Capstone Project

Final SUBMISSION

Team:

Srividya Ravichandran
Abhijith N V

Key Objectives

- Understand the given Demographic and Credit Bureau data
- Perform data analysis and cleansing
- Perform WoE and IV to handle outliers, missing data and also identify important variables
- Build models and chose the right model to predict the default
- Run the model on the rejected data and get the probability score and compare it with the accepted applicants
- Build an application scorecard with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points
- On the basis of the scorecard, identify the cut-off score below which you would not grant credit cards to applicants
- Check the accuracy, sensitivity of the model
- Finalize the model and give recommendations on the variables that influences the default

What Data we have

- Demographic Data of the applicant with the below details about the applicant
  - Age
  - Gender
  - Marital Status
  - No of dependents
  - Income
  - Education
  - Profession
  - Type of residence
  - No of months in current residence
  - No of months in current company

What Data we have

Credit Bureau data explaining the financial details of the applicant with the below details

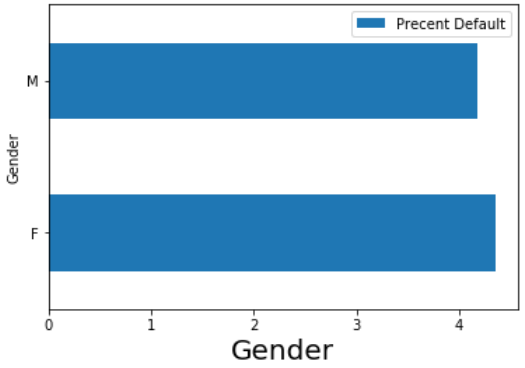
- Number of times customer has not payed dues since 90days in last 6 months
- Number of times customer has not payed dues since 60 days last 6 months
- Number of times customer has not payed dues since 30 days last 6 months
- Number of times customer has not payed dues since 90 days last 12 months
- Number of times customer has not payed dues since 60 days last 12 months
- Number of times customer has not payed dues since 30 days last 12 months
- Average utilization of credit card by customer
- Number of times the customer has done the trades in last 6 months
- Number of times the customer has done the trades in last 12 months
- No of PL trades in last 6 month of customer
- No of PL trades in last 12 month of customer
- Number of times the customers has inquired in last 6 months
- Number of times the customers has inquired in last 12 months
- Is the customer has home loan (1 represents "Yes")
- Outstanding balance of customer
- Number of times the customer has done total trades
- Is the customer has auto loan (1 represents "Yes")

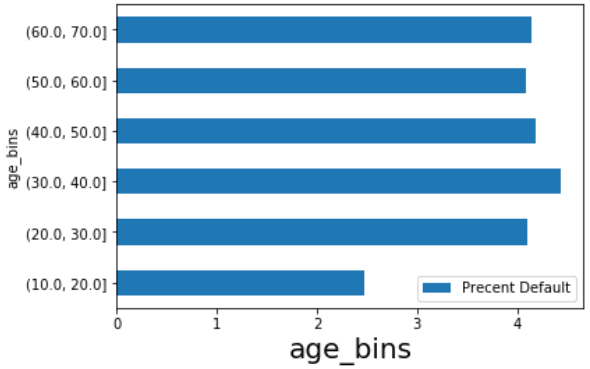
Data Preparation and Analysis

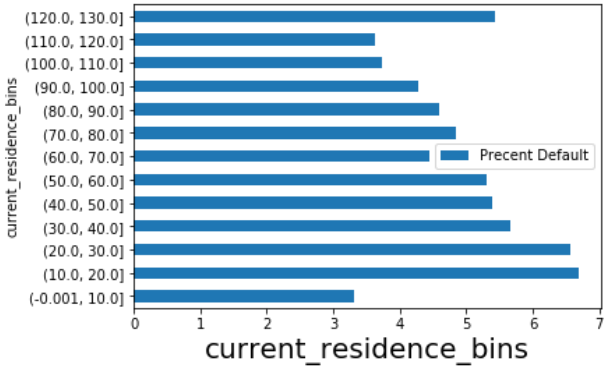
- Filter the Data with performance Tag as null as they are rejected applications and are not useful for model building
- Age has data less than 18 years, default them to 18 years as that is the minimum age for issuing credit card
- We will assign null values in the categorical columns to ‘unk’, as they will be handled through WoE
- Identify duplicates on both date sets based on application ID and clean the same
- Bin the following continuous variables across Demographic and Credit Bureau data
  - Age
  - Income
  - # of months in current residence
  - # of months in current company
  - Average CC utilization in the last 12 months
  - Outstanding balance
- Check Performance Tag values between Demographic and credit bureau to identify mismatch to ensure there is no data quality issues
- Merge the two data sets, Demographic data and Credit Bureau data on Application ID for further model building

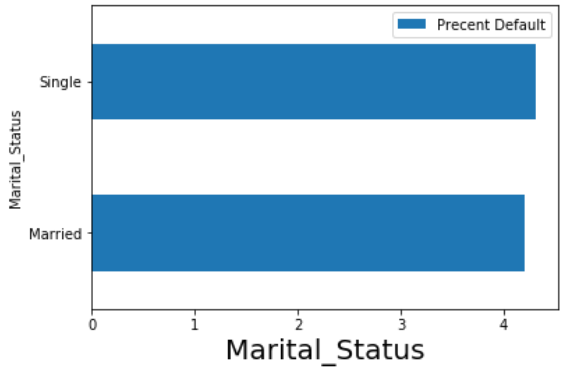
Data Exploration

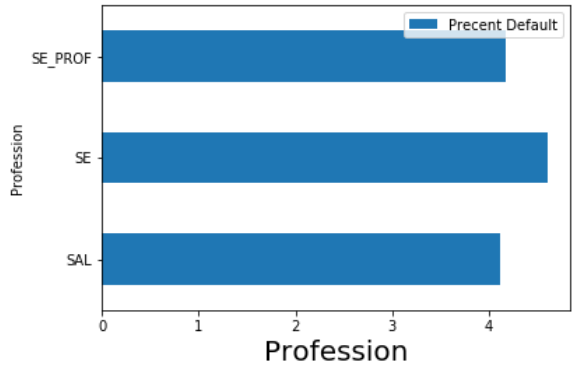
Analyze Default percentage by various Demographic data

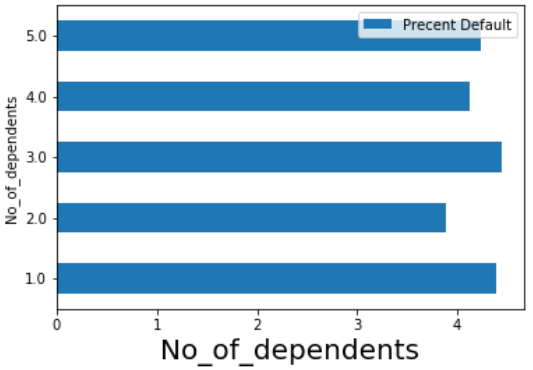






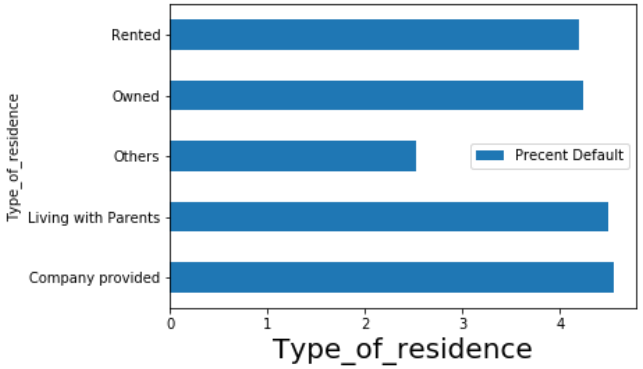


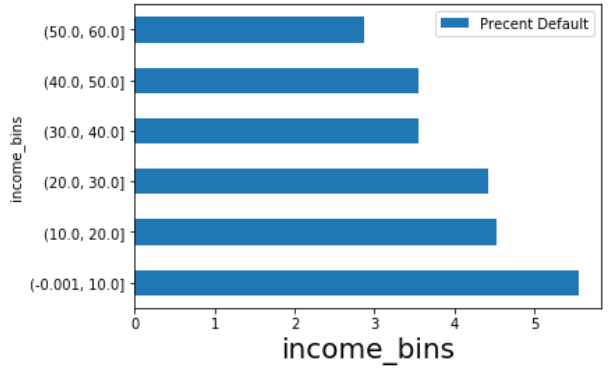


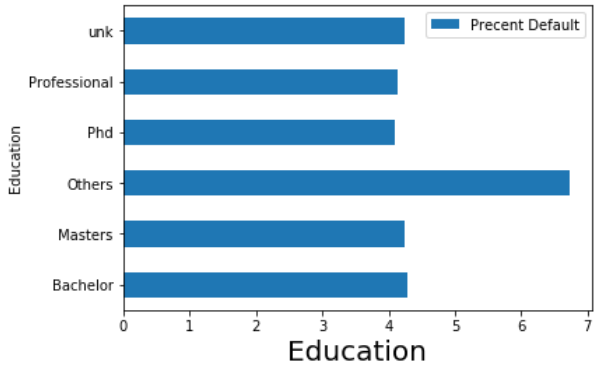


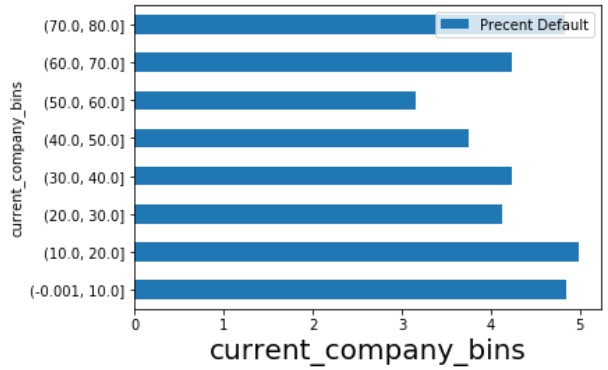
Data Exploration

Analyze Default percentage by various Demographic data, Contd…







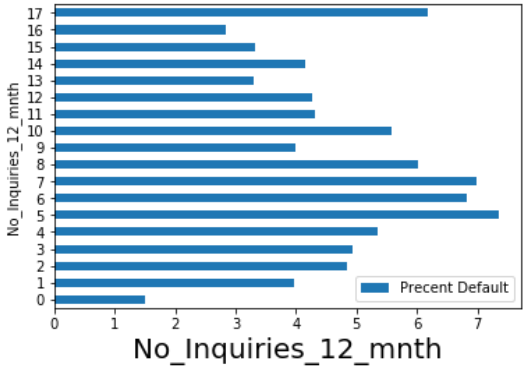


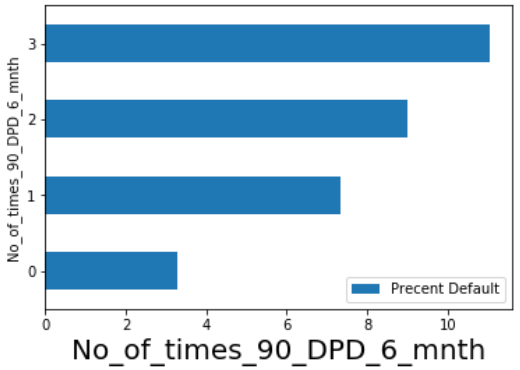
Inferences

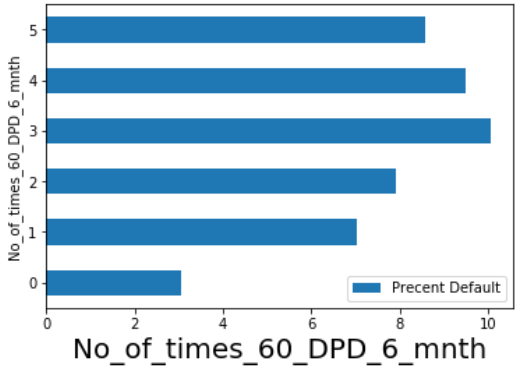
- Applicants between 30-40 years has the highest default percentage, whereas applicants who are less than 20 years has the least default percentage
- Gender, Marital Status, No of Dependents are not a significant factor in default
- Lower the income higher than chances of default
- People with Others as the education has the highest default percentage

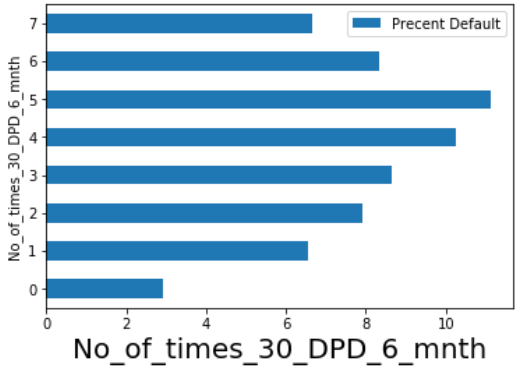
Data Exploration

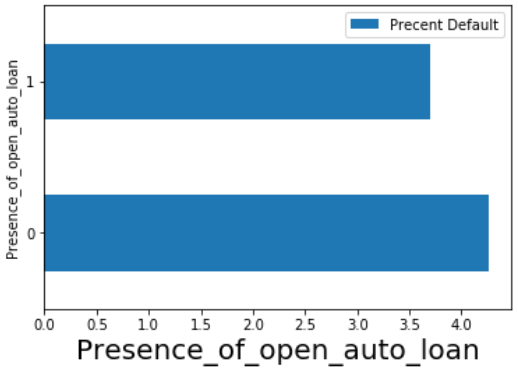
Analyze Default percentage by various Credit Bureau data

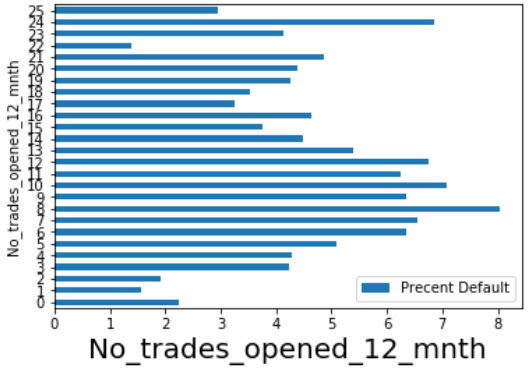


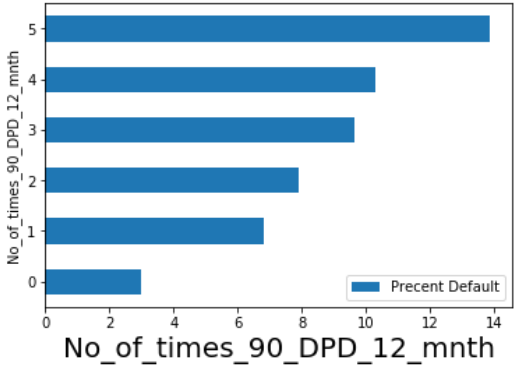


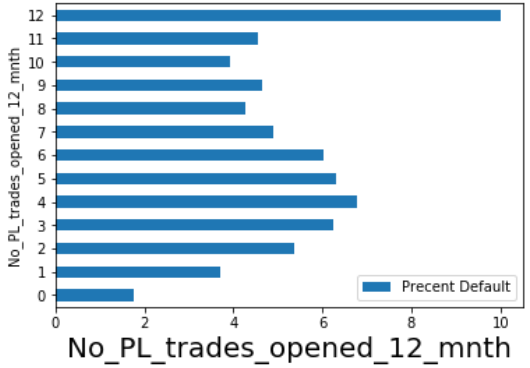


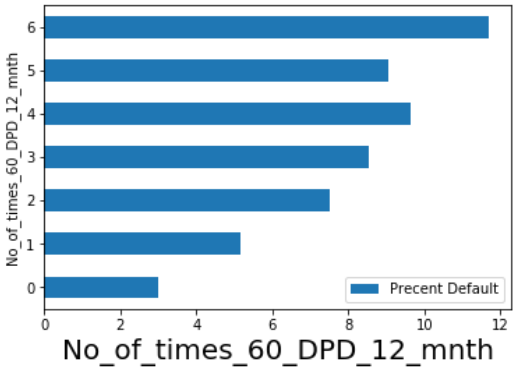


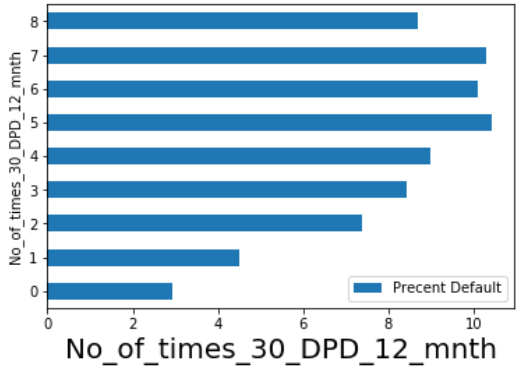


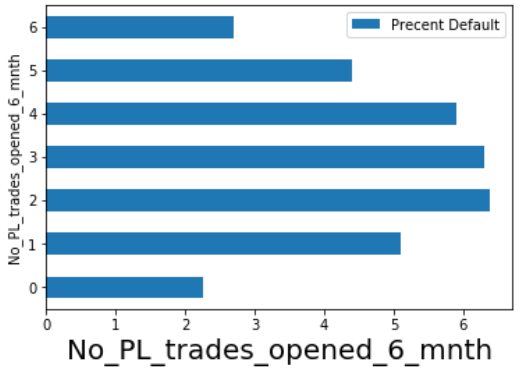


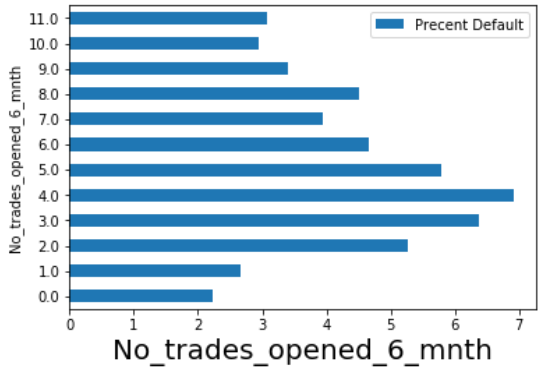






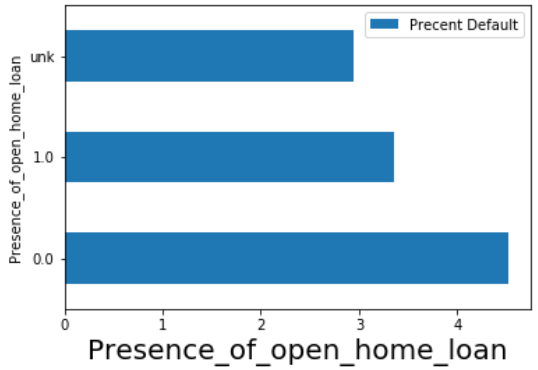


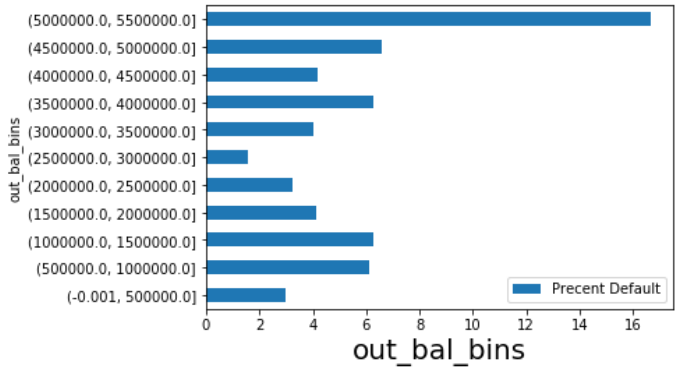


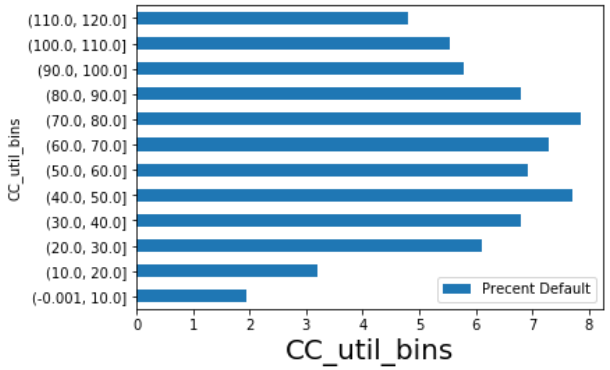


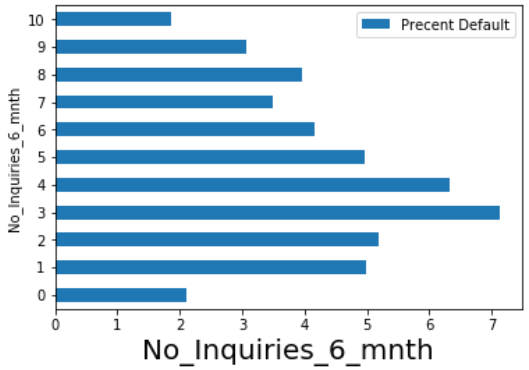
Data Exploration

Analyze Default percentage by various Credit Bureau data, Contd….





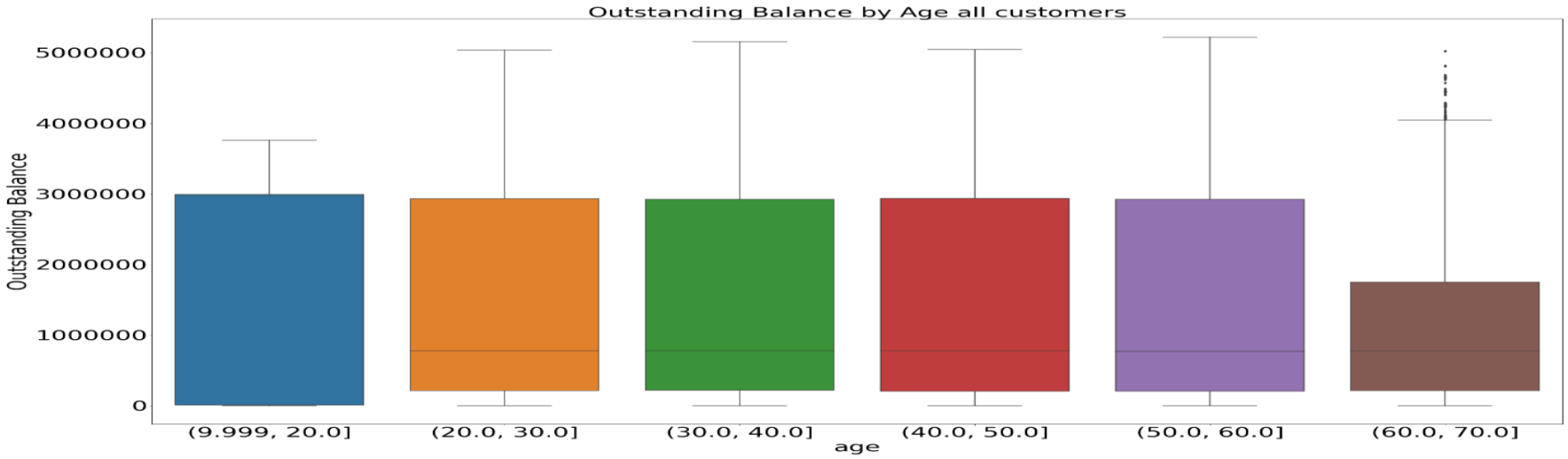


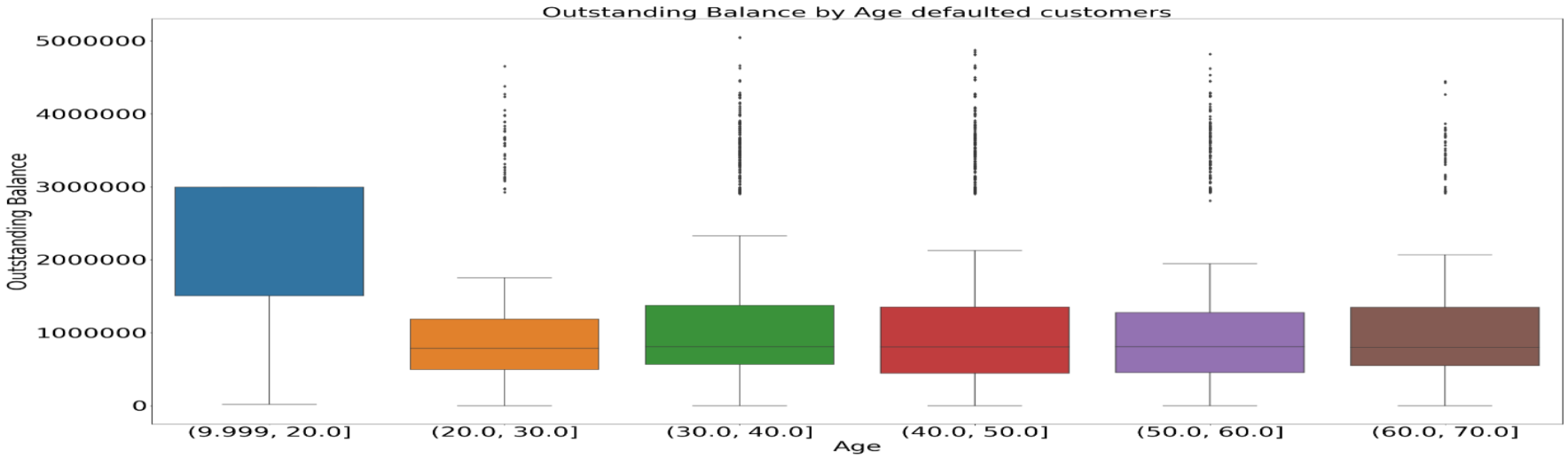


Inferences

- Higher the times past due in the last 3,6,12 months more likely the default
- People with Home loan are less likely to default
- As the outstanding balances increase the likelihood off default also increases
- People with 40-80% CC utilization have a tendency to default more

Data Exploration

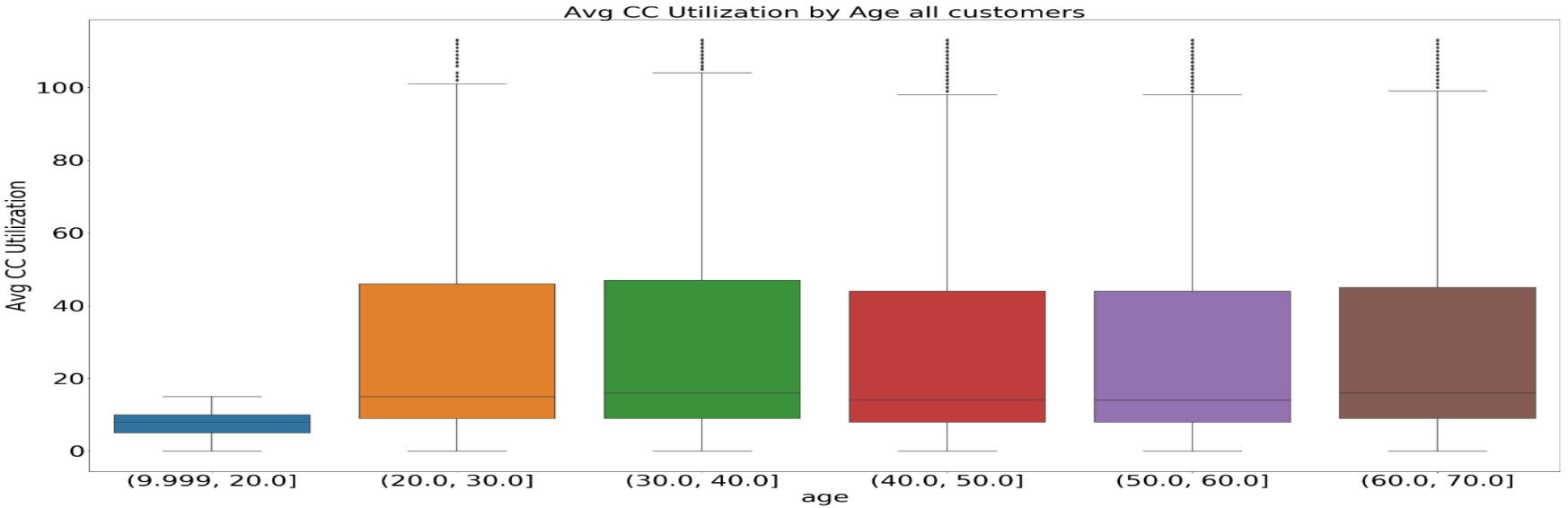


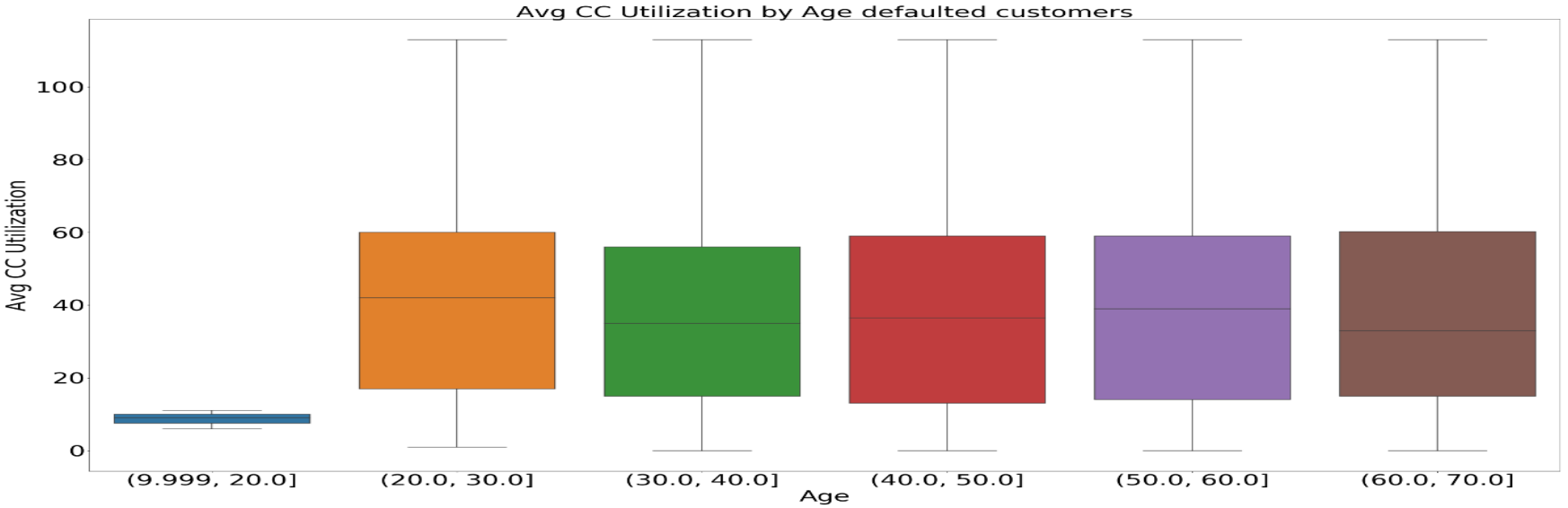


Inferences

- Avg outstanding balance is same across all ages

Data Exploration





Inferences

- Avg CC utilization is same across all ages for all customers
- Avg CC utilization is higher for people with age category 30-40 years for defaulted customers

Weight of Evidence and IV

- Perform WoE on the Demographic data and combine data set
- WoE will ensure missing values and outliers are handled
- Replace WoE values in the main data set before model building

Weight of Evidence and IV

Below is the Details of the variables and no. of bins created so that the WOE is monotonic.

- No_trades_opened_12_mnth - bins=8
- No_PL_trades_opened_12_mnth - bins=7
- No_Inquiries_6_mnth - bins=6
- No_Inquiries_12_mnth - bins=9
- Age - bins=3
- Income - bins=7
- No_of_months_in_current_residence - bins=8
- No_of_months_in_current_company - bins=6
- Avgas_CC_Utilization_12_mnth - bins=5 (null bins)
- Outstanding_Balance - bins=10 (+ null bin; Wavy but monotonic)
- Total_No_of_Trades - bins=7

All categorical variables have one bin for each category

['Gender', 'Marital_Status','Education', 'Profession',
'Type_of_residence','Presence_of_open_home_loan','Presence_of_open_auto_loan','No_of_dependents','No_of_times_90_DPD_6_mnth',
'No_of_times_60_DPD_6_mnth','No_of_times_30_DPD_6_mnth', 'No_of_times_90_DPD_12_mnth','No_of_times_60_DPD_12_mnth',
'No_of_times_30_DPD_12_mnth','No_trades_opened_6_mnth', 'No_PL_trades_opened_6_mnth']

The WOE on the Rejected dataset is the WOE value based on the bins created on the accepted dataset.

Important Variables

Demographic Data

Variable	IV	Predictor Type
current_residence_bins	0.074498	Weak Predictor
income_bins	0.040351	Weak Predictor
No_of_months_in_current_company	0.022663	Weak Predictor
No_of_dependents	0.002818	Useless for Prediction
Profession	0.002221	Useless for Prediction
age_bins	0.001628	Useless for Prediction
Type_of_residence	0.000942	Useless for Prediction
Education	0.000783	Useless for Prediction
Gender	0.000568	Useless for Prediction
Marital_Status	0.000146	Useless for Prediction

Inferences

- For the Demographic data alone none of the variables indicate to be strong variable for prediction
- We have to use RFE to see what can be used for building the model for prediction
- Current Residence and Income Seems to be good variables for prediction

Important Variables

Combined Demographic & Credit Bureau Data

Variable	IV	Predictor
CC_util_bins	0.315854	Strong Predictor
No_trades_opened_12_mnth	0.293562	Medium Predictor
No_PL_trades_opened_12_mnth	0.258546	Medium Predictor
Outstanding_Balance	0.246663	Medium Predictor
No_Inquiries_12_mnth	0.245249	Medium Predictor
No_of_times_30_DPD_6_mnth	0.24425	Medium Predictor
Total_No_of_Trades	0.232262	Medium Predictor
No_PL_trades_opened_6_mnth	0.22422	Medium Predictor
No_of_times_30_DPD_12_mnth	0.21861	Medium Predictor
No_of_times_90_DPD_12_mnth	0.215653	Medium Predictor
No_of_times_60_DPD_6_mnth	0.211274	Medium Predictor
out_bal_bins	0.196406	Medium Predictor
No_trades_opened_6_mnth	0.191843	Medium Predictor
No_of_times_60_DPD_12_mnth	0.188231	Medium Predictor
No_of_times_90_DPD_6_mnth	0.162659	Medium Predictor
No_Inquiries_6_mnth	0.11337	Medium Predictor

Inferences

- Credit Card Utilization is a strong predictor for Default
- No of Trades opened in last 12 months and No of PL trades opened in last months are the next 2 variables which are medium predictors for default behavior

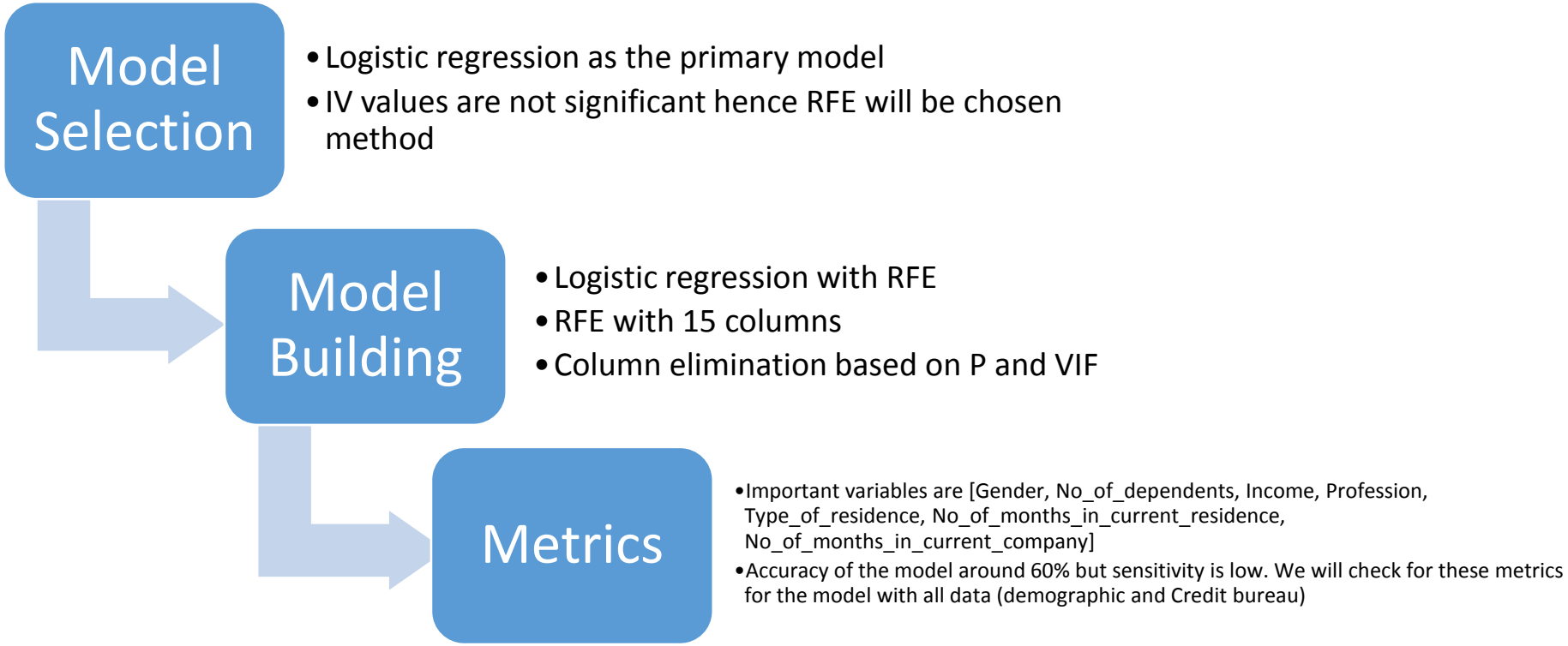
Next Steps and Model Building – Demographic data

- Replace WoE values to the main data frame (without the rejected application)
- Perform Test and Train split of the data for the above data frame
- Perform RFE for the Demographic data to identify variables
- Perform Logistic regression to and check the accuracy and sensitivity
- If required decide to perform Ridge/Lasso regression and other higher models like Random Forests/SVM to achieve better higher accuracy and sensitivity

Next Steps and Model Building – Combined Data

- Replace WoE values to the main data frame (without the rejected application)
- Perform Test and Train split of the data for the above data frame
- Use important variables as part of IV and perform logistic regressions
- Perform RFE if required and continue for logistic regressions
- Check the accuracy and sensitivity
- If required decide to perform Ridge/Lasso regression and other higher models like Random Forests/SVM to achieve better higher accuracy and sensitivity
- Use the model to predict the odds for rejected applicants and compare with approved applicants and make observations on various variables that impacted the decision
- Chose optimal cutoff in such a way the credit loss is minimized and the company can rely on the model to accurately predict model and ensure the right applicants are selected

Model Building – Demographic Data



Model Building – Combined Data

Model Selection

- Try Logistic regression as the initial model with following feature selection options
  - IV Columns
  - RFE
  - PCA
- Logistic regression with Cross Validation
- Decision tree model
- Random Forests
- SVM

Model Building

- Logistic regression with RFE
- RFE with 15 columns
- Column elimination based on P and VIF

Metrics

- Important variables are Income, No of months in current residence, No of months in current company, Gender, No of dependents, Profession, Type of residence
- Accuracy of the model around 80% but sensitivity is low. We will check for these metrics for the model with all data (demographic and Credit bureau)

Model Building – Combined Data

Model Selection

- Try Logistic regression as the initial model with following feature selection options
  - IV Columns
  - RFE
  - PCA
- Logistic regression with Cross Validation
- Decision tree model
- Random Forests
- SVM

Model Building – Combined Data

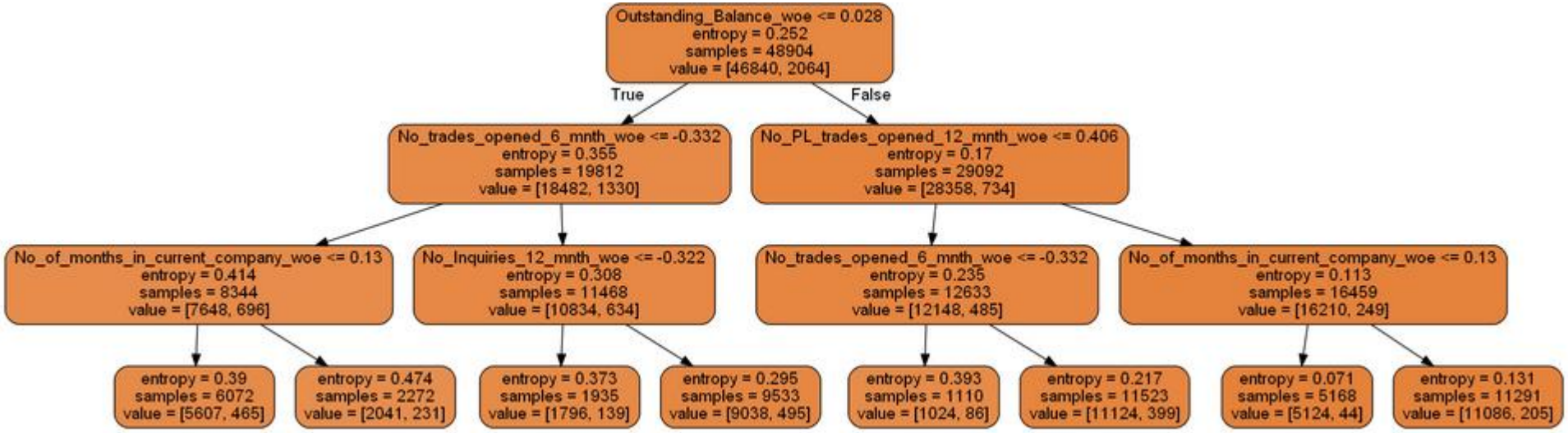
Model Building & Metrics - LR

- Logistic Regression model on IV columns which are medium to strong predictor (IV between 0.1 and 0.5)
- Results are not great, accuracy is 75% but sensitivity is low at 6%
- Since IV is not yielding good fit, we tried RFE with 15 columns
- Results are not that great with accuracy of 67% with sensitivity of 10%
- Based on the RFE fit, we will try PCA to seek better fit
- PCA is a better fit in Logistic regression
- 14 Principle components explain over 95% of the variables.
- LR over PCA has an accuracy of 60% and sensitivity of 67%
- The probability cutoff for logistic regression over the RFE variables, IV variables and the PCA has been 0.5 where the sensitivity, specificity and accuracy curves cut each other.

Model Building – Combined Data

Model Building & Metrics – Decision trees

- Tried Decision trees and Random Forests to fit the data
- Decision Trees - Hyper parameters tuning for Criterion, min_samples_leaf , min_samples_split and max_depth.
- Grid search and ran the model for best hyper parameters [Criterion = entropy, Max_depth = 3, min_samples_split = 50, min_samples_leaf = 18]
- The decision tree does not classify any application as default therefore not a useful model. This has a good accuracy 95% , very high precision of 96% but very low recall for default applicants which is required to be high.



Model Building – Combined Data

Model Building & Metrics – Random Forest

- Grid search and run the model for best hyper parameters for Random Forests
- Random Forests with
    - 'max_depth': [7],
    - 'min_samples_leaf': range(100, 150, 170),
    - 'min_samples_split': range(300, 350, 400),
    - 'n_estimators': [500, 900],
    - 'max_features': [8, 13, 14]
- RF Accuracy and sensitivity of 95% and 0.5% respectively therefore a bad model to choose.

Model Building – Combined Data

- This takes a very long time to run on this dataset.

Model Building & Metrics – SVM

Model Conclusions

Conclusions

- PCA with Logistic regression is the best model
- Accuracy and sensitivity is 60 & 67% respectively
- Important predictor variables are
    - No of enquiries in the last 6 months
    - Avg Credit card utilization
    - Presence of open home loan
    - Income
    - No of months in current company
- We need higher sensitivity with a good accuracy hence we chose PCA
- Our goal is to correctly predict the defaults hence sensitivity is a key aspect for our modeling
- Cut off is chosen at 0.5 probability

Financial Benefit Analysis

Steps for Application scorecard

- Apply the model to the whole data set without performance tag
- Apply the model to whole data set with the performance tag (without test train split and smote)
- Calculate the score based on the below formula for both data sets
  - Odds = pred_prob / (1 - pred_prob)
  - Log_odds = np.log(odds)
  - 400 + (20 * (log_odds/ np.log(2)))
  - Positive likelihood = sensitivity / (1 - specificity) – for accepted dataset
  - Likelihood for rejected dataset = p1*p2*p3*(1-p4)*(1-p5)*(1-p6) [where p1, p2, p3 are defaulted data point probabilities, p4,p5,p6 are not defaulted data point probabilities]
- Build application scorecard for each application and assign scores
- Apply scores to both rejected and accepted application data sets

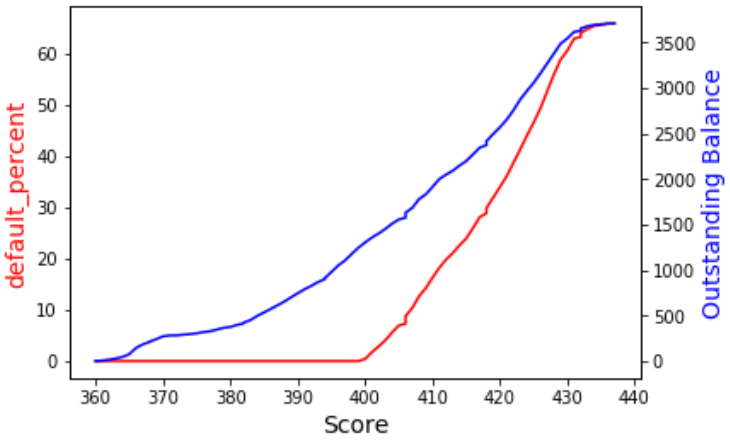
Financial Benefit Analysis

Assumptions

- Model is built using the data set of the accepted applicants only, i.e Performance tag is not null
- Model is built using the combined data set of Demographic and Credit Bureau data
- Some of the categorical variables are assigned unk where null values existed as WoE will scale that data
- Financial benefit analysis is made on the factor off how much outstanding balance could have been saved if the model was applied on the accepted candidates
- Another factor to determine the score would be to understand what percentage of applicants would be rejected based on the cutoff

Financial Benefit Analysis

Final Conclusions



- We have to chose a right score cut off to ensure the outstanding balance is minimal for the accepted candidates who default
- Also the cut off should ensure the right mix of reject applicants are rejected
- Based on the above criteria we have chosen the Score cut off of 430
- If we use the model to automate the application selection/rejection then
  - @ 430 cut off for the accepted applicants we would have recovered 3,483 Million in outstanding balance
  - This is about 95% of the total outstanding balance of the defaulters from the accepted applicants
  - 60% of the Actual defaulter application being rejected by the model
  - Applying the same cutoff to rejected applicants close to 75% of the applicants would have been rejected by the model
  - Positive likelihood on accepted dataset is 1.64.