A REPORT ON CUSTOMER CHURN PREDICTION FOR CONNECT TELECOM COMPANY

ConnectTel is a leading telecommunications company, a trusted provider of reliable voice, data, and Internet services. They offer telecommunications solutions, including mobile networks, broadband connections, and enterprise solutions.

PROBLEM DEFINITION

The problem we aim to solve is customer churn prediction for ConnectTel Telecom Company. Customer churn, also known as customer attrition, is where customers cease their relationship with a company or switch to a competitor. This poses a significant threat to its business sustainability and growth. High churn rates can lead to revenue loss and decreased profitability.

With the use of data analytics and machine learning techniques on the customer churn data, our objective is to develop a customer churn prediction system. This system will enable ConnectTel to:

1. Identify customers who are at risk of churning.
2. Understand the factors and behaviors that contribute to churn.
3. Predict future churn events with high accuracy.
4. Implement retention strategies to retain valuable customers.
5. Enhance customer loyalty, satisfaction, and long-term profitability.

EXPLORATORY DATA ANALYSIS:

We conducted exploratory data analysis to understand the characteristics and distributions of variables in the dataset.

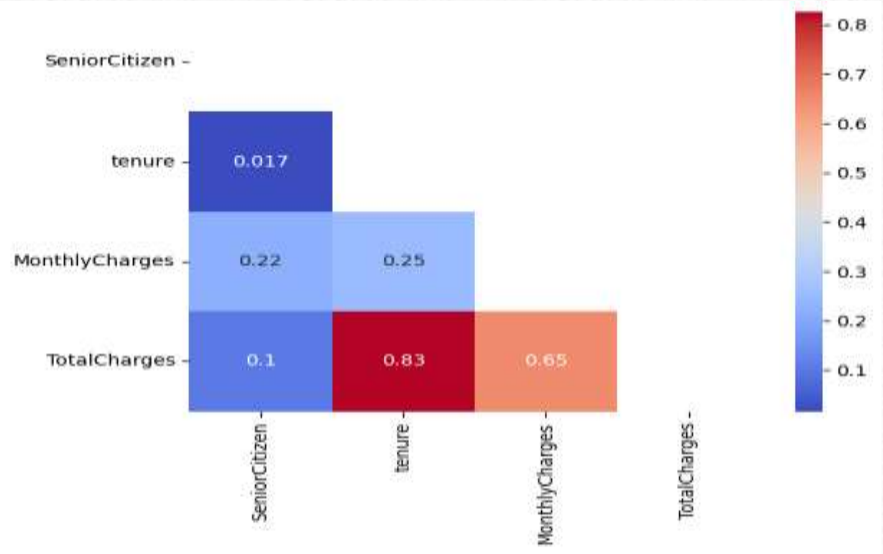
To begin the customer Churn prediction, we load the Connect Tel customer data set using Python via Jupyter Notebook, we display basic information about the dataset, and we show the summary statistics for numerical columns. Also, we check the first few rows of the DataFrame to understand its structure and contents, then we look for any missing values, we check for duplicates, and Finally, we visualize the distribution of the target variable 'Churn' and 'Churn' relationship with some key feature e. g Contract and Tenure.

Churn rate
No = 73.5%
Yes = 26.5% This shows that overall we have a Low Churn rate.
<u>Contract</u> = Monthly (55%), one year(20.9%), and Two years(24.1%), meaning we have a Large number of Customers on Monthly subscriptions.
<u>For tenure</u>, we have a visualization display that shows that a Large number of New Customers left within a short period.

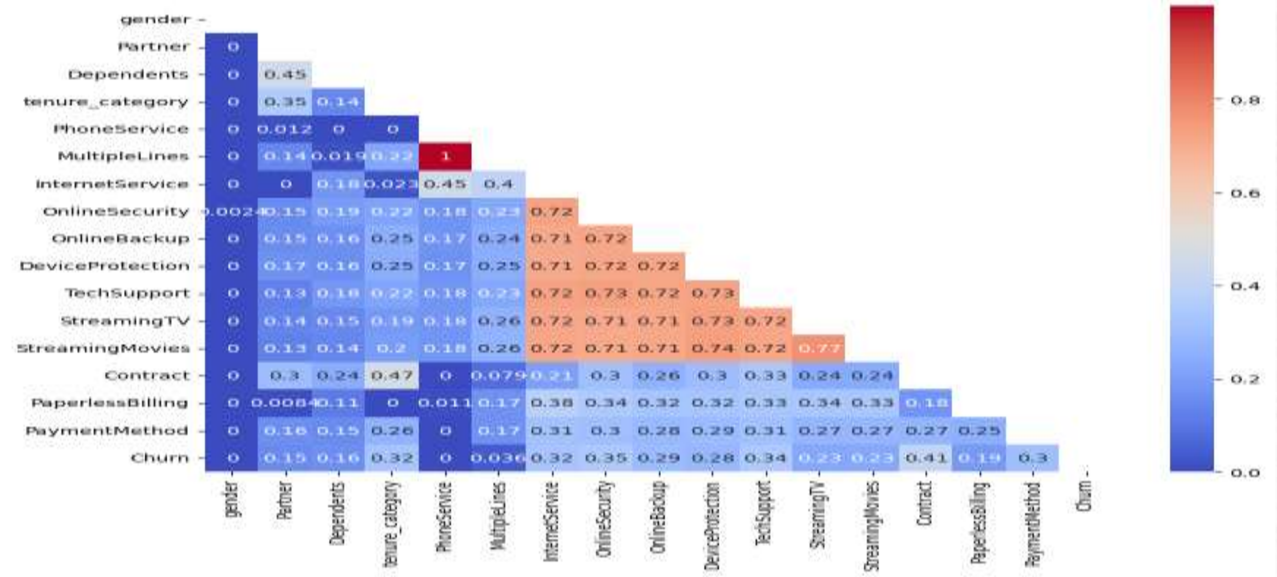
EXPLORE CORRELATIONS

Correlation refers to the statistical relationship between two entities.

For the customer Churn data, we explored correlations for numerical & categorical variables by selecting numerical features, categorical features and plotting a correlation matrix



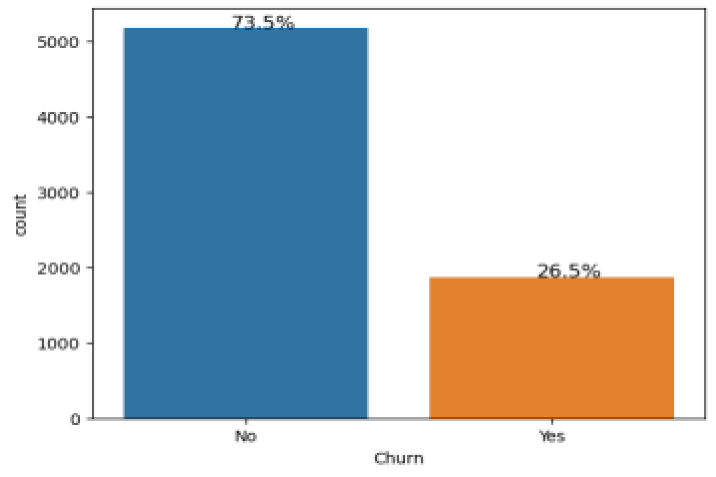
From the Numerical correlation matrix, we were able to show that features like TotalCharges, Tenure and MonthlyCharges are highly correlated to each other.

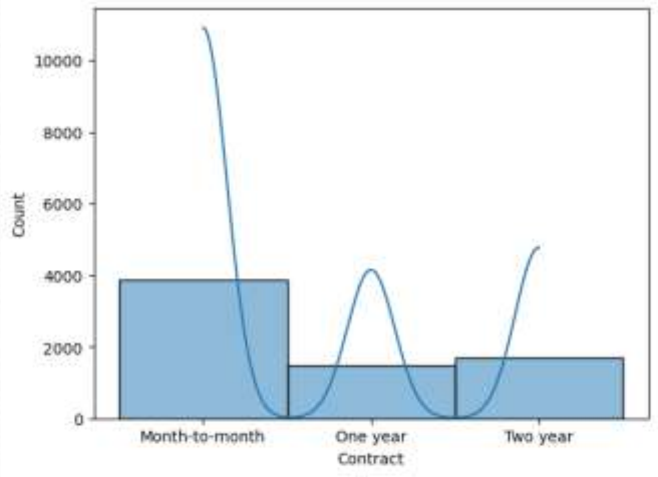


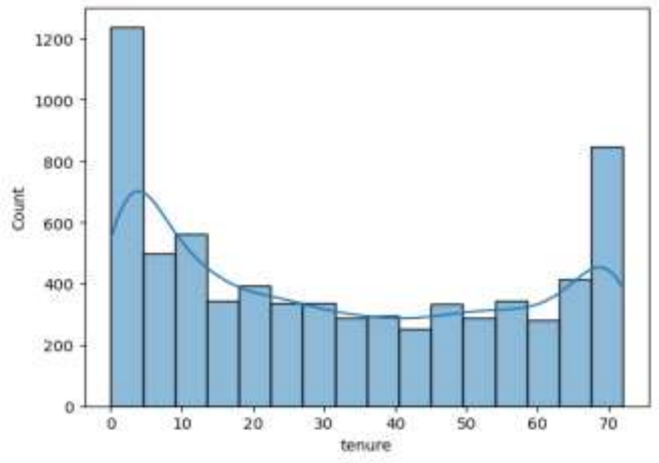
From the Categorical correlation matrix, features like Multiple_lines, Phone_services, Online_security, Online_backup, Device protection, Tech Support, Streaming_movies, Streaming_tv are highly correlated to each other.

UNIVARIATE ANALYSIS

In order to describe a type of data which consist of observations on a single attribute we use univariate analysis. For Connect Tel churn dataset, we observed the following







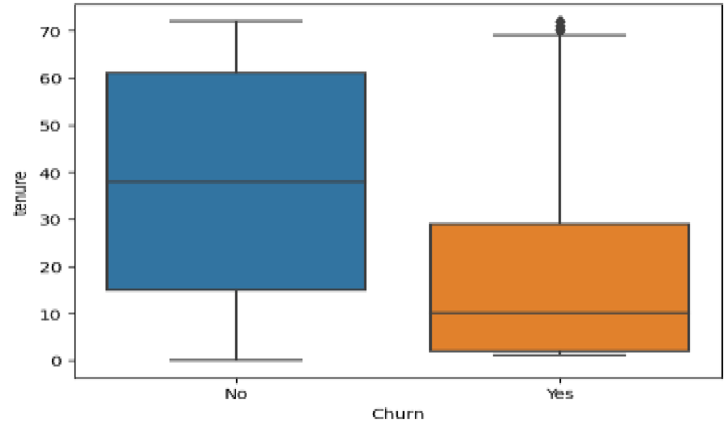
The bar plot displays the total count of customers who are still loyal to the brand and their services and those who have left the brand

A visual display of the contract type that majority of customers are subscribed to.

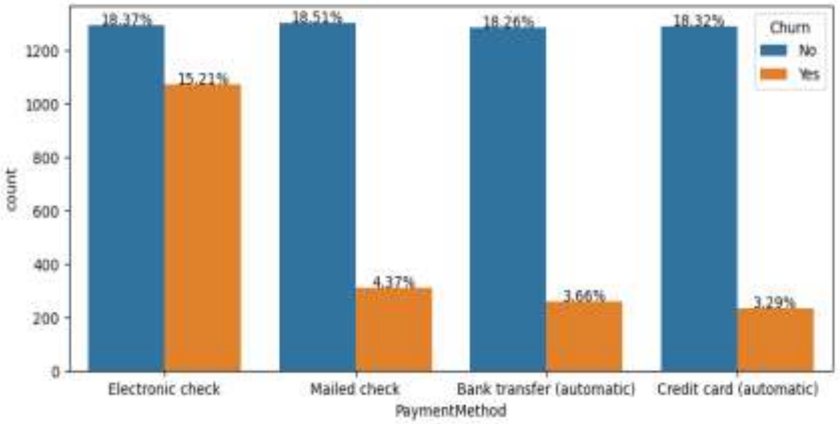
Tenure means longevity and its is obvious from the histogram plot above that a very high amount of New subscribers left and some Long-time subscribers to.

BIVARIATE ANALYSIS

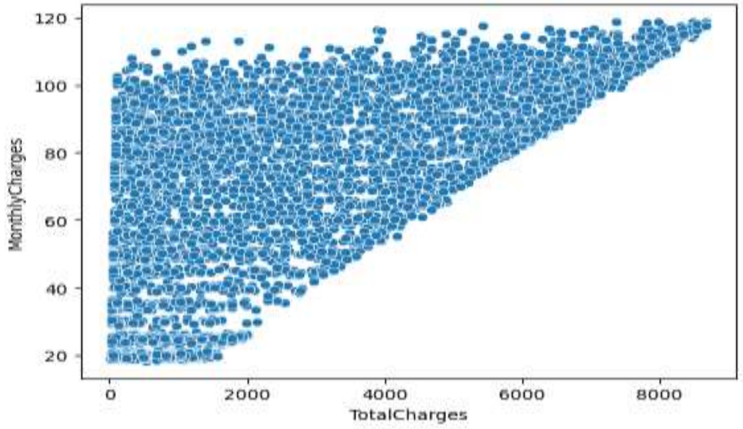
In order to describe a type of data which consist of observations on a two attributes we use bivariate analysis. For Connect Tel churn dataset, we analyzed various Categories and we observed the following



We plotted a box plot to show customers relative to Churn. It was observed that most customers stayed Loyal and fewer customers left within a short tenure with Connect Tel.



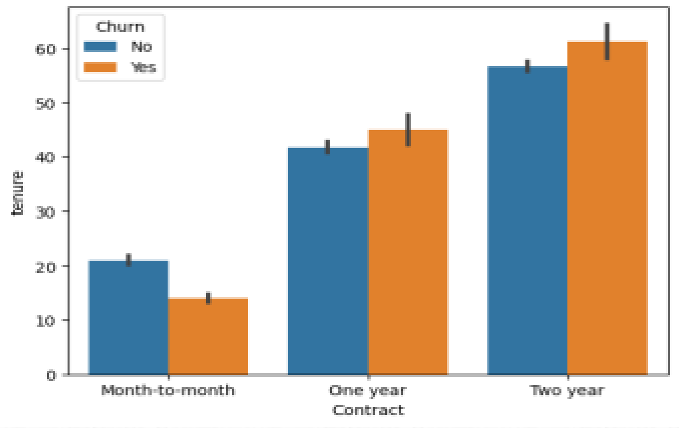
For the payment method to Churn, a lot of Customers who use Electronic Check Churned more when compared to other payment methods.



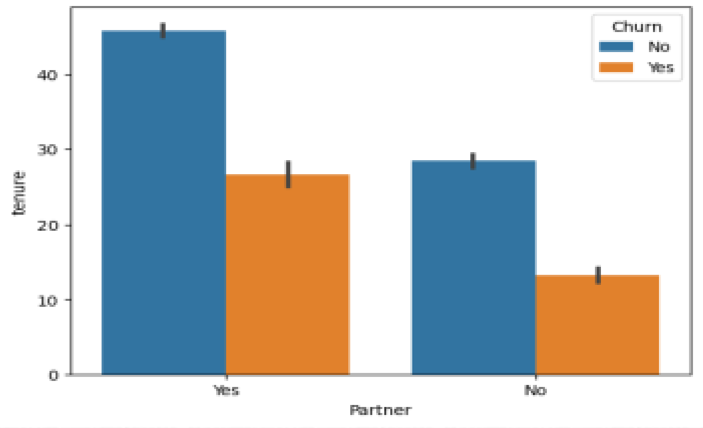
This scatterplot shows that the higher the Monthly charges payed by the customers the Higher the TotalCharges.

MULTI-VARIATE ANALYSIS

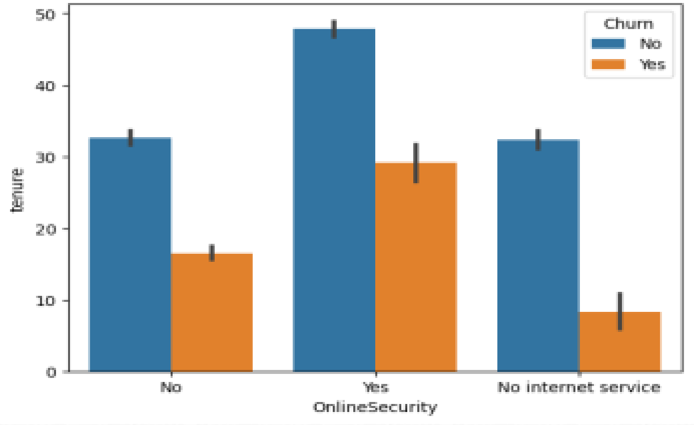
To describe a type of data which consists of observations on more than two attributes we use Multivariate analysis. The Multivariate tells a different story compared to others and we were able to observe the following



We made a plot of Contract and Tenure against Churn and we noticed that customers on a year and two year contract churn more when viewed over a very Long period of time.



Plotted Partner and Tenure against Churn and we noticed that customers with partner churn more when viewed over the years.



We also plotted Online Security and Tenure against Churn and we noticed that customers with existing online security churn more when viewed over the years compared to customers with No online security & No Internet service.

FEATURE ENGINEERING

Feature engineering is a machine learning technique for creating new features or variables that are not in the existing training set.

We applied this technique and created a new feature called 'tenure_category' from 'tenure’. The tenure_category was segmented into three i.) New customer ii.) mid-level customer and iii.)Longtenure_customer

This Segmentation was used to capture underlying patterns or relationships in the data. it will also help to improve model interpretability and lead to improved model performance.

ENCODING CATEGORICAL VARIABLES

Encoding categorical variables is the process of converting categorical data into numerical format, which can be used as input for machine learning algorithms.

There are several methods for encoding categorical variables

1. Ordinal Encoding: Assigning integer values to categories based on their natural ordering.
2. One-Hot Encoding: Creating binary (0/1) dummy variables for each category.
3. Label Encoding: Assigning integer labels to categories, each category is mapped to a unique integer value.
4. Target Encoding (Mean Encoding): Replacing categories with the mean of the target variable.

For this data set, we used **One-hot encoding** to create dummy variables for each category as input for machine learning algorithms. # Perform one-hot encoding: df = pd.get_dummies(data,dtype="int64",drop_first=True)

MODEL SELECTION, TRAINING AND VALIDATION

Model selection refers to the process of selecting the best machine-learning model or algorithm for a given task
For this project, we took the following steps for the process of model selection:

1.After encoding, we assign the data a variable name(df) and our target variable ‘Churn’ to y. then we scale the data. Scaling features ensures that all features are penalized equally, making the model feature importance scores more interpretable. For this project, the Minmax scaler was used to scale and transform each feature to a given range, typically between 0 and 1.

2. Splitting Data: The preprocessed data is typically divided into training, validation, and test sets. The training set is used to train the models, the validation set is used to tune hyperparameters and evaluate model performance during training, and the test set is used to evaluate the final selected model's performance.
(# X_train,X_test,y_train,y_test = train_test_split(scaled_df,y,test_size=0.2, random_state=42)

The split data was trained on 9 Supervised learning algorithms
1.) LogisticRegression()
2.) RandomForestClassifier()
3.) SupportVectorClassifier()
4) Naives Bayes
5) SGD Classifier
6) Random Forest
7) XGB Classifier
8) KNNeighbors Classifier
9) Decision Tree

MODEL EVALUATION

This involves fitting the model to the training data and adjusting its parameters.

After fitting the data into the 9 machine-learning algorithms, the Logistic Regression algorithm performed better with the best accuracy, precision, and F1 score

	Logistic Regression	LinearSVC	SVC	SGD Classifier	Random Forest	XGB Classifier	KNNeighbors Classifier	Decision Tree	Naives Bayes
Accuracy Score	**81.97%**	81.26%	81.05%	80.98%	79.63%	78.07%	77.22%	70.69%	68.2%
Precision Score	**68.89%**	67.99%	68.28%	67.44%	65.81%	60.6%	57.69%	44.68%	44.88%
Recall Score	58.18%	55.23%	53.08%	54.42%	47.99%	49.06%	52.28%	45.04%	**88.2%**
F1 Score	**63.08%**	60.95%	59.73%	60.24%	55.5%	54.22%	54.85%	44.86%	59.49%
ROC Score	74.36%	72.93%	72.1%	72.48%	69.51%	68.79%	69.24%	62.48%	**74.6%**

MODEL EVALUATION (Optimization/ Hyperparameter tunning)

After evaluation, we carried out model Optimization/ Hyperparameter tunning for the 9 Machine-learning algorithms USING GRID SEARCH and it was observed that SGD Classifier performed better with the best accuracy, recall and ROC score

index	Logistic Regression	LinearSVC	SVC	SGD Classifier	Random Forest	XGB Classifier	KNeighbors Classifier	Decision Tree	Naive Bayes
Accuracy Score	81.83%	81.26%	81.26%	**82.19%**	80.91%	81.41%	77.71%	77.29%	68.2%
Precision Score	68.34%	68.11%	**69.4%**	69.06%	68.57%	68.69%	58.7%	58.98%	44.88%
Recall Score	58.45%	54.96%	52.28%	59.25%	51.47%	54.69%	53.35%	46.65%	**88.2%**
F1 Score	63.01%	60.83%	59.63%	**63.78%**	58.81%	60.9%	55.9%	52.1%	59.49%
ROC Score	74.35%	72.85%	71.99%	**74.85%**	71.49%	72.86%	69.92%	67.48%	74.6%

MODEL EVALUATION (Optimization/ Hyperparameter tunning)

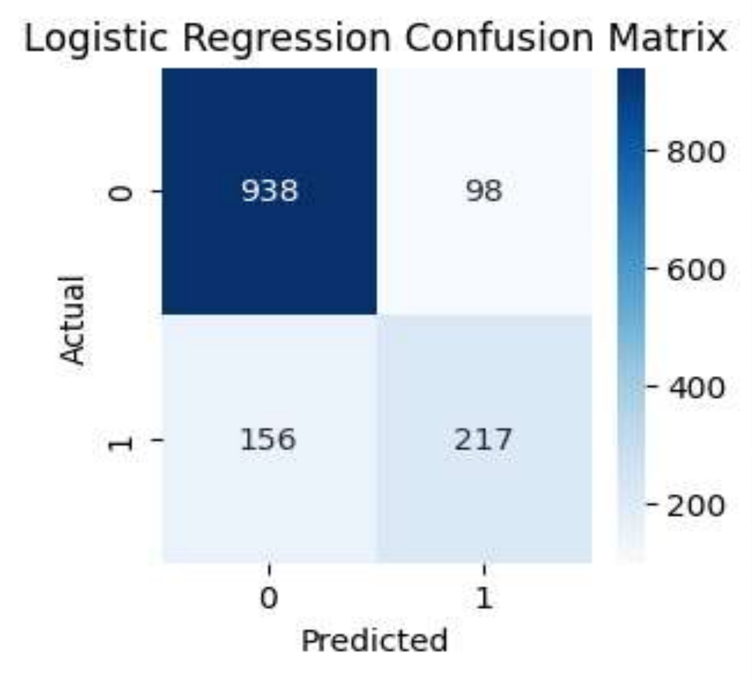
An ENSEMBLE METHOD FOR HYPER-PARAMETER TUNNING was carried out using the Voting Classifier with RandomizedSearchCV.

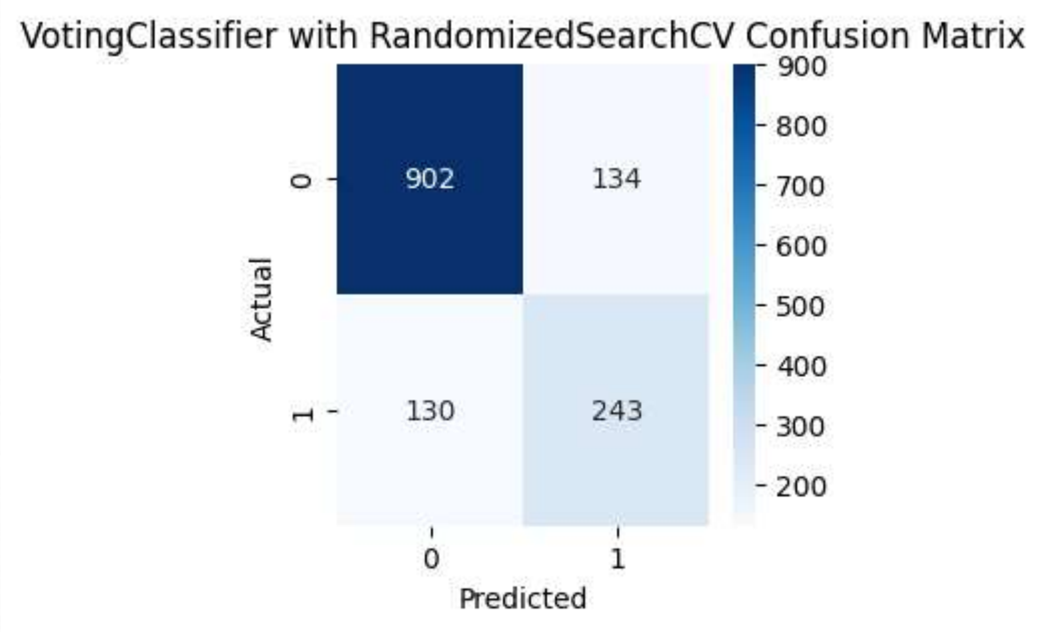
A Voting Classifier is an ensemble machine-learning model that combines predictions created by various individual models to build a final prediction to make a better-performing combination. A combination of 7 algorithms `LogisticRegression, SVC, RandomForestClassifier, XGBClassifier, KNeighborsClassifier, DecisionTreeClassifier, and GaussianNB were used to make this prediction.`

Metrics/ best model	Voting Classifier
Accuracy Score	81.26%
Precision Score	81.33%
Recall Score	81.26%
F1 Score	81.3%
ROC Score	76.11%

CONFUSION MATRIX

The result below shows the confusion matrix for Logistic regression and Voting Classifier with Randomized Search.





CONCLUSION

From the confusion matrix for LogisticRegression
The no of actual positives correctly predicted (TRUE POSITIVE) is 938 & the no of actual Negatives correctly predicted(TRUE NEGATIVE) is 217 while the no of (FALSE POSITIVE) is 98 & FALSE NEGATIVE) is 156

After we carried out model Optimization/ Hyperparameter tuning using the Voting Classifier
The no of actual positives correctly predicted (TRUE POSITIVE) is 902 & the no of actual Negatives correctly predicted(TRUE NEGATIVE) is 243 while the no of (FALSE POSITIVE) is 134 & FALSE NEGATIVE) is 130

The voting Classifier had a lower FALSE NEGATIVE (130) compared to Logistic Regression (156)
Minimizing false negatives can in turn improve user satisfaction and retention.

RECOMMENDATION

1.Retention Strategies:

Implement proactive measures such as personalized offers, discounts, loyalty programs, and improved customer service to enhance customer satisfaction and loyalty.

2. Service Improvement:

a.) Identify areas for service improvement based on customer feedback and complaints.

b.) Address issues related to service quality, network reliability, billing transparency, and customer support to reduce dissatisfaction and churn.

3. Communication and Engagement: Provide relevant and timely information about new services, upgrades, and promotions to keep customers informed and engaged.

4. Feedback from customers through surveys, feedback forms, and customer support interactions to gauge satisfaction and identify areas for improvement.

5. Regularly monitor churn rate, customer satisfaction scores, and customer lifetime value to assess the impact of retention initiatives.