# PREDICTING CUSTOMER CHURN & CUSTOMER VALUE: A MACHINE LEARNING APPROACH

**C. U. AMARASINGHA**

# TABLE OF CONTENTS

# TABLE OF FIGURES

# 1. INTRODUCTION

In today's highly competitive business landscape, understanding and mitigating customer churn is crucial for maintaining a stable and profitable customer base. Customer churn refers to the phenomenon where customers stop doing business with a company over a given period. High churn rates can significantly impact a company's revenue and growth potential, making it essential for businesses to identify the underlying causes and predict which customers are at risk of leaving.

This report focuses on the application of classification techniques to predict customer churn. The aim is to develop a model that can accurately identify customers who are likely to churn based on historical data.

By accurately predicting customer churn, companies can allocate resources more efficiently, tailor their marketing efforts, and develop personalized interventions to retain valuable customers.

# 2. CLASSIFICATION

This report discusses the classification model developed in various stages, including data collection, preprocessing, model training, hyperparameter tuning, and evaluation. Several machine learning algorithms, such as logistic regression, random forests, Gradient Boost, and stacking, are employed and compared to determine the most effective model for predicting churn. For this analysis, a laptop with 16GB RAM and a GPU was used.

## 2.1 Data Selection

The dataset for this task was sourced from the UC Irvine Machine Learning Repository (UCI Machine Learning Repository, 2020). It's a dataset that is randomly collected from an Iranian telecom company over 12 months. It provides 3150 records of customer data which can be used for prediction. An overview is shown in Figure 1.

| | Call Failure | Complains | Subscription Length | Charge Amount | Seconds of Use | Frequency of use | Frequency of SMS | Distinct Called Numbers | Age Group | Tariff Plan | Status | Age | Customer Value | Churn |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 8 | 0 | 38 | 0 | 4370 | 71 | 5 | 17 | 3 | 1 | 1 | 30 | 197.640 | 0 |
| 1 | 0 | 0 | 39 | 0 | 318 | 5 | 7 | 4 | 2 | 1 | 2 | 25 | 46.035 | 0 |
| 2 | 10 | 0 | 37 | 0 | 2453 | 60 | 359 | 24 | 3 | 1 | 1 | 30 | 1536.520 | 0 |
| 3 | 10 | 0 | 38 | 0 | 4198 | 66 | 1 | 35 | 1 | 1 | 1 | 15 | 240.020 | 0 |
| 4 | 3 | 0 | 38 | 0 | 2393 | 58 | 2 | 33 | 1 | 1 | 1 | 15 | 145.805 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 3145 | 21 | 0 | 19 | 2 | 6697 | 147 | 92 | 44 | 2 | 2 | 1 | 25 | 721.980 | 0 |
| 3146 | 17 | 0 | 17 | 1 | 9237 | 177 | 80 | 42 | 5 | 1 | 1 | 55 | 261.210 | 0 |
| 3147 | 13 | 0 | 18 | 4 | 3157 | 51 | 38 | 21 | 3 | 1 | 1 | 30 | 280.320 | 0 |
| 3148 | 7 | 0 | 11 | 2 | 4695 | 46 | 222 | 12 | 3 | 1 | 1 | 30 | 1077.640 | 0 |
| 3149 | 8 | 1 | 11 | 2 | 1792 | 25 | 7 | 9 | 3 | 1 | 1 | 30 | 100.680 | 1 |

3150 rows × 14 columns

*Figure 1: Dataset Overview*

## 2.2 Exploratory Data Analysis

An EDA was carried out to gain a clear understanding of the data. The plot in Figure 2 shows the distribution of churn customers on the basis of complaints. From this, it can be seen that customer churn is higher from customers who have complained. This can be inferred as those who complain regarding the service are more likely to stop mobile services.

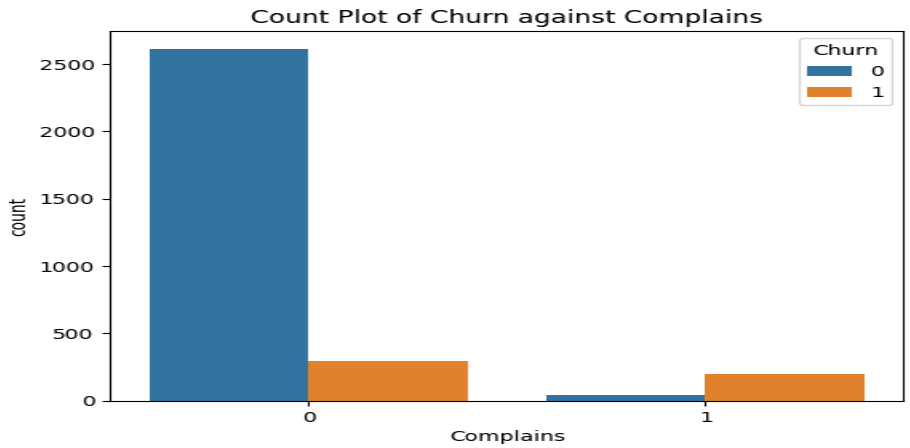


*Figure 2: Churn Distribution Among Complaints*

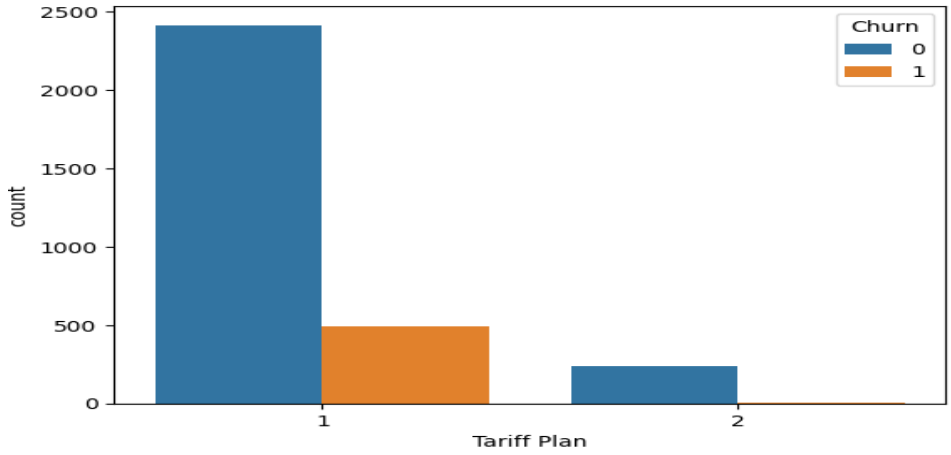


*Figure 3: Churn Distribution among Tariff Plan*

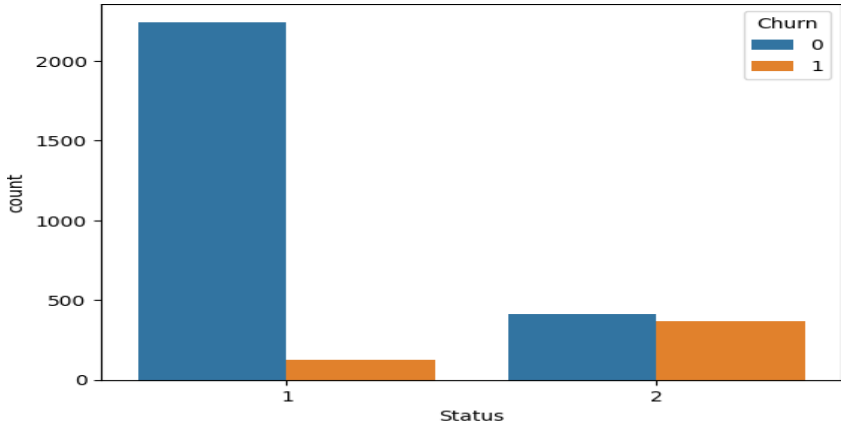


*Figure 4: Customer Distribution Against Status*

Figure 3 shows the distribution of customers among tariff plans. It can be seen that customers who stop services mostly use the Pay As You Go Plan. Figure 4 also indicates that the status also influences the customer attrition as churning is more frequent in inactive customers.

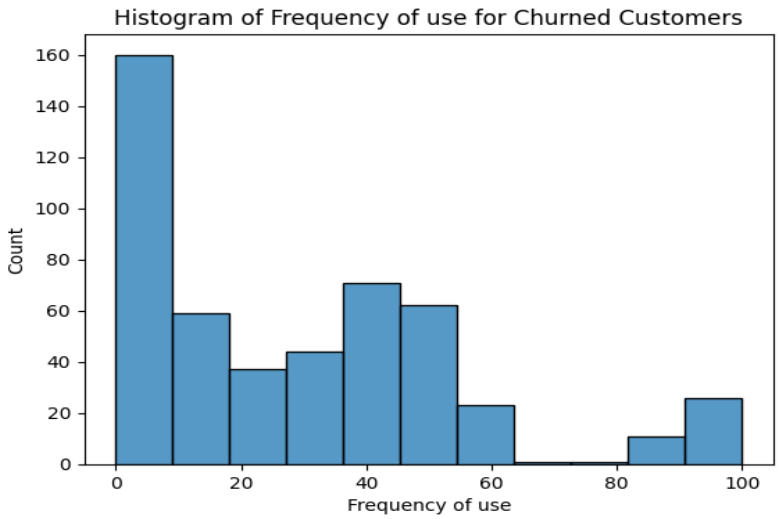


*Figure 5: Histogram of Frequency of Use for Churned Customers*

Figure 5 shows the distribution of frequency of use for customers who have stopped services. It can be seen that churned customers are highest in the lower categories indicating that the lower the use, the higher the likelihood of customer attrition. Figure 6 shows how customer value is distributed for both types of customers. Lower customer value is seen for churned customers compared to existing customers.

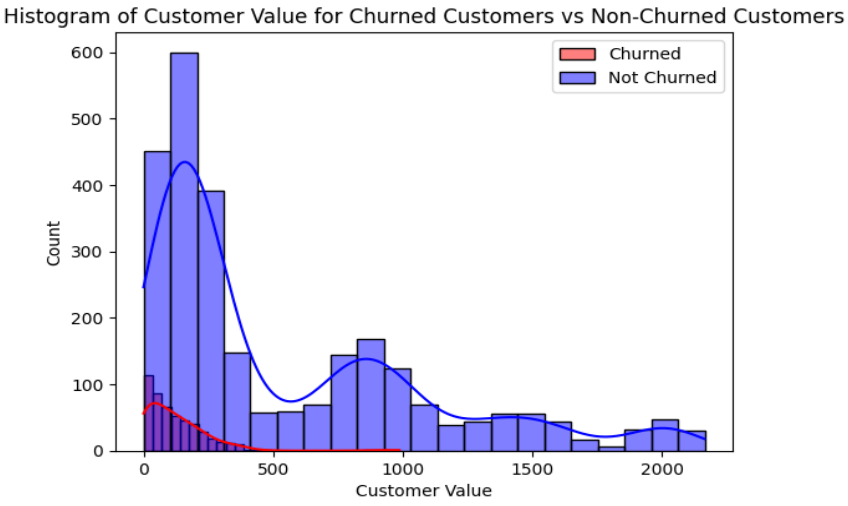


*Figure 6: Histogram of Customer Value*

In consideration of the EDA, it can be seen that several features such as ‘complains’, ‘tariff plan’ and ‘frequency of use’ influence the customer behaviour.

## 2.3 Data Preprocessing

Before models are developed, the data must be cleaned to ensure the accuracy and integrity of it. This helps in building better models with reliable predictive power. First, each feature was examined to identify its data type. There are several features such as ‘complains’, ‘age group’, and ‘tariff plan’ which are categorical in nature. These were converted to ensure they were recognized as such. Next, the dataset was checked for missing values. As there were no missing values, no treatment was applied in this case.

After, this summary statistics were generated which is shown in Figure 7. From this, it can be seen that the numerical features are in different scales. For example, “seconds of use” are in 1000s while the ‘frequency of use’ are in 100s. They also have different units. Using these features as it is can affect the performance of models through overfitting. Additionally, the boxplots shown in Figure 8, show many outliers present. This too can affect the performance of the model through overfitting. It was decided to leave the outliers as it is and address them through standardization. Robust Scaler was chosen for this due to its robustness against outliers. It uses the median and interquartile range in scaling.

| | Call Failure | Subscription Length | Seconds of Use | Frequency of use | Frequency of SMS | Distinct Called Numbers | Age | Customer Value | Churn |
|---|---|---|---|---|---|---|---|---|---|
| count | 3150.000000 | 3150.000000 | 3150.000000 | 3150.000000 | 3150.000000 | 3150.000000 | 3150.000000 | 3150.000000 | 3150.000000 |
| mean | 7.627937 | 32.541905 | 4472.459683 | 69.460635 | 73.174921 | 23.509841 | 30.998413 | 470.972916 | 0.157143 |
| std | 7.263886 | 8.573482 | 4197.908687 | 57.413308 | 112.237560 | 17.217337 | 8.831095 | 517.015433 | 0.363993 |
| min | 0.000000 | 3.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 15.000000 | 0.000000 | 0.000000 |
| 25% | 1.000000 | 30.000000 | 1391.250000 | 27.000000 | 6.000000 | 10.000000 | 25.000000 | 113.801250 | 0.000000 |
| 50% | 6.000000 | 35.000000 | 2990.000000 | 54.000000 | 21.000000 | 21.000000 | 30.000000 | 228.480000 | 0.000000 |
| 75% | 12.000000 | 38.000000 | 6478.250000 | 95.000000 | 87.000000 | 34.000000 | 30.000000 | 788.388750 | 0.000000 |
| max | 36.000000 | 47.000000 | 17090.000000 | 255.000000 | 522.000000 | 97.000000 | 55.000000 | 2165.280000 | 1.000000 |

*Figure 7: Summary Statistics*

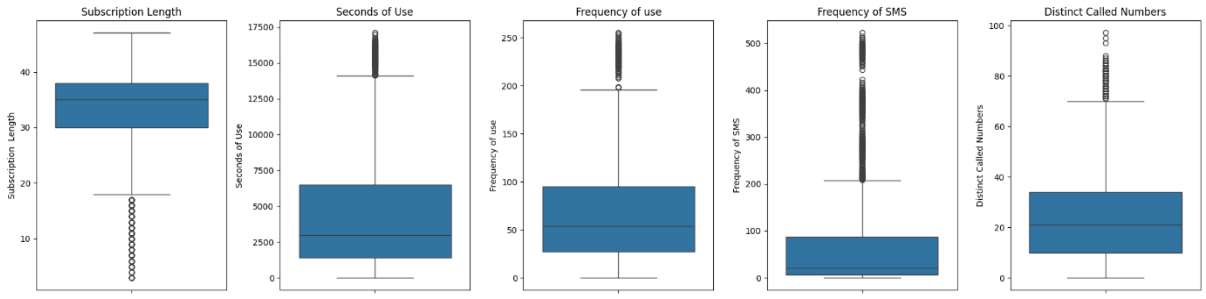


*Figure 8: Boxplots of Numerical Features*

Next, pairwise correlation was compared for each feature which is shown in Figure 9 and Figure 10.

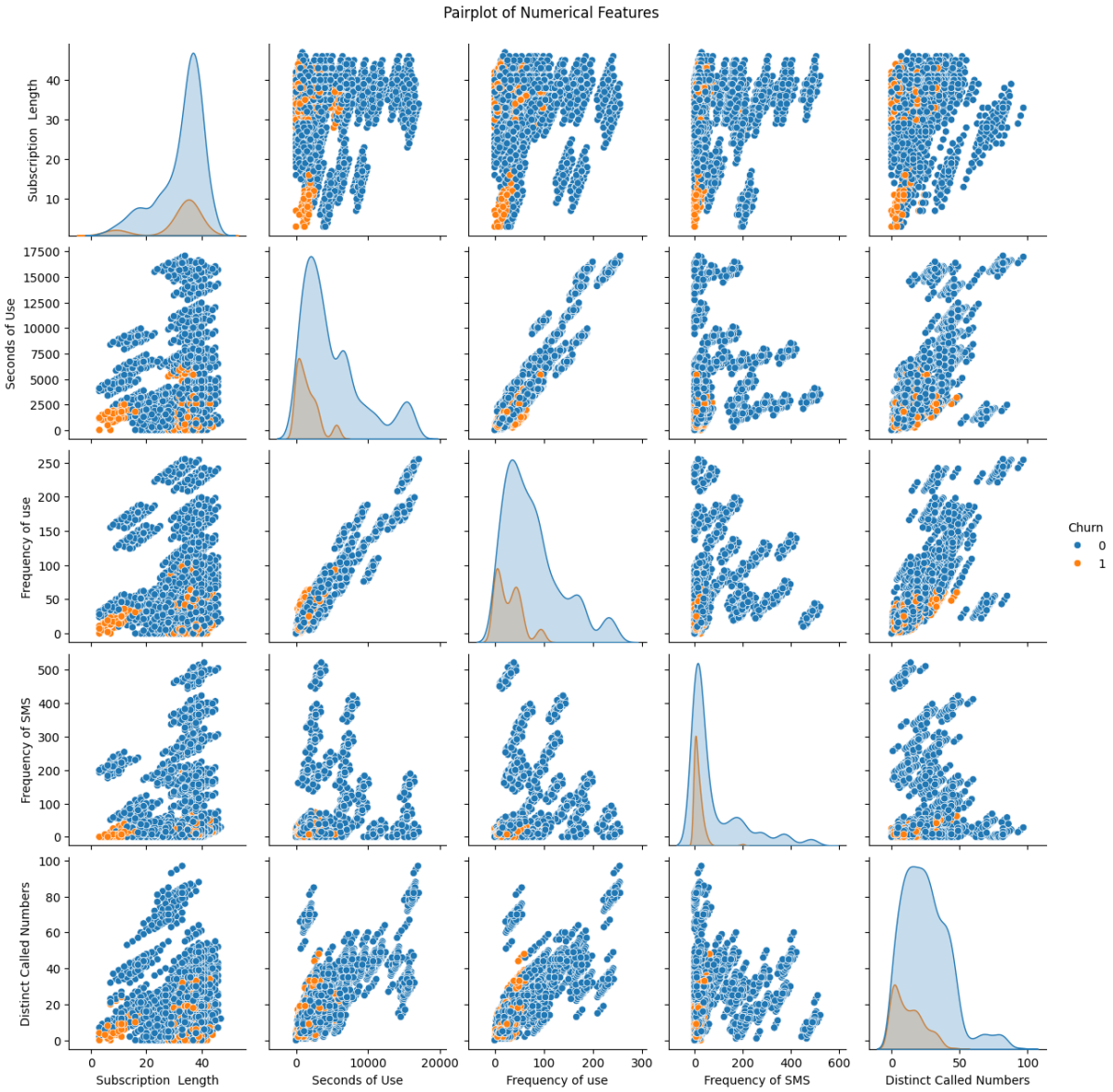


*Figure 9: Pairplots of Numerical Features*

The scatterplots show varying relationships between the features. A strong positive relationship can be observed between 'frequency of use' and 'seconds of use'. This can be verified by observing the correlation matrix. It also shows that most of the variables are skewed; often to the right with the exception of 'subscription length'.

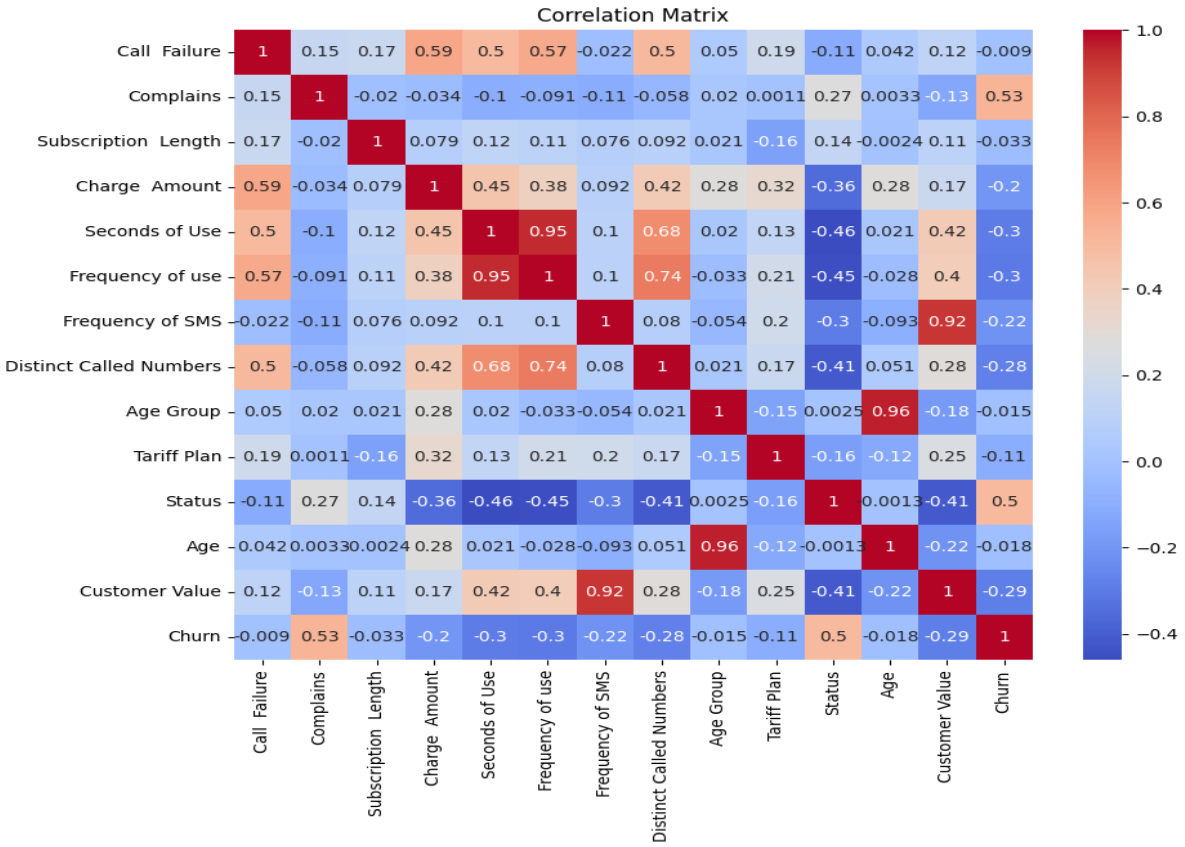


*Figure 10: Correlation Matrix*

The correlation matrix shows varying correlations between the features. It can be observed that the highest correlation for churn is seen with the ‘complains’ and ‘status’ features. Both show a positive relationship indicating that if a customer is inactive within the 9 months or has made complaints, they are highly likely to end services. Additionally, if a customer frequently uses the mobile service, they are less likely to end services (-0.3 correlation). Moreover, we can observe highly correlated features such as the ‘seconds of use’ and ‘frequency of use’ pair, ‘age’ and ‘age group’ pair, and ‘frequency of sms’ and ‘customer value’. When certain features are highly correlated, the additional feature doesn’t contribute new information. Therefore, having highly correlated features increases the variance and noise of the dataset. To avoid this, the features of ‘age group’ and ‘seconds of use’ are dropped.

Certain algorithms cannot handle categorical variables directly. They must be converted to a numerical output in these cases. In this task, all algorithms used except Logistic Regression are capable of handling categorical variables directly. However, categorical feature encoding is unnecessary in this case as all features (even ones recognized as categorical) are in a numerical format. Thus, no separate encoding was performed.

Finally, the dataset was examined for class imbalance. This is shown in Figure 11. It can be seen that the dataset is highly imbalanced with fewer instances for the churned customers. This imbalance could negatively affect the performance model where they would perform well with the majority class but perform poorly with the minority class. It was decided to address this data imbalance by adjusting the class weights in each algorithm used.

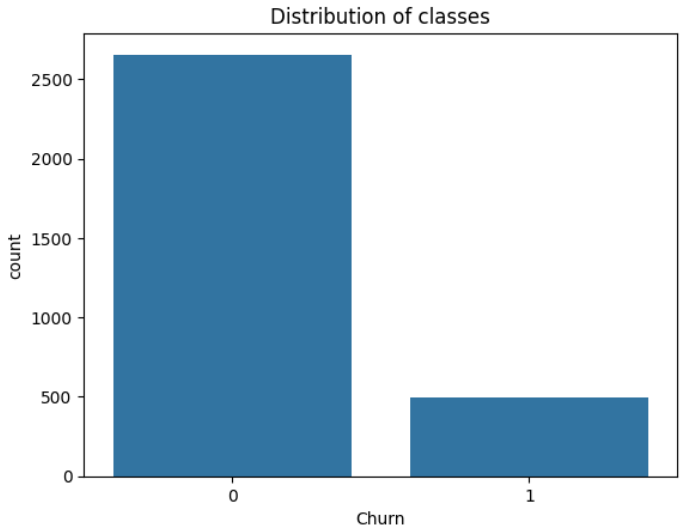


*Figure 11: Class Imbalance in Data*

## 2.4 Model Development

In order to train models for the classification task at hand, 4 algorithms were chosen. They are shown in Figure 13.

- Logistic Regression

  A linear model used primarily for binary classification that uses a logistic function to predict probabilities. This model was chosen for its simplicity which makes it easy to understand. Additionally, it also acts as a good baseline model to compare performances.

- Random Forest

  An ensemble learner which uses bagging to combine multiple decision trees. This model was chosen as it's highly robust against noise as it averages predictions of many trees. It also provides insights on feature importance which is useful during feature selection for model training. This is shown in Figure 12.

  It can be seen that the most importance features are 'call failure', 'complains' and 'subscription length'.

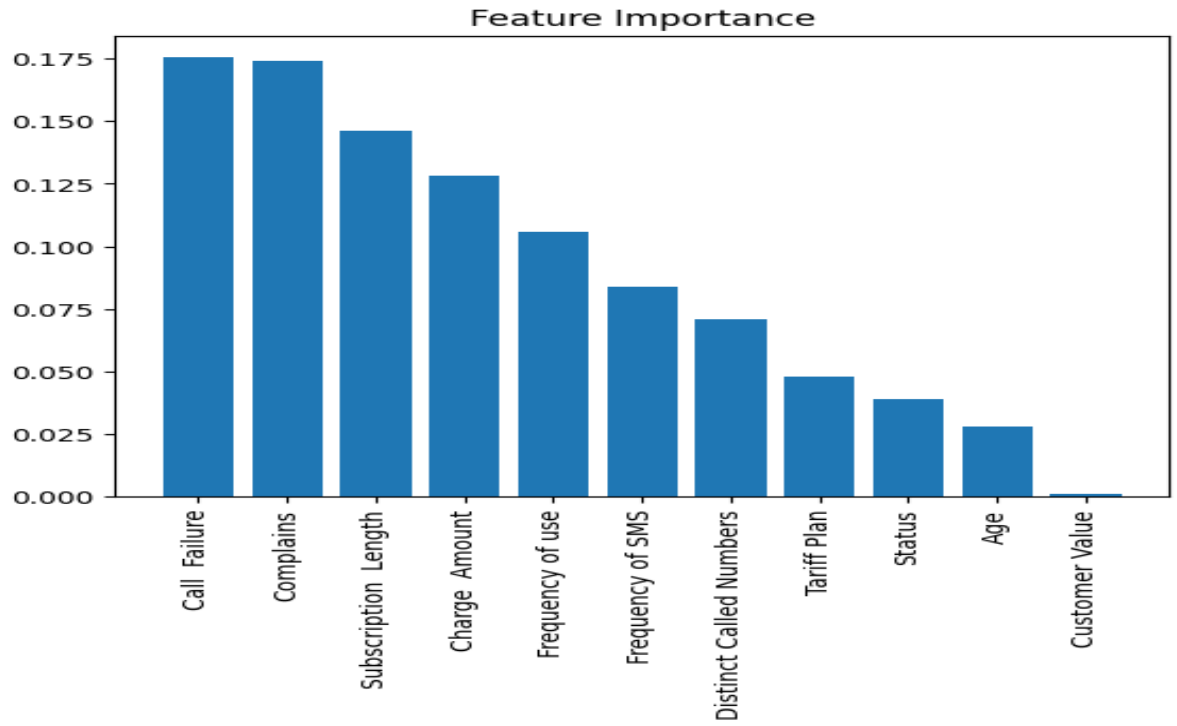


*Figure 12: Feature Importance*

- LightGBM

  A gradient boosting algorithm that uses decision trees. This model was chosen due to its efficiency (faster training speed). It can also handle class imbalance and categorical features.

- Stacking Classifier

  An ensemble learner that combines multiple classification models. Here, we use the Logistic Regression and Random Forest models created in the stack and use LightGBM as the meta-classifier. This was chosen as it leverages the strength of different models by combining them which leads to improved performance. They also usually lead to better generalization.

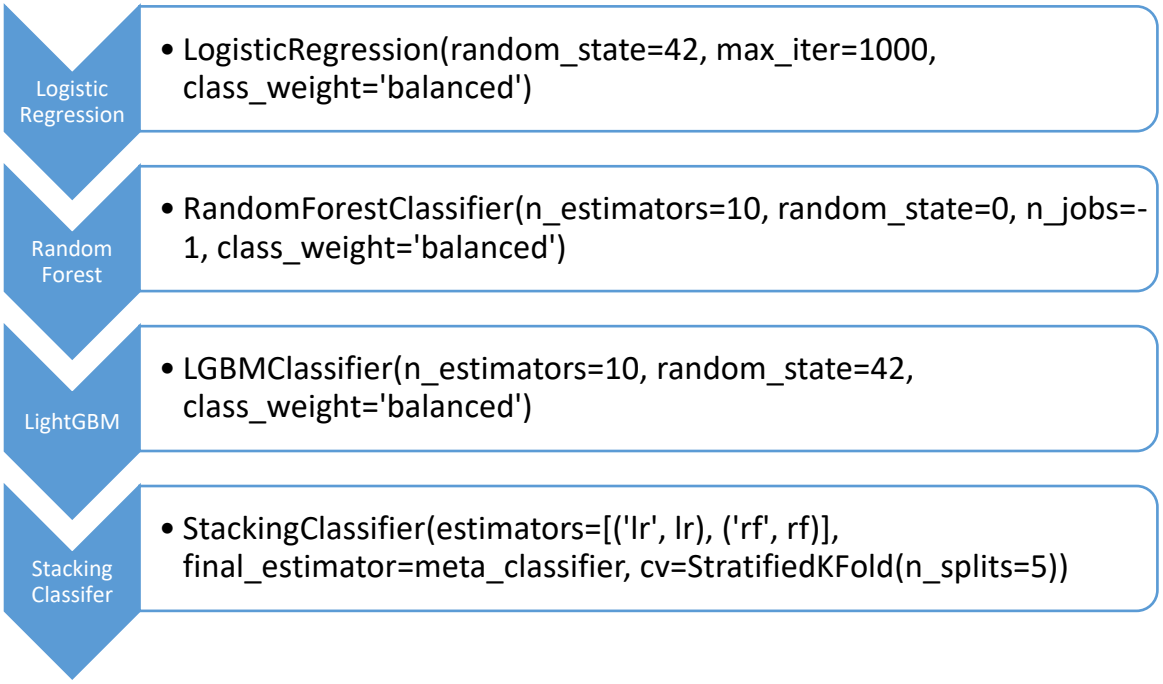


*Figure 13: Developed Models*

### 2.5 Performance Measurement

To develop an adequate classification model that is capable of predicting customer churn, several models must be created and tested.

By comparing the performance of the models created, one can find the best model that generalizes to the patterns that exist within the data. For this purpose, all the models were developed after splitting the dataset into two as training and test data. This allows us to identify how the trained model performs with unseen data. Thus, 80% of the data was used for training with the remaining used for validation.

There are many ways in which performance can be measured for classification such as accuracy, precision, recall, area under the curve, confusion matrix etc. Usually, the final model is chosen based on the accuracy. However, using accuracy creates a misleading image, especially when the dataset is heavily imbalanced. Therefore, model evaluation is conducted by generating many of the above mentioned metrics while the final model selection will be done based on the F1 score. This is a metric which calculates the harmonic mean between precision and recall.

Precision is the fraction of correctly predicted positive instances out of all predictions while recall is the fraction of correctly predicted positive instances out of all positive instances. Since there is a data imbalance present with the minority class in positive instances, both precision and recall must be relatively high for the model. Thus, the weighted average F1 score would be used primarily for model evaluation.

## 2.6 Hyperparameter Tuning

Hyperparameter tuning is the process of optimizing the parameters of a machine learning model that are not learned from the training data but are set prior to the training process. This is used to find the optimal combination of hyperparameters that yield the best performance of the model on a given task. This process is crucial as it can significantly enhance the accuracy, efficiency, and generalization ability of the churn prediction model. This can lead to improved model performance, prevention of overfitting or underfitting, and better utilization of computational resources.

Here, after assessing the initial performance of each model, different hyperparameters were tested for each algorithm using GridsearchCV. Using this, every combination of the hyperparameter specified could be explored. They are shown in Table 1. This led to improvements in certain models which is discussed in the next section.

| **Model** | **Hyperparameter Tuned** |
|---|---|
| Logistic Regression | C, solver, penalty |
| Random Forest | Number of estimators, minimum samples split |
| LightGBM | Number of estimators, learning rate |

*Table 1: Hyperparameters*

## 2.7 Model Evaluation

| Performance Metric | Model | | | |
|---|---|---|---|---|
| | Logistic Regression | Random Forest | LightGBM | Stacked Classifier |
| **Accuracy** | 0.8397 | 0.954 | 0.8968 | 0.9476 |
| **F1 Score** | 0.8556 | 0.9535 | 0.9052 | 0.9497 |
| **Precision** | 0.4942 | 0.8723 | 0.6104 | 0.7705 |
| **Recall** | 0.8687 | 0.8283 | 0.9494 | 0.9495 |
| **Area Under the Curve** | 0.9234 | 0.9841 | 0.964 | 0.9821 |

*Table 2: Model Performance Before Hyperparameter Tuning*

Overall, it can be seen that all four models show adequate performance of over 80% accuracy. However, since the dataset used for the training is highly imbalanced, the models are evaluated primarily based on the F1 score. Since the imbalance is addressed through the class weight adjustment in each algorithm, all models show promising performance. The logistic regression model acts as a baseline model with an F1 score of 0.8556. However, its precision is quite low indicating that it's less able to predict positive instances out of all predictions. The best performance is given by the bagging approach with Random Forest which gives an F1 score of 0.954. The confusion matrix of this model is shown in Figure 14. It can be seen that the model shows is adequate in predicting for both classes.

After hyperparameter tuning, performance has improved in certain models. Logistic Regression shows no change while Random Forest and LightGBM show significant results. It can be seen that the F1 score of Random Forest has increased by 0.54% while it has increased by 6.08% in LightGBM. Therefore, we can conclude that through hyperparameter tuning, we were able to improve the performance of the models created. After tuning, LightGBM shows the most promising performance across many metrics. The confusion matrix for this is shown in Figure 15. It can be seen that the prediction of the positive class has increased from 82.8% to 96% after hyperparameter tuning. This shows better predictions than what was shown with Random Forest. Therefore, this is chosen as the final model which will be used for customer churn prediction.

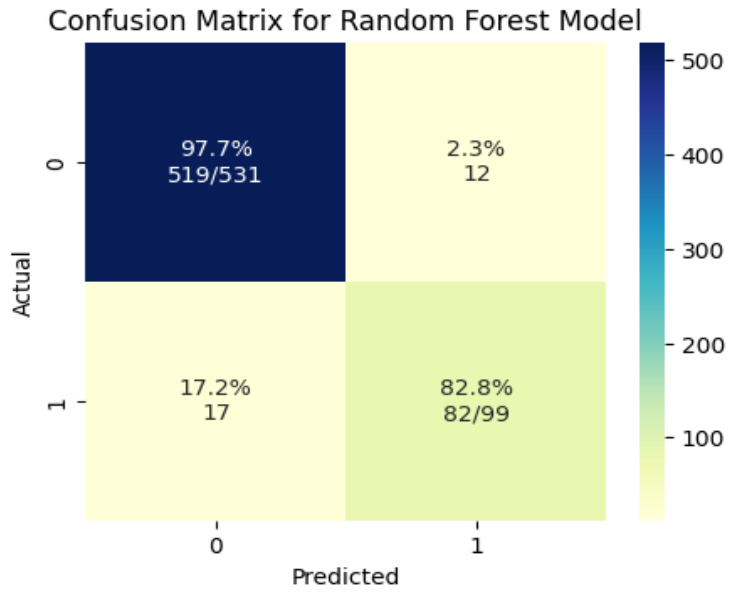


*Figure 14: Confusion Matrix for Random Forest Model*

| **Performance Metric** | **Model** | | |
|---|---|---|---|
| | **Logistic Regression** | **Random Forest** | **LightGBM** |
| **Accuracy** | 0.8397 | 0.9587 | 0.9651 |
| **F1 Score** | 0.8556 | 0.9589 | 0.966 |
| **Precision** | 0.4943 | 0.8614 | 0.8407 |
| **Recall** | 0.8687 | 0.8788 | 0.9596 |
| **Area Under the Curve** | 0.926 | 0.9851 | 0.9917 |

*Table 3: Model Performance After Hyperparameter Tuning*

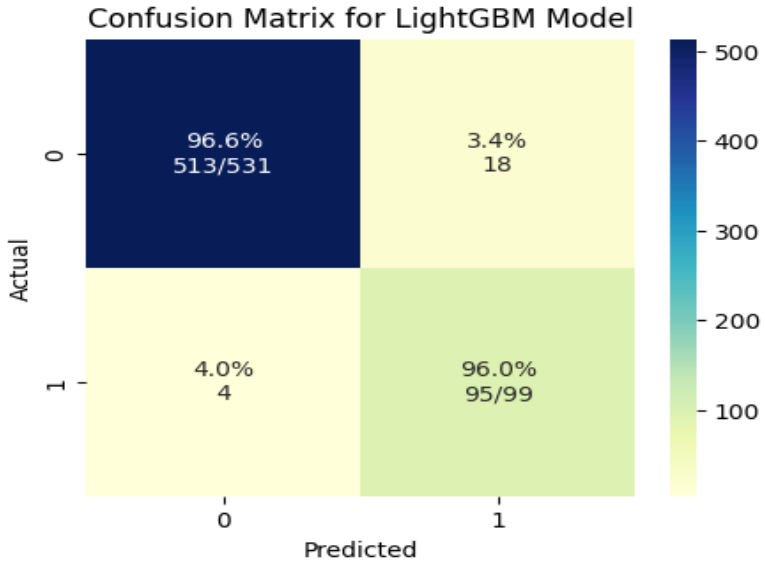


*Figure 15: Confusion Matrix for LightGBM Model*

# 3. REGRESSION

In today's competitive landscape, understanding and predicting customer value is crucial for sustaining growth and enhancing customer satisfaction. It can provide a strategic advantage by enabling targeted marketing, personalized service offerings, and efficient resource allocation. Customer value prediction involves estimating the future revenue a customer will generate, which is vital for optimizing business strategies and improving overall profitability.

The benefits of predicting customer value are multifaceted. Firstly, it allows the company to identify high-value customers who are more likely to contribute significantly to revenue. By focusing marketing efforts and personalized services on these high-value customers, the company can improve customer retention and loyalty. Additionally, customer value prediction helps in identifying at-risk customers who might churn, allowing the company to implement proactive retention strategies. This not only reduces churn rates but also enhances customer satisfaction by addressing issues before they escalate.

Thus, this report also examines the prediction of customer value through regression techniques.

## 3.1 Model Development

In order to train models for the prediction task at hand, 3 algorithms were chosen.

- Linear Regression
  A model which predicts a target variable by finding the best-fit linear relationship between the variables. This model was chosen as it is easy to understand, and use. It's also highly interpretable. Additionally, it can be used as a baseline model for model comparison.
- Random Forest
  An ensemble learner that combines multiple decision trees to predict. This was chosen due to its capability of capturing complex interactions among variables and controlling overfitting. Additionally, it provides insights as to what features contribute towards to the model performance. This is shown in Figure 16.

It can be seen that the most important features are ‘call failure’, ‘complains’, and ‘subscription length’. They contribute the highest to the model performance.

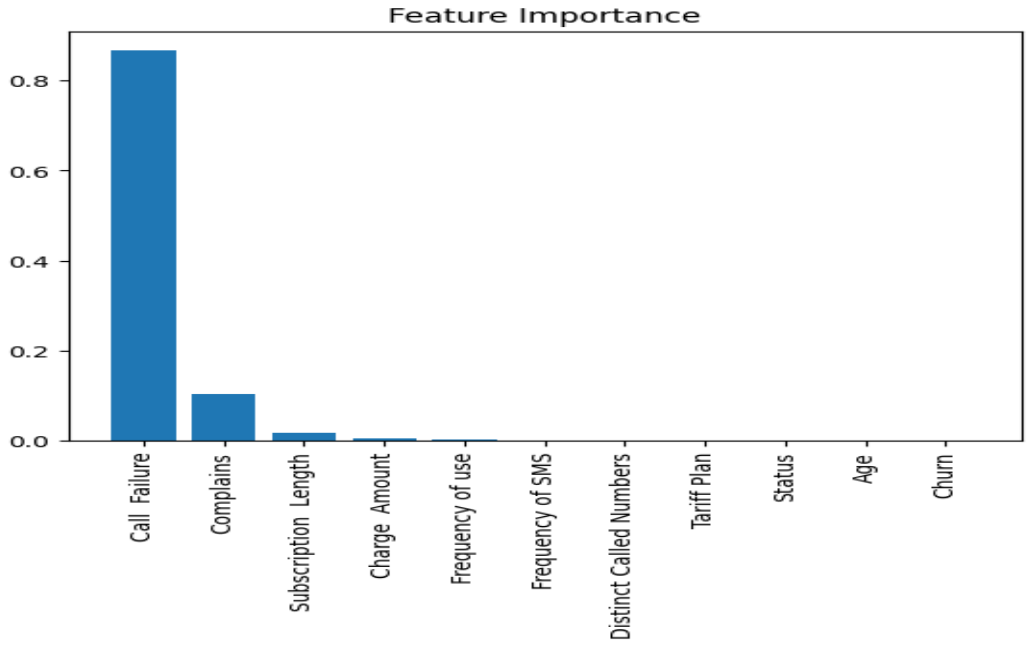


*Figure 16: Feature Importance of Random Forest Regression*

- LightGBM

  A gradient boost based ensemble learner. This model was chosen due to its speed, and scalability while maintaining good prediction accuracy.

The chosen models were trained after splitting the dataset to two as train and test data on a 60:40 ratio.

## 3.2 Hyperparameter Tuning

The initial models fitted showed promising results. However, hyperparameter tuning was still conducted to identify whether performance can be improved. This was conducted through GridsearcCV which finds the best parameters after searching through all the combinations of parameters specified. The hyperparameters are shown in Table 4.

| Model | Hyperparameter Tuned |
|---|---|
| Linear Regression | C through ridge regression |
| Random Forest | Number of estimators, minimum samples split, maximum depth |

*Table 4: Hyperparameters Tuned for Regression*

### 3.3 Model Evaluation

There are many ways in which performance can be measured for regression such as mean absolute error(MAE), mean squared error (MSE), root mean squared error and R squared. All the models which were developed were assessed based on these.

Table 5 shows the performance metrics after the initial model development. Generally, lower values indicate better performance in error metrics. It can be seen that Random Forest has a better MAE than other algorithms. However, LightGBM models provides better metrics in other evaluation criteria. It provides the lowest MSE, and RMSE. It also has the highest R squared which indicates that the model has a good fit to the data.

| Performance Metric | Model | | |
|---|---|---|---|
| | **Linear Regression** | **Random Forest** | **LightGBM** |
| **MAE** | 56.42 | 17.79 | 18.31 |
| **MSE** | 6974.01 | 2344.43 | 1493.55 |
| **RMSE** | 83.51 | 48.42 | 38.65 |
| **R Squared** | 0.9728 | 0.9908 | 0.9941 |

*Table 5: Initial Performance Metrics for Regression*

Table 6 shows the performance metrics after hypeparameter tuning. It can be observed that there has been no significant change in the performance of the models through hyperparameter tuning. Therefore, LightGBM remains the best model which provides the best predictions.

| Performance Metric | Model | |
|---|---|---|
| | **Linear Regression** | **Random Forest** |
| **MAE** | 56.43 | 17.79 |
| **MSE** | 6977.09 | 2344.43 |
| **RMSE** | 83.53 | 48.42 |
| **R Squared** | 0.9727 | 0.9908 |

*Table 6: Performance Metrics for Regression after Hyperparameter Tuning*

# 4. CONCLUSION

In this customer churn classification task, we evaluated the performance of four different models: Logistic Regression, Random Forest, LightGBM and stacking classifer. Each model was assessed based on key metrics including Accuracy, F1 Score, Precision, Recall, and Area Under the Curve (AUC).

The LightGBM model emerged as the top performer with an accuracy of 0.9651, an F1 score of 0.966, and an AUC of 0.9917. LightGBM's efficient handling of large datasets, speed, and advanced capabilities in capturing complex patterns in the data, particularly with imbalanced classes, made it the most effective model for predicting customer churn.

In this customer value prediction task, we evaluated the performance of three different models: Linear Regression, Random Forest, and LightGBM.

The LightGBM model emerged again as the top performer with an RMSE of 38.65, an R squared of 0.9942.

The benefits of accurately classifying customer churn are substantial. Effective churn prediction allows businesses to proactively identify at-risk customers and implement targeted retention strategies, thereby reducing customer attrition rates. This leads to increased customer lifetime value, improved customer satisfaction, and ultimately, higher profitability.

In summary, while each model has its strengths, LightGBM provided the best overall performance and is recommended for deployment in predicting customer churn and customer value. Its superior metrics highlight its capability to accurately identify customers at risk of churning, as well reliably predict customer. This can be crucial for maintaining a loyal customer base and enhancing business growth as well.