LENDING CLUB CASE STUDY

SUBMISSION

Name: Arindam Ghosh & Karthik Chandran Pillai

Summary – Problem Statement

- The problem statement is to identify risky loan applicants for one of the largest online loan marketplace which facilitates personal loans, business loans and financing of medical procedures.
- Typically for lending companies, lending loans to ‘risky’ applicants is the largest financial loss (credit loss) wherein borrowers do not repay back the loan and default on the repayments.
- At the same time, if the company rejects a loan applicant who is likely to pay back the loan, it would lead to loss of business to the company
- Our objective is to understand:
  - How loan attributes influence tendency of default?
  - How consumer attributes influence tendency of default?

Problem solving methodology

Data Cleaning: Load the Loan data available to us and clean data for missing fields, dropping empty columns, converting certain variables to numeric values and imputing missing values

Outlier Detection: Then we describe the data and identify that few fields have outliers, hence we treat data and only retain values where Z-score < |3|

Univariate and Bivariate Analysis I – We try to plot bar graphs for most categorical and continuous variables and try to identify which variables could have a meaningful impact on loan default behavior

Univariate and Bivariate Analysis II – From this analysis, we drop certain variables basis observations and select few variables for further analysis (both categorical and continuous variables)

Derived Variable - Basis external study we identified that the Credit scoring model in the BFSI Industry uses a model to identify 'good' and 'bad' loan customers using WoE (Weight of Evidence) and IV (Information Value)

We then proceed to calculate the WoE and IV for each of the variables shortlisted post our earlier Analyses (4 Categorical and 9 Continuous) and plot the graphs of each variable by their WoE values to discern any trends and their relation to loan default probability

Finally we plot the IV value of all variables in an ascending order to identify the variables which have medium and strong prediction power (values above 0.1)

Conclusion – From the variables We finally can conclude the variables that:

- Have a high IV value
- Which show a strong trend w.r.t loan default rate for customers

Analysis – Key Observations I

Primary Data observations

- Post scanning through the loan data file – companies and rounds, we notice that there are close to 111 columns (data headers) and 39717 rows (data observations).
- Post treating the data for missing values, empty columns, outliers etc. we are left with 32964 rows and 38 columns

Univariate & Bivariate Analysis

- Categorical variables
  - We divide data into two data frames by their loan status (Fully Paid customers and Charged Off/Defaulted customers)
  - We then calculate the proportion of counts for each variable in both data frames and plot their bar graphs
  - If there is any significant difference in the trends of proportion between Full paid and Charged Off bar graphs, we are to conduct further analysis on the respective variable
  - We decide to proceed with 4 categorical variables for further analysis basis our findings namely: Loan Grade, Loan Sub Grade, Purpose and Address State
- Continuous Variables
  - We plot 2 subplots of 12 continuous variables by both their Mean and Median values for Loan Status
  - From the subplots we shortlist 9 variables for further analysis where we can see a discernible difference in the mean/median between plot of Fully Paid customers and Charged Off customers namely: Annual income, Interest rate, Loan term, Debt to Income ratio, Total Credit Accounts, Revolving Utilization, Last Payment Date, Loan funded amount and Last credit pull date

External Study – Credit Scoring Model

In our study from external sources, we were able to find that the Credit Risk Industry uses multiple methods to profile Credit risk of customers. Two of the available approaches are:

- Judgemental Approach (qualitative, expert based approach based on business experience)
- Statistical Scoring (data based statistical approach via optimal multi-variate analyses)

Based on our understanding of data available to us, we decide to proceed with the statistical approach that is used to build a Credit Risk model. The most common method used to build such a model is **Logistic Regression (LR)** and we have zeroed in on the Information Value (IV) criterion and Weight of Evidence (WOE) methodology under this model as they provide a great framework for exploratory analysis and variable screening for binary classifiers.

According to Baesens et al. (2016) and Siddiqi (2012), WOE and IV analysis enable one to:

- Consider each variable's independent contribution to the outcome.
- Detect linear and non-linear relationships.
- Rank variables in terms of "univariate" predictive strength.
- Visualize the correlations between the predictive variables and the binary outcome.
- Seamlessly compare the strength of continuous and categorical variables without creating dummy variables.
- Seamlessly handle missing values without imputation.
- Assess the predictive power of missing values.

Source: *RPubs by RStudio*

WOE and IV Methodology

Post the Bivariate Analysis we are able to have an understanding as to which independent variables might impact our Credit Risk Analysis. In order to calculate the impact of each independent variables we calculate their Weight of Evidence and Information Value.

The weight of evidence tells the predictive power of an independent variable in relation to the dependent variable. Since it evolved from credit scoring world, it is generally described as a measure of the separation of good and bad customers. “Bad Customers” refers to the customers who defaulted on a loan. and “Good Customers” refers to the customers who paid back loan.

Information value is one of the most useful technique to select important variables in a predictive model. It helps to rank variables on the basis of their importance. Formulae for both parameters are as below:

$$WoE = [ln(\frac{\text{Relative frequecy of Goods}}{\text{Relative frequecy of Bads}})] * 100$$

$$IV = \sum(DistributionGood_i - DistributionBad_i) * WoE_i$$

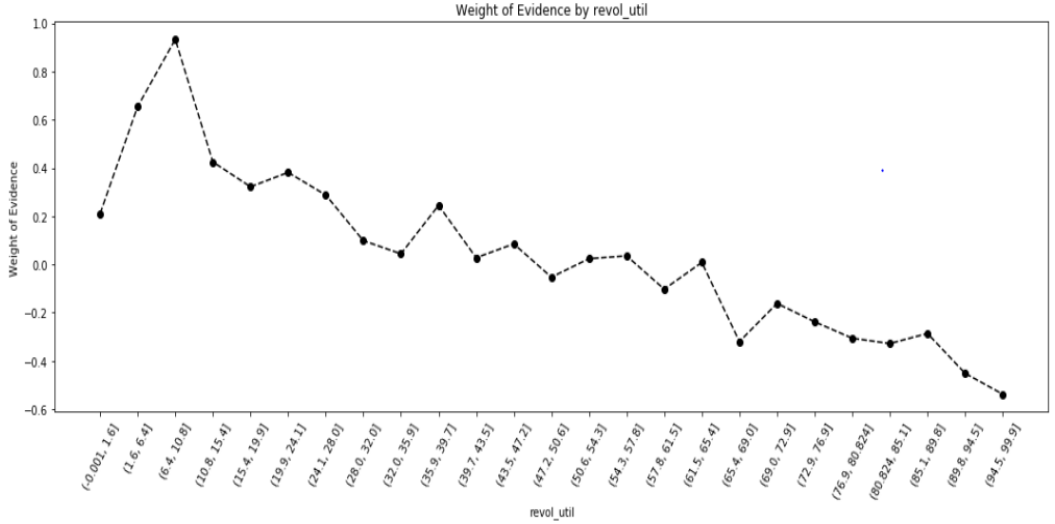
WoE and IV formulas

Once we have the IV of the variable, we can check against this table to see the predictive power of the variable.

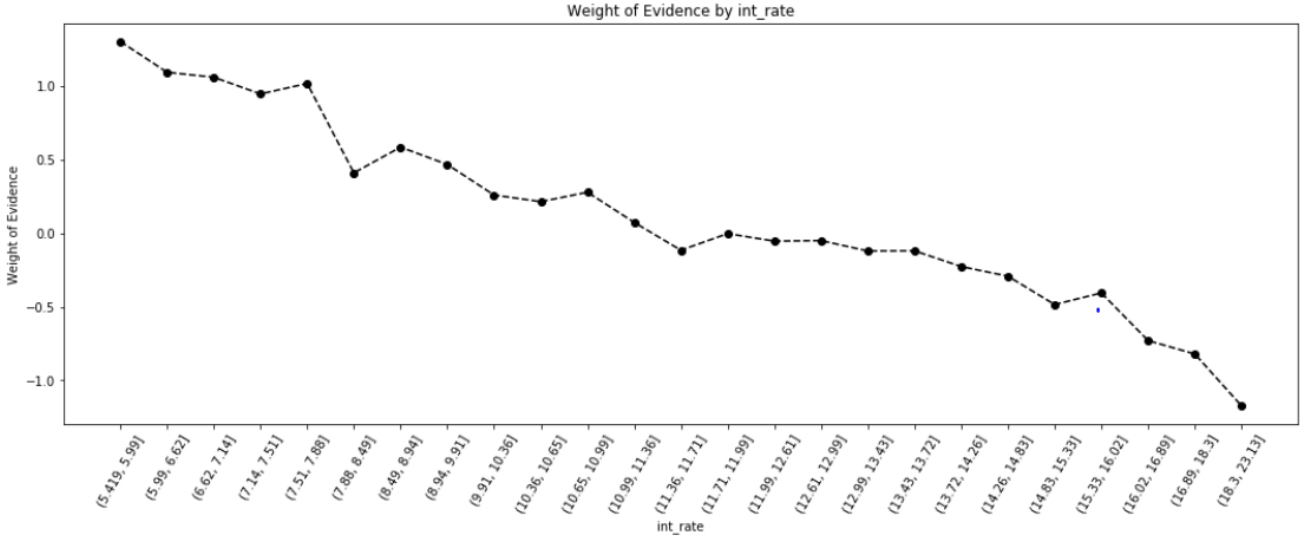
Information Value	Predictive Power
< 0.02	useless for prediction
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
0.3 to 0.5	Strong predictor
>0.5	Suspicious or too good to be true

Analysis – Key Observations II (WoE plots)

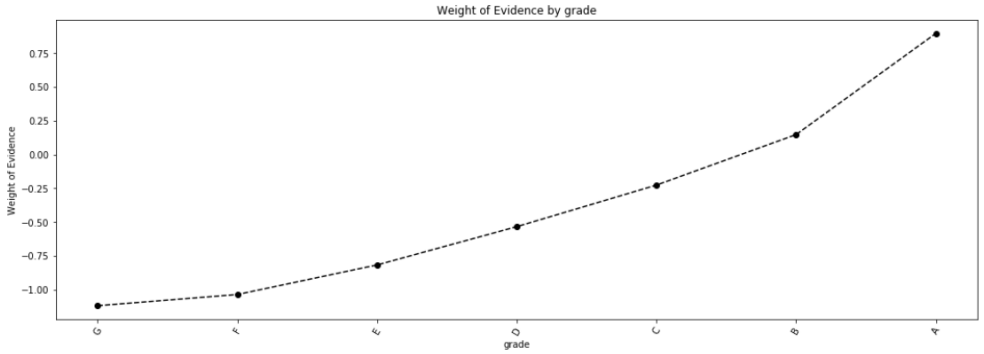
As the WoE score becomes more negative it signifies the chances of loan default increasing , similarly as the WoE score improves the Credit worthiness also improves. With these parameters , we plot the WoE scores of the top variables and see how they impact our Credit Analysis



Inference: There is significant evidence which says that with increase in revolving utilization rate, the loan default possibility increases

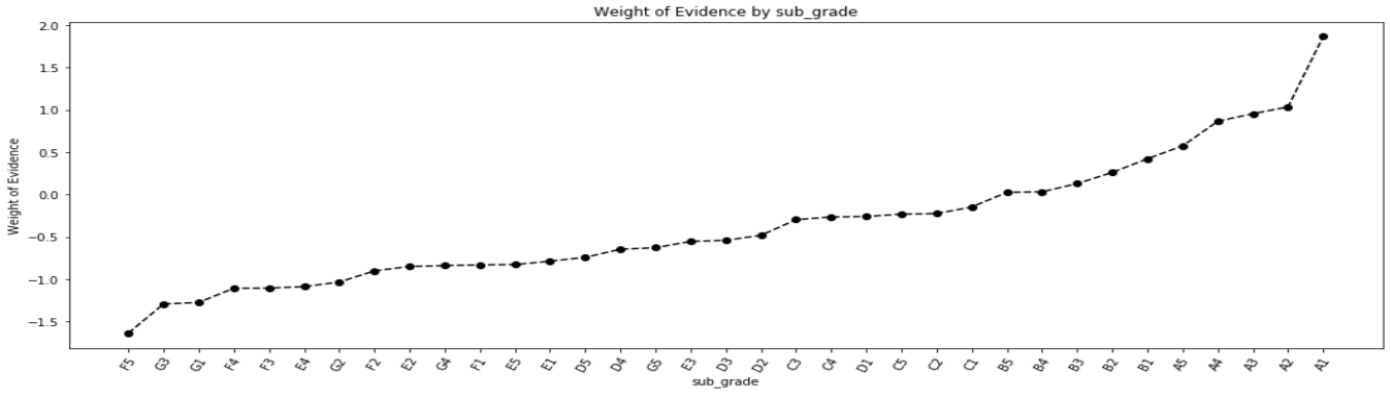


Inference: As the interest rate moves to the higher buckets the WoE reduces which basically provides evidence that the probability of default increases



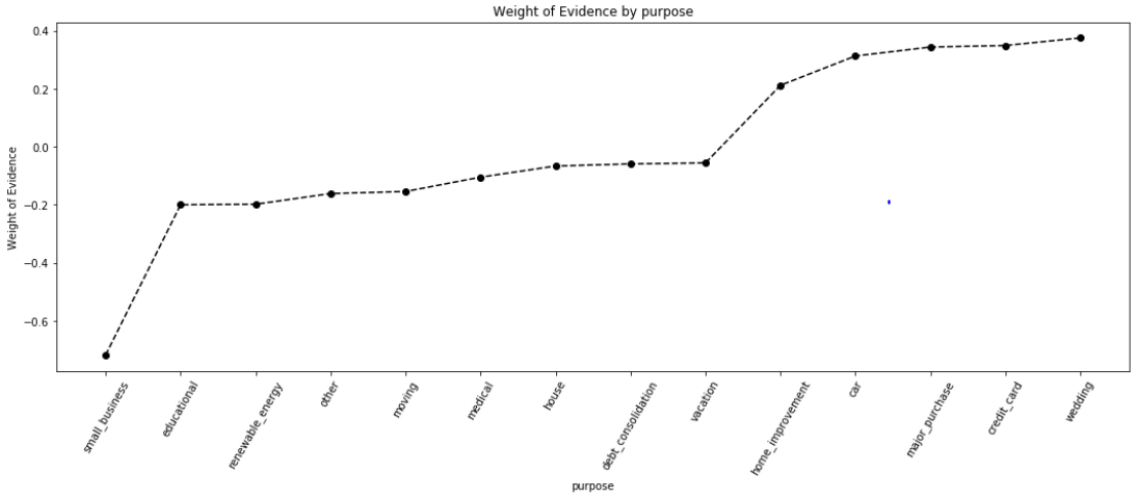
Inference :

Grade has a very strong IV , which indicates that it has a strong prediction value in forecasting whether the customer will default or not. As per the graph, we can conclude: Loans classified as A has the highest WoE and thereby least chance of default and subsequently with decrease in loan grade from A to G the WoE becomes negative which means increasing probability of default

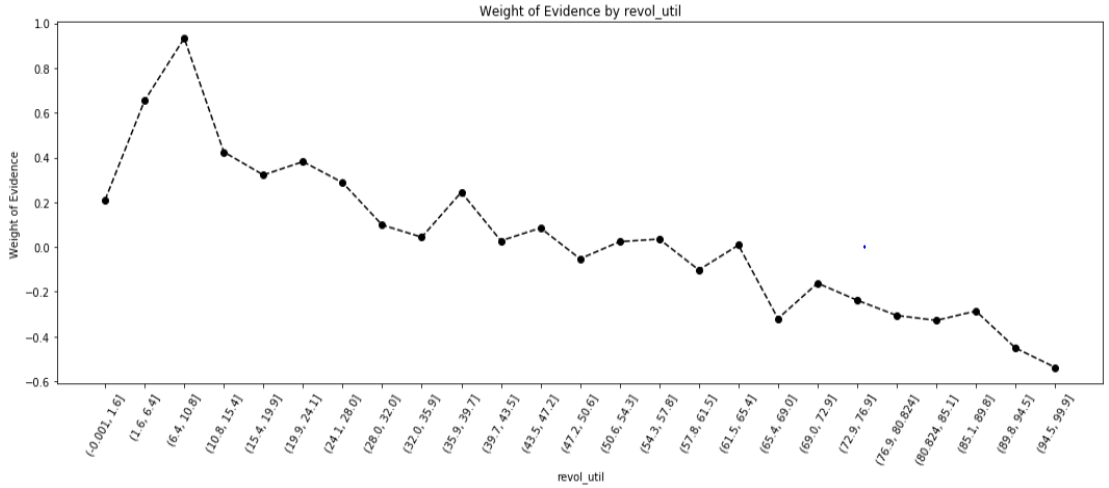


Inference: There is a clear WoE trend which shows that subgrades of A and B have the highest WoE reflecting least possibility of default for such loans, post C2 there is a declining trend for WoE

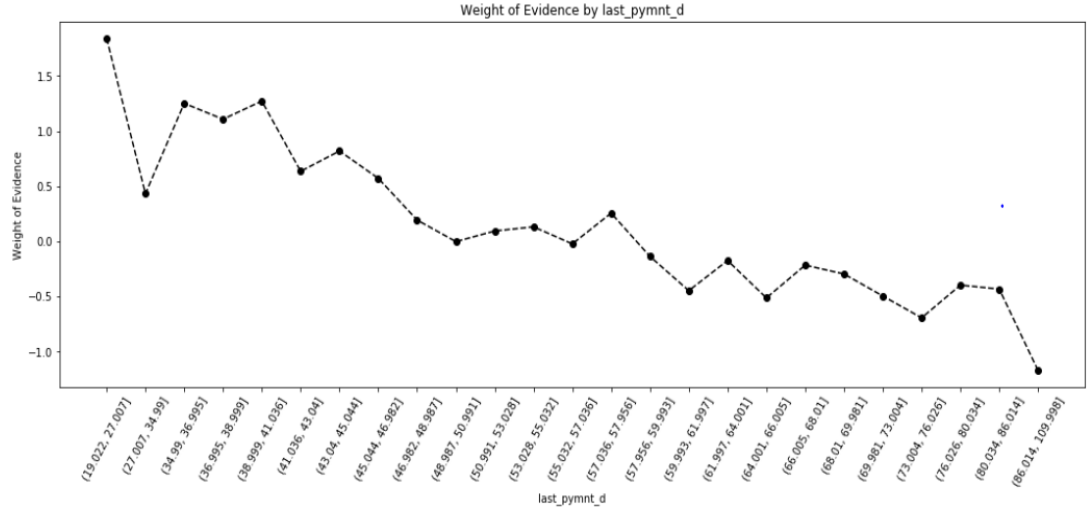
WoE Plots (Contd.)



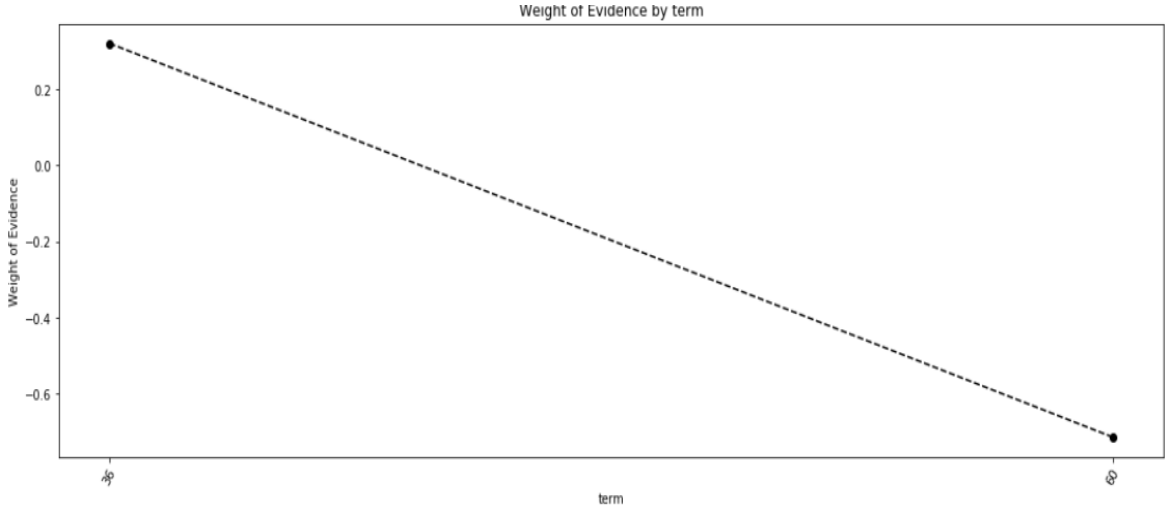
Inference: People value loans taken for big occasions such as Weddings, Major Purchase, Car and have a higher tendency to repay them whereas small business loans, educational loans etc. have a higher tendency of default



Inference: There is significant evidence which says that with increase in revolving utilization rate, the loan default possibility increases

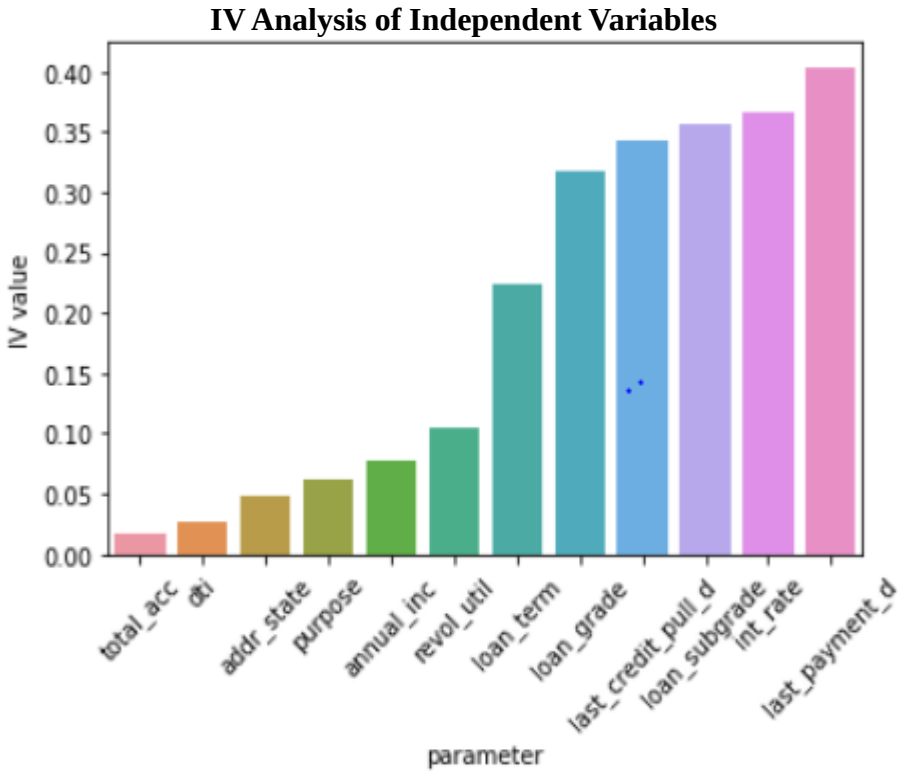


Inference: As the number of days from last payment date increases, there is a strong trend that shows increased loan default rate



Inference: As the term length of the loan increases the probability of loan default increases as well

Results



- On computing the IV values, we get 12 feature variables having IV values greater than .02
- These are the 12 variables which must be considered before approving credit to an individual.
- Among these features , we find a few like revol_util (Revolving Utilization rate), loan term , loan grade , loan subgrade , last credit pull date , interest rate and last payment date have a high IV value and hence are more significant in determining individual’s credit worthiness.
- The WoE plots also show trends for each of these variables giving us an idea how they impact loan default probability for each customer

Conclusions

From our Analysis of the loan data made available to us, we were able to conclude the following:

- Some variables have a key impact in determining the probability of loan default by a customer which are basically sub categorized into loan attributes and consumer attributes
- Loan attributes which impact loan default probability are:
  - Loan Grade
  - Loan Subgrade
  - Loan Term
  - Loan Interest Rate
- Consumer attributes which impact loan default probability are:
  - Revolving utilization rate
  - Last payment date
  - Last credit pull date
  - Annual Income

The lending marketplace will hence be able to predict loan default rate using these key variables and build a Logistic Regression model for Credit Risk analysis and then decide to lend to customers accordingly depending on the categorization provided by the model as ‘good’ or ‘bad’ customer.

The more accurate their model, the better they will be at averting financial loss through loan defaults from future customers and will also be able to identify potential customers whom they can disburse loans to, without rejecting them wrongly.