Table of Contents

Introduction

The 2008 Global Financial Crisis highlighted the significance of credit risk faced by major financial institutions. Assessing the credit worthiness of individual borrowers and businesses are crucial to maintain a stable balance in the loan book to reduce the accumulation of bad debt. In particular, the advent of peer-to-peer (P2P) lending platforms represents a new frontier in financial technology, disrupting the lending market. It provides an online platform to directly connect potential borrowers and investors, without a central financial intermediary (i.e. bank). This empowered investors to have easy access to lend money to borrowers or small businesses beyond the traditional financial institutions. Therefore, credit risk assessment and management is critical to attract high quality customers through the analysis of big data.

The paper focuses on Lending Club, the first P2P lender to have registered its offerings in the United States. The company enables investors to independently evaluate the credit risk of listed loans based on the borrowers' profiles. These come in various grade schemes and interest rates, where a higher interest tends to lead to a poorer credit grade. Using the publicly available dataset (George, 2019), machine learning models and classifiers are built to identify characteristics common amongst groups of borrowers. These findings will help financial firms anticipate the possibility of bad loans and provide more accurate credit profiling for good customers to enhance borrower satisfaction and trust.

Literature Review

This section highlights prior literature in adopting machine learning models to improve the loan prediction processes in minimising potential default rates. Zhang (2023) utilised the Random Forest classification algorithm to classify loans, attaining a prediction accuracy of over 85%. Suhaolnik et al (2023) similarly outlined a series of machine learning algorithms and utilised the logistic regression, decision tree and XGBoost models. The study concluded that XGBoost was a relative outperformer in accurately predicting loan default rates. However, these research studies utilised existing features in the dataset to train the models. As such, this paper aims to iterate upon previous literature to introduce customer profiling to segment the different borrowers and create new features for analysis.

Data

Based on the publicly available Lending Club data, this study examines the loans made between 2017 and 2018. There are a total of 2,260,701 loan records consisting of the list of features for each loan created. Before commencing on the analysis, data cleaning was performed to curate a more robust and reliable dataset.

Feature Preprocessing

1. **Removing rows and columns with NaN values**

Columns with more than 30% of null values will be removed as the significant number of missing values can introduce bias and affect the accuracy of statistical analyses and machine learning models. As such, this enables a more complete and representative dataset for analysis. Rows with missing loan amounts were also dropped as it will be challenging to make accurate predictions.

2. **Conditions for Treating NaN values**

For the remaining features with missing values, they were categorised into 3 cases:

a. **Mean Value:** For these features with missing values, the average was taken for the column. This includes the Debt to income ratio (DTI), delinquency amount, mortgage amounts as the missing data is likely to be random and not systematically related to any specific patterns.
b. **Median Value**: The median was used to fill in the missing values for columns including annual income and employment length. These employment information are likely to have a skewed distribution and thus, the median will be robust to outliers.
c. **'-' / Zero Value:** The missing values in columns such as number of recorded bankruptcies were replaced with 0. This signifies that no bankruptcies were recorded. Similarly, for other categorical features, the missing values were all replaced with 0 since they did not have a specific meaning in the data.

3. **One hot representations**

Categorical features were then replaced with their one-hot representations into a numerical format for the machine learning models to utilise these features effectively.

4. **Updating Loan status**

To create binary variables on the loan status where we are interested in whether the user successfully paid off or defaulted, I treated "Fully Paid" as a positive label, and "Charged Off", "Late", "Grace" as the negative label. Other values such as "Current" were removed from the sample as the statuses of these loans have not been finalised, and were thus, irrelevant.

5. **Normalisation**

All columns with numerical data were then normalised to have a similar scale to improve the performance of the subsequent machine learning models. This is due to the wider scales for certain features which may dominate and thus exert a disproportionate influence on the models.

6. **Imbalanced Data Set**

The dataset now contains 1,382,351 loan records, 78.04% of them are Fully Paid (positive labels) while the remaining 21.96% are Defaults (negative labels). This presents a highly skewed class proportion where the majority class are good loans. Yet, the objective of this study is to analyse default loans and provide predictive analytics in identifying these risky loans. Given the large amount of data, downsampling of profiles with good loans was performed to reduce information loss from the minority class of risky loans. As a result, both types of loan profiles were equally represented with 50% for each label. The data was then split in the ratio of 70 : 30 for the training and test data set respectively for the subsequent machine learning applications.

Types of Classification Methods Used

In this section, the paper delves into the application of machine learning techniques for the classification of loans as either 'good' or 'bad'. Specifically, K means clustering was used to profile borrowers into different customer segments while the baseline classifier: Logistic Regression, Random Forest Classifier and XGBoost were used to determine the effectiveness of default predictions.

K Means Clustering

K Means clustering is a form of cluster analysis which partitions a dataset into K clusters where the sum of the squared Euclidean distances between each data point and their assigned cluster mean (i.e. centroid) is minimised (Ikotun et al, 2023). Every point within an identified cluster will be similar to each other, but different from other neighbouring clusters.

To determine the optimal number of clusters, represented by the value of K, the elbow method is used. It involves running the K-means algorithm multiple times with a different K to determine the total variance (Humaira & Rasyidah, 2018). The results are then plotted on a graph - x axis represents the number of clusters and the y axis represents the total variance for each cluster. As K increases, the total variance should gradually decrease as the points are increasingly closer to their respective centroids. Thus, the elbow point represents the ideal K, where adding more clusters does not significantly reduce the total variance and provide value to the model.

Logistic Regression

The logistic regression model is a supervised machine learning algorithm for classification. It can adopt different forms of classifiers which are trained on a set of input data to make a binary decision (Jurafsky & Martin, 2024). In this paper, the sigmoid classifier will be introduced. As such, the result suggests the probability that the borrower belongs to a particular category based on the set of dependent variables.

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{1n}}}$$

P calculates the probability of the borrower's repayment of the loan, while the X variables refer to the features in the dataset that describes the borrower's profile. The value of P ranges from 0 to 1, where 1 suggests a perfect credit status, with no probability of default.

Random Forest Classifier

The random forest classification is an ensemble model that combines multiple individual trees while randomly varying their design (Biau, 2012). This utilises a top-down, greedy approach through recursive binary splitting as it picks the best feature-threshold combination that results in the optimal objective function. The classifier also employs an array of hyperparameters for tuning the training result to address potential issues from overfitting.

Parameters	**Range of Values**
max_depth	2, 5
min_samples_leaf	5, 10
n_estimators	5, 10

XGBoost Classifier

XGBoost leverages on the existing gradient boosting algorithms which group several decision trees together to form a strong classifier. It primarily trains the base learners to learn the negative radiant of the current loss function of the ensemble. As a result, each addition to the ensemble directly reduces overall training error based on the errors made by prior ensemble members.

Parameters	**Range of Values**
num_class	3, 5
max_depth	3, 5
learning_rate	0.01, 0.1
subsample	0.5, 1.0
colsample_bytree	0.5, 1.0
n_estimators	10, 50, 100

For these classifiers, the Gridsearch function was used for hyperparameter tuning.

Customer Segmentation

Qualitative Assessment

a. Socio Economic Class

The dataset has also provided information on the socio-economic class of each borrower which are valuable in understanding the type of customer through a single comprehensive metric. These include:

- **FICO scores (high and low ranges):** The scores highlight the borrower's creditworthiness and risk of default through these two ranges. A higher score has historically shown to have a lower rate of default based on the average of the high and low FICO scores for each borrower.

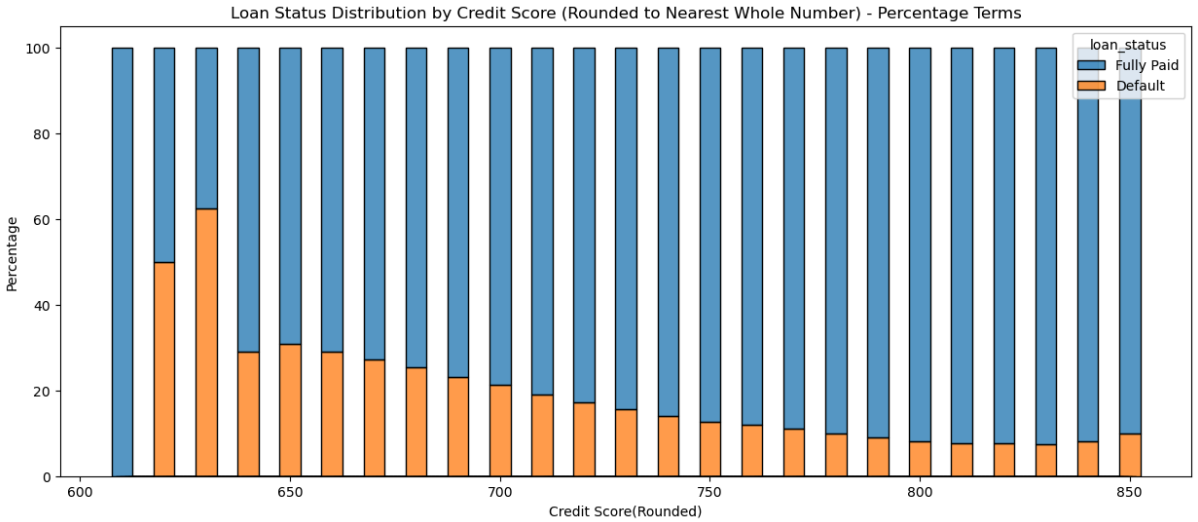


Fig 1: Bar chart on the distribution of loan status by the nearest rounded credit score (tens)

- **Employment Indicator:** The employment length indicates a borrower's job stability and potentially financial stability. Longer employment duration is often correlated with a lower risk of unexpected financial distress, making it a valuable component of the socio-economic profile.
- **Home Ownership:** The home ownership feature signifies a borrower's wealth status as it highlights if they are either on mortgage, renting their home or having complete ownership. In particular, a borrower under mortgage suggests a potentially high credit utilisation which is a potential tell tale indicator on their credit status.

Employment Indicator **Home Ownership**

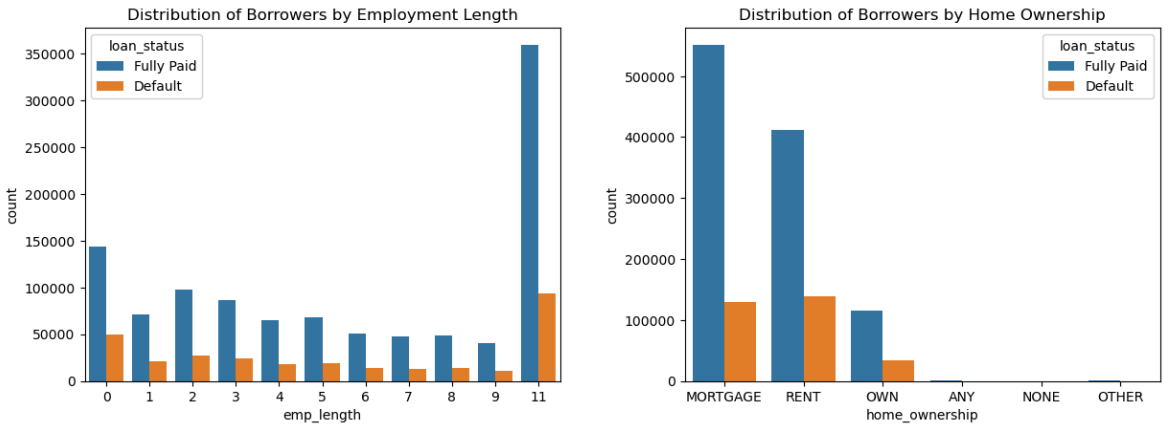


Fig 2: Bar chart on the distribution of loan status by employment length and home ownership

The K-means algorithm was applied to segment the data into k groups of equal variance. Utilising the K-means++ initialization, the borrowers were clustered into individual segments based on the averages recorded for each socio-economic feature in the cluster.

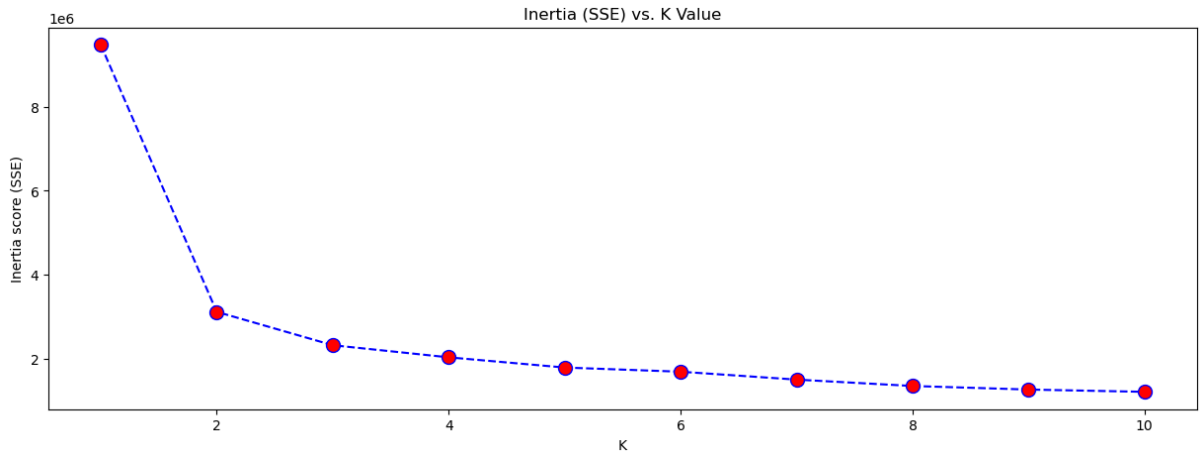


Fig 3: Elbow Joint plot of the clusters for socio-economic labels

Cluster ID	**Customer Profile**
0 (Low Risk)	Borrowers exhibit a relatively higher score that the other segments with a longer employment length. They have moderate home ownership rates and likely represent financially stable individuals with established credit histories and stable employment
1	Borrowers have a lower FICO score and shorter employment length relative to cluster 0. This might include individuals with relatively stable finances but shorter job tenures.
2 (High Risk)	Borrowers exhibit the lowest FICO scores and employment length amongst the clusters. However, it has the highest proportion of home ownership which means that these individuals could be in less stable financial situations, but with some wealth in their homes.

.

Table 1: Interpretation of Customer Profile by socio-economic clusters (Results in Appendix)

b. Loan Attributes

Similarly, clusters were formed based on the loan attributes to understand the borrowing behaviours and potential risk profiles of each borrower. The features involved include:

- **Loan amount:** Higher loan amounts recorded may indicate a stronger credit profile and income stability for the borrower to be eligible. However, it also provides a potential risk factor from default given the larger sum involved.
- **Term structure:** The loan duration indicates the borrower's ability to commit to the financial obligations. A longer duration poses higher risks for lenders due to the uncertainties over extended periods.
- **Loan grades:** LendingClub has classified each loan into different grades based on the borrower's risk profile. These loans were graded from A to G with sub categories within each grade. The higher the grade, the higher the risk the loan poses to the investor. This aligns with the relatively higher rates of default in Figure 4, where customers with "C" scores and below were more likely to default on their loans.
- **Interest rate:** The rates provided are also a key indicator of credit risk. Higher rates are typically associated with higher risk borrowers and loans. This can be seen from Figure 4 where default rates are empirically higher as the interest rates increase.

Grade of Loan **Interest Rate (Rounded)**

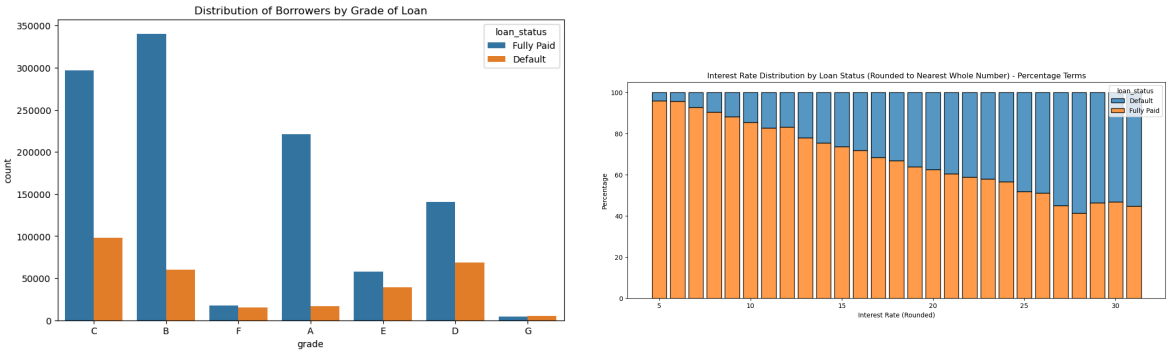


Fig 4: Bar chart on the distribution of loan status by grade of loan and interest rates

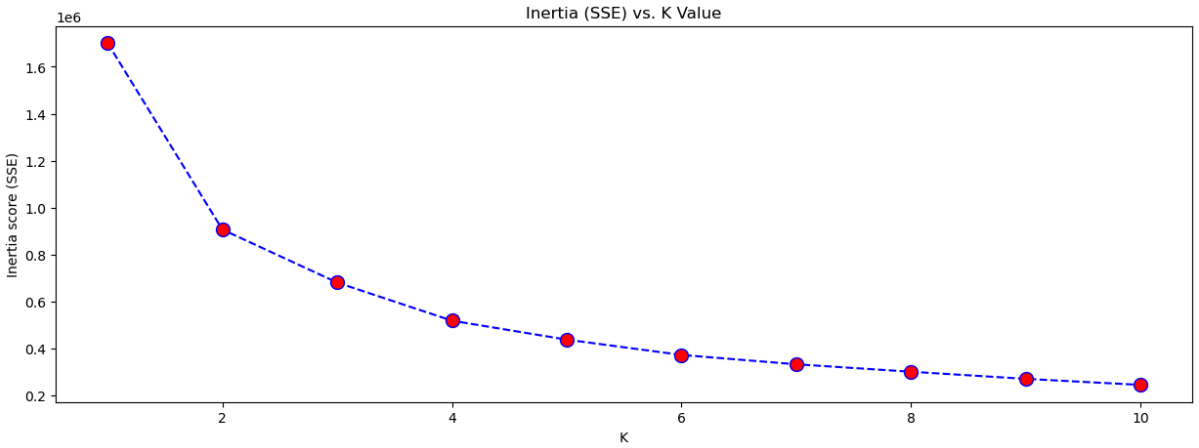


Fig 5: Elbow Joint plot of the clusters for loan attributes

Cluster ID	Customer Profile
0 (Low Risk)	Cluster 0 represents borrowers who take smaller than average loans at a slightly higher interest rate. The term of their loans is average, and the grade is the lowest among the three clusters, which suggests these might be the safest loans.
1	Cluster 1 likely represents the most creditworthy borrowers who take slightly less than average loan amounts, enjoy the lowest interest rates, prefer shorter terms, indicating low to moderate risk.
2 (High Risk)	Cluster 2 seems to consist of borrowers who take out the largest loans at the highest interest rates, with longer terms, and have the lowest grade (*since higher number represents a lower grade)*, suggesting these loans are the riskiest.

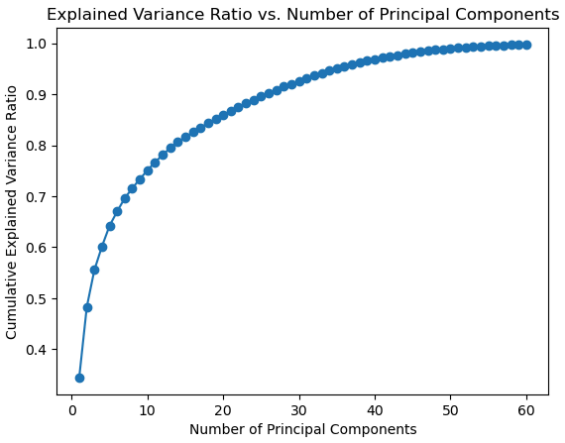
Table 2: Interpretation of Customer Profile by loan attribute clusters (Results in Appendix)

After the model has been fitted on the training data, the coordinates of the cluster centroid are saved and predicted on the test data. This will provide the additional features of ‘cluster_ids_socio_economic’ and ‘cluster_ids_loan_profile’ in both train and test data.

Principal Component Analysis

Given the large amount of data with 79 features, feature selection was applied to reduce the size of the dataset for analysis using Principal Component Analysis (PCA). PCA is a dimensionality reduction method that transforms the existing features into a new, smaller set while minimising the loss of information. Introduced in by Karl Pearson (1901), the algorithm finds new features that reflect directions of maximal variance in the data while being mutually uncorrelated. It uses the linear combinations of the existing features and projects them into the principal component space. For this paper, I adopted the sklearn.decomposition.PCA implementation and used a threshold of 95%. This means that the top components were selected till they are able to explain up to 95% of the variance of the data. A total of 37 features were then extracted, significantly reducing the size of the dataframe from an original 79 features.

Explained Variance Ratio vs Number of Principal Components



Feature Importance

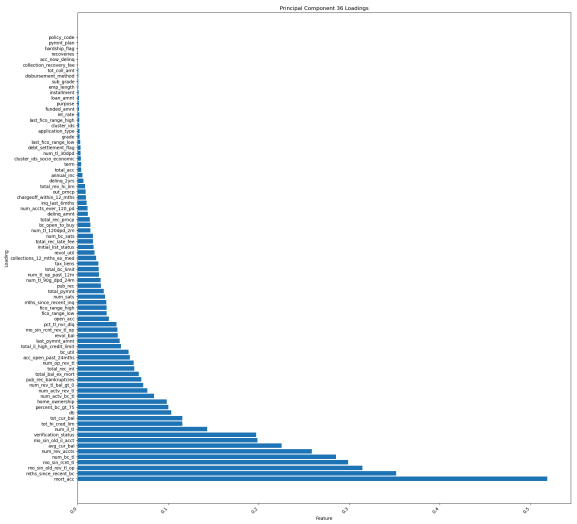


Fig 6: Principal Components and Relative Feature Importance

K Means Clustering on Overall Data

Using the 37 most important features, the K-means algorithm was executed on the train data set to determine the number of clusters to generate. This was done using the elbow plot method where the data was iterated through to generate clusters for different numbers of k from 2 to 10. This then returns the sum of squared distances of each data point from the centroid of its assigned cluster - inertia score, and plotted against the k value. The elbow point in the plot represents the k value where the reduction in inertia by increasing k is negligible and thus reached an optimal point.

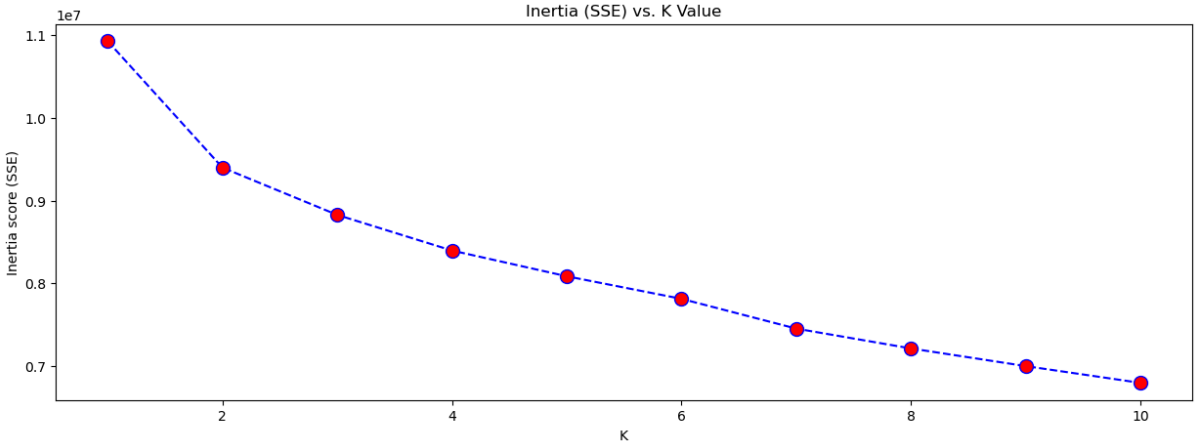


Fig 7: Elbow Joint plot of the clusters for overall dataset

In this case, the optimal number of clusters is 3. To help with cluster profiling, the pandas describe() method was used to determine the average for each feature per cluster. This new cluster ID was introduced as a new feature to the dataset.

Cluster ID	Customer Profile
0 (Low Risk)	This profile suggests financial savviness and stability, with a disciplined approach to managing a complex array of credit accounts. It can be seen from the high activity in revolving credit accounts, established credit histories, higher total credit limits and lower than average values in their credit delinquencies.
1	This cluster represents individuals with potentially lower credit engagement or newer credit history, indicated by fewer accounts, lower balances, and limits, but more recent bankcard openings and inquiries, suggesting they might be in the early stages of building or rebuilding their credit.
2 (High Risk)	This cluster appears to consist of individuals with more recent credit activity, including opening new accounts and inquiries, but also higher delinquencies, bankruptcies, and accounts ever past due, indicating a riskier credit profile with recent and potentially aggressive credit seeking behaviour or financial distress.

Table 3: Interpretation of Customer Profile by Overall Clusters

Evaluation Metrics

To evaluate the model performance, the AUC-ROC (Area of Curve under the Receiver operating Curve) is adopted. ROC curve is a graph which shows the performance of a classification model at all thresholds. It is represented based on the percentage of true positive (True Positive Rate) and false positive (False Positive rates). AUC then measures the area underneath the ROC curve, representing the degree of separability. It reflects an area between 0 and 1, where a higher value near to 1 represents a strong model while a lower value to 0 shows a poor model. These prediction results are summarised in a confusion matrix to evaluate the performance of the classification model. This is a table consisting of 4 different combinations of actual versus predicted values - True positive, True Negative, False Positive and False Negative.

Using the confusion matrix, the results were also evaluated using the following metrics:

- Precision refers to the actual number of correct predictions divided by the total predictions made by the model. This measures the number of correct positive predictions made by the model.
  $$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- Recall refers to the number of true positives over the total number of true positives and false negatives. This measures the number of positive class samples in the dataset which were correctly identified by the model.
  $$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- Accuracy is defined as the total number of correctly classified samples over the total number of classified samples.

  $$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

- F1 score combines the precision and recall scores of the model by computing the number of occurrences the model made a correct positive prediction across the dataset.

  $$F - 1\ Score = \frac{2 * True\ Positive}{2 * True\ Positive + False\ Positive + False\ Negative}$$

Therefore, a good model should present results with an AUC close to 1, alongside a higher accuracy and F1 score.

Empirical Results

The section provides the results for the selected classifiers and discusses these empirical findings. Each classifier was evaluated based on the aforementioned metrics and the importance of the features were obtained. These were then used to create new clusters of borrower segments based on the results, using 0.1 as the threshold. Finally, this new dataset with identified borrower profiles were tested on the best performing model.

Summary of Results

	AUC Score	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.933	0.93	0.93	0.93
Random Forest Classifier	0.932	0.93	0.93	0.93
XGBoost	0.945	0.95	0.95	0.95
Combined Classifier (XGBoost) with Customer Segments	0.946	0.95	0.95	0.95

Table 4: Results for Classifiers Used

Random Forest Classifier

Based on the Random Forest Classifier, the features were ranked based on their importance. Figure 9 indicates that 'last_pymnt_amnt' and 'last_fico_range_low' are the two features that stand out as significantly more important than the others in predicting loan default rates.

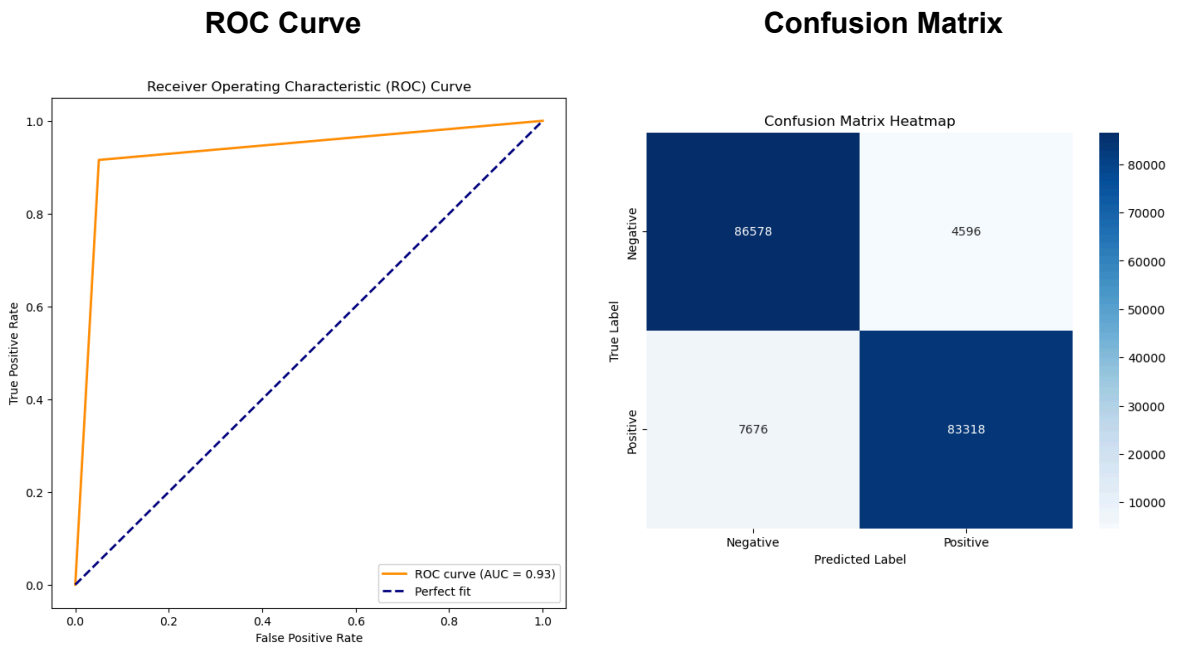


Fig 8: ROC Curve and Confusion Matrix (Random Forest Classifier)

This feature is likely to be highly indicative of a borrower's current financial status and their ability to repay the loan. A higher last payment amount could suggest that the borrower is financially stable and making an effort to repay the loan, which could correlate with a lower probability of default. Conversely, a low last payment amount might indicate financial stress or a lack of commitment to repayment, potentially leading to a higher default risk.

The FICO score is a well-established credit scoring system that is widely used to assess a borrower's creditworthiness. The 'last_fico_range_low' represents the lower end of the borrower's FICO score range at the last update. FICO scores take into account various factors such as payment history, credit utilization, length of credit history, new credit, and types of credit used. A lower FICO score is typically associated with higher credit risk, which means there's a greater likelihood that the borrower may default on their loan. It is interesting to note that the lower end was ranked higher than the ‘last_fico_range_high’.

Feature Importance **Elbow Joint Plot**

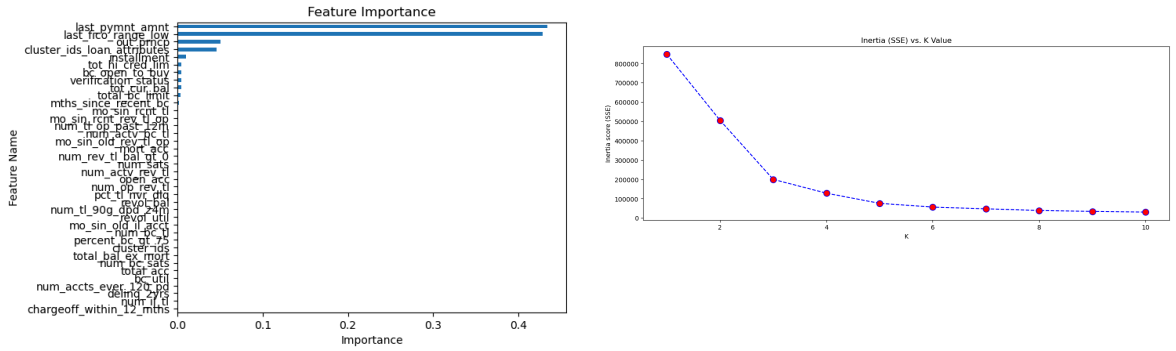


Fig 9: Elbow Joint plot of the clusters for the FICO score and last payment amount (Random Forest Classifier) K= 4 is the optimal number of clusters.

Characteristics of each cluster ID can be summarised below:

Cluster ID	**Customer Profile**
1 (Low Risk)	These borrowers make larger payments, which might suggest they are more financially secure and actively paying down their debts. Their moderately high FICO scores also indicate relatively responsible creditworthiness.
3	Their slightly lower payments are compensated by the high FICO scores, indicating a strong credit history and good credit hygiene. Therefore, they pose less of a risk to defaults.
0	Borrowers in this cluster tend to make smaller payments which could be attributed to higher levels of debt relative to their incomes. A lower FICO score corroborates the plausibility that they have a history of credit issues.
2 (High Risk)	With one of the lowest payments and FICO scores, this cluster could reflect borrowers with financial issues and significant negative credit events, with possibly delinquent accounts. They might be in a financially precarious position and represent a high risk for lenders.

Table 5: Interpretation of Customer Profile by Random Forest Classifier (Results in Appendix)

XGBoost Model

Based on the “Information Gain”, the XGBoost classifier ranked ‘last_fico_range_low’, ‘out_prncp’ and ‘last_pyment_amount’ as the highest features based on their importance. ‘out_prncp’ refers to the outstanding loan amount available while the latter highlights the amount last paid by the borrower. These two can be interpreted as complementary indicators

in highlighting the debt servicing capability of the borrower. A borrower who consistently pays more than the minimum due could be seen as less risky, since they are actively reducing their debt burden. This pattern could be captured by observing a decrease in the 'out_prncp' over time with corresponding significant 'last_payment_amount' figures. Likewise, a large last payment could suggest that the borrower has a healthy cash flow that allows them to pay off their debts promptly, while a smaller payment might raise questions about their liquidity profile.

ROC Curve **Confusion Matrix**

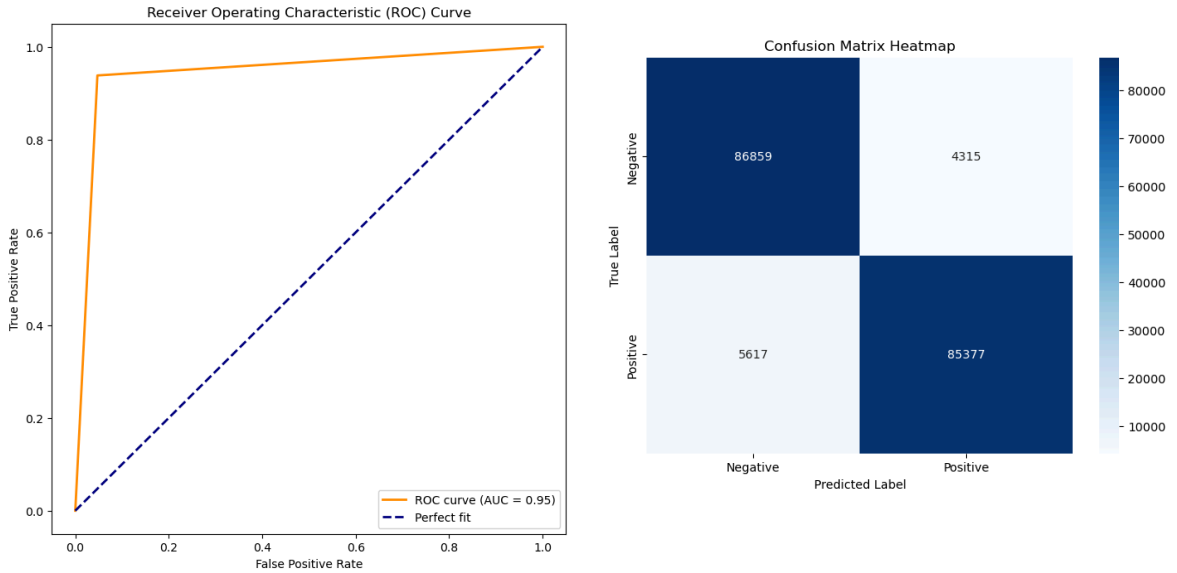


Fig 10: ROC Curve and Confusion Matrix (XGBoost Classifier)

Feature Importance **Elbow Joint Plot**

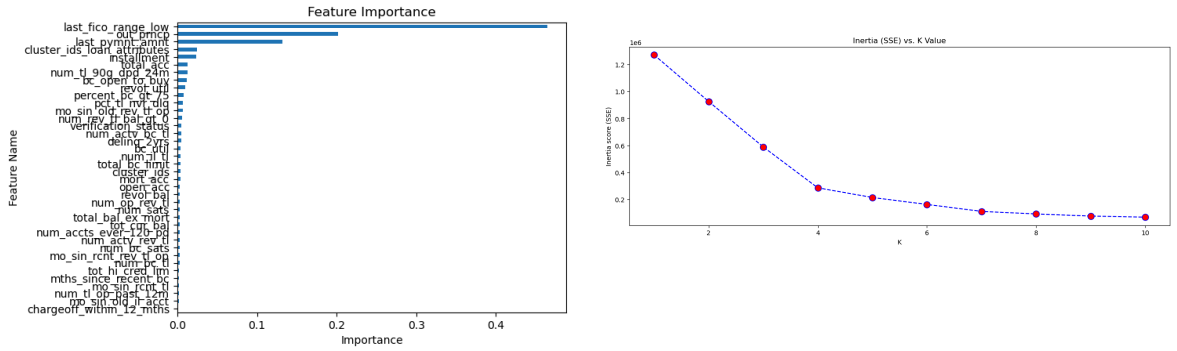


Fig 12: Feature Importance based on the classifier's results. Elbow Joint plot of the clusters for the outstanding principal left (XGBoost Classifier)

Characteristics of each cluster ID can be summarised below:

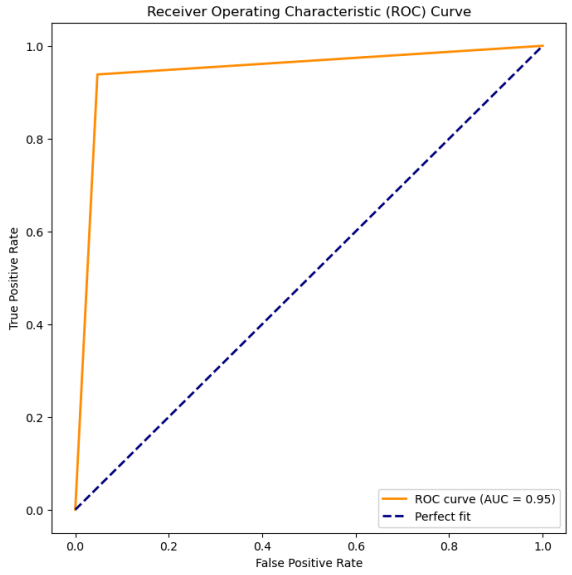
Cluster ID	Customer Profile
0 (Low Risk)	Borrowers in this profile made the highest average last payment amount, indicating that they paid their debts substantially. This is boosted by the lower than average outstanding principal, which could indicate that they have either been paying off their loans or have smaller loan amounts. It corroborates with the above average FICO score, suggesting good creditworthiness.
1	Their below average last payments amounts are complemented by the slightly lower amount of outstanding principal. With their average FICO score, this is potentially a credit risk profile of borrowers.
3	Borrowers in this cluster made a lower payment previously, albeit with a lower outstanding principal. However, their much lower FICO scores pose a riskier profile.
2 (High Risk)	This group of borrowers have one of the lowest average payments made but an outstanding amount of loans left. With their relatively low FICO scores, it is likely that these customers are in a riskier position given the significant debt remaining.

Table 6: Interpretation of Customer Profile by XGBoost Classifier (Results in Appendix)

Logistic Regression

The logistic regression returned only 1 feature - 'out_prncp' which can be found in those identified by the XGBoost Classifier. In summary, Cluster 0 appears to have the lowest risk profile with the least outstanding principal amount left to pay. This was followed by Cluster 2, 1 and 3 which poses the highest risk of default with an abnormally larger amount left.

ROC Curve



Feature Importance

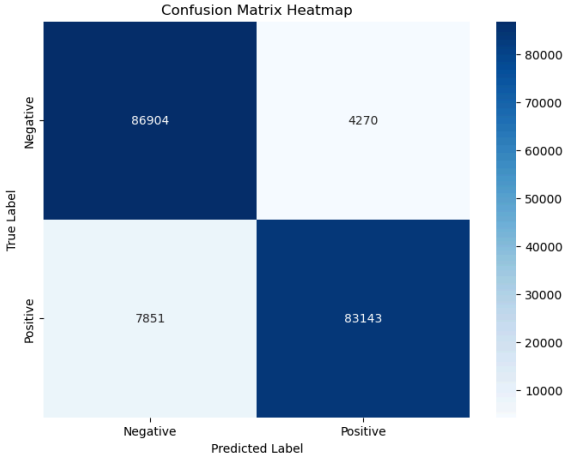
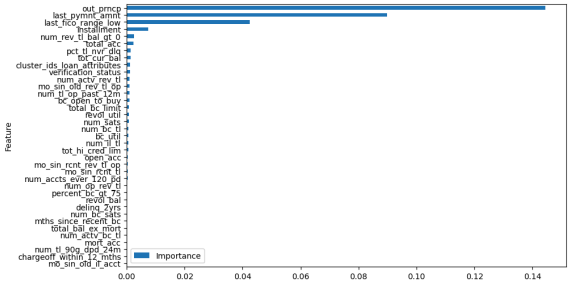


Fig 11: ROC Curve and Confusion Matrix (Logistic Regression Classifier)

Feature Importance



Elbow Joint Plot

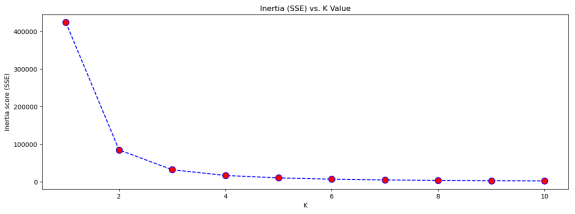
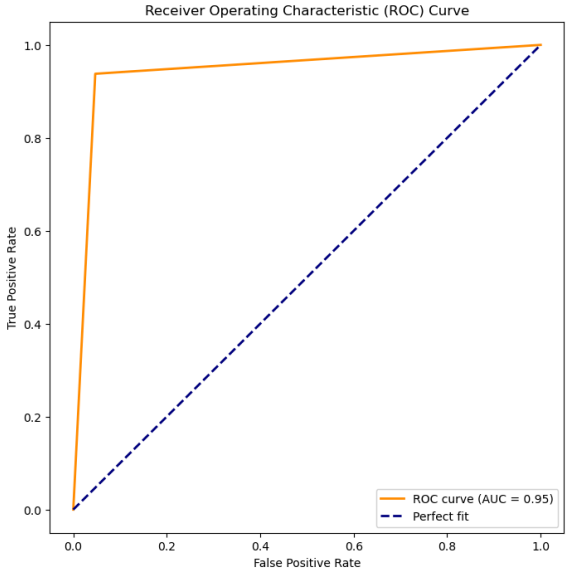


Fig 12: Feature Importance based on the classifier's results. Elbow Joint plot of the clusters for the number of revolving balances and total current balance (Logistic Regression Classifier)

Overall Classifier

Using these new attributes and clusters identified using the aforementioned techniques, the XGBoost model was selected to test the influence of these new features on predicting loan default rates. This was due to the highest AUC score and other metrics when performed on the original dataset. Hyperparameter tuning was applied once again using this new model.

ROC Curve



Confusion Matrix

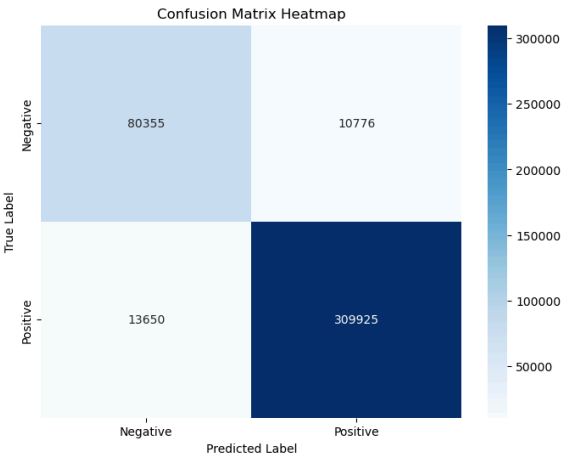


Fig 13: ROC Curve and Confusion Matrix (Overall Classifier)

While the accuracy scores did not feature significant improvements from the initial XGBoost model, the AUC score was marginally higher. Two observations can be made:

- Low FICO, outstanding principal and last payment amount made have been consistent attributes that appeared at the top as the most important features. These features are inherently linked to the borrower's financial stability and behavior, making them significant predictors in models related to credit risk and loan defaults. In fact, the latter two features could be strong indicators of a borrower's likelihood to repay their loans at each stage. A lower principal and higher payment amount potentially suggests one who is less likely to default. Their consistent appearance as top features underscores their importance in the decision-making process and their strong predictive power in assessing the likelihood of a borrower defaulting on a loan.
- This was supported with the appearance of 'cluster_ids_rf' that had a relatively moderate feature importance, with the former ranked second. A plausible reason could be due to it being a composite measure of the 3 aforementioned attributes, thus holding some predictive abilities.

To test this further, only the 4 attributes were isolated for predicting default rates. Surprisingly, this yielded an AUC of 0.934, indicating that they are strong signals for predicting loan default rates. Furthermore, simpler models using fewer features are less likely to overfit the data and the model showed that it could learn the most critical patterns associated with these variables.

Another plausible reason could be also inferred from the initial outperformance of XGBoost relative to the other 2 classifiers. Due to the inherent nature of the model, it includes built-in regularisation to prevent overfitting which could occur with the Random Forest if not properly tuned. In particular, as financial data often have features that are highly correlated, regularisation helps by shrinking the coefficients of less important features. Thereby, this reduces their impacts on the model and prevents the model from fitting noise in the training data. It also performed better than the logistic regression as the latter might be too simplistic.

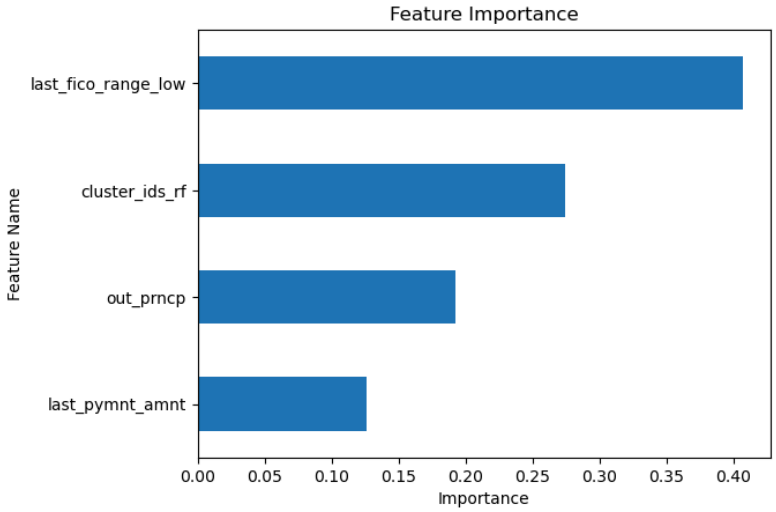


Fig 14: Feature Importance based on the classifier's results (Overall Classifier)

Conclusion

Leveraging different classifiers for predictive analytics is a powerful tool for financial institutions to gain valuable insights into their borrower profiles. In this study, apart from feature engineering and machine learning models, the paper deliberately downsampled the imbalance data set and introduced various classification algorithms to. Using the K Means clustering technique, it profiled borrowers into various profiles based on their similar characteristics. These results provide a preliminary perspective for lenders to quickly understand the characteristics and behavior of their borrowers.

However, while clustering provides a high-level overview by grouping similar data points, individual data points can indeed be very informative, especially in financial contexts where outliers or exceptional cases can hold significant insights. These single data points can represent unique or rare borrower profiles that might not fit neatly into larger clusters but are crucial for understanding the spectrum of borrower behaviour.

Appendix

Categorical Variables: One-hot encoding

Variable	Number of Terms	Mapping
term	2	{' 36 months': 0, ' 60 months': 1}
home_ownership	6	{'MORTGAGE': 0, 'RENT': 1, 'OWN': 2, 'ANY': 3, 'NONE': 4, 'OTHER': 5}
verification_status	3	{'Not Verified': 0, 'Source Verified': 1, 'Verified': 2}
pymnt_plan	2	{'n': 0, 'y': 1}
purpose	14	{'debt_consolidation': 0, 'small_business': 1, 'home_improvement': 2, 'major_purchase': 3, 'credit_card': 4, 'other': 5, 'house': 6, 'vacation': 7, 'car': 8, 'medical': 9, 'moving': 10, 'renewable_energy': 11, 'wedding': 12, 'educational': 13}
initial_list_status	2	{'w': 0, 'f': 1}
application_type	2	{'Individual': 0, 'Joint App': 1}
hardship_flag	2	{'N': 0, 'Y': 1}
disbursement_method	2	{'Cash': 0, 'DirectPay': 1}
debt_settlement_flag	2	{'N': 0, 'Y': 1}

Customer Segmentation

Socio-Economic Class

	Cluster		
column	**0**	**1**	**2**
last_fico_range_high	-0.0137	-0.0273	0.0422
last_fico_range_low	-0.0223	-0.0299	0.0521
grade	3.00	2.992	2.980
emp_length	5.811	1.267	10.791
home_ownership	0.643	0.733	0.538

Loan Attributes

	Cluster		
	0	**1**	**2**
loan_amnt	-0.155	-0.148	0.663
int_rate	0.190	-0.932	1.503
term	0.306	0.102	0.696
grade	3.306	1.665	5.024

Random Forest

	Cluster			
	0	**1**	**2**	**3**
last_pymnt_amnt	-0.494	2.344	-0.486	-0.030
last_fico_range_low	-0.196	0.592	-3.626	0.642

XGBoost

	Cluster			
	0	**1**	**2**	**3**
last_pymnt_amnt	1.993	-0.356	-0.473	-0.486
out_prncp	-0.193	-0.146	5.166	-0.161
last_fico_range_low	0.606	0.133	0.096	-3.626

Logistic Regression

	Cluster			
	0	**1**	**2**	**3**
out_prncp	-0.186	4.748	2.112	8.316

Optimal Parameters

Random Forest

Feature	Optimal Parameter	Parameters
max_depth	5	2, 5
min_samples_leaf	10	5, 10
n_estimators	10	5, 10

XGBoost

Feature	Optimal Parameter	Parameters
colsample_bytree	1.0	0.5, 1.0
learning_rate	0.1	0.01, 0.1
max_depth	5	3, 5
n_estimators	100	10, 50, 100
subsample	0.5	0.5, 1.0

Logistic Regression

Feature	Optimal Parameter	Parameters
solver	newton-cg	newton-cg
class_weight	balanced	balanced
max_iter	500	500, 1000, 2000

References

Biau, G. (2012, March 26). *Analysis of a random forests model*. arXiv.org. https://arxiv.org/abs/1005.0208

George, N. (2019, April 10). *All lending club loan data*. Kaggle. https://www.kaggle.com/datasets/wordsforthewise/lending-club

Humaira, H., & Rasyidah, R. (2020b). Determining the appropriate cluster number using elbow method for k-means algorithm. *Proceedings of the Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia*. https://doi.org/10.4108/eai.24-1-2018.2292388

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (April, 2023). K-means Clustering Algorithms: A comprehensive review, variants analysis, and advances in the era of Big Data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Jurafsky, D., & Martin, J. H. (2024, February 3). Logistic regression. https://web.stanford.edu/~jurafsky/slp3/5.pdf

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Suhadolnik, N., Ueyama, J., & Da Silva, S. (2023). Machine Learning for Enhanced Credit Risk Assessment: An empirical approach. *Journal of Risk and Financial Management*, *16*(12), 496. https://doi.org/10.3390/jrfm16120496