Loan Repayment Prediction : Case Study

Rushi Pankade

Problem Statement

Build a credit score using Machine Learning and that indicates the likelihood of loan repayment

Credit Scoring – What , Why & How

What : Methodology leveraged by Financial Institutions to determine the risk of non payment associated with loans

Why is it used?

- Credit Scoring enables decision making at all customer lifecycle stages
- Removes the need to manually examine each loan customer
- Clear understanding of Denial or Approval reasons leading to sound business approach

How is it used?

- Credit Scores are used to determine the following areas under Loan portfolios
  - **Approval** – Should the Loan be approved?
  - **Pricing** – What is the right Interest ?
  - **Cross Sell** – Can we sell another loan ?
  - **Refinance** – Should there be a change in Interest?

Scorecards – How to they look like ?

Scorecards are deployed to take Model Driven decisions in an intuitive manner

Why is it used?

Variable	Value Bucket	Scorecard Point
AGE	<22	100
AGE	22<=AGE<26	120
AGE	26<=AGE<29	185
AGE	29<=AGE<32	200
AGE	32<=AGE<36	210
AGE	36<=AGE<42	225
AGE	>=42	250
HOME STATUS	OWN	225
HOME STATUS	RENT	110
INCOME	<10000	120
INCOME	10000<=INCOME<17000	140
INCOME	17000<=INCOME<25000	180
INCOME	25000<=INCOME<35000	200
INCOME	35000<=INCOME<58000	225
INCOME	58000<=INCOME<100000	230

Steps to build a Scorecard

- Exploratory Data Analysis
- Variable Discretization
- WOE Substitution and IV Screening
- Model Fitting and Validation
- Score Scaling from Model Parameters
- Drive Decisions through Scores

Dataset and Description

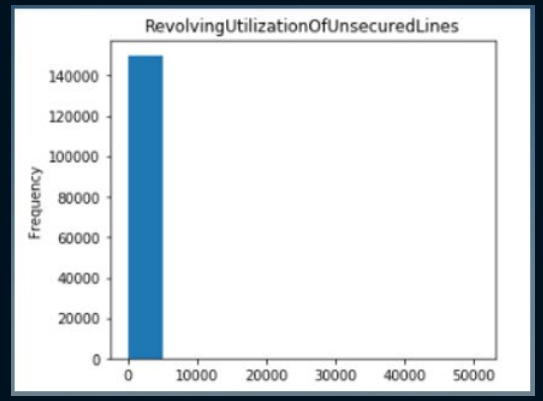
VARIABLE	DESCRIPTION
Application	The month which the loan was funded
	The number of payments on the loan. Values are in months and can be either 36 or 60.
Bureau	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
	The total number of credit lines currently in the borrower's credit file
Collections	Number of collections in 12 months excluding medical collections
	post charge off collection fee
Credit Line	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
	The month the borrower's earliest reported credit line was opened
Deliquency	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
	The number of months since the borrower's last delinquency.
Demographics	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
	The state provided by the borrower in the loan application
Employment	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
	The job title supplied by the Borrower when applying for the loan.*
Income	Indicates if income was verified by LC, not verified, or if the income source was verified
	The combined self-reported annual income provided by the co-borrowers during registration
Outstanding	Remaining outstanding principal for portion of total amount funded by investors
	Remaining outstanding principal for total amount funded
Payment	Last total payment amount received
	Next scheduled payment date
Pricing	Interest Rate on the loan
	Interest received to date
Risk Score	The lower boundary range the borrower’s FICO at loan origination belongs to.
	The lower boundary range the borrower’s last FICO pulled belongs to.
Risky Behavior	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
	Total credit revolving balance
Status of Loan (Dependent)	Status of Loan

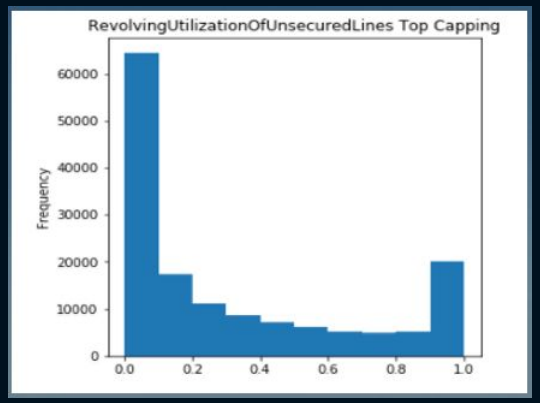
Exploratory Data Analysis

Defining Dependent Variable and Cleaning of Independent Variables

- Check for variability in the data. Lack of variability reduces the explanatory power over dependent variable
- Treating of outliers : Flooring and Capping
- Missing Value Imputation Methods :
  - Central Tendency (Replace with Median – Continuous)
  - Central Tendency (Replace with Mode for Character)
  - Class Mean Substitution (Replace with closet Bin Mean)

Example of Outlier Treatment





Binning and Discretization

What is the nature of Bivariate Relationship of variables

- Create buckets of independent variables based on ranking methods
- This process allows us to understand feature performance better
- The insights from this part of the analysis can be useful in devising portfolio risk strategies

Example of Binning Method

```
Import pandas as pd
df ['var_bin']=pd.qcut(df['var'].rank(method='first').values,10,duplicates='drop').codes+1
df.groupby('var_bin')['Dependent'].mean()
```

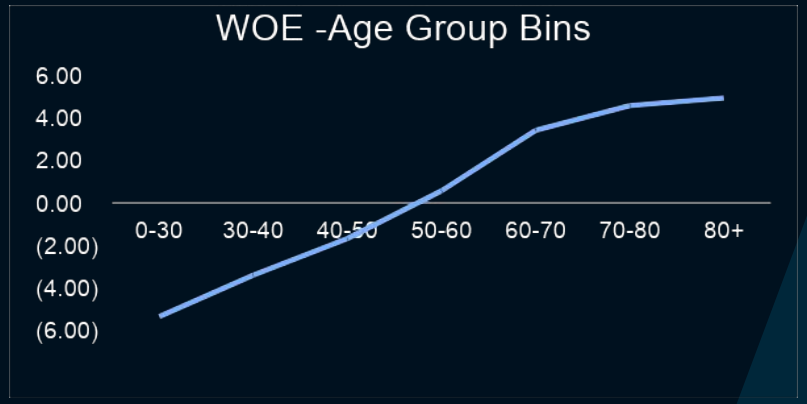
Capped_Age	Cappped_Age_Bin
21	A:0-22
23	A:22-26
27	A:26-29
28	A:29-32
33	A:32-36
34	A:36-42
60	A:42+

Weight of Evidence Substitution

Weight of Evidence is used for substituting in place of categorical values for Variable Bins

Calculation Logic

$$WOE = \text{Ln}\left[\frac{\text{Distribution of Good Cases}}{\text{Distribution of Bad Cases}}\right] \times 100$$



Uses and Interpretation

- WOE measures the strength of a bin in differentiating the Good and Bad accounts
- WOE values are substituted in place of the Bin Labels across variables
- WOE <0 indicates that the variable bin is captures higher proportion of bad accounts
- The overall trends for WOE across should be logical and we should ideally observe a monotonic trend

Information Value of Variables

Information value is the measure of overall predictive power of the variables and is very useful for feature selection

Calculation Logic

IV = Sum (Good Case % – Bad Case %)| WOE

IV	Predictive Power
<0.02	Not Useful
0.02 to 0.1	Weak Predictor
0.1 to 0.3	Medium Predictor
0.3 to 0.5	Strong Predictor
>0.5	Suspicious

Example (WOE & IV)

Age Bins	Obs Count	Bad	Good	Bad Distribution	Good Distribution	WOE	IV
0-30	10,758	1,244	9,514	7%	7%	-5.36	0.02
30-40	24,339	2,390	21,949	16%	16%	-3.4	0.02
40-50	35,037	2,893	32,144	23%	23%	-1.68	0.01
50-60	34,806	2,149	32,657	23%	23%	0.56	0
60-70	27,424	952	26,472	18%	19%	3.4	0.02
70-80	12,700	298	12,402	8%	9%	4.56	0.02
80+	4,447	89	4,358	3%	3%	4.91	0.01

IV =0.09

Model Fitting on prepared data

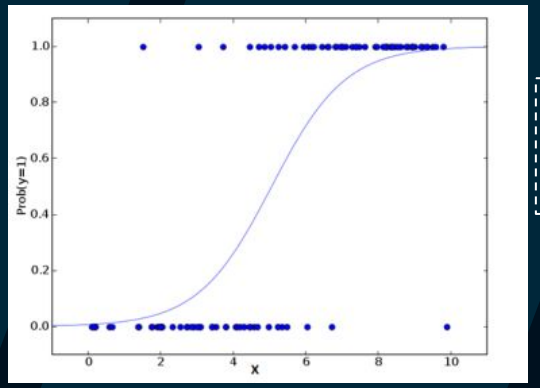
Parametric Approach : Logistic Regression Classifier to be used to fit the training data and model to be finalized post iterations

Logistic Regression Model

- Dependent Variable – Categorical (Good/Bad)
- Relates Log of Odds to a Linear Combination of Predictors
- Final model have statistically significant predictors
- LR Coefficients to be estimated from the training data to be used for developing Scorecard
- The Scores are further scaled to fall within a certain range of values

Model Structure

$$\log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 \cdot x_{i2} + \beta_2 \cdot x_{i2} + \_ + \beta_p \cdot x_{in}$$



Predicted Probabilities Describe a Sigmoidal Curve

How to Measure Classification Performance

Confusion Matrix

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

- Accuracy

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- Precision

$$\frac{TP}{TP + FP}$$

- Recall

$$\frac{TP}{TP + FN}$$

F Score

$$\text{F-measure (F)} = \frac{1}{\left(\frac{1/r + 1/p}{2}\right)}$$

$$\frac{2TP}{2TP + FP + FN}$$

Score Scaling

Scores for each variable bin is calculated as follows

$$\text{Score} = \left[\beta * \text{WOE} + \alpha / N\right] * \text{Factor} + \text{Offset}/N$$

Where

β	Logistic Regression coefficient of the variable
α	Logistic Regression Intercept
N	Number of Variables in the Model
Factor	Points to double the odds / Ln (2)
Offset	Score – { Factor * ln (Odds) }

- We choose to scale the points such that a total score of 600 points corresponds to good/bad odds of 50 to 1
- An increase of the score of 20 points corresponds to a doubling of the good/bad odds.

Factor – 20/ln(2) = 28.85

Offset – 600 – 28.85 * Ln (50) = 28.85

Decision Making from Scorecard

Considerations for Score Usage

- Choosing a score cut off that is appropriate for the Loan Product
- For Example : Secured Loans may be approved for a lower score than unsecured loans
- Scores not only are used for assessment but also for predictive purposes
- The Model scores combined with business considerations are used for final decision making

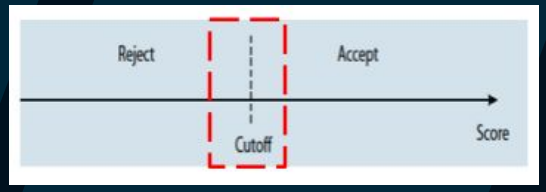
An Example

Characteristic	Attribute	Scorecard Points
AGE	<22	100
AGE	22<=AGE<26	120
AGE	26<=AGE<29	185
AGE	29<=AGE<32	200
AGE	32<=AGE<37	210
AGE	37<=AGE<42	225
AGE	>=42	250
HOME	OWN	225
HOME	RENT	110
INCOME	<10000	120
INCOME	10000<=INCOME<17000	140
INCOME	17000<=INCOME<28000	160
INCOME	28000<=INCOME<35000	200
INCOME	35000<=INCOME<42000	225
INCOME	42000<=INCOME<58000	230
INCOME	>=58000	260

Let cutoff=600

So, a new customer applies for credit.....

AGE	35	210 points
INCOME	$38K	225 points
HOME	OWN	225 points
Total		660 points
Decision:		GRANT CREDIT



Data Set Link

https://www.kaggle.com/wendykan/lending-club-loan-data

Loan Data (887K x 75)

THANK YOU!

Rushi Pankade

Phone
+91 9075853157

Email
rvpankade@gmail.com