Final Report: Customer Churn Analysis

1. Objective

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

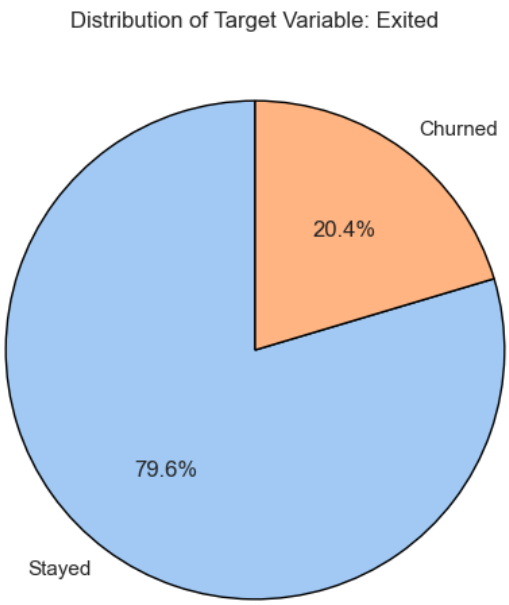
2. Exploratory Modeling and Evaluation

Dataset Overview

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

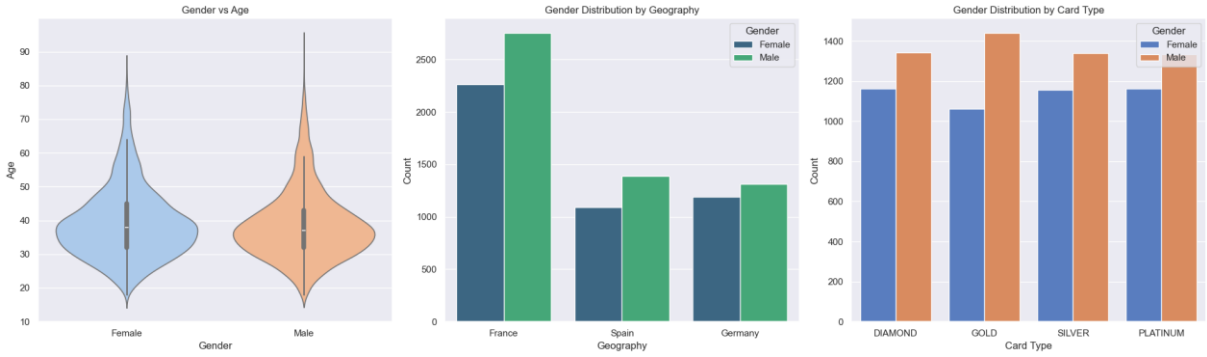
Key Observations

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



2. **Feature Distribution and Correlation**:

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



3. **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.

3. 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.

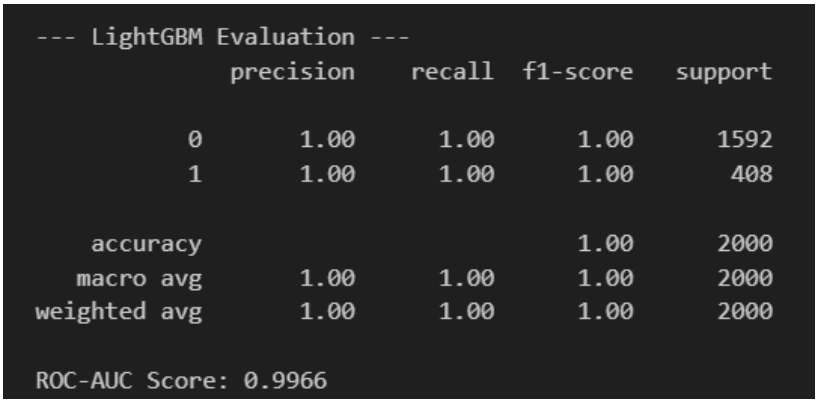
Evaluation Metrics

- **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.

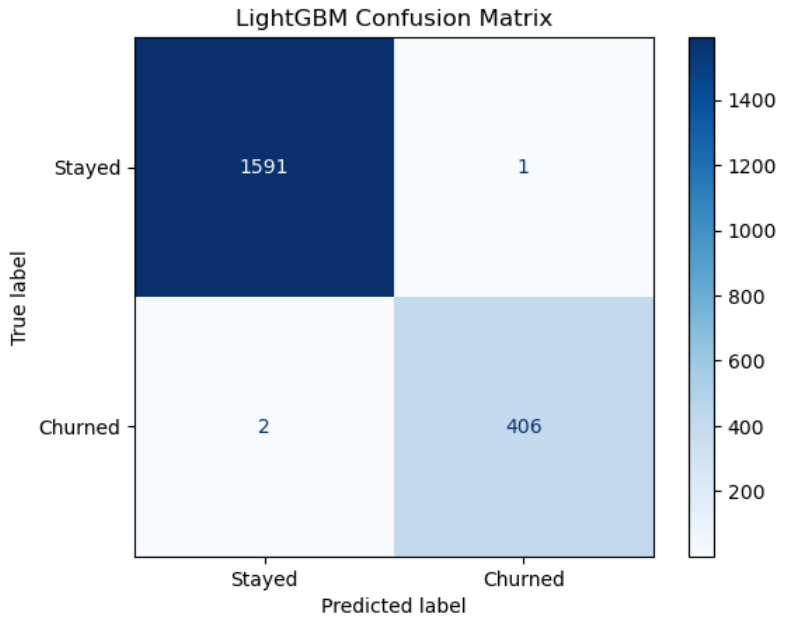


```
--- LightGBM Evaluation ---
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1592
           1       1.00      1.00      1.00       408

    accuracy                           1.00      2000
   macro avg       1.00      1.00      1.00      2000
weighted avg       1.00      1.00      1.00      2000

ROC-AUC Score: 0.9966
```



XGBoost

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

KNN

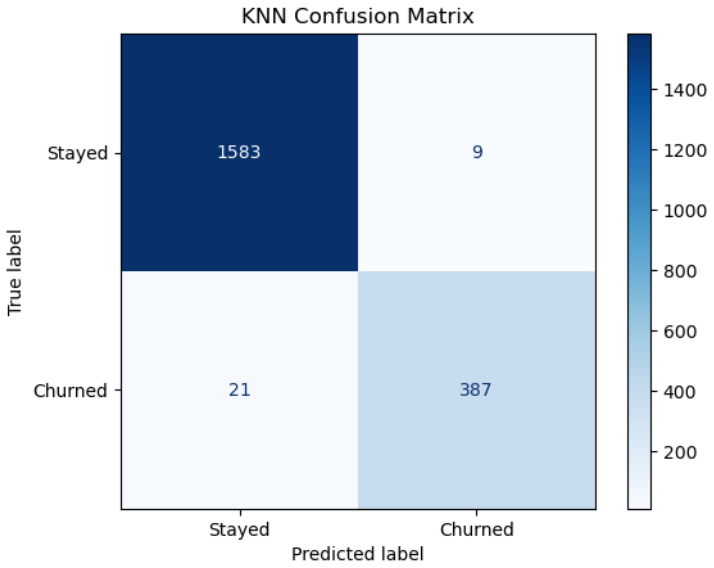
- **Performance**: Weaker compared to tree-based models.
  - **Accuracy** = 98%.
  - **ROC-AUC** = 0.9957.
- **Key Issues**: Sensitivity to feature scaling, data dimensionality, and class imbalance.

```
--- KNN Evaluation ---
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1592
           1       0.98      0.95      0.96       408

    accuracy                           0.98      2000
   macro avg       0.98      0.97      0.98      2000
weighted avg       0.98      0.98      0.98      2000

ROC-AUC Score: 0.9957
```



4. 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.

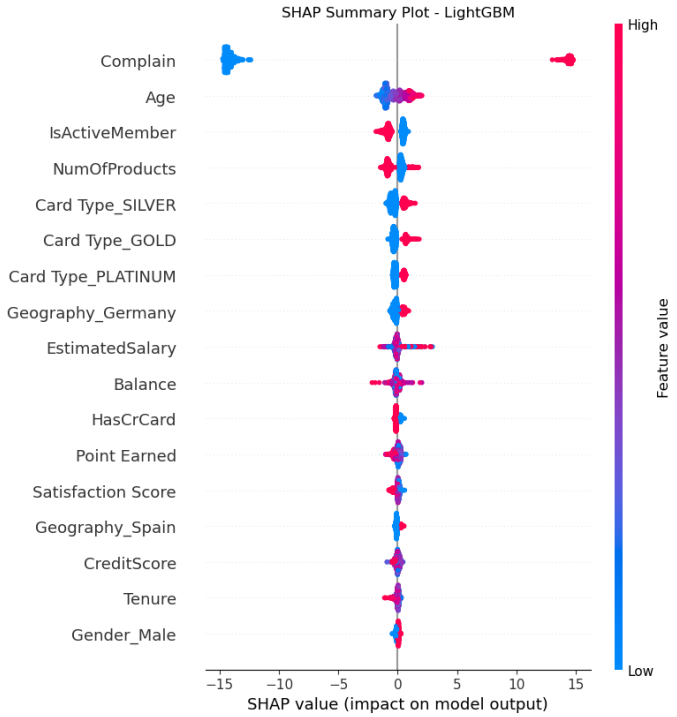
```
Confusion Matrix:
[[1591    1]
 [   2  406]]
```

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.



5. 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.**