CARNEGIE MELLON UNIVERSITY

COURSE: PROGRAMMING FOR DATA ANALYTICS

COURSE CODE: 04-638-A

INSTRUCTOR: PROFESSOR GEORGE OKEYO

ASSIGNMENT: FINAL PROJECT: CUSTOMER SEGMENTATION AND CLASSIFICATION USING MACHINE LEARNING

REPORT TITLE: A DATA-DRIVEN EXPLORATION OF SEGMENTATION AND PREDICTIVE MODELING FOR CONSUMER PRODUCT RECOMMENDATIONS

PREPARED BY

Name: Samalie Piwan

AndrewID: spiwan

Date: 17th December 2023

ABSTRACT

In this project, we address the challenges of understanding and predicting customer behavior for a business. The context involves a customer base with varying transaction patterns. The goal of this analysis identify distinct customer segments and predict which transaction tier a customer belongs to, based on their transaction history and in so doing, enhance marketing strategies. The approach includes exploratory data analysis, feature engineering, and the application of for classification and prediction machine learning algorithms. Results showcase robust customer segments and predictive models, contributing to more targeted marketing initiatives. In conclusion, this project empowers businesses to make data-driven decisions, fostering customer-centric strategies and ultimately improving overall operational efficiency.

BACKGROUND AND PROBLEM STATEMENT

In the rapidly evolving landscape of business, understanding and responding to customer behavior are pivotal for sustained success. This project stems from the recognition that a diverse customer base necessitates an analytical approach to marketing consumer products.

The challenge lies in effectively classifying and predicting customer behavior to consumer products. Using purchasing patterns and distinct customer segments is crucial for optimizing marketing efforts and addressing the overarching need for a more targeted and adaptive business approach.

APPROACH

i. Data Preparation

The dataset required for this analysis was contained in a CSV file named ‘CC General.csv’. This file was loaded into a pandas dataframe for preparation and analysis

ii. Exploratory Data Analysis

To start, the information about the names, datatypes and nullability of the columns in the dataframe was described. The dataset was then checked for missing values which would cause our model training to fail. Values found were handled using imputation with the mean value in each column. The dataset was also checked for outliers which would skew the training and test data. Values found were replaced with the lower quartile and upper quartile values in each column. Finally, the distribution of column data was visualized to determine the central tendency and variability of the data.

iii. Performance Metrics

The performance of the unsupervised model was evaluated using the silhouette score that evaluates cluster cohesion and separation, and the Calinski and Harabasz score that evaluates between-cluster dispersion and within-cluster dispersion. The performance of the supervised model was evaluated using accuracy, precision, recall and F1 scores. A classification report was also generated for each model, to determine the performance metrics of the model on each cluster generated.

iv. Unsupervised Model Building

For the unsupervised model, a scaled version of the main dataset was used. Scaling was done using the StandardScaler to normalize the data and reduce variance. KMeans Clustering was selected as the unsupervised model for this project. The KElbowVisualizer was used to select the optimal number of customer clusters, and the scaled dataset run against the KMeans model to generate the customer clusters. Principal Component Analysis was then leveraged to visualize the customer clusters.

v. Supervised Model Building

The clusters generated in step iv were appended to the main dataframe, so that each customer had a customer segment assigned. This dataframe was saved to a file ‘spiwan-cc-labeled.csv’, which was then loaded to build the supervised model. The generated customer clusters were selected as the dependent variable, and all the other features in the dataset as the independent variable. The DecisionTreeClassifier was selected as the supervised learning module. To train and test the model, the dataset was split into test and train data using the train_test_split() function.

vi. Model Debugging

To debug the model for overfitting and underfitting, the scikit learn ‘learning_curve’ and ‘validation_curve’ functions were used. ‘learning_curve’ evaluates a model's performance on both the training set and a cross-validated test set, where a significant gap between the training and cross-validated test curves suggests overfitting while high errors that do not show significant improvement with additional data suggest underfitting. ‘validation_curve’ generates a validation curve by varying one hyperparameter while keeping others constant, where a significant difference between training and test performance for certain hyperparameter values indicates overfitting while high errors low performance for both training and test sets across all hyperparameter values suggests underfitting.

vii. Model Deployment

The final benchmark model selected after feature selection and hyperparamterization was saved into a .pkl file. Next, a Flask application was created to host the frontend of the application that receives the user’s input and returns the predicted customer cluster. On the application frontend, only the optimal features selected after feature selection were required as user input.

User input was then sent to the application backend, which run a prediction based on the final benchmark model and gave a resulting predicted customer cluster.

These results were then sent to the application frontend and displayed back to the user. All results under cluster 0 were labelled ‘Gold Tier Customer’ while cluster 1 were labelled ‘Silver Tier Customer’.

RESULTS AND DISCUSSION

i) Exploratory Data Analysis results

Dataset Information

Excluding the ‘CUST_ID’ column, the dataset was found to contain columns with only float and int values

Missing Values

Two columns MINIMUM_PAYMENTS' and 'CREDIT_LIMIT’ where found to have missing values. These were handled using imputation with the mean value for each column

Data Distribution

The data is generally right skewed, except for 'PURCHASES_FREQUENCY' column where the tail is skewed to the left. This indicates that lower amounts across all columns for example purchases, cash advances, credit limits, etc. 'TENURE' is the only left-skewed column in the dataset

Outliers

The 'PURCHASES', 'ONEOFF_PURCHASES', 'CASH_ADVANCE', 'PAYMENTS' and 'MINIMUM_PAYMENTS' columns contain outliers. This indicates that some customers are transacting with significantly larger amounts than the general customer base.

ii) Evaluation results for unsupervised model

a) Silhouette Score evaluation

The unsupervised KMeans model has a good silhouette score of approximately 0.435. This indicates that the clusters formed by the model are on average well-separated, and the data points within each cluster are relatively close to each other compared to points in other clusters. This suggests a reasonable degree of cluster cohesion and separation, showing that our model can correctly identify the underlying patterns in the data

b) Calinski and Harabasz Score evaluation

The unsupervised KMeans model has a Calinski and Harabasz score of 7912.5. This suggests that the clusters formed by the model have a high degree of separation and compactness. Our score indicates strong cluster structures, showing that our model can correctly identify the underlying patterns in the data

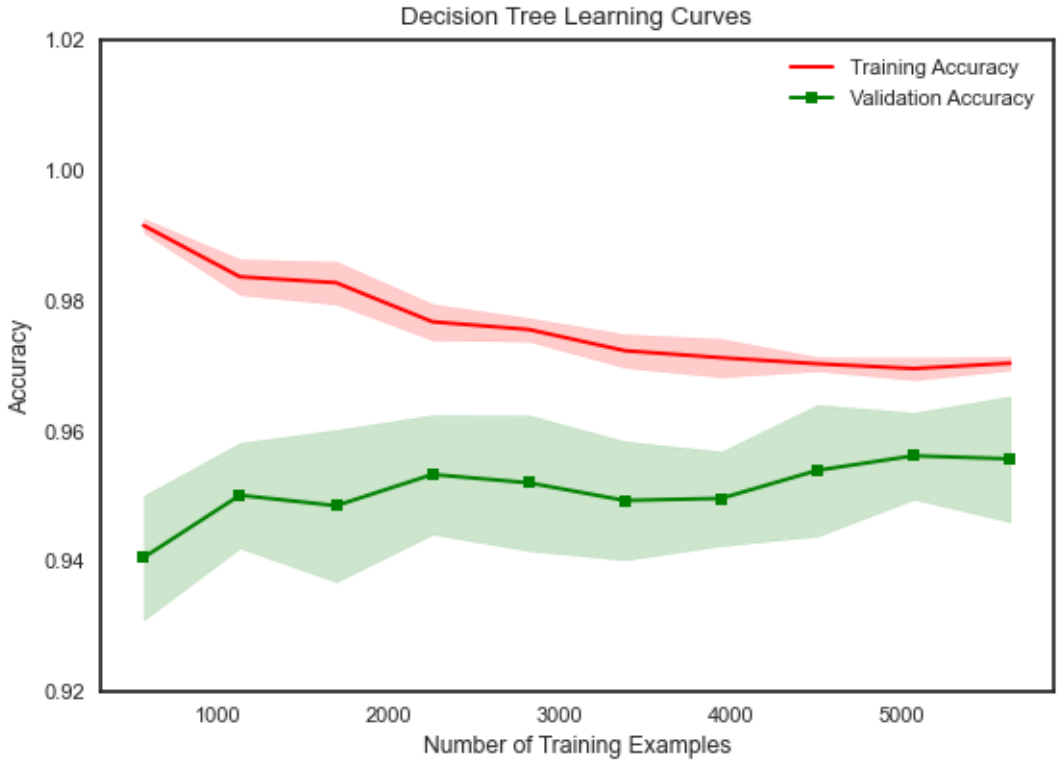
iii) Evaluation results for all supervised models

The first model that used all features and no hyper-parameters had an overall accuracy of 0.95, meaning it correctly predicted 95% of the customer clusters.

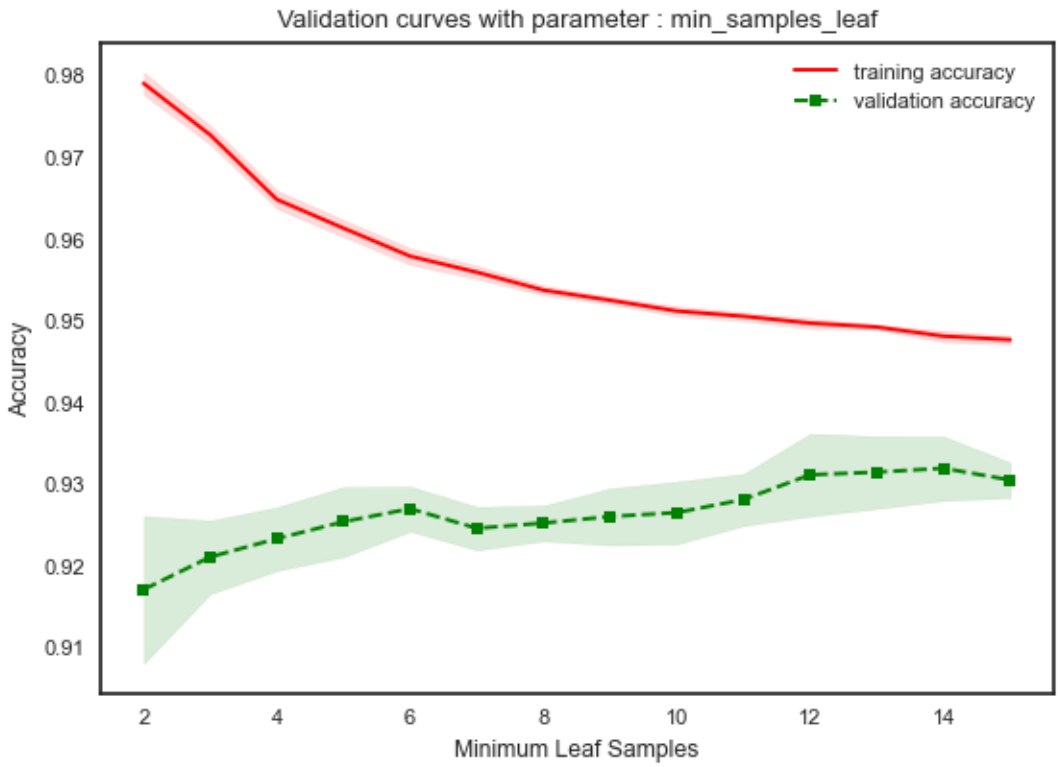
After feature selection, the second model has an overall accuracy of 0.92, which is lower than the initial accuracy of 0.95. While the model performance has lowered, this could be attributed to a reduction in the overfitting after feature selection. Therefore, we can conclude that the robustness of the model has improved.
The third hyper-parametrized model has an accuracy score of 0.94, which was is higher than the score of the second model and nearly as high as that for the first model.
Therefore, because the hyper-parametrized model considers both optimally selected features and the best hyperparameters, we concluded that this model had the best overall performance and was selected as the final benchmark model.

iv) Results of model debugging

The learning curves show a significant gap between the training accuracy curve and validation accuracy curve with a low number of training examples, which indicated overfitting. However, this gap reduces as number of training examples increase, which indicates improvement in performance and a reduction in overfitting.

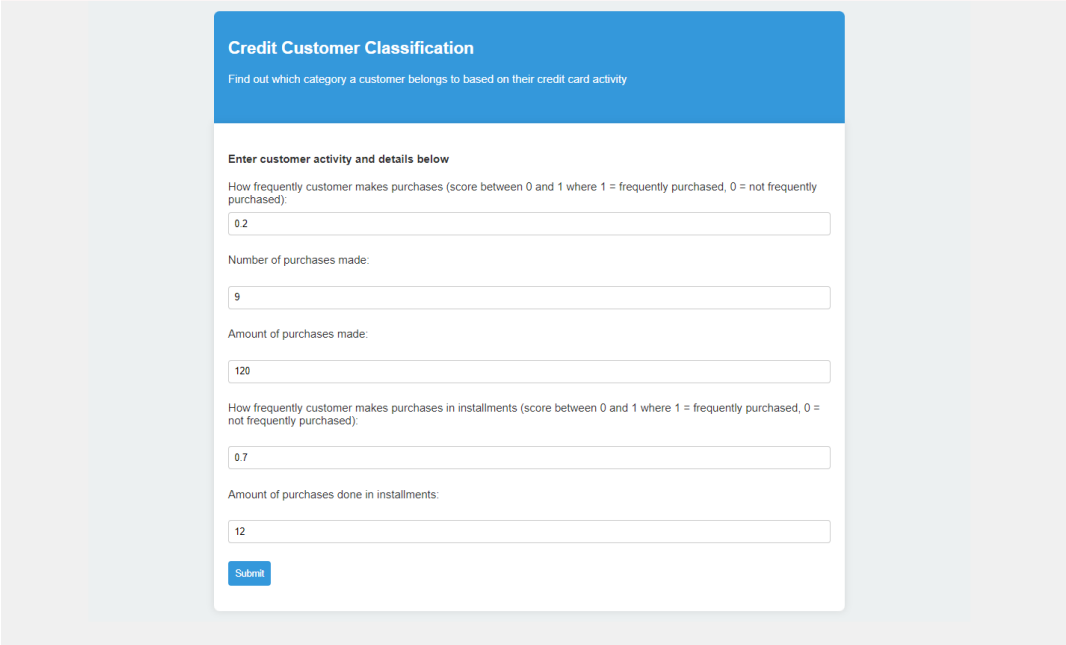


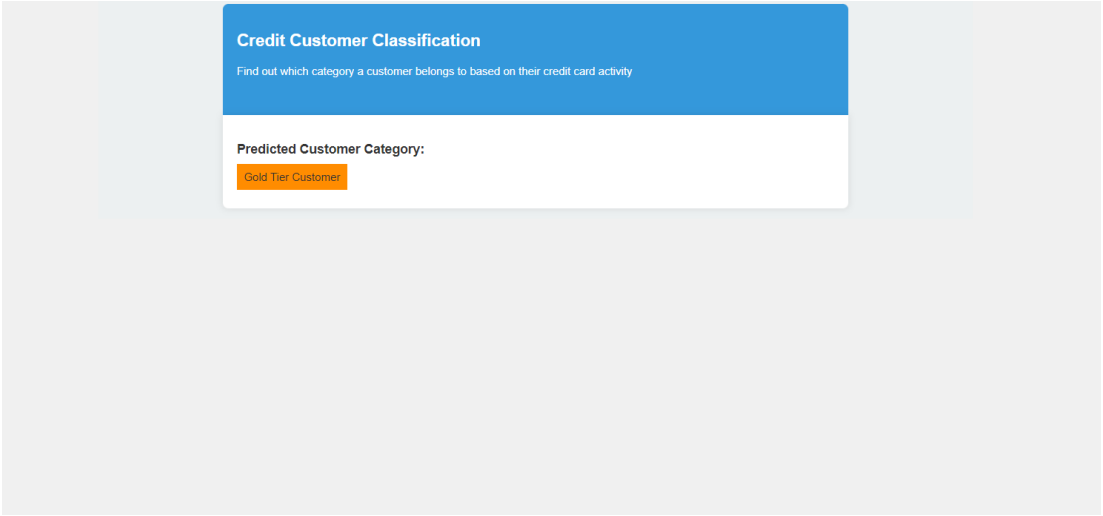
The validation curves are generated considering ‘min_samples_leaf’ as the parameter. The curves show a significant gap between the training accuracy curve and validation accuracy curve with a low number of leaves, which indicated overfitting. However, this gap reduces as number of leaves increase, which indicates improvement in performance and a reduction in overfitting.



v) Web application screenshots

The screenshots below demonstrate the model's ability to classify both cluster 0 (Gold Tier) and cluster 1 (Silver Tier) customers based on the customer's transaction details.





vi) Discussion

The project's findings illuminate key aspects of customer behavior, providing a foundation for strategic discussions.

We discover that customers fall into two main segments and can predict what segment a customer may fall into based on their transaction behavior.

The identified customer segments offer a data-backed understanding, allowing for targeted customer segmentation. The predictive models showcase the potential for anticipating future customer actions, enabling proactive decision-making.

Discussions around implementing these insights into business practices should focus on refining marketing strategies, improving customer engagement, and fostering adaptability in response to evolving market dynamics.

CONCLUSION

In conclusion, this project delivers actionable insights for businesses by uncovering distinct customer segments and developing predictive models. With a focus on personalized marketing, and adaptability, the project enhances overall efficiency, fostering a customer-centric approach to drive sustained growth and satisfaction.