**Report 3 – LightGBM (Gradient Boosting Decision Tree)**

**Objective**

This phase of the project aimed to further optimize the LightGBM model through **hyperparameter tuning** and assess its **generalizability** using **cross-validation**. Additionally, model interpretability was explored using SHAP values to identify key features driving churn predictions.

**1. Hyperparameter Tuning**

Hyperparameter tuning was conducted using **Grid Search**, varying the following parameters:

* num\_leaves: Maximum number of leaves in a tree.
* learning\_rate: Step size for model updates during boosting iterations.
* n\_estimators: Number of boosting iterations.
* max\_depth: Maximum depth of a tree.
* min\_child\_samples: Minimum number of samples required in a leaf node.
* subsample: Fraction of samples used for each boosting iteration.
* colsample\_bytree: Fraction of features considered during boosting.

The **best-performing hyperparameters** were identified as:

Best Parameters for LightGBM: {'colsample\_bytree': 0.6, 'learning\_rate': 0.05, 'max\_depth': 10, 'min\_child\_samples': 20, 'n\_estimators': 500, 'num\_leaves': 40, 'subsample': 0.6}

2. Model Performance

Despite hyperparameter tuning, the model's performance remained consistent, achieving **near-perfect accuracy**. The confusion matrix showed **1 false negative** and **2 false positives**, identical to the initial model evaluation.

A screenshot of a computer

Description automatically generated

3. SHAP Values and Feature Importance

**SHAP values** were used to interpret the predictions of the LightGBM model, providing insights into the most influential features affecting customer churn.

**SHAP Interpretation:**

* **X-axis**: Represents SHAP values – feature contributions to the model's prediction.
  + Positive SHAP values: Push predictions toward the **positive class** (Churned).
  + Negative SHAP values: Push predictions toward the **negative class** (Stayed).
* **Color**: Indicates the magnitude of feature values.
  + Red: High feature values.
  + Blue: Low feature values.

**Key Insights:**

1. **Complain:** Customers who filed a complaint (Complain = 1) have a positive SHAP value, indicating a higher likelihood of churn. Conversely, those who did not file a complaint (Complain = 0) are less likely to churn.
2. **Age**: Older customers tend to have a **negative impact** on churn predictions, meaning they are more likely to stay.
3. **IsActiveMember** and **NumOfProducts**:
   * Active customers and those with **more products** are less likely to churn.
4. **EstimatedSalary, Balance, and Point Earned**: Moderate importance with less clear directional influence compared to **Complain** and **Age**.

A graph with different colored lines

Description automatically generated

**4. Summary and Conclusions**

The project successfully identified the most effective machine learning algorithms for predicting customer churn:

* **LightGBM** and **XGBoost** performed equally well, achieving near-perfect accuracy.
* **KNN** was excluded due to its lower performance.
* Given resource constraints and compatibility issues, further refinement focused exclusively on **LightGBM**.

The interpretability analysis using SHAP values revealed **Complain** and **Age** as the most influential predictors of churn.

**5. Business Perspective**

If this were a real-world scenario, the analysis suggests that European banks should **prioritize addressing customer complaints** to mitigate churn. Additionally, targeted retention strategies for older customers and proactive engagement with **active customers** or those with **multiple products** could enhance customer retention.

**Future Steps**

* Investigate the relationship between **Complain** and **SatisfactionScore**. These features are theoretically correlated, but the dataset appears to have been pre-processed to remove linear correlations.
* Conduct further analysis using real-world, non-manipulated datasets for deeper business insights.