**Exploratory Modeling and Evaluation**

**1. Goal of the Analysis**

The objective of this phase was to identify the most effective machine learning algorithms for predicting **customer churn** in a highly imbalanced dataset.

**2. Steps Taken**

**a) Model Selection**

Three machine learning algorithms were initially selected:

* LightGBM (Gradient Boosting Decision Tree)
* XGBoost (Extreme Gradient Boosting)
* K-Nearest Neighbors (KNN)

Tree-based models were chosen for their robustness and suitability for handling class imbalance, while KNN was used as a baseline comparison.

**b) Handling Class Imbalance**

The dataset exhibited significant class imbalance, particularly in:

* **Target variable**: Majority of customers stayed (79.6%), while 20.4% churned.
* **Geography** and **Gender** distributions also showed imbalances.

To address this:

1. **SMOTE (Synthetic Minority Oversampling Technique)** was applied to oversample the minority class (churned customers) in the training dataset.
2. **Stratified Sampling** was used during the train-test split to ensure consistent class proportions across training and test sets.

**c) Model Evaluation**

The three models were trained on the balanced dataset and evaluated on the test set using the following metrics:

* **Accuracy**
* **Precision**
* **Recall**
* **F1-Score**
* **ROC-AUC Score**
* **Confusion Matrix**

**3. Key Results**

**a) LightGBM**

* **Performance**: Nearly perfect with an accuracy of **100%** and an ROC-AUC score of **0.9966**.
* **Confusion Matrix**: Only **3 misclassifications** (1 false positive, 2 false negatives).

A screenshot of a computer

Description automatically generated

A diagram of a blue box

Description automatically generated with medium confidence

**b) XGBoost**

* **Performance**: Similar to LightGBM with an accuracy of **100%** and an ROC-AUC score of **0.9971**.
* **Confusion Matrix**: Identical misclassifications as LightGBM.

A screenshot of a computer program

Description automatically generated

A diagram of a blue box

Description automatically generated with medium confidence

**c) K-Nearest Neighbors (KNN)**

* **Performance**: Weaker compared to tree-based models.
  + Accuracy: **98%**
  + ROC-AUC: **0.9957**
  + Recall (Class 1): **0.95** (21 false negatives)
* **Confusion Matrix**: KNN produced **more false positives and false negatives** compared to LightGBM and XGBoost.

A screenshot of a computer program

Description automatically generated

A diagram of a number of blue squares

Description automatically generated with medium confidence

**4. Analysis of KNN Performance**

KNN’s underperformance compared to tree-based models can be attributed to:

1. **Curse of Dimensionality**: KNN struggles in high-dimensional spaces, where distances become less meaningful.
2. **Data Imbalance**: While SMOTE balanced the target variable, KNN remains sensitive to class imbalances.
3. **Scaling Sensitivity**: KNN relies on distance metrics (e.g., Euclidean distance), making it more sensitive to feature scales and variations compared to tree-based models.

**5. Next Steps**

1. **Model Refinement**:
   * Given its superior performance and stability, **LightGBM** will be the sole focus for model refinement and evaluation moving forward.
   * XGBoost will be excluded from further refinement due to its higher computational cost, despite achieving performance similar to LightGBM.
   * Perform hyperparameter tuning to further optimize their performance.
2. **Feature Interpretability**:
   * Use SHAP (SHapley Additive Explanations) to better understand feature contributions for both LightGBM.
3. **Deployment Preparation**:
   * Validate model using cross-validation.
   * Evaluate generalizability on unseen data.