**Final Report: Customer Churn Analysis**

# **Objective**

The primary goal of this project is to predict customer churn using machine learning techniques and identify key factors contributing to customer attrition.

# **Exploratory Modeling and Evaluation**

**Dataset Overview**

* The dataset includes 10,000 rows and 15 features, encompassing both numerical and categorical variables.
* Target variable: **Exited** (1 = Churned, 0 = Stayed).

**Key Observations**

1. **Target Variable Distribution**:
   * 79.6% of customers stayed.
   * 20.4% of customers churned.

A pie chart with a number of numbers

Description automatically generated

1. **Feature Distribution and Correlation**:

* **Age** and **EstimatedSalary** exhibit normal distributions.
* Categorical variables such as **Geography** and **Gender** are imbalanced.

A graph with green and blue bars

Description automatically generated

1. **Initial Insights**:

* The dataset appears manipulated, as linear relationships between features are weak or non-existent.
* The imbalance in the **target variable**, **Geography**, and **Gender** suggests that mitigation techniques will be necessary.

# **Machine Learning Models Selected**

1. **LightGBM**: Gradient Boosting Decision Tree.
2. **XGBoost**: Extreme Gradient Boosting.
3. **KNN**: Baseline distance-based model.

**Handling Class Imbalance**

* **SMOTE**: Synthetic Minority Oversampling Technique was applied to oversample the minority class in the training set.
* **Stratified Train-Test Split**: Ensured consistent class proportions.

**Evaluation Metrics**

* **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **ROC-AUC**.
* Confusion matrices were used for a detailed breakdown of model predictions.

**Key Results**

**LightGBM**

* **Performance**: Near-perfect results with **Accuracy** = 100% and **ROC-AUC** = 0.9966.
* **Confusion Matrix**:
  + 1 False Positive and 2 False Negatives.

A screenshot of a computer

Description automatically generated

A diagram of a blue box

Description automatically generated with medium confidence

**XGBoost**

* **Performance**: Similar to LightGBM, with **Accuracy** = 100% and **ROC-AUC** = 0.9971.
* **Confusion Matrix**: Identical results as LightGBM.

**KNN**

* **Performance**: Weaker compared to tree-based models.
  + **Accuracy** = 98%.
  + **ROC-AUC** = 0.9957.
* **Key Issues**: Sensitivity to feature scaling, data dimensionality, and class imbalance.

A screenshot of a computer program

Description automatically generated

A diagram of a number of blue squares

Description automatically generated with medium confidence

# **Conclusion on Models: LGBM**

* **LightGBM and XGBoost** outperformed KNN, achieving near-perfect results.
* Given computational efficiency and compatibility issues, **LightGBM** was selected for further refinement.

**Model Refinement and Interpretability**

**Hyperparameter Tuning**

* **Grid Search** was performed to optimize LightGBM parameters.
* **Best Parameters**:

{'colsample\_bytree': 0.6, 'learning\_rate': 0.05, 'max\_depth': 10,

'min\_child\_samples': 20, 'n\_estimators': 500, 'num\_leaves': 40, 'subsample': 0.6}

**Performance After Tuning**: Identical to the initial evaluation.

A black background with white text

Description automatically generated

**Cross-Validation**

* **Cross-Validation Scores (ROC-AUC)**:  
  [1.0, 0.9999, 0.9999, 0.9998, 0.9999]
* **Mean ROC-AUC**: 0.99993, indicating strong generalizability.

**Feature Importance and SHAP Interpretation**

* **SHAP Summary Plot** provided insights into feature contributions:
  + **Complain**: Customers who filed complaints are more likely to churn.
  + **Age**: Older customers are less likely to churn.
  + **IsActiveMember** and **NumOfProducts**: Active customers with multiple products have reduced churn risk.

A graph with different colored lines

Description automatically generated

# **Key Insights**

1. **Best Model**: LightGBM provided robust and near-perfect results, making it the final choice for deployment.
2. **Key Predictors**:
   * Address customer complaints promptly to reduce churn.
   * Focus on retention strategies for younger customers.
   * Encourage product diversification among customers.

**Limitations**

* The dataset appeared heavily manipulated, limiting real-world generalizability.
* XGBoost and KNN were excluded due to computational cost and underperformance, respectively.

**Future Steps**

* Validate LightGBM on unseen, real-world data.
* Investigate potential relationships between **Complain** and **Satisfaction Score.**