Loan applications analysis

Nazar Liubas

September 2025

Contents

1 Abstract

In the dynamic and expanding borrowing market, accurately assessing loan default risk and identifying distinct borrower segments are critical for ensuring lender profitability and financial stability. This project uses loan application data to comprehensively analyze these key risks.

The analysis is divided into two parts:

- Supervised Learning: This phase focuses on building predictive models. We train and evaluate Random Forest and Logistic Regression to classify loans as likely to default or not. This approach also allows for an in-depth analysis of the specific borrower characteristics that have the most significant impact on default risk.
- Unsupervised Learning: Here, we apply Principal Component Analysis (PCA) and k-means clustering to identify distinct borrower segments, each with its own unique risk profile. This segmentation provides valuable insights for targeted risk management and strategic decision-making.

The developed code can be found on GitHub repository [2].

2 Data and preprocessing

The dataset used in this study is the Lending Club dataset available on Kaggle [1]. LendingClub is an American peer-to-peer lending company that publicly released data on loan applications, loan statuses, repayments, and related information until they discontinued this practice. The dataset spans applications from 2007 to 2020 and contains approximately 3 million records across 150 variables.

Given the size of the dataset, training on the entire sample was computationally infeasible. Therefore, the analysis focuses on applications from 2019, the most recent pre-pandemic year unaffected by the economic disruptions of 2020.

Regarding features, many columns capture operational data (e.g., the total amount repaid to date) or information assigned by the company after loan approval (e.g., credit grades, interest rates). Since these variables are not available at the moment of application, they were excluded from the analysis. Instead, only borrower-provided information available at application time was retained for modeling.

The variables used in the analysis were: Loan Status, Loan Amount, Employment Length, Annual Income, Home Ownership Status, Verification Status, Debt-to-Income Ratio, FICO Score (a credit score developed by the Fair Isaac Corporation), Loan Term, and Title.

To prepare these variables for modeling, several preprocessing steps were applied:

- **Loan Status**: mapped to numeric values in order to distinguish between "good" and "bad" loans. Specifically, [Fully Paid] = 1, while [Late (31–120 days), Late (16–30 days), Charged Off, Default] = –1.
- **Employment Length**: categorical text values such as "<1 year", "1 year", . . . , "10+ years" were mapped to corresponding numeric values 0.5, 1, . . . , 10.
- **Home Ownership Status**: categorical values "OWN", "RENT", "MORTGAGE" were converted into dummy variables.

- **Term**: loan term values ("36 months", "60 months") were converted into dummy variables.

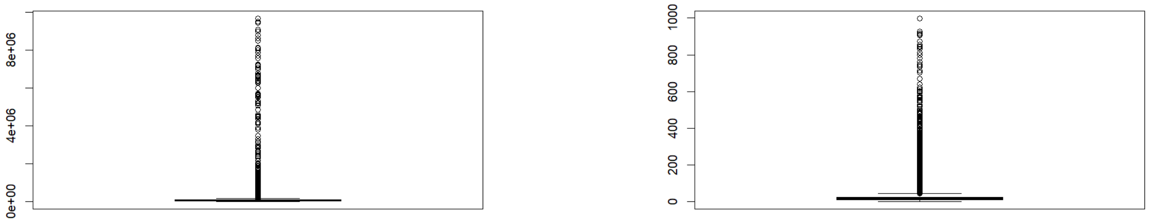
- **Verification Status**: categorical values "Not Verified", "Source Verified", and "Verified" were converted into dummy variables.

- **Title**: the most frequent 12 categories of loan titles were converted into dummy variables.

- **Address**: state codes, which were converted into dummy variables.

- **Loan Amount and FICO Score**: retained in their original numeric form.

- **Annual Income and Debt-to-Income Ratio**: boxplot inspection of Fig. 1 revealed outliers concentrated in the upper tail of the distribution. To address this, the top 1% of values were removed.



(a) Annual income boxplot

(b) Debt-to-Income ratio boxplot

Figure 1: Boxplots for defining outliers

Additionally, all missing values in the above variables were dropped.

Since the dataset was imbalanced (approximately 56k "positive" loans vs. 18k "negative" loans), random undersampling of the majority class ("positive") was applied, resulting in a balanced dataset of 17,799 records for each class, i.e., 35,598 loans.

The correlations between continous variables, as well as their distributions can be seen on Fig. 2.

3 Supervised learning

3.1 Goal statement

The primary objective of this analysis is to develop an effective classifier that can predict whether a loan will default or not. Beyond building a predictive model, the study also aims to identify which borrower characteristics are most strongly associated with default risk.

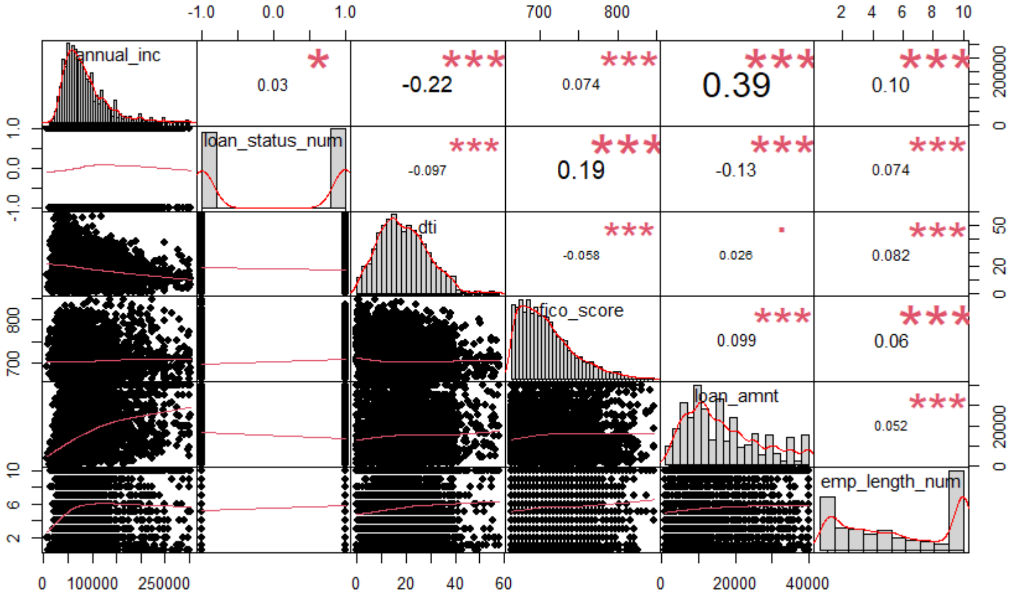


Figure 2: Variables relations

3.2 Methodology and results

Random Forest and Logistic Regression were employed as the primary classification models.

For the Random Forest, hyperparameter tuning was performed on a subset of 5,000 loans. The data was split into training, validation, and test sets in proportions of 56%, 14%, and 30% respectively. A grid search over the parameter combinations shown in Table 1 was conducted, and models were compared using the F1 score on the validation set. The best-performing configuration (highlighted in the table) was then selected. With these parameters, the final Random Forest model was retrained on the combined training and validation sets of the full dataset and subsequently evaluated on the test set.

Parameter	Candidate values
`ntree`	100, 200, **300**
`mtry`	3, 5, 7, **8**
`nodesize`	1, **5**, 10

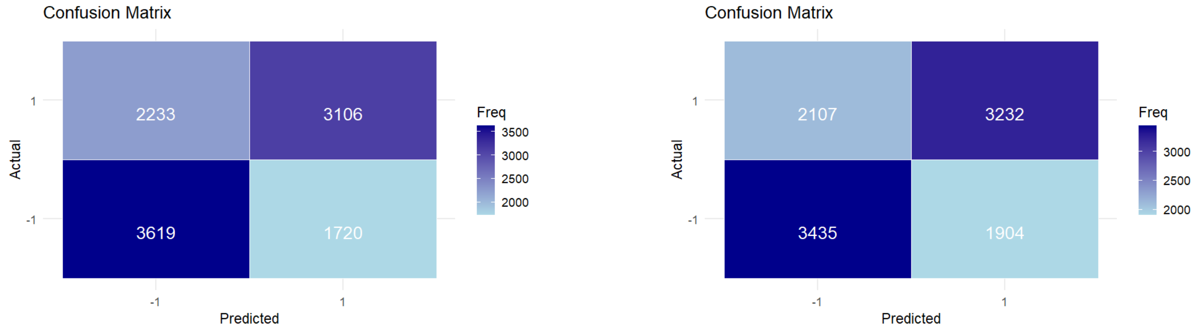
Table 1: Grid of Random Forest hyperparameters considered for tuning.

For Logistic Regression, the model was trained on the combined training and validation sets and tested on the held-out test set. Unlike Random Forest, Logistic Regression required additional care in handling categorical variables: for home ownership, term, and verification status, one category was dropped from each set of dummies to avoid perfect multicollinearity. In addition, variables with a large number of pivot categories, such as address and title, were excluded entirely.

The comparative results of both models on the test set are presented in Figure 3. Overall, the performance was very similar: Logistic Regression slightly outperformed Random Forest, achieving an F1 score of 0.617 compared to 0.611. Additionally, Logistic

Regression has benefit of better interpretability and less features used for training, so it is decided to be better model in this case.

Importantly, both models demonstrated a relatively low rate of false positives (i.e., classifying truly "bad" loans as "good"), which is the most costly error type from a business perspective.



(a) Random Forest, F1 = 0.611

(b) Logistic Regression, F1 = 0.617

Figure 3: Performance comparison of Random Forest and Logistic Regression on the test set.

Additionally, analysis of variables was conducted on Tab. 2. In addition to common columns, as Estimate, $\Pr(> |z|)$, and Significance code, Odds Ratio (OR) was added here, which is calculated as exponential of the Estimated coefficient.

Variable	**Estimate**	**Odds Ratio**	**$\Pr(> \lvert z \rvert)$**	**Signif.**
(Intercept)	-6.876	0.0010	$< 2e\text{-}16$	***
loan_amnt	-0.000027	0.9999	$< 2e\text{-}16$	***
annual_inc	0.0000011	1.0000	0.00151	**
dti	-0.0167	0.9835	$< 2e\text{-}16$	***
fico_score	0.0106	1.0106	$< 2e\text{-}16$	***
emp_length_num	0.03841	1.0392	$< 2e\text{-}16$	***
home_ownershipMORTGAGE	0.3076	1.3602	4.60e-13	***
home_ownershipRENT	-0.2528	0.7766	5.26e-09	***
term.60.months	-0.4112	0.6629	$< 2e\text{-}16$	***
verification_statusSource.Verified	-0.04027	0.9605	0.16613	
verification_statusVerified	-0.2807	0.7552	5.13e-12	***

Table 2: Logistic regression results with odds ratios. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

It shows the following:

- Loan amount has a small but significant negative coefficient, with an odds ratio very close to 1 (0.9999). This means that larger loan amounts are associated with slightly lower odds of loan being "good". Although the effect per unit is very small, it can accumulate for large loan sizes.

- Annual income has a positive effect (0.0000011, OR = 1.0000). This implies that higher incomes increase the odds of a loan being good, although the effect is economically modest due to the scaling of income.

- Debt-to-income ratio (DTI) has a negative effect (-0.0167, OR = 0.9835), suggesting that higher debt burdens reduce the probability of a loan being good. Each additional unit of DTI reduces the odds by about 1.65%.
- FICO score is one of the predictors with positive coefficient (0.0106, OR = 1.0106). Higher credit scores increase the likelihood of repayment, which aligns with expectations. Each additional FICO point increases the odds of repayment by about 1%.
- Employment length also contributes positively (0.3841, OR = 1.0392). Longer employment increases repayment likelihood, with about a 4% increase in odds per year.
- Home ownership shows clear effects:
    - Borrowers with a mortgage have higher odds of repayment (OR = 1.36), meaning they are significantly more reliable than those in the reference category (ones who own).
    - Borrowers who rent have lower odds (OR = 0.78), suggesting greater risk of default compared to owners.
- Loan term is very impactful: choosing a 60-month loan decreases the odds of repayment (OR = 0.66) compared to 36-month, meaning long-term loans are substantially riskier.
- Verification status shows mixed results: Source Verified is not statistically significant, implying no effect. While Verified has a significant negative effect (-0.28, OR = 0.76), surprisingly suggesting that verified borrowers are riskier than unverified ones. This might reflect that verified borrowers often require extra checks because their applications raise concerns.

3.3 Conclusions

The primary objective of this analysis was to develop an effective loan default classifier and to identify the key borrower characteristics most strongly associated with default risk. To achieve this, two classification models, Random Forest and Logistic Regression, were evaluated. The Logistic Regression model was found to be the better-performing model, achieving a slightly higher **F1** score of **0.617** compared to Random Forest's **0.611**. Crucially, both models demonstrated a desirable low rate of false positives, which is the most significant type of error from a business perspective. The decision to select Logistic Regression was further supported by its superior interpretability and efficiency, as it required fewer features for training.

The subsequent variable analysis using the Logistic Regression model showed several key insights into the factors influencing loan default:

- **Positive Indicators**: Higher FICO scores, longer employment lengths, and higher annual incomes were all found to be positively correlated with a loan being "good," aligning with common financial intuition. Borrowers with a mortgage were also found to be significantly more reliable than those who own their homes outright.

- **Negative Indicators**: Higher debt-to-income (DTI) ratios and larger loan amounts were associated with a higher likelihood of default. The loan term was a particularly strong predictor, with 60-month loans being substantially riskier than 36-month loans.

- **Unexpected Findings**: The analysis revealed that borrowers with "Verified" status were surprisingly riskier than their unverified counterparts, suggesting that this group may have undergone extra check due to initial concerns about their applications. Conversely, "Source Verified" status did not show a statistically significant effect.

In conclusion, the analysis successfully identified a robust model and key borrower characteristics for predicting loan defaults, providing actionable insights for risk assessment.

4 Unsupervised learning

4.1 Goal statement

This section sets a goal to explore unsupervised methods to identify borrower segments with distinct risk profiles using Principal Component Analysis (PCA) for dimensionality reduction and k-means clustering for grouping borrowers. The clusters are analyzed for differences in loan characteristics and default risk.

4.2 Methodology and results

The analysis was conducted using a set of scaled variables, including Loan amount, Annual income, Debt-to-income ratio, Employment length, FICO score, and Loan Status.

Principal Component Analysis (PCA) was first performed to explore the underlying structure of the data. The first two components explained approximately 45% of the total variance, and their loadings, visualized in Figure 4, reinforced the findings from the supervised model. This analysis reveals a positive relationship between Loan Status, FICO scores, and years of employment, while showing a negative correlation with loan amount and the Debt-to-Income ratio.

Following the PCA, the k-means clustering algorithm was applied to define distinct groups of borrowers. The optimal number of clusters, k, was determined by evaluating two key methods:

- The Elbow method, which plots the within-cluster sum of squares (WCSS) against the number of clusters and identifies the "elbow" point where the rate of decrease in WCSS significantly diminishes.

- The Silhouette method, which measures how well each data point fits within its assigned cluster, with the optimal k being the one that yields the highest average silhouette score.

Based on the results shown in Figure 5, an optimal k of 3 was selected. The final clusters are visualized in Figure 6, with a summary of key variables for each segment provided in Table 3. These segments represent the main groups of borrowers identified in the dataset:

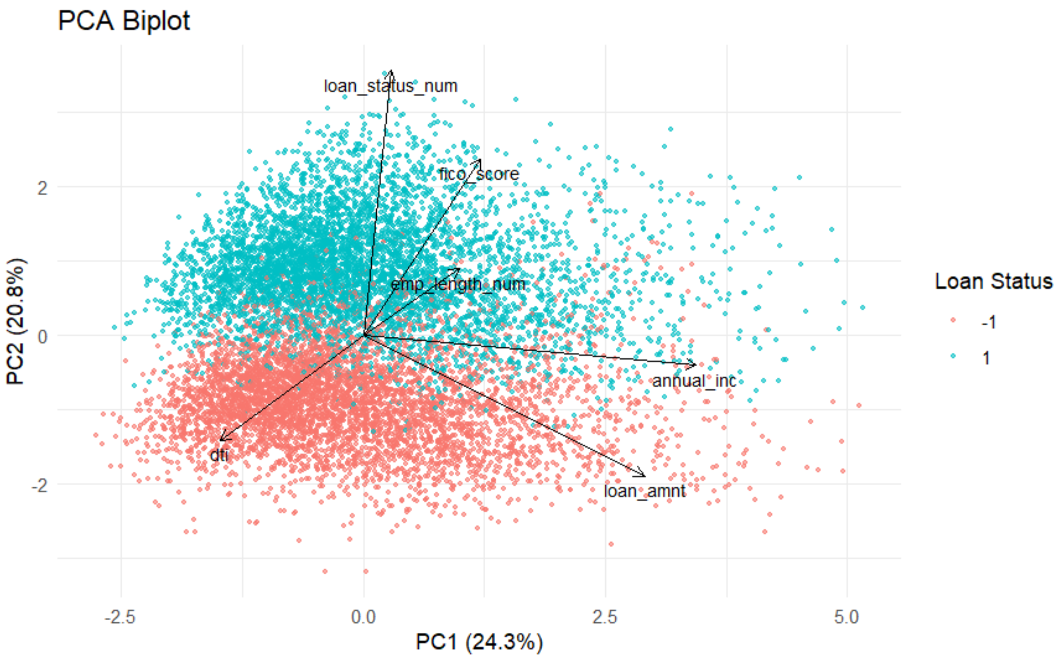
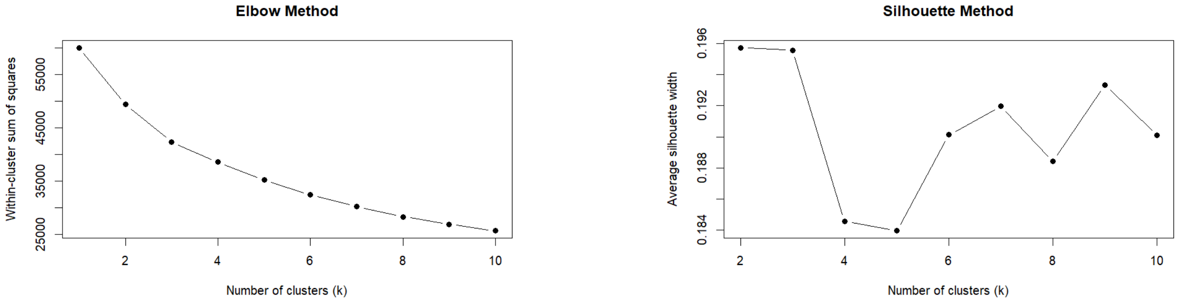


Figure 4: PCA visualisations and loadings



(a)

(b)

Figure 5: Choosing the proper K for K-means algorithm

- Cluster 1: how-risk borrowers with mid-tier Income ($70,964) and Loan Size ($11,401), suggesting that they are taking out modest, manageable loans relative to their solid income. Additionally, they have good, but not perfect, financial metrics, such as DTI (18.74) and FICO (711.5). In general, they should be actively targeted and offered competitive rates. They are highly profitable and low-risk customers.
- Cluster 2: high-risk borrowers with very high default rate of 99%. They show the lowest income and the highes debt-to-income ratio. So, despite a lower than average loan amount, they are spending the highest proportion of their income on debt payments. Additionally, they show the lowest FICO score and employment history, which suggests less job stability. In general they are source of financial loss for the lending company.
- Cluster 3: moderate risk borrowers, who have the highest income, the best FICO score, dti ratio and employment history, while taking the biggest loans. Usually more detailed analysis of such profiles should be done, in particular of the purpose

of the loan, since it may be investments with a high risk of default.

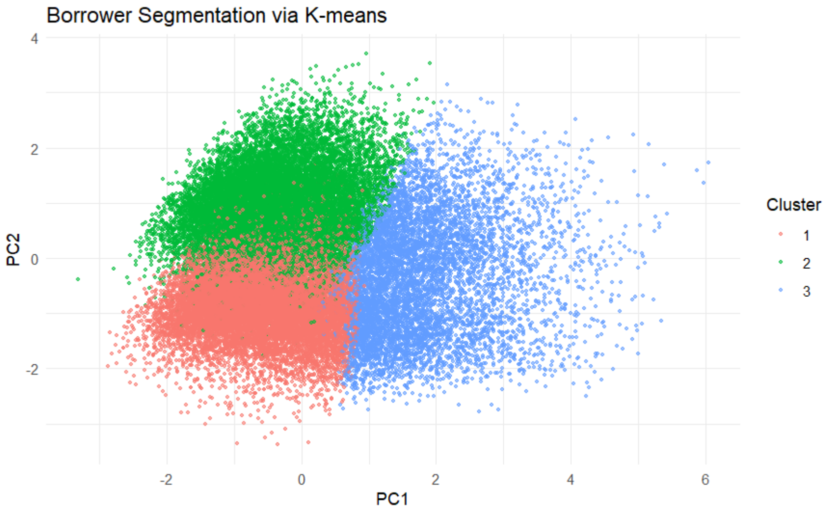


Figure 6: Clusters visualization

	Cluster 1	Cluster 2	Cluster 3
avg_loan	11,401.59	13,865.41	28,739.11
avg_income	70,964.44	65,883.39	142,191.62
avg_dti	18.74	21.33	15.57
avg_fico	711.49	696.53	718.04
avg_emp_length	5.71	5.06	5.91
default_rate	0.00028	0.99993	0.55068

Table 3: Clusters statistics

4.3 Conclusions

This analysis successfully used PCA and k-means clustering to identify three distinct borrower segments with different risk profiles:

- Cluster 1 (Low-Risk): Borrowers with moderate income and manageable debt exhibit a near-zero default rate, representing ideal, profitable customers.
- Cluster 2 (High-Risk): Borrowers with low income and a high debt-to-income ratio show a near-certain default rate (100%), identifying them as a significant source of financial loss.
- Cluster 3 (Complex Risk): Affluent borrowers with the best credit profiles but the largest loans present a paradox with a comparatively high default rate (55%), suggesting factors like strategic default or investment risk not captured by traditional metrics.

In summary, unsupervised learning revealed critical insights for targeted risk assessment, showing that default risk is driven by a combination of financial capacity and borrower behavior beyond usual credit scores.

References

[1] ethon0426. *Lending Club 2007–2020Q1 Loan Data*. Kaggle dataset. Available at `https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1`. 2020.

[2] Liubas N. *Loan applications analysis*. `https://github.com/liubas3171/loan_applications_analysis`. Accessed: 2025-09-07. 2025.