Home Exam by Eden Malka 31994040 - Loan Recommendation System

Data Overview

I work with a dataset of 1,000 loan applications that includes 14 variables. Out of these, 6 are categorical (such as *term*, *purpose*, and *home_ownership*) and 8 are numerical (including *loan_amount*, *interest_rate*, and *annual_income*).

I recode the outcome variable, **loan_status**, into a binary format where 0 represents fully repaid loans and 1 represents charged-off (defaulted) loans. I explicitly define this variable as the **target variable** for all subsequent analyses and modeling.

I also check for missing data and find that about 1.4% of the dataset is incomplete. The missing entries appear in four variables : *interest_rate*, *annual_income*, *employment_length*, and *revol_util* , each with roughly 5% missing data. All the other variables are complete. I decide to handle these missing values later using appropriate imputation techniques before training any predictive models.

** For this part of the analysis, I rely on the **Feature Statistics** generated in Orange, which are also included in the submitted files.*

Here are the variables and their descriptions as provided:

- **loan_id** – Unique identifier for each loan application
- **loan_amount** – Total amount requested for the loan
- **term** – Repayment duration of the loan (36 or 60 months)
- **interest_rate** – Annual interest rate applied to the loan
- **employment_length** – Borrower's employment duration
- **home_ownership** – Borrower's home ownership status
- **annual_income** – Annual income of the borrower in USD
- **purpose** – Primary reason for taking the loan
- **addr_state** – US state of residence of the borrower
- **dti** – Debt-to-income ratio (monthly debt payments / income)
- **delinq_2yrs** – Number of delinquencies in the past 2 years
  revol_util – Utilization of revolving credit lines (in %)
- **total_acc** – Total number of credit accounts of the borrower
- **loan_status** – Final loan outcome: Fully Paid or Charged Off

Descriptive Statistics

	loan_amount	interest_rate	annual_income	dti	revol_util	total_acc
Valid	1000	950	950	1000	950	1000
Missing	0	50	50	0	50	0
Median	19938.500	11.990	69482.450	17.935	50.750	35.000
Mean	19706.983	12.079	69450.319	17.841	50.123	35.107
Std. Deviation	11413.777	3.514	21315.413	5.041	19.810	14.713
Minimum	1009.000	2.180	12009.720	0.870	−4.130	10.000
Maximum	39952.000	28.532	169257.584	34.040	105.690	59.000
25th percentile	9424.000	9.730	55916.325	14.503	35.920	22.750
50th percentile	19938.500	11.990	69482.450	17.935	50.750	35.000
75th percentile	28716.000	14.218	82253.518	21.155	63.350	48.000

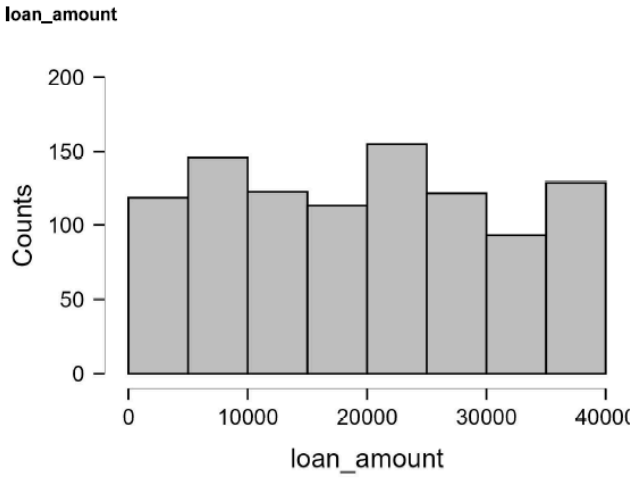
Looking at the table, I first notice that not all variables have complete data. Specifically, *interest_rate*, *annual_income*, and *revol_util* each have 50 missing values, while the other variables are fully observed. This aligns with what I already mentioned in the data overview. In terms of distributions, most variables look reasonable, but there are a couple of red flags. For example, the variable *revol_util* has a minimum value of -4.13, which is not possible since utilization of revolving credit lines cannot be negative. This is most likely a data entry or recording error that should be corrected or removed before modeling. Aside from that, the ranges generally make sense: loan amounts go from around $1,000 to nearly $40,000, interest rates range between 2% and 28%, and annual incomes span a realistic spread from about $12,000 up to $169,000. The standard deviations also highlight substantial variability across applicants, particularly for *annual_income* and *loan_amount*. Overall, the table helps me confirm both the scale and spread of each variable, while also flagging the missing values and the implausible negative observation in *revol_util* that I will need to address later.

**I chose to present only the numerical variables here in order to analyze them in detail.*

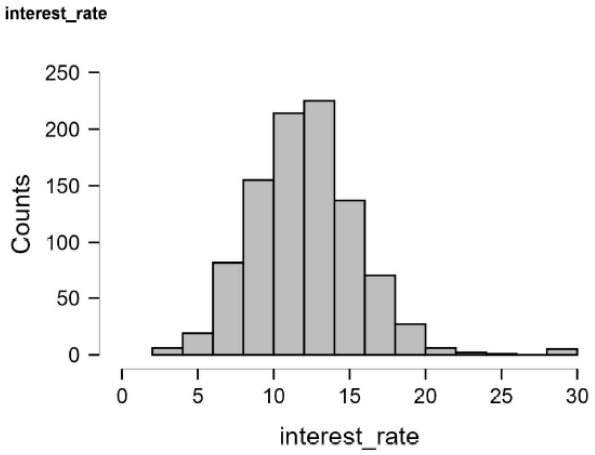
** As The variable employment_length is not included in this table but also contains missing values that will need to be handled later.*

Distribution Plots

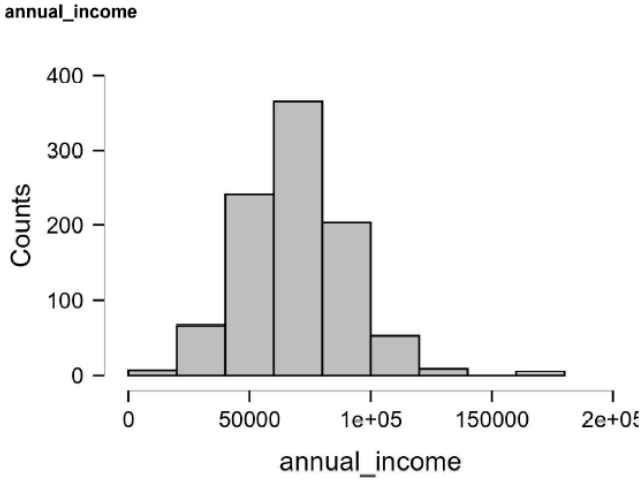
To better understand the structure of the dataset, I examine the distributions of both numerical and categorical variables. These plots allow me to identify patterns such as skewness, concentration of values, and potential outliers that may influence later modeling. Since presenting all variables would be redundant, I focus here on a selected subset of features that provide the most meaningful insights: **interest_rate, dti, annual_income, loan_amount, term, loan_status**. The full set of plots is included in the appendix.



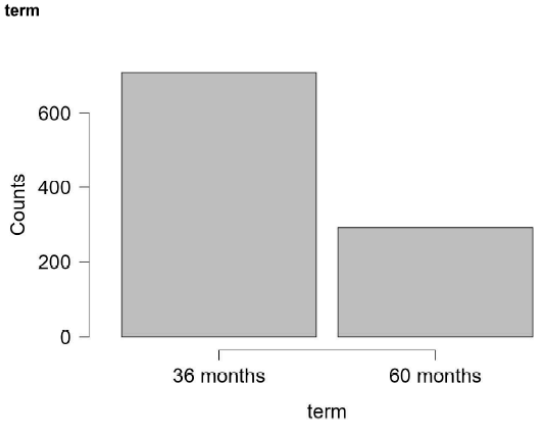
The distribution of loan amount appears fairly spread out across the possible range, without a clear single peak. Most loans fall between $5,000 and $35,000, with no sharp concentration in one area. This indicates that the bank grants loans of varying sizes quite evenly, rather than focusing on a specific segment.



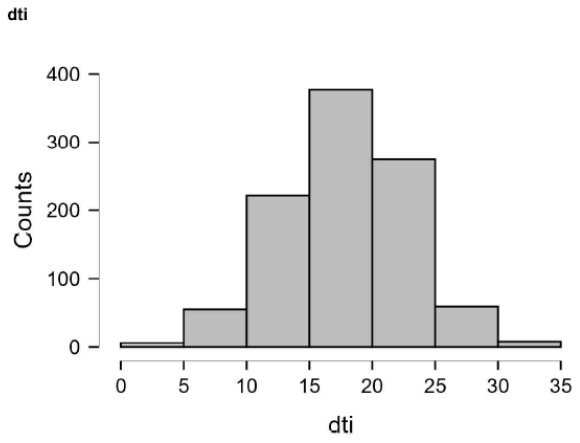
The distribution of interest rates is approximately bell-shaped, concentrated between 10–15%. Still, clear outliers appear at both ends: a small group of loans with rates around 30%, which likely reflect very risky clients with high default probability.



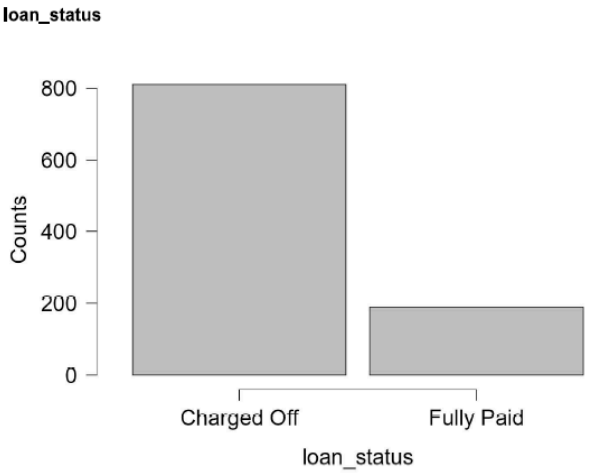
As we can see, the distribution of annual income is roughly bell-shaped, with most borrowers earning between $60,000–$80,000. Still, there are clear outliers on the upper end, above $150,000–$200,000. From a business perspective, such high earners may represent low-risk clients, but since they are rare, I will deal with them in the next part.



As we see, most loans are for 36 months, while fewer are for 60 months. This reflects business logic: shorter terms carry lower risk and are more common, while longer terms involve higher default risk.



The DTI distribution is fairly symmetric and centered around 15–20, with no clear outliers. This indicates most borrowers hold a reasonable balance between income and debt

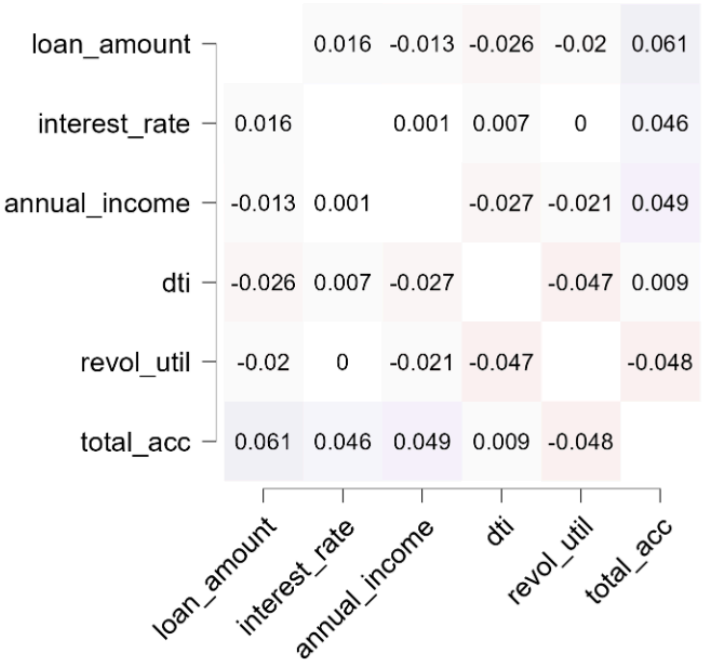


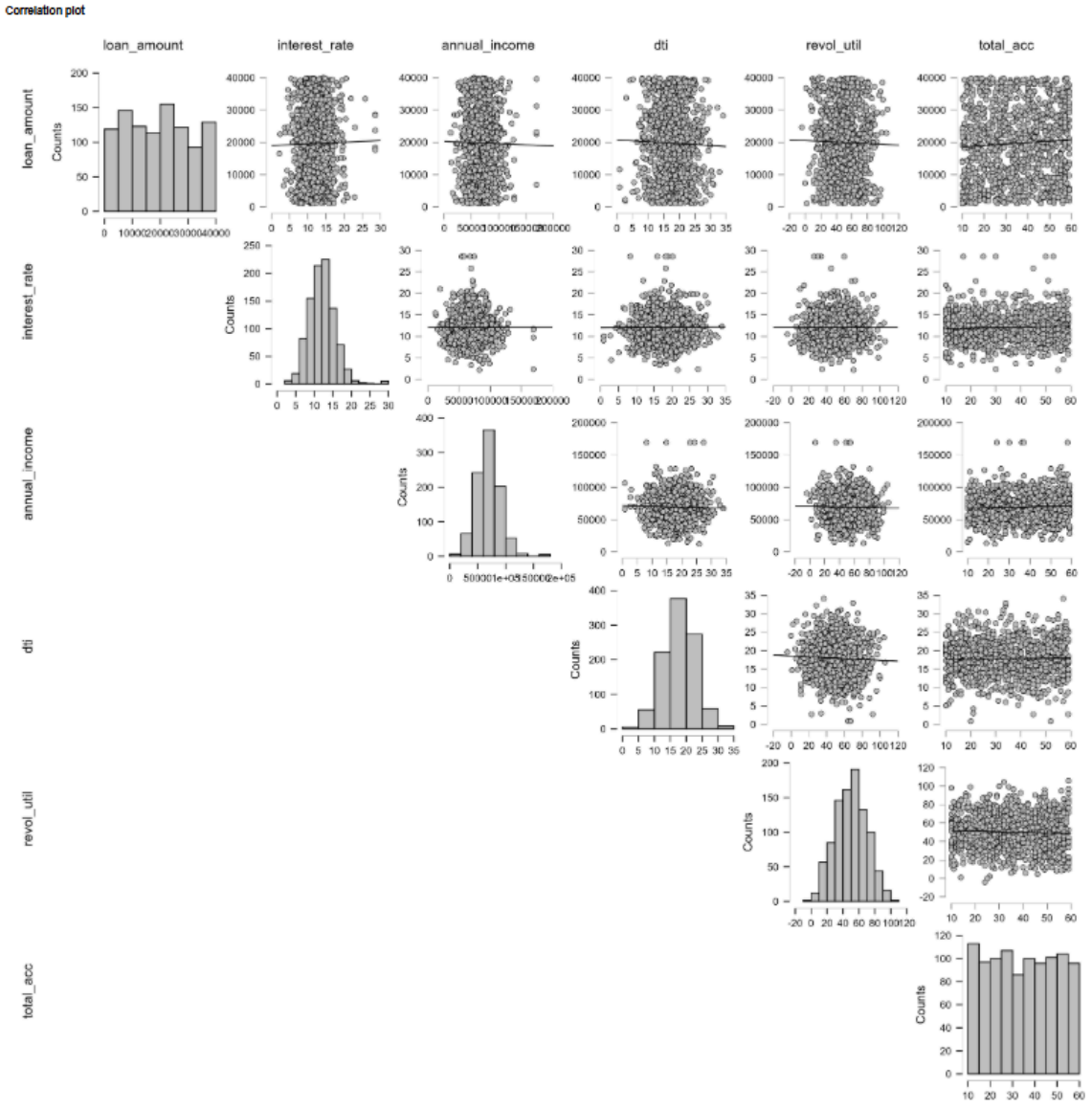
As we see, most loans in the dataset are charged off**,** while only a minority are fully paid. This imbalance is critical, as it shows that defaults dominate the data. In the next part, when I build the model, I will need to address this issue to ensure it can correctly identify risky borrowers while still approving safe clients, which is essential for the bank's profitability.

Correlation Analysis (Heatmap) & Correlation Plot

To explore the relationships between numeric variables, I looked at both a Pearson correlation heatmap and a correlation plot. Pearson's r measures the strength and direction of a linear relationship between two variables, where values close to +1 mean a strong positive link, values close to –1 mean a strong negative link, and values near 0 mean almost no linear relationship. In my heatmap, all correlations are close to zero, which shows that the features are mostly independent. The correlation plot helps me read this visually: the diagonal shows histograms of each variable, while the scatterplots show pairwise relationships. For example, annual_income and loan_amount have only a very slight positive trend, and interest_rate and DTI look almost random with no clear structure. This visual confirmation matches the heatmap values.

For my analysis, this is useful because it shows that the variables are not strongly correlated, which means I can include all of them in the loan approval model to capture different aspects of borrower risk. From a business perspective, this helps me isolate the true drivers of default and confirms that a reliable system must combine several financial indicators such as income, interest rate, and debt-to-income ratio.


loan_amount
interest_rate
annual_income
dti
revol_util
total_acc
0.016 -0.013 -0.026 -0.02 0.061
0.016 0.001 0.007 0 0.046
-0.013 0.001 -0.027 -0.021 0.049
-0.026 0.007 -0.027 -0.047 0.009
-0.02 0 -0.021 -0.047 -0.048
0.061 0.046 0.049 0.009 -0.048


Correlation plot
loan_amount
interest_rate
annual_income
dti
revol_util
total_acc
Counts

Factors Influencing Loan Default

After preparing and cleaning the dataset, I moved on to explore how different factors may influence the likelihood of loan default. To do this, I relied on **box plots** to visually compare the distributions of key variables between borrowers who repaid their loans and those who defaulted

Loan Amount and Default Status

When I examined the loan amount, I found that the average requested sum was very similar between borrowers who repaid their loans (≈19,635 USD) and those who defaulted (≈19,724 USD). The box plot shows almost identical distributions, and the t-test confirmed this with t=0.091 and p=0.927. This high p-value (greater than 0.05) means that the difference is not statistically significant. Moreover, the very small t-value indicates that the observed difference in loan amounts between the two groups is essentially negligible, supporting the null hypothesis that there is no meaningful difference. In other words, the loan amount itself does not appear to affect the probability of default.

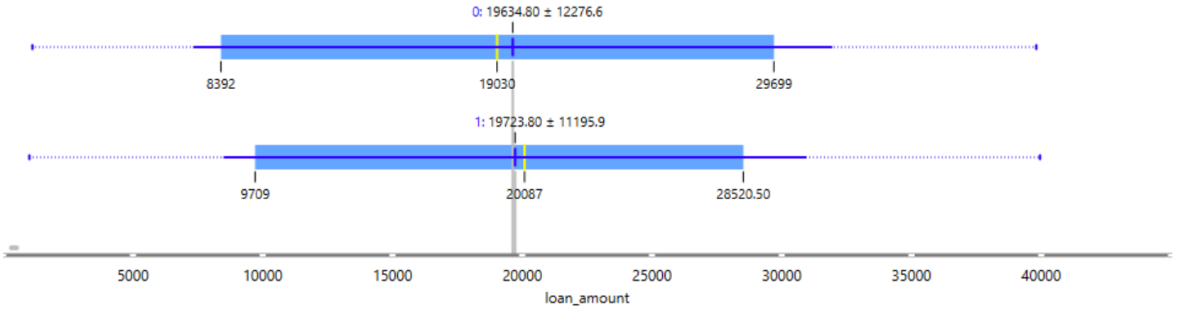


Figure 2. Distribution of loan amount by loan status (Box Plot)

Annual Income and Default Status

Looking at annual income, I observed a clearer separation. Borrowers who successfully repaid had an average income of about **73,815** USD, compared to roughly **68,458** USD among those who defaulted. This difference was statistically significant (t=2.93, p=0.004), which suggests it is unlikely to be due to random variation.

In practice, this indicates that borrowers with lower incomes are more vulnerable to default, which makes annual income an important factor to monitor during credit approval.

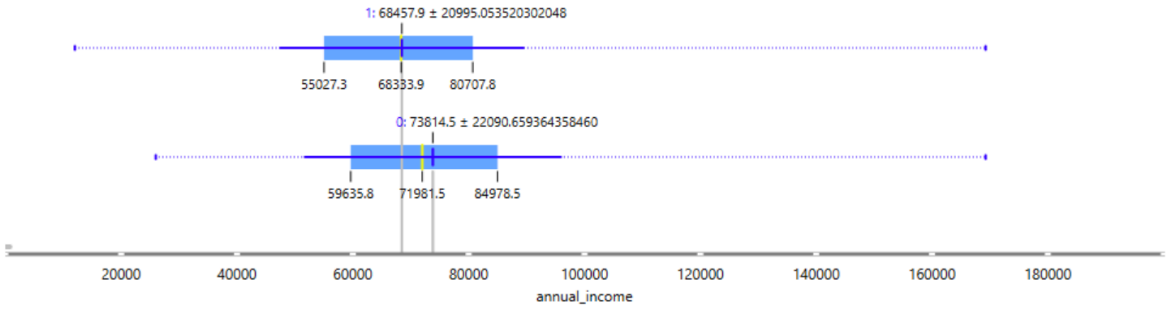


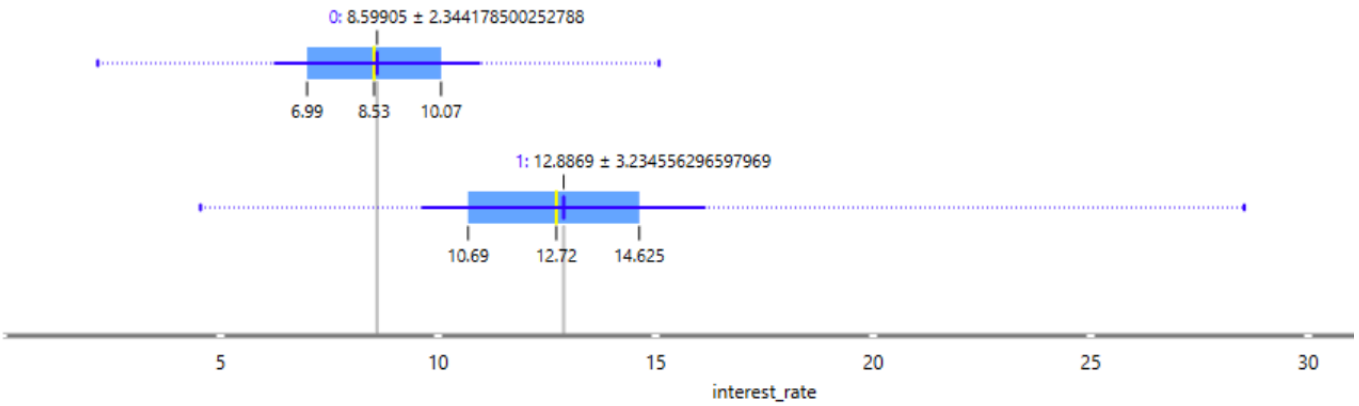
Figure 3. Distribution of annual income by loan status (Box Plot)

Interest Rate and Default Status

The difference in interest rates was striking. Loans that were fully repaid carried an average interest rate of about 8.6%, while defaulted loans had a much higher average of 12.9%. The test results (t=20.379, p<0.001) confirmed that this gap is highly significant, as we can see visually in the graph. The high t-value also means that the difference between the two groups is large compared to the variability in the data.

This finding suggests that higher interest rates, which are typically assigned to riskier borrowers, are strongly associated with default. In other words, the pricing of loans already reflects underlying risk and successfully identifies borrowers more likely to fail repayment.

Figure 4. Distribution of interest rate by loan status (Box Plot)

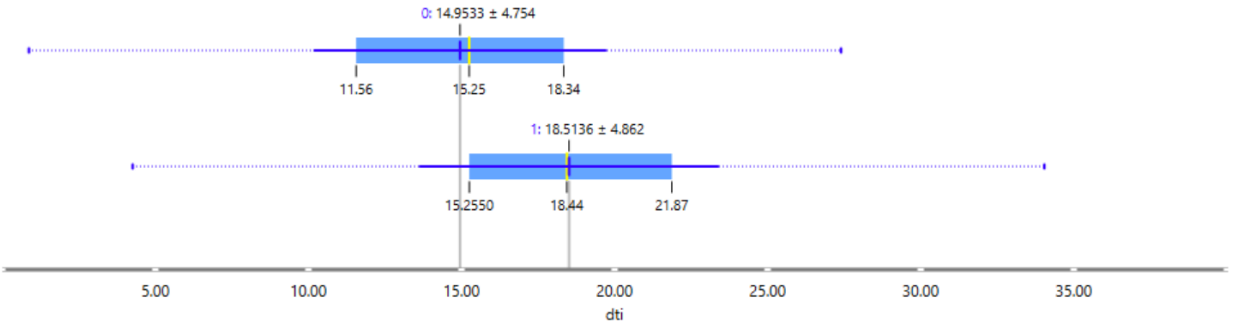


Debt-to-Income Ratio and Default Status

When I compared the debt-to-income ratio (DTI) between borrowers who defaulted and those who did not, I saw a clear difference. The average DTI for borrowers who fully repaid their loans was approximately 14.95, while for those who defaulted the average was higher

at about 18.51. Visually, the box plot shows that the distribution for defaulted loans is shifted to the right, with generally higher values and a wider spread. This suggests that borrowers who defaulted tended to carry heavier debt burdens relative to their income. The statistical test supports this observation: t = 9.231, p < 0.00. (The very low p-value indicates that the difference is highly significant, and the relatively large t-value shows that the gap between the two groups is substantial compared to the overall variability in the data).

In short, borrowers with higher debt-to-income ratios are more likely to default, which makes dti a potentially important factor to consider when assessing loan risk.



(Figure 5. Distribution of debt-to-income ratio by loan status (Box Plot

Loan Purpose and Default Status

I wanted to examine whether certain loan purposes, such as debt consolidation, are associated with a higher risk of default. To explore this, I analyzed the variable *purpose* using both a distribution plot and a Pivot Table grouped by loan status. Visually, the chart shows that defaults (in blue, corresponding to loan_status = 1) occurred across all loan purposes, though with some variation in frequency. To support this observation, I calculated the default rate for each category using the Pivot Table.

For instance, debt consolidation loans had a default rate of about **83.8%** (160 out of 191), while credit card loans showed a lower rate of approximately **78.7%** (159 out of 202). Home improvement loans also stood out with a default rate of **78.7%** (163 out of 207). In contrast, major purchase loans (81.7%) and other purposes (82.7%) showed intermediate levels of default.

Overall, these results suggest that while defaults are common across all purposes, debt consolidation and "other" loans tend to have slightly higher default rates. This indicates that loan purpose, while not the strongest predictor on its own, can still provide useful context for identifying riskier applications during the loan approval process.

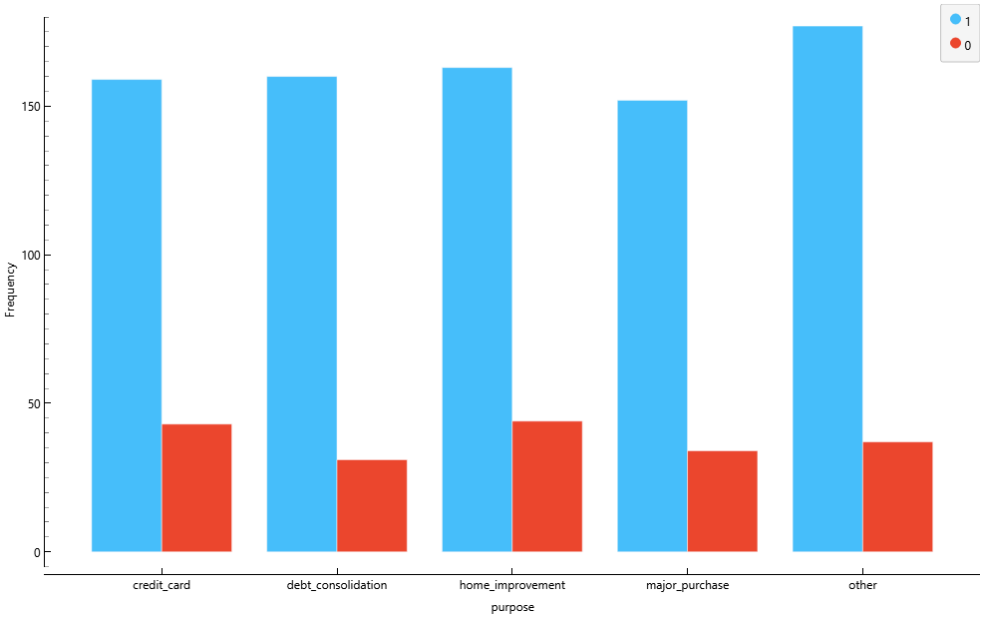


Figure 7. Distribution of loan purpose by loan status

**loan_status = 0* represents fully repaid loans, while loan_status = 1 represents defaulted loans*

	loan_status		
Count	**1**	**0**	**Total**
credit_card	159.0	43.0	**202.0**
debt_consolidation	160.0	31.0	**191.0**
home_improvement	163.0	44.0	**207.0**
major_purchase	152.0	34.0	**186.0**
other	177.0	37.0	**214.0**
Total	**811.0**	**189.0**	**1000.0**

purpose

Table 1. Pivot Table of loan status by loan purpose with default rates

Employment Length and Default Status

Looking across employment-length groups, defaults (blue, coded as “1”) dominate in every category, but the rates do vary. Borrowers with 1–3 years and 4–6 years of employment show the highest default rates, both at 83.1% (206 out of 248 in each group). The 10+ years group is slightly lower at 80.6% (154/191), and 7–9 years drops further to 78.9% (131/166). The lowest rate appears among borrowers with less than one year of employment, at 77.3% (75/97).

Overall, there isn’t a clean linear trend, but the 1–6 year range stands out as somewhat riskier, whereas <1 year shows the lowest default share in this dataset.

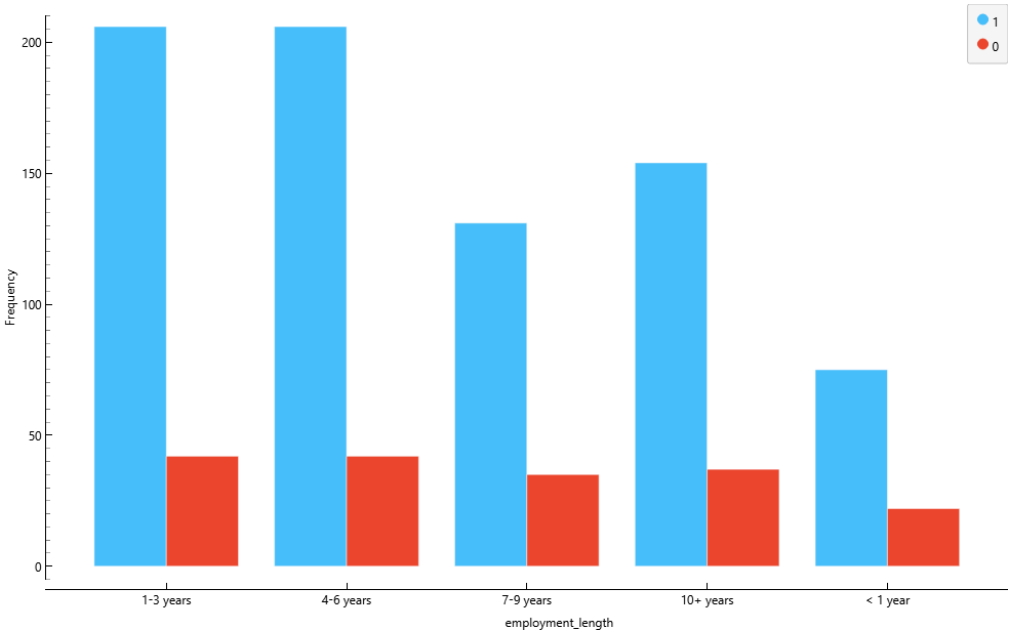


Figure 8. Distribution of loan status by employment length

loan_status

employment_length	Count	1	0	Total
	1-3 years	206.0	42.0	**248.0**
	4-6 years	206.0	42.0	**248.0**
	7-9 years	131.0	35.0	**166.0**
	10+ years	154.0	37.0	**191.0**
	< 1 year	75.0	22.0	**97.0**
	Total	**772.0**	**178.0**	**950.0**

Table 2. Pivot Table showing default rates across employment length groups

Home Ownership and Default Status

Looking at home ownership, default patterns show some interesting differences. Among renters, about **80.9%** of borrowers defaulted (327 out of 404), which is the highest rate across the groups. Borrowers with a mortgage followed closely, with a default rate of **82.1%** (325 out of 396). Homeowners without a mortgage had the lowest default rate, at **79.5%** (159 out of 200).

While the differences aren't huge, the trend suggests that owning a home outright is linked with slightly lower risk, whereas renters and those with mortgages appear more prone to default.

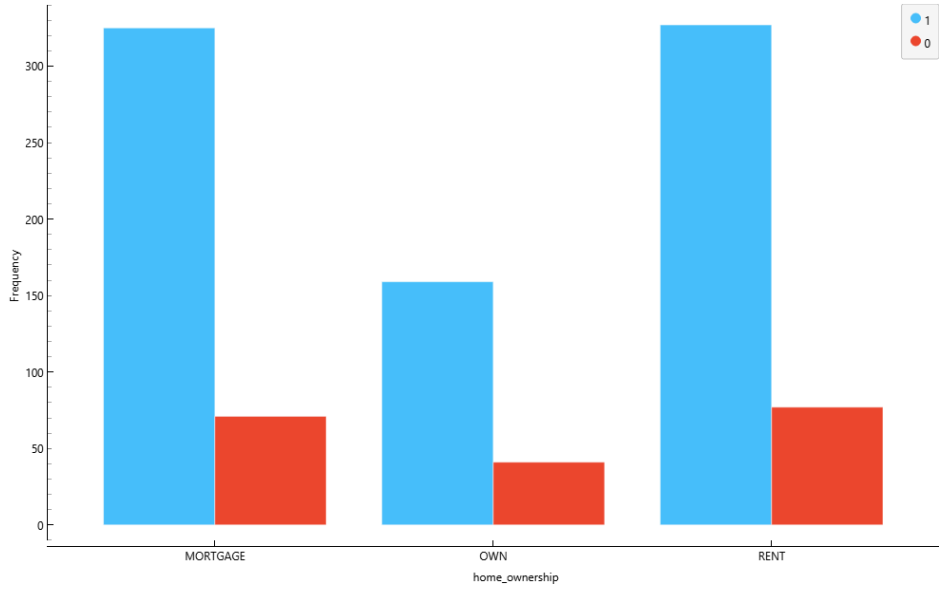


Figure 9. Distribution of home ownership by loan status

home_ownership	loan_status		
Count	**1**	**0**	**Total**
MORTGAGE	325.0	71.0	**396.0**
OWN	159.0	41.0	**200.0**
RENT	327.0	77.0	**404.0**
Total	**811.0**	**189.0**	**1000.0**

Table 3. Pivot Table of loan status by home ownership type

Interest Rate vs. Debt-to-Income Ratio (Scatter Plot)

In this part, I wanted to see whether a combination of interest rate and debt-to-income ratio (DTI) could help explain loan defaults. The scatter plot shows blue points for loans that defaulted and red points for those that were fully repaid.

The plot shows that defaults (blue) are spread across a wide range of values but appear more common among borrowers with mid-level interest rates and moderately high DTI. In contrast, many of the repaid loans (red) are clustered around lower interest rates and relatively moderate DTI values.

This pattern might indicate that borrowers with less favorable terms—neither the lowest nor the highest rates—struggle the most. On the other hand, those facing very high interest rates may have been more carefully screened or more conscious of their risk, which could explain why they show up less frequently among the defaults.

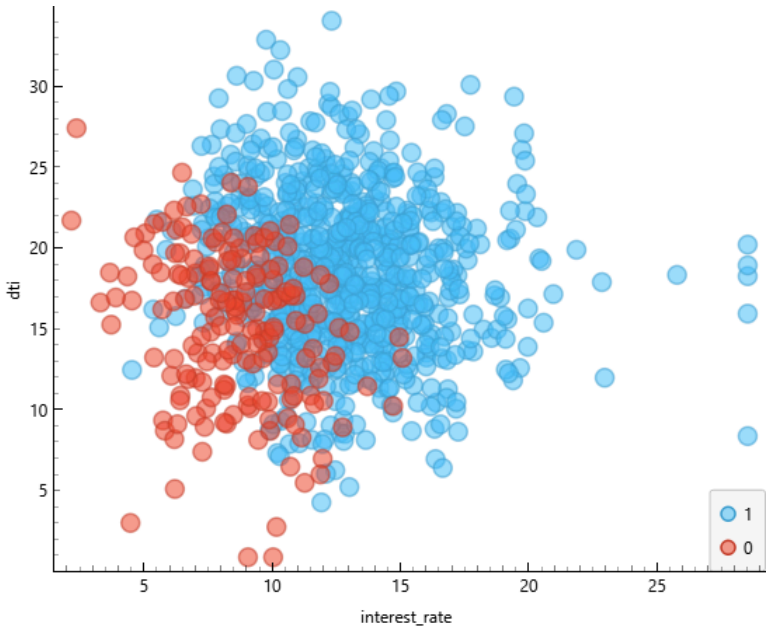


Figure 10. Interest Rate vs. Debt-to-Income Ratio by Loan Status

Key Factors Associated with Loan Default

After analyzing the dataset, several factors clearly stand out as drivers of loan performance. Among the numeric features, **interest rate** plays a central role: loans with lower rates were much more likely to be fully repaid, while defaults clustered among borrowers with mid-range interest rates. Interestingly, very high interest rates did not always correspond to the highest default risk, suggesting that these borrowers may have been more carefully screened or more cautious about repayment.

Debt-to-income ratio (DTI) also matters. Borrowers with higher DTIs were more prone to default, reflecting the burden of heavy financial commitments. However, the scatter plot of interest rate against DTI showed that the highest concentration of defaults occurred in a "middle zone" — borrowers facing both moderate-to-high DTIs and mid-level interest rates, rather than the extreme cases.

On the categorical side, **home ownership status** was strongly associated with outcomes. Renters and those with mortgages had the highest default proportions (over 80%), while borrowers who fully owned their homes showed slightly lower risk (about 79.5% default). This pattern suggests that having stable housing and fewer ongoing obligations may improve repayment capacity.

Other features, such as **loan purpose**, also displayed patterns: loans taken for credit card refinancing or home improvement tended to default more often compared to other purposes. Employment length showed only weak associations, though very short job tenure (less than one year) was linked to slightly higher risk.

Overall, the strongest drivers of loan default appear to be **interest rate, debt-to-income ratio, and home ownership status**, with loan purpose adding additional explanatory power. Importantly, the interaction between interest rate and DTI highlights that default is not simply

a matter of “high versus low risk,” but often concentrated in the middle-risk profiles where borrowers may be stretched but not obviously overextended.

Part B: Developing the Powerful Model

Data Preparation for Modeling

As part of preparing the data for modeling, I first addressed the issue of missing values. For the numeric variables with missing entries (*interest_rate*, *annual_income*, *revol_util*), I used the average of the existing values so that the replacements reflect realistic central values in the data. For the categorical variable *employment_length*, the imputation was done using the most frequent category, since this represents the typical borrower profile. In addition, I decided not to drop the outliers, because even rare clients are still potential borrowers. Instead, I capped extreme values (e.g., annual income above 200K) and treated them as missing, which means they were replaced by the average income from normal observations. For *revol_util*, I set the minimum value to 0 in order to remove the negative entry that was clearly a data entry error. As a result, the model can now focus on the true financial patterns that drive loan performance, rather than being distorted by rare errors or unrealistic cases.

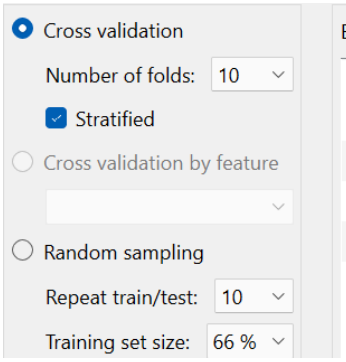
A. The Predictive Model

To develop a robust predictive model for loan recommendations, a selection of machine learning algorithms was rigorously evaluated. The primary goal was to identify a model that not only delivers high predictive accuracy but also satisfies the crucial business requirement for explainability and transparency. The evaluation was conducted using a 10-fold stratified cross-validation method. This technique ensures that our performance estimates are reliable and generalizable by training and testing each model on ten different subsets of the data, providing a comprehensive assessment of its capabilities.

After a thorough comparison, the Logistic Regression model emerged as the superior choice for our Loan Recommendation System. As shown in the evaluation results table, this model achieved an outstanding balance between strong predictive performance and the interpretability necessary for a banking environment. It registered an Area Under the Curve (AUC) of 0.931, signifying its excellent ability to distinguish between applicants who are likely to repay and those who are likely to default. Furthermore, its overall accuracy (CA) of 0.889 means it correctly classifies nearly 90% of loan applications.

From a business perspective, two metrics are particularly vital: Precision {TP / (TP + FP)} and Recall {TP / (TP + FN)}. Our Logistic Regression model scored a high precision of 0.917, indicating that when it predicts a loan will be repaid, it is correct in almost 92% of cases, minimizing the risk of approving bad loans. Critically, it also achieved a recall of 0.949. This high recall is essential for the bank, as it means the model successfully identifies nearly 95% of all actual defaulters, directly addressing the core objective of reducing risk by flagging potentially problematic loans. The strong F1-Score (0.933) and Matthews Correlation Coefficient (MCC) of 0.618 further confirm that the model provides a reliable and balanced predictive performance across both classes.

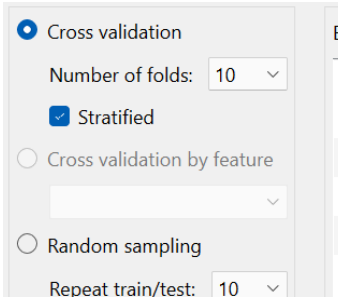
While other complex models like Gradient Boosting and Random Forest also demonstrated high performance, with Random Forest even achieving a slightly higher recall of 0.964, they were ultimately deemed less suitable. These models function as "black boxes," where the internal logic behind a decision is opaque. In a setting where loan officers must be able to explain the reasoning for an approval or denial to both customers and regulators, this lack of transparency is a significant drawback. Conversely, the kNN model, despite a high recall, was clearly unreliable, as evidenced by its extremely low AUC (0.545) and MCC (0.057), suggesting its predictions are not much better than random chance. Therefore, Logistic Regression stands out as the optimal solution, offering a powerful predictive engine whose decisions are straightforward to interpret, thereby building the trust necessary for successful implementation. Below is the table with a summary of the results:



Evaluation results for target 1

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.931	0.889	0.933	0.917	0.949	0.618
Gradient Boosting	0.918	0.884	0.930	0.912	0.948	0.599
Neural Network	0.897	0.859	0.915	0.897	0.933	0.510
Random Forest	0.858	0.862	0.919	0.878	0.964	0.487
kNN	0.545	0.788	0.879	0.816	0.953	0.057

To further validate the stability of the results, I applied a second evaluation method by splitting the dataset into 75% for training and 25% for testing using the Data Sampler. This approach simulates the real-world scenario where the model is trained on historical data and then applied to new, unseen applications. The results shows in the table below:



Evaluation results for target 1

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.929	0.887	0.931	0.920	0.942	0.617
Gradient Boosting	0.916	0.877	0.927	0.899	0.956	0.564
Neural Network	0.890	0.848	0.907	0.897	0.918	0.486
Random Forest	0.845	0.844	0.909	0.859	0.965	0.398
kNN	0.524	0.781	0.876	0.811	0.952	0.003

The performance patterns were consistent with the cross-validation results: Logistic Regression again delivered the best balance, with an AUC of 0.929, accuracy of 88.7%, precision of 0.920, and recall of 0.942. These numbers confirm that the model not only maintains its ability to identify nearly all potential defaulters but also continues to minimize false approvals. Gradient Boosting and Random Forest achieved slightly higher recall values (0.956 and 0.965, respectively), but as noted earlier, their lack of transparency makes them less practical for a banking context. kNN, on the other hand, showed unstable results (AUC 0.524, MCC 0.003), reinforcing why it was discarded. Overall, this second validation step strengthens the choice of Logistic Regression as the most robust and business-relevant solution.

Multicollinearity Check

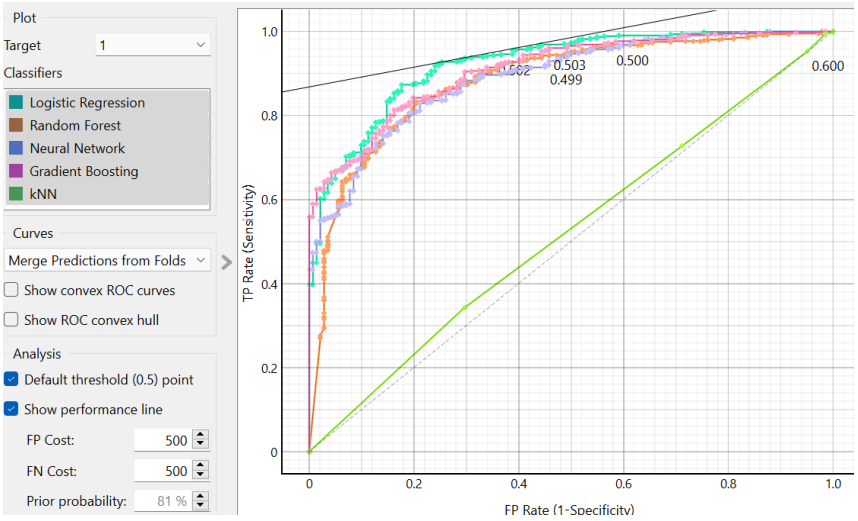
Since the logistic regression model was chosen as the final approach, I also tested for multicollinearity to ensure its reliability. In Part A, I already observed through the correlation matrix that no variables were highly correlated, but to validate this more rigorously within the model I calculated both the Variance Inflation Factor (VIF) and the Tolerance for each predictor. VIF measures how much a variable is explained by the other variables in the model, while Tolerance is simply 1 − R^2 of that regression. As we learned in the course, VIF values above 5 would indicate a potential multicollinearity problem, while Tolerance values below 0.1 would suggest a serious issue. In my results, all VIF values were well below 5 (ranging between 1.1 and 1.9) and all Tolerance values were far above 0.1, which confirms that the predictors are not strongly correlated with each other. This means the model is stable and each financial factor contributes independently, making the loan approval recommendations easier for officers to interpret and trust.

Multicollinearity Diagnostics

	Tolerance	VIF
loan_amount	0.890	1.124
interest_rate	0.534	1.874
revol_util	0.788	1.270
annual_income	0.846	1.182
total_acc	0.924	1.082
term	0.862	1.160
employment_length	0.720	1.390
home_ownership	0.875	1.143
purpose	0.694	1.441
addr_state	0.561	1.783
dti	0.670	1.493
delinq_2yrs	0.660	1.516

I also performed an ROC curve analysis to get a final visual comparison of the models. The plot shows the trade-off between correctly identifying defaulters (True Positive Rate) and wrongly flagging good borrowers as risky (False Positive Rate).

In the chart, we can see that the blue line, representing my Logistic Regression model, was consistently higher and closer to the top-left corner. This visually confirms its superior performance, which is also captured by its excellent AUC score of 0.931. From a business standpoint, this is exactly what the bank needs. It means my model strikes the best balance between minimizing losses from bad loans and avoiding the rejection of good, creditworthy customers, which is the foundation of a profitable lending system.



B. Prioritize explainability

As a loan officer, you don't just want the model to tell you "approve" or "reject" without context. What logistic regression does is calculate the probability that a borrower will default, based on the information in the loan application. Each piece of information has a weight (a coefficient) that shows whether it makes default more or less likely. A positive weight means the factor pushes the risk of default up, while a negative weight pushes it down.

To make these effects easier to interpret, I exponentiated the coefficients into odds ratios (OR). An OR above 1 indicates increased risk of default, while an OR below 1 means reduced risk. This makes the model's recommendations clear and actionable for loan officers. We can see it in the table below:

Variable	Coefficient	OR(= exp(Coefficient))	Interpretation
Interest Rate	-0.944	0.389	Each unit increase in the interest rate reduces the odds of default to 39% of the previous level (risk decreases).
Employment length (1–3 years)	1.454	4.28	Borrowers with 1–3 years of employment are about 4.3 times more likely to default compared to the baseline.
Term = 60 months	2.697	14.83	Loans with a 60-month term are nearly 15 times more likely to default compared to 36-month loans.
Home ownership = Mortgage	1.886	6.59	Borrowers with a mortgage are about 6.6 times more likely to default compared to the baseline group.

For example, the coefficient for interest rate corresponds to an odds ratio of 0.389, meaning that with each unit increase in the rate, the odds of default become only 39% of what they were before (i.e., the risk is reduced). In contrast, borrowers with 1–3 years of employment are about 4.3 times more likely to default compared to the baseline, and loans with 60-month terms are nearly 15 times more likely to default. This transparency makes the model a tool you can trust when making daily credit decisions.

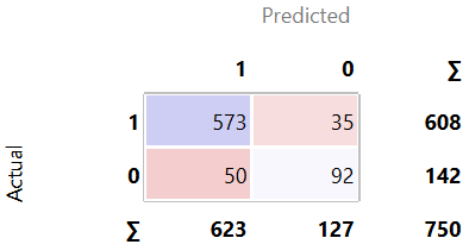
**note that in the distribution plots we observed the highest concentration of defaults in the mid-range interest rates (around 10–15%), while at the very high end (25–30%) the pattern was less pronounced. Since logistic regression assumes a linear relationship with the log-odds, it tries to approximate this non-linear, U-shaped trend with a straight line. As a result, the coefficient for interest rate may appear negative, even though the descriptive plots suggest a different underlying pattern.*

So I can conclude that as a loan officer, you can clearly see which factors drive each recommendation, making the model a transparent and practical tool rather than a "black box." It supports your daily work by helping you justify decisions to clients, regulators, and management with confidence.

C. Evaluate the model.

My goal is to see how the model helps the bank minimize losses from bad loans while still maximizing the number of good loans approved. In the following analysis, I demonstrate how the results of the model directly support these business goals.

Confusion Matrix



To really understand the confusion matrix, it helps to go through each cell. The top-left (573) shows the True Positives – borrowers who actually defaulted and were correctly flagged by the model. The top-right (35) are the False Negatives – risky borrowers who defaulted but were mistakenly approved. These are the most costly mistakes, since they mean real financial losses; for example, if the average loan is $10,000, that could be around $350,000 lost. On the other side, we had 142 good borrowers: 92 of them were correctly approved (True Negatives), while 50 were mistakenly flagged as defaulters (False Positives). These cases don't create direct losses, but they do mean missed profit opportunities and weaker customer relationships. When we look at the metrics, Recall is defined as:

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \text{Recall}$$

is $573/(573 + 35) = 0.94 = 94\%$. This tells us that the model catches almost all the risky borrowers, which is crucial for protecting the bank from major losses.

Precision is defined as:

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = \text{Precision}$$

Here,thia iis $573/(573 + 50) = 0.92 = 92\%$. This means that when the model rejects a loan, it is correct the vast majority of the time..

Together, these two metrics show a strong balance: the bank is shielded from nearly all bad loans while still approving most good ones. In practice, this translates into fewer unpaid loans, more sustainable profitability, and greater trust in the system.

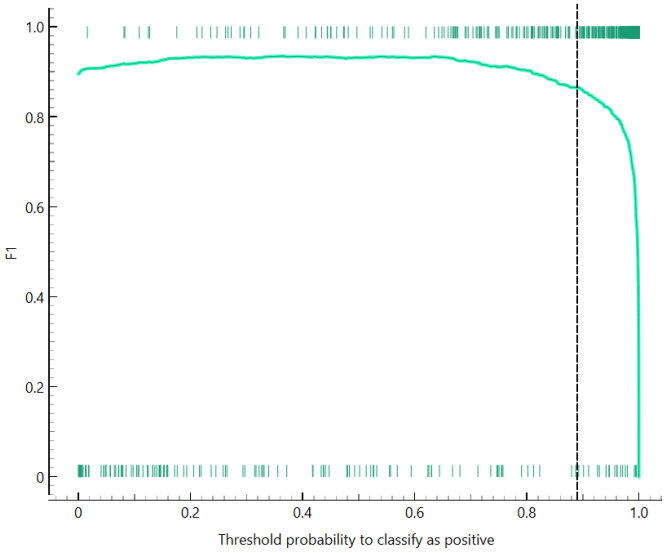
Managing the Trade-off: Minimizing Losses vs. Maximizing Good Loans

Another important aspect of evaluating the model is the choice of the probability threshold. By default, logistic regression uses 0.5 as the cutoff, but when we examined the performance curve, we saw that the best balance of performance occurs at a lower threshold, around 0.3.

Mathematically, this means that the model classifies a borrower as “likely to default” even if the predicted probability is only 30%.This directly connects to the bank’s two main goals: minimizing losses and maximizing good loans. Lowering the threshold increases recall, which minimizes losses by catching almost every potential defaulter and preventing risky approvals. At the same time, it slightly reduces precision, which means some good customers are rejected unnecessarily, this lowers the number of approved profitable loans.

The calibration plot below confirms this result by showing how the F1 score – which combines precision and recall into a single measure – peaks around the same range (0.3–0.4). The F1 score is defined as:

$$\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \cdot 2 = F1$$



This demonstrates that the model achieves its best overall balance when risky borrowers are flagged early, reinforcing the conclusion that a lower threshold better serves the bank’s strategy.

Moreover, the value of the model is that management can decide where to set this balance. In times of higher economic uncertainty, the bank may prefer to minimize losses and set a stricter threshold. In more favorable conditions, it may raise the threshold closer to 0.5, approving more good loans and maximizing revenue. In this way, the model is not just a predictor, but a flexible decision-making tool that allows the bank to actively control the trade-off between risk and profitability.

Model Stability and Business Impact

Finally, it is important to emphasize that the model’s strong performance is not a one-off result. In the first part of the analysis, I compared models using two different validation methods: 10-fold cross-validation and a 75/25 train–test split. Both approaches confirmed that logistic regression consistently delivers high accuracy, precision, and recall. This consistency indicates that the model is not overfitting to a single dataset but instead strikes a healthy balance between bias and variance. In other words, it is complex enough to capture

meaningful patterns, but not so complex that it loses its ability to generalize. This stability shows that the model is reliable and can be trusted to perform well on new loan applications.