Holly Cohan,Yesh Onipede, Yu Xia

Week 14 Peer Review: Santander's Product Recommendation System - Maria Alice Fagundes Vieira and Aritra Ray

Link to GitHub: https://github.com/mariaalice-fv/applied_analytics_project

Github

- **Repository**: The repository is well-documented, with a clear README file that outlines the objectives, dataset.
- **Folder Structure**: The folder structure is logical, using cookiecutter folder structure, with clear separations between data, code, and outputs, making it easy to navigate.
- **Suggestions**: Consider giving the project a more compelling title and adding visual elements to the README for more storytelling

Code/notebook

- **Code**: The code is clean and well-commented, with intuitive variable names that enhance readability.
- **Markdown:** Markdown cells are effectively used to explain each step, providing context and improving the notebook's flow.
- **Suggestion**: It might be better to consider adding subheadings in the notebook that follows the MDLC for easier navigation.

Code output

- **Data Understanding and Exploratory Data Analysis (EDA):** The EDA section is thorough and well-documented. Highlighting demographic distributions, purchasing trends, and the geographic spread of customers provides a solid understanding of the dataset's characteristics. These findings help connect the data to the problem objectives effectively. The details about the skewed distribution in gender and the age ranges associated with specific product categories are useful for understanding the dataset. The visuals chosen provide graphical evidence for the observed trends and patterns.
- **Graph Background**: The LIME example is very informative in terms of feature importance. It might be better to add a white background to the cell output for better readability for audiences who are using dark mode in their development environment.
- **Environment Dependencies**: It might be helpful to include the list of environment dependencies reflecting planning for reproducibility and deployment of the report in the notebook as well.

Report

Your report is a comprehensive and well-structured overview of the Model Development Life Cycle. It is clear that significant time and effort was put into documenting each stage. The level of detail provided for

each section demonstrates a clear understanding of the challenges involved in developing a machine learning model for personalized recommendations.

- **Problem Definition and Objective:** The problem statement is clearly articulated and aligns well with the business objective of improving Santander Bank's product recommendation system. The focus on customer satisfaction, product adoption, and operational efficiency sets a strong foundation for the project. Framing the challenge as a predictive task based on historical customer behavior and transaction data is an effective approach that provides a clear path for the rest of the MDLC.
- **Data Preparation:** The data preparation steps are well-explained, and your choices for handling missing values, encoding categorical variables, and scaling numerical features are appropriate. Using strategies such as mean or median imputation for numerical variables and mode imputation for categorical variables maintains the data's integrity. Your engineered variables seem to capture additional important relationships existing in the data. The decision to retain outliers is a justifiable approach since it avoids potentially removing meaningful data points that could influence the model's performance. The stratified data splitting ensures balanced representation across training and test sets, which supports the model's generalizability.
- **Model Development:** The rationale for selecting Collaborative Filtering as the final model is well-reasoned and aligns with the problem's objectives. Comparing the performance of Collaborative Filtering with other models like XGBoost and Random Forest highlights a methodical approach to model selection. Highlighting evaluation metrics such as ROC AUC and F1 Score emphasizes your focus on assessing the model's ability to generalize and handle class imbalances. The explanation of hyperparameter optimization is clear, even though minimal tuning was required. Prioritizing a simpler model with consistent performance makes sense given the dataset and business needs. The customer behavior insights collected from Collaborative Filtering add an extra layer of interpretability to the model.
- **Model Evaluation and Validation:** The evaluation section provides a detailed analysis of how the model performs on unseen data, focusing on metrics like ROC AUC and F1 Score. The explanation of why Collaborative Filtering outperformed other models is logical, especially in the context of leveraging customer-product interaction data. Acknowledging the model's limitations, such as its challenges with class imbalance during training, reflects that your assessment is realistic. The improvements made to improve the model, such as retaining duplicate rows and incorporating the date column as a feature, are logical and well-supported by reasoning. They appear to have contributed to a noticeable improvement in F1 Scores.
- **Deployment and Maintenance:** The deployment plan is practical and aligns well with the project's requirements. Using a batch processing approach for monthly and biweekly updates maintains computational efficiency while maintaining timely recommendations. The daily batch predictions enhance the system's adaptability. The monitoring plan, with defined thresholds for performance metrics, ensures successful ongoing evaluation. Including business-focused metrics like conversion rates in addition to model metrics will help maintain model relevance and accuracy. The maintenance plan outlines an effective strategy for addressing potential data and concept drift.
- **Suggestions for Improvement:**

- **Visuals**: While the report includes multiple useful visuals, providing more detailed explanations of how these visuals are relevant may be useful and could strengthen the EDA section. It also might be useful to move some of the visuals to an Appendix to keep the report a bit more concise and readable.
- **Hyperparameter Optimization**: While minimal optimization was necessary for the chosen model, a more detailed explanation of other potential hyperparameters that were tested for other models might add value to your report.
- **Ethical Considerations**: Expanding on the ethical considerations mentioned in the conclusion would strengthen the report. It may be useful to include a separate dedicated section on bias detection and mitigation strategies.

Overall Assessment:

This report demonstrates a strong understanding of the Model Development Life Cycle and provides a comprehensive approach to addressing Santander's business problem. The structure is logical and thorough, with justified decisions made throughout the process. Although there are areas where additional details or metrics could enhance the report, the report effectively finds a balance between business relevance and using proper technical methods.