# Deployed Machine Learning Based Card Holder Segmentation System

*Author:*
Rahman MABANO
*(rmabano)*

## Abstract

This technical report delves into the analysis of credit card usage data, utilizing a combination of unsupervised and supervised machine learning techniques to achieve nuanced customer segmentation. The study begins with meticulous data preparation, followed by an extensive exploratory data analysis (EDA) to uncover underlying patterns and relationships. Employing a Random Forest classifier with carefully tuned hyperparameters, the report demonstrates how quality data is more instrumental than complex models in predictive accuracy. A comparative analysis of models, including learning curve assessments, reveals the intricacies of overfitting and the importance of feature selection. The culmination of this study is the deployment of the most effective model in a Flask-based web application, showcasing real-time customer segmentation. This report underscores the significance of methodical data analysis and model tuning in deriving actionable business insights, particularly in customer behaviour prediction and strategic decision-making in the financial sector.

## I. Background & Problem Description

Analysing credit card usage data is pivotal for financial institutions aiming to refine their customer engagement and risk management strategies [1]. The dataset at hand encapsulates a wealth of behavioural information for around 9000 active credit card users over a six-month period. It provides detailed insights into the balances, purchase behaviours, cash advances, payment regularity, and credit tenure among other key behavioural indicators. By dissecting this data, credit card issuers can segment customers effectively, personalize credit offerings, anticipate credit risks, and tailor customer-centric strategies to enhance loyalty and retention.

Through segmentation, the issuer can pinpoint distinct customer groups, enabling the crafting of targeted marketing campaigns and product offerings. For instance, specific segments may benefit from customized credit limits or tailored interest rates, driving both customer satisfaction and financial health. Additionally, patterns unearthed from the data can fortify risk assessment models, ensuring early identification of potential defaults, and facilitating proactive risk mitigation.

The analytical journey culminates in the development and comparison of two machine learning models – an initial model encompassing the full range of features and a refined model focusing on a subset of optimal features. Validation curves underscore the efficacy and robustness of these models, underpinning their potential to be deployed for real-time customer segment prediction. This transition from analysis to application signifies a strategic move towards leveraging data-driven insights for operational decision-making and offering a more personalized customer experience.

## II. Approach

Throughout this process, the choices made were driven by the goal of achieving a nuanced understanding of customer behaviour, leading to actionable insights for strategic decision-making. Each method was carefully selected to align with the dataset's characteristics and the business objectives at hand, ensuring that our approach was not only technically sound but also strategically focused. In addressing the credit card dataset, the main approach was methodical, encompassing several key stages of data-driven analysis and model development. Here's a succinct yet comprehensive description of the process:

- **Data Preparation**: Initiated with a thorough cleansing of the dataset, addressing any missing values and standardizing the data to ensure consistency and quality, thereby laying a solid foundation for the analytical modelling that followed. Missing values were handled using imputation, replacing missing values with mean values.
- **Exploratory Data Analysis (EDA):** An extensive EDA was conducted, leveraging both visual and statistical methodologies to identify patterns, outliers, and structural nuances within the data. This critical phase was instrumental in shaping the subsequent modelling direction and extracting meaningful insights. Outliers were eliminated with the log transformation method. Log transformation was a suitable initial approach due to the nature of the data and the goal of retaining as much information as possible.
- **Label Encoding and Scaling:** The dataset contains only one categorical column, CUST_ID, which is an identifier and not a feature for modelling. Hence, no label encoding is required for the dataset. All other numerical features where scaled.

- **Performance Metrics:** Key metrics such as accuracy, precision, recall, F1 score, and AUC-ROC were employed to provide a comprehensive assessment of the model's predictive performance, ensuring a balanced evaluation of both accuracy and error sensitivity. When comparing models, it provided sufficient conviction to evaluate models using validation curves as well.
- **Unsupervised Model Building**: A clustering model was developed using the K-Means algorithm, known for its effectiveness in distinguishing 2 distinct customer segments based on variables related to purchasing behaviour and transaction frequency, one lower than the other. Ultimately, the optimum k was found using methods such as the elbow method, and silhouette score analysis [2]. PCA clustering was also carried out to learn the demarcation between client types [3], which are explained in depth in the notebook. Two methods to evaluate the model were utilised to ascertain the performance (Silhouette and Calinski Harabasz Score)
- **Supervised Model Building**: Following the unsupervised phase, the dataset was labelled according to the derived clusters, which are 0 and 1 under a column called Clusters, which facilitated the training of supervised models, and was our target/prediction column. The Random Forest Classifier was selected for its proficiency in managing complex, high-dimensional data and its inherent resistance to overfitting.
- **Model Debugging Using Learning Curves:** The application of learning curves provided valuable insights into the model's learning trajectory, pinpointing instances of overfitting or underfitting and guiding the fine-tuning process to enhance model performance. A learning curve was used to analyse the first random forest model and helped us understand whether the model was overfitting or not. Valuation curces were also used when more random forest classifiers were constructed.
- **Model Deployment**: The culmination of the process was the deployment of the final benchmark model via the Flask web framework. This step transformed the model into an interactive web application, capable of delivering customer segment predictions in real-time based on new input data.

## III. Results and Discussion

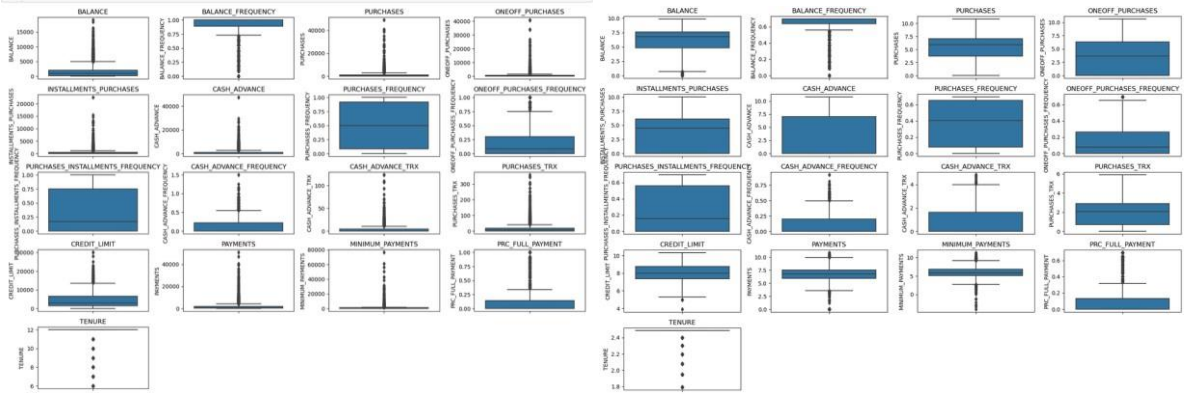
### 1. Exploratory Data Analysis (EDA) Results

**Missing values**

There was a total of 314 missing values, 313 belonging to the MINIMUM_PAYMENTS column and 1 the CREDIT_LIMIT column. All were replaced by mean values of those respective columns.

**Outliers**

Using log transformation, the outliers were eliminated, while still retaining essential dataset information. The images below show how the columns structure changes after getting rid of them.




### 2. Unsupervised Model Evaluation Results

For a K-Means model, an optimum k value was sought using the elbow method, then further evaluated using the silhouette score. A 2D PCA showing the 2 clusters/types of clients was then plotted, showing that most clients are in the second cluster, with higher and more lucrative spending patterns. To evaluate the K-means Model, we used the Silhouette score as well as the Calinski Harabasz Score. For both, it gave a high score, meaning that it can easily distinguish between client clusters, as follows:

Silhouette Score: 0.45611435050073795

Calinski Harabasz Score: 8141.7078231420755

The 2 clusters represent types of clients. One was named ‘Premium Client’ while the other was named 'Regular Client’ as will be seen in the web deployment result section.

### 3. Evaluation Results For all Supervised Learning Models.

For several reasons ranging from its ability to handle nonlinear relationships, robustness, versatility and performance, a Random Classifier Model was chosen for this section. Several modifications were made to model to find the best model as follows:

- **Initial Model**

Considering all features in the dataset, and not altering any of the hyperparameters, and initial model is constructed. It had the following classification report, after a 10-fold cross validation test:

```
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.95      0.97       609
           1       0.98      0.99      0.98      1181

    accuracy                           0.98      1790
   macro avg       0.98      0.97      0.97      1790
weighted avg       0.98      0.98      0.98      1790
```

- **Model with Selected Features**

Using Random Forrest feature importance in build function, 6 of the most important features were used to make predictions, instead of the whole dataset. The features are [PURCHASES_FREQUENCY, PURCHASES_TRX, PURCHASES, CASH_ADVANCE, CASH_ADVANCE_FREQUENCY, CASH_ADVANCE_TRX. The classification report of this new model came us as follows:

```
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.96      0.95       609
           1       0.98      0.97      0.98      1181

    accuracy                           0.97      1790
   macro avg       0.96      0.97      0.96      1790
weighted avg       0.97      0.97      0.97      1790
```

- **Model with Tuned Hyperparameters**

The model with selected features was then used to tune certain Random Forrest Hyperparameters as shown below:

```
RandomForestClassifier(max_depth=10, min_samples_leaf=2, min_samples_split=5,
                       n_estimators=50, random_state=42)
```

As a result, a much better model was built than the one with the 6 selected features (which had become the benchmark model) [4] as shown in the classification report shown below:

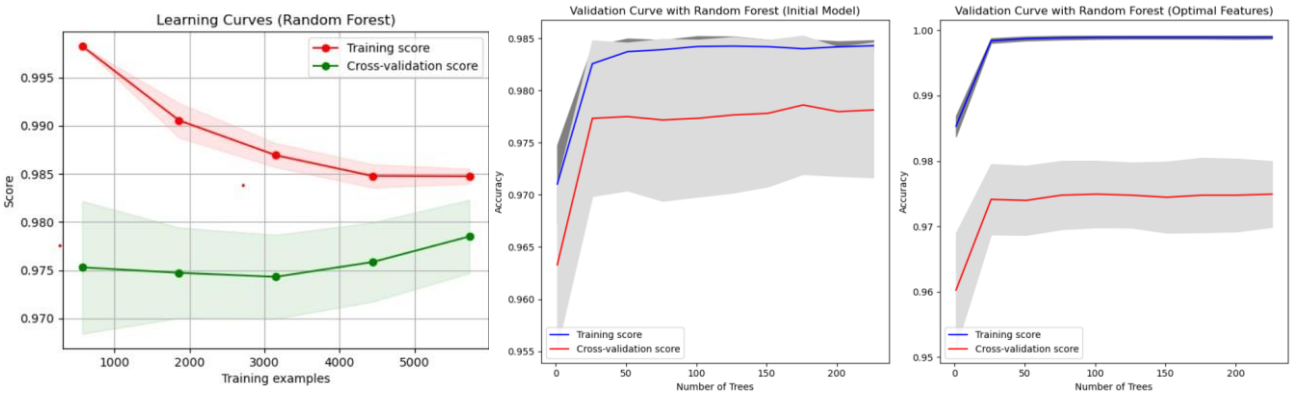
```
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.95      0.96       609
           1       0.98      0.98      0.98      1181

    accuracy                           0.97      1790
   macro avg       0.97      0.97      0.97      1790
weighted avg       0.97      0.97      0.97      1790
```

## 4. Results of model Debugging

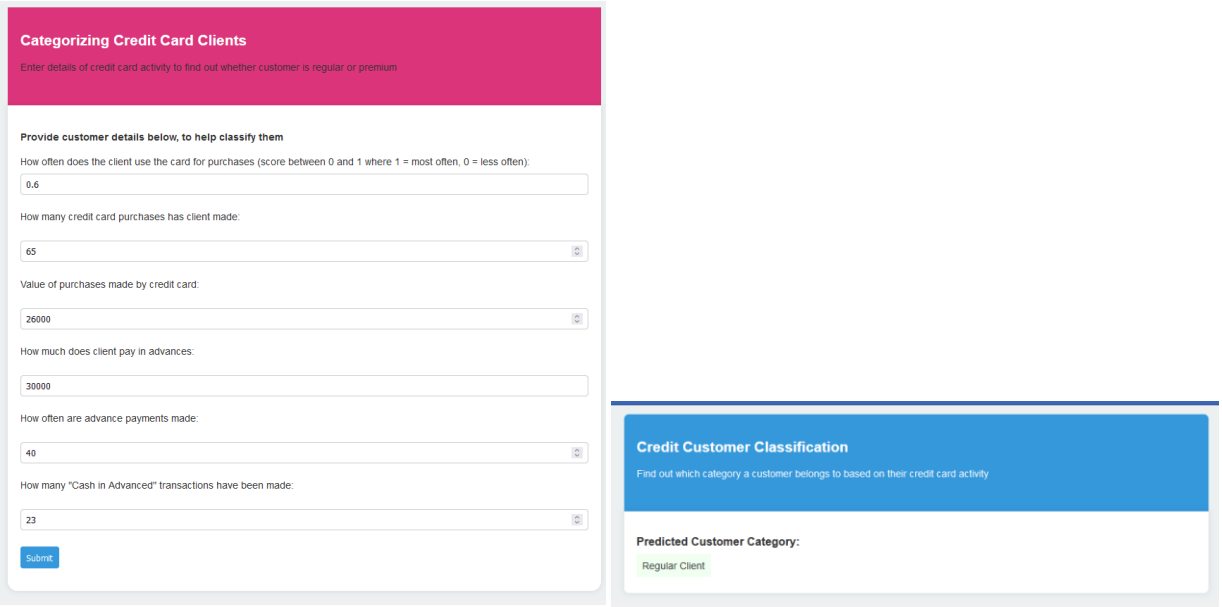
After the first model is built and evaluated using a learning curve to check for overfitting/underfitting as shown below, and then also compared to the model with only chosen features as shown below:




the learning and validation curves helped inficated the progress of the training and testing datasets, to ensure that no mistakes where made during fitting the models that may result in overfitting/underfitting.

5. **Web Application**

With the help of Flask library and HTML frameworks, an interactive web application was constructed as shown in the snapshots below [5]:




## IV. Conclusion

Supervised learning methods offer precise metrics to evaluate model performance, essential for fine-tuning applications. In this study, the Random Forest model with tuned hyperparameters emerged as superior, demonstrating effectiveness upon deployment in a web application. Quality data proved paramount, as evidenced by the slight performance drop across supervised classification models, likely due to the dataset's high quality and minimal missing data. Initial model accuracy suggested potential overfitting, a hypothesis supported by learning curve analysis. Conversely, feature engineering highlighted the importance of feature selection in preventing overfitting. While the unsupervised model excelled in cluster creation, enhancing its performance further presents certain challenges.

**References**


1) Shefrin, H., & Nicols, C. M. (2014). Credit card behaviour, financial styles, and heuristics. *Journal of Business Research*, *67*(8), 1679-1687.
2) Vlachos, Andreas. "Evaluating unsupervised learning for natural language processing tasks." *Proceedings of the First workshop on Unsupervised Learning in NLP*. 2011.
3) Faradayan, Mohammad, Faramarz Safi-Esfahani, and Zahra Beheshti. "Combining hierarchical clustering approaches using the PCA method." *Expert Systems with Applications* 137 (2019): 1-10.
4) Malakar, Preeti, et al. "Benchmarking machine learning methods for performance modelling of scientific applications." *2018 IEEE/ACM Performance Modelling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)*. IEEE, 2018.
5) Singh, Pramod. "Deploy Machine Learning Models to Production." *Cham, Switzerland: Springer* (2021).