**Exploratory Data Analysis Report**

**1. Introduction: Nature of the Problem**

Customer churn is framed as a binary classification problem, with the objective of predicting whether a customer will churn (leave) or stay. The target variable, Exited, is binary:

* 1 (Yes): The customer churned (left the bank).
* 0 (No): The customer did not churn (remained a customer).

**2. Preprocessing**

* The dataset contains no missing values or duplicates. Data types for all variables are correctly set.
* The following columns were removed as they do not contribute to the analysis or modeling process:
  + RowNumber: Simply a row identifier.
  + CustomerId: A unique customer identifier, not useful for prediction.
  + Surname: Unrelated to customer behavior.

**3. Treatment of Variables**

Here is an overview of the variables and their types.

A screenshot of a computer screen

Description automatically generated

**3.1 Analysis of Categorical Variables**

**Geography, Gender, and Card Type**

* **Geography:**
  + The dataset is dominated by French customers, accounting for over half of the entries. Customers from Spain and Germany are less represented. This imbalance could bias the model toward patterns observed in French customers, potentially reducing its generalizability.
  + Proposed Mitigation: Use stratified sampling during train-test splits or apply techniques like balanced class weights or undersampling.

A graph showing distribution of countries/regions

Description automatically generated

**Gender**:

* Males slightly outnumber females in the dataset. However, the churn rate differs significantly between genders:
  + **Churn rate for females**: 25.1%
  + **Churn rate for males**: 16.5%
* Although females are underrepresented, they have a higher churn rate. For now, no mitigation is deemed necessary, but further checks such as feature importance analysis, model bias assessment, and fairness metrics will be conducted later.

A graph showing distribution of gender

Description automatically generated

**Card Type**:

* The dataset includes four card types: Diamond, Gold, Silver, and Platinum. The distribution of card types is perfectly balanced, with no dominant type.

A chart of different colored rectangular shapes

Description automatically generated

**Relationships Between Categorical Variables and Age**

* **Gender vs. Age** (Violin Plot):
  + Both genders show similar age distributions, with a peak density in the 30-40 age range.
  + Male data points show a wider spread at older ages.
* **Gender Distribution Across Geography**:
  + Males are slightly more dominant in each country.
* **Gender Distribution by Card Type**:
  + Males dominate slightly across all card types, particularly for Gold cards.

A graph with green and blue bars

Description automatically generated

**3.2 Analysis of Numerical Variables**

**3.2.1 Continuous Variables: CreditScore, EstimatedSalary, Point Earned**

* **CreditScore**:
  + The range of CreditScore values (350–850) does not align with international standards like FICO (300–850) or SCHUFA (Germany’s system). This range likely reflects internal standardization.
  + The distribution is approximately bell-shaped, resembling a normal distribution.
* **EstimatedSalary** and **Point Earned**:
  + Both are uniformly distributed across their respective ranges.

A graph showing a distribution of credit score

Description automatically generated

A graph showing the amount of salary

Description automatically generated

A graph showing a distribution of points

Description automatically generated

**Implications for Machine Learning Models:**

* Normal distributions like CreditScore might favor models sensitive to distribution (e.g., Logistic Regression, LDA, Naive Bayes), but due to uniform distributions in other features, distribution-insensitive models (e.g., Tree-Based Models, KNN, SVM, Neural Networks) are more appropriate.

**Outlier Analysis: CreditScore and Churn**

* **T-Test Results**:
  + Null Hypothesis (H₀): "There is no significant difference in the mean credit scores of churned and non-churned customers."
  + Results: T-statistic = -2.678, P-value = 0.007
  + Conclusion: There is a statistically significant difference in the mean CreditScore between churned and non-churned customers. Outliers in CreditScore will be retained, as they appear valid and contribute to model learning.

***3.2.2 Correlation Analysis***

* *A heatmap was generated to assess relationships among numerical variables (including discrete and binary ones).*

A graph with red squares and white text

Description automatically generated

**Findings**:

* No strong linear correlations were observed among the variables. For instance, variables like EstimatedSalary and CreditScore, which are often correlated in real-world datasets, showed no significant relationship.
* **Possible Reasons**:
  1. The dataset may have been pre-processed or anonymized to weaken or remove linear relationships.
  2. Relationships might be non-linear, which standard correlation coefficients cannot detect.

Pairplots were explored to investigate potential non-linear patterns, but most plots revealed random distributions, supporting the hypothesis of dataset manipulation.

Finally, an exploration of the distribution of target variable “Exited” was conducted.

A pie chart with a number of numbers

Description automatically generated

The graph highlights an **imbalance in the dataset** with respect to the target variable (Exited). Since most machine learning algorithms assume balanced class distributions, this imbalance can negatively affect model performance by biasing predictions toward the majority class.

To mitigate these effects, several strategies will be employed:

1. **Synthetic Minority Oversampling**:
   * Oversampling the minority class (Churned) using techniques like SMOTE (Synthetic Minority Oversampling Technique). This approach is preferred over downsampling the majority class to avoid losing valuable information from the dataset.
2. **Selection of Robust Algorithms**:
   * Algorithms such as **Tree-Based Models** (e.g., Decision Trees, Random Forest, Gradient Boosting) and **K-Nearest Neighbors (KNN)** are better equipped to handle class imbalances and will be prioritized.
3. **Class Weights**:
   * Incorporating class weights into the model, where a higher weight is assigned to the minority class (Churned). This penalizes the model more heavily for misclassifying the minority class, encouraging better predictions for it.
4. **Stratified Sampling**:
   * During train-test splits, stratified sampling will ensure that the training and testing datasets maintain the same class proportions as the original dataset, preventing further imbalance during evaluation.

These strategies aim to address the imbalance and ensure fair and effective model performance on both majority and minority classes.

**4. Conclusions and Next Steps**

1. **Dataset Limitations:**
   * The dataset appears heavily manipulated, with weak or non-existent linear relationships. As such, the focus will be on technique and process rather than real-world predictive power.
2. **Imbalanced Data:**
   * Geographic, gender, and target variable (Exited) imbalances will be addressed using mitigation techniques such as stratified sampling, balanced class weights, and oversampling methods like SMOTE to ensure robust model performance.
3. **Machine Learning Model Selection:**
   * The selection of machine learning algorithms will focus on addressing these imbalances and leveraging non-linear relationships in the data.
   * The preliminary selection of algorithms will be:
     1. LightGBM (Light Gradient Boosting Machine): A gradient boosting algorithm optimized for speed and efficiency, well-suited for imbalanced datasets.
     2. XGBoost (Extreme Gradient Boosting): Handles imbalanced data effectively and excels in performance for non-linear relationships.
     3. K-Nearest Neighbors (KNN): Suitable for imbalanced datasets when paired with appropriate scaling and distance weighting.
4. **Feature Importance and Fairness:**
   * Further evaluation of feature importance, model bias, and fairness metrics will guide model interpretability and ethical considerations, ensuring insights are actionable and equitable.