Group 1.
Mathias Ducatillon,
Shopica Venkataramanan,
Hunter Lewis
Marcos Lopez

PROGRESS REPORT

PROJECT SUMMARY

REPORT DATE	PROJECT NAME	PREPARED BY
24/04/2023	E-commerce Platform for Identifying Repeat Buyers	Group 1

STATUS SUMMARY

This report provides an update on the progress made on a project aimed at identifying repeat buyers on an e-commerce platform. The project involves data visualization, feature engineering, dataset statistics and feature ranking, prediction model, and model evaluation.

PROJECT OVERVIEW

TASK	% DONE	NOTES
Data Visualization	**50%**	To visually analyze the dataset, several visualization techniques were used. The Box plot was used to show distributions of numeric data values, especially when comparing them between multiple groups. Other visualization techniques used include histograms, and distribution plots. Visualizations also showed the data to be very imbalanced, with 90% of one-time purchases, and 10% of repeated purchases. The goal was to provide insights from the data, and the visualizations created were able to achieve this.
Feature Engineering	**70%**	From the given information, new features were created to correlate users and merchants. 41 new features were created. These features were derived from existing information provided in the dataset and from our own understanding and creativity. Some of the features created include quantity of seen items, frequency of actions (such as click, add-to-cart, purchase, and add-to-favorite), time spent on the platform, and correlations between brands and sellers.
Dataset Statistics and Feature Ranking	**40%**	The dataset's statistical summary was provided, and feature ranking was performed using SHAP. The top features in order of importance were the quantity of clicks made on the platform, and total actions made on each seller. These features were also used for PCA feature reduction, where the optimal number of features was identified to be 3.
Prediction Model	**60%**	Different combinations of features were iterated through to identify the optimal features and remove potential correlated features. The Bayes Classifier was used to identify customers who would be repeat buyers or not. The model was trained on 80% of the data and tested on 20% of the data. The Bayes classifier obtained a 85% accuracy score. Non-parametric techniques like nearest neighbor and parametric techniques such as MLP (Multi-Layer Perceptron) were also used for classification. A comparative study of performance analysis was performed, concluding that nearest neighbor outperformed the Bayes and MLP classifier by obtaining 97.5% accuracy score.
Model Evaluation	**50%**	The performance of each classification technique was evaluated using metrics such as accuracy, precision, recall, F1 score, ROC curve, and confusion matrix. Each

		technique's results were compared, and insights were provided on which method works best for the given classification problem. Recommendations were made for the e-commerce platform, including offering personalized recommendations to customers, improving the user interface, and investing in customer retention strategies.

CONCLUSION

The project has made considerable progress in identifying repeat buyers on an e-commerce platform. Data visualization, feature engineering, dataset statistics and feature ranking, prediction model, and model evaluation were all mostly successfully carried out. The next step for the project is to finish all these tasks and propose techniques to convert one-time buyers to loyal customers.