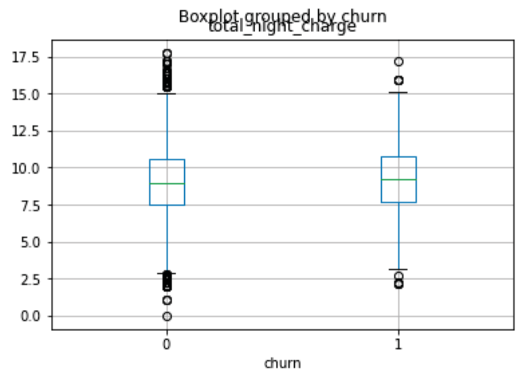

Fig. 7 Total Night Charge vs Churn

From the statistical analysis, we can notice that calls are still made more often by loyal customers. The improvement of the charge pricing might lead to more loyal customers, as illustrated in Fig. 7.

C. Data Prepossessing

1) Feature Engineering: The churn customer dataset contains several categorical features that may provide valuable information about customers, so they are important for building a model using those features. It is possible to remove features with null and low importance from the dataset but removing these categorical features with high importance could negatively impact the model. As machine learning cannot understand and handle these variables as they are, these variables must be transformed so that they can be read by machines in a way that increases model performance. Label encoding for a categorical variable is a good technique for processing categorical variables such as area codes and state features and converting "Yes" and "No" in the voice mail plan, international plan, and churn features into 1 for "Yes" and 0 for "No". Subsequently, in this study, feature significance is assessed for all features and based on this, features are included or excluded for the prediction process. In addition, our dataset has 15 continuous variables, all of which have different value ranges. Having values in different scales presents problems for machine learning algorithms. Therefore, there is a need to put all these values in the same range to build an effective and unbiased model. This can be achieved by normalizing all continuous features to a certain value. The normalization done by Minimum–Maximum value is helpful in preserving the relationships between the original input data, unlike normalization methods which depend on the mean and standard deviation of the data, as these values may change over time. One of the main drawbacks is that this method is sensitive to outlier values present in un-normalized data [18], as well as the out-of-boundary issue that occurs in cases where test data values are outside of the normalized range [19]. Min-Max normalization is a method of scaling feature values between zero and one by taking the maximum and minimum values of the feature variables and scaling them accordingly, illustrated as:

$$z_i = \frac{(x_i - min(x))}{(max(x) - min(x))}$$

where z_i: The i normalized value in the dataset; x_i: The i value in the dataset; $min(x)$: The minimum value in the dataset; $max(x)$: The maximum value in the dataset.

2) Feature Selection: Using feature selection algorithms can help increase the performance of a model, though traditionally a model requires a large number of features to be accurate [20]. Feature selection is often used to construct a model using various subsets of data to create a dependable and precise model. The technique determines which features are most likely to be significant in predicting churn.

The study uses two techniques for selection: the Pearson Correlation Coefficient (PCC), the Chi-Squared (CHI) method for categorical features, and the T-Test for numerical ones.

As a statistical term, the Pearson Correlation Coefficient (PCC) refers to the degree to which two features have a linear relationship with each other [21]. The correlation scores range between -1, which indicates a negative correlation, and +1, which indicates a positive correlation, with 0 being no correlation. A feature with a high correlation is linearly dependent, so its effect on the target variable (churn) is almost the same. If two features have a high correlation, one of them can be deleted [21]. Fig. 8 illustrates the correlation examination of features to each other; we notice that the features: charge in the day, charge in the evening, charge at night, and international charges are 100% correlated to total minutes in a day, evening, night, and minutes of international. Therefore, to effectively build a better model, it will be useful

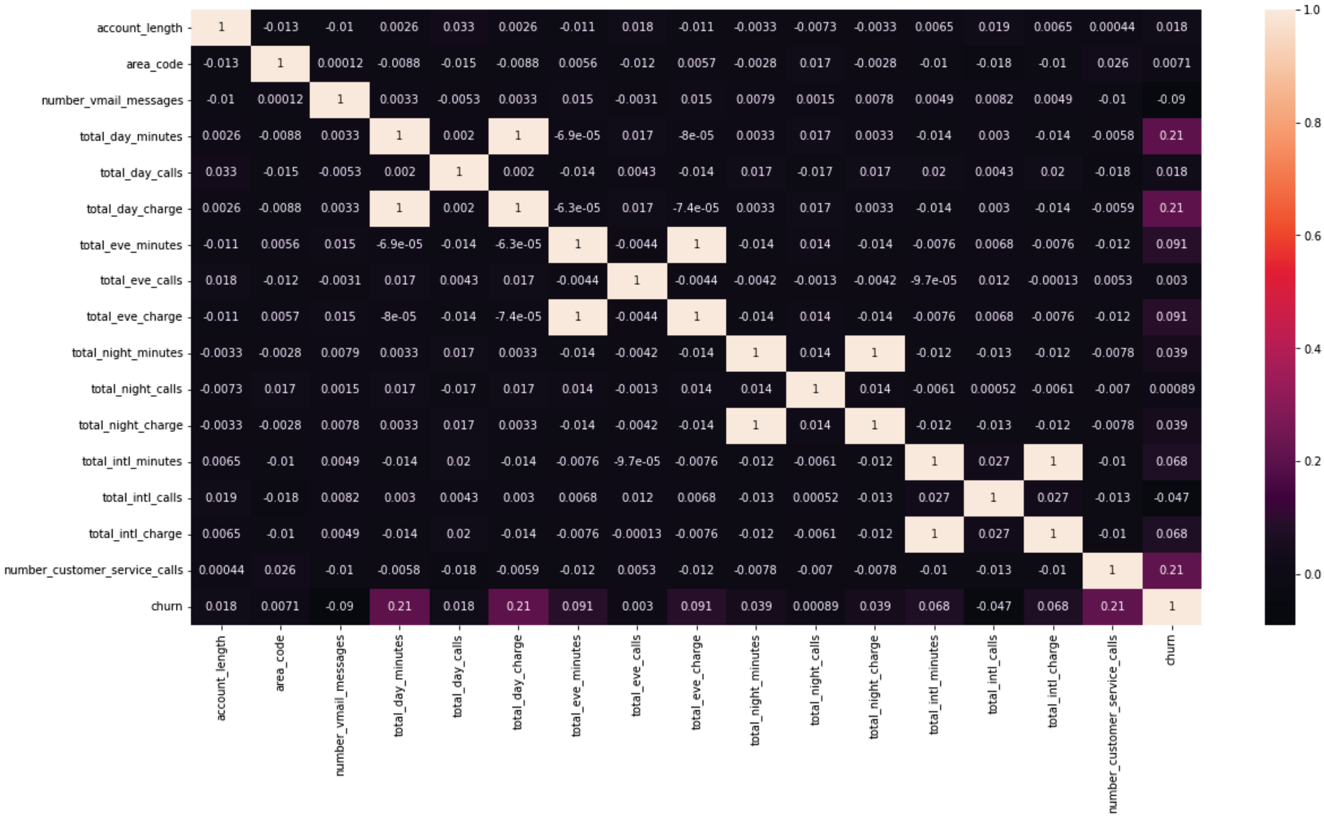

Fig. 8 Pearson correlation coefficient displayed in a heat map

to remove these features from the dataset and focus on the remaining features that correlate with the target class, such as minutes in a day, charge in the day, and total customer service calls, which are related to customer churn with a 0.21 score.

A Chi-square test can be used to test whether row and column entities are associated in a two-way table (contingency table). The null hypothesis (H0) assumes that there is no association or interdependence between the variables, in other words, one variable does not vary depending on the other variable, while the alternative hypothesis (Ha) assumes that there is an association or interdependence between them. Basically, the chi-square test assesses the difference between the observed and expected values under the null hypothesis of no interdependence or association. For this test calculation, it is essential to calculate the expected values based on the data; a cell in a contingency table has an expected value equal to (row total multiplied by column total) divided by n, where n is the number of observations included in the table [22], [23].

A t-test compares the averages of two groups using a statistical test. It is used to test a hypothesis, which is a process used to determine if one group affects the entire group of interest, or if two groups are different from one another [24]. T-test and Chi-square test are calculated as the following equations:

Chi-test equation:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i= Observed value (actual value); E_i= Expected value.

T-test equation:

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where $\bar{x}$ = Observed Mean of the Sample; μ = Theoretical Mean of the Population; s = Standard Deviation of the Sample; n = Sample Size.

The t-test and chi-square test depend on the p-value, which stands for the probability value, and it indicates whether the result could have occurred by chance alone. If the p-value is smaller, it is more likely that the feature should be included in the model [25].

Based on the Chi-Test results illustrated in Fig. 9, the Area Code feature has been removed since its P-value is greater than the significance value, thus failing to reject the null hypothesis that the feature is significantly essential. Almost all features are statistically significant as shown in Fig. 10 via a T-test, except for total evening calls and total night calls. Therefore, we removed them in order to reduce dimensionality and improve performance.

D. Modeling

In this paper, four models are analyzed as classifiers to predict customer churn: Random Forest, Logistic Regression, Decision Tree, and Gradient Boosting. The dataset has been shuffled and split, with 80% for the training dataset and 20%

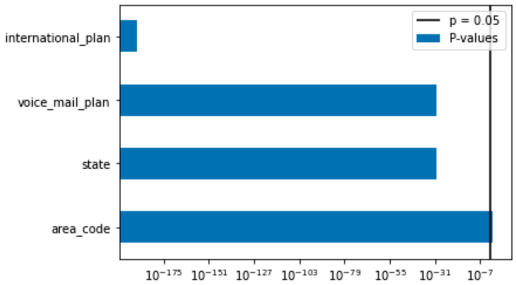

Fig. 9 Chi-Test Results

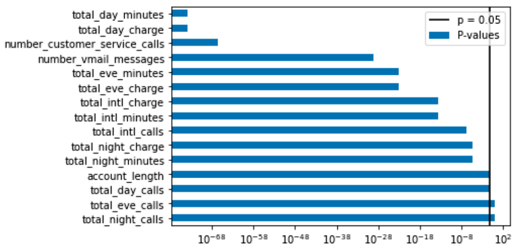

Fig. 10 T-Test Results

for the testing dataset, using the rule of thumb. The models are as follows:

1) A random forest is an ensemble of decision tree classifiers. It is a collection of classifiers called a ”forest”. These trees are generated using a random selection of features at each node to determine the split process, and the values of a random vector determine the split, sampled randomly from each tree in the collection of classifiers with the same distribution. During classification, each tree votes, and the most occurring class is returned [26]. It could work with thousands of parameters without deletion or destruction, and it can also handle missing values in the data set, so it can be trained to obtain an efficient model [9].
2) Gradient boosting produces competitive, robust, and interpretable procedures, especially for classification problems [27]. It is a technique for developing ensemble models based on fitting an initial model, such as a tree model or linear regression, to the dataset, and then building another model to accurately predict the cases where the initial model fails. Combining two models will lead to better results than boosting one model alone many times. Every subsequent model tries to correct for the shortcomings of the boost of the ensemble of all previous models [28].
3) Logistic Regression is a probabilistic model used for binary classification of a categorical value that depends on one or more parameters. It is a statistical process for determining how variables are related to each other [9]. In dealing with the customer churn prediction problem, data must be transformed from its initial form in order to achieve better performance, sometimes performing as well as decision trees. In the same way that linear regression can model numeric responses, logistic regression can model categorical responses or response variables that have been transformed in some way [29].
4) The decision tree is one of the predictive statistics and modeling approaches used in the fields of machine learning and data mining. Models of decision trees that target a variable which can take a finite set of values are called classification trees. In this tree structure, class labels are represented via leaves and branches which represent combinations of features that contribute to those classes. The main parameter to set the root node of the tree is gain. Decision tree classifiers have been shown to have better accuracy than other classification methods, according to Sharma’s survey [30]. Its process is based on identifying ways to split a dataset based on different events. It is one of the most widely used and practical methods for machine learning [31].

As mentioned earlier, the targeted churn customers for prediction are 14% of the total customers, which is too small to train the model on. This is known as an imbalanced dataset, and due to this difference in each class, the algorithms will tend to be biased towards the majority class, which in our case are the non-churner customers. This will not result in a good performance on the minority class. Therefore, handling the imbalanced dataset by using grid search with hyperparameters has been fine-tuned using 5-fold cross-validation and a class weight equal to ”balanced” to prevent overfitting, and to give different weights to both the majority and minority classes. This difference in weights will resist the biased classification process of the classes through the training process, balancing the misclassification made by the minority class by setting a higher class weight and reducing weight for the majority class [32]. Afterwards, the model with the best scores is chosen to make predictions in new data, referred to as a test set.

E. Evaluation

In each model, precision, accuracy, F1 score, recall, and ROC-AUC were calculated. The dataset was shuffled according to the rule of thumb of 80% for the training dataset and 20% for the test dataset. As a measure of imbalanced learning, recall typically measures how well the minority class is covered, aiming to improve recall without affecting precision. Unfortunately, these goals can sometimes be inconsistent with each other, so in order to increase the true positive rate for minority classes, the number of false positives is also oftentimes increased, contributing to reduced precision [33]. However, rather than picking one measure or the other, we can choose a new score that combines both precision and recall: the F1 score.

The F1 score provides a harmonic mean of recall and precision, which is the average of them. As all of the precision and recall are rates, it makes logical sense to use the harmonic mean, meaning it will give equal weight to precision and recall. If both precision and recall are high, the F1 score will also be high, and if the F1 score is low, it means that both precision and recall are low. A medium F1 score indicates that one of precision and recall is low and the other is high. The F1 score is calculated as illustrated as:

F1 Score:

$$2\times\frac{Precision \times Recall}{Precision + Recall}$$

The second performance metric is an area under the Receiver Operating Characteristic (ROC) curve. ROC AUC can be promising on severely imbalanced classification problems with small examples of the target class [34]. It is one of the most popular evaluation metrics for binary classification problems. If a classifier performs well, then it should readily increase the true positive rate and the area under the curve should be close to 100%. When the classifier performs similarly to the random classifier and the area under the curve is about 50% or less, the true positive rate increases linearly with the false positive rate. The ROC curve is a plot of the False Positive Rate (x-axis) vs. the True Positive Rate (y-axis), calculated as shown as:

True Positive Rate(Y):

$$\frac{TP}{TP+FN}$$

False Positive Rate (X):

$$\frac{FP}{FP+TN}$$

The plot of the ROC AUC is the rate of correct predictions of the target class (y-axis) versus the rate of errors in the negative class (x-axis). Ideally, getting the fraction of correct target class predictions to be 1 should be at the top left corner of the graph and the rate of incorrect negative class predictions should be 0 to the left of the graph. Thus, the best possible classifier to achieve ideal performance is at the top left of the plot (coordinate (0,1)) [34].

IV. Results

In this section, comparisons among four experiments using the Orange dataset are presented. These algorithms deal with unbalanced datasets, and the performance is compared based on the results of two experiments. In the first experiment, no feature selection technique was performed, and all 20 features were included. In the second experiment, feature selection technique was performed using feature significance tests and PCC with 13 features included. All models were evaluated using the F1-score and area under the curve (AUC-ROC) and compared to each other, as well as with similar existing models described in the related work. The results showed that the performance of the models with feature selection was better than that of standard models, and all ensemble models proved to be persistent against the dimensionality of features and an imbalanced dataset, since there was not much difference in their results in both scenarios. All models showed good results in terms of F1 scores and AUC: gradient boosting outperformed all other models with 99% F1 score and AUC; logistic regression provided poor results with 49% F1 score and 77% AUC; decision tree performed well with 98% AUC and 96% F1 score; and random forest performed well, achieving 95% for both metrics. Tables II and III demonstrate the results of both scenarios.

TABLE II
Performance Without using Feature Selection

Model	F1-score	AUC
Logistic Regression	48%	76%
Random Forest	95%	95%
Decision Tree	95%	97%
Gradient Boosting	99%	99%

TABLE III
Performance With using Feature Selection

Model	F1-score	AUC
Logistic Regression	49%	77%
Random Forest	95%	95%
Decision Tree	96%	98%
Gradient Boosting	99%	99%

V. Conclusion and Future Work

In conclusion, the study applied four experiments to predict customer churn. The results showed that the Gradient Boosting model with feature selection tasks performed well in comparison to the other three experiments and previous models. Additionally, the study demonstrated the impact and value of feature engineering and selection in increasing and improving the performance of the models. Furthermore, once the customer has been identified as churning, alternative strategies should be presented to deal with the churn rate and prioritize it. There is still potential for improving the performance of the results by exploring deep learning models over the traditional models, and the data are continuously generated on a daily basis, therefore further experimentation remains necessary.

References

[1] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS quarterly*, pp. 1165–1188, 2012.
[2] T. Landis and S. Philips, "Customer retention marketing vs. customer acquisition marketing," Apr 2022. [Online]. Available: https://www.outboundengine.com/blog/customer-retention-marketing-vs-customer-acquisition-marketing/
[3] S. Yadav, A. Jain, and D. Singh, "Early prediction of employee attrition using data mining techniques," in *2018 IEEE 8th International Advance Computing Conference (IACC)*, 2018, pp. 349–354.
[4] B. J. Ng), "Classification analysis on telco customer churn," Nov 2021. [Online]. Available: https://medium.com/swlh/classification-analysis-on-telco-customer-churn-a01599ad28d7
[5] C.-P. Wei and I.-T. Chiu, "Turning telecommunications call details to churn prediction: a data mining approach," *Expert systems with applications*, vol. 23, no. 2, pp. 103–112, 2002.
[6] K. Coussement and D. Van den Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," *Expert systems with applications*, vol. 34, no. 1, pp. 313–327, 2008.
[7] H. Abbasimehr, M. Setak, and M. Tarokh, "A neuro-fuzzy classifier for customer churn prediction," *International Journal of Computer Applications*, vol. 19, no. 8, pp. 35–41, 2011.
[8] H. Jain, A. Khunteta, and S. Srivastava, "Churn prediction in telecommunication using logistic regression and logit boost," *Procedia Computer Science*, vol. 167, pp. 101–112, 2020.
[9] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing*, pp. 1–24, 2021.

[10] M. Karanovic, M. Popovac, S. Sladojevic, M. Arsenovic, and D. Stefanovic, "Telecommunication services churn prediction-deep learning approach," in *2018 26th Telecommunications Forum (TELFOR)*. IEEE, 2018, pp. 420–425.

[11] G. Esteves and J. Mendes-Moreira, "Churn perdiction in the telecom business," in *2016 Eleventh International Conference on Digital Information Management (ICDIM)*. IEEE, 2016, pp. 254–259.

[12] H. Jain, A. Khunteta, and S. P. Shrivastav, "Telecom churn prediction using seven machine learning experiments integrating features engineering and normalization," 2021.

[13] P. Swetha, S. Usha, and S. Vijayanand, "Evaluation of churn rate using modified random forest technique in telecom industry," in *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 2018, pp. 2492–2497.

[14] O. company, "Customer signature for churn analysis - dataset by bob-wakefield," Sep 2019. [Online]. Available: https://data.world/bob-wakefield/call-center-data

[15] C. MAWER, "The value of exploratory data analysis," Sep 2017. [Online]. Available: https://www.svds.com/value-exploratory-data-analysis/

[16] s. babu, "Eda for machine learning: Exploratory data analysis in python," Apr 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/04/rapid-fire-eda-process-using-python-for-ml-implementation/

[17] A. Ultsch, "Emergent self-organising feature maps used for prediction and prevention of churn in mobile phone markets," *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 10, no. 4, pp. 314–324, 2002.

[18] G. Dougherty, *Pattern recognition and classification: an introduction*. Springer Science & Business Media, 2012.

[19] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, 2020.

[20] A. Al Marouf, M. K. Hasan, and H. Mahmud, "Comparative analysis of feature selection algorithms for computational personality prediction from social media," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 3, pp. 587–599, 2020.

[21] V. R, "Feature selection - correlation and p-value," Apr 2022. [Online]. Available: https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf

[22] K. Pearson, "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[23] M. Nikulin, "Chi-squared test for normality," in *Proceedings of the International Vilnius Conference on Probability Theory and Mathematical Statistics*, vol. 2, no. 1, 1973, pp. 119–122.

[24] R. Bevans, "An introduction to t-tests," Dec 2020. [Online]. Available: https://www.scribbr.com/statistics/t-test/

[25] A. A, "Statistical tests: Feature selection using statistical tests," Jun 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/feature-selection-using-statistical-tests/

[26] J. Han, M. Kamber, and J. Pei, "Data mining: Concepts and techniques third edition [m]," *The Morgan Kaufmann Series in Data Management Systems*, vol. 5, no. 4, pp. 83–124, 2011.

[27] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[28] J. Hoare, "Gradient boosting explained - the coolest kid on the machine learning block - displayr," https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/, (Accessed on 05/05/2022).

[29] U. Ş. Gürsoy, "Customer churn analysis in telecommunication sector," *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, vol. 39, no. 1, pp. 35–49, 2010.

[30] H. Sharma and S. Kumar, "A survey on decision tree algorithms of classification in data mining," *International Journal of Science and Research (IJSR)*, vol. 5, no. 4, pp. 2094–2097, 2016.

[31] P. Yadav, "Decision tree in machine learning — by prince yadav — towards data science," https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96, 11 2018, (Accessed on 05/05/2022).

[32] K. Singh, "How to dealing with imbalanced classes in machine learning," Oct 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/

[33] H. He and Y. Ma, *Imbalanced learning: foundations, algorithms, and applications*. Wiley-IEEE Press, 2013.

[34] J. Brownlee, "Roc curves and precision-recall curves for imbalanced classification," https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/, Jan 2020, (Accessed on 05/05/2022).