PURCHASE PATTERN ANALYSIS

Project 2

Supervised Learning Report

Introduction

The goal of this project is to help identify popular product bundles and predict the next items that customers are likely to purchase. By doing this, marketing strategies such as targeted promotions and product bundling can be improved, ultimately aiming to increase the average order value (AOV).

To achieve this, patterns in customer behavior are being studied. Associations between products that are frequently bought together are being explored, and models are being used to predict what customers might buy next. This work supports more personalized shopping experiences and better marketing decisions.

For this part of the project, supervised machine learning methods were used to predict the target variable 'reordered' i.e. whether a product will be included in a customer's next order. Various features based on purchase history and user behavior were analyzed to create an effective model. The results can be applied to make shopping more engaging for customers while boosting sales.

Initial Model Selection:

As the dataset is large only **30%** sample of the preprocessed data using stratified sampling, has been used for initial model selection using cross validation. Then the data is split into training and testing set with a test size of **20%**. The supervised models that have been used initially for model selection are:

1. **Logistic Regression Variants:**
    - L1 Regularization: For feature selection and interpretability.
    - L2 Regularization: To prevent overfitting.
    - ElasticNet Regularization: Combines L1 and L2 benefits for balanced feature selection and regularization.
2. **Linear Discriminant Analysis(LDA)**: Chosen for its efficiency in classifying features while reducing dimensionality.
3. **Quadratic Discriminant Analysis(QDA):** Included for cases where classes have differing variances

4. **Decision Tree:** Used for its simplicity, interpretability, and ability to capture non-linear relationships.
5. **Bernoulli Naive Bayes:** Used for probabilistic classification
6. **Random Forest:** An ensemble model chosen for its robustness, non-linear modeling capability
7. **XGBoost:** Preferred for its efficiency and strong performance on both linear and non-linear relationships.
8. **LightGBM:** Selected for faster training, memory efficiency, and native handling of categorical features.

- **<u>Cross Validation and Evaluation</u>:**
  A **5-fold cross-validation** was conducted to ensure robust performance evaluation. The **F1 score** was used as the primary metric to balance precision and recall, crucial for imbalanced datasets. Each model's average F1 score across the folds was calculated to identify the best-performing model for further refinement. The results are:

Model	Cross_Val_Scores	Mean_Cross_Val
lightGBM	[0.84799736 0.84925832 0.84755292 0.84800397 0.84740065]	0.8480426445
xgboost	[0.84479638 0.8467846 0.84451508 0.84514197 0.84372626]	0.8449928584
Random Forest	[0.82979977 0.8312377 0.8290435 0.82920699 0.81302982]	0.8264635576
Naive Bayes	[0.81615423 0.81740037 0.81529755 0.81576622 0.81608204]	0.8161400825
Logistic Regression with l2 regularization	[0.81503533 0.81589514 0.81499544 0.81561126 0.81444871]	0.8151971756
Logistic Regression with l1 regularization	[0.81503164 0.81589349 0.81500064 0.81560676 0.8144509]	0.8151966847
Logistic Regression with elastic net regularization	[0.81503164 0.81589063 0.81499995 0.81560758 0.81444871]	0.8151957032
Decision Tree	[0.81333755 0.81463314 0.81252061 0.8130407 0.81236145]	0.8131786915
Linear Discriminant Analysis	[0.8080748 0.80827769 0.80753246 0.80795217 0.80671055]	0.8077095344
Quadratic Discriminant Analysis	[0.7466837 0.75823929 0.80011634 0.7499824 0.73532752]	0.7580698504

- **Model Comparison:**
  - **Logistic Regression Variants (L1, L2, ElasticNet)**:
    - **Performance**: Performed moderately, with F1 scores slightly lower than ensemble models. Provided interpretability, particularly useful for feature selection (L1) and understanding relationships in the data.
  - **Linear and Quadratic Discriminant Analysis (LDA & QDA)**:
    - **Performance**: LDA performed better than QDA due to its ability to reduce dimensionality efficiently.LDA is computationally efficient, but it assumes linear boundaries, which limits its ability to capture non-linear dependencies.
  - **Decision Tree**:
    - **Performance**: Offered decent performance but was prone to overfitting, as shown by its inability to generalize well across folds. Highly interpretable and captured non-linear relationships.
  - **Bernoulli Naive Bayes**:
    - **Performance**: Achieved a competitive F1 score of 81.61%.Fast and effective for probabilistic predictions, making it a good baseline.
  - **Random Forest**:
    - **Performance**: Robust performance with an F1 score of 82.81%.Naturally handles overfitting by averaging multiple decision trees.
  - **XGBoost and LightGBM**:
    - **Performance**: Both models achieved the highest F1 scores among all tested models (84.49% and 84.80%, respectively).
- **Conclusion:**
  After cross validation on each model **four** best models are selected on the basis of their average **F1 Score**. In this dataset, **LightGBM**, **XGBoost**, **Random Forest** and **Naive Bayes** are the four best performing models with their respective average F1 scores which are **84.80%**, **84.49%**, **82.81%** and **81.61%.**

Final Model Selection

After the initial model selection, the **100%** of the preprocessed data has been used for final model selection. The supervised models that are used in final model selection are **LightGBM**, **XGBoost**, **Random Forest** and **Naive Bayes.** The data is split into training and testing sets with a test size of **20%**.

- **Cross Validation and Evaluation:**
  A **5-fold cross-validation** was also conducted in final model selection to ensure robust performance evaluation. The **F1 score** was used as the primary metric to balance precision and recall, crucial for imbalanced datasets. Each model's

average F1 score across the folds was calculated to identify the best-performing model for further refinement. The results are:

Model	Cross_Val_Scores	Mean_Cross_Val
lightGBM	[0.8479707 0.84852718 0.84841494 0.84799102 0.84759148]	0.8480990646
xgboost	[0.84505531 0.84539022 0.84529326 0.84475559 0.84403358]	0.844905592
Random Forest	[0.82916487 0.82318999 0.82998041 0.82951172 0.82923899]	0.8282171967
Naive Bayes	[0.81574358 0.81635277 0.81638002 0.81566897 0.81556617]	0.8159423027

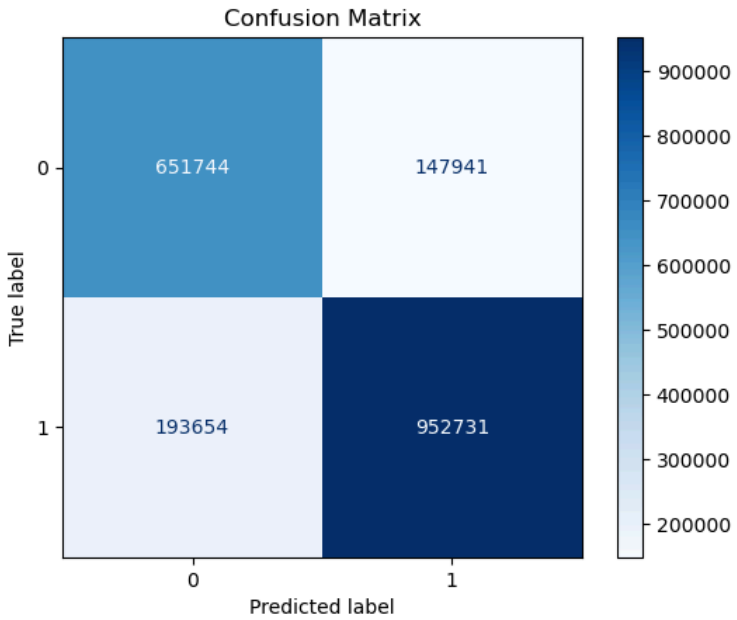
- **Model Comparison:**
  - **LightGBM**:
    - **Performance**: Outperformed all other models with the highest F1 score (84.80%) on the full dataset. Its ability to handle categorical data natively and efficiently process large datasets made it the best fit.
  - **XGBoost**:
    - **Performance**: Achieved the second-highest F1 score (84.49%), close to LightGBM.
  - **Random Forest**:
    - **Performance**: Achieved a slightly lower F1 score (82.81%), still showing robust performance.
  - **Bernoulli Naive Bayes**:
    - **Performance**: Had the lowest F1 score (81.61%) among the final models but remained competitive given its simplicity. Extremely fast and computationally efficient, making it a useful lightweight alternative.
- **Conclusion:**
  After applying cross validation on the four initially selected models on the whole dataset, **LightGBM** is performing best out of those four models with the maximum average **F1 Score 84.80%.**

Final Model Evaluation

The LightGBM model was selected as the final model based on its superior performance during the evaluation phase. The following metrics were used to assess the model's effectiveness on the test set:

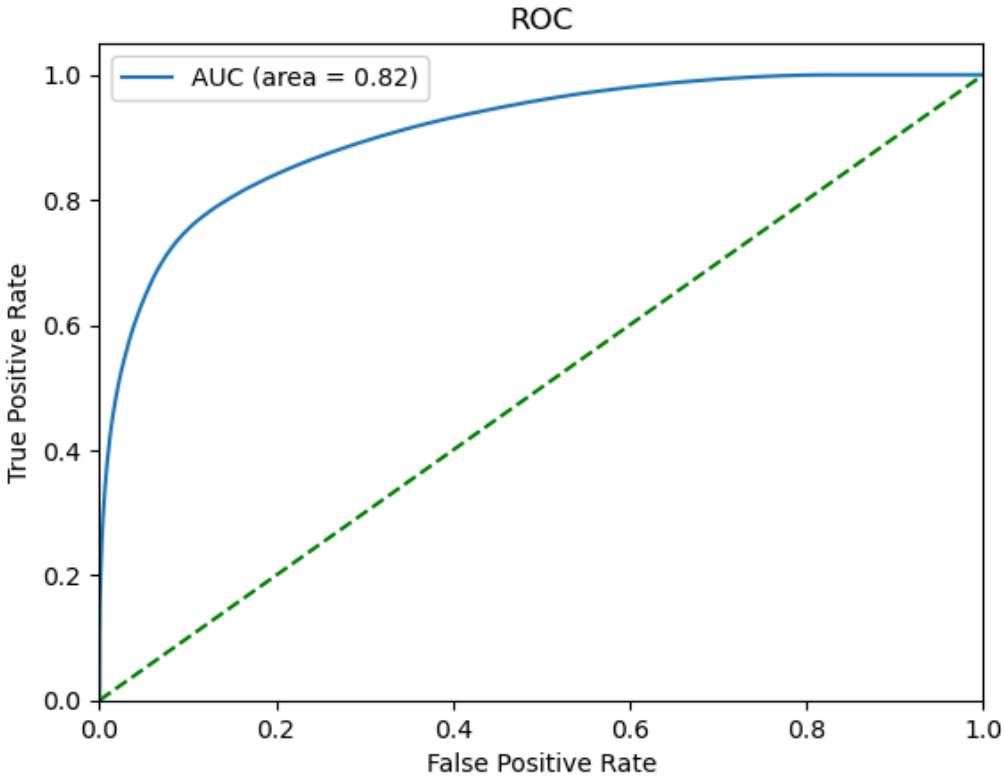
	precision	recall	f1-score	support
0	0.77	0.82	0.79	799685
1	0.87	0.83	0.85	1146385
accuracy			0.82	1946070
macro avg	0.82	0.82	0.82	1946070
weighted avg	0.83	0.82	0.83	1946070

- **Accuracy Score**: **0.8244**
  The model correctly classified 82.44% of all instances, demonstrating strong overall performance.
- **Precision Score**: **0.8656**
  LightGBM achieved a high precision, indicating that a significant proportion of predicted positive cases were indeed true positives. This is particularly important in reducing false positives.
- **Recall Score**: **0.8311**
  The model captured 83.11% of actual positive cases, reflecting its effectiveness in identifying true positives, even at the cost of some false negatives.
- **F1 Score**: **0.8480**
  The F1 score, which balances precision and recall, highlights the model's robustness, particularly in managing the trade-off between false positives and false negatives.
- **Confusion Matrix:**



- **True Label 0**: The model correctly predicted 651,744 instances as class 0, and incorrectly predicted 147,941 instances as class 1.
- **True Label 1**: The model correctly predicted 952,731 instances as class 1, and incorrectly predicted 193,654 instances as class 0.
- The matrix displays the raw counts of the predictions, which is useful for understanding the model's performance.

- **ROC-AUC Curve:**



- The Area Under the Curve (AUC) is 0.82, which indicates that the model has a reasonably good performance. AUC values range from 0 to 1, with 1 representing a perfect model and 0.5 indicating a random classifier.
- The closer the ROC curve is to the top-left corner (0, 1), the better the model's performance, as it indicates a high TPR and low FPR. From ROC-AUC curve it can be said that the model has room for improvement

Future Work

Model Optimization:

- Perform hyperparameter tuning for LightGBM to further optimize its performance
- Experiment with ensemble techniques, combining predictions from LightGBM, XGBoost, and Random Forest to create a meta-model.

Conclusion

This project successfully leveraged supervised learning methods to predict reordered items, with LightGBM emerging as the best-performing model based on F1 score, precision, and recall metrics. The findings can support data-driven marketing strategies such as targeted promotions and product bundling to enhance customer satisfaction and boost sales. Despite its effectiveness, the model leaves room for improvement in areas such as recall and overall robustness, as indicated by the ROC-AUC curve. Future enhancements, including advanced feature engineering and model optimization, can further refine performance and ensure scalability for the real-world.