Indian Statistical Institute, Kolkata

Development of a comprehensive a Credit Risk Model to Predict Expected Loss in Individual Lending

A dissertation submitted in fulfillment of the requirements for the award of
Master of Technology
in
Cryptology and Security
By:Disha Khutia
CrS2204

Primary Supervisor:
Anand Kumar
Deloitte India

Secondary Supervisor:
Dr. Debrup Chakraborty
Indian Statistical Institute,Kolkata

CERTIFICATE

This is to certify that the dissertation entitled **"Development of a comprehensive a Credit Risk Model to Predict Expected Loss in Individual Lending"** submitted by **Disha Khutia** to Indian Statistical Institute, Kolkata, in fulfillment for the award of the degree of **Master of Technology in Cryptology and Security** is a bonafide record of work carried out by her under my supervision and guidance. The dissertation has fulfilled all the requirements as per the regulations of this institute and, in my opinion, has reached the standard needed for submission.

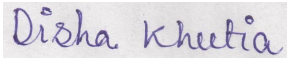
.....

Dr. Debrup Chakraborty,
Indian Statistical Institute, Kolkata

Acknowledgements

This internship has proved to be a great opportunity for me to expand my learning and professional horizon. I consider myself very lucky and honoured to have so many wonderful people lead me through, in this project.

I would like to show my cordial gratitude to my advisors, Anand Kumar, Deputy Manager ,Deloitte and Dr. Debrup Chakraborty, ISI Kolkata, for their guidance and support and encouragement throughout my internship. I would also like to thank my teammates in Deloitte for their valuable suggestions and discussions.They have made my works a lot easier to me with their important suggestions. Finally, I am thankful to my family and friends for their guidance.



Indian Statistical Institute
Kolkata - 700108

Contents

Chapter 1

Introduction

1.1 Overview

During my internship at Deloitte India, I had the opportunity to work on a pivotal project for a client bank, focusing on the development of an advanced credit risk model. My primary responsibility was developing the Probability of Default (PD) model, which is crucial for enhancing the bank's risk management strategies by accurately predicting the likelihood of borrowers defaulting on their loans.

The Probability of Default (PD) model forms a fundamental part of the broader credit risk assessment framework, which also includes Loss Given Default (LGD) and Exposure at Default (EAD). By precisely estimating the PD, the model provides key insights into the potential risk associated with lending to individual borrowers. This, in turn, enables the bank to make informed lending decisions, optimizing the balance between risk and return, and ultimately supporting the institution's financial stability.

The development phase of the model was followed by rigorous validation and continuous monitoring processes to ensure its accuracy, reliability, and robustness. These measures are essential to maintaining the integrity of the model over time, adapting to evolving market conditions, and ensuring its effectiveness in real-world applications.

This project report outlines the methodologies used in developing the credit risk model, discusses the challenges faced during the process, and describes the solutions adopted to address these challenges. Additionally, it emphasizes the crucial role this model plays in improving the bank's decision-making capabilities, thus contributing significantly to its long-term financial health and stability.

1.2 Scope of the project

- Understanding the features ,identifying all potential features that could influence credit risk,Assessing the relevance of each feature. Not all collected features will necessarily contribute to the model.
- Conduct detailed exploratory data analysis to identify patterns, relationships, and key insights in the data.

- Develop and train a logistic regression model, carefully selecting the features to optimize predictive accuracy and robustness of the PD model.
- Evaluate the model using key metrics such as accuracy, ROC-AUC, and confusion matrix, ensuring thorough validation .
- Monitoring the PD model via assessing population stability.

1.3 Key Questions

1. What are the key features that influence a borrower's likelihood of default?
2. How can we preprocess and transform the data to make it suitable for logistic regression modeling?
3. What are the most effective ways to handle categorical variables and continuous variables within the model?
4. How do we ensure the model's predictions are accurate and reliable through appropriate evaluation metrics and validation techniques?

Chapter 2

Background

Credit risk involves the potential for a borrower to default on their loan or fail to make timely repayments, resulting in financial losses for the bank. To manage this risk, traditional banking uses various credit risk management techniques, including credit scoring, underwriting, and loan monitoring. These methods evaluate the creditworthiness of borrowers and ensure that loans are appropriately structured and priced according to the associated risk level.
Credit risk for individual retail borrowers is calculated using a combination of quantitative and qualitative methods.
The primary techniques include:

Credit Scoring Models:
Statistical models, such as FICO or VantageScore, that use historical data and various borrower attributes to predict the likelihood of default. Scores are derived from algorithms that analyze factors like payment history, credit utilization, and length of credit history.

Probability of Default (PD):
Estimating the likelihood that a borrower will default on their loan. This can be calculated using logistic regression models or other predictive modeling techniques. PD is a key input in the calculation of expected loss.

Loss Given Default (LGD):
The portion of the loan that the lender expects to lose if the borrower defaults. It is influenced by the presence and quality of collateral. LGD is calculated as (1 - Recovery Rate), where the Recovery Rate is the proportion of the defaulted loan that can be recovered.

Exposure at Default (EAD):
The total value that the bank is exposed to when the borrower defaults. It includes the outstanding loan balance and any undrawn credit lines. EAD is a critical component in determining the total potential loss.

Expected Loss (EL):
The product of PD, LGD, and EAD, representing the average loss the lender can expect over a certain period.

$$EL = PD \times LGD \times EAD$$

By carefully analyzing these components, banks can effectively assess the credit risk associated with individual retail borrowers, ensuring that loans are appropriately structured and priced according to the risk level. This comprehensive evaluation helps in mitigating potential losses and maintaining financial stability.

Chapter 3

Approach

3.1 Why Logistic Regression

Initially, we implemented logistic regression for the following reasons.It provides clear and interpretable results, ideal for binary classification tasks such as predicting defaults, and it offers probabilistic outputs, aiding in nuanced decision-making for risk management.

- **Interpretability:**
  - Logistic regression provides clear and interpretable results. The coefficients can be directly understood in terms of the log-odds of the default probability, which is crucial for stakeholders who require transparency in model decisions.
  - It allows us to interpret the impact of each predictor variable on the probability of default, aiding in better understanding and trust in the model.
- **Binary Outcome Suitability:**
  - The problem of predicting defaults is inherently binary (default vs. no default). Logistic regression is naturally suited for binary classification tasks, making it an ideal choice for this application.
- **Probabilistic Output:**
  - Logistic regression produces probabilities as outputs, providing more nuanced information compared to models that only provide class labels. This probabilistic output is essential for risk management, enabling more informed decision-making.
- **Implementation Simplicity:**
  - The implementation of logistic regression is straightforward, and it is supported by numerous statistical and machine learning libraries. This ease of implementation allows for rapid development and deployment.
  - Its simplicity also makes it easier to debug and validate, ensuring that the model performs as expected.

3.2 Input Data

The dataset under consideration contains extensive information on over 800,000 consumer loans issued by the bank, spanning from 2014 to 2022. This dataset encompasses a wide range of attributes, offering a comprehensive perspective on each loan and borrower.

For model development, we are partitioning the data into training and testing sets. Data from 2014 to 2020 will be used for training the model, while data from 2020 to 2022 will be reserved for testing its performance.

Below is a detailed description of the most significant columns in the dataset.

1. **Loan Information**
    - **Funded Amount**: The total amount funded for the loan.
    - **Term**: The loan term in months (either 36 or 60 months).
    - **Interest Rate**: The interest rate assigned to the loan.
2. **Lender Information**
    - **Grade**: External credit agency grade for the loan, ranging from A to G.
    - **Sub-Grade**: More granular rating within each grade.
    - **Loan Status**: Indicates the current status of the loan (e.g., fully paid, charged off).
3. **Borrower Information**
    - **Employment Length**: Number of years the borrower has been employed.
    - **Home Ownership**: The home ownership status of the borrower (e.g., RENT, OWN, MORTGAGE, OTHER).
    - **Annual Income**: Self-reported annual income of the borrower.
    - **Purpose**: The purpose of the loan (e.g., debt consolidation, home improvement).
    - **Debt-to-Income Ratio (DTI)**: Ratio of the borrower's total monthly debt payments to their monthly income.
    - **Address**: Borrower's address information, including state and zip code.
    - **Credit Inquiries**: Number of credit inquiries in the past 6 months.
    - **FICO Range**: The borrower's FICO score range at the time of loan origination.
4. **Repayment and Recovery Information**
    - **Recoveries**: Amount recovered after the borrower has defaulted.
    - **Total Recovered Principal**: Total principal amount recovered.
    - **Collection Recovery Fee**: Fees associated with collecting on the defaulted loan.

3.3 General Pre-processing

We will begin the preparation and pre-processing of the data specifically for building the Probability of Default (PD) model. This model will utilize logistic regression to predict the likelihood of a borrower defaulting on their loan.

3.3.1 Handling Missing Values

Missing values of the following features were handled by filling them with 0 ,because number of missing value present in the dataset was extremely low.

- `months_since_earliest_cr_line`
- `acc_now_delinq`
- `total_acc`
- `pub_rec`
- `open_acc`
- `inq_last_6mths`
- `delinq_2yrs`
- `emp_length`

The missing values of the feature **"total revolving high credit/credit limit"** was replaced by **"funded_amount"** which can serve as a proxy for the credit limit, as the funded amount might be closely related to the borrower's credit limit.

3.3.2 Converting Data Types

This steps involved converting text-based numeric and date variables into actual numeric values,then calculating the time difference with reference date 2023-12-01 in months. This step is done for the features "term","emp_length".

3.3.3 Converting Date Variables

`date_of_earliest_credit_line` and `loan_issue_date` were converted from string representations to numeric values representing the number of months since these events.

Example: Calculating `months_since_earliest_cr_line`

- Converted the string to a timestamp using `pd.to_datetime` with a specified format.
- Calculated the difference between the timestamp and a reference date (e.g., December 2023).
- Converted the difference from days to months .

3.3.4 Dummy Variables for Categorical Features

Categorical variables need to be converted into dummy variables to be used in predictive models.
One hot encoding was done on **Grade,sub_grade,home_ownership,addr_state, verification_status,loan_status,purpose,,initial_list_status.**.For example:

```
loan_data_dummies = pd.get_dummies(loan_data['loan_status'],
                                   prefix='loan_status',
                                   prefix_sep=':')
```

This results in a DataFrame with columns like:

- loan_status:Charged Off
- loan_status:Current
- loan_status:Default
- loan_status:Does not meet the credit policy. Status:Charged Off
- loan_status:Does not meet the credit policy. Status:Fully Paid
- loan_status:Fully Paid
- loan_status:In Grace Period
- loan_status:Late (16-30 days)
- loan_status:Late (31-120 days)

Each row in the DataFrame will have a 1 in one of these columns, indicating the presence of the respective category, and 0 in the others.

3.3.5 Dependent Variable: Good/ Bad(Default)

According to the loan status of the individuals we have classified them as good and bad borrowers(means they might default).Individuals with loan status as **'Charged Off','Late (31-120 days)','Does not meet the credit policy. Status Off'** are identified as bad and given the binary value 0 and rest of them are good borrowers with binary value 1.

3.4 Fine classing, coarse classing

Fine and coarse classing are two approaches to binning or discretizing continuous variables. These approaches differ in the level of granularity or detail in which the continuous variable is to be divided into bins or categories.

3.4.1 Fine classing

Fine classing involves dividing the continuous variable into a larger number of smaller bins or categories, resulting in a more detailed representation of the data. This approach provides a higher level of granularity and can capture finer patterns or variations within the data. Fine classing is suitable when the data has a wide range and requires a more precise representation.

3.4.2 Coarse classing

Coarse classing, on the other hand, involves dividing the continuous variable into a smaller number of larger bins or categories, resulting in a less detailed representation of the data. This approach provides a broader overview and simplifies the data by grouping similar values together. Coarse classing is suitable when the data has a large range and a higher level of detail is not necessary or practical. Monotonic binning is a technique commonly used in credit scoring to discretize continuous variables while preserving their monotonic relationship with the target variable (e.g., default or non-default). The objective is to create bins that exhibit a consistent trend or monotonic pattern with the target variable across the range of values.

3.4.3 Bin evaluation

In credit risk assessment, Information Value (IV) serves as a crucial metric for evaluating the predictive power of variables within different bins or categories. IV quantifies how well an independent variable distinguishes between good and bad credit risks across these bins.

$$\text{IV:} \qquad \sum_{i=1}^{n} (\text{Distr Good}_i - \text{Distr Bad}_i) * ln\left(\frac{\text{Distr Good}_i}{\text{Distr Bad}_i}\right)$$

Higher IV values indicate that a variable and its corresponding bins effectively discriminate between borrowers likely to default and those who are not. This information helps in prioritizing variables for model development and feature selection.

By convention, the values of the IV statistic in credit scoring can be interpreted as follows:

Information Value	Variable Predictiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive Power
0.1 to 0.3	Medium predictive Power
0.3 to 0.5	Strong predictive Power
>0.5	Suspicious Predictive Power

We then transform all the independent variables using the weight of evidence (WoE) method.It quantifies the predictive power of each bin or category by comparing the proportion of defaulters (bad applicants) to non-defaulters (good applicants). It seeks to

establish a monotonic relationship between the independent variables and the likelihood of default.

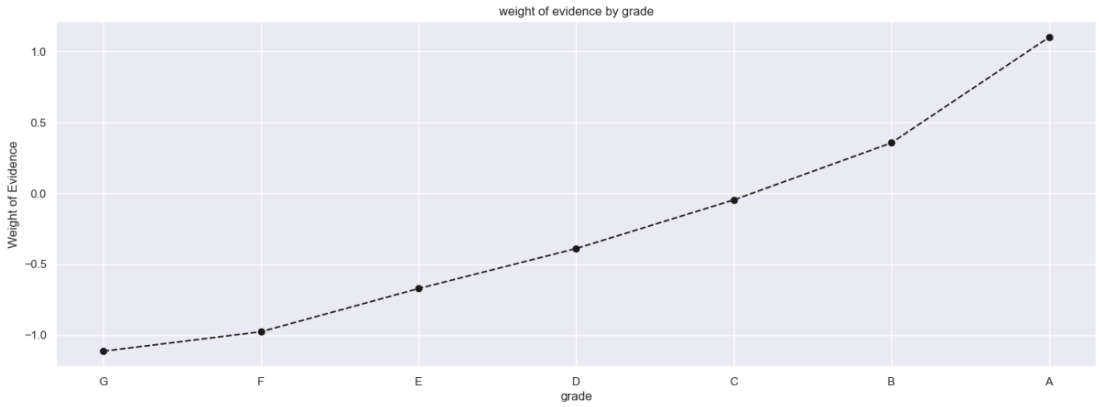
$$\text{WoE:} \qquad [\, ln\left(\frac{\text{Distr Good}}{\text{Distr Bad}}\right)] \times 100.$$

3.4.4 Creating dummies:discrete

Based on WOE we decide how to organize the original categories of the discrete variables into the dummy variables of PD model.

Weight of Evidence Analysis for the Grade Variable

The grade variable reflects the external credit ratings of all variables. It is essential for determining the credit risk associated with each loan. The Weight of Evidence (WoE) for the grade variable should exhibit very good patterns, providing clear insights into the risk levels.



- As illustrated in the chart, the Weight of Evidence (WoE) increases almost monotonously with the increase in external credit rating grades, from the worst grade 'G' to the best grade 'A'.
- This pattern indicates that loans with higher external ratings are generally better on average. Specifically, the greater the grade, the higher the WoE.
- We put grade:G as reference category.Which will be removed later to avoid dummy variable trap.

Weight of Evidence Analysis for the addr_state Variable

we tackle the more complex address_state variable, which has 51 categories including all U.S. states and Washington D.C. We calculate the weight of evidence for each state, revealing significant variations.

- For instance, Nebraska and Iowa show much lower weights, while Maine and Idaho exhibit higher ones. We propose combining states with similar weights into broader categories, such as grouping states with the highest risk like North Dakota, Nebraska, and Iowa.
- Conversely, states like California and New York, which have many observations and unique weights, are kept in separate categories.
- This method ensures a conservative risk management approach, particularly when data is sparse or absent.
- Finally, we create dummy variables for these new categories, integrating them into our regression model to better manage and predict credit risk.

In similar way coarse classing on "verification_status", "purpose","initial_list_status" was done.

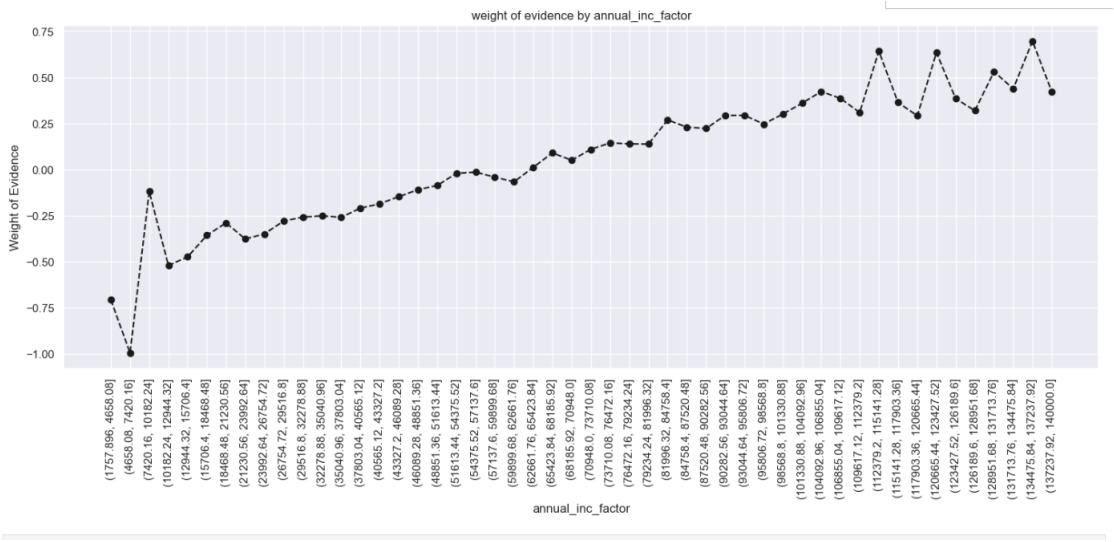
3.4.5 Creating dummies:continuous

The PD model must be interpretable, using only dummy variables as predictors.

- To handle continuous variables, we first split them into categories, similar to how discrete variables were handled.
- Start by performing fine classing to create many categories encompassing the range of continuous variable values.
- we treat continuous variables similarly to discrete ones but ensure their categories are ordered quantitatively to maintain the integrity of the data's natural order.

This approach ensures that the PD model remains interpret-able and compliant with regulatory requirements.

Weight of Evidence Analysis for the Annual Income Variable



- Initially attempted with 50 categories but found that insufficient due to large imbalances in the number of observations per category.
- Adjusted to 100 categories but still found imbalance issues, particularly with a large number of observations in the lower income categories.
- Created a category for annual incomes greater than $140,000, representing about 5% of observations.
- For incomes less than or equal to $140,000, further categorized into narrower income ranges (e.g., $20,000 intervals from $20,000 to $100,000, specific categories for$100,000 to $120,000 and $120,000 to $140,000).

  Creating Dummy Variables: Utilized np.where method to create dummy variables based on income categories for modeling purposes.

similarly created dummies for other continuous variables like "DTI", "months_since_last_record.

3.5 Selecting the features(escaping the dummy variable trap)

To avoid the dummy variable trap and simplify our model, we have deleted reference categories from our set of dummy variables . The dummy variable trap occurs when all dummy variables for a categorical feature are included, leading to multicollinearity and unstable coefficients. By excluding one dummy variable for each original categorical variable as the reference category, we ensure the model remains stable and interpretable.

3.6 Model Training

Using logistic regression, we train a model to predict the probability of borrower default based on historical data of borrower characteristics and loan outcomes. The model calculates the log-odds of default probability, interpreting how each input feature influences

this likelihood. After training, the model provides straightforward probabilities ranging from 0 to 1, facilitating easy interpretation for stakeholders unfamiliar with data science.

3.7 Model Performance

To evaluate the accuracy of our logistic regression model, we need to determine how well it classifies good and bad borrowers. We do this by applying a cutoff probability, TR, to the estimated probabilities. Borrowers with estimated probabilities less than or equal to TR are classified as bad (0), and those with probabilities greater than TR are classified as good (1).
Initially, we set TR to 0.5 and use a **confusion matrix**, which provides detailed insights into the model's classification abilities. The **confusion matrix** shows the number of true positives (good borrowers correctly classified), true negatives (bad borrowers correctly classified), false positives (bad borrowers incorrectly classified as good), and false negatives (good borrowers incorrectly classified as bad).

For example, with a threshold of 0.5, the model achieves an overall accuracy of 89.07%, but it also generates many **false positives**, meaning many bad borrowers are incorrectly classified as good. This could lead to granting loans to risky borrowers.

To mitigate this, we set a more conservative threshold, such as 0.9. At this threshold, the model correctly identifies more bad borrowers but **reduces the number of true positive classification**s, which could result in fewer approved loans overall.

ROC Curve and AUC

Overall accuracy is not the only metric to consider. The rates of true positive predictions (sensitivity) and false positive predictions (specificity) are crucial. We can visualize these rates using the **Receiver Operating Characteristic (ROC) curve,** which plots the true positive rate against the false positive rate for various thresholds. The Area Under the Curve (AUC) measures the model's performance; an AUC of **70.21%** indicates our model is fair.
A good PD model minimizes the risk while maximizing the number of approved loans. Thus, different thresholds and multiple competing models need to be tried to find the best balance between accuracy and business needs.

3.8 Scorecard

3.8.1 Interpreting PD Model Coefficients

In a logistic regression model for predicting the probability of default (PD), each coefficient represents the change in the logarithm of the odds of the outcome when the corresponding dummy variable changes from 0 to 1. For a dummy variable x with coefficient β, the difference in the log-odds between $x = 1$ and $x = 0$ is β, and the exponentiated coefficient e^{β} represents the ratio of the odds for $x = 1$ to the odds for $x = 0$.

For example, if the coefficient for a dummy variable indicating higher education is 0.24, the log-odds of non-default for borrowers with higher education compared to those without is 0.24, and the odds ratio $e^{0.24} \approx 1.27$, meaning the odds of non-default for borrowers with higher education are 1.27 times (or 27%) higher than for those without higher education.

interpretting the PD model and estimate the probability of default (PD) and non-default for individual accounts is achieved by applying the `predict` method to the test dataset. Let's quickly revise the process.
First, we multiply the values of a borrower on the independent dummy variables by the respective model coefficients. Raising an exponent to this result gives the odds for being good versus bad. From there, the estimated probability of being good can be calculated. For each observation from all dummy categories of the original independent variable, only one of the dummy variables can have a value of one; the rest are zero. Therefore, multiplying by zero results in zero, and multiplying by one yields the same result. Thus, calculating the power to which the exponent should be raised to obtain the odds is simplified to summing the regression coefficients for all dummy categories to which an observation belongs.

Practical Example

We take an observation (index 362514):

- Intercept: -1.338497
- External grade C: 0.689618
  Sum: -1.338497 + 0.689618 = -0.648879
- Home ownership (Mortgage): 0.106222
  Sum: -0.648879 + 0.106222 = -0.542657
- Location (California): 0.064466
  Sum: -0.542657 + 0.064466 = -0.478191
- Verification status (Verified, reference category): 0
  Sum remains: -0.478191
- Loan purpose (Major purchase, car, home improvement): 0.265831
  Sum: -0.478191 + 0.265831 = -0.21236

Continuing this process until all relevant coefficients are added results in a final sum of -2.50279. These are the log odds. Raising an exponent to this power, we get the odds: $e^{-2.50279} = 12.216531$.
The estimated probability of being a good borrower is:

$$\frac{12.216531}{12.216531 + 1} = 0.924337$$

Therefore, the probability of non-default is 92.43%.

3.8.2 Creating a Scorecard

PD models are often turned into scorecards for easier interpretation. We store the coefficients of our PD model in a `summary_table` DataFrame. It contains regression coefficients for all independent dummy variables except for the reference categories.
To make the scorecard interpretable, we include the names of the original independent variables, which can be extracted from the dummy variable names.
We aim to create a scorecard with a minimum score of 300 and a maximum score of 850, similar to a typical FICO score. This requires rescaling the creditworthiness assessment produced by our model.

Rescaling to Credit Scores

To rescale the coefficients to scores:

$$score = \left(\frac{coefficient - minimumsumofcoefficients}{maximumsumofcoefficients - minimumsumofcoefficients} \right) \times (850-300)+300$$

This transformation ensures that the scores range from 300 to 850.

Scorecard Example

For instance, the score corresponding to the intercept is around 311.87, close to the minimum desired score of 300. We adjust and round scores for ease of interpretation.

3.8.3 Calculating Individual Credit Scores

Calculating an individual credit score involves summing the scores corresponding to the respective dummy categories. For example, for the borrower with index 362514:

- Intercept: 312
- External grade C: 53
  Sum: 312 + 53 = 365
- Home ownership (Mortgage): 8
  Sum: 365 + 8 = 373
- Location (California): 5
  Sum: 373 + 5 = 378
- Continue similarly for all relevant categories, resulting in a final score of 608.

3.8.4 Setting Cutoffs for Loan Approval

To decide who gets a loan, we set a cutoff point based on the estimated probabilities of default. A specific cutoff impacts both the number of approved/rejected borrowers and the quality of the loans granted.
We can also set cutoffs using credit scores, which are more interpretable. For example, a cutoff score of 585 results in an approval rate of 53.86
.

Chapter 4

Monitoring the PD model

4.1 Population Stability Index (PSI) Analysis

4.1.1 Overview of PSI

The Population Stability Index (PSI) is a metric used to evaluate how much a new data population has deviated from the original data population on which a predictive model was trained. It is essential to ensure that the model remains accurate and reliable over time.
PSI is calculated using the following formula:

$$PSI = \sum (ActualProportion - ExpectedProportion) \times \log\left(\frac{ActualProportion}{ExpectedProportion}\right)$$

A PSI value interpretation is as follows:

- $PSI < 0.1$: Little or no change, model remains reliable.
- $0.1 \leq PSI < 0.25$: Moderate change, further investigation required.
- $PSI \geq 0.25$: Significant change, model redevelopment likely needed.

4.1.2 Data Used for PSI Calculation

Two datasets were utilized for this analysis:

- **Original Population:** Data used to build the initial PD model.
- **New Population:** Data from new loan applicants in 2023.

Both datasets were preprocessed to ensure they have the same structure and features, facilitating a valid comparison.

4.1.3 PSI Calculation Results

The PSI was calculated for various features as well as the dependent variable (score). Below are the findings:

$PSI < 0.1$ **(Stable)**

- Accounts Now Delinquent
- Address State
- Annual Income
- Home Ownership
- Inquiries in the Last 6 Months
- Interest Rate
- Months Since Earliest Credit Line

These features exhibited little to no change, indicating stability between the original and new populations.

Features with 0.1 $\leq PSI < 0.25$**(Moderate Change)**

- Debt to Income (PSI = 0.08, close to 0.1 but still considered stable)

Features with PSI ≥ 0.25**(Significant Change)**

- Initial List Status (PSI = 0.33)
  - Original population: 35% W category, 65% F category.
  - New population: 63% W category, 37% F category.
- Months Since Issue Date (PSI = 0.3)
  - Reflects the natural time difference between original and new datasets rather than borrower characteristics.

Dependent Variable (Score)

- PSI = 0.3
  - Significant difference indicating a notable change in the distribution of credit scores between the original and new populations.

4.1.4 Implications of PSI Results

The substantial differences identified, especially in the Initial List Status and Months Since Issue Date, suggest a need for further investigation. These shifts may not solely be due to changes in borrower characteristics but could also reflect changes in operational practices or the passage of time.
The PSI value for the dependent variable (score) exceeding 0.25 is particularly concerning, indicating that the model's output distribution has changed significantly. This suggests the PD model may no longer be as reliable as it was when first developed.

4.1.5 Recommendations

Based on the PSI analysis, the following steps are recommended:

- Conduct a detailed investigation into the changes observed in Initial List Status and Months Since Issue Date to understand the root causes.
- Evaluate the impact of these changes on the model's performance and outcomes.
- Consider redeveloping the PD model with the new data to maintain its predictive accuracy and reliability.
- Implement a regular schedule for PSI analysis to continually monitor population stability and ensure ongoing model performance.

The PSI analysis has highlighted significant changes between the original and new populations, particularly in terms of credit scores and certain features. These findings necessitate further investigation and potentially redeveloping the PD model to ensure it continues to provide accurate risk assessments.

Chapter 5

Conclusion & Future Work

5.1 Conclusion

In this project, we developed a Probability of Default (PD) model using logistic regression, focusing on ensuring the model's interpretability and compliance with regulatory requirements. Our approach involved pre-processing both discrete and continuous variables, transforming them into categorical dummy variables that facilitate straightforward interpretation.
For discrete variables, we fine classed them, then evaluated and grouped them based on their weights of evidence (WoE). This step ensured that the categories were representative of varying levels of credit risk. We plotted these WoE values to visually inspect and validate our groupings.
Handling continuous variables required an additional step: fine classing them into categories that reflected their quantitative differences. We adapted our existing functions to calculate and plot the WoE for these categories while maintaining their natural order. This adaptation allowed us to apply a consistent methodology across both discrete and continuous variables, ensuring the model's robustness and interpretability.
The logistic regression model was trained using these dummy variables, providing clear insights into how different factors contribute to the probability of default. The weight of evidence plots, along with the logistic regression coefficients, allowed us to identify and quantify the impact of various predictors on default risk.
Overall, our model demonstrated strong predictive performance while adhering to the regulatory requirement of interpretability. By transforming variables into dummy categories and using weights of evidence, we ensured that the model's predictions are not only accurate but also transparent and easy to understand. This approach enhances the model's utility for risk management and decision-making in credit assessment.

5.2 Future Work

While the development of the Probability of Default (PD) model using logistic regression has provided a robust and interpretable framework, there are several areas for future work that can enhance and extend the model's capabilities:

- **Feature Engineering and Selection:** Explore additional features that might have predictive power in determining defaults. This can include macroeconomic indicators, borrower behavior data, or alternative data sources.

- **Model Comparison and Ensemble Methods:** Compare the logistic regression model with other machine learning algorithms such as Decision Trees, Random Forests, Gradient Boosting Machines, and Neural Networks to assess performance improvements. Investigate the use of ensemble methods, which combine multiple models to improve prediction accuracy and robustness. Techniques such as stacking, bagging, or boosting could be particularly beneficial.

- **Temporal Stability and Model Updating:** Ensure the model remains accurate over time by implementing a framework for periodic retraining and validation using the latest data. This will help in capturing any shifts in the underlying data distribution and maintaining model performance. Analyze the temporal stability of features to understand how their predictive power changes over time and adjust the model accordingly.

- **Integration with Business Processes:** Integrate the PD model into existing credit risk management systems and workflows to enable seamless decision-making. Develop user-friendly dashboards and reporting tools that visualize the model's predictions and provide actionable insights for business users.

- **Regulatory Compliance and Ethical Considerations:** Enhance the model by incorporating scenario analysis and stress testing capabilities to evaluate how economic shocks or changes in borrower behavior impact the probability of default. Use these insights to develop strategies for mitigating potential risks under different economic conditions.

References

[1] https://www.eba.europa.eu/regulation-and-policy/model-validation/guidelines-on-pd-lgd-estimation-and-treatment-of-defaulted-assets

[2] https://www.bis.org/bcbs/index.htm

[3] Deloitte Training Sessions.