Fundamentals Of Data Science

Topic : Bank Customer Churn Analysis

Name
Ashish Kumar
Pratik Nikhade
Riya Kumari
Yash Doshi

Index

Abstract:-

For companies, customer churn prediction proves to be an important strategy that can be deployed to combat both high cost and low profits. In this study we have brought forward customer churn in the banking industry, examining its depths with the new machine learning models and employing them in order to retain customers longer using strategies of targeting the behavior of the customers. Among the study aims is not only gaining information how customers avoid further purchase but also developing prediction models on customers' avoidance of further purchases, and hence the need to have this information timely and in real time.

The abstract will state the research problems connected to this study and will place them in the proper context according to their significance from the standpoint of preventing consumer turnover which may lead humanizing a situation that is harmful for business. Through these drivers, we understand how churn can happen and we adjust our processes of customer retention with the resources allocation scrutinized, which improves efficiency

The phase of data understanding is the stage where the firm goes through the dataset in huge detail and looking out for attributes like customer demographics, account attributes, and transaction history for example. Explore the data and through data analysis, unravel some hidden trends, patterns and relationships between different data points that are used to model those earlier observations.

In phase of data modelling supervised machine learning algorithms such as logisitc regression and Random forests will be most likely applied. The main reason is for churn prediction of these prediction algorithms.

Moreover, the model is analysed that gives the answers of the factors most likely to have a positive or negative impact on the customers' behavior. Companies can exploit model coefficients of regression and multicollinearity based features for the assessment of asset retention strategies, and to do the evaluation accurately advises focusing on time- and cost-efficiency grading approaches.

This study finally shows the great value of the customer churn prediction in banking industry and indicates the efficiency of machine learning algorithms on users that may jump towards a new organization. By implementing predictive analytics, enterprises can strategically deal with customer churn, boost client gratification and gain a stake in business profitability for the long run.

Introduction:-

Customer retention remains be the keystone to unwavering profits and the success of banking and other industries. A client's disparition or the phenomenon of attrition, which implies the end of a client's dealings with the bank, is now one of the crucial difficulties banks worldwide face. In the wake of recognizing customer retention as the world's most extraordinary tool bans are migrating to data-driven approaches by means of this they can predict and mitigate churn. This opens the discussion of the issue of customer migration within financial industry and offers a resolution using information technology tools and machine learning technology.

Problem Statement:

The objective of this project is to develop a predictive model that can identify customers at risk of leaving the bank based on various features and details provided in the data set. The dataset includes information such as Customer ID, surname, credit score, and other relevant bank details.

Solution:

The solution that the study offers to address the issue of customer churn is the one which is knowledge based and it hinges on the machine learning algorithms and predictive analytics, for solving the problem. Through getting data from historical customer data containing information on demographics, transaction history as well as interaction patterns, banks can spot the trends and patterns in customer behavior. The bases for building up the models based on which churn prediction can be done at the level of single customer are collected right from this stage.

The solution involves several key steps: The solution involves several key steps:

1.Data Collection and Preparation: Assemble a complete data of customers coming from various places including transactional log, survey of customer, census data and demographic information. Render the data clean and formed in order for it to get reported with greater accuracy and uniformity.

2. Feature Selection and Engineering: Emphasize on relevant attributes which impact churn rate and next do an operation on features in order to transform current ones or create new ones.

3. Model Development: Employ machine learning algorithms like logistic regression, or ensemble methods such as random forest in the design and buildup of predictive models. Train the models by historical data techniques and score them through the correct measures.

Data Understanding Phase:

Dataset Name: Bank Customer Churn Analysis

Domain: Banking and Finance

Source of data:

https://www.kaggle.com/datasets/shantanudhakadd/bank-customer-churn-prediction?select=Churn_Modelling.csv

Atrributes:-

RowNumber ,CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary

Target Variable:- Exited

Data Preprocessing(Python)

For Data preprocessing following functions were used to process the dataset:

1. **Visualization of Target Variable Distribution:**
    - The code creates a count plot to visualize the distribution of the target variable 'Exited' using seaborn's **countplot()** function.
    - Each bar represents the count of customers who exited (1) and those who did not exit (0).
    - Annotations are added to each bar to display the exact count value.
    - This visualization helps understand the distribution of the target variable and identifies any class imbalances that may exist.

Number of Non-churned = 7963

Number of Churned = 2037

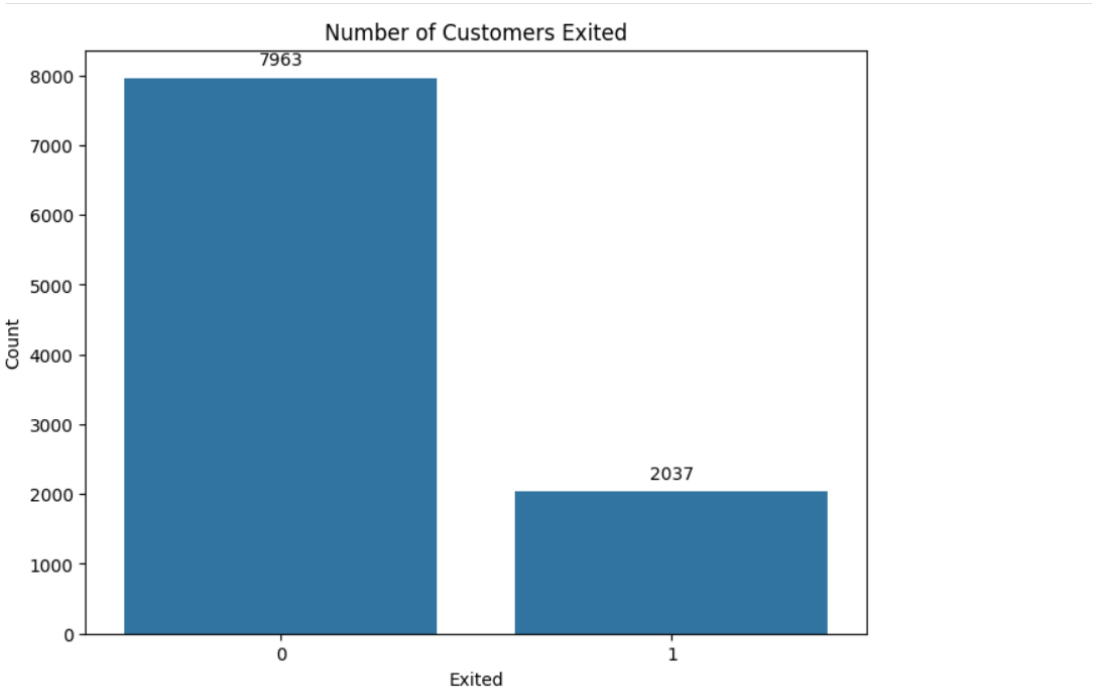

Fig.4 Non-Churned vs Churned

This code snippet demonstrates several steps of data preprocessing:

1. **Dropping Unneeded Columns**:
    - The code uses data.drop(columns=['RowNumber','CustomerId','Surname'], inplace=True) to remove columns named 'RowNumber', 'CustomerId', and 'Surname' from the dataset.
    - These columns likely contain unique identifiers or irrelevant information for the analysis.
2. **Encoding Categorical Variables**:
    - In the 'Geography' column, the code replaces the categorical values 'Germany', 'France', and 'Spain' with numeric values '0', '1', and '2', respectively.
    - In the 'Gender' column, the code replaces 'Female' and 'Male' with '0' and '1', respectively.
    - This transformation is essential because machine learning algorithms typically require numeric inputs and cannot directly process categorical data.
3. **Converting Data Types**:
    - After encoding, the code converts the 'Geography' and 'Gender' columns from object data type to integer using pd.to_numeric().
    - This conversion ensures that the columns are in a numeric format suitable for analysis and modeling.

In [168…

```
#Data Preprocessing
#delete unneeded columns
data.drop(columns=['RowNumber','CustomerId','Surname'],inplace=True)

#Change value in country column
data['Geography'] = data['Geography'].replace(['Germany'],'0')
data['Geography'] = data['Geography'].replace(['France'],'1')
data['Geography'] = data['Geography'].replace(['Spain'],'2')

#Change value in gender column
data['Gender'] = data['Gender'].replace(['Female'],'0')
data['Gender'] = data['Gender'].replace(['Male'],'1')

#convert object data types column to integer
data['Geography'] = pd.to_numeric(data['Geography'])
data['Gender'] = pd.to_numeric(data['Gender'])
data.dtypes
```

Out[168…

```
CreditScore          int64
Geography            int64
Gender               int64
Age                  int64
Tenure               int64
Balance            float64
NumOfProducts        int64
HasCrCard            int64
IsActiveMember       int64
EstimatedSalary    float64
Exited               int64
dtype: object
```

Fig.5 Data Pre-processing

Correlation Matrix:

The relation between the attributes was found out using correlation matrix in python. The correlation matrix helps to understand relationships between different features in the dataset, helping to identify which features are strongly correlated with each other. This information can guide feature selection steps before building predictive models

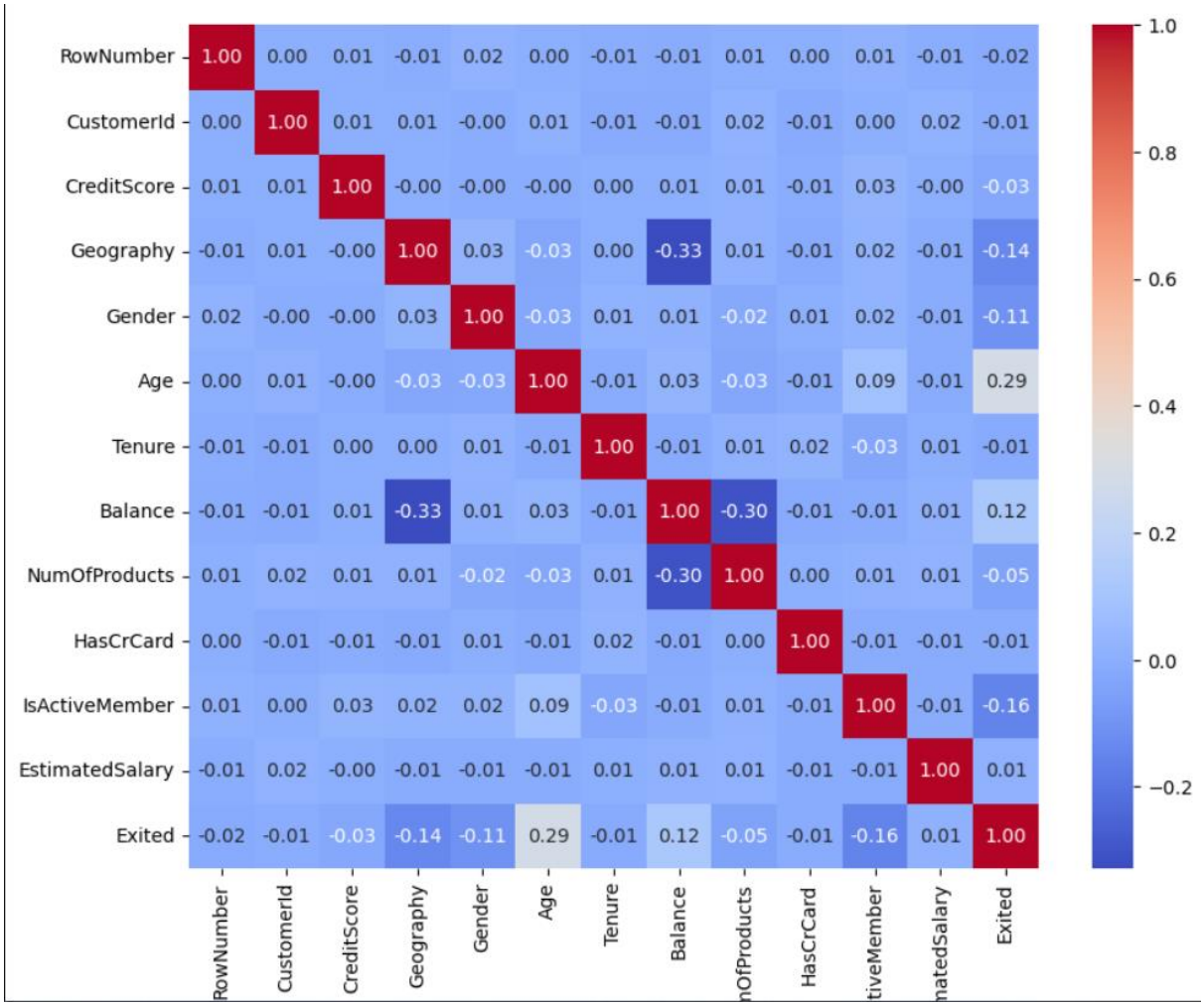

Fig. Correlation Matrix

Interpretation:

- **CreditScore vs. Exited**: There is a weak negative correlation (-0.027) between CreditScore and the likelihood of a customer exiting the bank. This suggests that customers with higher credit scores are slightly less likely to exit the bank.
- **Geography vs. Exited**: The correlation coefficient is -0.139, indicating a weak negative correlation between the geographic location and the likelihood of a customer exiting the bank. This suggests that customers from certain geographical locations might be slightly less likely to exit the bank compared to others.

- **Gender vs. Exited**: The correlation coefficient is -0.107, indicating a weak negative correlation between gender and the likelihood of a customer exiting the bank. This suggests that gender has a slight influence, with certain genders being slightly less likely to exit the bank.
- **Age vs. Exited**: There is a moderate positive correlation (0.285) between age and the likelihood of a customer exiting the bank. Older customers tend to be more likely to exit the bank compared to younger ones.
- **Tenure vs. Exited**: The correlation coefficient is -0.014, indicating a very weak negative correlation between tenure (the length of time the customer has been with the bank) and the likelihood of exiting the bank. This suggests that tenure has minimal impact on customer churn.
- **Balance vs. Exited**: There is a weak positive correlation (0.119) between the account balance and the likelihood of a customer exiting the bank. Customers with higher balances are slightly more likely to exit.
- **NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary vs. Exited**: These features show weak correlations with the likelihood of a customer exiting the bank. The correlations are relatively low and do not strongly indicate a significant relationship with customer churn.

Data Modelling:

Data Modelling often entails using both unsupervised and supervised learning techniques. Because we have tagged data, this dataset is intended for supervised learning.

1. Logistic Regression:

Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e. binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. The independent variables can be nominal, ordinal, or of interval type.

Confusion Matrix:-

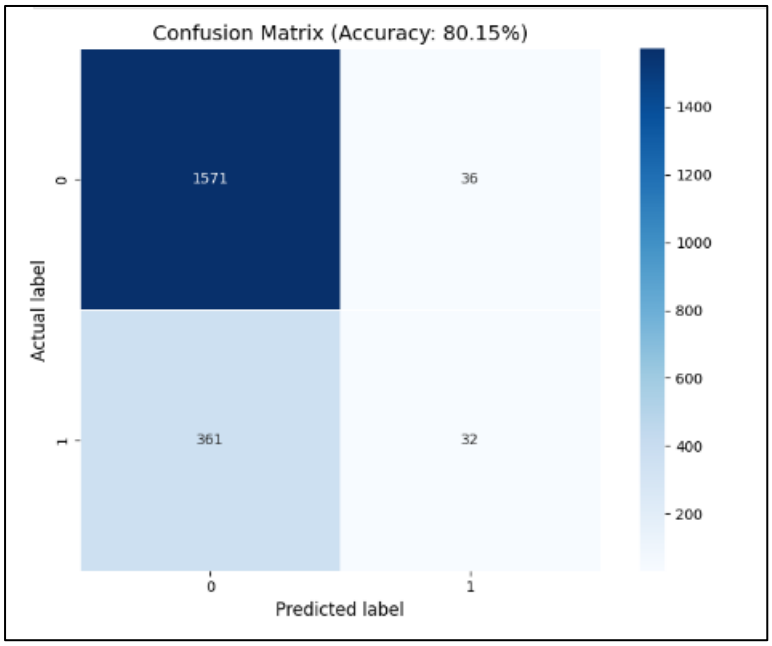

Fig.7 Logistic Regression

Logistic Regression Confusion Matrix (cm3):

- True Negatives (TN): 1571
- False Positives (FP): 36
- False Negatives (FN): 361
- True Positives (TP): 32

Interpretation:

Logistic Regression has the highest number of true negatives (1571), indicating its effectiveness in correctly identifying customers who do not churn.

Result :

Classification report:

	precision	recall	f1-score	support
0	0.81	0.98	0.89	1607
1	0.47	0.08	0.14	393
accuracy			0.80	2000
macro avg	0.64	0.53	0.51	2000
weighted avg	0.75	0.80	0.74	2000

Fig.8 Classification report of logistic regression

Interpretation:-

Class 0 (no churn) precision stands at 81%, which indicates that 81% of class 0 instances which were classified using the model were correctly predicted. The accuracy of churn class is found to be 47%, means that 47% of all instances corresponding to the class 1 were actually the class 1.

-T he classifier first accuracy of 98%, thereby correctly identified 98% of all actual class 0 instances. Class 1 recall is at 8%, so to make things clearer, that means only 8% of instances considered actual class 1 were captured by the classifier in this case.

- 89% is class 0's F1-score, and is 14% for class 1. A score gauge formed above this line helps to understand the balance between precision and recall for each class.

- The model accuracy of overall has 80% showing that it has capability to do the correct classification of instances across the both classes.

- The macro precision, recall and F1 measurements stand at 64%, 53%, and 51%, respectively, which means that precision, recall, and F1-score are the average over all classes.

- The weighted average precision, recall, and F1-score are been found to be 75%, 80%, and 74 considering the contribution of each class's support.

Random Forest:

Random Forest is an ensemble learning method that constructs various such decision trees as unsupervised processes and their predictions are used to obtain accuracy and robustness.

Confusion Matrix:

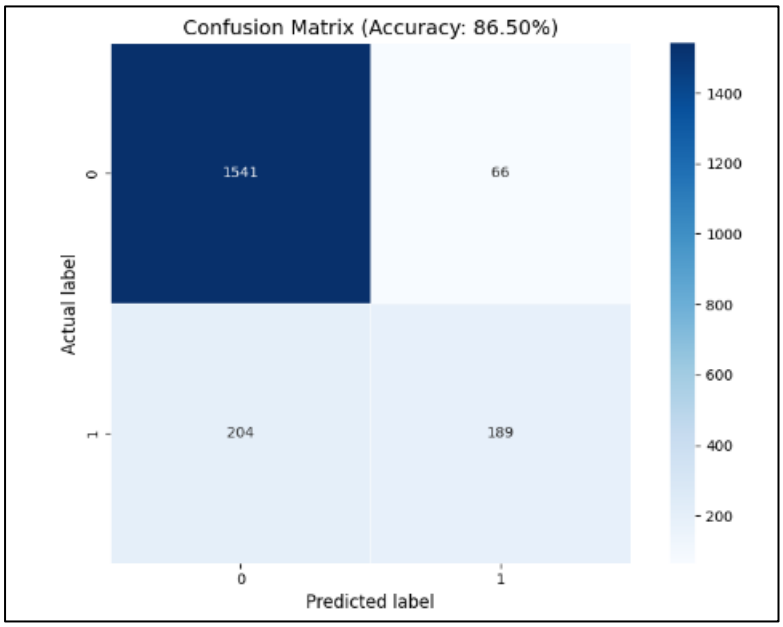

Fig.9 Random Forest

Random Forest Classifier Confusion Matrix (cm1):

- True Negatives (TN): 1547
- False Positives (FP): 60
- False Negatives (FN): 211
- True Positives (TP): 182

Interpretation:
The Random Forest classifier has the highest number of true positives (182) among the three models, indicating its effectiveness in correctly identifying customers who churn.

Result:

Classification report:

	precision	recall	f1-score	support
0	0.88	0.96	0.92	1607
1	0.74	0.48	0.58	393
accuracy			0.86	2000
macro avg	0.81	0.72	0.75	2000
weighted avg	0.86	0.86	0.85	2000

Fig. 10 Classification Report of Random Forest

Interpretation:-

- The precision for class 0 (non-churn) is 88%, this figure shows 88% of the class 0 instances that were predicted as class 0 were non-churns. class 1 (churn) with precision=0.74, i.e., 74% of the instances that were class 1 in actuality were correctly predicted 74% of the time.

- The recall for class 0 is 96% which performs as an acceptance inched by correctly identifying 96% of instances in class 0. As for the class 1 recall, it is 48%, which means that it captured 48% of all the actual class 1 instances.

- Binomials (F1-score for class₀) 92% and (F1-score for class₁) 58%. The total of the three scores show how precisely and accurately the prediction is made for every class.

- All in all the accuracy of the model as it pertains to both classes averages at 86% which shows it can correctly classify instances.

- The macro precision, recall, and F1-scores are 81%, 72%, and 75%. Thus the average performance for class-1 and class-2 together is 75%.

- The weighted average precision, the recall, and the F1-score (precision weights the class's support) score are 86%, 86%, and 85% respectively.

Evaluation and Comparison of Models :

Models	**Accuracy**
Logistic Regression	80.15%
Random Forest	86.50%

Table- Accuracy of 3 algorithms

Based on the accuracy scores obtained from the evaluation of the three models:Based on the accuracy scores obtained from the evaluation of the three models:

1. The classifier's Random Forest had the highest of accuracy of **86.50%.**
2. Logistic Regression, which is regarded, reached the accuracy of **80.15%.**

From reviewing Logistic Regression, Random Forest, and based on the dataset, it is clear that Random Forest and makes the best prediction with the highest prediction of 86.50%. This shows the highest accuracy result reaching up to 82.95% in the case of both classes while Logistic Regression (80.15%) are considered to be less effective.

Discussion

Using feature importance yielded that Age proved to be the factor with the highest impact in predicting the variable that is being targeted. This phenomenon puts into showcase the importance of age in decision making between whether the customer will leave or stay with the service. The ability of the model to suggest which feature most contributes to the prediction and thereby guides feature selection efforts, isolation of factors that are major drivers of the outcome accomplished and also improving model interpretability requires understanding which features contribute most to the model's predictions.

Consequently, the Random Forest model performance outshine and age as a affecting feature is important thereby it is recommended to over prioritize age-related insights by integration the Random Forest model into the decision process. Organizations can capitalize on the feature advantages of random forest, namely, having a high tolerance to overfitting, the ability to handle high dimensionality and ability to identify intricate non-linear relationships. Using this technique, they can arrive at insightful decisions based on their understanding of the reason for customer churn and accordingly develop targeted retention strategies that would enhance customer satisfaction with the aim of improving general business performance.

ROC Curve:-

A ROC curve is a line chart which illustrates the diagnostic capability of binary classifier systems and their performances with regard to discrimination threshold variations. It is plotted when a rope is used to suspend the body over the edge of the cliff.

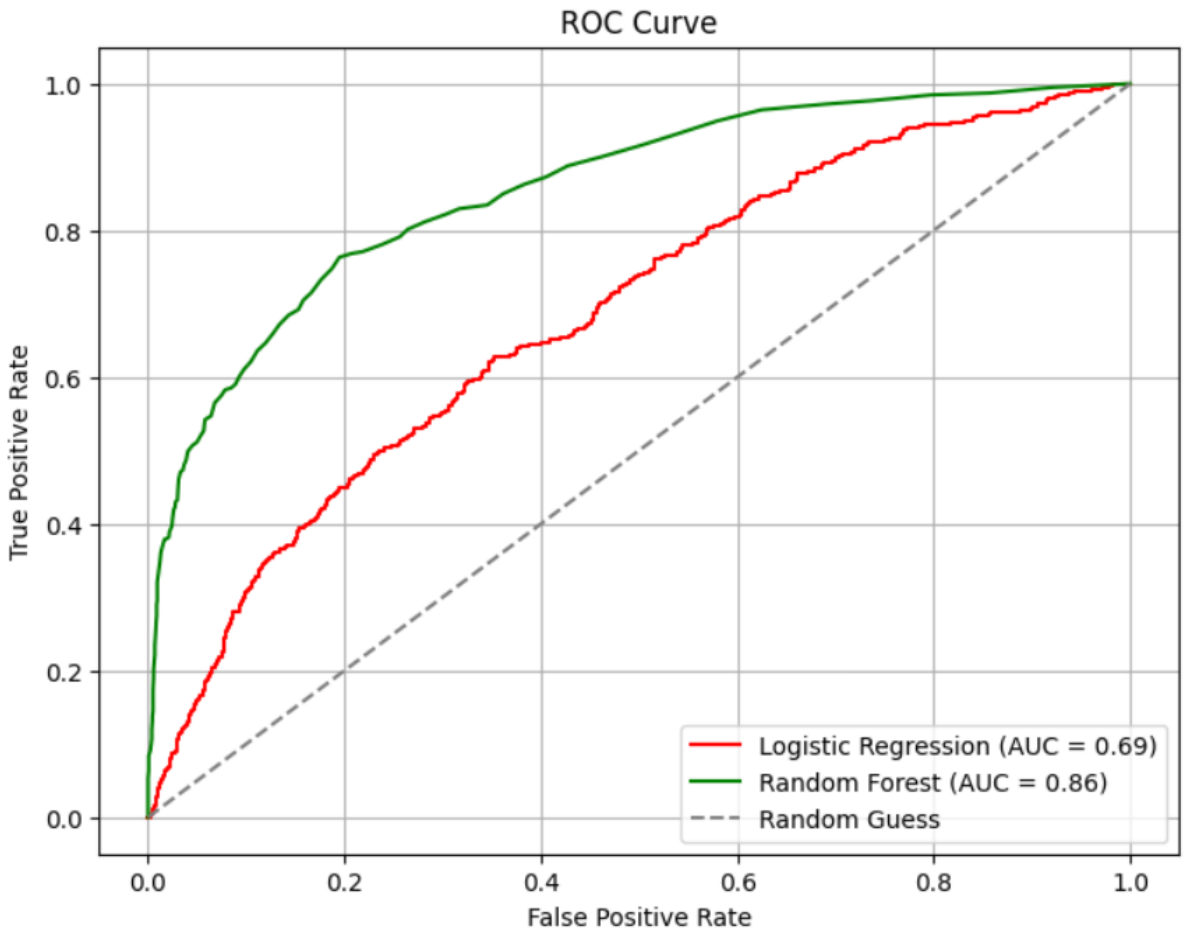

Fig.13 ROC curve

ROC curve is the model usefulness criteria for the two class problems. As y-axis is TPR, it plots the TPR on y-axis against the FPR on the x-axis.

The random forest model is the one with the highest AUC (Area Under the Curve) and our AUC(AUC) is 0.86. This means that random forest model is a top-performer among other models with regard to its ability to discern between positive and negative cases. statistically speaking it is the second-highest figure, in front of logistic regression model with AUC score 0.69.

The ROC curve, in general, reveals that random forest is on top of others at distinguishing between cases that are positive and cases that are not, followed by the logistic regression model.

Feature importance

Providing feeding and educational programs will help parents understand the importance of a nutritious diet.
Feature subset selection is a vital process of machine learning having multiple features. First of all, it solves the problem of dimensionality quandary, which means the more features there are, the more complex models become and the more computational cost is spent. As the space to explore is reduced by focused feature selection, overfitting is managed, and the model can generalise unseen data. Using Python script, we found out that: -

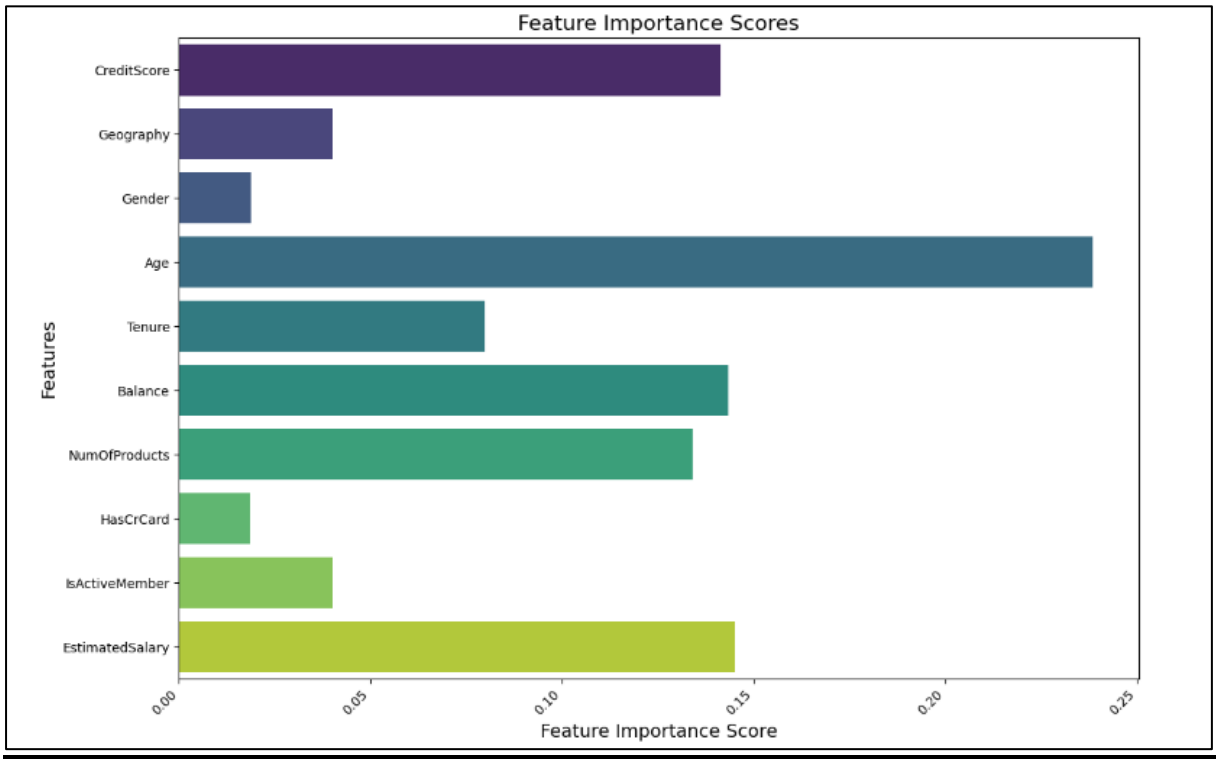

Fig.6 Feature Importance Score

Interpretation

The array feature_importances represents the importance scores assigned to each feature in a machine learning model, i.e, Random Forest classifier.

- CreditScore: 0.1414

- Geography: 0.0401
- Gender: 0.0187
- Age: 0.2387
- Tenure: 0.0798
- Balance: 0.1434
- NumOfProducts: 0.1341
- HasCrCard: 0.0186
- IsActiveMember: 0.0401
- EstimatedSalary: 0.1451

Each value represents the relative importance of the corresponding feature in predicting the target variable ,i.e, 'Exited'. Higher values indicate features that have a stronger influence on the model's predictions.

Interpretation:

- **Age** has the highest importance score (0.2387), indicating that it is the most important feature for predicting customer churn in the Random Forest model.
- **EstimatedSalary** also has relatively high importance (0.1451), suggesting that it contributes significantly to the model's predictions.
- **Balance** (0.1434) and **CreditScore (**0.1414) are also important features in predicting customer churn.
- **NumOfProducts** and Tenure have moderate importance scores (0.1341 and 0.0798, respectively).
- **Geography, Gender, HasCrCard, and IsActiveMember** have relatively lower importance scores compared to other features in the model.

Conclusion:

The evaluation of classification models is a very crucial part in the making of strategic choices by businesses for customer churn analysis. The contrast between three prevailing models, Random Forest, Logistic Regression, bring some distinct concepts of their prediction power, which can very successfully be used by the organizations aiming to suppress client attrition.

The Random Forest model outperforms the other two approaches, showing an outstanding evaluation issue of 86.50%. This is an indication that Random Forest is able to correctly assign classifications related to both churn and non-churn instances, and therefore it is indeed a heavyweight model for churn prediction duties. Then, analysis of ROC curve supports the fact that Random Forest emerge as the winner once again by demonstrating the highest Area Under Curve (AUC) of 0.86. The ROC analysis gives a clear picture of Random Forest's great discriminatory power, which is a strong point of this machine learning algorithm in segregating the churn cases with positive outcomes from those with negative.

This is where the Logistic Regression models have their edge over other models because their accuracy lies between 80.15% -77.75%, but they have moderate discriminatory power at 0.69 and 0.68 AUC respectively. Even if Logistic Regression models have high explanatory skills in identifying customer churns, it is the Random Forest model that is second to none in the area of its predictive power.

In addition, the importance of the features shows that the age is the most crucial characteristic to consider as a variable for client abandonment. Age is the one feature that is the standout influencing factor in comparison to the other features displayed, thus falling more into determining the behavior and churn probability of customers. Via identifying the importance of age and other meaningful features, companies can develop retention tactics that will cover the particular needs and views of different customer segments, hence increasing the customer satisfaction rate which is leading to an increase in their loyalty.

Therefore, it is observed that the outcomes shed light on the need to utilize high-grade machine learning methods such as Random Forest that enable the rate prediction to be done exactly.

Reference: -

1. https://www.kaggle.com/datasets/shantanudhakadd/bank-customer-churn-prediction?select=Churn_Modelling.csv
2. https://www.analyticsvidhya.com/blog/2022/09/bank-customer-churn-prediction-using-machine-learning/