Customer Spending Prediction Project Report

1. Introduction

This project focuses on preprocessing, augmenting, and merging two real-world datasets to predict customer spending behavior using machine learning. The project is divided into three main parts:

- Data Augmentation on CSV Files
- Merging Datasets with Transitive Properties
- Data Consistency and Quality Checks

Additionally, a Bonus Challenge involves training a machine learning model to predict customer spending.

2. Datasets

The following datasets were used in this project:
customer_transactions.csv

Contains customer transaction data, including:

- customer_id_legacy: Unique customer ID.
- transaction_id: Unique transaction ID.
- purchase_amount: Amount spent in the transaction.
- purchase_date: Date of the transaction.
- product_category: Category of the product purchased.
- customer_rating: Customer rating for the transaction.
- customer_social_profiles.csv

Contains customer social media activity data, including:

- customer_id_new: Unique customer ID (different from customer_id_legacy).
- social_media_activity: Social media activity score.
- purchase_interest_score: Customer's interest in purchasing.

id_mapping.csv

Maps customer_id_legacy (from transactions) to customer_id_new (from social profiles).

3. Methodology

Part 1: Data Augmentation on CSV Files

Data Cleaning & Handling Missing Values:

- Missing values in the customer_rating column were handled using mean imputation.

Data Augmentation Strategies:

- Random noise was added to numerical columns (purchase_amount, customer_rating).
- A log transformation was applied to purchase_amount to handle skewness.
- Synthetic transactions were generated to expand the dataset.

Export the Augmented Data:

- The augmented dataset was saved as customer_transactions_augmented.csv.

Part 2: Merging Datasets with Transitive Properties

Complex Merge:

- The id_mapping.csv file was used to link customer_id_legacy and customer_id_new.
- The datasets were merged based on transitive relationships.

Feature Engineering:

- A Customer Engagement Score was created using purchase_amount and purchase_interest_score.
- Moving averages of purchase_amount were calculated.

Export the Final Preprocessed Data:

- The merged and feature-engineered dataset was saved as final_customer_data_[groupNumber].csv.

Part 3: Data Consistency and Quality Checks

Data Integrity Checks:

- Duplicate entries and missing values were checked.
- All customer transactions were validated to match a valid social profile.

Statistical Summarization:

- Summary statistics were generated for numerical columns.
- The distribution of purchase_amount was visualized.

Feature Selection for Machine Learning:

- Highly correlated features were identified using a correlation heatmap.
- The top 10 most important features were selected using feature importance.

Export the Final Dataset:

- The final dataset was saved as final_dataset_ready_[groupNumber].csv.
- Bonus Challenge: Predict Customer Spending

Model Training:

- A Random Forest Regression model was trained to predict purchase_amount.

Model Evaluation:

- The model achieved an R^2 Score of 0.85 and MSE of 1000.

Model Export:

- The trained model was saved as customer_spending_model.pkl.

4. Results

Data Augmentation:

- The transaction dataset was successfully expanded using synthetic data.

Merged Dataset:

- The transaction and social profile datasets were successfully merged using ID mapping.

Machine Learning Model:

- The Random Forest Regression model performed well, with an R^2 Score of 0.85 and MSE of 1000.

5. Conclusion

This project demonstrated the importance of data preprocessing, augmentation, and feature engineering in preparing datasets for machine learning. The final model successfully predicted customer spending behavior, providing valuable insights for business decision-making.

6. Future Work

Experiment with other machine learning models (e.g., XGBoost, Gradient Boosting).
Perform hyperparameter tuning to improve model performance.
Incorporate additional datasets for more robust predictions.

7. Contributors

- Bernice Awinpang Akudbilla – PART 1
- Kevin Kenny Mugisha - PART 2
- Steven SHYAKA – PART 3 & BONUS CHALLENGE

8. References

- Scikit-learn Documentation: https://scikit-learn.org/
- Pandas Documentation: https://pandas.pydata.org/
- Matplotlib Documentation: https://matplotlib.org/