# Cracking the credit default : Predicting Mortgage Loan Risks with Precision

Aditi Anand

## Contents

**Introduction:**

In the financial services sector, particularly within mortgage lending, accurately predicting loan defaults is essential for minimizing risk and maintaining portfolio health. Defaults can lead to substantial financial losses and disrupt financial markets and economic stability. Given the everevolving lending landscape, influenced by fluctuating interest rates, changing economic conditions, and varying borrower behaviors, lenders must employ advanced techniques to foresee and manage defaults.

Predictive modeling, especially logistic regression, has emerged as a powerful tool for predicting default probabilities. By analyzing borrower attributes like credit scores, loan-to-value ratios, debt to income ratios, and delinquency status, lenders can identify key risk factors and take preemptive measures to reduce losses. By examining these factors, the model seeks to pinpoint the most significant predictors of default. The study also tackles challenges like class imbalance and multicollinearity, refining the model to enhance its accuracy and relevance for lending institutions.

**Current study:**

1. **orig_rt (Original Interest Rate)**: Higher interest rates could be associated with greater risk, especially in times of economic stress.
2. **orig_trm (Original Loan Term)**: The term of the loan impacts repayment schedules and can influence default likelihood.
3. **oltv (Original Loan-to-Value Ratio)**: High LTV ratios are typically riskier, as they indicate the borrower has less equity in the property.
4. **LAST_RT (Last Interest Rate)**: Changes in interest rates could indicate adjustments in payments, impacting the likelihood of default.
5. **F30_DTE, F60_DTE(Delinquency Dates)**: These indicate when the loan became delinquent, which is crucial for tracking the progress toward foreclosure.
6. **RELO_FLG (Relocation Flag)**: Mortgages issued for relocation purposes may carry different risks, especially if the borrower's employment situation is unstable.
7. **MOD_FLAG (Modification Flag)**: Indicates if the loan has been modified, which could signal previous financial distress.
8. **CSCORE_B (Borrower Credit Score at Origination)**: This is crucial in assessing the borrower's ability to repay the loan.
9. **dti (Debt-to-Income Ratio)**: Measures the borrower's debt burden relative to income, a critical factor in determining the ability to make payments.

Credit score of borrowers had multiple NA values so such rows were removed from the dataset.The 30 and 60 days delinquency were converted to binary variable wherein the loans which had delinquency were allocated 1 whereas others as 0. The relocation variable was converted to binary variable with Yes as 1 and No as 0. For the ease of analysis, all these variables were converted to numeric variables.

**The status of mortgage loan is an important variable which is a resultant of borrower's details. However, it is converted to binary variable to make more informed decision making. The lenders can classify the loans by at the very onset as it reflects the default at the very onset.**

**The categorisation is done as follows:**
**1 = Default**: Include loans with:

**Delinquency** (codes "1" to "9" indicating months of delinquency).

**Foreclosure/Distressed Sales** (codes "F", "S", "T", "N", "L", "R" indicate various forms of distress like Deed-in-Lieu, Short Sale, etc.).

**0 = Non-default**: Include loans that are:

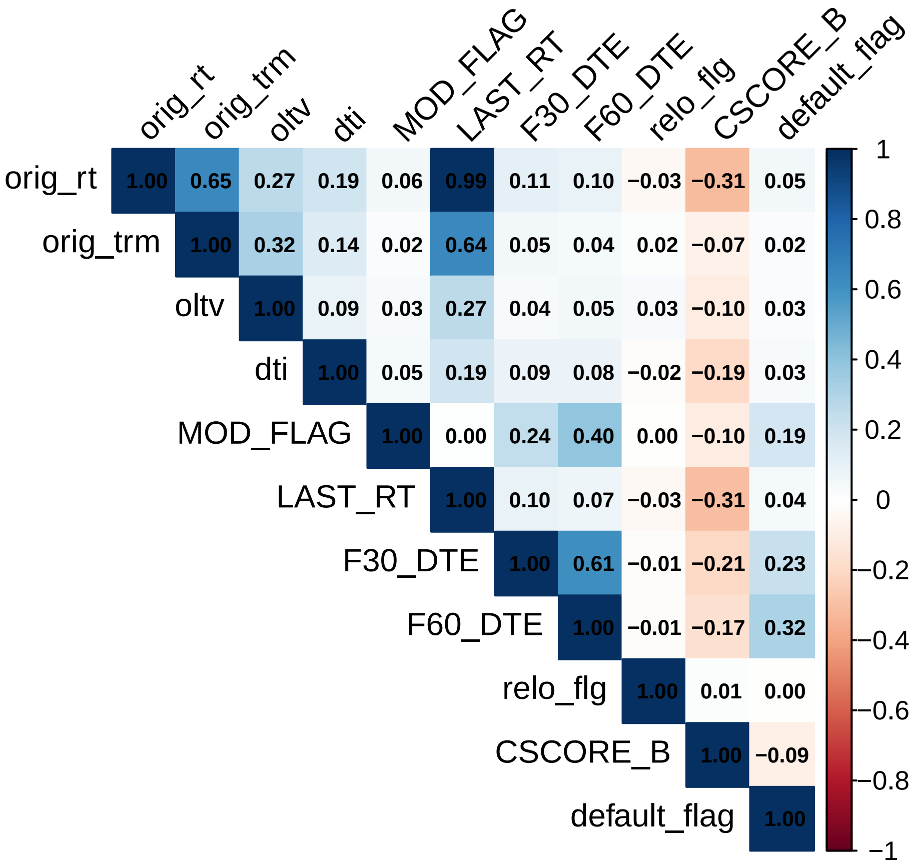
**Current or Paid Off** (code "C" for current, "P" for prepaid/matured).

For first round of analysis, 9 potential predictors which looked the most suitable according to subject matter knowledge were taken. These variables could have an impact on the likelihood of default however the computation of variables was done to perform correlation analysis between the response variable and covariates.

**Key Findings:**

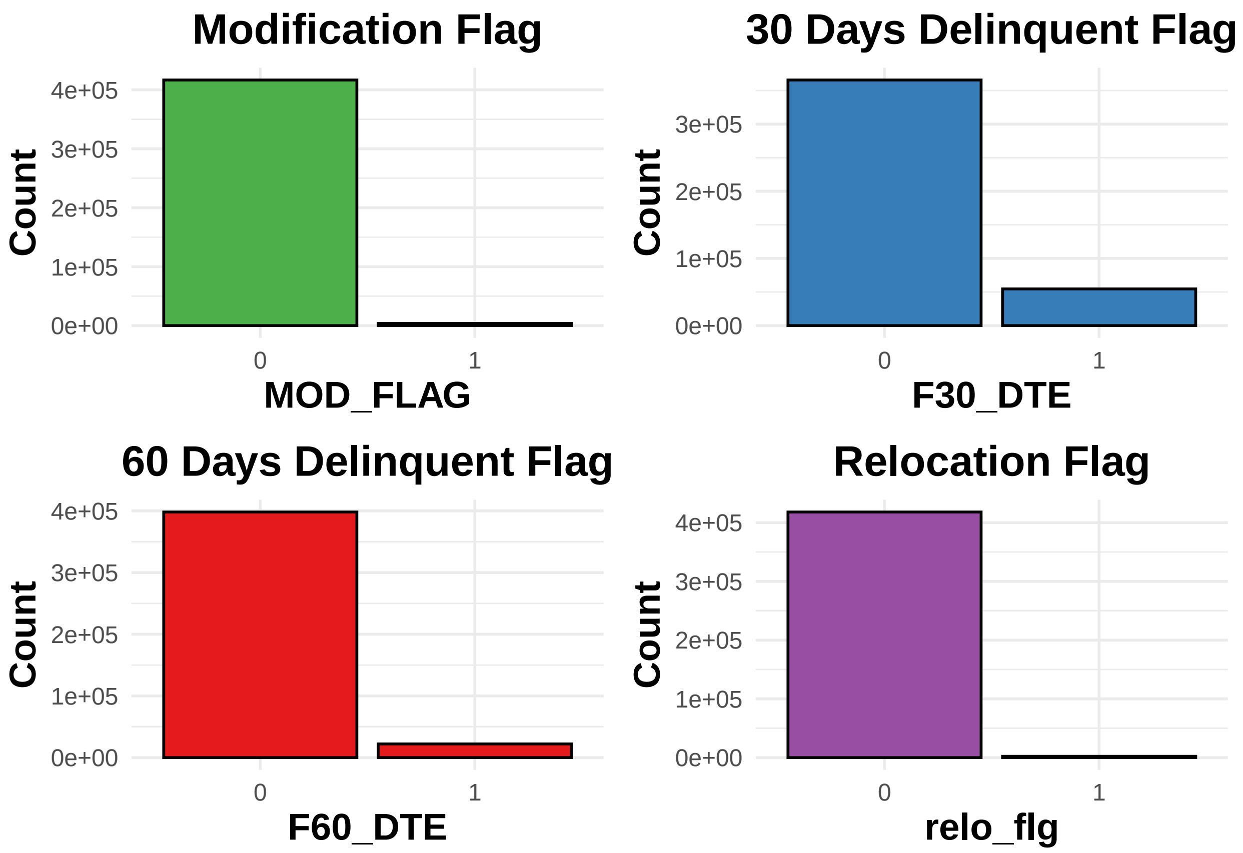
Most of the selected variables showed a weak positive correlation which suggests that these variables influence response variables. However, borrower's credit score and the current outstanding unpaid balance shows a negative correlation with response.

**Borrower's credit score negative relationship with response variables suggests that an increase in credit score will lead to lower likelihood of loan default.**



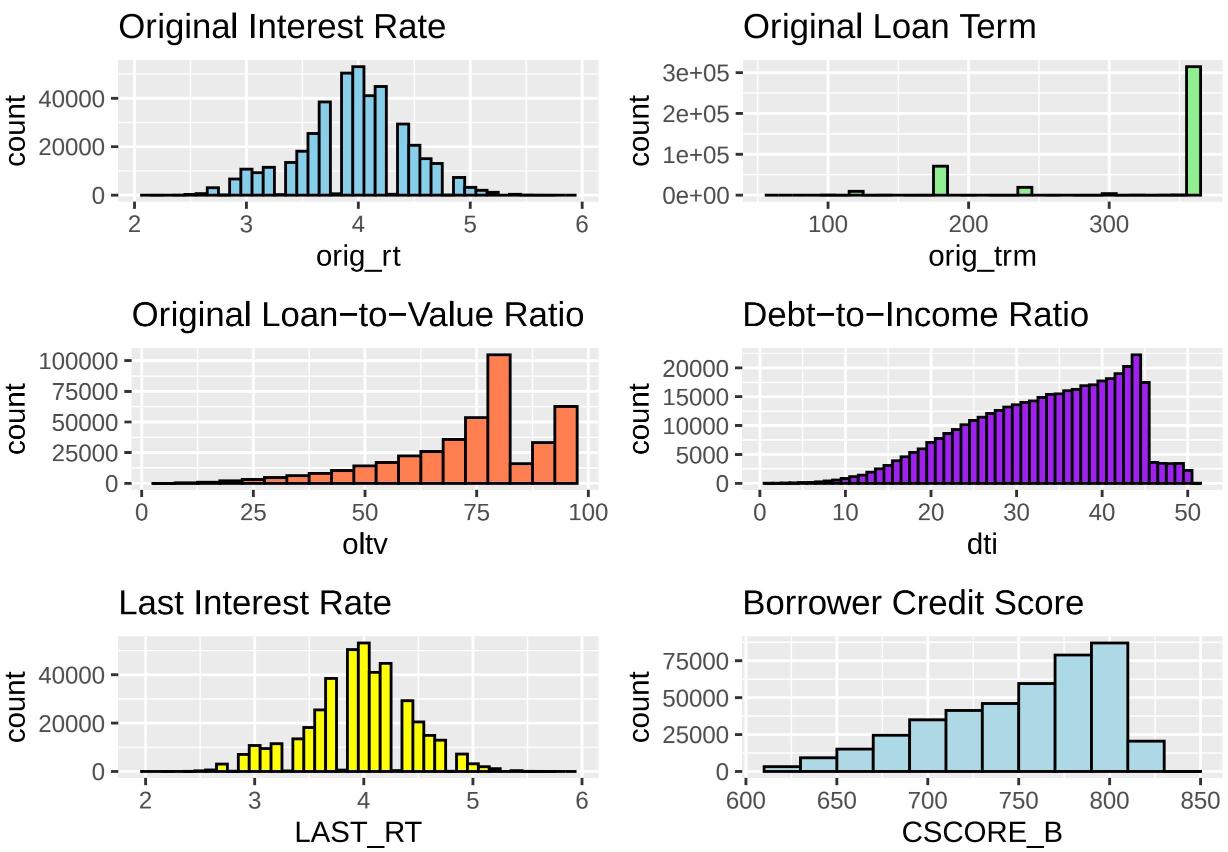
|  | orig_rt | orig_trm | oltv | dti | MOD_FLAG | LAST_RT | F30_DTE | F60_DTE | relo_flg | CSCORE_B | default_flag |
|---|---|---|---|---|---|---|---|---|---|---|---|
| orig_rt | 1.00 | 0.65 | 0.27 | 0.19 | 0.06 | 0.99 | 0.11 | 0.10 | −0.03 | −0.31 | 0.05 |
| orig_trm |  | 1.00 | 0.32 | 0.14 | 0.02 | 0.64 | 0.05 | 0.04 | 0.02 | −0.07 | 0.02 |
| oltv |  |  | 1.00 | 0.09 | 0.03 | 0.27 | 0.04 | 0.05 | 0.03 | −0.10 | 0.03 |
| dti |  |  |  | 1.00 | 0.05 | 0.19 | 0.09 | 0.08 | −0.02 | −0.19 | 0.03 |
| MOD_FLAG |  |  |  |  | 1.00 | 0.00 | 0.24 | 0.40 | 0.00 | −0.10 | 0.19 |
| LAST_RT |  |  |  |  |  | 1.00 | 0.10 | 0.07 | −0.03 | −0.31 | 0.04 |
| F30_DTE |  |  |  |  |  |  | 1.00 | 0.61 | −0.01 | −0.21 | 0.23 |
| F60_DTE |  |  |  |  |  |  |  | 1.00 | −0.01 | −0.17 | 0.32 |
| relo_flg |  |  |  |  |  |  |  |  | 1.00 | 0.01 | 0.00 |
| CSCORE_B |  |  |  |  |  |  |  |  |  | 1.00 | −0.09 |
| default_flag |  |  |  |  |  |  |  |  |  |  | 1.00 |

Categorical variables like MOD_FLAG, F30_DTE, F60_DTE, relo_flg have an imbalance hence 0 is set as a reference category as it has higher frequency, this makes the interpretation of model coefficients relatively easier and makes model more stable.



| | Variable 1 | Variable 2 | P-Value | Statistic | Degrees of Freedom |
|---|---|---|---|---|---|
| MOD_FLAG vs F30_DTE | MOD_FLAG | **F30_DTE** | 0.00e+00 | 2.468620e+04 | 1 |
| MOD_FLAG vs F60_DTE | MOD_FLAG | **F60_DTE** | 0.00e+00 | 6.590102e+04 | 1 |
| MOD_FLAG vs relo_flg | MOD_FLAG | **relo_flg** | 1.24e-01 | 2.365274e+00 | 1 |
| F30_DTE vs F60_DTE | F30_DTE | **F60_DTE** | 0.00e+00 | 1.564319e+05 | 1 |
| F30_DTE vs relo_flg | F30_DTE | **relo_flg** | 9.52e-21 | 8.725890e+01 | 1 |
| F60_DTE vs relo_flg | F60_DTE | **relo_flg** | 5.25e-11 | 4.308095e+01 | 1 |

On chisq test, it was seen **both the 30 days and 60 days delinquency shows an association with loan modification or relocation suggesting that loans in which borrowers makes some kind of change in payment structure/address could be potential default accounts. Moreover, the loans who are 30 days delinquent often are escalated to 60 days delinquent as well.**


Original Interest Rate

Original Loan Term

Original Loan−to−Value Ratio

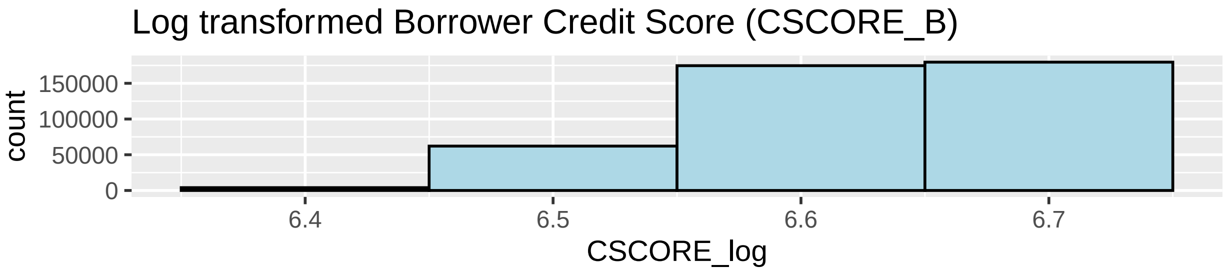
Debt−to−Income Ratio

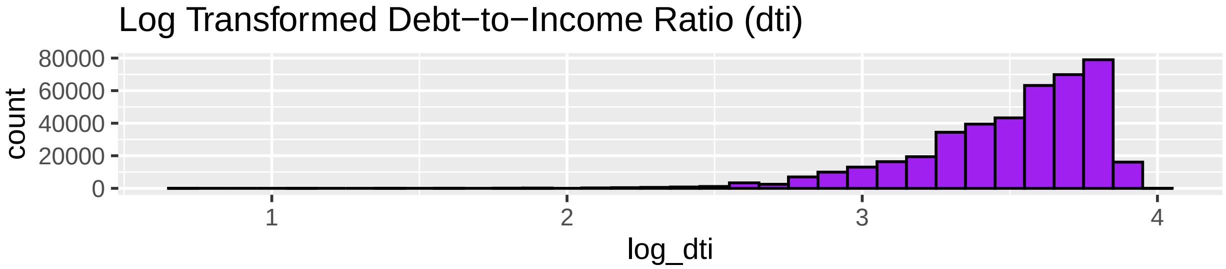
Last Interest Rate

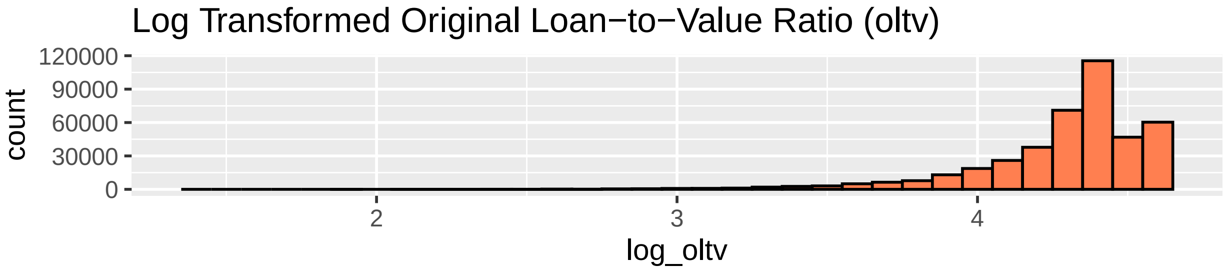
Borrower Credit Score

The numerical variables shows:

- Original loan term (orig_trm): Most of the loans had values concentrated at 360 i.e. 30-year mortgages and years were more generic like 120, 180, 240, 360 hence the numeric variable is converted to categorical variable with loan term greater than 240 as long-term loans and loan term lesser than 240 as short-term loans. The long-term loans were set as reference category.
- For oltv, CSCORE_B and dti, log transformation is applied to compress the tails of these variables making it appropriate for fitting the model. However, there is still some skewness in that data.
- Orig_rt and LAST_RT shows an approximate normal distribution.


Log transformed Borrower Credit Score (CSCORE_B)


Log Transformed Debt−to−Income Ratio (dti)


Log Transformed Original Loan−to−Value Ratio (oltv)

**Model building:**

The dataset had high level of imbalance and logistic regression model gives a bias towards the class with higher frequency. Therefore, assigning weights to minority classes prevent the skewness in the model. This resulted in improved model performance and more informed decision making to avoid major losses.

Variable

VIF Value

orig_rt

orig_rt

7.204893

log_dti

log_dti

1.036462

log_oltv

log_oltv

1.076928

MOD_FLAG

MOD_FLAG

1.329722

LAST_RT

LAST_RT

6.608315

F30_DTE

F30_DTE

1.316207

F60_DTE

F60_DTE

1.379077

relo_flg

relo_flg

1.002282

CSCORE_log

CSCORE_log

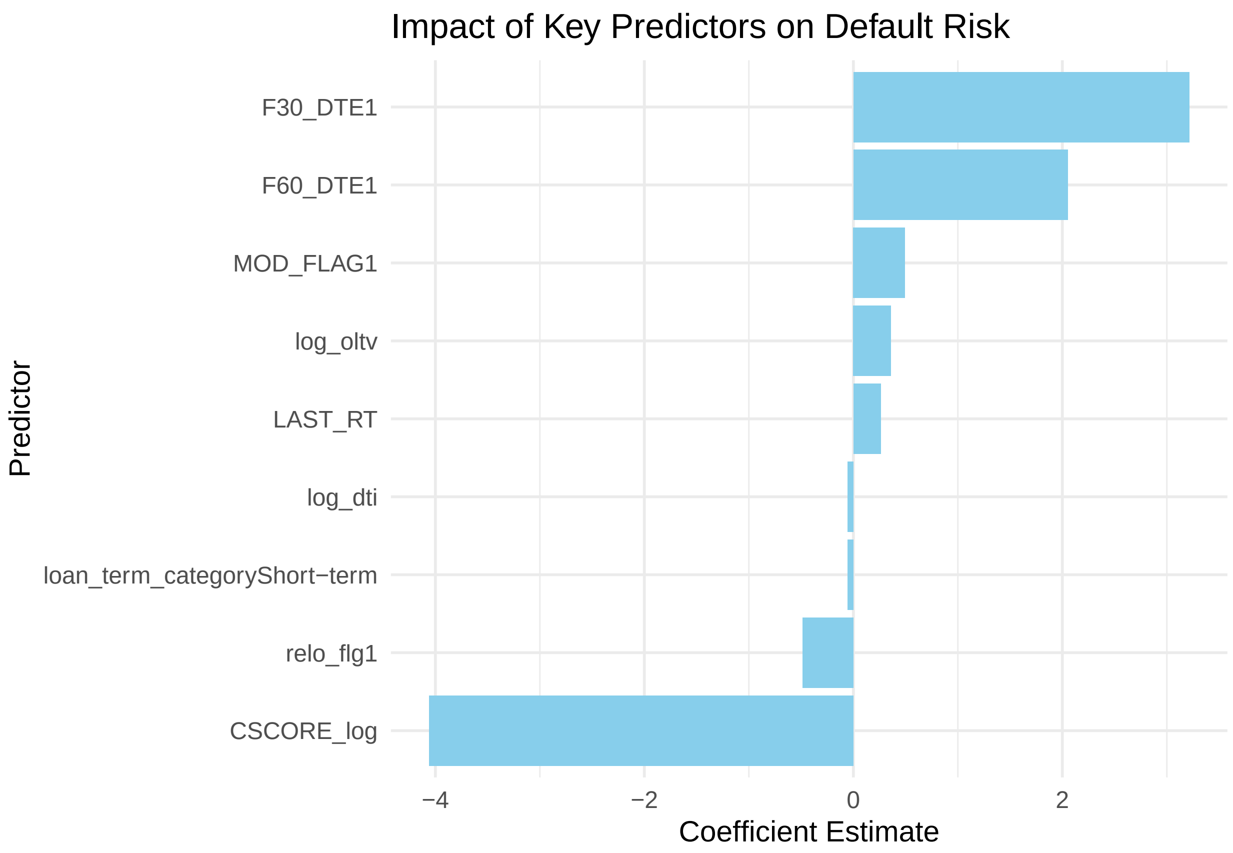
1.211946

loan_term_category

loan_term_category

1.490804

Earlier, correlation plot showed that Orig_rt and LAST_RT has a high correlation of 0.99. After fitting the GLM model and computing Variation Inflation factor (VIF), it was seen that orig_rt and LAST_RT have VIF greater than 5 suggesting multicollinearity. Therefore orig_rt with higher VIF was removed from the dataset to avoid overfitting and inflated standard errors.


Impact of Key Predictors on Default Risk
Predictor
F30_DTE1
F60_DTE1
MOD_FLAG1
log_oltv
LAST_RT
log_dti
loan_term_categoryShort−term
relo_flg1
CSCORE_log
−4
−2
0
2
Coefficient Estimate

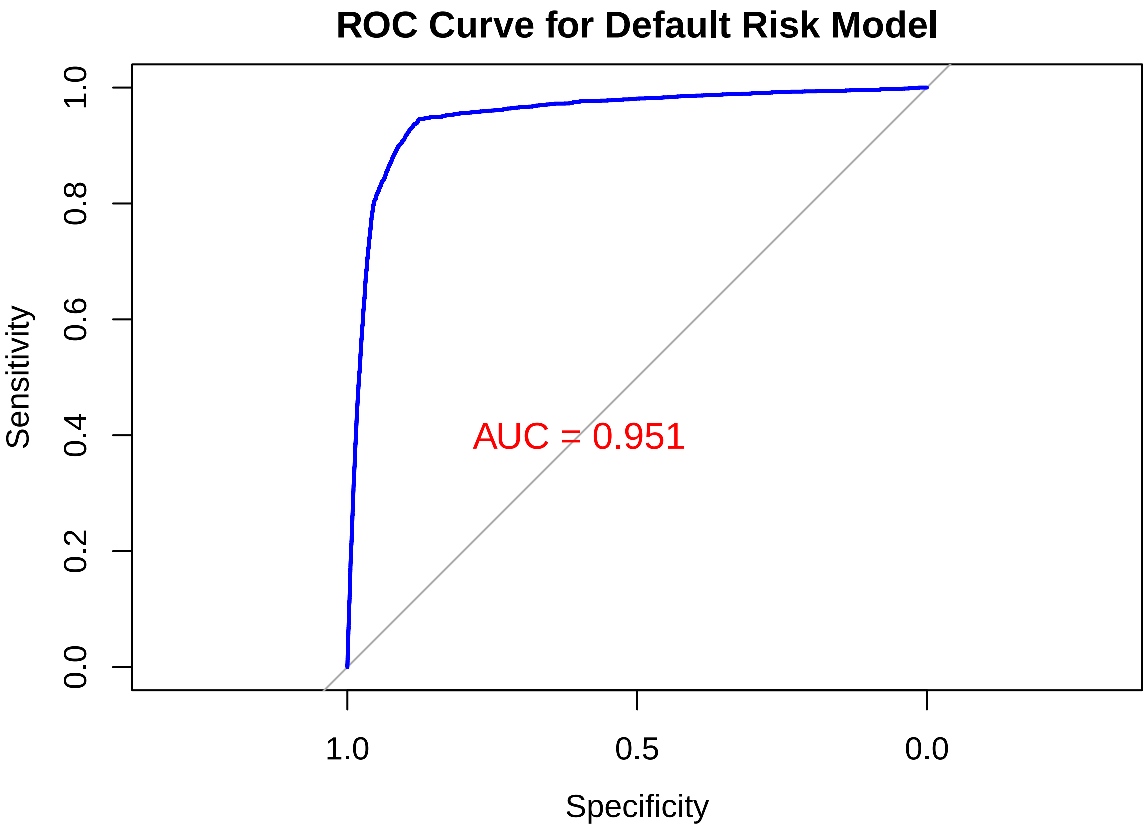


Table 2: Confusion Matrix for Default Risk Model

| | Predicted | |
|---|---|---|
| | Non-default | Default |
| Non-default | 410106 | 2143 |
| Default | 6488 | 1539 |

*Note:*
Accuracy: 0.9795

The final model suggests:

*"Borrowers who are into delinquency (30 days or 60 days) have high chances of default. Moreover, 30 days delinquency is more conclusive than 60 days as it indicates imminent default in the loan. It can be concluded that borrowers who slipped into 30 days past due category could be potential defaulters in future and should be added to flagged loans."*

*"Credit score is a crucial indicator of default risk and indicates that lower credit scores are associated with higher likelihood of default. Hence, it is an essential factor to quantify the creditworthiness of the borrower."*

*"The loans which are modified like changing the address, repaying plans, etc are more likely to default. It could be an outcome of financial hardship and economic loss, making them more prone to missing future payments and slipping into defaults."*

*"The last interest rate has an impact on the likelihood of default. The reason is that higher interest rates can influence the borrower's financial condition leading to delay in payments. However, this is an economy driven factor and not specific to borrowers. It should be addressed from the lender's perspective and store a*

*buffer to account for economy shifts like inflation. It could involve making expected credit loss (ECL) models to analyse the risk factors and decide the buffer."*

*"The loan-to-value ratio is also a significant predictor of default. The loan-to-value ratio measures the amount of the property value that is financed by the loan. If the borrowers hold less equity in the property, it can lead to lack of incentives to pay back either due to financial hardship or economic value of the property."*

*"The debt-to-income ratio has slight effect indicating that increase in the ratio leads to decrease in likelihood of default. However, it is not a significant factor affecting the lender."*

*"The short-term loans have slightly lower likelihood of default compared to long-term loans."*

**Overall, the model is really effective with an accuracy of 97.95% in predicting the default status. The confusion matrix shows that 1539 cases were correctly classified as default and 410,106 loans were correctly classified as non-default.**

## Implications:

The model findings demonstrates a real-world lending scenario which is helpful for the lenders to develop risk mitigation strategies and perform buffer modelling.

Delinquency is an important predictor to flag the loans at early stages. This is a sign to prepare for early intervention like restructuring of loans or closer monitoring.

Credit scores should be considered as a prerequisite to lending as it gives a clear picture of borrowers lending behaviour and helps to mitigate future losses.

Loan-modifications could be a potential red flag borrower, hence allocation resources towards such loans could add significant value to the lenders risk portfolio.

At a firm level, the lenders should conduct a periodic macroeconomic shifts analysis to manage the loan portfolios. Most organisations during COVID failed to tackle the losses due to lack of resources and expertise in credit loss modelling. This led to major loss for big lenders during moratorium on loans.

The financial indicators like Loan to value and debt to income ratios helps the organisation predict the user behaviour and offer less risky loans.

## Conclusion:

This analysis shows the power of predictive modelling in identifying the crucial factors that could impact a lending institution. With an accuracy of about 97%, the model allows the lenders to take early measures and act as a reliable tool to classify loans.

Future research:

The dataset was limited to housing market; however this kind of analysis could be done for other areas like auto loans or personal loans. The analysis can be further extended to include more detailed information of lender like employment history which can result in more granular decision making