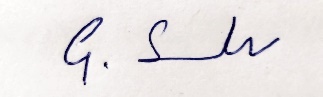


**Project Summary**

|  |  |
| --- | --- |
| Batch details | DSE- FT-Chennai, May 2024 – G5 |
| Team members | Ashwin G  Chitralekha G  Karthick V  Mohamed Umar Farook A A  Satheesh Kumar G |
| Domain of Project | BFSI |
| Proposed project title | Bank Customer Churn Prediction |
| Group Number | 5 |
| Team Leader | Satheesh Kumar G |
| Mentor Name | P V Subramanian |

Date:



Signature of the Mentor Signature of the Team Leader

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl NO** | **Topic** | **Page No** |
| 1 | ABSTRACT | 3 |
| 2 | INTRODUCTION | 3 |
| 2.1 OBJECTIVE | 3 |
| 2.2 CURRENT CHALLENGES | 3 |
| 3 | PROPOSED BUSINESS PROBLEM STATEMENT | 4 |
| 4 | INDUSTRY OVERVIEW | 4 |
| 4.1 BACKGROUND INDUSTRY REVIEW | 4 |
| 4.2 LITERATURE SURVEY | 5 |
| 5 | DATASET DESCRIPTION AND DOMAIN | 6 |
| 6 | DATA EXPLORATION (EDA) | 8 |
| 7 | FEATURE ENGINEERING | 11 |
| 8 | BASE MODEL | 13 |
| 9 | FUTURE ENHANCEMENT | 14 |
| 10 | CONCLUSION | 14 |

# **ABSTRACT**

This project focuses on predicting customer churn for an European MNC bank operating in France, Germany, and Spain. The goal is to identify factors contributing to customer attrition and build predictive models to forecast churn. By leveraging historical banking data, including customer demographics, transaction history, and account activity, machine learning algorithms will be applied to develop a churn prediction model. The insights gained will help the bank implement targeted retention strategies, reduce churn rates, and enhance customer satisfaction across the three key markets.

# **INTRODUCTION**

### **OBJECTIVE**

The objective of this project is to develop a robust customer churn prediction model for the European MNC bank operating in France, Germany, and Spain. By analysing historical data, including customer demographics, transaction patterns, and product usage, the bank aims to accurately predict which customers are at risk of leaving. The model will help the bank implement proactive retention strategies, such as personalized offers, targeted communication, and improved customer service, to reduce churn rates. Ultimately, the goal is to increase customer retention, enhance customer loyalty, minimize revenue loss, and strengthen the bank's competitive position in these markets.

### **CURRENT CHALLENGES**

# **PROPOSED BUSINESS PROBLEM STATEMENT**

The European MNC bank faces a significant challenge with customer churn across its branches in France, Germany, and Spain. High churn rates lead to revenue losses and increased customer acquisition costs. Understanding why customers leave is critical to addressing this issue. Factors such as competition, customer dissatisfaction, changing banking needs, and poor service could contribute to churn. The bank seeks to predict which customers are likely to leave to take proactive steps, such as offering personalized services, improving engagement, or addressing pain points. By reducing churn, the bank aims to improve customer retention and strengthen its market position.

# **INDUSTRY REVIEW**

## **Industry Review**

#### **Current Practices**

In the banking industry, customer churn prediction has become a critical focus as banks aim to retain customers in a highly competitive market. Current practices to predict and prevent customer churn generally involve the use of historical data, customer behaviour analysis, and targeted marketing strategies.

* **Historical Data Analysis**: Banks commonly rely on historical customer transaction data to detect patterns that indicate potential churn. Metrics such as account tenure, frequency of account usage, product holdings, and transaction volume are analysed to understand which customers are more likely to leave.
* **Segmentation and Profiling**: Many banks segment their customers based on demographics, transaction behaviour, and product usage. This segmentation allows banks to personalize retention efforts, targeting high-risk segments with tailored offerings.
* **Customer Feedback and Sentiment Analysis**: Banks frequently analyse customer feedback obtained through surveys, complaints, and support interactions. By analysing customer sentiment, banks can proactively address concerns and improve customer experience.
* **Predictive Analytics Models**: Traditional predictive models like logistic regression, decision trees, and random forest classifiers are used by banks to predict churn based on historical data. These models are often deployed to flag customers with a high likelihood of leaving.
* **Personalized Retention Programs**: Once high-risk customers are identified, banks offer personalized retention programs. These programs include loyalty rewards, product bundling, or discounts on fees, aimed at improving customer satisfaction and retention.

#### **Background Research**

Research into customer churn prediction has led to advancements in machine learning and data analytics methods specifically applied to the banking sector. Some key areas of background research include:

* **Machine Learning Techniques**: The banking industry has seen a shift from traditional statistical methods to more sophisticated machine learning approaches. Algorithms like Support Vector Machines (SVM), Gradient Boosting, and Neural Networks have shown higher accuracy and robustness in predicting churn as they can capture complex relationships within customer data.
* **Feature Engineering and Importance**: Research has highlighted the importance of feature engineering to improve model performance. Features such as customer lifetime value, monthly transaction counts, average balance, and even non-traditional features like web/app activity have been found useful in predictive models.
* **Customer Lifecycle Value Models**: A growing area in churn research is understanding customer lifecycle value. Retaining high-value customers is especially critical, so models that predict not only churn but also expected lifetime value help banks prioritize retention efforts effectively.
* **Time-Series Analysis**: Time-series analysis has become a popular approach as it considers changes in customer behavior over time. By applying time-series techniques, banks can detect trends, seasonality, and unusual patterns that may precede churn.
* **Sentiment and Text Mining**: In recent years, text mining and sentiment analysis of unstructured data, such as customer feedback and social media interactions, have provided valuable insights into customer emotions and satisfaction levels. This approach allows banks to address issues that might lead to churn more effectively.
* **Churn Prevention Strategies**: Background research also includes effective churn prevention strategies. Studies have shown that combining predictive analytics with proactive customer service measures—such as early intervention for at-risk customers and personalized retention programs—significantly reduces churn rates.

## **Literature Survey**

#### 1) Predictive Model Development

Early churn models relied on logistic regression for interpretability. However, machine learning models such as decision trees, random forests, and gradient boosting have since outperformed traditional methods in capturing complex patterns (Verbeke et al., 2012). Deep learning models, like LSTMs, effectively detect sequential behaviours in time-series banking data (Shaikh et al., 2019).

#### 2) Feature Engineering

Research identifies customer demographics, account balance, and transaction frequency as core features, with recency-frequency-monetary (RFM) values shown to be reliable predictors (Buckinx & Van den Poel, 2005). More recent studies use NLP for sentiment analysis from feedback, which improves prediction accuracy by capturing customer satisfaction levels (Ahmad & Baig, 2020).

#### 3) Time-Series and Sequential Analysis

Time-series models (e.g., ARIMA) and sequence-based deep learning (LSTMs, CNNs) have advanced churn predictions by leveraging temporal patterns in customer behaviour (Kumar & Ravi, 2016; Zhang et al., 2021).

#### 4) Sentiment Analysis Integration

Incorporating customer sentiment from reviews and complaints via NLP enhances prediction accuracy. Studies show that sentiment is strongly tied to churn rates and enriches models when combined with structured data (Chen et al., 2018).

#### 5) Retention Strategies and Cost-Sensitive Learning

Retention strategies prioritize high-value at-risk customers. Cost-sensitive models improve accuracy and profitability by emphasizing correct classifications of high-value customers (Neslin et al., 2006; Liu & Shi, 2020).

#### 6) Model Evaluation and Interpretability

Besides accuracy, interpretability tools like SHAP and LIME are essential for stakeholder transparency in financial predictions (Ribeiro et al., 2016). These methods clarify which features drive churn predictions, crucial for decision-making in banking.

# **Dataset and Domain**

## **Data Dictionary**

Please refer to the appendix for more details.

## **Variable categorization (count of numeric and categorical)**

**Numerical variables**

**There are 14 numerical variables, inclusive of 1 Target Variable and listed below:**

**1 ) RowNumber**

**2 ) CustomerId**

**3 ) CreditScore**

**4 ) Age**

**5 ) Tenure**

**6 ) Balance**

**7 ) NumOfProducts**

**8 ) HasCrCard**

**9 ) IsActiveMember**

**10 ) EstimatedSalary**

**11 ) Exited (Target Variable)**

**12 ) Complain**

**13) Satisfaction Score**

**14 ) Point Earned**

**Categorical variables**

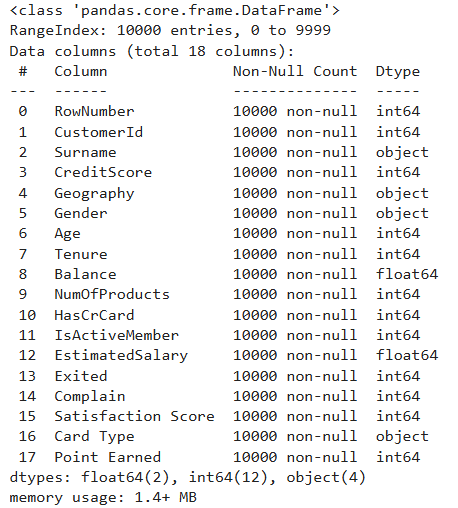
**There are 4 categorical variables and listed below:**

**1 ) Surname**

**2 ) Geography**

**3 ) Gender**

**4 ) Card Type**



.

## **Pre-Processing Data Analysis (count of missing/ null values, redundant columns, etc.)**

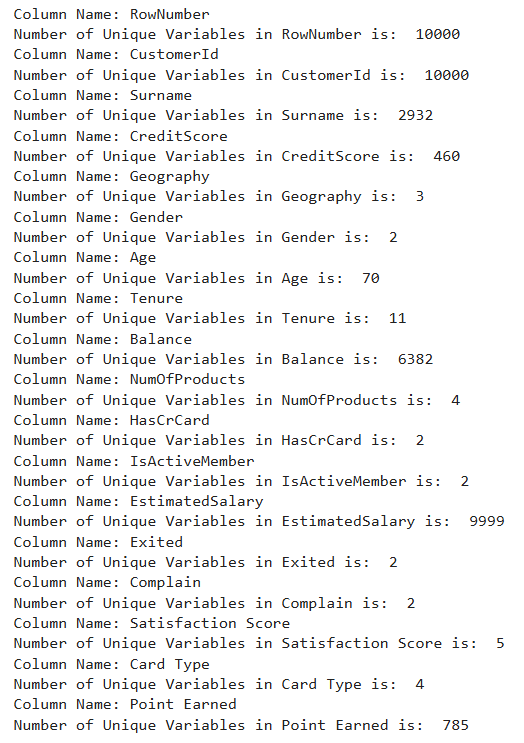
### **Count of Missing values**

There are 0 columns that have missing values.



### **Redundant columns or unwanted columns**

Redundant columns are RowNumber, CustomerId and Surname and they are dropped.



### **Duplicate rows**

There are no duplicates in our dataset.

## **Alternate sources of data that can supplement the core dataset (at least 2-3 columns)**

The additional sources may include Price Perception, Trust.

## **Project Justification - Project Statement, Complexity involved, Project Outcome –Commercial, Academic or Social value**

**Project Statement:** To develop a robust customer churn prediction model for the European MNC bank operating in France, Germany, and Spain.

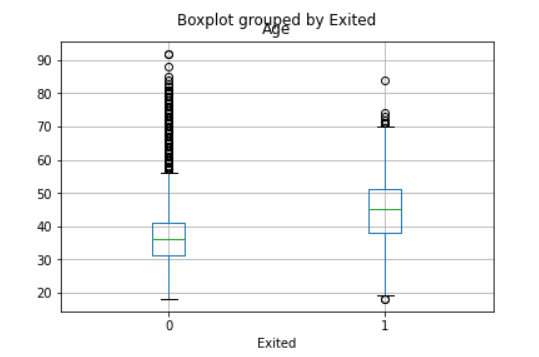
**Complexity involved:** Low

**Project Outcome –Commercial, Academic or Social value:** Commercial

# **Data Exploration (EDA)**

## **Relationship between variables**

Increase in Age are more likely to Churn.



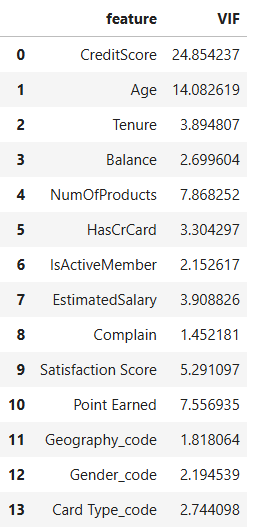
## **Check for**

* 1. **multi-collinearity**

The higher the value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high.

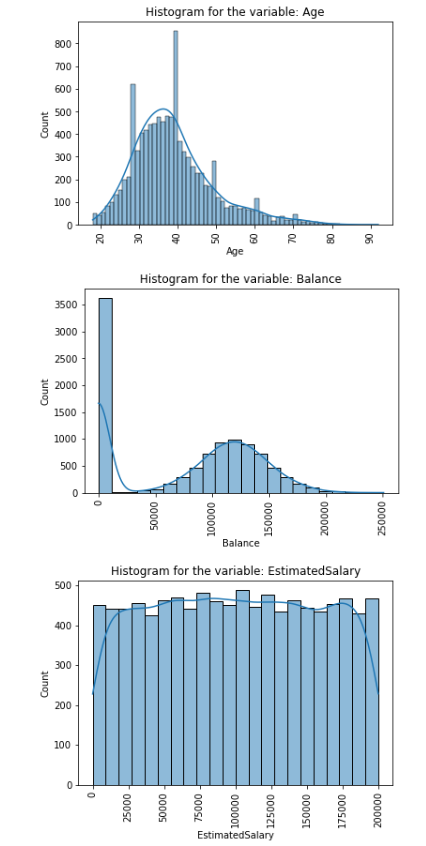
1. CreditScore, 2) Age, 3) NumOfProducts, 4) Satisfaction Score and 5) Point Earned.

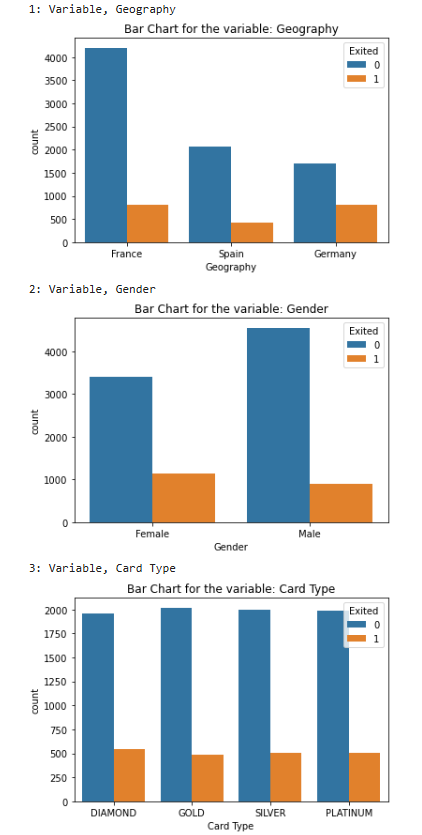
We need to remove variables having greater than the threshold value of 5 from our dataset.

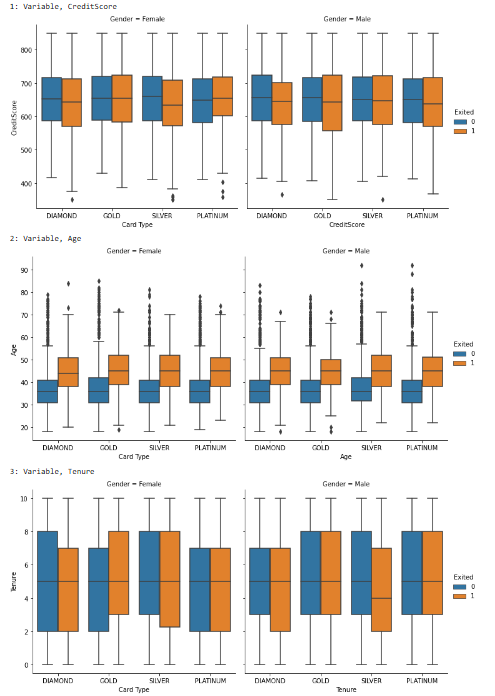


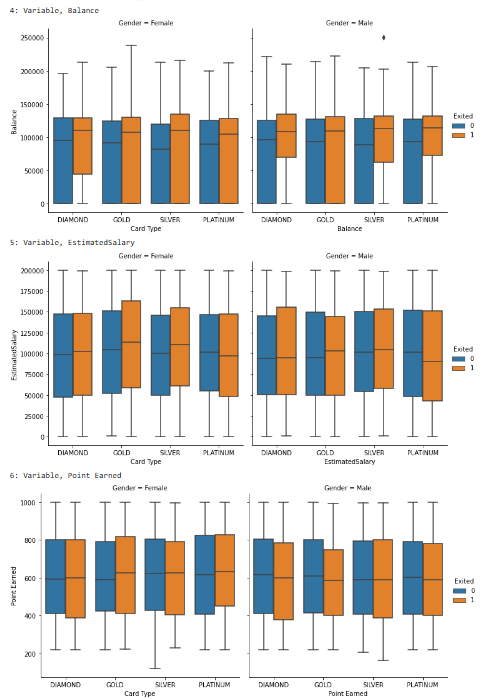
* 1. **distribution of variables**

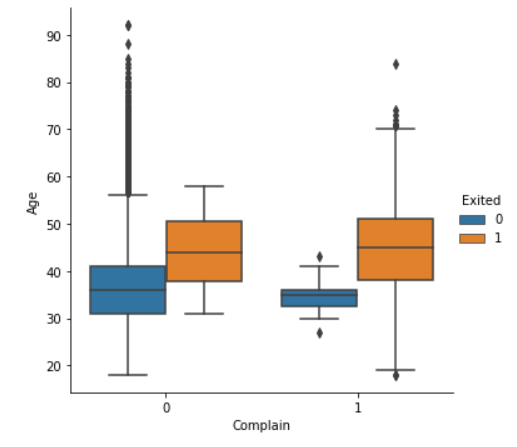
None of the continuous numerical variables ('Age', 'Balance', 'EstimatedSalary') are normally distributed.









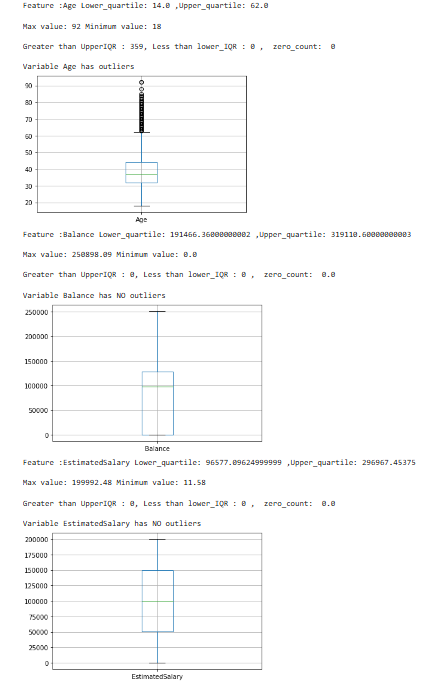


* 1. **Presence of outliers and its treatment**

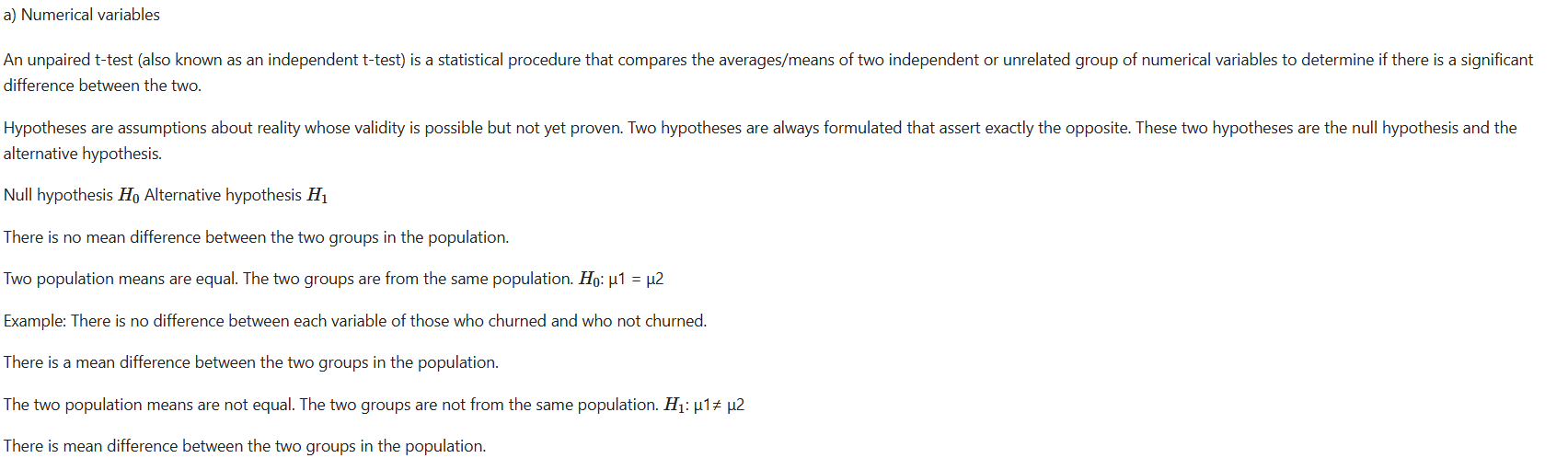
Outliers badly affect mean and standard deviation of the dataset. · It increases the error variance and reduces the power of statistical tests. By applying outlier treatment, machine learning practitioners can handle extreme values effectively. The primary goals of outlier treatment are: Identifying Outliers: Through various statistical methods, such as visualizations and mathematical approaches, outliers can be detected within a dataset.

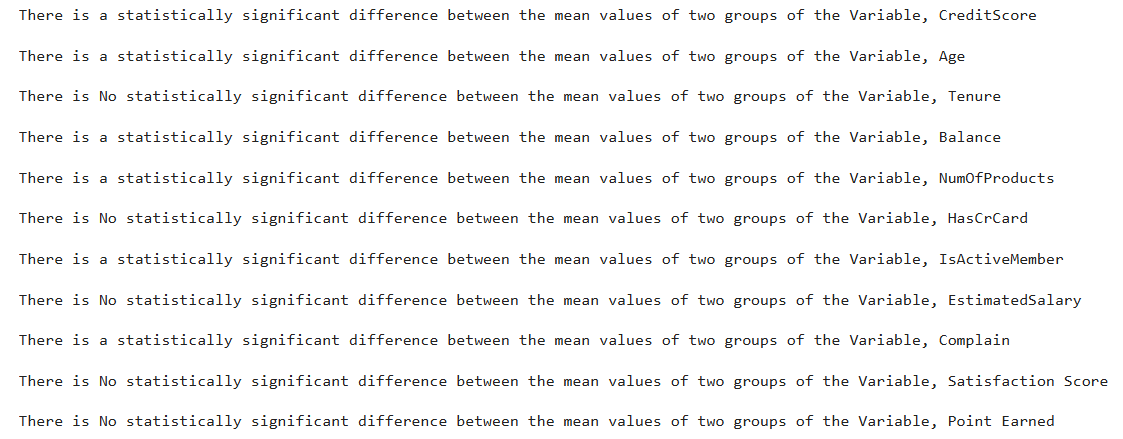
We are interested to identify the outliers in our continuos numerical variables such as 'Age', 'Balance', 'EstimatedSalary' that affects the mean & standard deviation rather than the discrete numerical variables. Discrete variables are typically categorical, meaning they take on a limited number of values or categories.

### However, if the outlier is physically possible you should consider it.



* 1. **statistical significance of variables**

****

****

1. **Normality Tests**

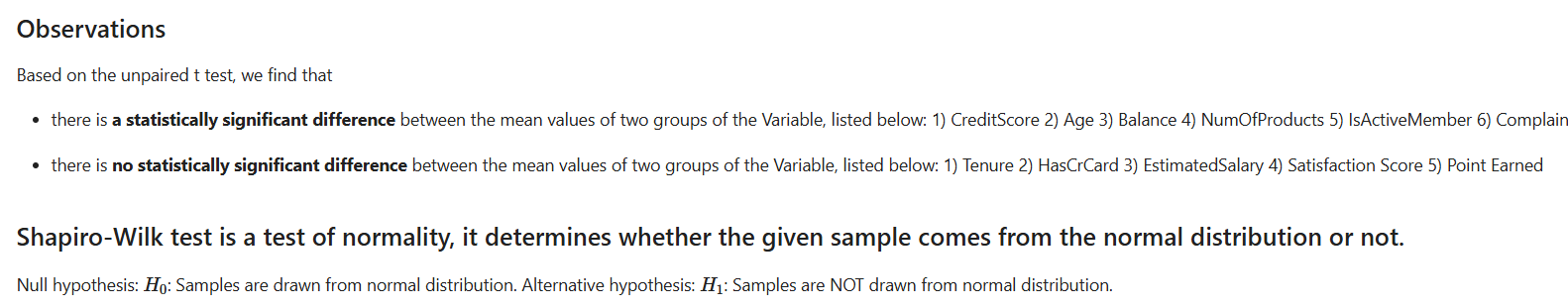
**Shapiro-Wilk Test:** Tests whether a data sample has a Gaussian distribution.

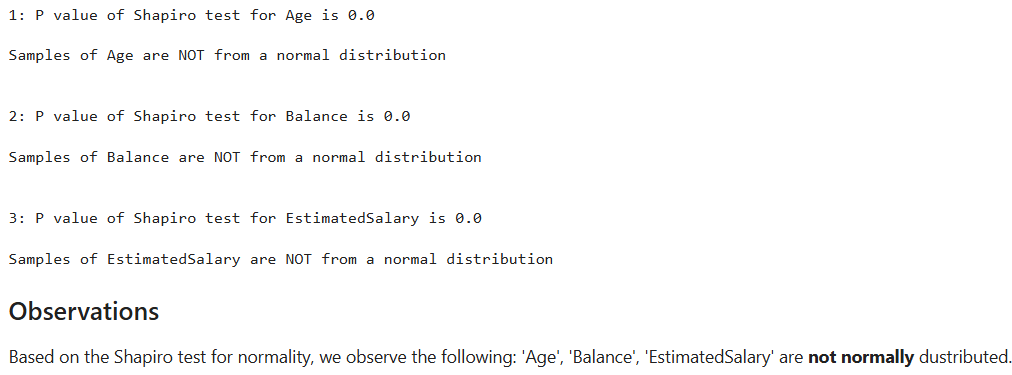
**Assumptions:** Observations in each sample are independent and identically distributed (IID).

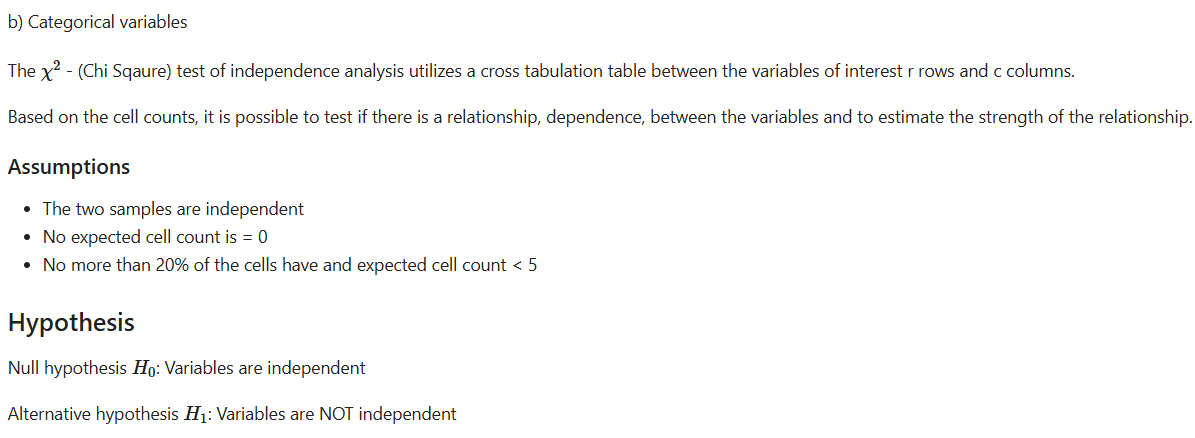
**Interpretation**

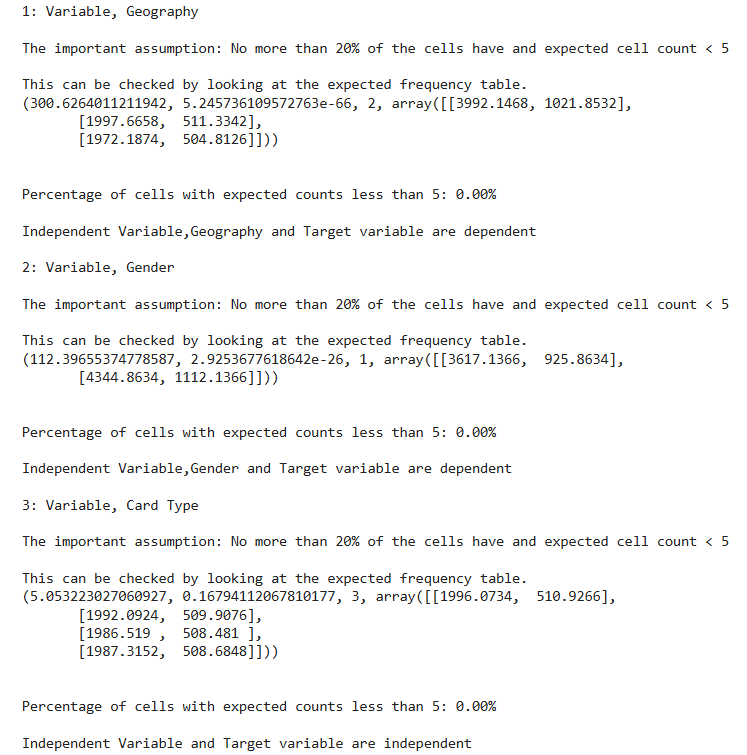
H0: the sample has a Gaussian distribution.

H1: the sample does not have a Gaussian distribution.









Observations:

We have checked each categorical independent variable with our target categorical variable using Chi Square test of independence. We observed the assumption for chi-square test of independence *(No more than 20% of the cells have and expected cell count < 5)* is satisfied.

**Each of the independent variables, 'Geography', 'Gender' and the target variable, Exited are dependent.**

**Independent variable, 'Card Type' and the target variable, Exited are independent.**

1. **Calculate Correlation Between Continuous & Binary Target Variable**

Point biserial correlation is used to calculate the correlation between a binary categorical variable (a variable that can only take on two values) and a continuous variable and has the following properties:

Point biserial correlation can range between -1 and 1. For each group created by the binary variable, it is assumed that the continuous variable is normally distributed with equal variances. For each group created by the binary variable, it is assumed that there are no extreme outliers.

The hypotheses for point biserial correlation thus result in:

Null hypothesis: The correlation coefficient r = 0 (There is no correlation)

Alternative hypothesis: The correlation coefficient r ≠ 0 (There is a correlation)

There is no correlation between the following variables and the Target binary variable:

There is a correlation between the following variables and the Target binary variable:

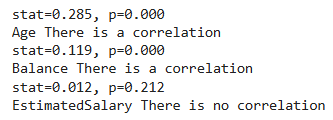
**Chi square test of independence**

The objective is to determine whether the association between two qualitative variables is statistically significant. The formulation of the hypotheses for this statistical analysis is something like this.

Null Hypothesis (H0): There is no substantial relationship between the two variables (in case of independence test), or there is no difference

Alternative Hypothesis (H1): There is a substantial relationship between the two variables (in case of independence test), or there is a difference

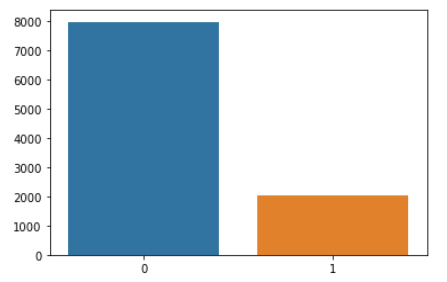
Read more at: <https://analyticsindiamag.com/ai-mysteries/how-to-use-the-chi-square-test-for-two-categorical-variables/>



* 1. class imbalance and its treatment

So, we need to choose the model performance measure carefully to avoid bias to the majority class.

Precision and recall are common metrics used when evaluating classification models for detection of a certain important class. Recall represents how many samples are important. class was discovered by the model of all the samples in the class, while precision represents the accuracy of predictions for that certain class. In this application, the important class is "Suspected Fraud". True positives are the number of correctly identified data points of the important class, False positives is the number of data points incorrectly identified as important and False negatives is the number of data points incorrectly identified as not important.



**SMOTE**

The problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary.

One way to solve this problem is to oversample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model.

An improvement on duplicating examples from the minority class is to synthesize new examples from the minority class. This is a type of data augmentation for tabular data and can be very effective.

Perhaps the most widely used approach to synthesizing new examples is called the Synthetic Minority Oversampling Technique, or SMOTE for short. This technique was described by Nitesh Chawla, et al. in their 2002 paper named for the technique titled “SMOTE: Synthetic Minority Over-sampling Technique.”

SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbors for that example are found (typically k=5). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space.

We shall apply SMOTE on training data.

# **Feature Engineering**

## **Whether any transformations required**

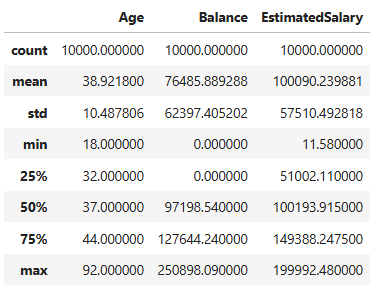
Data transformation is used when data needs to be converted to match that of the destination system.

We have performed label encoding to make our data suitable for model building.

## **Scaling the data**

Data scaling is applied to numeric columns. In our dataset we have three continuous numerical columns:

1. Age
2. Balance
3. EstimatedSalary



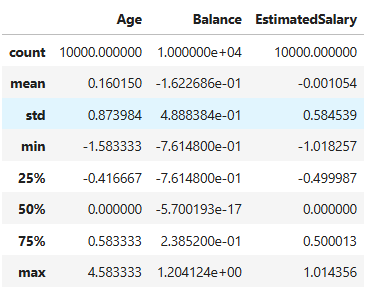
### Robust scaling

Both standard and robust scalers transform inputs to comparable scales. The difference lies in how they scale raw input values.

Standard scaling uses mean and standard deviation. Robust scaling uses median and interquartile range (IQR) instead.

Robust scaling answers a simple question. How far is each data point from the input’s median?

### The fact that robust scaling uses median and IQR makes it resistant to outliers.



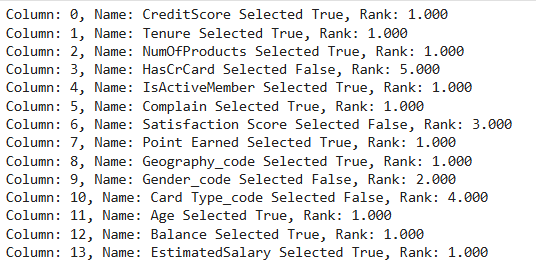
## **Feature selection**

### **Automatically select the number of features**

The RFE method is available via the RFE class in scikit-learn.

RFE is a transform. To use it, first the class is configured with the chosen algorithm specified via the “estimator” argument and the number of features to select via the “n\_features\_to\_select” argument.

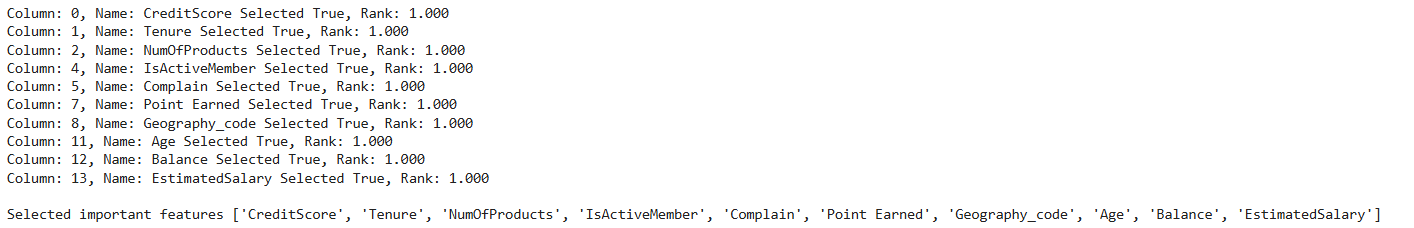
The algorithm must provide a way to calculate important scores, such as a decision tree. The algorithm used in RFE does not have to be the algorithm that is fit on the selected features; different algorithms can be used.



Once configured, the class must be fit on a training dataset to select the features by calling the fit() function. After the class is fit, the choice of input variables can be seen via the “support\_” attribute that provides a True or False for each input variable.

It can then be applied to the dataset by calling the transform() function.

We can see the RFE that uses a Random Forest and selects 10 features and then fits a model on the selected features achieves a balanced accuracy of about 99.8 %.



**Balanced accuracy** in binary and multiclass classification problems to deal with imbalanced datasets. It is defined as the average of recall obtained on each class.

Fit an RFE model on the whole dataset and selects five features, then reports each feature column index (0 to 9), whether it was selected or not (True or False), and the relative feature ranking.

The “support\_” attribute reports true or false as to which features in order of column index were included and the “ranking\_” attribute reports the relative ranking of features in the same order.

Selected important features to predict the target variable:

We shall use these features in our model building.

## **Dimensionality reduction**

Since we have selected top 10 variables affecting the dependent variable, our dataset is Not Huge. **We are not going to apply dimensionality reduction such as Principal Component Analysis or Factor Analysis etc**. in our project.

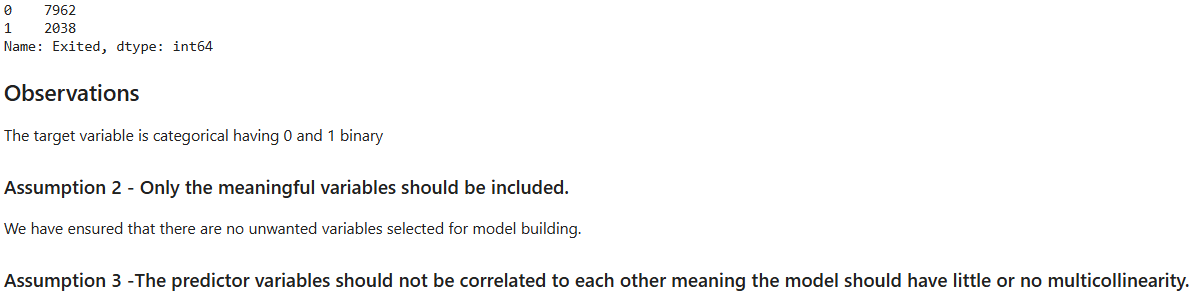
We use PCA when you have high-dimensional data to reduce its dimensionality while preserving most of the variance, simplifying analysis and visualization.

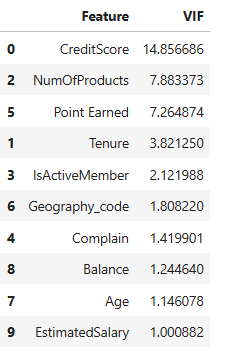
# **Assumptions**

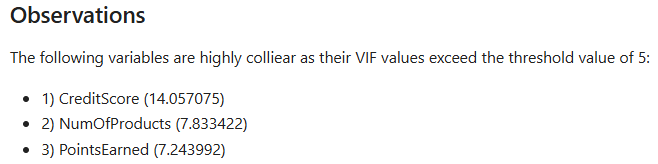
* Check for the assumptions to be satisfied for each of the models in
  + Classification – Decision Tree, Random Forest, Bagged and boosted models

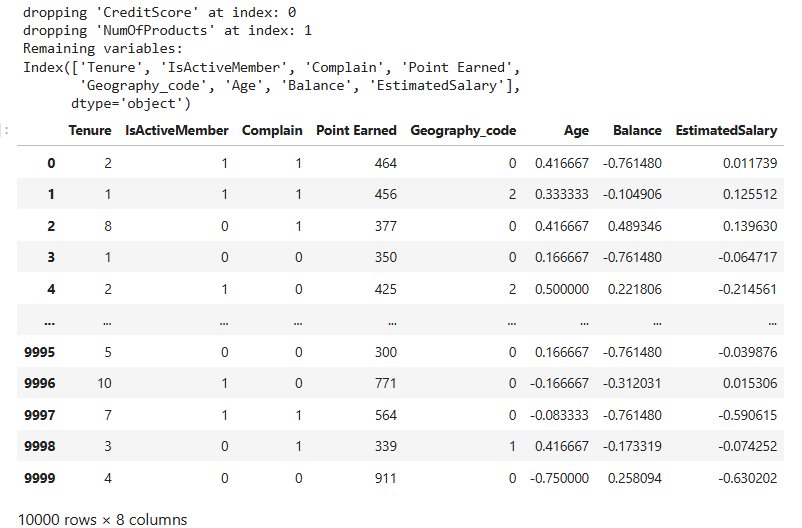
**Base Model**

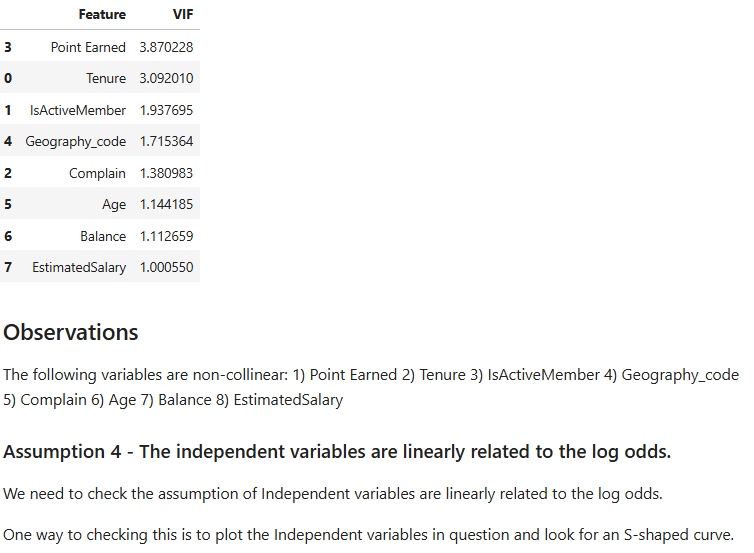
Logistic regression is widely used as a base model in customer churn prediction due to its simplicity, interpretability, and effectiveness for binary classification problems. It models the relationship between a set of independent variables and a binary outcome—in this case, whether a customer will churn (1) or stay (0).

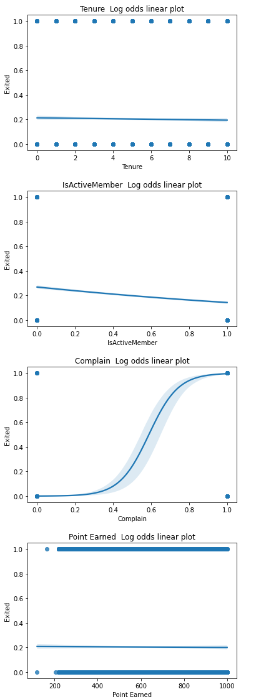
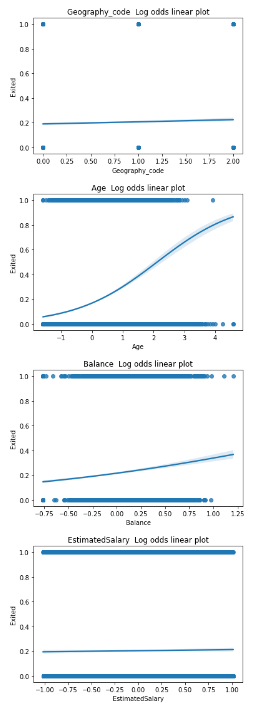






****

****

****

**Assumption 5 - Logistic regression requires quite a large number of observations.**

Number of events: 2038

Number of predictor variables: 10

Events per predictor: 203.8

Explanation:

We calculate the number of events by summing the 'Exited' column, which represents the cases where the outcome of interest occurs.

We calculate the number of predictor variables by counting the number of columns in the DataFrame and excluding the outcome variable.

We divide the number of events by the number of predictor variables to get the events per predictor.

We can then compare the calculated events per predictor with the recommended guideline of 10-20. If the ratio is below this guideline, it may indicate a potential violation of the assumption of a sufficiently large sample size.

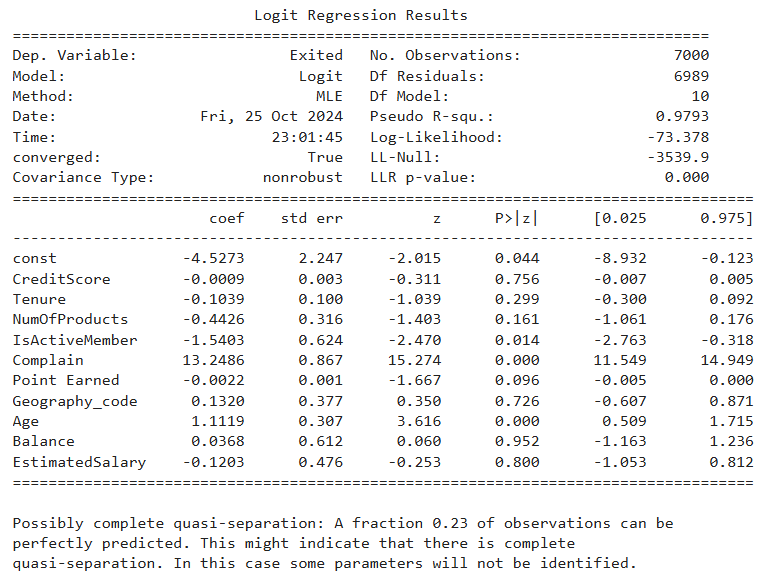
Observations:

With 2038 events and 10 predictor variables, the calculated number of events per predictor is approximately 203.8. This exceeds the commonly recommended guideline of having at least 10-20 events per predictor variable.

Inference: The dataset appears to meet the assumption of having a sufficiently large sample size for logistic regression.

Having a high number of events per predictor variable suggests that there should be adequate statistical power and precision in estimating the model parameters, enhancing the reliability of the logistic regression analysis. Therefore, the dataset likely provides a robust basis for fitting a logistic regression model and conducting statistical inference.

**Report Psuedo R-square, model coefficients and p-value**

****

**Observation**

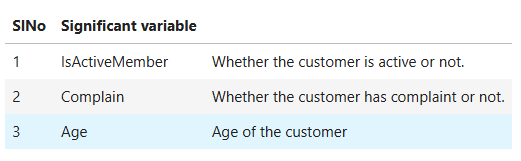
We observe that the McFadden R square (Pseudo R square) is 98.20 % and the model fitness is very good. This McFadden approach is one minus the ratio of two log likelihoods. The numerator is the log likelihood of the logit model selected and the denominator is the log likelihood if the model just had an intercept.

A goodness of fit using McFadden‟s pseudo r square (ρ^2) is used for fitting the overall model. McFadden suggested ρ^2 values of between 0.2 and 0.4 should be taken to represent a very good fit of the model (Louviere et al.,2000). <http://www.lifesciencesite.com/lsj/life1002/286_B01288life1002_2028_2036.pdf>

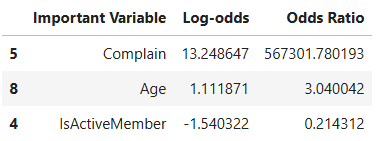
**List the significant variables at 5% level of significance**

Observation

The following variables are significant at 5 % level of significance:

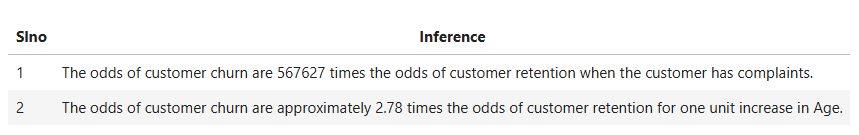


**Get Odds ratio**



**Odds Ratio Interpretation for significant variables**

Holding other things constant:

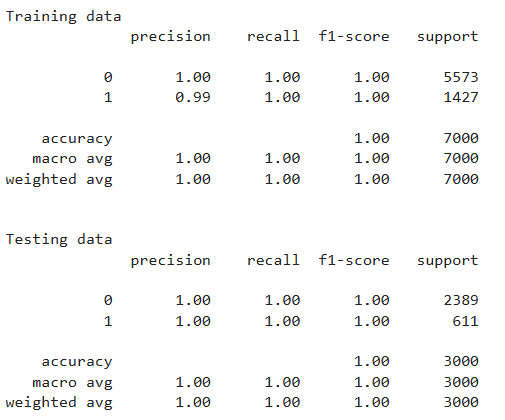


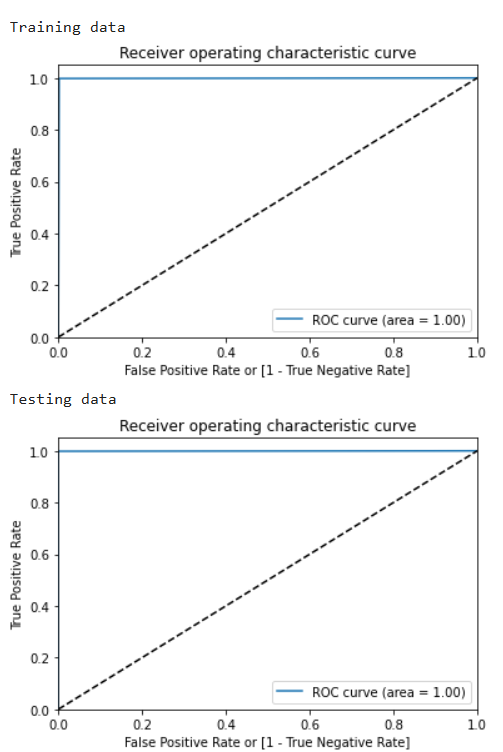
**Binary Classification**  
Logistic regression is designed to estimate the probability of a binary outcome. The model calculates the probability of a customer churning based on the given features, typically producing a probability score that is converted into a binary prediction.

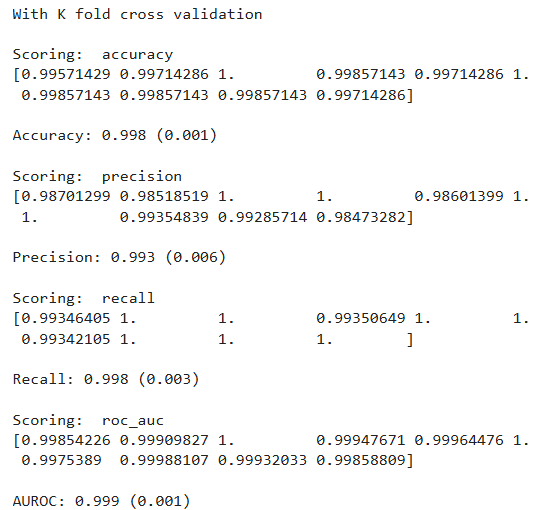
**Model Evaluation Metrics**  
Logistic regression’s performance is often evaluated with metrics like:

* **Accuracy**: The overall correctness of predictions.
* **Precision and Recall**: Particularly important in churn prediction to assess how well the model identifies true churners.
* **ROC-AUC**: The Area Under the Receiver Operating Characteristic curve helps in understanding the model's ability to distinguish between churners and non-churners.









In banking churn prediction, logistic regression is valuable for its straightforward interpretability, which is crucial for decision-making. It provides insights into how factors such as customer tenure, account activity, and product usage impact churn likelihood. By understanding these drivers, banks can develop targeted strategies to retain customers.

### **9) Future Enhancements**

### **Incorporation of Advanced Machine Learning Models** Beyond logistic regression, more complex algorithms such as Random Forest, Gradient Boosting, or even neural networks could be introduced to capture complex, non-linear relationships within customer data. Ensemble methods, in particular, could improve prediction accuracy by combining the strengths of multiple models.

### **10) Conclusion**

The bank customer churn prediction model is a strategic tool designed to help financial institutions retain valuable customers by identifying those at risk of leaving. By leveraging data-driven insights, the model provides a systematic approach to understanding customer behaviour, identifying factors that contribute to churn, and enabling proactive retention strategies. Through the use of logistic regression, we have established a strong baseline model that offers reliable predictions on customer churn risk. This initial model demonstrates the potential of data analytics to inform decision-making and improve customer relationship management.

**APPENDIX:**

**DATA DICTIONARY:**

| **FIELDS** | **DESCRIPTION** |
| --- | --- |
| **RowNumber** | **Corresponds to the record (row) number** |
| **CustomerId** | **contains random values** |
| **Surname** | **the surname of a customer** |
| **CreditScore** | **Credit Score of a Customer** |
| **Geography** | **Customer’s Location** |
| **Gender** | **Gender of a Customer** |
| **Age** | **Age of a Customer** |
| **Tenure** | **the number of years that the customer has been a client of the bank** |
| **Balance** | **Balance of a Customer** |
| **NumOfProducts** | **the number of products that a customer has purchased through the bank** |
| **HasCrCard** | **denotes whether a customer has a credit card** |
| **IsActiveMember** | **denotes whether a customer is Active** |
| **EstimatedSalary** | **Estimated Salary of a Customer** |
| **Exited** | **whether or not the customer left the bank. ( Target variable )** |
| **Complain** | **customer has complaint or not** |
| **Satisfaction Score** | **Score provided by the customer for their complaint resolution.** |
| **Card Type** | **type of card holds by the customer** |
| **Points Earned** | **the points earned by the customer for using credit card** |