# BFS CAPSTONE PROJECT

*Business presentation*

# Objective – Business Understanding

CredX is a leading credit card provider which experienced an increase in credit loss. The CEO believes that the best strategy to mitigate credit risk is to ‘acquire the right customers’.

Identify the right customers to acquire using predictive models.

Determine the factors affecting credit risk by applying predictive models.

Build an application scorecard and identify the cut-off score below which you would not grant credit cards to applicants.

Assess the financial benefit of our predictive models.

# Problem Solving Methodology

We followed CRISP-DM framework in the below sequence:

- Load and understand the data sets provided
- Exploratory Data Analysis(EDA)
  - Missing value Analysis
  - Outlier analysis and treatment
  - Bi-Variate Analysis (correlation and default percentage)
  - WOE  and IV analysis
- Model Building and Evaluation
  - Models – Logistic Regression, Decision Tree, Random Forest and SVM
  - Evaluation – Accuracy, Sensitivity, Specificity, KS Statistic and ROCR
- Application Score card generation
- Financial Impact Analysis

# Data Understanding

**Demographic data:** This is obtained from the information provided by the applicants at the time of credit card application and contain data like age, gender, marital status, income etc.

**Credit bureau data**: This is taken from the credit bureau and contains variables related to credit card usage, customer inquiries, transactions etc.

**Summary:**

- 71295 records in both the files with 12 columns(variables) in demographic data and 19 columns(variables) in credit bureau data.
- Application ID is the unique identifier i.e., primary key in both the files which will be used for merging.
- All the application ID's present in demographic data are part of bureau data and vice versa.
- Performance Tag is the target variable which says if customer is default(1) or not (0).
- 71292 unique records are present in both the data sets and 3 duplicate records(having same application ID) are found in both the data sets, which need to be removed.

## Missing/ NA values:

| Variable | Missing Percentages | |
|---|---|---|
| | Count | Percentage |
| Gender | 2 | 0.0028% |
| Marital Status | 6 | 0.0084% |
| No of dependents | 3 | 0.0042% |
| Education | 119 | 0.1669% |
| Profession | 14 | 0.0196% |
| Type of residence | 8 | 0.0112% |
| Avgas CC Utilization in last 12 months | 1058 | 1.4840% |
| No of trades opened in last 6 months | 1 | 0.0014% |
| Presence of open home loan | 272 | 0.3815% |
| Outstanding Balance | 272 | 0.3815% |
| Performance Tag | 1425 | 1.9988% |

1425 rows with no performance tag. We can assume that the applicant is not given credit card, hence they will be separated as rejected data.

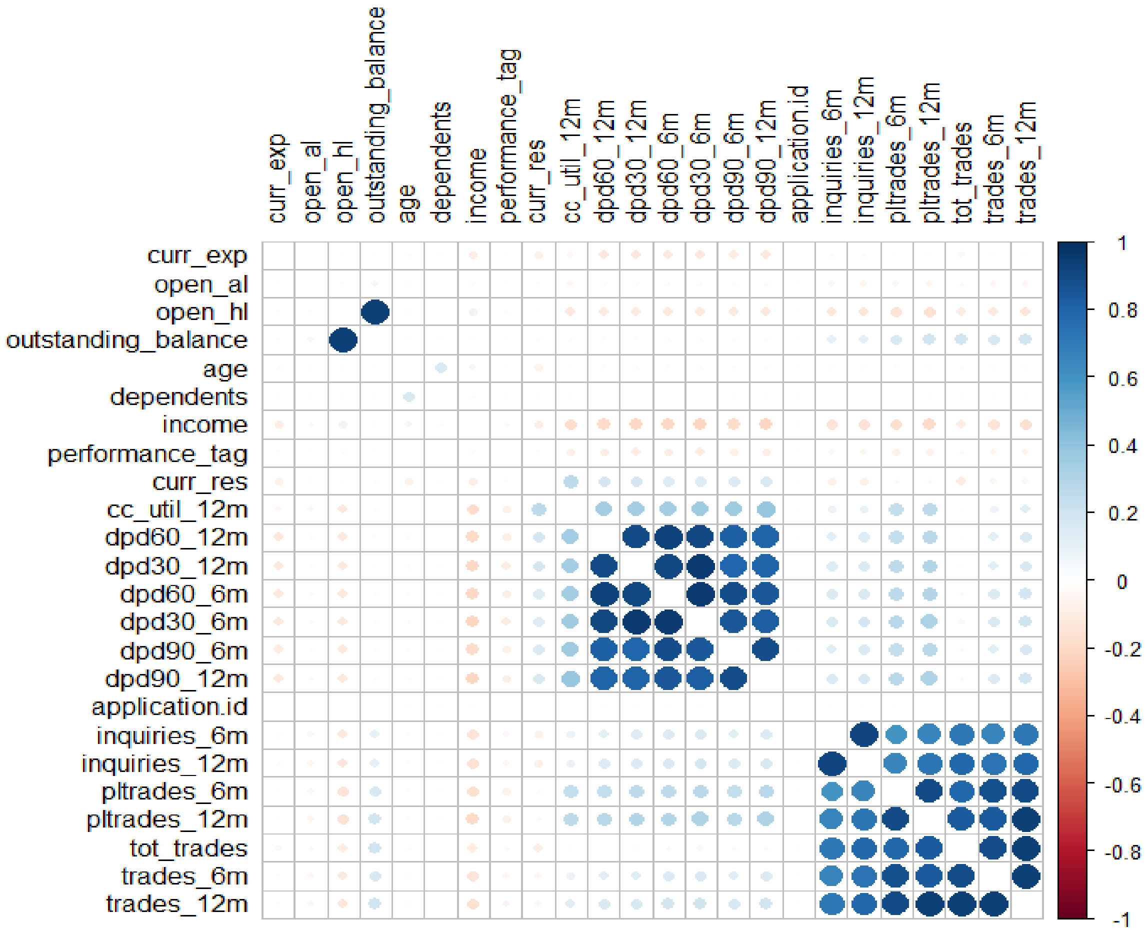
## Outliers:

Outlier treatment is done using box plots and quantile analysis for below variables:

- Age – less than 18 are fixed as 18 as the minimum age required is 18 for credit card issuing.
- Experience in current company capped to 72 months at 98th Percentile value.
- Income – negative income values are replaced with minimum value of 4.5
- 30 DPD in last 6 months capped to 6 as it is possible only 6 times in 6 months

Outliers are observed in some variables related to DPD and inquiries, but not doing any treatment as the values are practically possible to achieve.

# Univariate Analysis



High correlation is observed in below category of variables:

- All the DPD variables(30, 60 & 90) is last 6 months and 12 months
- Inquiries and trades(Normal trades, PL trades and total trades) in last 6 months and 12 months.
- DPD variables, inquiries and trades are moderately correlated to credit card utilization.
- Open home loans and outstanding balance are highly correlated to each other.

# Bi-Variate Analysis - Correlation

# Bi-Variate Analysis –

*Default Percentage, WOE & IV Analysis*

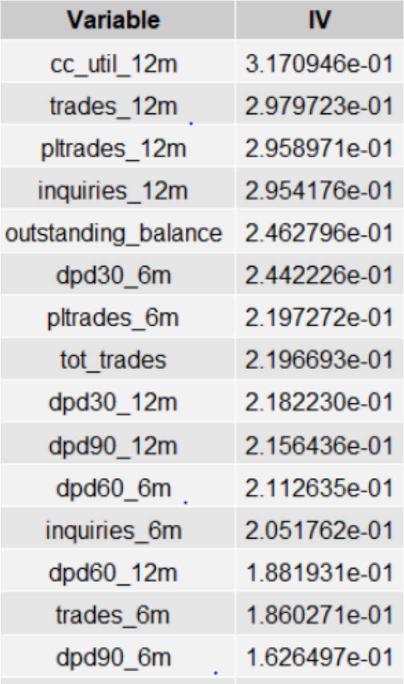
## Default Percentage Analysis

- Significant change in default rate is observed in Income, Months in current residence and company variables of demographic data.

- In credit bureau data, default rate is increasing with increase in below variables:
  - DPD (30, 60 and 90) in last 6 and 12 months
  - Average credit card utilization
  - Number of trades, PL trades and inquiries in last 6 and 12 months
- In other variables there is no consistent trend in the default ratio across categories

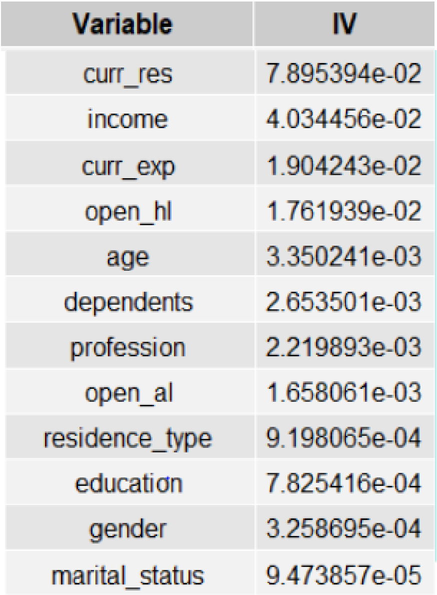
## WOE & IV Analysis

Strong and Medium Predicters



| Variable | IV |
|---|---|
| cc_util_12m | 3.170946e-01 |
| trades_12m | 2.979723e-01 |
| pltrades_12m | 2.958971e-01 |
| inquiries_12m | 2.954176e-01 |
| outstanding_balance | 2.462796e-01 |
| dpd30_6m | 2.442226e-01 |
| pltrades_6m | 2.197272e-01 |
| tot_trades | 2.196693e-01 |
| dpd30_12m | 2.182230e-01 |
| dpd90_12m | 2.156436e-01 |
| dpd60_6m | 2.112635e-01 |
| inquiries_6m | 2.051762e-01 |
| dpd60_12m | 1.881931e-01 |
| trades_6m | 1.860271e-01 |
| dpd90_6m | 1.626497e-01 |

Weak Predicters



| Variable | IV |
|---|---|
| curr_res | 7.895394e-02 |
| income | 4.034456e-02 |
| curr_exp | 1.904243e-02 |
| open_hl | 1.761939e-02 |
| age | 3.350241e-03 |
| dependents | 2.653501e-03 |
| profession | 2.219893e-03 |
| open_al | 1.658061e-03 |
| residence_type | 9.198065e-04 |
| education | 7.825416e-04 |
| gender | 3.258695e-04 |
| marital_status | 9.473857e-05 |

# Models Built & Evaluation

- Logistic Regression:
  - Model 1: Demographic data
  - Model 2: WOE transformed combined data with only IV predicted variables(Strong and medium)
- Decision Tree model on combined data
  - Model 1: WOE transformed data
  - Model 2: Combined data with missing value imputation by traditional approach*
- Random Forest model on combined data
  - Model 1: WOE transformed data
  - Model 2: Combined data with missing value imputation by traditional approach*
- SVM Model on combined data
  - Linear Kernel
  - Radial Kernel
  - Polynomial Kernel

* Calculated the percentage of missing values and imputed them by deleting the rows as the % is very low(< 1.5%)

## Model Evaluation:

| Model | Acc. | Sens. | Spec. | KS | ROCR |
|---|---|---|---|---|---|
| Logistic Regression with WOE | 0.640 | 0.633 | 0.640 | 0.274 | 0.637 |
| Decision Tree with WOE | 0.614 | 0.617 | 0.563 | 0.180 | 0.590 |
| Decision Tree with Normal Data | 0.569 | 0.620 | 0.566 | 0.186 | 0.593 |
| Random Forest with WOE | 0.580 | 0.623 | 0.578 | 0.201 | 0.601 |
| Random Forest with Normal Data | 0.604 | 0.604 | 0.598 | 0.202 | 0.601 |

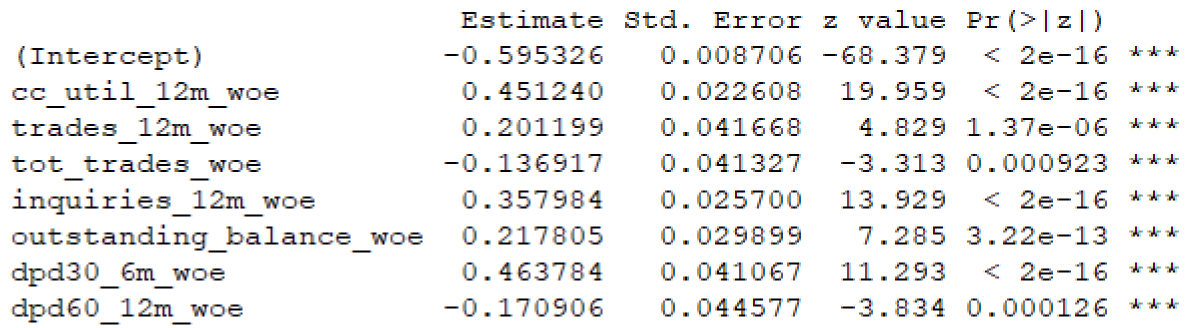
Based on the evaluation parameters **Logistic regression** with WOE and IV selected variables is considered as the final and selected model.

**Note:**

- It is very hard to conclude anything from SVM model as the evaluation parameters are not consistent and modelling is done with a very small data set. Hence not considered for final evaluation.
- Logistic Regression with demographic data is not considered for evaluation as the important predictors returned from the model are only two and evaluation parameters are having a very low values.

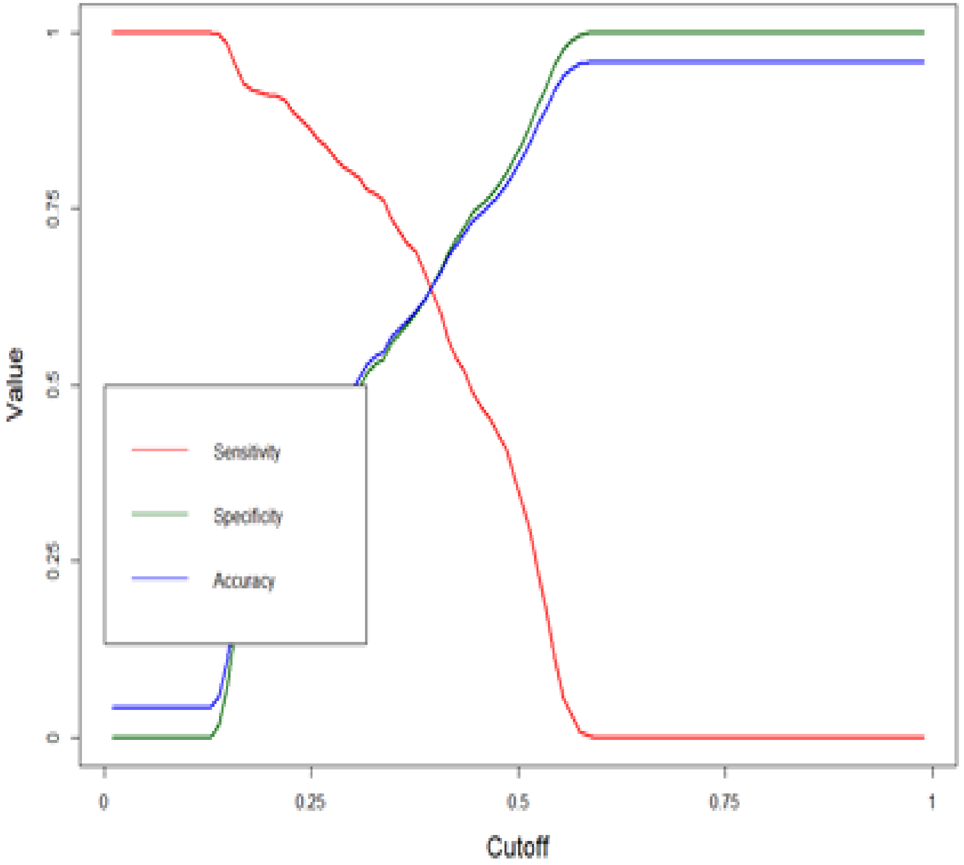
# Selected Model Analysis

Logistic regression Model on combined data with WOE transformation is selected as the approved model as the evaluation parameters are better when compared with other models.



```
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.595326   0.008706 -68.379  < 2e-16 ***
cc_util_12m_woe         0.451240   0.022608  19.959  < 2e-16 ***
trades_12m_woe          0.201199   0.041668   4.829 1.37e-06 ***
tot_trades_woe         -0.136917   0.041327  -3.313 0.000923 ***
inquiries_12m_woe       0.357984   0.025700  13.929  < 2e-16 ***
outstanding_balance_woe 0.217805   0.029899   7.285 3.22e-13 ***
dpd30_6m_woe            0.463784   0.041067  11.293  < 2e-16 ***
dpd60_12m_woe          -0.170906   0.044577  -3.834 0.000126 ***
```

**Variables that define default**



- Customers with predicted probability greater than or equal to 0.396 are most likely to default.
- Any increase in the credit card utilization, inquiries and Trades in last 12 months, 30dpd in last 6 months and outstanding balance, increases the probability of customer becoming default.

# Application Score card

Using the logistic regression model application scorecard is built with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points.

**Steps:**

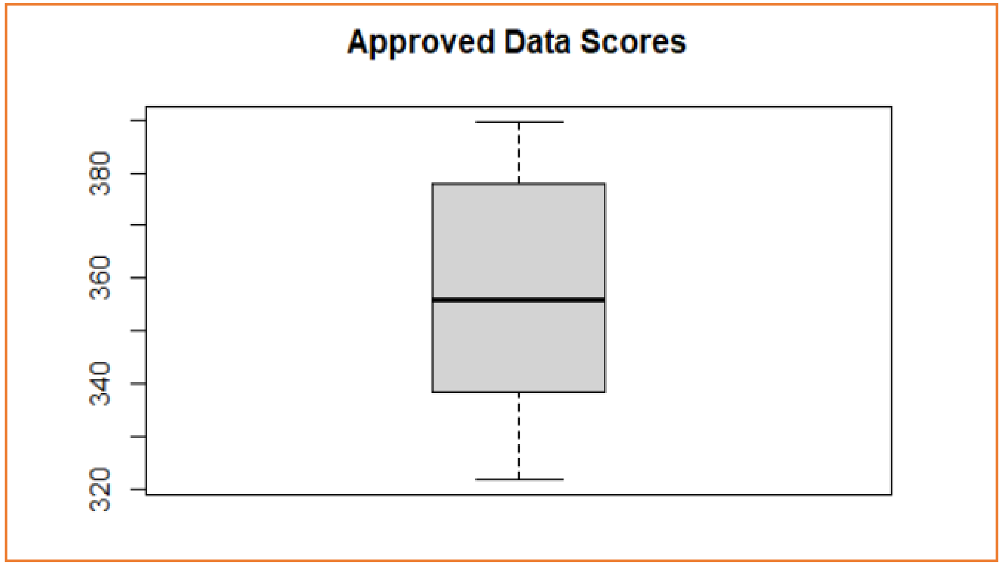
1. Probability of default for all applicants is calculated(P) for both accepted and rejected customers
2. Probability of non default is calculated (1-P)
3. Odds were calculated for all customer [(1-P)/P]
4. Used the following formula for computing application score card:

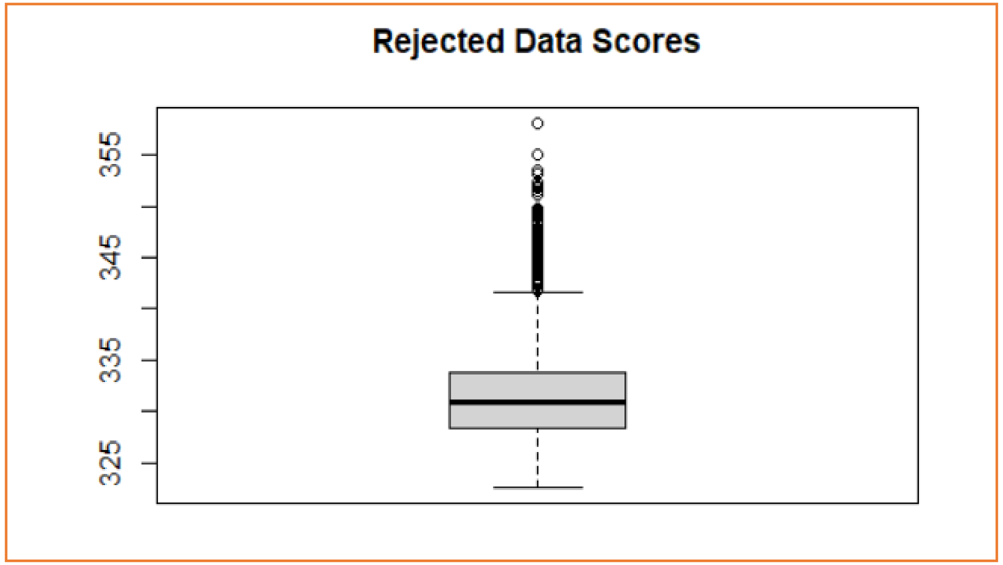
   Score = 400 + slope * log(odds) where slope is 20/(ln(20)-ln(10))
5. Using the above formula Scores are calculated for both accepted and rejected.
6. Using the same formula, cut off score is identified at cut off probability

**Summary:**

- **Cut off score is identified as 345.73 – below which credit card will not be granted.**
- Scores are ranging from 321.8 to 389.7 in approved data with median of 356.
- Scores are ranging from 322.5 to 358.2 in approved data with median of 330.8
- Based on score card, in rejected data 1385 are default and 40 are non-default.

**Model is correctly identifying 97% of the rejected customers correctly.**





# Financial Benefit Analysis

|  | Actual |  |
|---|---|---|
| **Predicted** | Default | Non-Default |
| Default | 1805 | 23969 |
| Non-Default | 1142 | 42951 |

Total : 69867
Actual Defaulters : 1805 + 1142 = 2947
Actual Non-Default : 23969 + 42951 = 66920
Rejected – Model : 1805 + 23969 = 25774
Accepted by Model : 1142 + 42951 = 44093

**Assumptions:**

## Avg credit loss per default : 1500$
## Acquisition cost per customer : 5 $
## Avg yearly revenue from non default customer: 100$
## Yearly profit = Revenue - total acquisition cost - total credit loss

**Model Approval Rate = 44093/69867 = 63.1%**

**Financials with out model:**

Acquisition Cost(A) = 69867 * 5 = 349,335 $
Credit Loss (B) = 2947 * 1500 = 4,420,500 $
Revenue (C) = 66920 * 100 = 6,692,000 $
Profit (P1 = C-A-B) = 1,922,165 $

**Financials with model:**

Acquisition Cost(D) = 44093 * 5 = 220,465 $
Credit Loss (E) = 1142 * 1500 = 1,713,000 $
Revenue (F) = 42951 * 100 = 4,295,100 $
Profit (P2 = F-E-D) = 2,361,635 $

**Model Implications:**

Credit Loss Saved (B – E) = 2,707,500 $
Revenue Loss (C – F) = 2,396,900 $
Profit Gain = P2 – P1 = **439,470 $**
Profit Gain % = [(P2 – P1)/P1] * 100 = **22.86%**

**Even though there is a considerable revenue loss, CredX will have an increased profitability by 22.86%.
This is achieved because of savings in acquisition cost and potential credit loss saved by using model automation.**

# Conclusion/Recommendations

We recommend CredX to implement our logistic regression model to mitigate credit risk by acquiring the right customers which would result in:

- Increase in company profit by  22.86%
- Reduced acquiring costs and credit loss

# Thank You!!!