Predicting Customer Churn in Telecommunications: A Machine Learning Approach for Mogadishu, Somalia

Zakaria Ali Hassan
Department of computer application ,jamhuriya university of science and technology
Mogadishu, Somalia
zakariyea153@gmail.com

Adan Saleban Ali
Department of computer application, ,jamhuriya university of science and technology
Mogadishu, Somalia
adan.sleban@gmail.com

Abshir Muse Ahmed
Department of computer application ,jamhuriya university of science and technology
Mogadishu, Somalia
capitalyare@gmail.com

Abdijibar Jama Mohamed
Department of computer application ,jamhuriya university of science and technology
Mogadishu, Somalia
kaapearaysane@gmail.com

***Abstract*—With the use of machine learning techniques, this paper offers a thorough method for forecasting customer attrition in Mogadishu, Somalia's telecom industry. The rapidity at which consumers stop using a telecom provider is known as customer churn, and it presents a serious financial risk. The objective of this project is to improve prediction accuracy and provide guidance for targeted client retention efforts by utilizing a range of machine learning models, including Support Vector Machines, Random Forest, Decision Trees, and Logistic Regression. The study utilizes a strong approach that includes feature selection, data preparation, and model assessment in order to successfully handle the subtleties of client behavior patterns. The results show that Random Forest models outperform other models in anticipating possible customer attrition, which makes proactive client retention initiatives possible. This effort advances not only the theoretical**

Keywords—customer churn prediction, accuracy, decision tree, Random Forest, model selection, data preprocessing

I. INTRODUCTION

The telecommunications sector has become one of the main industries in developed countries. Companies are working hard to survive in this competitive market depending on multiple strategies. Three main strategies have been proposed to generate more revenues acquire new customers, upsell the existing customers, and increase the retention period of customers. However, comparing these strategies taking the value of return on investment (RoI) of each into account has shown that the third strategy is the most profitable strategy, proves that retaining an existing customer cost much lower than acquiring a new one[1].

One of the biggest issues facing the telecommunications sector is customer attrition. The purpose of this study is to determine the variables that affect customer attrition, create a predictive model for customer attrition, and offer the best analysis of data visualization outcomes. The suggested approach for analyzing churn prediction consists of many stages: pre-processing the data, analysis, application of machine learning algorithms, assessment of the classifiers, and selection of the most effective one for prediction. The three main steps in the data preparation procedure were feature selection, data transformation, and data cleaning[2].

Customer churn detection is one of the most crucial study areas in the telecommunications industry that a firm must address in order to retain current clients. Churn refers to client attrition as a result of competitors' offers ending or maybe network problems. The lifetime value of a client is significantly impacted by churn rate as it influences both the duration of service and the company's future income. The businesses are searching for a model that can forecast client attrition as it has a direct impact on industry revenue. Random Forest with XGBoost [3].

The development of the sector depends on improved client requirements perceptions and higher-quality models and services. Companies that want to be profitable and competitive must priorities reducing customer attrition, which has a significant influence on their operations. Thus, a great deal of study has been done by academics all over the world to comprehend the mechanisms of client turnover. In order to determine the numerous churn variables and their intricate linkages in the telecom churn literature now in existence[4].

Churn prediction is a key predictor of the long-term success or failure of a business. In this research, Ubiquitous techniques like Random Forest Classifiers and SVMs are compared with relatively newer architectures like XGBoost and Deep Neural networks to classify if a customer will churn or not. The efficiency of these models is further explored by passing them through a grid search. From this experiment, it could be inferred the Random Forest model works best for this particular use case with a prediction accuracy of 90.96% on the testing data before grid search[5].

Every day, a sizable number of customers in the telecom sector create massive amounts of data. Churn, or the process of clients moving from one company to another within a predetermined period of time, in order to keep consumers' data from being churned, telecom management and analysts are attempting to figure out why subscribers are cancelling their contracts. This system gathers the reasons why consumers in the telecom business subscribe to leaves and utilizes classification algorithms to determine which customers have subscribed to leaves[6].

For the majority of businesses operating in low-cost switching sectors, customer turnover is their top priority. Of

the industries that have this problem, the telecommunications sector is thought to be the worst. With the aim of lowering the rate of customer attrition, mobile service providers have deployed CRM (Customer Relationship Management). Still, there is a significant incidence of employee turnover in the telecom sector[7].

Customer churn is the term used to describe the loss of important clients to rival companies by service providers, particularly those in the telecommunications industry. The telecom sector has seen significant transformation in the past few years, including new services, new technologies, and market liberalization that increased competition. Customer attrition results in a significant loss of telecom services, making it a very critical issue[8].

II. LITERATURE REVIEW

Customer churn is the term used to describe the loss of important clients to rival companies by service providers, particularly those in the telecommunications industry. The telecom sector has seen significant transformation in the past few years, including new services, new technologies, and market liberalization that increased competition. Customer attrition results in a significant loss of telecom services, making it a very critical issue[9].

It is especially crucial when calculating a business in industries where subscription-based income is generated, like banking, insurance, or telecommunications. Researchers have found that acquiring new clients in the cutthroat market of today might be up to ten times more expensive than keeping hold of current ones. It is an analysis technique used for things like figuring out current consumer profiles, examining customer escapes, and projecting customer escapes[10].

In today's corporate world, acquiring and keeping clients are the most important priorities. Every business's market is expanding quickly, which is increasing the number of subscribers. As a result, businesses now understand how critical it is to hold onto their current clientele. It is now required of service providers to lower churn rates because, from a big perspective, neglect could result in a decline in profitability. Churn prediction assists in determining which consumers are most likely to depart from a business. The problem of an ever-rising churn rate is one that the telecom industry is dealing with. By using data mining tools, these telecom businesses can develop efficient strategies to lower their turnover rate[11].

Customer attrition is the process of a customer switching from one business service to another. Customer Churn Prediction is a tool used to predict potential customers who may quit the firm before they do. In order to attract potential churners and keep them around, this phase assists the business in developing the necessary retention policies, which lowers the company's financial loss[12].

Many sectors are concerned about customer turnover, but highly competitive industries are more vulnerable to it. Losing clients increases the need to find new ones and results in financial loss due to lower revenues[13].

In every business, including banking, customer churn has grown to be a significant issue. To identify potential departing customers, banks have long attempted to monitor customer interactions. The basic goal of customer churn modelling is to identify high-risk consumers so that preventive measures can be taken[14].

According to earlier studies, there are two categories of focused approaches—proactive and reactive—for handling client attrition. Reactively, the business holds off on terminating service until the client requests it. The business looks for clients who are likely to leave in an effort to be proactive. Next, the business offers incentives in an effort to keep those clients. Customers will leave businesses if churn projections are off, so churn should be accurate to avoid wasting money[15].

To learn more about the data and the business challenge, data exploration is necessary. For the Data mining model, the CRISP-DM methodology is widely acknowledged[16].

The three different strategies for addressing class imbalance were the cost-sensitive strategy, the algorithm-level strategy, and the data-level strategy. The following techniques were used in the data level approach: under sampling, oversampling, and hybrid sampling. Under sampling results in the loss of potentially valuable data, whereas oversampling, when combined with massive amounts of data, causes overfitting and lengthens the learning process. Methodology at the algorithmic level the bagging and boosting techniques are used to address class imbalance. Decision tree (C4.5) and Random Forest algorithms were utilized for the bagging approach, while AdaBoost and SMOTEBOOST algorithms were used for the boosting method. The data-driven and algorithm-level approaches are both included in the cost-sensitive strategy[17].

Machine learning is mostly applied to difficult tasks or problems involving large amounts of data. For more complicated data, it is a suitable solution since it produces faster, more accurate findings. It aids a company in recognizing any unidentified hazards or lucrative opportunities. Two primary learning approaches are used in machine learning: Supervised Machine Learning and Unsupervised Machine Learning [18].

As a result of their inability to forecast which consumers will leave on schedule, many businesses often deal with the problem of losing clients. This research's primary goal is to give telecom providers a quick and practical method for identifying potential at-risk clients. In order to create our churn prediction model, we applied both logistic regression and logit boost[19]

The goal of this study is to determine how Particle Swarm Optimization determines the most appropriate K value parameters and how Z-Score normalizes the data to improve the performance of the K-Nearest Neighbor algorithm during classification. The produced accuracy of the categorization was checked using a confusion matrix. According to this study's findings, using Particle Swarm Optimization in conjunction with the K Nearest Neighbor algorithm and Z-score normalization can increase accuracy by up to 14%. After implementing Particle Swarm Optimization and Z-Score normalization, the accuracy increased from 68.5% to 72.5%[20].

III. METHODOLOGY

The telecom industry in Mogadishu, Somalia, has a lot of problems with customer churn, or the concept of customers continuously changing service providers. For telecom corporations to sustain a steady client base and guarantee long-term profitability, it is essential to comprehend and anticipate customer attrition. This paper offers a thorough process for creating a machine learning-based, accurate churn prediction system.

This churn prediction system was developed using an approach that includes many important phases. The dataset, which came from Kaggle, contains financial data, contract details, subscription services, and client attributes. In data preparation, duplicates are dealt with, missing values are addressed, categorical variables are encoded, and the data is divided into training, validation, and testing sets. Selecting the best algorithms, such as Support Vector Machine (SVM), Random Forest Classifier, Decision Tree Classifier, and so on, is part of the model selection process.

A. System Architecture

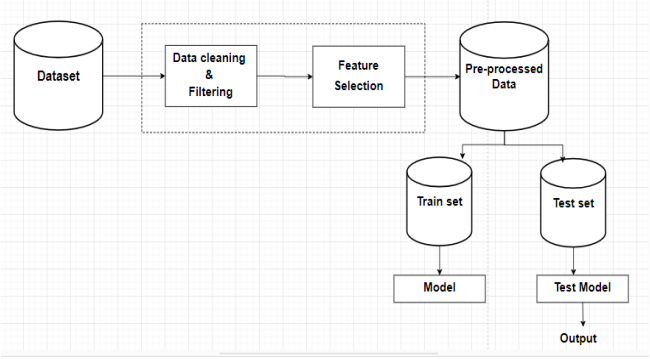

Fig. 1. System Architecture

B. System Features

The system has a powerful customer churn prediction module that uses machine learning techniques like Random Forest and Logistic Regression to determine how likely it is that a client would stop using the service. By providing real-time forecasts derived on customer data input, it improves the capacity to anticipate and mitigate churn threats.

The system also has data visualization capabilities that offer interactive visualizations across several categories, such as churn analysis, correlation charts, and confusion matrices. This facilitates a deeper comprehension and interpretation of the facts.

User control and security are essential system elements that guarantee important predictive functions and data are only accessible by authorized persons. Real-time statistics regarding the processed data, such as the total number of entries, churned customers, and retained customers, are shown on the statistical dashboard. Additionally, users can export data.

In order to guard against SQL injection and guarantee data integrity, the system also uses server-side data validation and secure session handling.

C. System Methodology

In the telecom sector, predicting customer loss using a methodical strategy is essential. The four main steps of this process are Model Acquisition, Model Selection, Model Implementation and Model Evaluation. Telecom firms can successfully detect and handle churn by carefully working through these stages, which will promote customer loyalty and long-term sustainability.

1) Model Acquisition

The dataset used in this study consists of a total of 7,043 rows and 21 columns. Each row in the dataset represents a unique customer and provides information about their characteristics, such as gender, age, partnership status, and dependents. It also includes details about the services they have subscribed to, including phone service, internet service, and various add-ons like online security, online backup, device protection, tech support, streaming TV, and streaming movies. The dataset also captures important contract-related information, such as the type of contract a customer has (month-to-month, one year, or two years), their paperless billing preference, and the payment method they use. Additionally, it includes financial information such as monthly charges and total charges incurred by the customers. To access the dataset, it can be obtained from the platform called Kaggle, which is a popular online community for data science and machine learning.

customer	gender	Senior	Citi	Partner	Dependen	tenure	PhoneSer	Multiple	Li	InternetSe	OnlineSec	OnlineBac	DevicePro	TechSupp	Streaming	Streaming	Contract	Paperless	Payment	MonthlyCl
7590-VHVI	Female	0	Yes	No		1	No	No phone	DSL	No	Yes	No	No	No	No		Month-to-	Yes	Electronic	29.85
5575-GNV	Male	0	No	No		34	Yes	No	DSL	Yes	No	Yes	No	No	No		One year	No	Mailed ch	56.95
3668-QPYI	Male	0	No	No		2	Yes	No	DSL	Yes	Yes	No	No	No	No		Month-to-	Yes	Mailed ch	53.85
7795-CFOI	Male	0	No	No		45	No	No phone	DSL	Yes	No	Yes	Yes	No	No		One year	No	Bank trans	42.3
9237-HQIT	Female	0	No	No		2	Yes	No	Fiber optic	No	No	No	No	No	No		Month-to-	Yes	Electronic	70.7
9305-CDSI	Female	0	No	No		8	Yes	Yes	Fiber optic	No	No	Yes	Yes	No	Yes		Month-to-	Yes	Electronic	99.65
1452-KIOV	Male	0	No	Yes		22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No		Month-to-	Yes	Credit car	89.1
6713-OKO	Female	0	No	No		10	No	No phone	DSL	Yes	No	No	No	No	No		Month-to-	No	Mailed ch	29.75
7892-POO	Female	0	Yes	No		28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes		Month-to-	Yes	Electronic	104.8
6388-TABI	Male	0	No	Yes		62	Yes	No	DSL	Yes	Yes	No	No	No	No		One year	No	Bank trans	56.15
9763-GRSI	Male	0	Yes	Yes		13	Yes	No	DSL	Yes	No	No	No	No	No		Month-to-	Yes	Mailed ch	49.95
7469-LKBC	Male	0	No	No		16	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne		Two year	No	Credit car	18.95
8091-TTVA	Male	0	Yes	No		58	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes		One year	No	Credit car	100.35
0280-XJGE	Male	0	No	No		49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes		Month-to-	Yes	Bank trans	103.7
5129-JLPI	Male	0	No	No		25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes		Month-to-	Yes	Electronic	105.5
3655-SNQ	Female	0	Yes	Yes		69	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes		Two year	No	Credit car	113.25
8191-XWS	Female	0	No	No		52	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne		One year	No	Mailed ch	20.65
9959-WOF	Male	0	No	Yes		71	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes		Two year	No	Bank trans	106.7
4190-MFL	Female	0	Yes	Yes		10	Yes	No	DSL	No	No	Yes	Yes	No	No		Month-to-	No	Credit car	55.2
4183-MYF	Female	0	No	No		21	Yes	No	Fiber optic	No	Yes	Yes	No	No	Yes		Month-to-	Yes	Electronic	90.05

Fig. 2. Dataset

Data preparation, also known as data pre-processing, is the process of cleaning, organizing, and transforming raw data into a format suitable for analysis or training machine learning models. It is a crucial step in the data analysis and machine learning pipeline, as the quality of the input data significantly impacts the performance and effectiveness of the models.

The main goals of our data preparation include:

• Handling Missing Data: Identifying and addressing missing values in the dataset. This can involve removing rows with missing values, imputing missing values based on statistical measures, or using advanced techniques to predict missing values.

• Handling Data Duplicates: Identifying and handling duplicate records in the dataset to avoid biases in model training and evaluation.

• Encoding Categorical Variables: Converting categorical variables (non-numeric) into a numerical format that machine learning models can understand. Common methods include one-hot encoding, label encoding, or using embeddings for more complex categorical data.

• Data Splitting: Dividing the dataset into training, validation, and testing sets. The training set is used to train the model, the validation set is used for fine-tuning and hyperparameter tuning, and the testing set is used to evaluate the model's performance on unseen data.

2) Model Selection

In customer churn prediction for telecommunications, we utilize a range of machine learning models, namely Decision Tree Classifier, Random Forest Classifier, Support Vector Machine (SVM), and Logistic Regression. These models are employed to forecast and identify customers who are likely to churn, enabling proactive measures to retain their loyalty.

a) Decision Tree Classifier:

A decision tree is a flowchart-like structure where internal nodes represent features or attributes, branches represent decisions, and leaf nodes represent outcomes. In this case, a decision tree classifier uses historical customer data to create a tree-like model that can predict whether a customer is likely to churn or not. It makes decisions based on a series of questions about the customer's characteristics, such as usage patterns, demographics, or service preferences.

b) Random Forest Classifier:

A random forest classifier is an ensemble learning method that combines multiple decision trees to make predictions. It generates a collection of decision trees, each trained on a random subset of the data and a random subset of features. When predicting churn, each decision tree in the random forest independently predicts the outcome, and the final prediction is determined by majority voting or averaging the individual tree predictions.

c) Support Vector Machine (SVM):

SVM is a supervised machine learning algorithm used for classification tasks. It works by finding an optimal hyperplane that separates different classes in a high-dimensional feature space. In the context of churn prediction, SVM can be trained on historical customer data, with features such as call duration, customer tenure, or usage patterns. SVM aims to find a decision boundary that maximizes the margin between churned and non-churned customers, allowing it to make predictions for new customers.

d) Logistic Regression:

Logistic regression is a statistical model used for binary classification problems, such as churn prediction. It models the relationship between the dependent variable (churn) and one or more independent variables (customer attributes). Logistic regression estimates the probabilities of churn by applying the logistic function to a linear combination of the input features. The resulting probabilities can then be used to classify customers as likely to churn or not, based on a predefined threshold.

3) Model Implementation

The process of developing, deploying, and transforming the system on for end users is known as implementation. In order to determine if a customer is churning or not, it involves integrating the learned machine learning models into a useful application that customers are able to interact with.

4) Model Evaluation

When we're trying to predict if customers will leave a telecom company, we use four different methods. First, we look at Decision Trees to see how well they spot potential leavers and understand their reasons. Then, we check Random Forests to see their overall accuracy and how they handle complex info. Next, Support Vector Machines (SVM) help us see if they can draw a line between staying and leaving customers accurately. We use terms like accuracy, precision, and recall to measure this. Lastly, Logistic Regression helps us estimate how likely someone is to leave and how easy it is to understand the reasons. By doing this, we find the best way to predict customer churn in telecom.

IV. PROPOSED WORK

First, we will download the dataset from Kaggle. Then, we will filter the data to find any null values. After that, we transformed all of the data into a format that was easier to comprehend and evaluate. Using Decision Tree Classifier and logistic regression, we attempt to develop a predictor model for the telecom firm. Here, we have a customer data set that we split into training and testing sections through feature selection and preprocessing. We have performed some feature engineering on this method in order to achieve more accurate and efficient results.

We are able to calculate the probability of happening at occurrence areas and create an exclusive probability classification with the use of Decision Tree Classifier. After the model has been trained, the data produces a result with all of its information, and we will use the remaining data to test the model. As a result, we will obtain an accuracy from the results that will allow us to forecast the customer.

churn and can provide a clear warning about the client, which can assist the business in taking certain actions to help prevent the current customer from leaving the service. (20% of the data were used for testing, and 80% for training).

In order to enable the model to learn from past data, we will attempt to fit the ytrain data and ytest data to the model fit by obtaining both findings. This involves using periods of time to force the model to repeatedly learn the same set of data. We displayed the performance of the models using the Bar chart, which allowed us to determine the model correctness of the generated data and produce a churn forecast (Fig. 11).

V. RESULTS

We tested our tool many times to make sure it works well. We used a lot of data from past customers to see if our tool could correctly predict who would leave. This included people from different backgrounds and with different usage patterns. We made sure the tool was tested in conditions that it would face in the real world, ensuring it can handle the complexities of an actual market environment. Our results were very promising. The model decision tree could predict with about 95% accuracy whether a customer would stay or leave. This means it was right 9.5 out of 10 times, which is very good for this kind of tool. We also measured how well it could identify customers who would definitely leave (sensitivity) and customers who would definitely stay (specificity), and the results were similarly high. Our findings show that telecom companies in Mogadishu could save a lot of money with this tool. By knowing who might

leave, they can try to keep these customers with special deals or improved services. This is cheaper than trying to find new customers. If our tool reduces customer loss by even 5%, it could mean millions saved in revenue.

```
print(classification_report(ylog_test, ylog_predict))
```

	precision	recall	f1-score	support
0	0.95	0.92	0.94	510
1	0.93	0.96	0.94	556
accuracy			0.94	1066
macro avg	0.94	0.94	0.94	1066
weighted avg	0.94	0.94	0.94	1066

Fig. 3. Logistic Regression Classification report for a balanced dataset

```
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.79	0.99	0.88	917
1	0.62	0.08	0.14	258
accuracy			0.79	1175
macro avg	0.71	0.53	0.51	1175
weighted avg	0.76	0.79	0.72	1175

Fig. 4. Logistic Regression Classification report for imbalanced dataset

```
print(classification_report(ydt_test, ydt_predict))
```

	precision	recall	f1-score	support
0	0.96	0.93	0.95	481
1	0.94	0.97	0.96	577
accuracy			0.95	1058
macro avg	0.95	0.95	0.95	1058
weighted avg	0.95	0.95	0.95	1058

Fig. 5. Decision Tree Classification for balanced dataset

```
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.87	0.90	0.88	917
1	0.59	0.50	0.54	258
accuracy			0.81	1175
macro avg	0.73	0.70	0.71	1175
weighted avg	0.81	0.81	0.81	1175

Fig. 6. Decision Tree Classification report for imbalanced dataset

```
print(classification_report(yr_test1, yr_predict1))
```

	precision	recall	f1-score	support
0	0.96	0.89	0.92	501
1	0.91	0.96	0.93	557
accuracy			0.93	1058
macro avg	0.93	0.93	0.93	1058
weighted avg	0.93	0.93	0.93	1058

Fig. 7. Random Forest Classification for balanced dataset

```
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.85	0.95	0.90	917
1	0.68	0.40	0.50	258
accuracy			0.83	1175
macro avg	0.77	0.67	0.70	1175
weighted avg	0.81	0.83	0.81	1175

Fig. 8. Random Forest Classification report for imbalanced dataset

```
print(classification_report(ysvm_test1, ysvm_predict1))
```

	precision	recall	f1-score	support
0	0.71	0.80	0.75	488
1	0.81	0.71	0.76	568
accuracy			0.75	1056
macro avg	0.76	0.76	0.75	1056
weighted avg	0.76	0.75	0.75	1056

Fig. 9. Random Forest Classification report for balanced dataset

```
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.79	0.99	0.88	917
1	0.62	0.08	0.14	258
accuracy			0.79	1175
macro avg	0.71	0.53	0.51	1175
weighted avg	0.76	0.79	0.72	1175

Fig. 10. SVM Classification report for imbalanced dataset

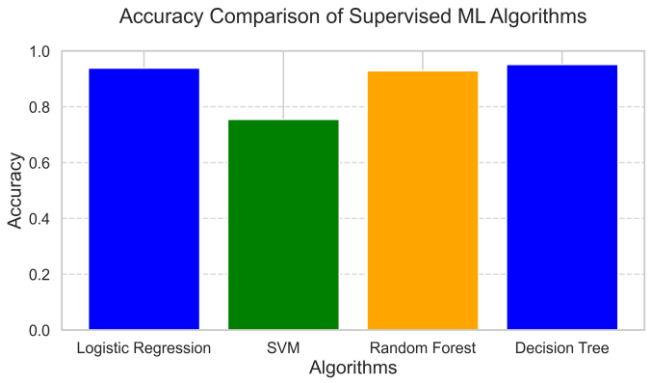

Fig. 11. Accuracy Comparison of Models

Four supervised machine learning algorithms are compared in the bar chart: Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression. Each bar shows the accuracy of a different algorithm; of the four, SVM has the lowest accuracy, Random Forest is the one that follows it most closely, and Logistic Regression and Decision Tree have the best accuracy. The accuracy ratings are displayed on the y-axis, and the algorithms are listed on the x-axis.

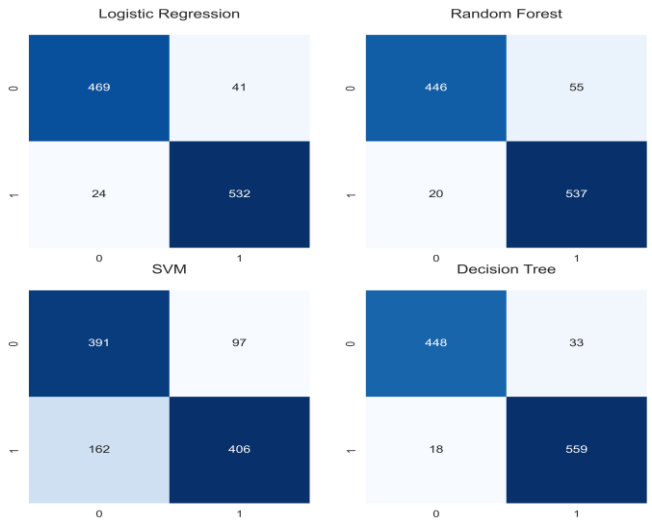

Fig. 12. Confusion Matrices

Four supervised machine learning algorithms' confusion matrices are shown in the image: Decision Tree, Random Forest, SVM, and Logistic Regression. The numbers for true positives, true negatives, false positives, and false negatives for each algorithm are displayed in each matrix, along with the labels 0 and 1, which represent successfully and wrongly identified instances of two classes, respectively. When compared to SVM and Random Forest, Decision Tree and Logistic Regression show higher true positive and true negative counts, indicating superior performance.

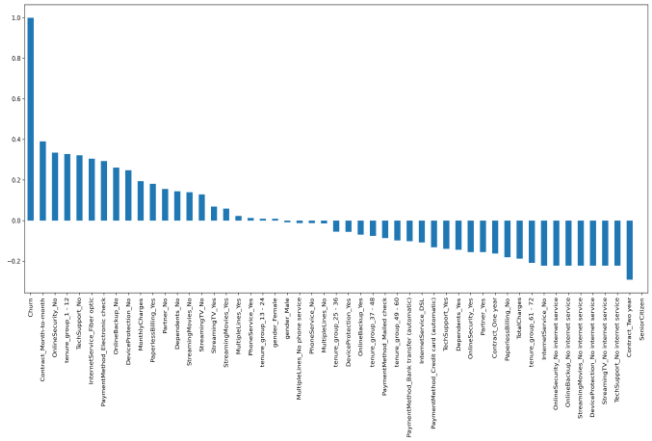

Fig. 13. corelation of all predictors with Churn

The correlation coefficients between different predictors and customer churn in a dataset are displayed in a bar chart. The y-axis displays the correlation values between the predictors, which range from -0.2 to 1.0, and the x-axis lists the predictors. Important determinants that show positive relationships with churn include "Contract_Month-to-month" and "OnlineSecurity_No," whereas "Contract_Two year" and "SeniorCitizen" show negative associations. Understanding the factors that contribute to customer churn is made easier by the chart, which shows which characteristics are most strongly linked to it.

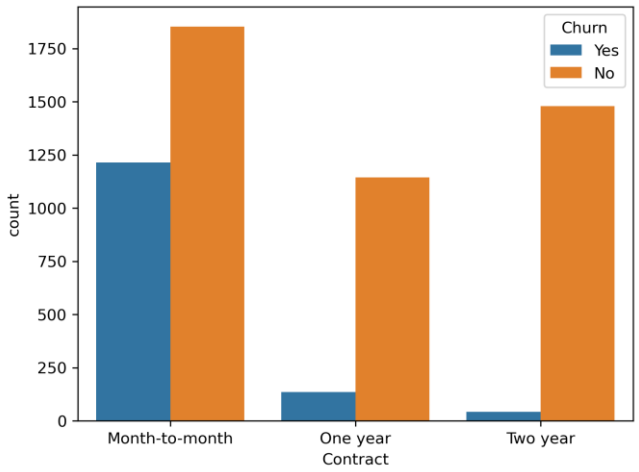

Fig. 14. Contract

The distribution of customer churn by contract type—month-to-month, one-year, and two-year—is depicted in the bar chart. The contract type is indicated by the x-axis, while the number of clients is represented by the y-axis. In contrast to consumers with one-year and two-year contracts, which have noticeably lower churn rates, those with month-to-month contracts had a greater churn rate (blue bars) in the chart.

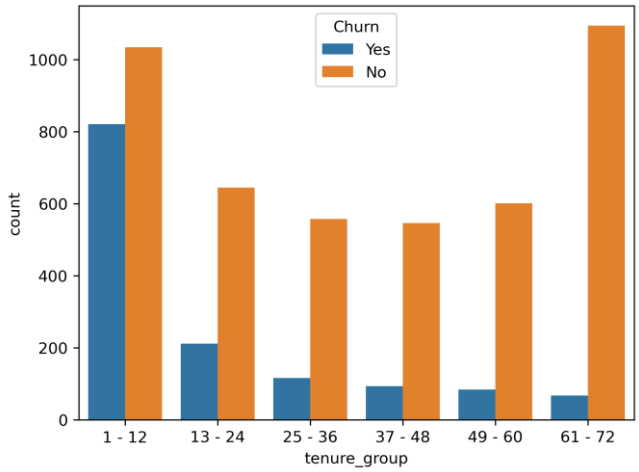

Fig. 15. Tenure group

The y-axis shows the number of customers, and the x-axis shows the tenure groups (in months) for the bar chart, which illustrates customer churn across various tenure groups. The turnover rate (blue bars) is higher for customers with shorter tenures (1–12 months) than it is for those with longer tenures (substantially lower). Customers in the 61–72-month tenure group have the lowest churn rate, suggesting that long-term customers have a greater retention rate.

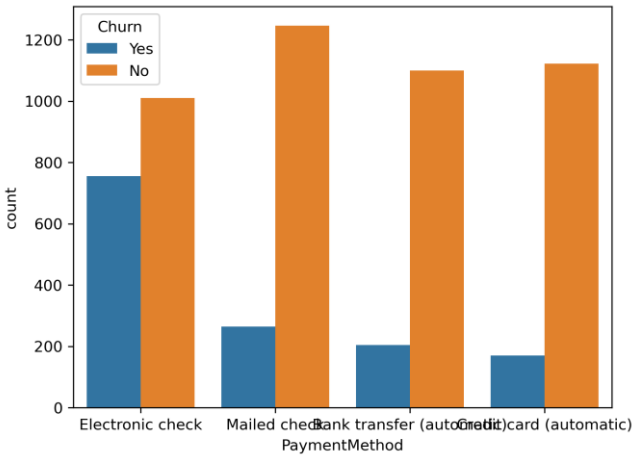

Fig. 16. Payment method

Customer churn is shown in the bar chart according to the following payment methods: credit card (automated), bank transfer (automatic), mailed check, and electronic check. The number of consumers is shown on the y-axis, while the payment options are listed on the x-axis. Consumers who use bank transfers and credit cards—both of which are automatic—had the lowest churn rates, while those who use electronic checks have the highest (blue bars), suggesting that customers who use automated payment methods are more likely to remain loyal.

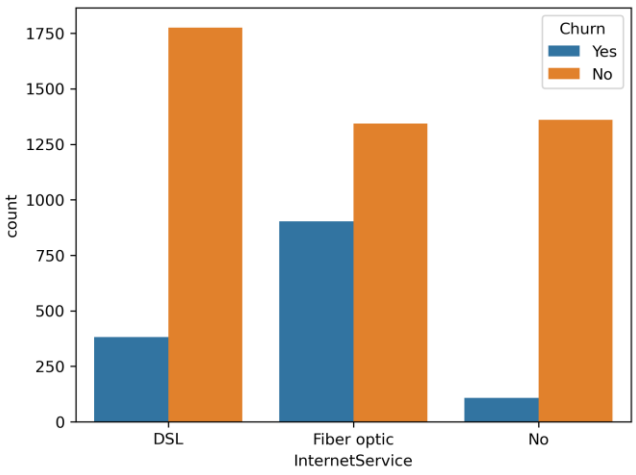

Fig. 17. Internet service

Based on the kind of internet service—DSL, fiber optic, and no internet service—the bar chart displays customer churn. The x-axis indicates the different categories of internet services, while the y-axis shows the number of customers. Clients with DSL service are the ones with the lowest churn rate (blue bars), next those with fiber optic service. Clients without internet access have the lowest rate of customer attrition, suggesting a greater level of customer retention.

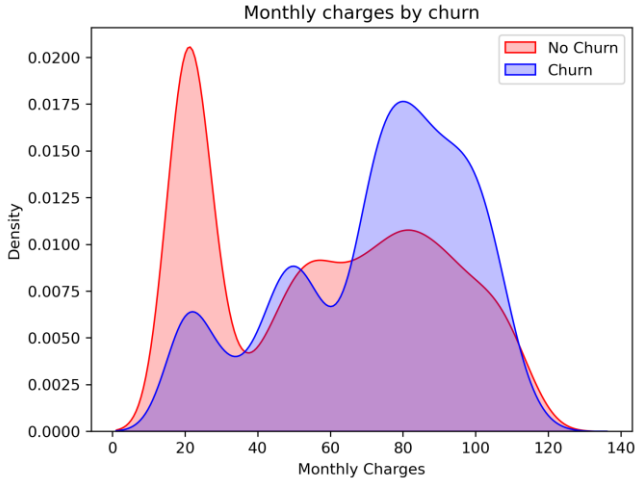

Fig. 18. Churn by Monthly Charges

The monthly fee distribution between churning customers (blue line) and non-churning customers (red line) is depicted in the density plot. The density is represented by the y-axis, while the monthly charges are displayed on the x-axis. Consumers who do not churn typically have lower monthly charges, with a notable peak at $20–$25, whereas customers who do churn typically have higher monthly charges, as seen by the blue peak around $70-$100. This implies a relationship between higher monthly fees and higher rates of customer attrition.

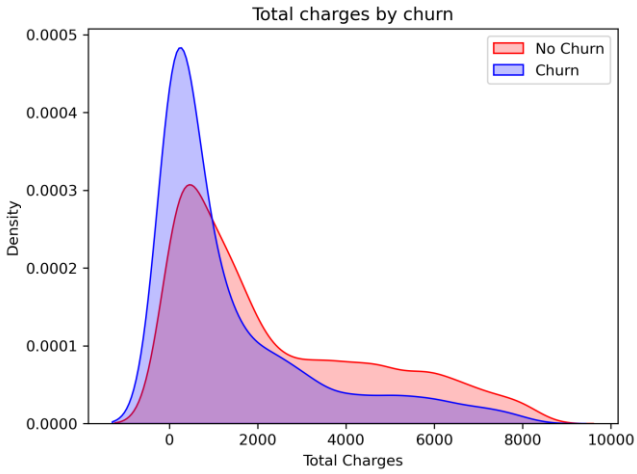

Fig. 19. Churn by Total Charges

The distribution of total charges between churning customers (blue line) and non-churning customers (red line) is depicted in the density plot. The density is shown on the y-axis, and the total charges are shown on the x-axis. The blue peak surrounding lower total charge values indicates that customers with lower total charges are more likely to experience customer attrition. The red line, on the other hand, peaks at higher total charge levels and taper down more gradually, indicating that consumers with greater total costs are less likely to churn.

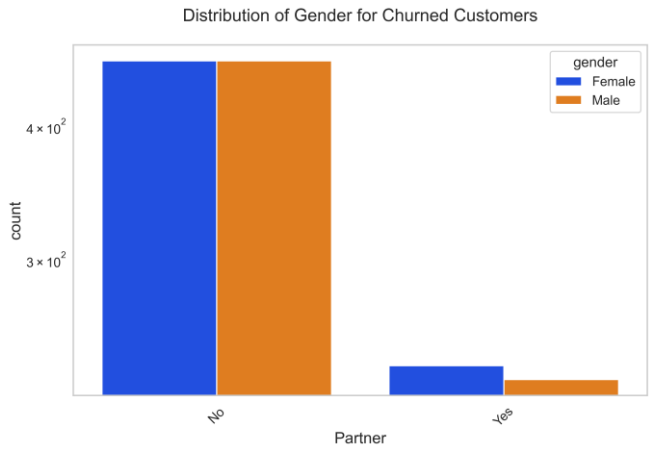

Fig. 20. Distribution of Gender for Churned Customers

The distribution of churned customers by gender (female in blue, male in orange) and whether or not they have a partner is depicted in the bar chart. The number of customers who have left is represented by the y-axis, and the partner status (Yes or No) is represented by the x-axis. According to the statistics, about equal numbers of males and females leave a business because they do not have a partner. Both genders' turnover rates are much lower for consumers who have a partner.

VI. CONCLUSION

Our research successfully developed a system that uses machine learning to predict when telecom customers in Mogadishu might stop using their services. The main goal was to create a tool that is both efficient and accurate in identifying customers likely to leave. Our system was trained with a lot of different data, including how long customers stay with the company, their service usage. This broad set of data helped the system learn better and make more accurate predictions. Furthermore, the results of our research demonstrated that our system significantly outperforms traditional predictive methods. Where previously companies relied on heuristic or simplified statistical methods, our machine learning-based approach provides a more dynamic and robust framework. It adapts to new data, continually improving its predictions as more information becomes available. This adaptability is particularly important in a rapidly changing market like Mogadishu's telecommunications sector. The results showed that our system is really good at predicting churn—better than the old methods where humans had to guess based on less information. This suggests that our tool could be a big help to telecom companies, giving them a way to identify at-risk customers early so they can try to keep them.

REFERENCES

[1] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J Big Data*, vol. 6, no. 1, p. 28, Dec. 2019, doi: 10.1186/s40537-019-0191-6.

[2] N. I. Mohammad, S. A. Ismail, M. N. Kama, O. M. Yusop, and A. Azmi, "Customer Churn Prediction In Telecommunication Industry Using Machine Learning Classifiers," in *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*, Vancouver BC Canada: ACM, Aug. 2019, pp. 1–7. doi: 10.1145/3387168.3387219.

[3] V. Kavitha, G. H. Kumar, S. M. Kumar, and M. Harish, "Churn prediction of customer in telecom industry using machine learning algorithms," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 5, pp. 181–184, 2020.

[4] V. Mahajan, R. Misra, and R. Mahajan, "Review on factors affecting customer churn in telecom sector," *IJDATS*, vol. 9, no. 2, p. 122, 2017, doi: 10.1504/IJDATS.2017.085898.

[5] P. Bhuse, A. Gandhi, P. Meswani, R. Muni, and N. Katre, "Machine learning based telecom-customer churn prediction," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE, 2020, pp. 1297–1301. Accessed: Feb. 04, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9315951/

[6] S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. H. Gawande, "Customer churn prediction in telecom sector using machine learning techniques," *Results in Control and Optimization*, vol. 14, p. 100342, 2024.

[7] S. Babu, D. N. Ananthanarayanan, and V. Ramesh, "A survey on factors impacting churn in telecommunication using datamininig techniques," *International Journal of Engineering Research & Technology (IJERT)*, vol. 3, no. 3, 2014, Accessed: Feb. 05, 2024. [Online]. Available: https://www.academia.edu/download/64723522/a_survey_on_factors_impacting_churn_IJERTV3IS031583.pdf

[8] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414–1425, 2012.

[9] B. Q. Huang, T.-M. Kechadi, B. Buckley, G. Kiernan, E. Keogh, and T. Rashid, "A new feature set with new window techniques for customer churn prediction in land-line telecommunications," *Expert Systems with Applications*, vol. 37, no. 5, pp. 3657–3665, 2010.

[10] Ö. Çelik and U. Osmanoğlu, "Comparing to Techniques Used in Customer Churn Analysis," Jul. 2019.

[11] N. Hashmi, N. A. Butt, and M. Iqbal, "Customer churn prediction in telecommunication a decade review and classification," *International Journal of Computer Science Issues (IJCSI)*, vol. 10, no. 5, p. 271, 2013.

[12] V. Umayaparvathi and K. Iyakutti, "Applications of data mining techniques in telecom churn prediction," *International Journal of Computer Applications*, vol. 42, no. 20, pp. 5–9, 2012.

[13] G. Xia and W. Jin, "Model of customer churn prediction on support vector machine," *Systems Engineering-Theory & Practice*, vol. 28, no. 1, pp. 71–77, 2008.

[14] A. O. Oyeniyi, A. B. Adeyemo, A. O. Oyeniyi, and A. B. Adeyemo, "Customer churn analysis in banking sector using data mining techniques," *Afr J Comput ICT*, vol. 8, no. 3, pp. 165–174, 2015.

[15] C.-F. Tsai and Y.-H. Lu, “Customer churn prediction by hybrid neural networks,” *Expert Systems with Applications*, vol. 36, no. 10, pp. 12547–12553, 2009.

[16] S. Zhang, C. Zhang, and Q. Yang, “Data preparation for data mining,” *Applied Artificial Intelligence*, vol. 17, no. 5–6, pp. 375–381, May 2003, doi: 10.1080/713827180.

[17] S. A. Bin-Nashwan and H. Hassan, “Impact of customer relationship management (CRM) on customer satisfaction and loyalty: A systematic review,” *Journal of Advanced Research in Business and Management Studies*, vol. 6, no. 1, pp. 86–107, 2017.

[18] H. Sayed, M. A. Abdel-Fattah, and S. Kholief, “Predicting potential banking customer churn using apache spark ML and MLlib packages: a comparative study,” *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 11, 2018, Accessed: Aug. 07, 2024. [Online]. Available: https://www.researchgate.net/profile/Manal-Abdel-Fattah-2/publication/329427276_Predicting_Potential_Banking_Customer_Churn_using_Apache_Spark_ML_and_MLlib_Packages_A_Comparative_Study/links/5c08086f4585157ac1aaf58e/Predicting-Potential-Banking-Customer-Churn-using-Apache-Spark-ML-and-MLlib-Packages-A-Comparative-Study.pdf

[19] S. Maheshwari, R. C. Jain, and R. S. Jadon, “A review on class imbalance problem: Analysis and potential solutions,” *International journal of computer science issues (IJCSI)*, vol. 14, no. 6, pp. 43–51, 2017.

[20] M. A. Imron and B. Prasetyo, “Improving algorithm accuracy k-nearest neighbor using z-score normalization and particle swarm optimization to predict customer churn,” *Journal of Soft Computing Exploration*, vol. 1, no. 1, pp. 56–62, 2020.