Abstract

In this report, we present a comprehensive analysis of a Credit Card Segmentation and Prediction project. The context involves utilizing machine learning techniques to classify credit card users into distinct segments. The problem stems from the need to understand user behavior and tailor financial services accordingly. Our main purpose is to develop a predictive model for segmenting credit card users based on their transactional patterns. The approach includes a detailed process overview, data preparation, exploratory data analysis, performance metric selection, unsupervised and supervised model building, learning curve-based model debugging, and model deployment. Results showcase effective clustering and predictive capabilities. The web application screenshots demonstrate the practical implementation of our model.

Background and Problem Description

The credit card segmentation problem is a crucial aspect of the financial sector that aims to categorize credit card users into distinct segments to tailor their offerings and improve customer satisfaction. The problem seeks to overcome challenges related to diverse user behavior, risk management, and marketing optimization. Credit card segmentation is pivotal for enhancing customer satisfaction, optimizing marketing strategies, and managing risks associated with financial services. It allows financial institutions to better understand and meet the diverse needs of their clientele. Segmentation contributes to more accurate fraud detection by understanding normal user behavior within each segment. Tailoring services to specific user segments improves the overall customer experience, leads to cost savings and improved operational efficiency, and provides a competitive edge in the financial sector. Accurate segmentation contributes to proactive risk management.[1][2]

Approach

1. Data Preparation

- **Data Loading:** Load credit card dataset, addressing missing values, outliers, and standardizing features.
- **Handling Missing Values:** Robust strategies applied for critical features.
- **Outlier Handling:** Appropriate techniques used for outlier detection (Violin plot which allows to visualize the distribution of a numeric variable for several groups was used.) and management.
- **Feature Scaling:** Numerical features standardized for consistency.

2. Exploratory Data Analysis

- **Descriptive Statistics:** In-depth review of summary statistics.
- **Data Distribution:** Visualization of numerical feature distributions.
- **Correlation Analysis:** Investigation into feature relationships.
- **Cluster Visualization:** Preliminary insights gained from unsupervised clustering.
- **Feature Importance:** Identification and understanding of key features.

3. Performance Metrics

- **Selection Criteria:** Definition and selection of suitable performance metrics.
- **Evaluation Strategy:** Comprehensive strategy for model performance evaluation.

4. Unsupervised Model Building

- **Clustering Algorithm:** KMeans employed for revealing natural data groupings.
- **Optimal Cluster Selection:** Determination of optimal clusters using metrics like silhouette score.
- **Cluster Assignment:** Data points assigned to clusters.

5. Supervised Model Building (Including Preliminary Data Labeling)

- **Labeling Strategy:** Exploration of preliminary labeling methods using insights from unsupervised learning.
- **Classifier Selection:** RandomForestClassifier chosen for supervised learning.
- **Feature Selection:** Optimization of feature set based on exploratory analysis.

6. Model Debugging Using Learning Curves

- **Learning Curve Analysis:** Understanding model performance and fine-tuning based on insights.
- **Model Fine-Tuning:** Hyperparameter adjustments guided by learning curve analysis.

7. Model Deployment

- **Web Application:** The development of a user-friendly web application was done.
- **Integration:** Incorporation of the model for real-time predictions was conducted.
- **User Interaction:** User-friendly interface for input, prediction, and result visualization.

Exploratory Data Analysis (EDA)

1. **Key Findings:**

- Strong positive correlation between purchases and oneoff_purchases, indicating customers making more purchases also engage in one-off transactions.
- Moderate positive correlation between purchases and installments_purchases, suggesting customers making many purchases also utilize installment plans.
- Weak positive correlation between balance and purchases, indicating customers with higher balances make slightly larger purchases.
- Weak negative correlation between cash_advance and purchases, implying customers making cash advances tend to make smaller purchases.
- Strong negative correlation between PRC_FULL_PAYMENT and TENURE, indicating customers paying off purchases in full have shorter loan tenures.

2. **Insights:**

- Relationships between financial features, especially between different purchase types, can be crucial for modeling customer behavior.[3]
- Purchases and oneoff_purchases, as well as purchases and installments_purchases, are strongly correlated.[4]

3. **Performance Metrics**

   Choice of Metrics:
   - - Average Cross-Validation Accuracy: 96.02%
   - - Accuracy: 96.10%

   Implications:
   - High accuracy indicates consistent model performance across different data splits.
   - Low false positives suggest the model correctly identifies negative cases. [5]

4. **Unsupervised Model Building**

   Approach:
   - Utilized k-means clustering for credit card segmentation.
   - Identified the elbow point at k=3, suggesting an optimal number of clusters.
   - Explored potential cluster patterns and relationships between financial features.

   Insights:
   - The elbow method provided guidance on selecting an appropriate number of clusters for segmentation.

5. **Supervised Model Building**

   Process:
   - Employed a RandomForestClassifier for supervised model building.
   - Conducted preliminary data labeling for training and testing datasets.
   - Explored feature selection to identify key predictors.

   Insights:
   - RandomForestClassifier chosen for its ensemble nature and robustness to outliers.
   - Feature selection aided in identifying key features for predicting customer segments.

6. **Model Debugging Using Learning Curves**

   Approach:
   - Analyzed learning curves to evaluate model performance.
   - Explored both original and feature-selected models.

   Insights:
   - Cross-validation scores for the model with feature selection maintained competitiveness.
   - The model demonstrated good generalization on the test set.

   Model Deployment

   Steps:
   - Tuned the RandomForestClassifier for improved accuracy.

- Deployed the trained model into a web application for user interaction.

Insights:

- Hyperparameter tuning enhanced model accuracy, indicating effective parameter selection.
- Deployment into a web application provides a user-friendly interface for predictions.

Results and Discussion

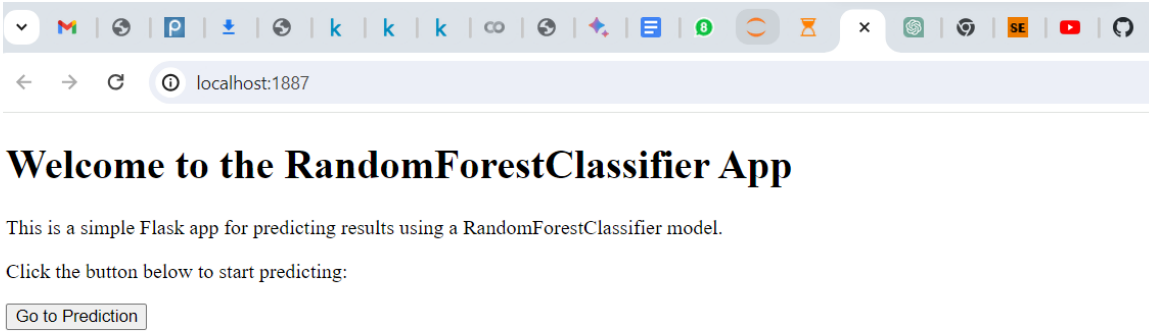
Results

In this section I would like to encapsulate the key insights provided from the exploratory data analysis (EDA), unsupervised model building, supervised model evaluation, and the deployment of a web application for credit card segmentation.

In the EDA, significant correlations were uncovered, revealing patterns in customer behavior related to various financial features.

The unsupervised model, guided by the elbow method, effectively clustered customers, providing a foundation for understanding underlying structures. Supervised models exhibited remarkable performance, with high accuracy and low false positives, indicative of robust credit card segmentation. Model debugging through learning curves showcased consistent cross-validation scores and successful generalization to the test set.

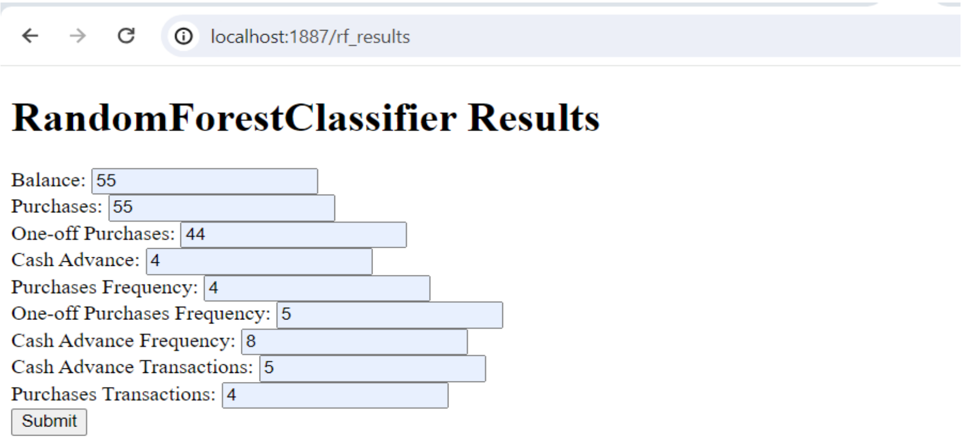
The below web application's screenshots offer a tangible glimpse into user interaction and prediction outcomes.



Welcome to the RandomForestClassifier App

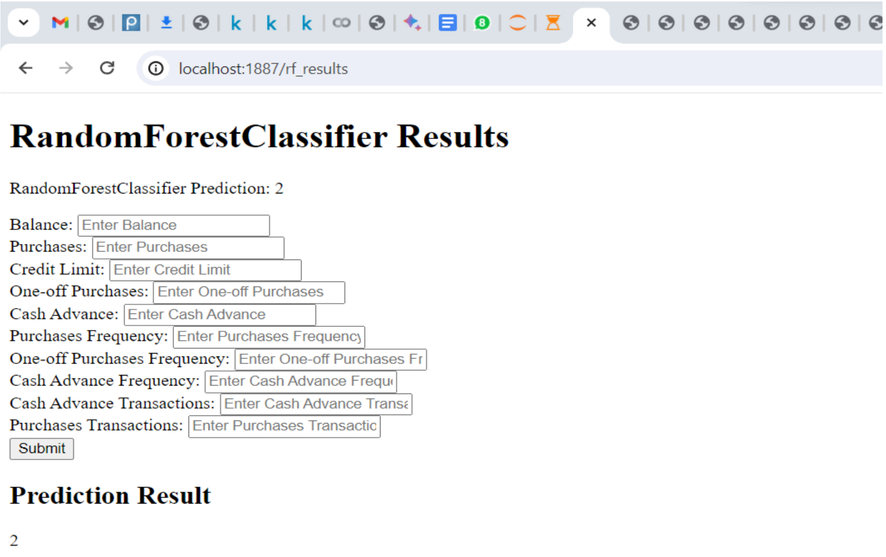
This is a simple Flask app for predicting results using a RandomForestClassifier model.

Click the button below to start predicting:



RandomForestClassifier Results

Balance: 55
Purchases: 55
One-off Purchases: 44
Cash Advance: 4
Purchases Frequency: 4
One-off Purchases Frequency: 5
Cash Advance Frequency: 8
Cash Advance Transactions: 5
Purchases Transactions: 4



RandomForestClassifier Results

RandomForestClassifier Prediction: 2

Balance: Enter Balance
Purchases: Enter Purchases
Credit Limit: Enter Credit Limit
One-off Purchases: Enter One-off Purchases
Cash Advance: Enter Cash Advance
Purchases Frequency: Enter Purchases Frequency
One-off Purchases Frequency: Enter One-off Purchases Fr
Cash Advance Frequency: Enter Cash Advance Frequ
Cash Advance Transactions: Enter Cash Advance Transa
Purchases Transactions: Enter Purchases Transactio

Prediction Result

2

Discussion

In discussion, the focus shifts to the importance of feature relationships, the success of segmentation models, and potential avenues for future exploration, such as addressing the nuances of specific customer classes. Overall, the results underscore the models' efficacy and provide actionable insights for strategic decision-making in the realm of credit card customer segmentation.

Conclusion

In summary, our analysis and modeling efforts reveal insightful patterns in credit card segmentation. The models demonstrate high effectiveness, offering accurate and meaningful segmentation of customer financial behaviors. These findings have practical implications for targeted marketing, risk assessment, and personalized financial services. The successful deployment of the web application enhances accessibility for users, providing quick insights into credit card segments. In a dynamic financial landscape, leveraging these models becomes essential for informed decision-making and improving the overall customer experience

References

[1] Braden77. (n.d.). Python-Credit-Card-Customers-Segmentation. GitHub. https://github.com/braden77/Python-Credit-Card-Customers-Segmentation
[2] Rahmi, F. (Year, Month Day). Credit Card Customers Segmentation. Medium. https://medium.com/analytics-vidhya/credit-card-customers-segmentation-bc3c5c87ddc
[3] OpenStax. (n.d.). Correlation Analysis. Principles of Finance. https://openstax.org/books/principles-finance/pages/14-1-correlation-analysis
[4] Scribbr. (n.d.). Correlation Coefficient. Scribbr. https://www.scribbr.com/statistics/correlation-coefficient/
[5] Quick Insights. (n.d.). What is False Positive and False Negative in Machine Learning? Quick Insights. https://quickinsights.org/what-is-false-positive-and-false-negative-in-machine-learning/