Credit Scoring

Table of Contents

1. Introduction

Credit is a very important product in banking and financial institutions. There is always a customer in need of a loan. Since Loans are always accompanied by risks, it is important to identify suitable applicants, and there has to be a means to determine and separate the good applicants from the bad. To solve this issue, financial institutions such as banks started developing credit scores. Using the customer’s credit scores lenders can define the risk of loan applicants. By calculating the credit score, lenders can make a decision as to who gets credit, would the person be able to pay off the loan and what percentage of credit or loan they can get.

Logistic regression can be used to predict default events and model the influence of different variables on a consumer’s creditworthiness. In this paper we use a logistic regression model to predict the creditworthiness of bank customers using predictors related to their personal status and financial history. We also compare and evaluate this algorithms based on the original dataset and transformed dataset using Weighted of Evidence to draw conclusions.

2. Theoretical Background

2.1. Logistic Regression

Logistic Regression sometimes called the logistic model or logit model, analyzes the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve.There are two models of logistic regression, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical. When the dependent variable is not dichotomous and consists of more than two categories, a multinomial logistic regression can be employed.

Advantages of logistic regression

1. The main advantage of logistic regression is that it is much easier to set up and train than other machine learning and AI applications.
2. Another advantage is that it is one of the most efficient algorithms when the different outcomes or distinctions represented by the data are linearly separable. This means that you can draw a straight line separating the results of a logistic regression calculation.
3. One of the biggest attractions of logistic regression for statisticians is that it can help reveal the interrelationships between different variables and their impact on outcomes. This could quickly determine when two variables are positively or negatively correlated, such as the finding cited above that more studying tends to be correlated with higher test outcomes.

2.2. Weight of Evidence

This method is commonly used alongside Logistic Regression for modeling Probability of default. WOE assesses the amount of information each attribute (category) of an independent variable has in predicting the class of a target variable. Mathematically, it is the natural log of the ratio of percentage distribution of non-defaulting customers to percentage of defaulting customers.

WOE = $ln(\frac{\%\, non-defaulting}{\%\, defaulting})$

Benefit of WOE

- It can treat outliers. These values (outliers) would be grouped to a class of .
- It can handle missing values as missing values can be binned separately.
- Since WOE Transformation handles categorical variables, there is no need for dummy variables.
- WoE transformation helps you to build a strict linear relationship with log odds. Otherwise it is not easy to accomplish a linear relationship using other transformation methods such as log, square-root etc.

2.3. Evaluation techniques

In order to evaluate the performance of a Machine Learning model, there are some metrics to know its performance and are applied for Regression and Classification algorithms.

Area Under the ROC Curve (AUC - ROC)

AUC is the area under the receiver operating characteristic curve, which measures the ability of a classifier to distinguish between classes across different threshold values. In consumer credit scoring, the area under the receiver operating characteristic curve (AUC) is one of the most commonly used measures for evaluating predictive performance.

Confusion matrix

Confusion matrix (average correct classification rate criterion) is one of the most widely used criteria in the area of accounting and finance (for credit scoring applications) in particular, and other fields, such as marketing and health in general. The average correct classification rate measures the proportion of the correctly classified cases as good credit and as bad credit in a particular data-set. A classification matrix presents the combinations of the number of actual and predicted. It consists of values like True Positive, False Positive, True Negative, and False Negative, which helps in measuring Accuracy, Precision, Recall, Specificity, Sensitivity, and AUC curve.

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

For this problem, false negatives are more severe than false positives, as it is more detrimental to the bank if a customer defaults, than for it to lose a customer that would not have defaulted. Also, a human can be in the loop to verify positive cases, hence the model can act as an initial filter for possible defaulters. As such, recall is more important than precision. Besides, poor metrics to use include accuracy, which is unsuitable due to the imbalanced nature of the target variable.

F-measure

F-measure is a best measure of the Test accuracy of the developed model. Higher the F1 Score, better the performance of the model. Using only recall will cause the model to be imprecise. To balance precision with recall, we can use the F-measure, including the F1 score, which is the Harmonic mean of Recall and Precision. It makes our task easy by eliminating the need to calculate Precision and Recall separately to know about the model performance.

3. Experimental Setup

3.1 Exploratory data analysis

3.1.1. Data collection

The dataset contains information of 150.000 customers. Each observation is characterized by 11 attributes. This is a moderate sized dataset with few features (more long than wide). Features can be classified into:

- Historical late repayments in the last 2 years(3 window periods of 30-59, 60-89,>=90)
- Financial Obligations (NumberOfOpenCreditLinesAndLoans, NumberOfDependents, DebtRatio, RevolvingUtilizationOfUnsecuredLines)
- Financial Capabilities (MonthlyIncome)
- Demographics (Age)

The columns provided in the dataset are as followed:

Variable Name	Description	Type

SeriousDlqin2yrs	Person experienced 90 days past due delinquency or worse	Y/N
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	percentage
age	Age of borrower in years	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times the borrower has been 30-59 days past due but no worse in the last 2 years.	integer
DebtRatio	Monthly debt payments, alimony,living costs divided by monthly gross income	percentage
MonthlyIncome	Monthly income	real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
NumberOfTimes90DaysLate	Number of times the borrower has been 90 days or more past due.	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times the borrower has been 60-89 days past due but no worse in the last 2 years.	integer
NumberOfDependents	Number of dependents in a family excluding themselves (spouse, children etc.)	integer

3.1.2 Imbalanced dataset

From the distribution of the target variable, this problem is a binary classification problem of classes 0 or 1 denoting if the customer will default in 2 years, with class 1 being the minority at 6.684%. This is an imbalanced learning problem.There are many techniques to overcome imbalanced datasets in classification, either by oversampling the minority class or undersampling the majority class. We also apply both of them to balance the dataset but it doesn’t bring good results. Therefore, we decided to keep the bias nature of the dataset towards a particular class (0) and using Precision, Recall, F1-score and AUC are the metrics to evaluate our predictive models.

3.1.3 Null values

Many learning algorithms cannot handle missing values, hence we need to handle missing values by dropping these observations, or imputing them. Imputing will increase the bias in our model, but dropping too many observations leads to a much smaller dataset for training. In our dataset, only 2 features have nulls (Monthly income and number of dependents). This could be due to customers not wanting to declare this personal information.

Our dataset has 29k observations with >= 1 feature having null values, corresponding to 20% of the dataset, which is quite significant. We will have to impute these null values rather than dropping observations with null. Monthly Income and NumberOfDependents have right tail skewed, so median imputation is preferred to be robust to outliers.

3.1.4 Outliers

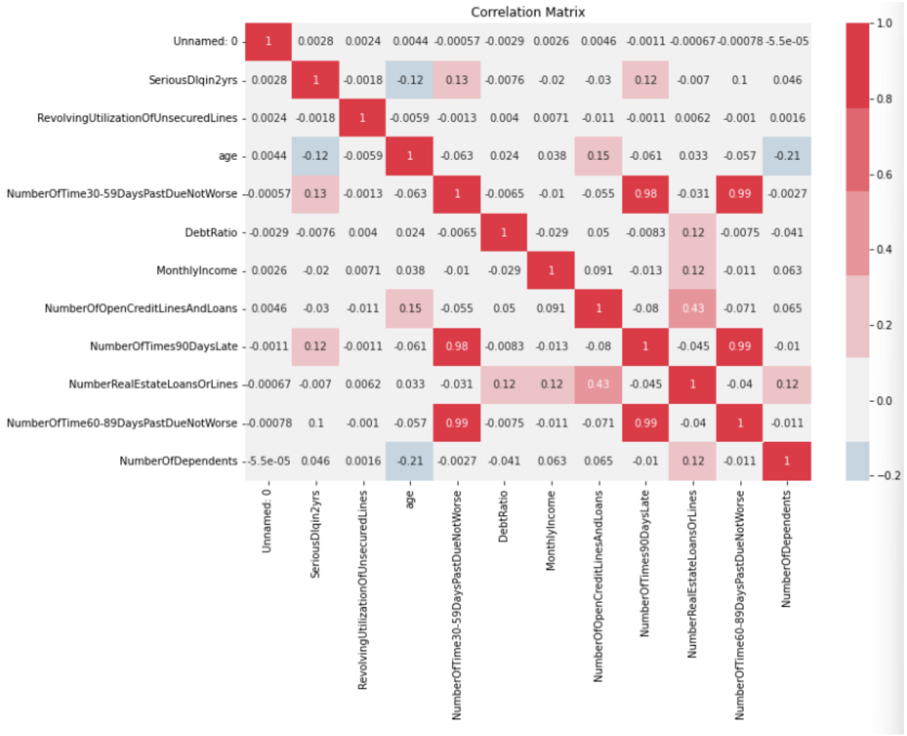
Many of the financial features have outliers (extreme outliers = > 3 x interquartile range), with a right-tail skew. Hence, we should either use models that are robust to outliers, or transform the features accordingly (e.g. Box-Cox transformation). Because this article uses Logistic Regression, we will use Box-Cox transformation features to detect outliers.

3.1.5 Duplicated

The dataset has not too many duplicates, hence will not likely bias the model much. These duplicates are only a problem if they came from non-random errors, e.g. duplicate entry into database. In a real scenario, we should dig deeper to identify if these are errors or coincidences. For this problem, we will not remove these duplicates and assume they are correct.

3.1.6 Correlation

From the Correlation matrix, we can see that 30, 60 and 90 days late payment features have slight positive correlation with the target variable (0.12), while age has slight negative correlation with the target variable (-0.12). These are likely useful features for predicting the target. Additionally, 30, 60 and 90 days past due are highly correlated with each other (close to 1). This may impact performance if the learning algorithm used assumes independence in dependent variables. We can either use models that are robust to multicollinearity or use feature selection / regularization methods to use just one of these features. We also try to apply regularization methods to combine Late repayment features by equal sum or weighted sum but it gives bad predicted results. Therefore, we keep original features in our model.



3.2 Feature engineering

In this section, we visualize and analyze each feature to decide which transformation method will be applied in our model.

Age

Age tends to have a somewhat reasonable distribution. There are a suspicious number of centenarians but plausible. The only certainly incorrect data is that there is one person in the dataset with age 0, and because infants are not legally permitted to take out loans, we will impute that to the next youngest person in the dataset.

RevolvingUtilizationOfUnsecuredLines

Approximately 98% of values of this Variable are between 0 and 1 with a well defined right-skewed distribution. Generally, Credit Utilization is expected to be within this region (0 - 1). Although, borrowers can sometimes spend beyond the credit limit. Values between 1 and 10 make up 2% of the dataset. Values beyond 10 are extremely big and they make up less than 0.5% of our data, these values would be dropped to prevent them from impacting our model.

Debt ratio

3750 records with DebtRatio > 3,500, only 185 of them have a value for monthly income. Further, the people who do have monthly income seem to either have a monthly income of either 1 or 0. 164 of them have the same value for 2 year default rate and monthly income, indicating that there is a data-entry error. And, despite owing thousands of times what they own, these people aren't defaulting any more than the general population. We can conclude that these entries must be data-entry errors, so we will remove them from our model.

Late Repayment
There are 267 instances where the three columns NumberOfTimes90DaysLate, NumberOfTime60-89DaysPastDueNotWorse, NumberOfTime30-59DaysPastDueNotWorse share the same values, specifically 96 and 98. We can see that sharing the same values of 96 and 98 respectively is not logical since trivial calculations can reveal that being 30 days past due for 96 times for a single person within a timespan of 2 years is not possible, which might indicate Data Entry error. However, they can't be dropped due to high information they possess in identifying defaulting members. Its best we keep them and replace all the 96/98s with 18s to make them not extreme outliers.

Monthly Income
Records with missing Monthly Income have high Debt Ratio (Median 1159).Summary Stat of Borrowers with high Debt Ratio shows that the Monthly Income of these Borrowers are 0. This could mean Borrowers with missing Monthly Income deliberately left the column blank because they are trivial workers not earning Monthly Income.The best method to handle this missing values is to replace it with 0.

Number of Dependents
Records with missing Number of Dependents occurred simultaneously with missing missing MonthlyIncome (i.e they share the same index).This shows that the same set of borrowers that left their Monthly Income blank also left Number of Dependents field Blank. Summary stat of Borrowers with missing monthly Income reveals they have no dependents. It's quite logical that this category of borrowers with little to no Income have no dependents.Thus, the best way to handle this missing values is to replace with 0

Number of open credit line and real estate loans
Two variables are right-skewed with no extreme values. Further preprocessing of this data would be aggregating similar Category (Fine Class) to a Coarse class during WOE Feature Engineering and Data Preprocessing.

Conclusion
After feature engineering, we have a new dataset with 146.021 rows and 11 features.

3.3 Convert variables to WOE

In this section, we convert raw data in to woe by following step:

- Fine Classing: All Continuous Variables would be binned into several categories based on its distribution. Any variable with more than 50 unique values is considered to be a continuous Variable. Other Numerical variable with less than 50 unique values would have each element as a separate category

- Coarse Classing: Categories with similar WOE value would be binned together. Percentage of observation would also influence coarse classing.
- Dummy variable would be created for each coarse class
- Each variable would have a reference attribute to avoid dummy variable trap

Our dataset for WOE have 93 features with the 146.021 rows (the same as data after dropping)

4. Finding and discussion

4.1. Evaluation

The model's evaluation criteria cannot be appraised solely on the basis of its accuracy due to the imbalance dataset and , hence F1 score, precision, recall of default and AUC are chosen for the evaluation.

In this section, the dataset is splitted into the training set and the test set, with a 4:1 ratio.

Training is carried out on the training set and then, the predictions are made on the test set. The evaluation results of Logistic Regression, Logistic Regression with WOE method are obtained as shown in table below.

Algorithm	F1 score	AUC score	Precision	Recall
Logistic Regression	0.97	0.858	0.54	0.17
Logistic Regression with WOE	0.97	0.866	0.59	0.19

As the table shown, all algorithms yields good results of over 0.8, in terms of AUC score and have the same high result in terms of F1 score (0.97). When compared on both metrics, it is shown that Logistic Regression with WOE performs better than traditional Logistic Regression model because it outperforms in both Precision and Recall of default . Finally, based on the above trials, we assess each model's stability and generalization capabilities and assign a credit score card in Logistic Regression with the WOE model.

4.2. Credit scorecard

The final part of this article is creating a simple, easy-to-use, and implemented credit risk scorecard that can be used by any layperson to calculate an individual's credit score given certain required information about him and his credit history.

The scorecard's score scale can be created by combining the logarithms of the default and non-default probability ratio logarithms (odds):

$$\text{Score} = A - B \times \ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$$

where:

- A represents the offset
- B represents the scale factor
- $\hat{p}$ represents the default probability.

with

$$A = bp - po \times \frac{ln(od)}{ln(2)} \text{ and } B = \frac{po}{ln(2)}$$

We assumed that the base point is 600 with od = 50 and po = 20. After calculation, an offset of 487.12 and factor of 28.85 are obtained. We get the scorecard of banking customers, as shown partially in the table below.

ID	Actual	Proba	Score
146670	0	0.030736	586.700376
312	0	0.016005	605.963345
2170	1	0.779367	450.709855
30008	0	0.018612	601.533395

Appendix

The code used to support the findings of this study are available here