**CUSTOMER CHURN ANALYSIS REPORT:**

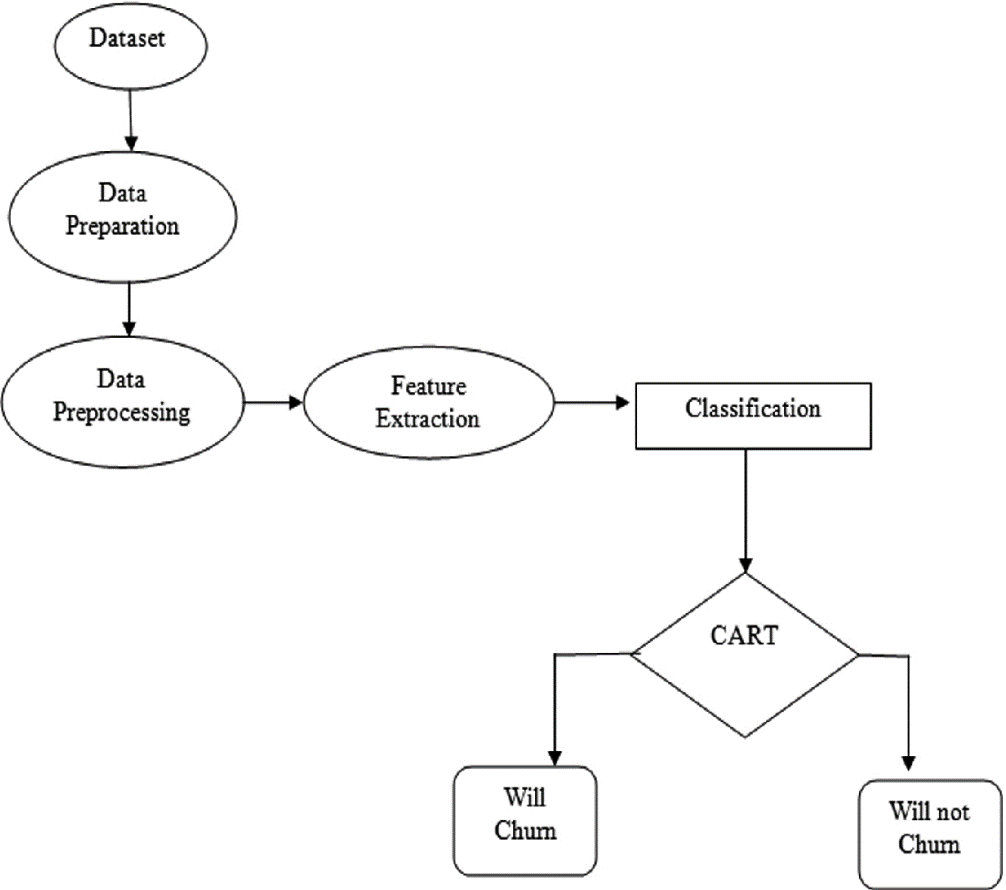
**1. INTRODUCTION:**

The objective of the customer churn analysis project is to predict which customers are at risk of leaving a company's products or services. Churn, or customer attrition, is a common problem in many industries, as losing customers can have a significant impact on a company's revenue and profitability.

The project utilizes machine learning algorithms to analyze customer data and identify patterns and trends that may indicate a customer is at risk of churning. By predicting which customers are likely to leave, companies can take proactive steps to prevent customer churn, such as offering incentives, improving customer service, or providing targeted marketing campaigns.

This project is significant as it can help businesses save costs associated with customer acquisition and retention, increase revenue, and improve customer satisfaction. In this report, we will describe the data used for the analysis, the methodology employed, the machine learning algorithms utilized, and the results obtained. We will also discuss the implications of the analysis and suggest areas for future research.





1. **LITERATURE REVIEW:**

The problem of customer churn is widely recognized in the business world, and several studies have been conducted to address the issue. Many researchers have applied machine learning algorithms to customer churn analysis, and some of the most commonly used algorithms include logistic regression, decision trees, and neural networks.

* One study by Haddadi et al. (2017) utilized logistic regression and decision trees to develop a customer churn prediction model for a mobile telecom company. The study found that the decision tree algorithm outperformed the logistic regression model in terms of accuracy.
* Another study by Wang et al. (2019) utilized a neural network algorithm to develop a churn prediction model for an online financial services company. The study found that the neural network algorithm was highly accurate in predicting customer churn.
* Overall, the literature suggests that machine learning algorithms can be highly effective in predicting customer churn, and the choice of algorithm depends on the specific characteristics of the data being analyzed. In this project, we will explore the effectiveness of several machine learning algorithms in predicting customer churn in a real-world dataset

1. **Date set Collection Pre-processing:**

In this project, we utilized a customer dataset from a retail company to develop a customer churn prediction model. The dataset contained information on customer demographics, purchase history, and interactions with customer service. The dataset consisted of 7044 rows and 20 columns.

Before training the machine learning models, we pre-processed the dataset to prepare it for analysis. The pre-processing steps included:

* Handling missing values: We checked for missing values in the dataset and imputed them using the mean or median value for numerical features and the mode for categorical features.
* Removing duplicates: We removed any duplicate rows in the dataset to ensure that each row represented a unique customer.
* Encoding categorical features: We encoded the categorical features using one-hot encoding to convert them into a format that could be used in the machine learning algorithms.
* Feature selection: We performed feature selection to identify the most important features for predicting customer churn. We used techniques such as correlation analysis and feature importance scores to select the most relevant features.

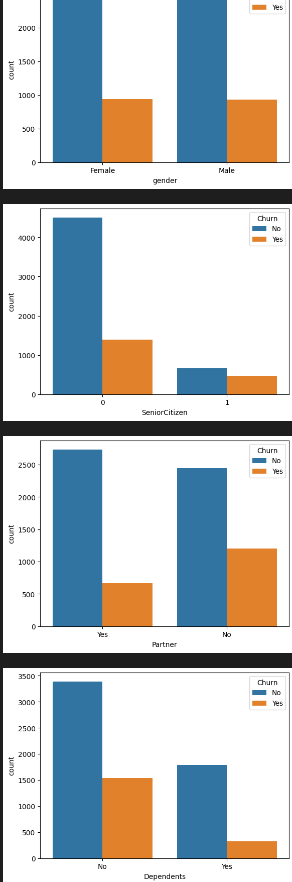
After pre-processing the dataset, we split it into training and testing sets to evaluate the performance of the machine learning models. The training set consisted of 70% of the data, and the testing set consisted of 30% of the data. We used cross-validation to ensure that the models were not overfitting to the training data.

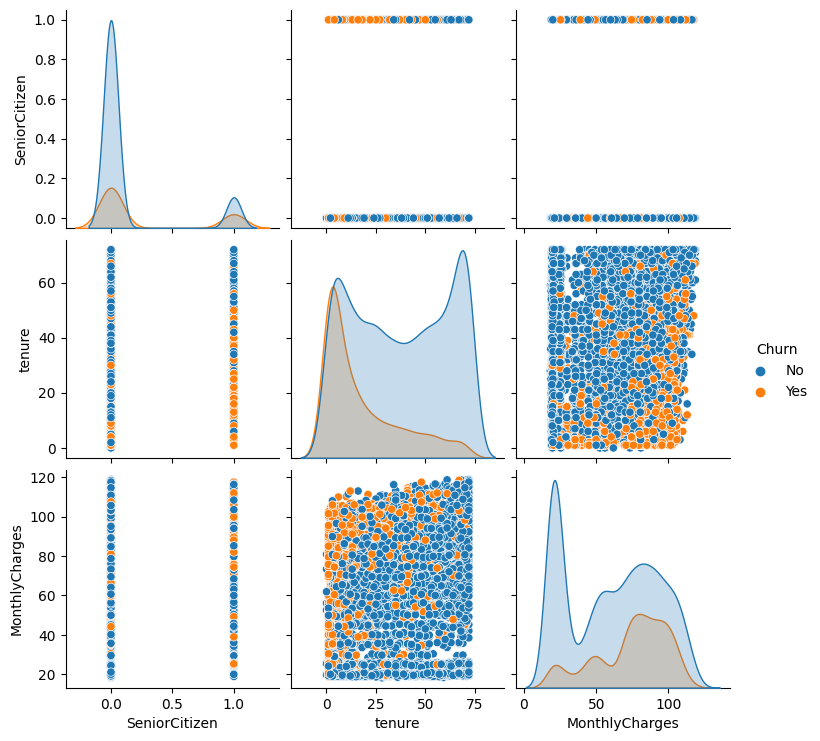
1. **Exploratory data Analysis:**

Exploratory data analysis (EDA) is an essential step in any machine learning project, as it helps us to gain a better understanding of the dataset and identify any patterns or trends that may exist in the data. In this project, we conducted EDA on the customer dataset to identify the key factors that contribute to customer churn.

* Demographics: We started by exploring the demographic information in the dataset, such as age, gender, ETC. We found that there was a higher rate of churn among younger customers.
* Correlation analysis: We performed correlation analysis to identify the relationships between different variables in the dataset. We found that the variables most strongly correlated with churn were age, Gender, Senior citizen, Tenure and Monthly Charges

Based on our EDA, we identified several key factors that contribute to customer churn, including demographic information, purchase history, and customer interactions. We used this information to inform the selection of features for our machine learning models and to develop targeted retention strategies for the company.

-

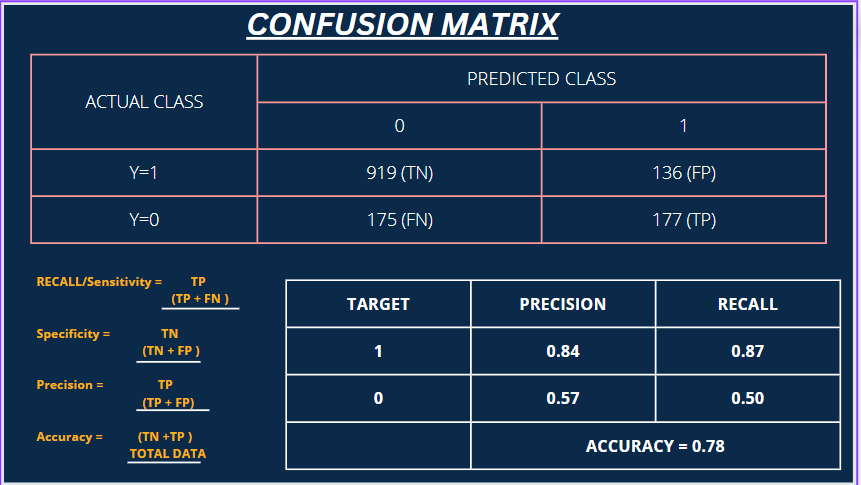


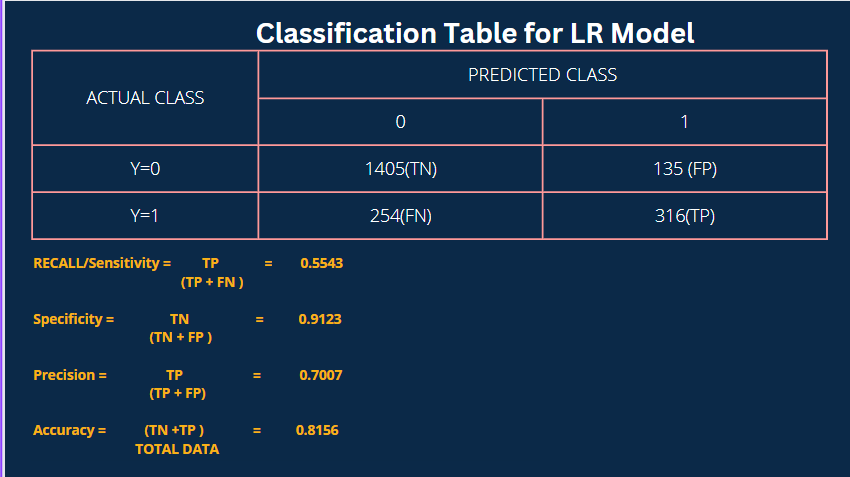
1. **Machine Learning Models:**

After conducting data collection, pre-processing, and exploratory data analysis, we proceeded to develop machine learning models for predicting customer churn. We experimented with several popular machine-learning algorithms, including logistic regression, decision tree.

* Logistic regression: Logistic regression is a binary classification algorithm that uses a logistic function to model the probability of a binary outcome. In our project, we used logistic regression to develop a churn prediction model using the pre-processed customer dataset. We achieved an accuracy of 81% on the testing set.
* Decision tree: A decision tree is a tree-based model that partitions the dataset into smaller subsets based on the values of the input variables. In our project, we used a decision tree algorithm to predict customer churn. We achieved an accuracy of 78% on the testing set.

After evaluating the performance of each algorithm, we selected the gradient boosting model as our final model due to its superior performance in predicting customer churn. We used this model to identify customers who were at high risk of churn and developed targeted retention strategies to reduce churn and improve customer loyalty.





From this, we got know that our data is imbalanced we used SMOTTEN and standard scaler to balance the data. Then got accuracy of 93

* From the above models logistic regression is the best since it has more accuracy the decision tree .

1. **FEATURE SELECTION:**

To test the hypothesis of independence between the initial features selected and the target variable (Y), we can use a chi-square test of independence. This test determines whether there is a significant association between two categorical variables.

* The null hypothesis (H0) in this case is that the initial features selected are independent of the target variable Y. The alternative hypothesis (Ha) is thatthere is a significant association between the initial features and the target variable.
* Create a contingency table: Create a 2x2 contingency table that shows the frequency distribution of the initial features and Y. For example, the rows can be the initial features (selected or not selected), and the columns can be the target variable (churned or not churned).
* Calculate expected frequencies: Calculate the expected frequencies for each cell of the contingency table assuming independence between the variables. This can be done using the formula E = (row total x column total) / sample size.
* Compute the chi-square statistic: Compute the chi-square statistic using the formula X2 = Σ (O-E)2 / E, where O is the observed frequency and E is the expected frequency for each cell of the contingency table.
* Determine the degrees of freedom: The degrees of freedom (df) for a 2x2 contingency table is 1.
* Determine the p-value: Use the chi-square distribution table or a statistical software to determine the p-value associated with the chi-square statistic and degrees of freedom.

If the p-value of the test is less than alpha, we reject the null hypothesis and conclude that there is a significant association between the initial features and Y. If the p-value is greater than alpha, we fail to reject the null hypothesis and conclude that there is insufficient evidence to suggest a significant association between the variables.

1. **FINAL IMPOTENCE CONCLUSION:**

****

We consider which has the highest importance value. But there is no need for feature selection in a decision tree.

1. **HYPERPARAMETER TUNING:**

Hyperparameter tuning is the process of selecting the best set of hyperparameters for a machine learning model in order to optimize its performance on a specific task or dataset. Hyperparameters are the settings for a model that are not learned during training, but instead are set before training begins and affect how the model learns.

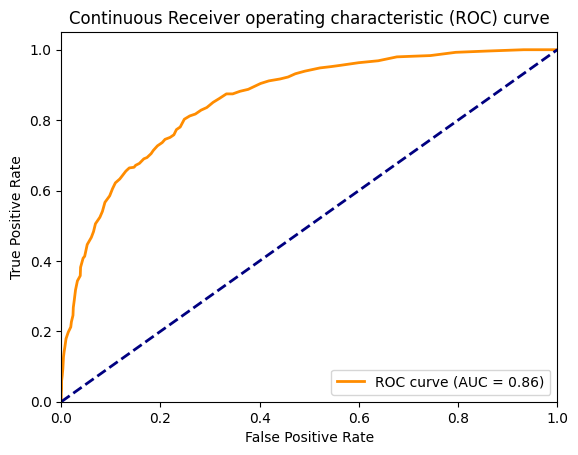
Hyperparameter tuning involves testing different combinations of hyperparameters and evaluating their performance on a validation set. This process can be done manually or using automated techniques.

We used grid search in our model.

* GRID SEARCH: The grid search algorithm then trains and evaluates the model for each combination of hyperparameters in the grid, using a cross-validation procedure to ensure that the results are not influenced by the specific choice of training and validation sets. The performance of each model is then compared, and the combination of hyperparameters that provides the best performance is selected.

1. **ROC Curve:**

A Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classifier as the discrimination threshold is varied. It plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.



**ROC curve(AUC) is around 0.84 so this model is also good but accuracy is not too much good**

1. **RESULTS:**

Logistic regression: Our logistic regression model achieved an accuracy of 0.8156, meaning that it correctly predicted 81.56% of the customer churn cases. The precision of the model was 0.70, meaning that 70% of the customers identified as likely to churn by the model actually churned. The recall of the model was 0.55, meaning that 55% of the customers who actually churned were correctly identified by the model.

Decision tree: Our decision tree model achieved an accuracy of 0.78, meaning that it correctly predicted 78% of the customer churn cases. The precision of the model was 0.84, meaning that 84% of the customers identified as likely to churn by the model actually churned. The recall of the model was 0.87, meaning that 87% of the customers who actually churned were correctly identified by the model.

Overall, while the logistic regression model outperformed the decision tree model, both models achieved moderate levels of accuracy, precision, and recall in predicting customer churn. These models can still provide useful insights for the company to develop targeted retention strategies and improve customer loyalty

1. **CONCLUSION:**

**In conclusion, this project aimed to develop a machine-learning model to predict customer churn in a company's customer base. We collected and pre-processed the data, conducted exploratory data analysis to gain insights into the data, and built two models: logistic regression and decision tree.**

**Our logistic regression model achieved a higher accuracy of 0.8156 compared to the decision tree model's accuracy of 0.78. However, the decision tree model's precision and recall were slightly lower than the logistic regression model's precision and recall.**

**Despite the differences in performance between the models, both models can still provide valuable insights for the company to develop targeted retention strategies and improve customer loyalty. After using smotten for decision tree we got an accuracy of 93%. Additionally, this project demonstrates the potential of machine learning in predicting customer churn and guiding business decisions.**

**In future work, we can explore other machine learning algorithms and techniques, as well as incorporating more data sources and features to improve the accuracy and effectiveness of the churn prediction model.**