BFSI Capstone

Cred-X Acquisition

CASE STUDY

FINAL-SUBMISSION

Group :
Praveen Nair
Himanshu Kunwar
Aditya Khajuria
Manish Bisht

Introduction

Problem Statement :

Cred-X, a leading credit card provider has experienced increased credit loss in recent years. To mitigate this risk, Cred-X wants to acquire the right customers.

Key Objective:

Help Cred-X identify the right customers using appropriate predictive models. That would involve below :

1. Using historical Data to determine factors affecting credit risk.
2. Crediting strategies to mitigate acquisition risk.
3. Assessing financial benefits of the project.

Data Description :

- Demographic Data : 71295 observations of 12 variables.
  - Contains data on customer age, income, gender, marital status etc.
- Credit Data : 71295 observations of 19 variables.
  - Data obtained from Credit Bureau, contains information on loans, outstanding balance, trades, DPD etc.

Assumptions:

1. The dependent variable “Performance Tag” is missing for 1425 observations , these are treated as applicants who have been rejected by CredX. We are using this for testing the cut-off for credit score.
2. There are some cases where all the variables in the credit burea data are zero and the credit card utilization is missing, hence this is missing data from Credit burea.
3. Cases where credit cards utilization is missing are customers without credit cards.

Approach

- **Data Preparation**

1. Check for Duplicates in Application ID and its uniqueness.
2. Checking for missing values in both the dataset.
3. Replaced the missing values with median.
4. Outlier Detection and treatment.
5. Creating derived variables.
6. Formatting and creating dummies for categorical data and scaling numerical data.

- **EDA**

1. Univariate, Bivariate, multivariate and correlation analysis of variables to determine which factors are likely to have influence on credit default.
2. WOE and Information Value Analysis to determine predictive value of variable.

- **Model Building and Selection**

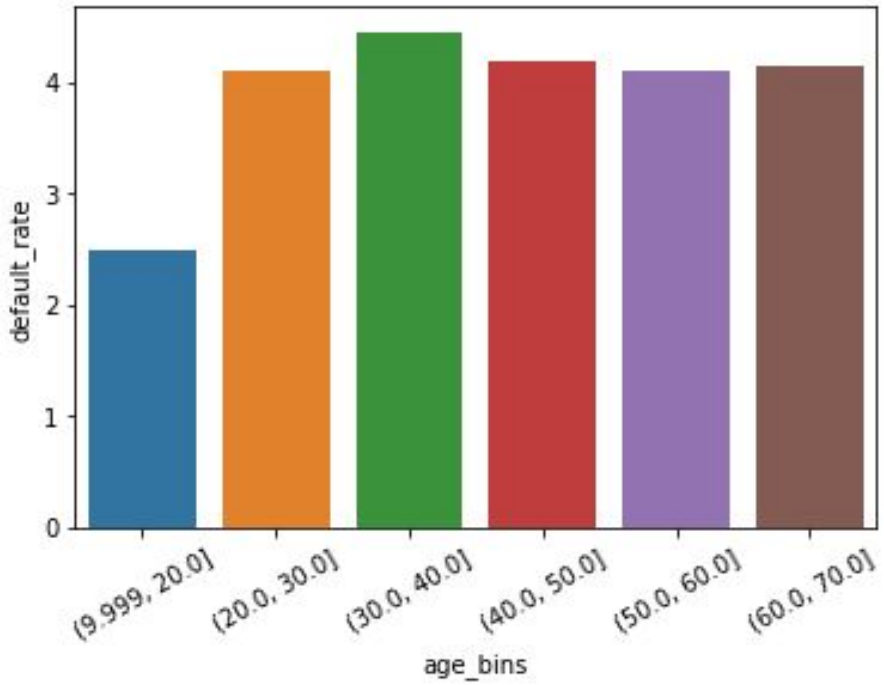
1. Divide data into training and test datasets in 7:3 ratio.
2. Iterative model building on training data set using
    1. Logistic Regression
    2. Decision Tree
    3. Random Forest
3. From above Models selection will be based on specific parameters and would be using that model to predict test data.
4. Evaluating model accuracy, sensitivity and specificity.
5. Plotting ROC curves.
6. Evaluating KS statistics.
7. Plotting Gain and Lift Charts.

- **Application Score Card and Financial Benefit.**

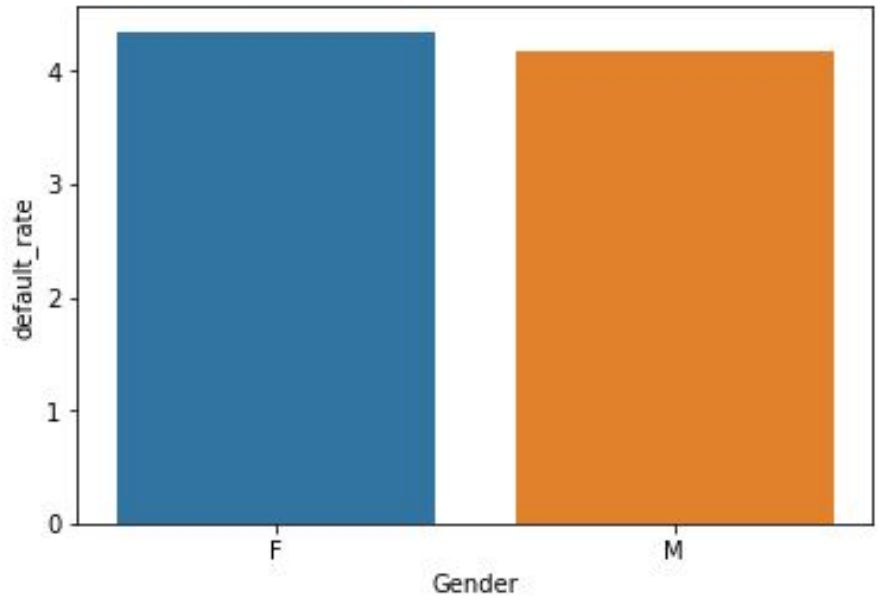
1. Chosen model from previous is used to build and application scorecard.
2. A cut-off is chosen below which the applicants will not be granted credit cards.
3. The financial benefit is accessed in terms of the credit loss minimized as well as in terms of the revenue maximized by acquiring right customers.
4. For score calculation, post model selection, we would create a scorecard based on below formula
   - Points to double the odds (pdo=20)
   - Factor = pdo/ln(2)
   - Offset = Score – {Factor * ln (Odds)}

$$\text{Score} = \sum_{i=1}^{n} \left(-\left(woe_i * \beta_i + \frac{a}{n}\right) * factor + \frac{offset}{n}\right)$$

- **EDA - Default % Across Categories.**

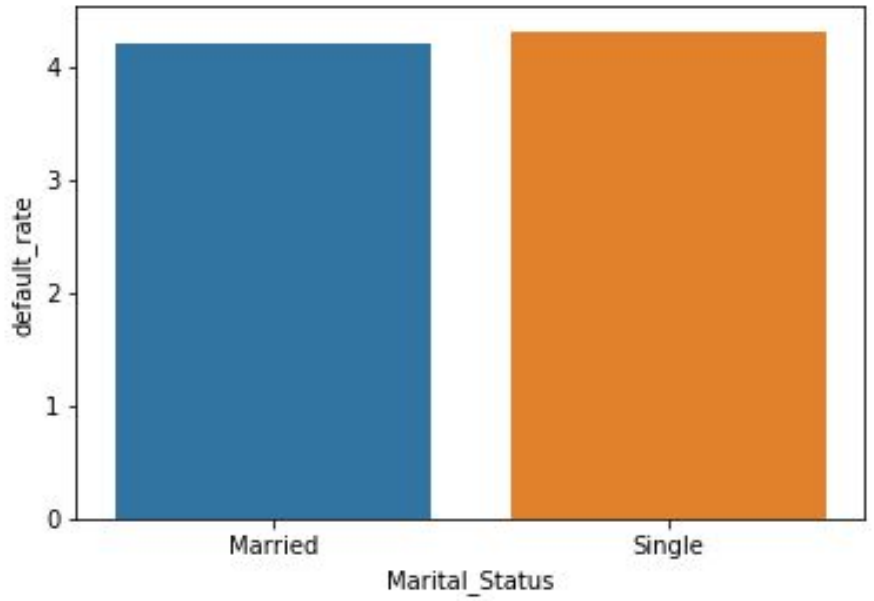


Defaulters % by Age Bins.

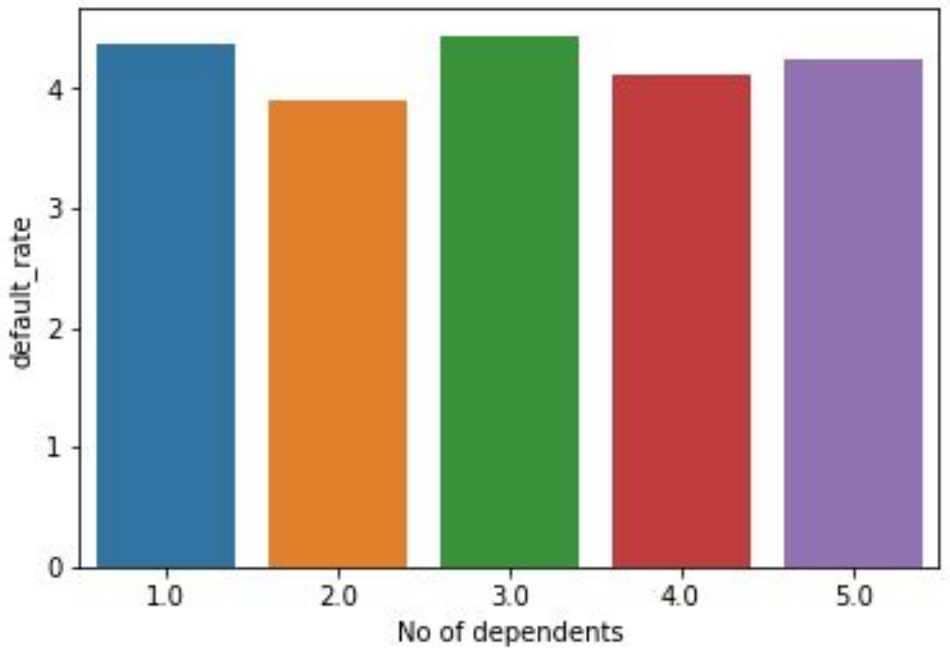


Defaulters % by Gender.

- **EDA - Default % Across Categories**.- *Contd..*

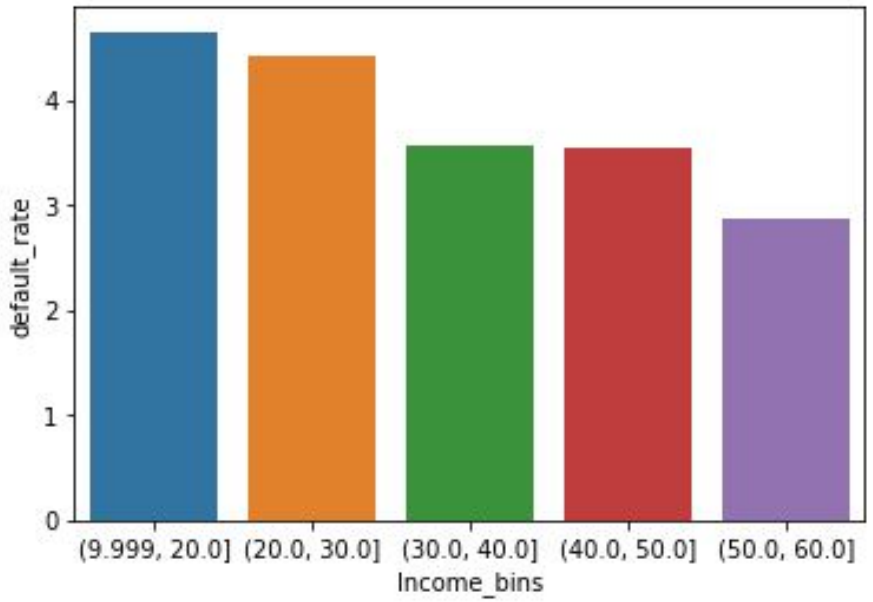


Defaulters % by Marital Status.



Defaulters % by No. of Dependents

• **EDA - Default % Across Categories**.- *Contd..*



Defaulters % by Income Bins.

- *Percentage of defaulters is approximately same across gender, marital status and no of dependents.*
- *Though it is slightly higher i.e. 5% for some ages, however there is no visible pattern for the same.*
- *Overall most approved candidates are male with professional or masters degree in a salaried profession married and living in a rented housing.*

- **Factors affecting Credit Risk- Demographic Data**

1. The dataset doesn't have many outliers.
2. Except for marginal differences in age and income bins among defaulters vs the entire data set of applicants, other demographic variables show no difference in pattern.
3. This indicates that demographic data are poor indicators and predictors of credit risk.

Demographic Data Information Value

	VAR_NAME	IV
1	Application ID	0.000010
6	No of dependents	0.000054
5	Marital_Status	0.000108
3	Gender	0.000316
0	Age	0.000665
2	Education	0.000783
10	Type of residence	0.000875
9	Profession	0.002354
7	No of months in current company	0.012736
4	Income	0.037297
8	No of months in current residence	0.052075

- We have Calculated WOE and IV to check the predictiveness.
- As presented in table, Demographic data has less predictiveness for credit risk.
- Only Income and No of months in current company has weak predictive power, other variable are not useful.

$$IV = \sum (\%\ \text{of non-events} - \%\ \text{of events}) * WOE$$

Information Value Formula

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

•Master file (Demographic + Credit Bureau data) Information Value

	VAR_NAME	IV
1	Application ID	0.000000000231
11	No of dependents	0.000077641468
6	Marital Status (at the time of application)	0.000099878726
4	Gender	0.000261017498
0	Age	0.000781389594
3	Education	0.000832848994
27	Type of residence	0.000865437100
23	Presence of open auto loan	0.001589716972
25	Profession	0.002073528659
22	Outstanding Balance	0.007021285035
12	No of months in current company	0.013047557872
24	Presence of open home loan	0.017256034251
19	No of times 90 DPD or worse in last 6 months	0.030353087486
5	Income	0.037108445436
13	No of months in current residence	0.052756189228
13	No of months in current residence	0.052756189228
17	No of times 60 DPD or worse in last 6 months	0.090450767034
8	No of Inquiries in last 6 months (excluding ho...	0.094721828935
21	No of trades opened in last 6 months	0.096411109104
18	No of times 90 DPD or worse in last 12 months	0.096707596621
10	No of PL trades opened in last 6 months	0.127259108587
16	No of times 60 DPD or worse in last 12 months	0.139655154451
15	No of times 30 DPD or worse in last 6 months	0.147097936707
26	Total No of Trades	0.148800028364
9	No of PL trades opened in last 12 months	0.180967485997
14	No of times 30 DPD or worse in last 12 months	0.191060159620
7	No of Inquiries in last 12 months (excluding h...	0.241818864439
20	No of trades opened in last 12 months	0.272303962910
2	Avgas CC Utilization in last 12 months	0.298301605246

- We have Calculated WOE and IV to check the predictiveness of whole Master data.
- Demographic data has less predictiveness where as Credit Bureau has better predictiveness.
- Variables having IV between 0.1 to 0.3 are showing good predictiveness.

Score Card - Calculating Scores

- Post model selection a score card is created using the formula specified above in the presentation.
- Cut-off selection - a cut off of 445 is selected based on the distribution of scores, because very high cut off will impact approval rate hugely.
- Using cut-off of 445 to reject applicants, we apply the scores to Rejected data (separated earlier for validation)-93% of these rejections are correctly classified by the scorecard.
- Overall 66% defaulters were filtered out using cut off of 445.

- **Data Manipulations**

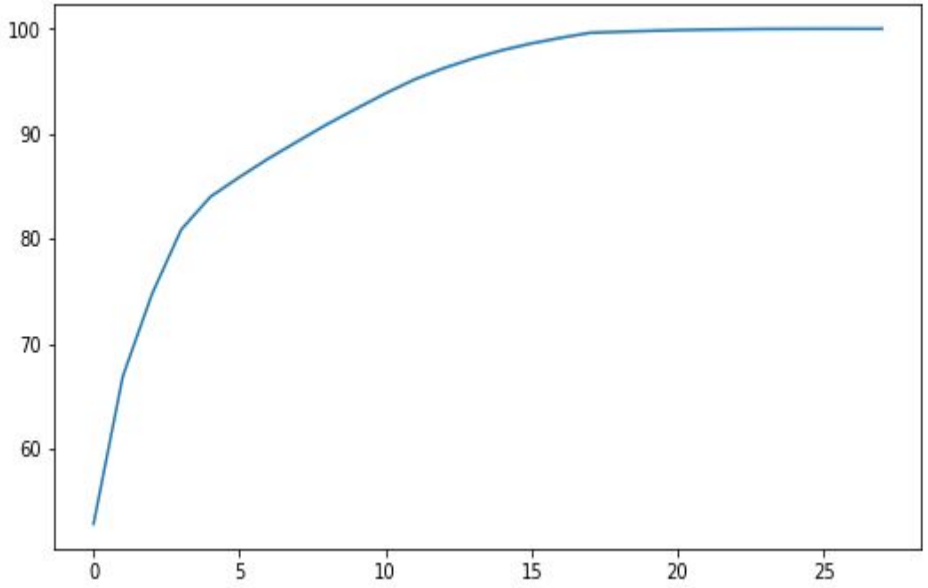
  - **Missing Value Imputation:**
    - With Median: No. of dependents, no of trades opened in last 6 months, Presence of open home loan, Outstanding balance.
    - With 0: NA values in
      - No of Dependents
      - Presence of Open Home loan
      - Outstanding Balance
      - Avg. CC utilization in last 12 months indicate no utilization of credit card by user.
    - Using Mode:
      - Gender : M
      - Marital Status at time of application : Married
      - Educational : Professional
      - Type of Residence : Rented.

- **Data Manipulations-** *Contd..*

- **Binning –**
  - Age : values less than 10 are imputed with median values.
  - Income: Negatives values have been imputed with median values.
- **Outlier Treatment** —
  - The outliers were treated within range of 20% -80 % for no of months in current company, Avg. CC utilization in last 12 months, No of trades opened in last 6 months and 12 months, No of PL trades opened in last 6 and 12 months, inquiries in last 6 and 12 months , total no of trades.
- **Dummy Variables –**
  - Dummy are created for all categorical variables : Gender, Marital Status, Education, Profession, Type of Residence.
  - Application Id wouldn’t be considered for model building.

Sampling, Model Building, Outcomes and Selection

- Only 4% of the data compromise defaulters, hence data is highly unbalanced
- Logistic regression is performed to predict the log odds of default depending upon categorical and numerical variables
- First model is created using glm() on all variables.
- After several iterations, final model is selected using p-values and VI
- Variables that have a negative impact on log odds of default are
  - Income
  - Average months in current company
  - Average credit card utilization
- Variables that have a positive impact on log odds of attrition are
  - No of times 90 DPD or worse in last 6 months
  - No of times 30 DPD or worse in last 6 months
  - No of times 90 DPD or worse in last 12 months
  - No of times 60 DPD or worse in last 12 months
  - No of times 90 DPD or worse in last 6 months
  - No of times 30 DPD or worse in last 6 months
  - No of PL trades opened in last 12 months
  - No of inquiries in last 12 months excluding home and auto loans
  - No of PL trades opened in last 6 months
- **Overall Accuracy at Sensitivity-71%, Specificity ~69%**
- **Area under the Curve is ~69%**



Plot feature variance

Financial Benefit Analysis

Objective

- From the financial perspective, we are trying to optimize the percentage of defaulters, while also taking care that we do not compromise revenue by reducing approved applicant percentage.
- The ultimate aim is to reduce credit loss for CredX, in terms of the outstanding balance of defaulters.

Benefits

- Originally ~ 4 to 4.2% of approved applicants were defaulters, with the model the number of defaulters has come down to 2.4%
- Total credit loss = defaulters outstanding = 3.96B
- Filtering out defaulters with score< 425, the new credit loss=1.3B, i.e. almost 3 times less.

Contd...

Potential Loss and Recommendations

- Original Approval rate- % of applicants granted credit (total applicants-Rejected Applicants)-98%while new approval rate is 60%.
- This implies that here is some potential loss in terms of rejecting credit worthy applicants-34%applicants who were non-defaulters (98%-4.2% defaulters -60% approved) will be rejected using this approach.
- However the company can decide to approve applicants in the low-medium score category by imposing a higher rate of interest on these applicants. In this way it can neutralize any losses expected from potential defaulters.

Recommendations for better Customer Selection

- As seen demographic data is not good predictor for potential default
- in addition, CredX should consider credit history variables, particularly
  - Average credit card utilization in last 12 months
  - No of trades opened in last 12 months
  - No of enquiries in last year
  - No of times 30 dpd or worse