Carnegie
Mellon
University
Africa

COURSE NAME: Programming for Data Analytics

COURSE CODE: 04638-A

INSTRUCTOR: Dr. George Okeyo

Assignment Title: Customer segmentation and classification using Machine Learning

Report title: Credit card customer segmentation and prediction

Done by: Leonard Niyitegeka

Andrew Id: lniyiteg

Submission date: 17 December 2023

1. Abstract

Credit card issuers face constant challenges in tailoring marketing and risk management strategies to diverse customer profiles. With traditional systems falling short, this study aims to develop a dynamic and accurate approach for customer segmentation and group prediction. We leverage K-means clustering to segment credit card customers based on their financial and transactional pre-processed data. Subsequently, we utilize Random Forest with cross-validation to predict the segment membership of new customers. The dataset used consists of 8950 credit card users and 18 variables. K-means clustering identified two distinct segments: active and inactive customers. Feature selection identified only seven essential features for prediction. Random Forest cross-validation achieved an F1 score of 98.7% which improved a bit after feature selection. These techniques provide a reliable and efficient method for credit card customer segmentation and group prediction. This offers valuable insights for targeted marketing and risk management strategies, tailored to specific customer profiles.

2. Background and problem description

Credit card issuers encounter ongoing challenges in customizing marketing and risk management strategies for diverse customer profiles[1]. Traditional credit card customer segmentation methods face challenges in scalability and adapting to evolving behaviours[2]. These approaches often struggle to handle large datasets efficiently and lack flexibility in capturing dynamic trends. As the number of credit card users grows and consumer behaviours change rapidly, traditional methods become less effective. To overcome these limitations, there is a need for more advanced and dynamic segmentation techniques, such as machine learning algorithms, which can provide nuanced insights, handle complex data, and adapt to evolving customer behaviours in real-time[3] . This study adopts a data-driven approach, utilizing methods such as K-means clustering and Random Forest, to enhance customer segmentation and predict group membership, offering valuable insights for targeted strategies in the credit card industry.

3. Approach

As it was mentioned above the purpose of this study was to perform clustering on credit card customers data and then the classification by using the best model with cross validation. As it obvious, the first step is to obtain the data. We read the data and then we checked the data to perform cursory examination and then we prepared the data for analysis. The data was originally found to be made up of 18 behavioural variables and 8950 entries of the data of credit card users behaviours. And there were some missing values in the dataset , 1 missing value in CREDIT_LIMIT and 313 missing values under MINIMUM PAYMENTS. The one missing value in credit limit was dropped and those in minimum payments was handled by filling them with the median of the values in the column due to the reason that the data in that column was positively skewness it was also one of the attempt to reduce the skewness. There were also outliers in the dataset. The only variables which was found not to have outliers were PURCHASES_FREQUENCY and PURCHASES_INSTALLMENTS_FREQUENCY. We also checked the correlation between the features and we found that there were some correlated features which was letter handled by using dimensionality reduction techniques. While we were checking for the distribution of the data in each column, we found that there were a lot of skewness in the data most of them were skewed on right hand side and we used log transformation to kind of reduce the skewness in the data. Next, we performed standardization by applying standard scaler on the data. Standard scaling is applied to credit card customer data to normalize features, ensuring equal contribution from each variable and enhancing the performance of machine learning algorithms like K-means clustering.

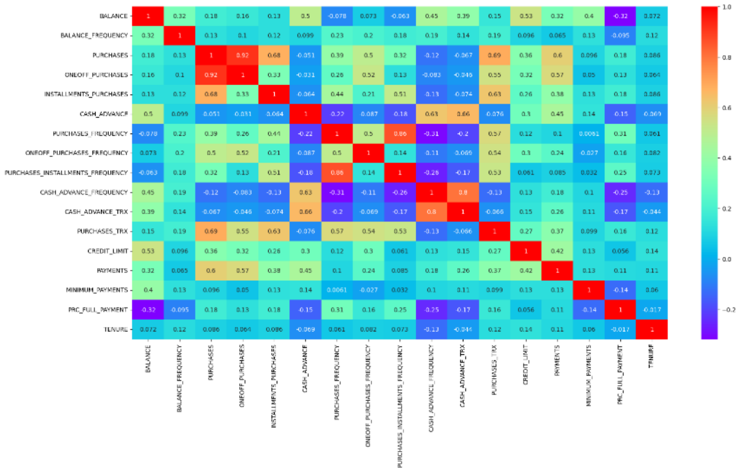
Next, we used the scaled data to perform clustering using K-means. Actually we tried different clustering methods but we decided to continue with K-means because it was the one which was giving the better silhouette score of about 0.517 and few optimal number of clusters which result in highest silhouette score. We first found the optimal number of cluster which was giving the highest silhouette score and we used that number of clusters to perform clustering which resulted in only 2 clusters which interpreted as active and inactive Credit card users. The after we used Random forest classifier with cross validation to predict the status of a customer whether in active customer group or in inactive customer group and then classification performance was assessed using F1 score, recall score and precision score. We used the above classification performance metrics because we had imbalanced data and those metrics are highly used for imbalanced data. Later, we created learning curves by using the learning curve function and then we saved the classification model (random forest classification) for later usage.

Next, we performed feature selection and dimensionality reduction by using random forest feature importance[4]. Only 7 features which had a significantly higher importance were selected and then they were used alone to build the same model, A random forest with cross-validation and the same performance metrics was used to check the performance of the model after feature selection. We then did hyper parameter tuning with using the model with only selected features because it was the one resulting high performance. We tuned that benchmark model by using GridSearchCV and set hyperparameters and then we saved the tuned model. The tuned model performance was also evaluated by using the performance metrics used earlier.

The last step in this study is model deployment. Model deployment is a critical step in making machine learning models accessible and usable and it allows the stakeholders to interact with the model without needing to understand the underlying complexity[5]. We built a Flask application that serves as a simple web-based interface for deploying a machine learning model. We developed html files both for the data value entry screen display (where the user enters the values for variables) and the other for prediction result display page. Note that the we used the model with the selected features because it was the one resulting in higher performance.

4. RESULTS AND DISCUSSION

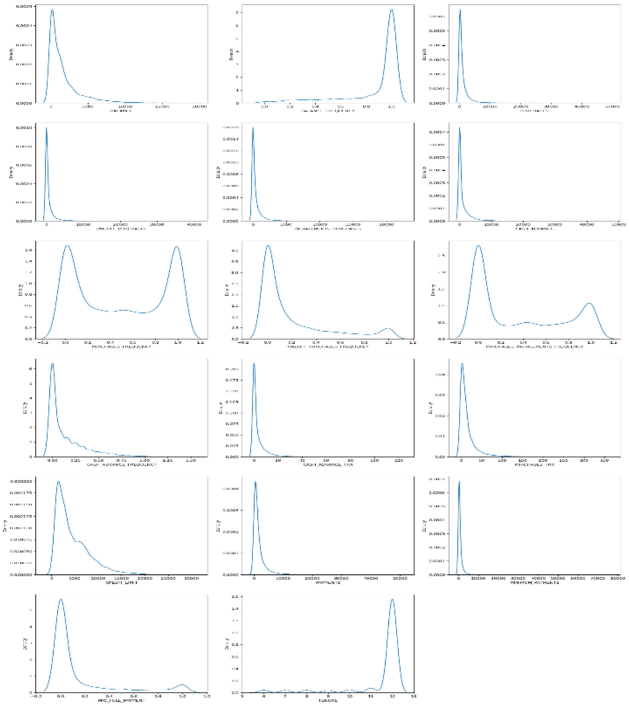
EDA Results: The dataset was found to possess missing values in 2 columns. As explained above missing values in one column (CREDIT_LIMIT) was handled by dropping it and missing values in MINIMUM_PAYMENTS was handled by filling with median due to the reasons communicated above. There were no duplicate data entry in the data. There were a lot of outliers, most of columns(variables) had outliers except 2 columns only.



Correlation :

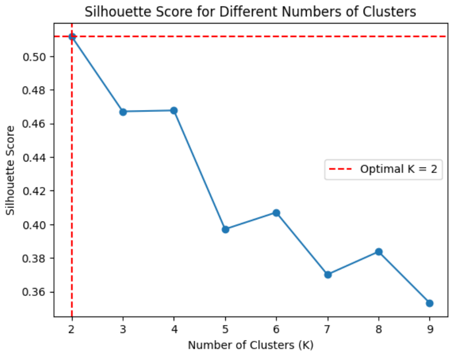
some data are highly correlated between them. This in turns, decreases the performance of the model in prediction thus this was eliminated by performing feature selection.

Distribution of variables' data



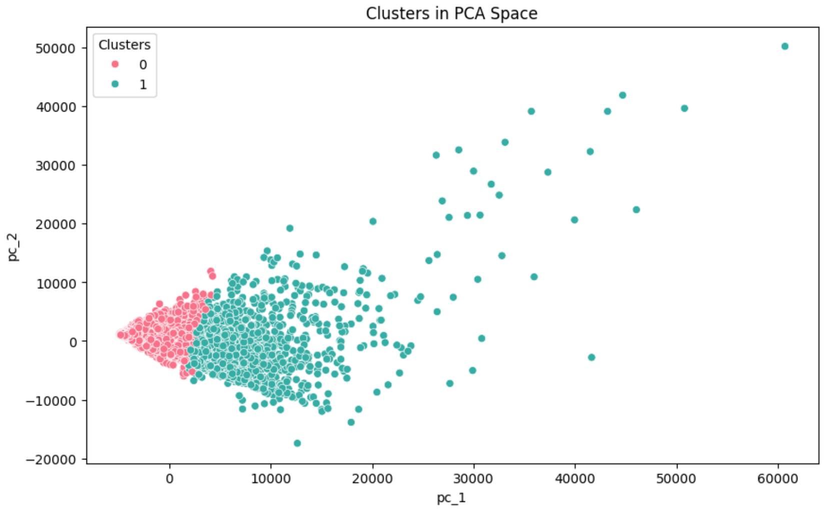
As it shown in the graph on the left side the data has a lot of skewness in different columns. The prevalent skewness type is the skewness on right (positively skewed).

<u>unsupervised learning results:</u>



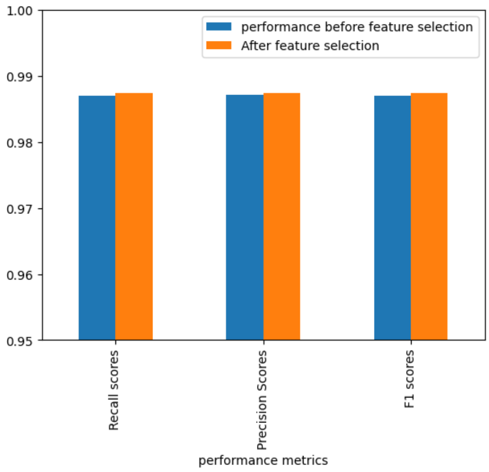
The optimal number of clusters (the clusters which results in highest silhouette score) is 2 . performance for unsupervised was evaluated by using silhouette score, and the evaluation result was found to be silhouette score of 0.517. This silhouette is not a bad score because it positive and above 0.5 and for silhouette score the one which is close to 1 the better thus the obtained silhouette score is good. Note that I have performed a try and error sequence to find the highest silhouette score this 0.517 is the one which was found to be the best.

as shown in the graph above so we performed clustering by using the the optimal number of clusters in PCA space and we obtained the following.



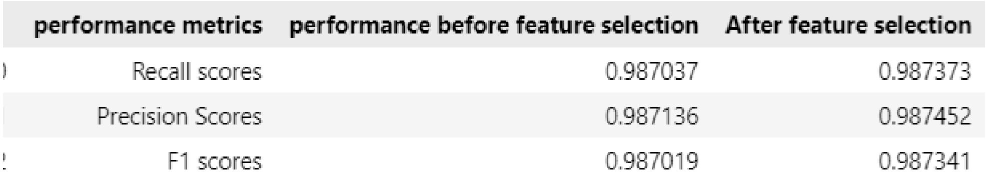
The graph on left side shows the clusters representation in PCA space using scatter plots. It shows the values in cluster 0 which we named unactive customers are very close to each other and their have lower first principal components. But cluster 1 (active) in green color some elements of the clusters are scattered away form the others (we can call them outliers) those who use the credit card most frequently than others.

Results for supervised learning evaluation



The graph on the left side shows the performance of random forest model with cross validation in prediction of the group of a customer in which customers fall in. It includes the performance of the model both before feature selection and after feature selection. The graph shows that the performance of model in all cases , f1 score, recall score and precision score, has improved after features selection, that is the performance of the model after feature selection is greater than the performance of the model before feature selection.

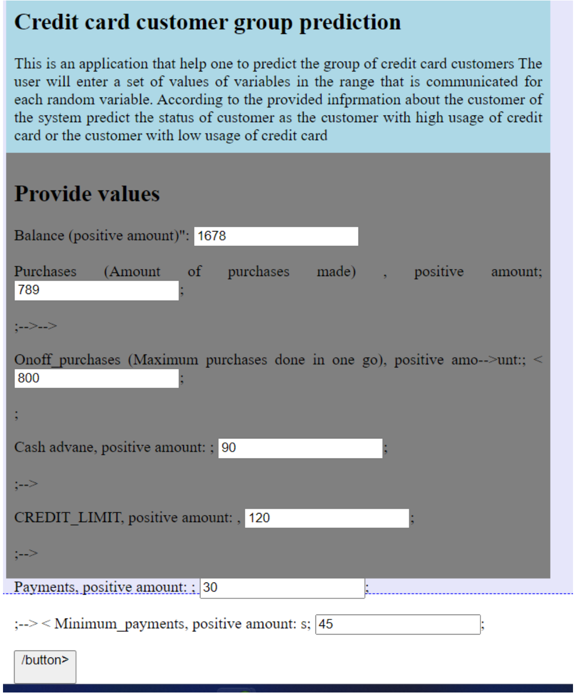
The model performance before and after feature selection is given below as



performance metrics	performance before feature selection	After feature selection
Recall scores	0.987037	0.987373
Precision Scores	0.987136	0.987452
F1 scores	0.987019	0.987341

Another important observation that we made is that hyperparameter tuning lead to a slight improvement of the model's performance as it can be observed on the graph below

Model deployment: model was deployed by using building a flask application and using html codes for interface.



Credit card customer group prediction

This is an application that help one to predict the group of credit card customers The user will enter a set of values of variables in the range that is communicated for each random variable. According to the provided infprmation about the customer of the system predict the status of customer as the customer with high usage of credit card or the customer with low usage of credit card

Provide values

Balance (positive amount)": 1678

Purchases (Amount of purchases made) , positive amount; 789 ;

;-->-->

Onoff_purchases (Maximum purchases done in one go), positive amo-->unt:; < 800 ;

;

Cash advane, positive amount: ; 90 ;

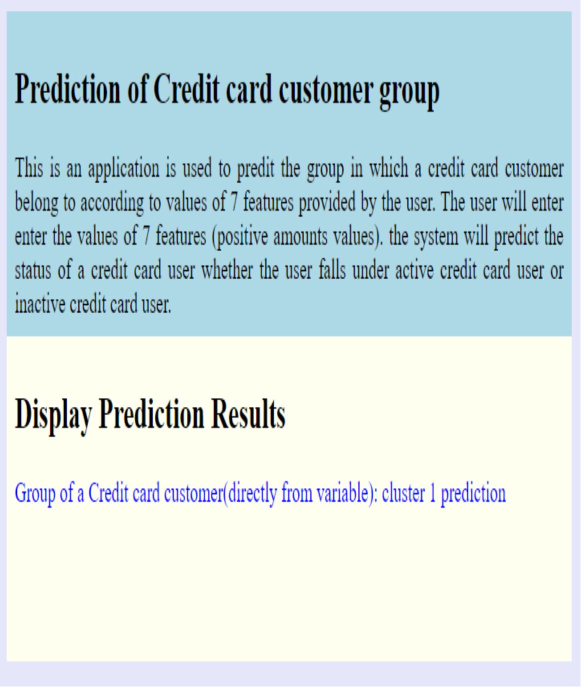
;-->

CREDIT_LIMIT, positive amount: , 120 ;

;-->

Payments, positive amount: ; 30 ;

;--> < Minimum_payments, positive amount: s; 45 ;



Prediction of Credit card customer group

This is an application is used to predit the group in which a credit card customer belong to according to values of 7 features provided by the user. The user will enter enter the values of 7 features (positive amounts values). the system will predict the status of a credit card user whether the user falls under active credit card user or inactive credit card user.

Display Prediction Results

Group of a Credit card customer(directly from variable): cluster 1 prediction

REFERENCES

[1] ‘What & Why Credit Risk Management for SMEs | HighRadius’. Accessed: Dec. 17, 2023. [Online]. Available: https://www.highradius.com/resources/Blog/what-why-credit-risk-management-sme/

[2] ‘(PDF) A review on customer segmentation methods for personalized customer targeting in e-commerce use cases’. Accessed: Dec. 17, 2023. [Online]. Available: https://www.researchgate.net/publication/371447632_A_review_on_customer_segmentation_methods_for_personalized_customer_targeting_in_e-commerce_use_cases

[3] A. Banduni, ‘CUSTOMER SEGMENTATION USING MACHINE LEARNING’.

[4] ‘Feature importances with a forest of trees’, scikit-learn. Accessed: Dec. 17, 2023. [Online]. Available: https://scikit-learn/stable/auto_examples/ensemble/plot_forest_importances.html

[5] ‘What is Model Deployment | Iguazio’. Accessed: Dec. 17, 2023. [Online]. Available: https://www.iguazio.com/glossary/model-deployment/