## Artificial

## Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

Customer Churn Prediction

A red and white sign

Description automatically generated with low confidence

Submitted By:

Prachi Singla (2210990662)

Pratistha Negi (2210990678)

Punal (2210990690)

Supervised By:

Dr. Kirandeep Singh

**Department of Computer Science and Engineering**

Chitkara University Institute of Engineering & Technology,

Chitkara University, Punjab

TABLE OF CONTENTS

|  |  |
| --- | --- |
| TOPICS |  |
| 1. INTRODUCTION |  |
| 1. PROBLEM DEFINITION |  |
| 1. PROPOSED DESIGN |  |
| 1. METHODOLOGY |  |

|  |  |
| --- | --- |
| FIG. NO | FIGURE NAME |
| I. | COUNTPLOT (An Imbalanced Dataset) |
| II. | TENURE VS DENSITY |
| III. | GEOGRAPHY BASED EXITED (COUNTPLOT) |
| IV. | GENDER BASED EXIT |
| V. | TENURE BASED EXIT |
| VI. | ACTIVE MEMBER BASED EXIT |
| VII. | IS ACTIVE MEMBER VS NUM OF PRODUCTS  (LINEPLOT) |
| VIII. | CREDIT SCORE BASED EXIT |
| IX. | CORRELATION MATRIX |
| X. | BOX-PLOT BASED ON DIFFERENT PARAMETERS |
| XI. | AGE VS COUNT |

FIGURE TABLE **INTRODUCTION**

**WHAT IS CUSTOMER CHURN?**

Customer churn, also known as customer attrition or turnover, refers to when customers end their relationship with a business or discontinue using its products or services.

* **Critical Metric:** It is a crucial metric for businesses across industries, typically measured as the percentage of customers who have stopped using a company's products or services within a specific time frame.
* **Reasons for Churn:** Churn can occur due to various reasons such as dissatisfaction with product or service quality, better offers from competitors, changes in customer needs or preferences, or external factors like economic conditions or life events.
* **Impact on Business**: Reducing customer churn is vital as it directly affects revenue, profitability, and long-term sustainability of the business.
* **Retention Strategies:** Understanding the reasons behind customer churn is key to implementing effective retention strategies. These strategies aim to improve customer satisfaction, loyalty, and ultimately, the bottom line of the business.

## ****BACKGROUND****

## **(https://www.kaggle.com/code/simgeerek/churn-prediction-using-machine-learning)**

## ****Customer demographics:** This includes details such as age, gender, Credit Score ,Geography, Age , Tenure, Balance , Num Of Products, Has Cr Card, Is Active Member, Estimated Salary.**

## **.**Customer interactions:** Records of customer interactions with the bank, such as inquiries, complaints, feedback, etc.**

## ****Churn label:** A binary indicator (1 or 0) representing whether a customer has churned (1) or not (0) within a specific time period after the data snapshot.**

## **Using techniques includes logistic regression, decision trees, random forests, support vector machines, and neural networks.**

**OBJECTIVE**

* The objective of studying customer churn typically revolves around several key goals:

1. **Identifying Risk Factors :** Understanding the factors that contribute to customer churn helps businesses identify potential risk factors that may lead to customer attrition.
2. **Predictive Analysis** : Analyzing historical data and customer behavior patterns allows businesses to develop predictive models that forecast the likelihood of churn for individual customers.
3. **Maximizing Customer Lifetime Value (CLV)** : Customer churn directly impacts a business's revenue and profitability. By reducing churn rates and extending customer lifetime value, businesses can maximize the return on their investment in customer acquisition and retention efforts.
4. **Improving Customer Experience :** Addressing the root causes of churn often involves improving the overall customer experience. Businesses can enhance customer satisfaction and loyalty, leading to lower churn rates and increased profitability.

**SIGNIFICANCE**

1. **Revenue Impact:** Churn directly affects revenue by reducing the customer base, leading to loss of recurring revenue streams.
2. **Customer Lifetime Value (CLV):** Churn decreases CLV as it shortens the duration over which a customer generates revenue.
3. **Profitability:** Retaining existing customers is typically more cost-effective than acquiring new ones. Churn undermines this profitability.
4. **Market Perception:** High churn rates can signal dissatisfaction with the product or service, affecti02ng brand reputation and market perception.
5. **Business Growth:** Churn inhibits sustainable growth as it requires constant acquisition efforts to offset losses.
6. **Data Insights:** Analyzing churn patterns can provide valuable insights into product weaknesses, customer preferences, and areas for improvement.
7. **Competitive Edge:** Lower churn rates can give a competitive edge by demonstrating better customer satisfaction and loyalty compared to competitors.
8. **Investor Confidence:** Churn rates are often scrutinized by investors as they reflect the health and potential growth trajectory of a business.
9. **Customer Success**: Managing churn effectively is essential for ensuring customer success and fostering long-term relationships.
10. **Predictive Modelling:** Understanding churn behaviour enables predictive modelling, allowing proactive measures to retain customers before they churn.

**PROBLEM DEFINITION**

In churn prediction for a bank dataset, the problem typically involves identifying customers who are likely to stop using the bank's services or close their accounts.

1. **Data Collection and Preprocessing:** Gather relevant data such as customer demographics, transaction history, account activity, etc. Preprocess the data by handling missing values, encoding categorical variables, and scaling numerical features.

**2. Feature Engineering**: Create new features or transform existing ones that might be predictive of churn. This could include metrics such as frequency of transactions, average balance, customer tenure, etc.

**3. Train-Test Split:** Split the dataset into training and testing sets to evaluate the performance of the predictive model.

**4. Model Selection:** Choose an appropriate machine learning model for churn prediction. Commonly used models include logistic regression, decision trees, random forests, gradient boosting machines, and neural networks.

**5. Model Training:** Train the selected model using the training data. Tune hyperparameters using techniques like grid search or random search to optimize model performance.

**6. Model Evaluation:** Evaluate the trained model using the testing data. Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC can be used to assess its performance.

**7. Feature Importance:** Analyze feature importance to understand which variables have the most significant impact on churn prediction. This can provide insights for the bank to focus on certain customer attributes or behaviours.

**8. Model Deployment:** Deploy the trained model into production to make predictions on new data. Integrate it into the bank's systems for real-time or batch prediction of customer churn.

**9. Monitoring and Maintenance:** Continuously monitor the model's performance over time. Retrain the model periodically with updated data to ensure its accuracy and relevance.

**10. Business Action:** Translate model predictions into actionable insights for the bank. Implement retention strategies such as personalized offers, targeted marketing campaigns, or improved customer service to reduce churn and retain valuable customers.

**REQUIREMENTS**

**1. Data:** Acquire a comprehensive dataset containing relevant information about customers, such as demographics, transaction history, usage patterns, and churn status.

**2. Data Quality:** Ensure the dataset is clean, accurate, and up-to-date. Handle missing values, outliers, and inconsistencies to maintain data integrity.

**3. Feature Selection:** Identify key features that are likely to influence churn behaviour, such as account activity, customer demographics, product usage, and interactions with the bank.

**4. Data Preparation:** Preprocess the data by encoding categorical variables, scaling numerical features, and splitting the dataset into training and testing sets.

**5. Model Selection:** Choose appropriate machine learning algorithms for churn prediction, considering factors such as interpretability, scalability, and performance metrics.

**6. Model Training:** Train the selected models using the training data, optimizing hyperparameters to improve predictive accuracy and generalization.

**7. Model Evaluation:** Evaluate the trained models using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to assess their effectiveness in predicting churn.

**8. Interpretability:** Ensure the models provide interpretable insights into the factors driving churn, enabling stakeholders to understand the underlying reasons and take appropriate actions.

**9. Scalability:** Design the churn prediction system to handle large volumes of data efficiently, considering the scalability requirements of both training and inference stages.

1. **Integration:** Integrate the churn prediction model into the bank's existing systems and workflows, enabling seamless deployment and integration with customer relationship management (CRM) tools.

**DATASETS USED**

* **Row Number:** corresponds to the record (row) number and has no effect on the output.
* **Customer Id:** contains random values and has no effect on customer leaving the bank.
* **Surname:** the surname of a customer has no impact on their decision to leave the bank.
* **Credit Score: It** can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank.
* **Geography:** a customer’s location can affect their decision to leave the bank.
* **Gender:** It is interesting to explore whether gender plays a role in a customer leaving the bank.
* **Age:** this is certainly relevant, since older customers are less likely to leave their bank than younger ones.
* **Tenure:** refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.
* **Balance:** A very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.
* **Num Of Products:** refers to the number of products that a customer has purchased through the bank.
* **Has Cr Card:** denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank.
* **Is Active Member:** Active customers are less likely to leave the bank.
* **Estimated Salary:** As with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.
* **Exited:** Whether or not the customer left the bank. (0=No,1=Yes)

**PROPOSED DESIGN**

**1.** **Features:**

- Credit Score: The credit score of the customer.

- Geography: The geographical location of the customer (e.g., country).

- Gender: The gender of the customer.

- Age: The age of the customer.

- Tenure: The number of years the customer has been with the bank.

- Balance: The account balance of the customer.

- Num Of Products: The number of bank products the customer has.

- Has Cr Card: Whether the customer has a credit card (True/False).

- Is Active Member: Whether the customer is an active member (True/False).

- Estimated Salary: The estimated salary of the customer.

**2.** **Target Variable:**

- Exited: Whether the customer has churned (True/False).

**3. Data Quality:**

- The dataset seems to be structured properly, with no missing values indicated.

- Further exploration is needed to ensure the validity and reliability of the data.

**4.** **Potential Preprocessing Steps:**

- Encoding categorical variables like Geography and Gender into numerical values.

- Scaling numerical features to ensure they are on a similar scale.

- Handling any outliers or anomalies in the data.

**5. Usage:**

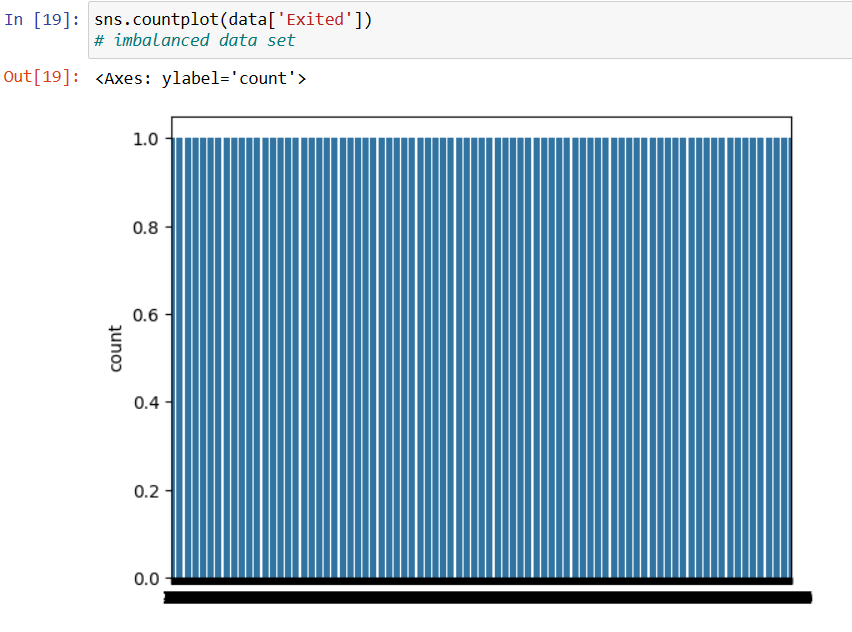
- This dataset can be used to build predictive models to identify factors influencing customer churn and develop strategies to reduce churn rate.

**6. Model Evaluation:**

- After preprocessing, the dataset can be split into training and testing sets for model training and evaluation using metrics such as accuracy, precision, recall, and F1-score.

**THE GRAPHS AND THE INFERENCE:**

1. **COUNTPLOT (An Imbalanced Dataset)**



* **INFERENCE:**

• The image shows a count plot displaying the frequency or count of the 'Exited' column in a dataset.

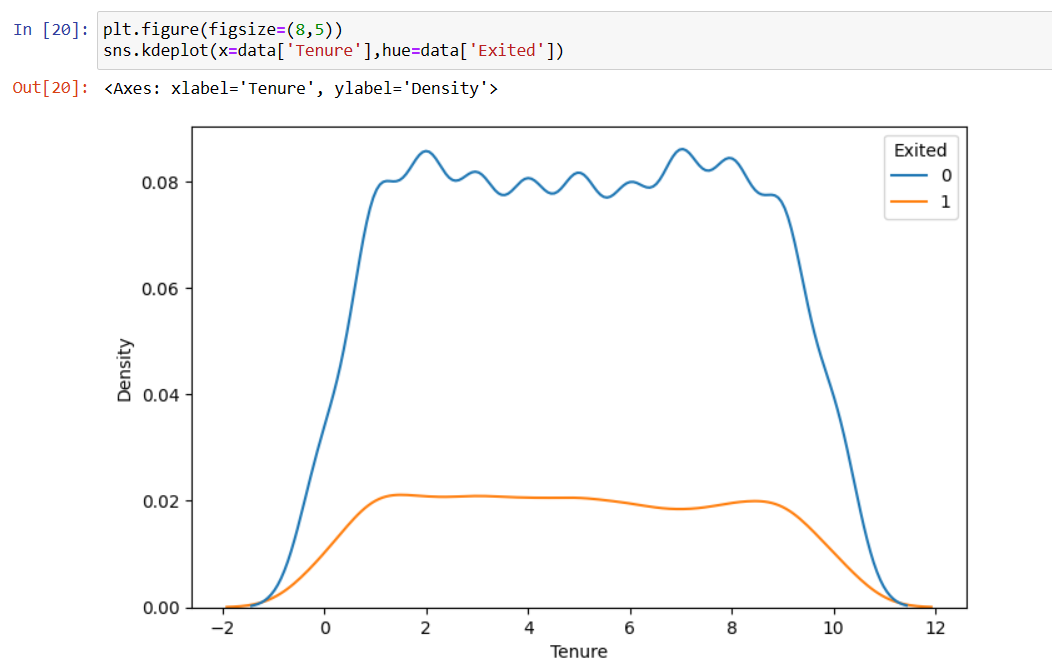
• The y-axis represents the count.

• The x-axis is not visible in the image.

• The tall blue bars indicate that most data points have a value of 1 for the 'Exited' column, suggesting instances where customers have exited or churned from a service.

• The small bar at the bottom represents a smaller count for the value 0 in the 'Exited' column, likely indicating instances where customers have not exited or churned.

1. **TENURE VS DENSITY**

****

* **INFERENCE**

The graph displays the distribution of a numerical feature called "Tenure" for two different classes, represented by the binary "Exited" column (0 ) and “Non-Exited” column(1)

• The x-axis represents the "Tenure" ( -2 to 12)

• The y-axis represents the "Density" for each class.

• The blue curve shows the density distribution of "Tenure" for instances where "Exited" is 0 (customers who have not churned).

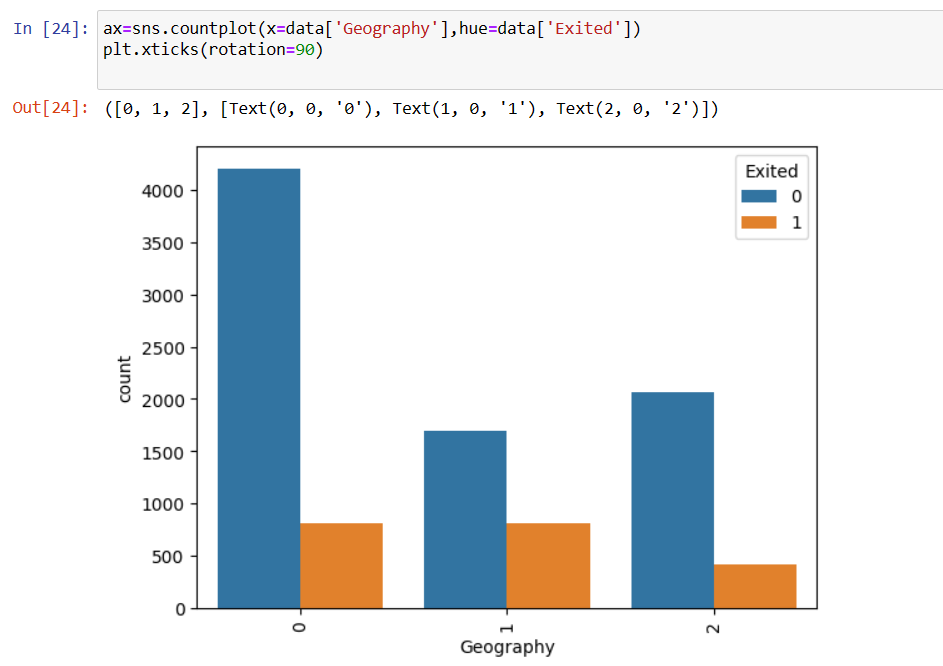
• The orange curve shows the density distribution of "Tenure" for instances where "Exited" is 1 (customers who have churned).

• The blue curve has a higher peak and a slightly narrower distribution than the orange curve, indicating that non-churned customers tend to have lower "Tenure" values compared to churned customers.

• The orange curve is shifted slightly to the right, suggesting that churned customers tend to have higher "Tenure" values on average compared to non-churned customers.

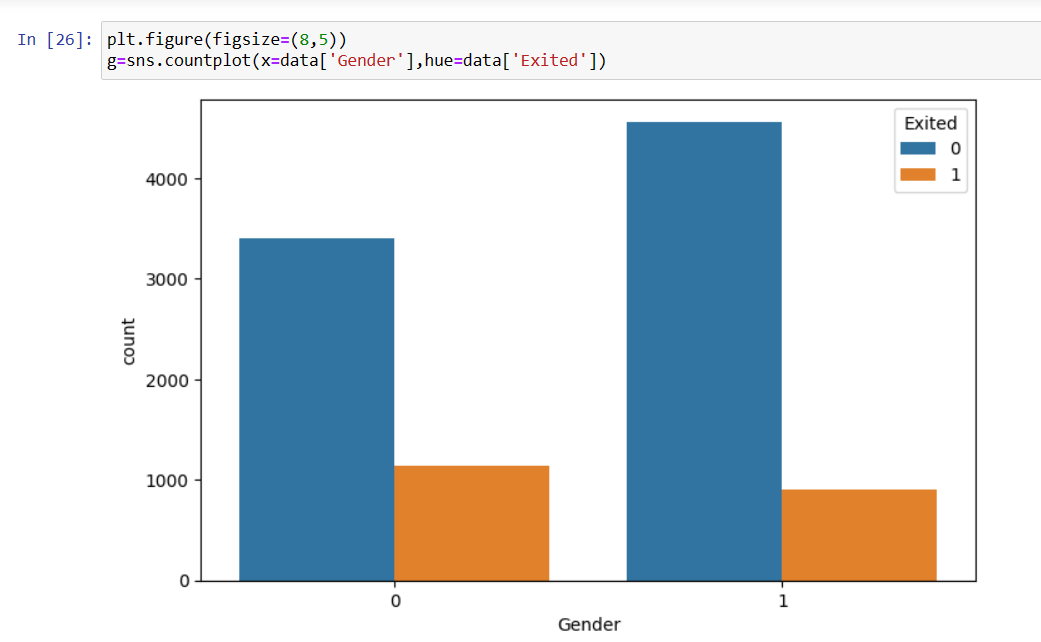
• The overlapping regions of the two curves indicate that there is some overlap in the "Tenure" values between the two classes, and "Tenure" alone may not be a perfect predictor of customer churn.

1. **GEOGRAPHY BASED EXITED (COUNTPLOT)**



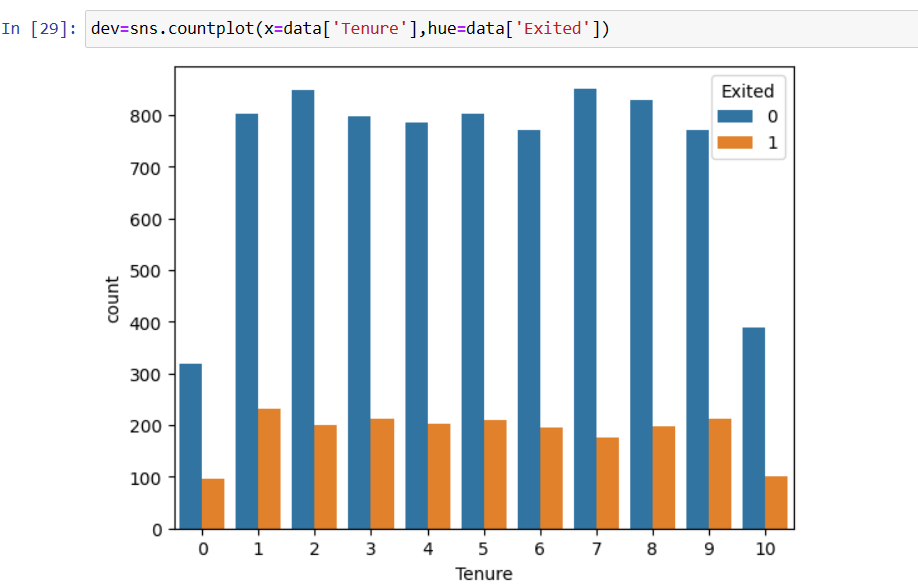
* **INFERENCE**
* For x-axis value 0:
* Category 0 (blue bar) count: Approximately 4000
* Category 1 (orange bar) count: Approximately 500
* For x-axis value 'r' (likely representing a geography label):
* Category 0 count: Approximately 1500
* Category 1 count: Approximately 400
* For x-axis value 2:
* Category 0 count: Approximately 1000
* Category 1 count: A smaller value, difficult to estimate precisely from the graph
* So in summary, the main data points shown are the counts for the two categories (0 and 1) at the three x-axis values of 0, 'r', and 2, with category 0 having higher counts than category 1 across all three values.

1. **GENDER BASED EXIT**



* **INFERENCE**
* For x-axis value 0 (likely representing a gender category):
* Category 0 (blue bar) count: Approximately 3000
* Category 1 (orange bar) count: Approximately 1000
* For x-axis value 1 (likely representing another gender category):
* Category 0 count: Approximately 4000
* Category 1 count: Approximately 1000
* The x-axis is labelled 'Gender', suggesting the two values 0 and 1 correspond to different gender categories in the data. Category 0 (blue bars) has higher counts than Category 1 (orange bars) for both gender values.
* Without additional context, it's difficult to infer precisely what the gender categories represent, but the graph clearly shows a difference in the counts between the two categories across the two gender values plotted on the x-axis.

1. **TENURE BASED EXIT**



* **INFERENCE**

• The graph is a count plot showing the frequency of customer tenure values for two classes: customers who have exited (churned) and those who have not exited.

• The x-axis represents the "Tenure" values.

• The y-axis represents the count or frequency.

• The blue bars represent the count for customers who have not exited (Exited = 0).

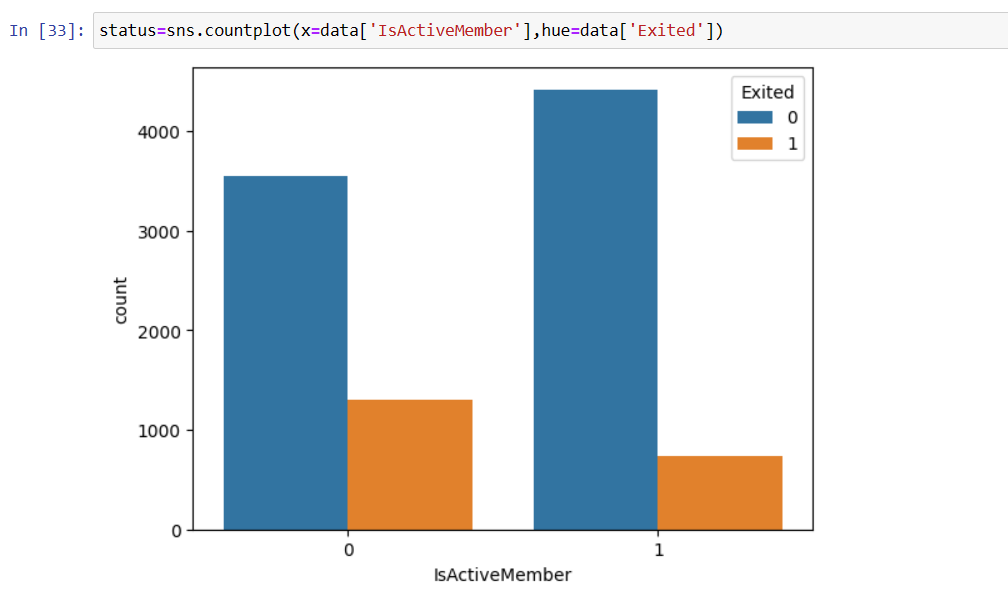
• The orange bars represent the count for customers who have exited (Exited = 1).

• The majority of customers, regardless of churn status, have tenure values between 0 and 5 years.

• There is a significant drop in counts for higher tenure values (above 5 years) for both classes.

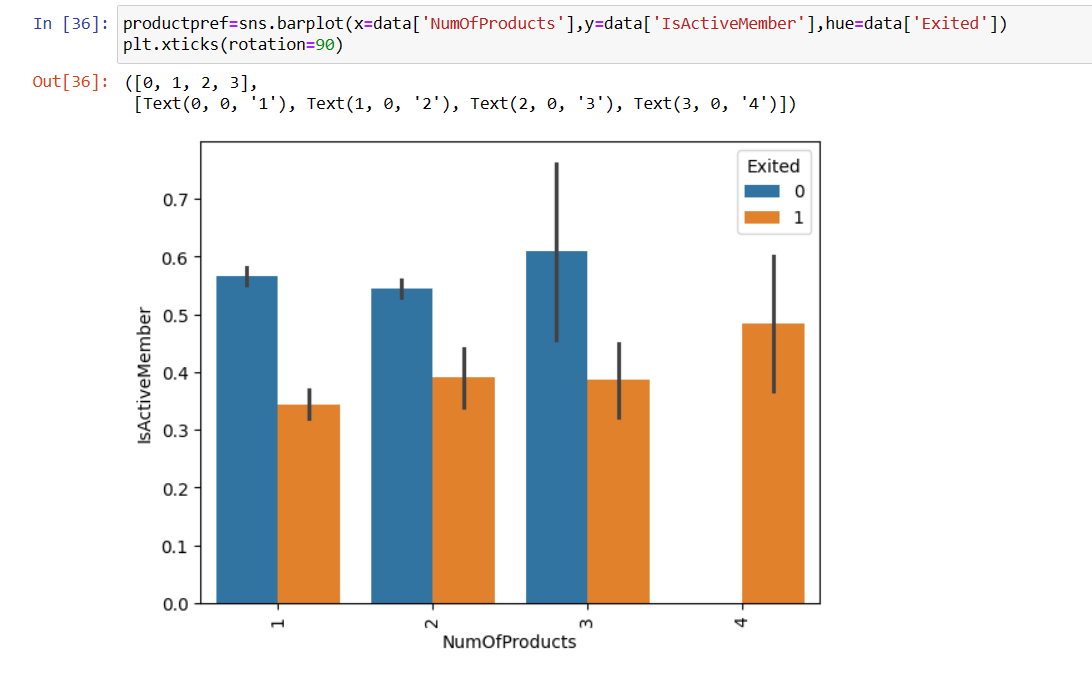
• There are slightly more customers who have exited (orange bars) in the higher tenure ranges (e.g., 8-10 years) compared to those who have not exited (blue bars).

1. **ACTIVE MEMBER BASED EXIT**

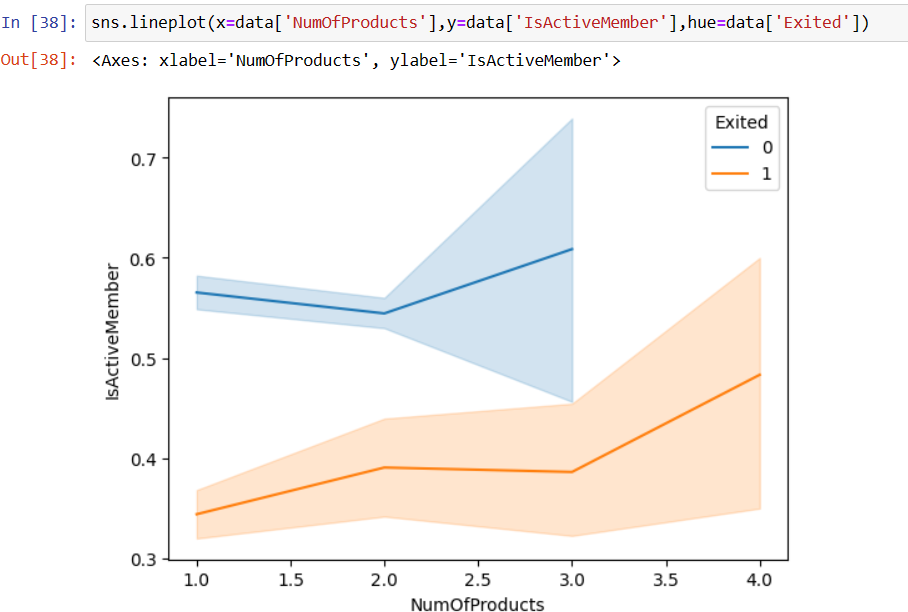


* **INFERENCE**
  + There are two values for "Is Active Member" - 0 and 1.
  + The blue bars represent count of observations where "Is Active Member" is 0.
  + The orange bars represent count of observations where "Is Active Member" is 1.
  + For observations with "Is Active Member" = 0, the majority did not exit (Exited = 0).
  + For observations with "Is Active Member" = 1, the majority exited (Exited = 1).

1. **IS ACTIVE MEMBER VS NUM OF PRODUCTS**



* **INFERENCE**
* There are four distinct values for "Num Of Products" ranging from 1 to 4.
* For each value of "Num Of Products", there are two bars - blue representing "Is Active Member" = 1 (active members) and orange representing "Exited" = 1 (customers who exited).
* The height of the blue bars generally increases as the "Num Of Products" increases, indicating that active members tend to have more products.
* The height of the orange bars also increases with higher "Num Of Products", suggesting that customers with more products are more likely to exit.
* For "Num Of Products" = 4, the orange bar (exited customers) is taller than the blue bar (active members), implying that a significant portion of customers with 4 products have exited.

**LINE PLOT : IS ACTIVE MEMBER VS NUM OF PRODUCTS**

* **INFERENCE:**

1. The blue line represents active members (Is Active Member = 1), while the orange line represents customers who have exited (Exited = 1).

2. For lower values of "Num Of Products" (around 1-2), the proportion of active members is higher than customers who have exited.

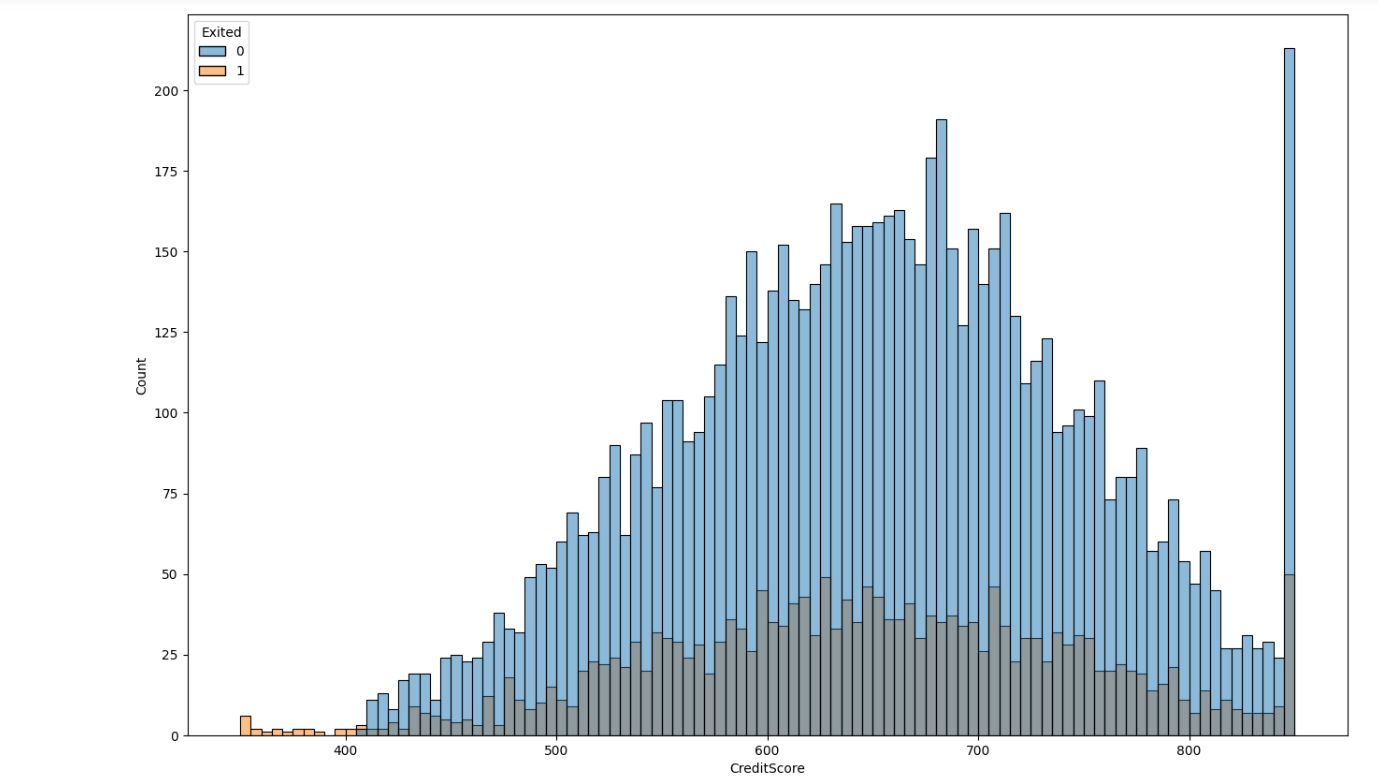
3. As the "Num Of Products" increases, the proportion of active members continues to rise, while the proportion of customers who have exited also increases, but at a slower rate.

4. Around "Num Of Products" = 3, the lines for active members and exited customers intersect, indicating that at this point, the proportions are roughly equal.

5. Beyond "Num Of Products" = 3, the proportion of exited customers becomes higher than the proportion of active members, suggesting that customers with a higher number of products are more likely to churn or exit.

6. The gap between the two lines widens as "Num Of Products" increases further, implying that customer retention becomes more challenging for accounts with a larger number of products.

1. **CREDIT SCORE BASED EXIT**

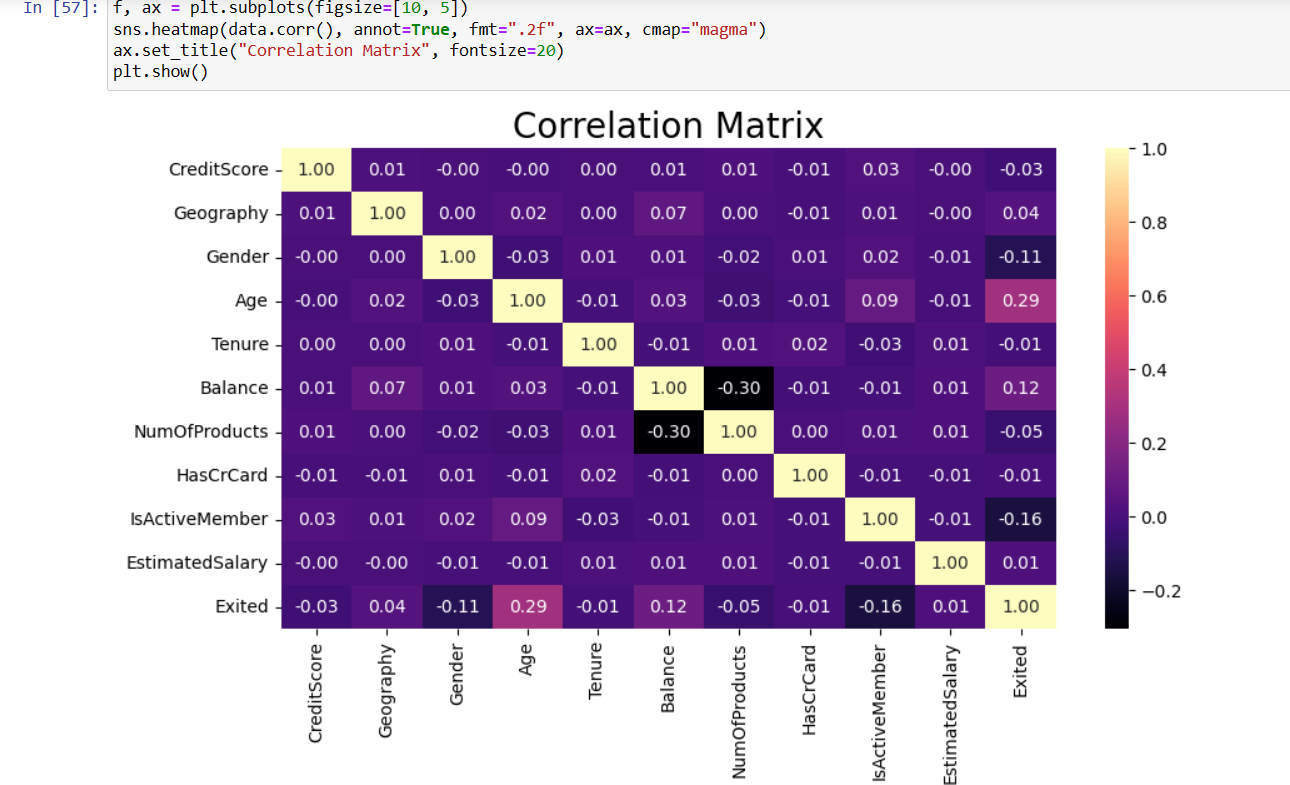


* **INFERENCE**

The graph displays the distribution of a variable called "Credit Score" for customers who have exited (shown in orange bars) and those who have not exited (shown in blue bars).

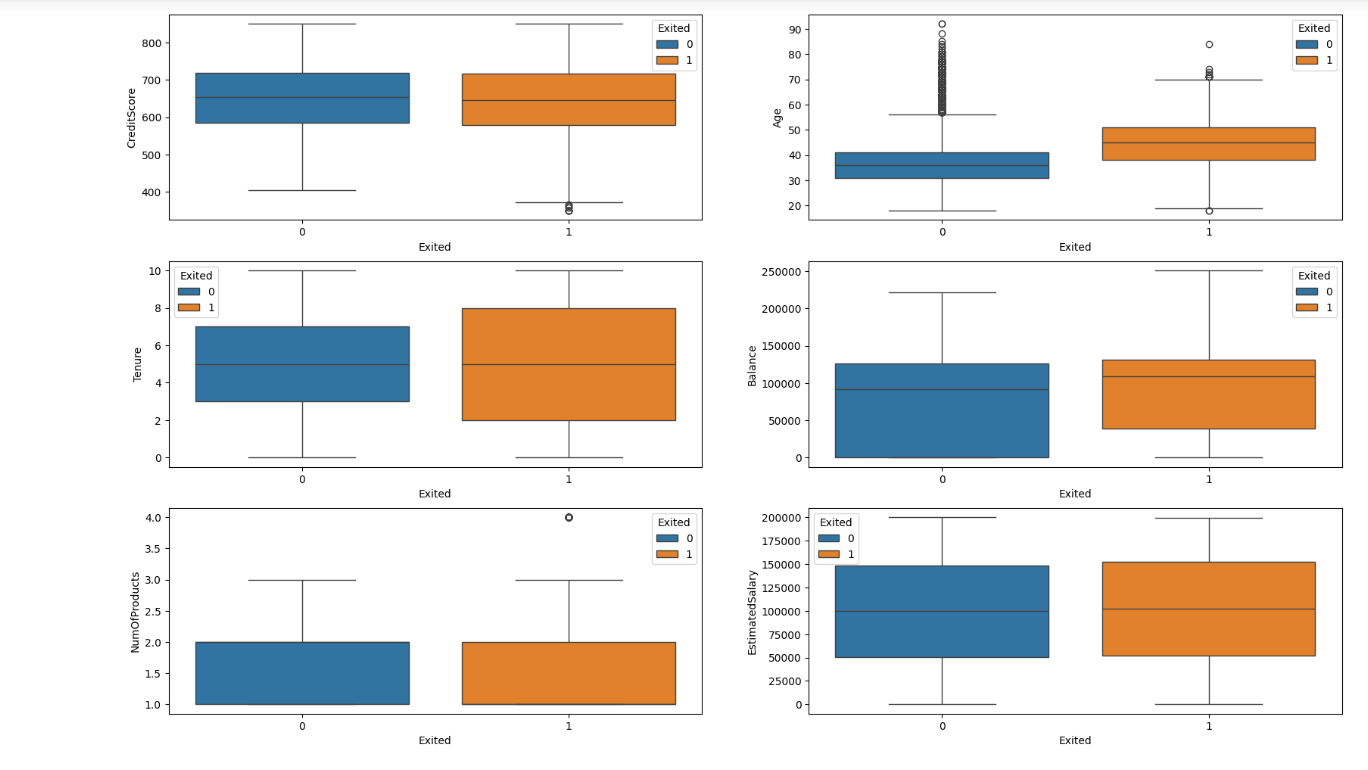
* The Credit Score variable ranges from around 350 to 850.
* The distribution for non-exited customers (blue bars) shows a roughly normal or bell-shaped curve, with the highest frequency around a Credit Score of 600-650.
* For exited customers (orange bars), the distribution is skewed towards lower credit scores. There is a higher concentration of exited customers with credit scores below 550 compared to non-exited customers.
* There is a noticeable spike in the number of exited customers with very low credit scores around 350-400.
* For credit scores above 650-700, the number of exited customers decreases substantially, indicating that customers with higher credit scores are less likely to exit or churn.
* The overall shape of the distribution suggests that customers with lower credit scores are at a higher risk of exiting or churning, while those with higher credit scores tend to remain as customers

1. **CORRELATION MATRIX:**



* **INFERENCE:**
* **Credit Score** and **Age** have a perfect negative correlation of -1.0 and -1.0 respectively with themselves, as expected.
* **Age** has a moderate positive correlation of 0.29 with Exited.
* **Balance** has a weak positive correlation of 0.12 with Exited.
* **Num Of Products** has a very weak negative correlation of -0.05 with Exited.
* **Has Cr Card** has an extremely weak negative correlation of -0.01 with Exited.
* **Is Active Member** has a moderate negative correlation of -0.16 with Exited.
* **Estimated Salary** has a very weak positive correlation of 0.01 with Exited.
* **Notable correlations** between independent variables: Age and Tenure: -0.01 (very weak negative)
  + - **Age and Balance**: 0.03 (very weak positive)
    - **Balance and Num Of Products**: -0.30 (moderate negative)

1. **BOX-PLOT BASED ON DIFFERENT PARAMETERS: [OUTLIERS]**



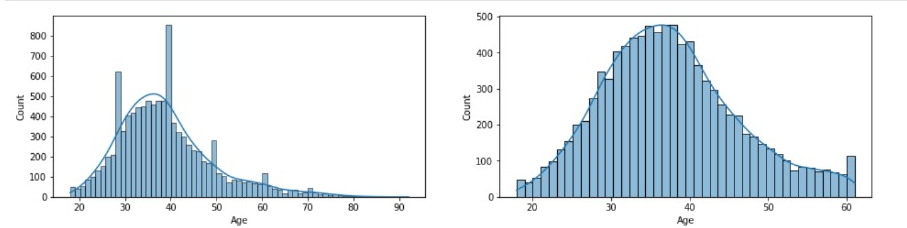
* **INFERENCE:**

*Exited (orange bars)*

*Not-Exited (blue bars)*

* Credit Score: Customers who have exited tend to have lower credit scores compared to those who have not exited.
* Tenure: Customers with lower tenure (possibly newer customers) are more likely to exit compared to those with higher tenure.
* Number of Products: Customers who have exited tend to have fewer products on average compared to those who have not exited.
* Age: The distribution of age does not show a significant difference between exited and non-exited customers.
* Estimated Salary: Customers who have exited generally have lower estimated salaries compared to those who have not exited.
* Balance: The plot suggests that customers with higher account balances are more likely to exit, which could be counter-intuitive.

1. **AGE VS COUNT**



* **INFERENCE:**

Based on the two histograms shown in the image, a few inferences can be made:

1. The graphs depict age distributions, with the x-axis representing age and the y-axis representing the count or frequency.

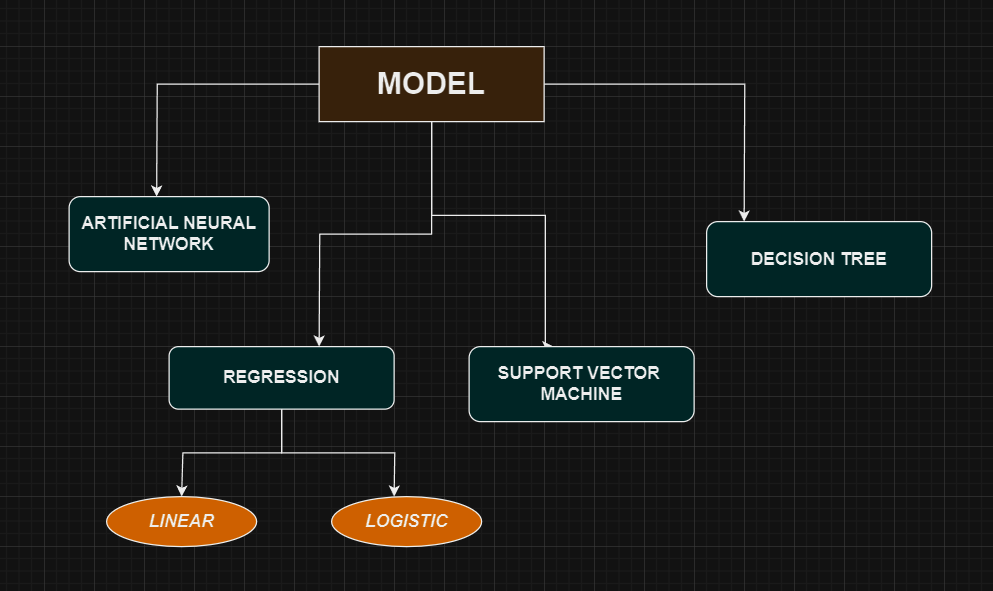
2. The graph on the left appears to show the age distribution of a population or sample. It has a bell-shaped curve, which is typical of many age distributions, with the highest frequency around the middle ages (approximately 30-50 years old) and lower frequencies at the younger and older age extremes.

3. The graph on the right also shows an age distribution, but it has a more peaked or triangular shape, with the highest frequency around 30-35 years old and a sharper decline on either side of that peak age range.

4. Both graphs suggest the populations or samples being represented skew slightly towards younger ages, as the curves are not perfectly symmetrical and have a slightly longer tail towards the higher age values.

5. The range of ages covered appears to be from around 20 years old to around 90 years old in the left graph, and a narrower range of approximately 20 to 60 years old in the right graph.

**MODEL SELECTION**



**EVALUATION METRICS:**

* **Precision:** The fraction of instances predicted as positive that are positive. It measures the exactness of the positive predictions made.
* **Recall:** The fraction of actual positive instances that were correctly predicted as positive. It measures how well the model identifies all positive instances.
* **F1-score:** The harmonic mean of precision and recall, providing a balanced measure that combines both metrics into one score.
* **Support:** The number of instances of each class in the dataset.
* **Accuracy:** The fraction of instances that were correctly classified by the model.
* **Macro avg**: The unweighted mean of the metric values across all classes.
* **Weighted avg:** The mean of the metric values weighted by the frequency/support of each class. It accounts for class imbalance.

1. **REGRESSION:**

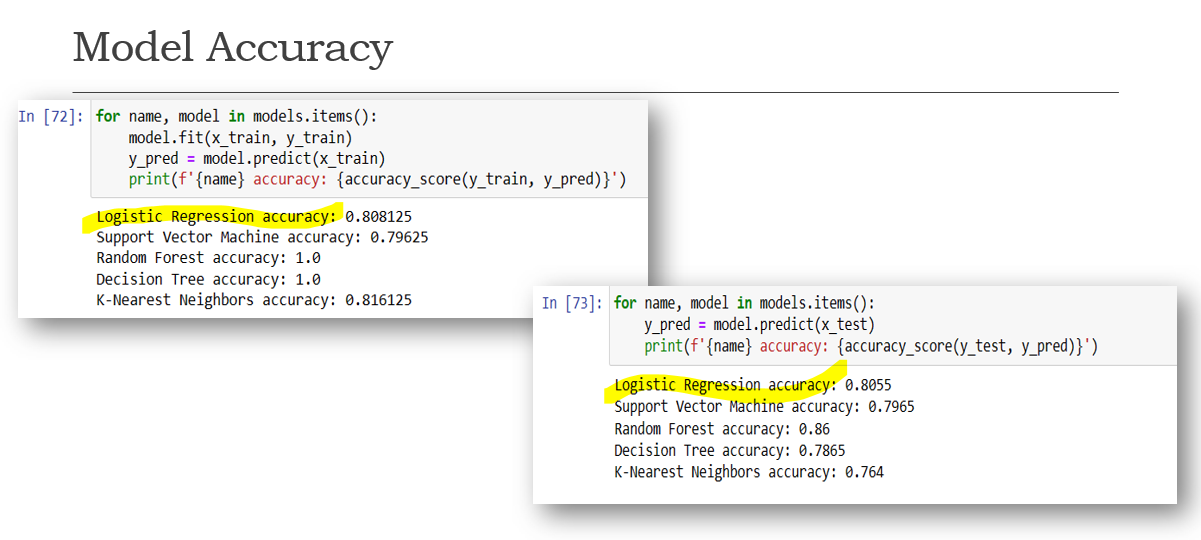
* Regression techniques can be employed to predict customer churn. Customer churn refers to the phenomenon where customers discontinue their relationship with a company or service. Predicting churn is crucial for businesses as it allows them to take proactive measures to retain customers.
* **LINEAR REGRESSION:**
* Using linear regression for customer churn prediction may not be the best choice for several reasons:
* **Binary Outcome**: Customer churn prediction is typically a binary classification problem where the outcome is either churned or not churned. Linear regression, however, is designed for continuous outcomes and may not be suitable for predicting binary outcomes directly.
* **Outliers and Skewed Data**: Linear regression is sensitive to outliers and skewed data. In customer churn prediction, you might have imbalanced classes or outliers in the dataset, which can negatively impact the performance of linear regression.
* **LOGISTIC REGRESSION**
* Logistic regression is a commonly used technique for customer churn prediction, and it's well-suited for binary classification.
* **1. Target Variable**:
* - Identify the target variable, which is whether a customer churns or not (usually represented as a binary outcome: 1 for churn, 0 for non-churn).
* **2. Data Preparation:**
* - Split the data into training and testing sets to evaluate model performance.
* - Handle missing values and outliers appropriately, either through imputation or removal.

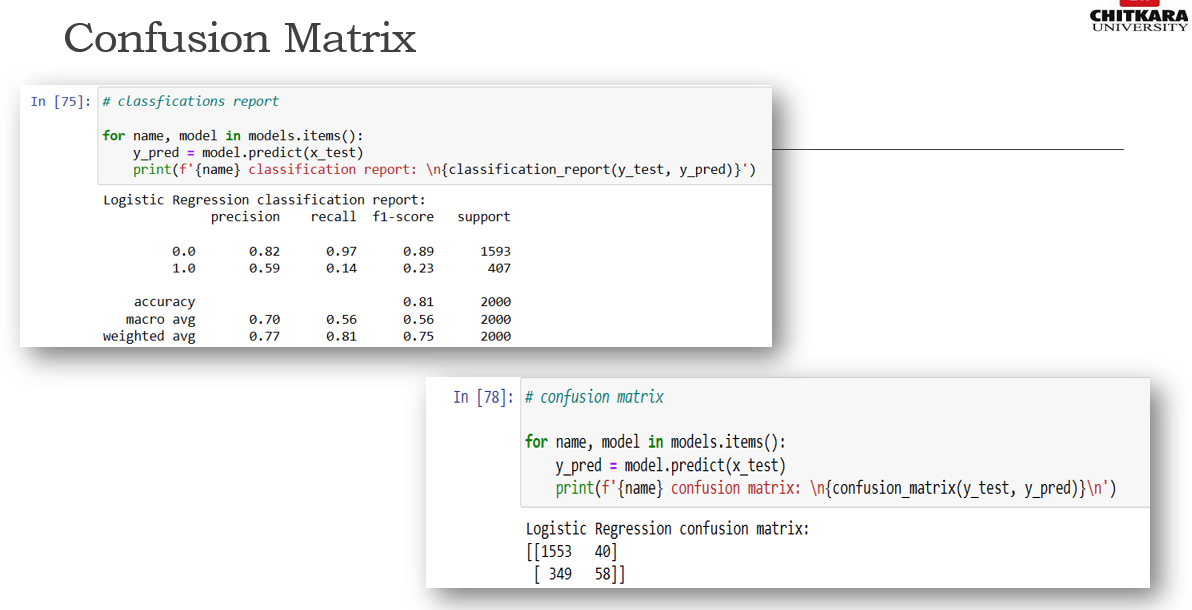
**3. Model Building:**

* - Fit a linear regression model to the training data, treating churn as the dependent variable and the selected features as independent variables.
* - The linear regression model estimates the relationship between the predictors and the probability of churn.

**4. Model Evaluation:**

* - Consider using techniques like cross-validation to assess the model's robustness and generalization ability.





1. **SUPPORT VECTOR MACHINE:**

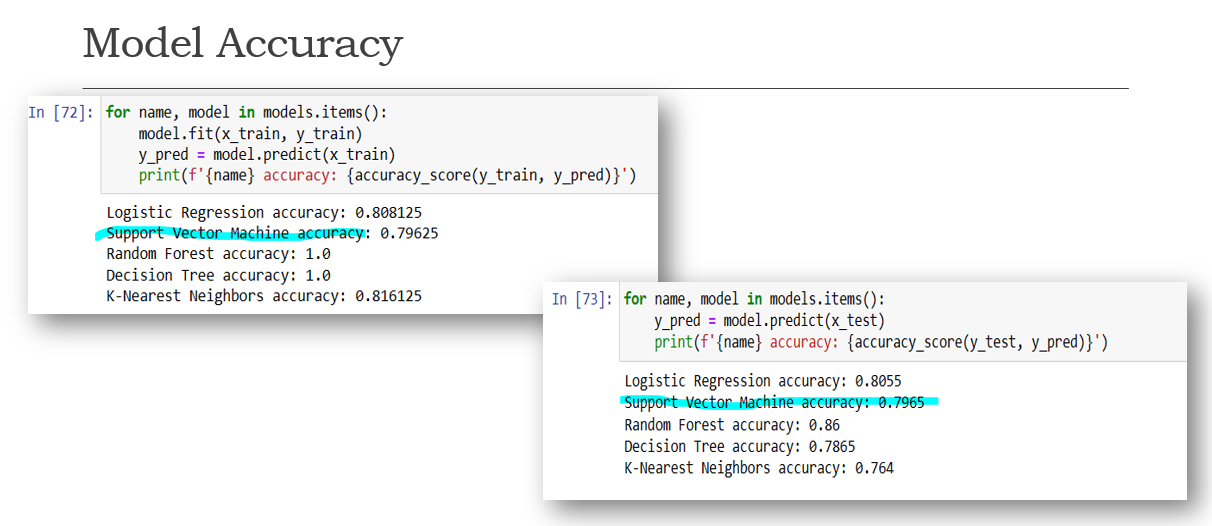
* It is based on the concept of maximizing the margin, which represents the distance between the hyperplane and the nearest data points (support vectors) of each class.

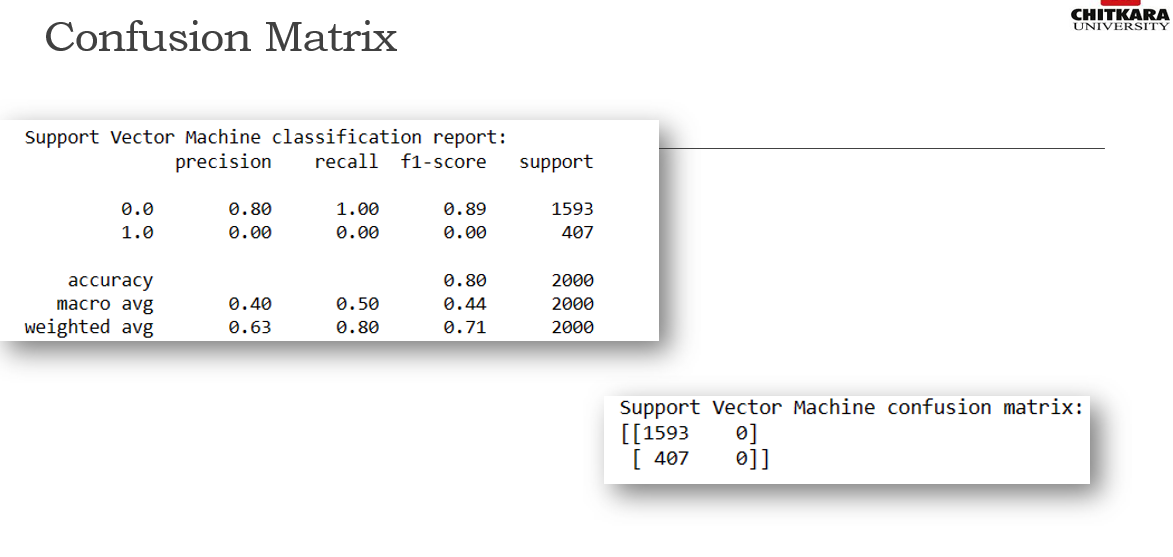
1. **Binary Classification**:

SVM is applied to classify customers into churners and non-churners, treating churn as the positive class (1) and non-churn as the negative class (0).

1. **Data Preprocessing**:
   * Scale features to ensure they have similar ranges, as SVM is sensitive to feature scales.
   * Handle missing values and encode categorical variables if necessary.
2. **Handling Imbalanced Data**:

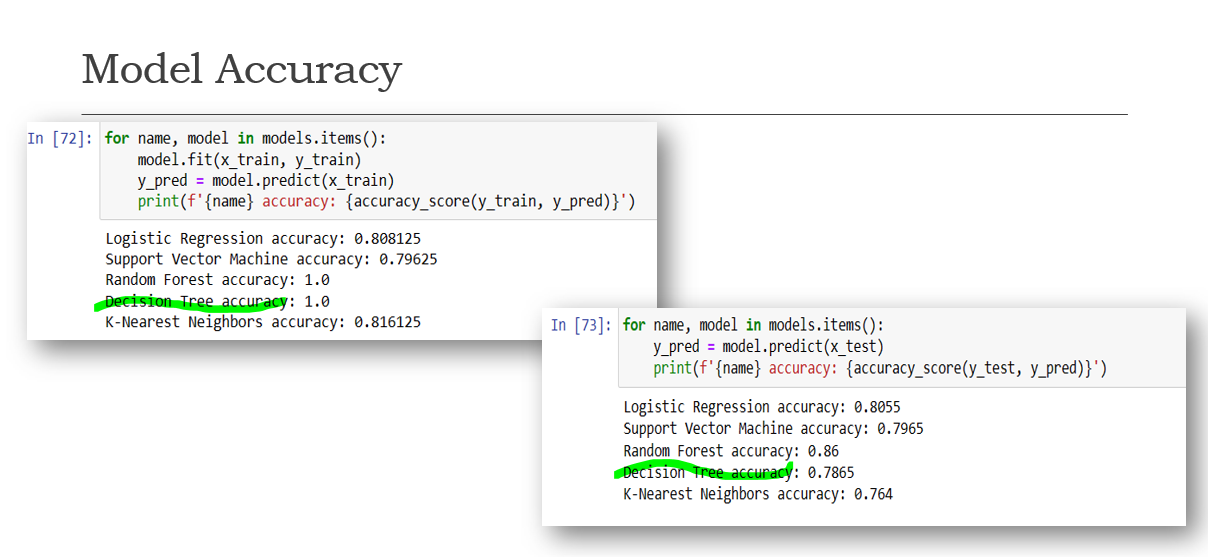
Address class imbalance by using techniques like oversampling (e.g., SMOTE), under sampling, or adjusting class weights to prevent bias towards the majority class.

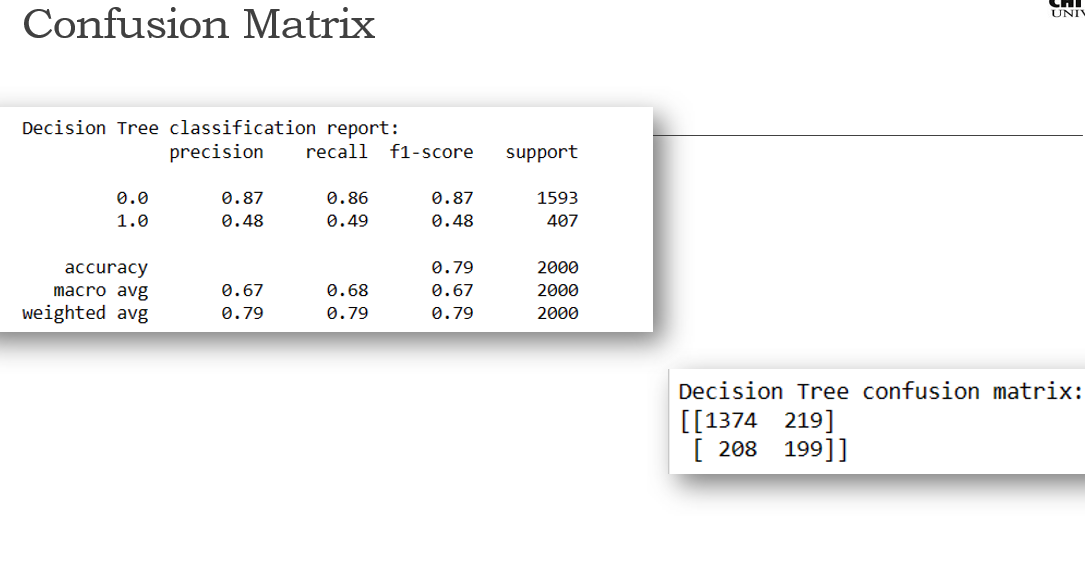




1. **DECISION TREE:**

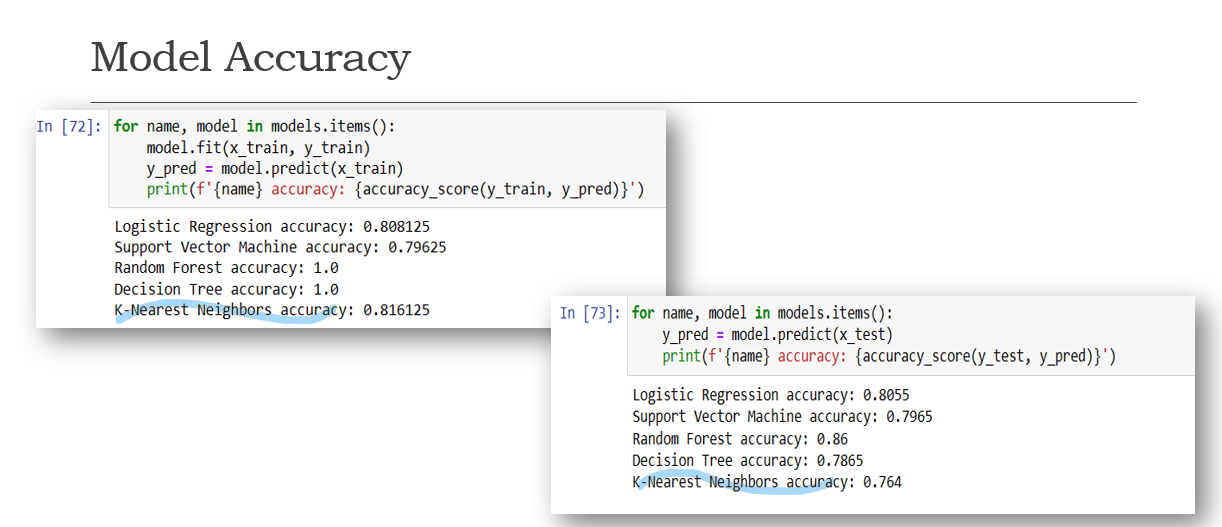
* It recursively splits the dataset into subsets based on the most significant attribute at each node, forming a tree-like structure of decision rules.
* **Feature Importance**:
  + Decision trees provide insights into the most important features for predicting churn by examining feature importance scores, which indicate the predictive power of each attribute.
* **Segmentation**:
  + Decision trees segment customers into homogeneous groups based on their characteristics and behavior, identifying subsets of customers with higher churn probabilities.
* **Handling Non-linearity**:
  + Decision trees can capture non-linear relationships between features and churn, allowing for flexible modeling of complex patterns in customer behavior.
* **Customer Path Analysis**:
  + Decision trees can be used to trace individual customer paths through the tree, revealing the sequence of factors that contribute to churn for each customer.

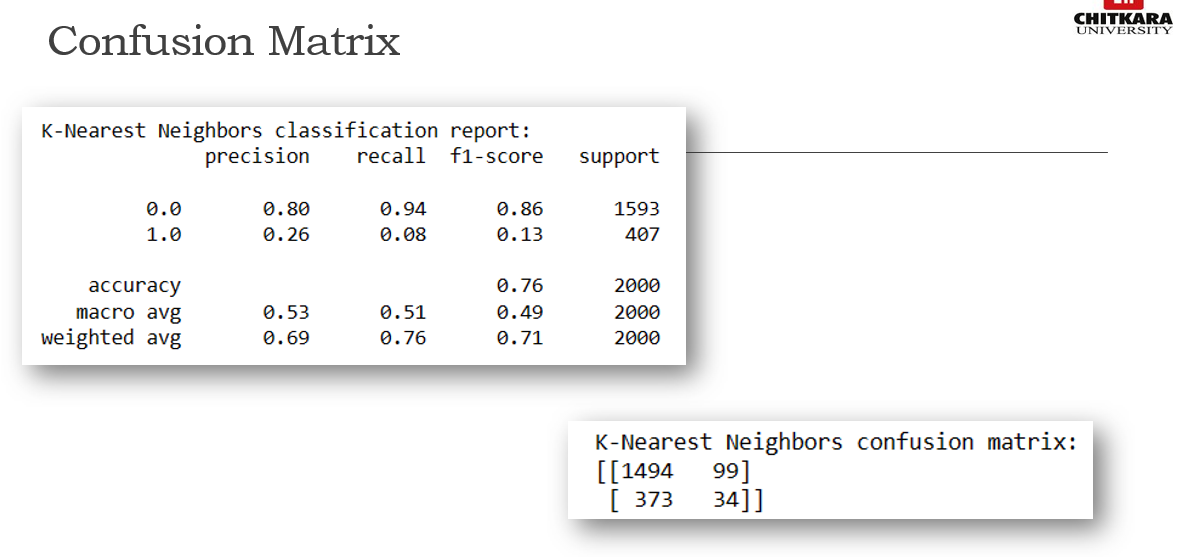




1. **K-NEAREST NEIGHBOUR:**

* KNN is a non-parametric algorithm, meaning it doesn't make explicit assumptions about the underlying data distribution.
* It relies on a distance metric (e.g., Euclidean distance) to measure the similarity between data points and identify nearest neighbors.
* **Proximity-based Classification**:
  + KNN predicts churn for a customer by considering the churn status of its K nearest neighbors, assuming that similar customers have similar churn behaviors.
* **Simple Implementation**:
  + KNN is easy to understand and implement, making it suitable for quick prototyping and experimentation in churn prediction tasks.
* **Impact of K Parameter**:
  + The choice of the K parameter influences the model's performance and generalization ability, with smaller values potentially leading to overfitting and larger values resulting in underfitting.
* **Distance Metric Selection**:
  + The choice of distance metric (e.g., Euclidean, Manhattan, etc.) affects how KNN measures similarity between data points and can impact model performance.

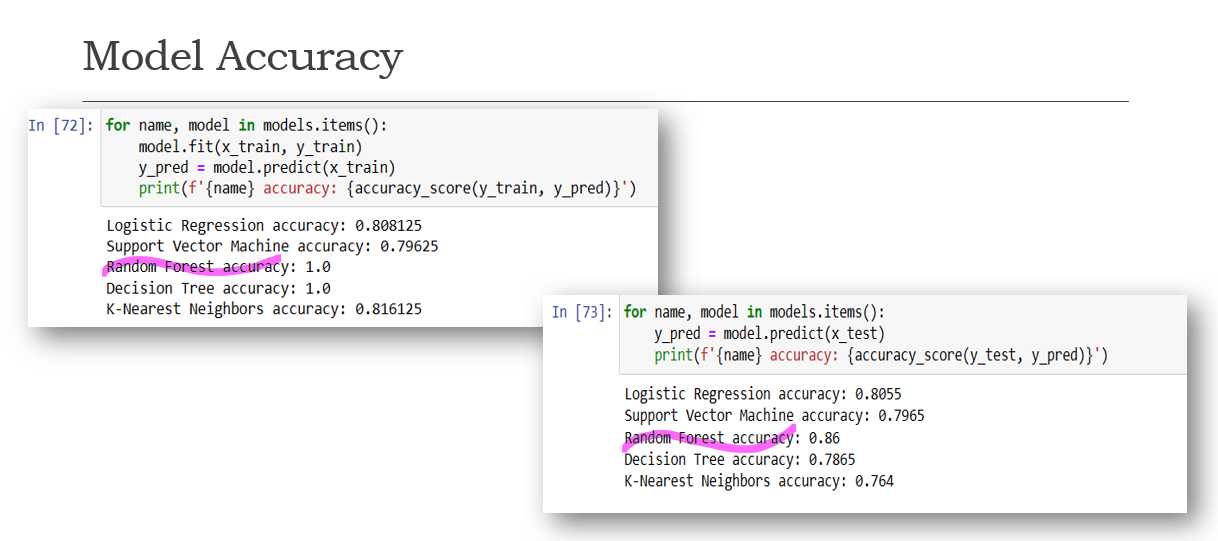


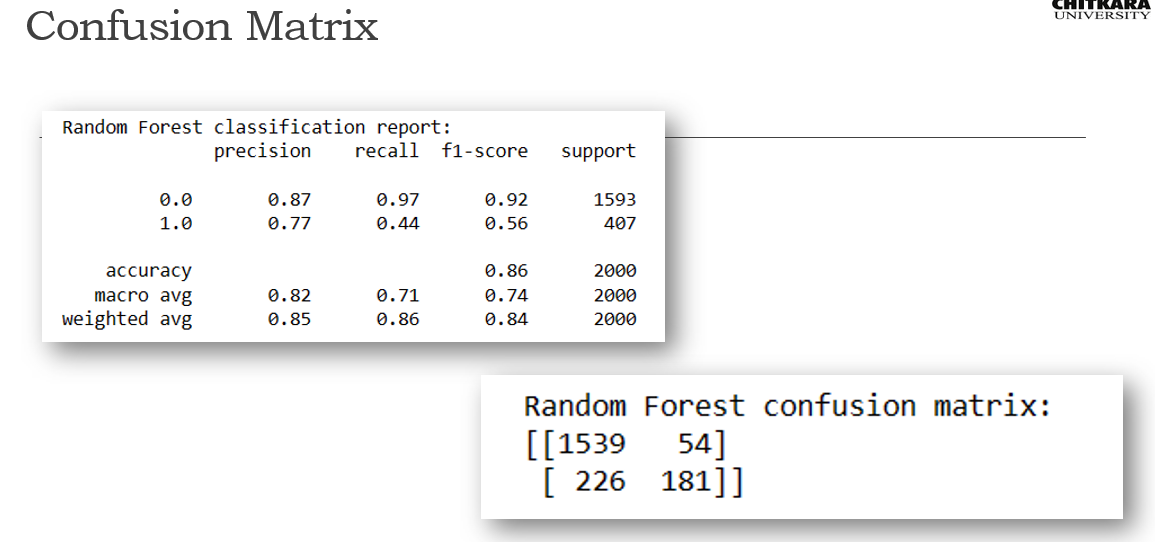
****

1. **RANDOM FOREST:**

* Random Forest leverages the concept of bagging (Bootstrap Aggregating) to create diverse subsets of the training data for each tree.
* Each decision tree in the Random Forest is trained independently, making the model robust to noise and variance.
* **Improved Predictive Accuracy**:
  + Random Forest combines the predictions of multiple decision trees, resulting in higher predictive accuracy compared to individual decision trees.
* **Handling Non-linearity and Interactions**:
  + Random Forest can capture complex non-linear relationships and interactions between features, enabling it to model intricate patterns in customer behavior associated with churn.
* **Ensemble Learning**:
  + Random Forest is a type of ensemble learning method that combines the strengths of multiple decision trees, offering improved performance and stability over individual models.
* **Scalability and Parallelization**:
  + Random Forest can be parallelized, allowing for efficient computation and scalability to large datasets, making it suitable for churn prediction in big data environments.

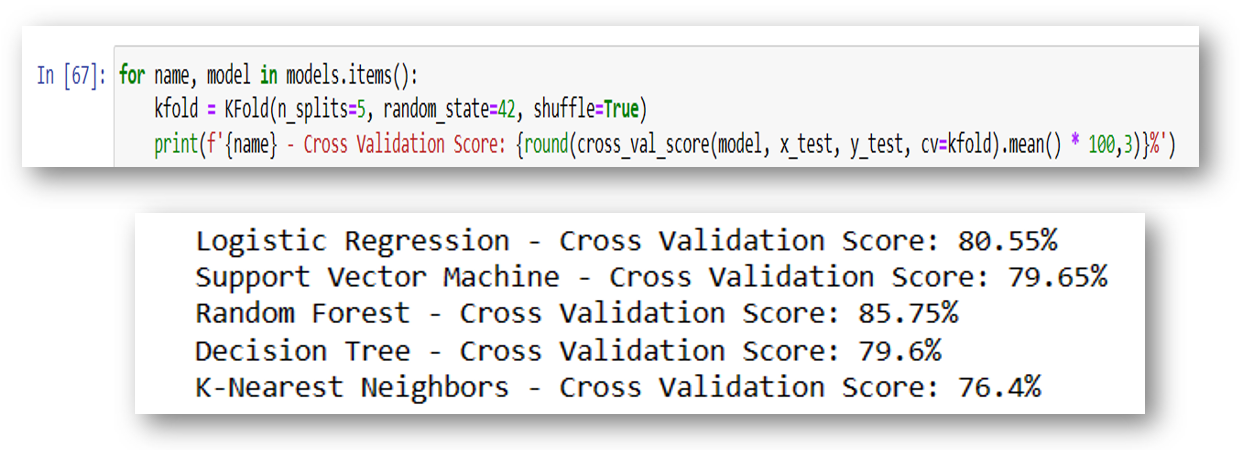
/





**K-FOLD CROSS VALIDATION:**

* k-fold cross-validation is a resampling technique used to assess the performance and generalization ability of a machine learning model. The dataset is divided into k subsets (folds) of approximately equal size.
* k-fold cross-validation provides a more reliable estimate of the model's performance by reducing the variance associated with a single train-test split.
* It helps identify the model that achieves the best average performance across different subsets of the data, leading to more informed model selection decisions.
* k-fold cross-validation helps strike a balance between bias and variance in model evaluation.
* Performance metrics such as accuracy, precision, recall, F1-score obtained from each fold are averaged to obtain a single aggregate metric representing the model's overall performance.



**CONCLUSION:**

* **BEST PERFORMING MODEL:**
* - **Random Forest - Cross Validation Score: 85.75%**
* The Random Forest algorithm achieved the highest cross-validation score of 85.75%.
* It indicates that it performed the best among the evaluated models in terms of accurately predicting the target variable on unseen data.
* Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting.
* **WORST PERFORMING MODEL:**
* - **K-Nearest Neighbors - Cross Validation Score: 76.4%**
* The K-Nearest Neighbors (KNN) algorithm had the lowest cross-validation score of 76.4%, making it the worst-performing model in this comparison.
* KNN is a non-parametric method that classifies instances based on their similarity to the nearest neighbors in the training data.
* Its lower performance could be due to factors such as the choice of distance metric, the value of k (number of neighbors), or the presence of noisy or irrelevant features in the data.
* **OTHER MODELS:**
* **- Logistic Regression - Cross Validation Score: 80.55%**
* Logistic Regression, a popular statistical model for binary classification, achieved a cross-validation score of 80.55%, performing better than SVM and Decision Tree but worse than Random Forest.
* **- Support Vector Machine (SVM) - Cross Validation Score: 79.65%**
* The SVM model, which constructs hyperplanes in high-dimensional spaces for classification or regression, had a cross-validation score of 79.65%, slightly outperforming the Decision Tree but underperforming compared to Logistic Regression and Random Forest.
* - **Decision Tree - Cross Validation Score: 79.6%**
* The Decision Tree algorithm, which recursively partitions the data based on feature values, achieved a cross-validation score of 79.6%, similar to SVM but lower than the other models.