8/1/2020

Predicting Delinquent Customer

Kamaldeep Kaur
Vinayak Balachandra Menon
Xinkai Zhou

Table of Contents

Executive Summary

It was August 10th, 2020. “What a hectic day,” Vinayak Menon said to himself as he sipped his afternoon tea. As he drank his tea, Menon recalled how much his company CredX has achieved in the past few years but still experiencing an increase in Credit loss. He called an urgent meeting with his data scientists Kamaldeep Kaur and Xinkai to analyze the losses in the company to mitigate the credit risk. As a CEO of a leading credit card provider CredX, He believes that the company should acquire the right customers to diminish the credit loss. He asked his data scientists to use some predictive models to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk, and assess the financial benefit of your project. He asked them to use the past data of the bank’s applicants. Two datasets need to be used in the analysis i.e. the information provided by the applicants at the time of Credit card application contains customer-level information on age, gender, income, marital status, etc, another one is data that is taken from Credit bureau and contains variables such as 'number of times 30 DPD or worse in last 3/6/12 months', 'outstanding balance', 'number of trades', etc. His team performed some data quality checks using woe and IV statistics, where they realized that the parameters in the demographic data don’t play much significant role in prediction and most of the significant variables are from Credit Bureau data. This information was not enough to determine the factors affecting credit loss so they move ahead with further analysis.

Understanding the data: There are two data sets in this project: demographic and credit bureau data.

Demographic/application data: This is obtained from the information provided by the applicants at the time of credit card application. It contains customer-level information on age, gender, income, marital status, etc.

1. It contains 71295 customers data including 12 variables.
2. The age group of most of the customers is between 35 to 55.
3. The number of male applicants in the bank is thrice than the number of female applicants i.e. 76% and 24% respectively.
4. Out of 71295 applicants, almost 60,000 applicants are married.
5. Almost 55000 applicants are staying at rented houses whereas approx. 15000 applicants owned their own houses.
6. There are 150 N/A values in the data and 3 non- unique Application ID with 6 records found.

Credit bureau: This is taken from the credit bureau and contains variables such as 'number of times 30 DPD or worse in last 3/6/12 months', 'outstanding balance', 'number of trades', etc.

1. This file also contains 71295 observations with 19 variables.
2. Both files contain a **performance tag,** which indicates whether the applicant has gone 90 days past due (DPD) or worse in the past 12 months (i.e. defaulted) after getting a credit card.
3. The application ID is a common ID between the two datasets.

Issues and Challenges:

Immediate issues facing the data scientists were data quality issues that need to fixed before performing the analysis on data.

- **Checking for duplicates in data:** 3 non-unique Application Id values (6 records) found. In the data, one customer has negative age (age was -3). 19 customers have age as 0. 45 customers have age less than 18 (assuming a person with age less than 18 is not eligible for applying for the card). The age for all the above record was set as 18.
- **Checking for N/A values:** There are 150 N/A values in demographic data and 1568 in Credit Bureau data. We need to perform the data quality checks using woe and IV Analysis.

WOE and IV Analysis:

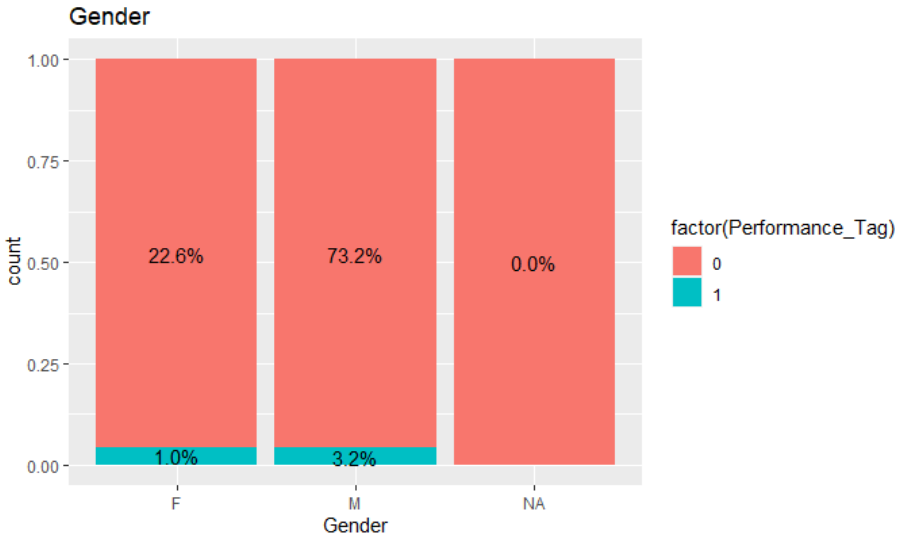
To remove the outliers and missing values, Data scientist team used the Woe function. WOE can handle the missing values by creating separate bins for missing values. There are several unavailable entries (NA) and outlier values within the data, WOE function will replace these values with the respective WOE values, indicating a relationship with the respective target value. For this, the scorecard package can be used.

According to the IV Analysis,They observed that demographic data is not that useful in the prediction as compared to credit bureau data. Generally, IV values more than 0.5 have strong predictive power.

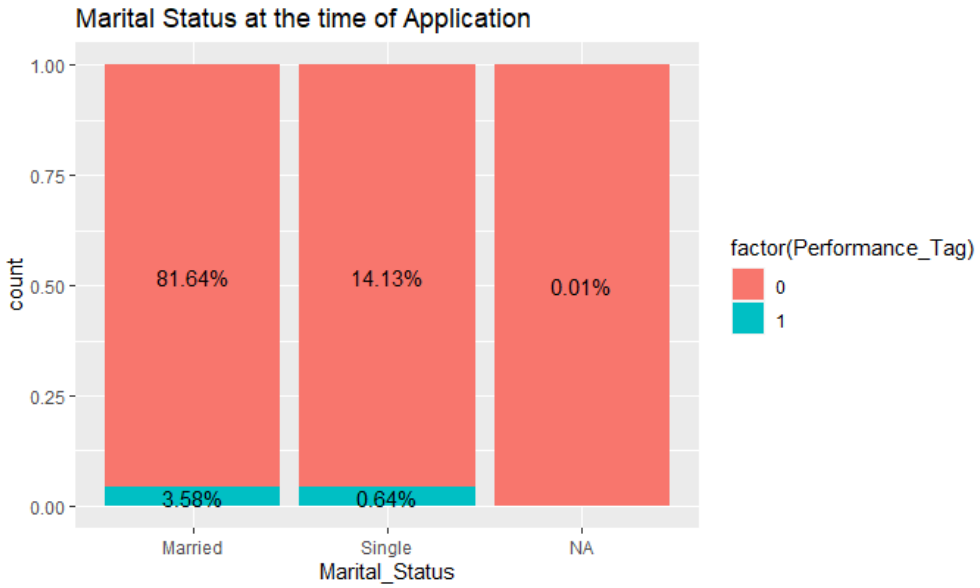
```
##                                                  Variable   IV
## 17                       Avgas.CC.Utilization.in.last.12.months 0.30
## 21                    No.of.PL.trades.opened.in.last.12.months 0.29
## 19                     No.of.trades.opened.in.last.12.months 0.27
## 23 No.of.Inquiries.in.last.12.months..excluding.home...auto.loans. 0.27
## 13                  No.of.times.30.DPD.or.worse.in.last.6.months 0.24
## 25                                      Outstanding.Balance 0.24
## 26                                     Total.No.of.Trades 0.24
## 16                 No.of.times.30.DPD.or.worse.in.last.12.months 0.22
## 20                       No.of.PL.trades.opened.in.last.6.months 0.22
## 12                  No.of.times.60.DPD.or.worse.in.last.6.months 0.21
## 14                 No.of.times.90.DPD.or.worse.in.last.12.months 0.21
## 15                 No.of.times.60.DPD.or.worse.in.last.12.months 0.19
## 22  No.of.Inquiries.in.last.6.months..excluding.home...auto.loans. 0.19
## 18                       No.of.trades.opened.in.last.6.months 0.18
## 11                  No.of.times.90.DPD.or.worse.in.last.6.months 0.16
## 9                         No.of.months.in.current.residence 0.09
## 5                                                 Income 0.04
## 10                      No.of.months.in.current.company 0.03
## 24                           Presence.of.open.home.loan 0.02
## 1                                                  Age 0.01
## 2                                               Gender 0.00
## 3                  Marital.Status..at.the.time.of.application. 0.00
## 4                                       No.of.dependents 0.00
## 6                                             Education 0.00
## 7                                            Profession 0.00
## 8                                        Type.of.residence 0.00
## 27                              Presence.of.open.auto.loan 0.00
```

Exploratory Data Analysis:
Demographic data:

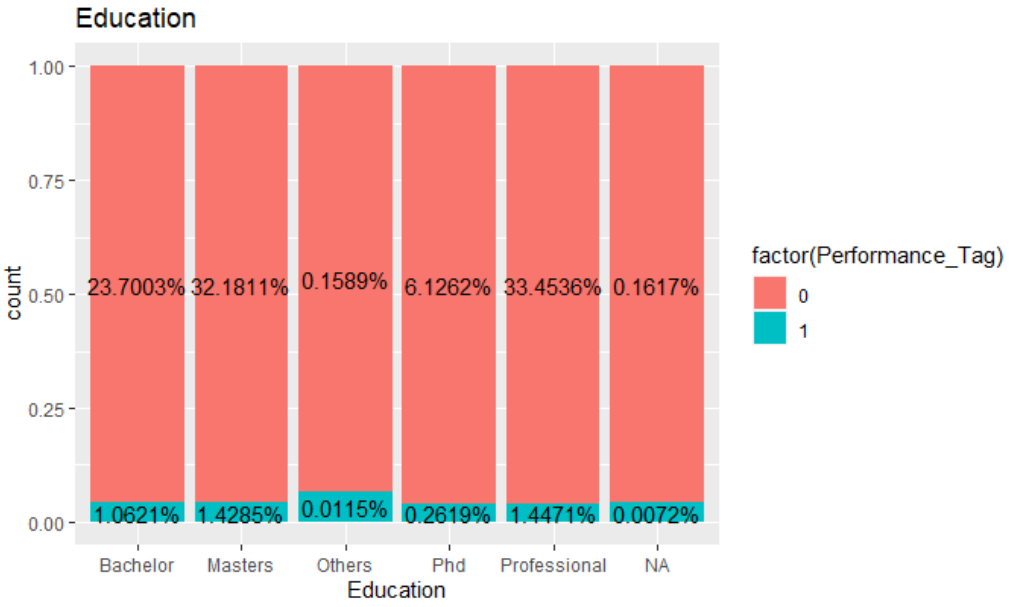
- There are almost 1425 customers who are defaulters in the data that give us a percentage of 4.2%.
- After checking the Gender Variable, It is clear that male applicants are more than female applicants but out of 24% female applicants, 1% of female applicants are defaulters.
- There is no such change in the defaulters in the male and female categories if we compare the count of both the customers.
- Out of 76%, 3.2% of male customers are defaulters.



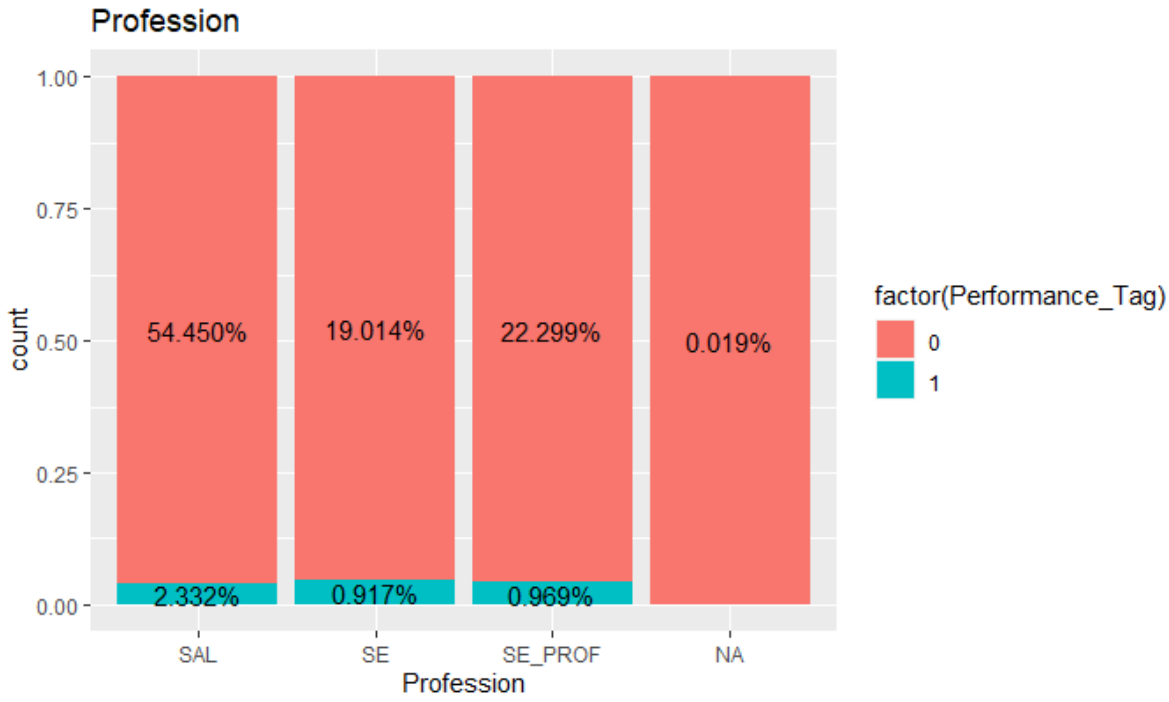
- After comparing the single and married data, 0.64% of single customers are defaulters whereas 3.58% are married defaulters.



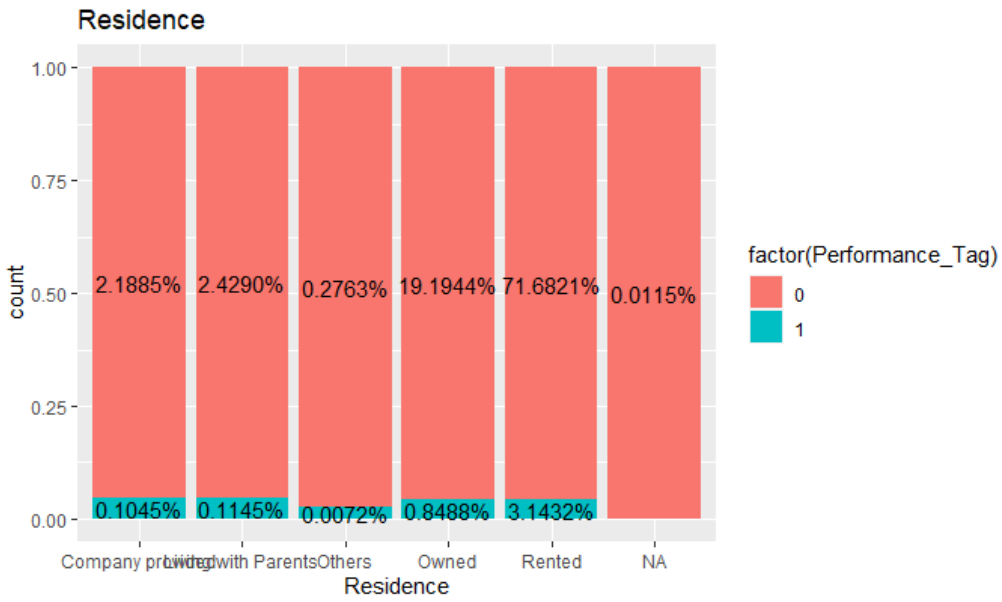
- Percentage of defaulters in Bachelors, Masters, Ph.D., Professional, and other customers are almost the same. Most of the N/A values are in education data i.e. 118.



- Most of the professional customers are salaried. 2.332% are defaulters inSAL, 0.917% in SE and 0.969% in SE_Prof.

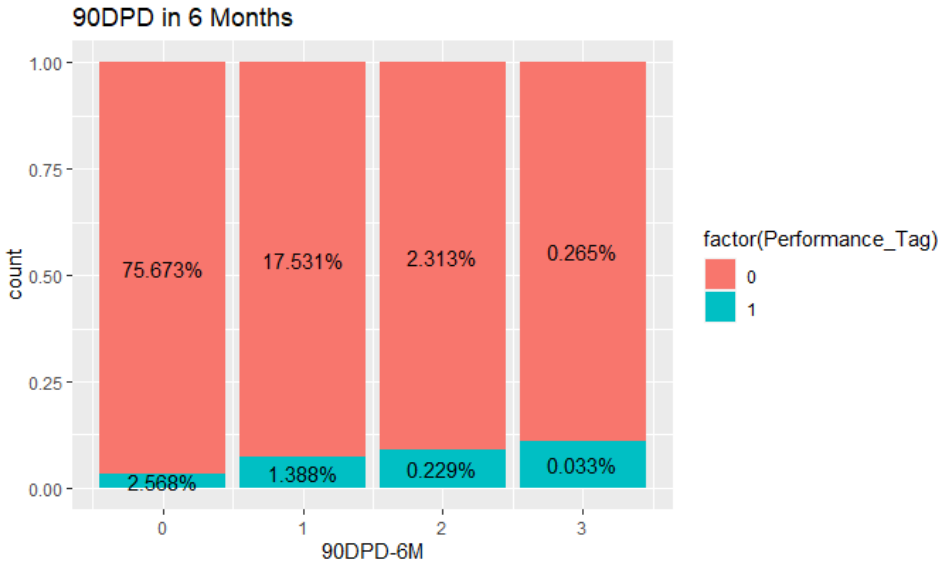


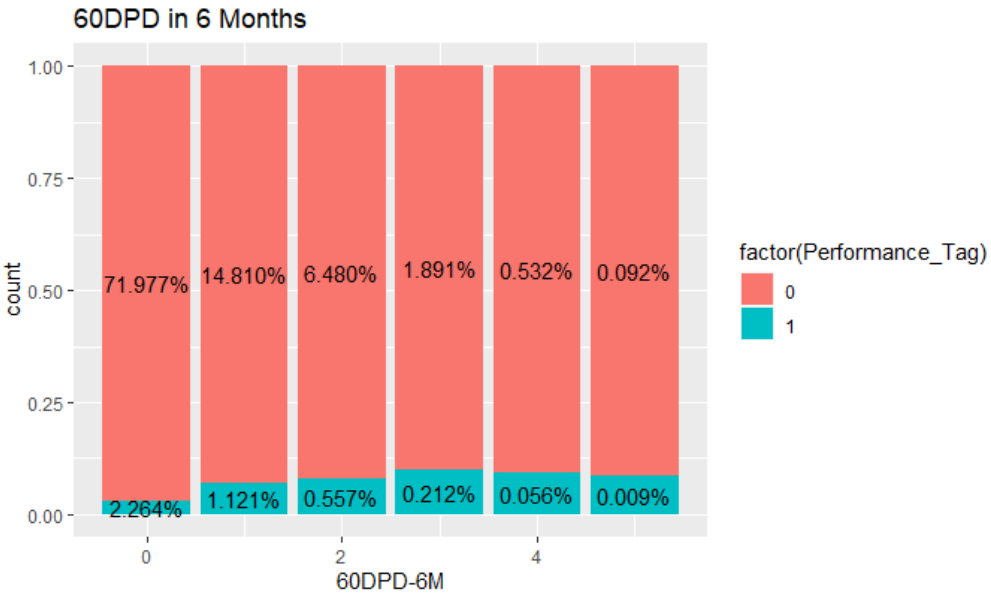
- Most of the customers are staying at rented homes whereas very few customers owned their houses.
- Most of the defaulters are in the rented house data that includes 3.14% of the defaulters.

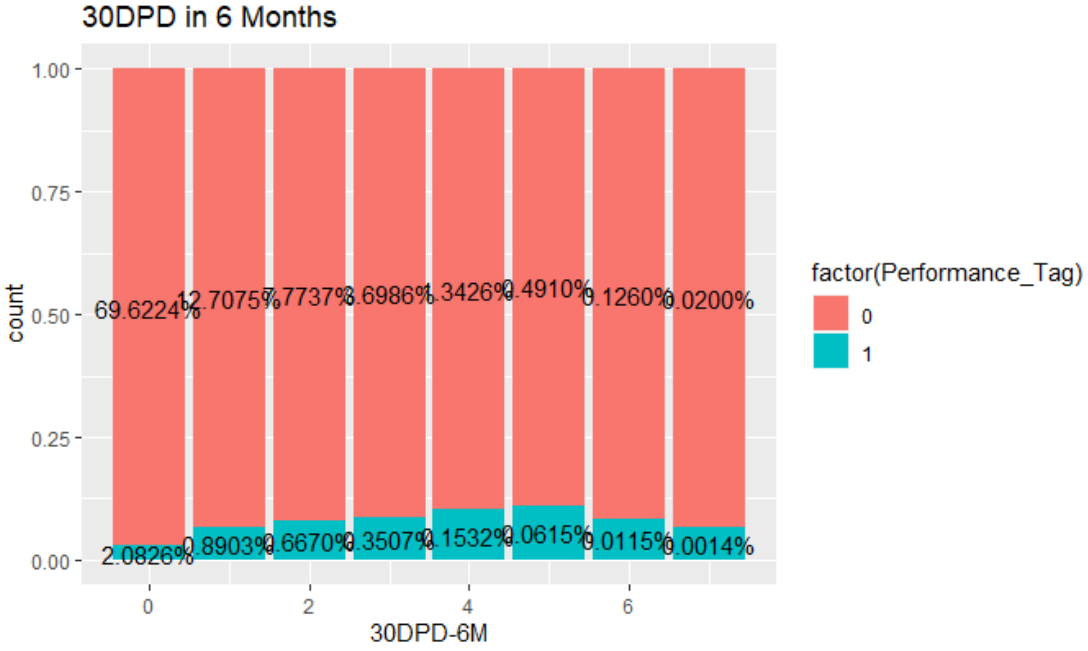


Credit Bureau Data:

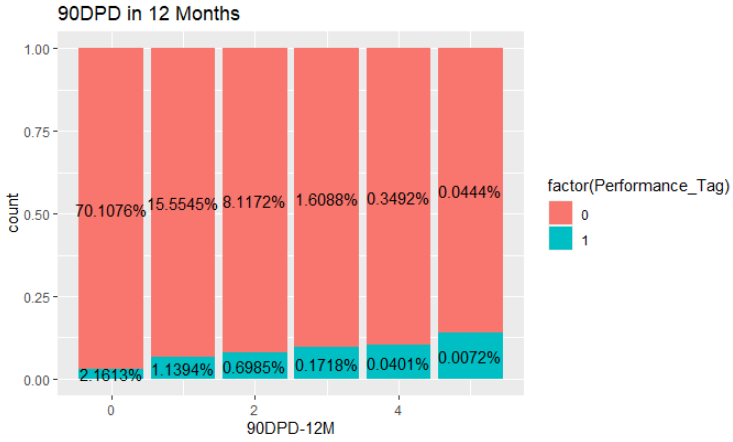
- Most of the default customers are those whose payment has gone 3 times 90 days past due in 6 months that gives 0.033% of the data.
- The percentage of the customers whose payment never gone past 90/60/30 days in 6 months are more and the order is increasing gradually.

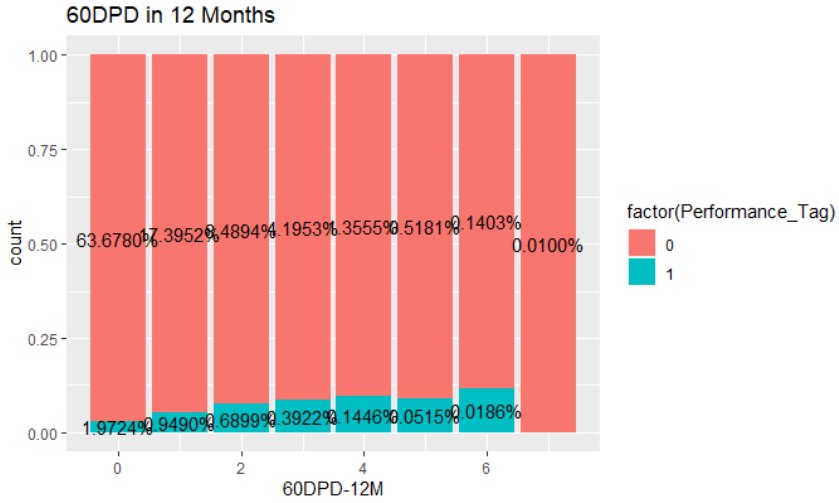


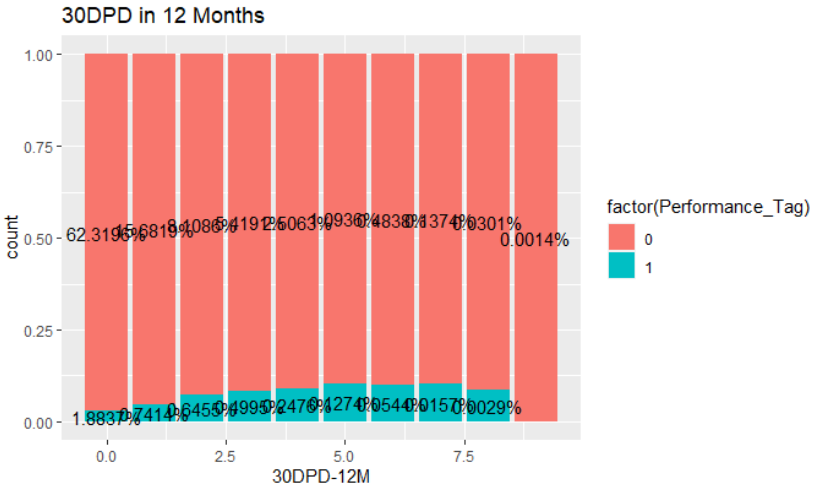

60DPD in 6 Months
count
1.00
0.75
0.50
0.25
0.00
71.977%
14.810%
6.480%
1.891%
0.532%
0.092%
2.264%
1.121%
0.557%
0.212%
0.056%
0.009%
0
2
4
60DPD-6M
factor(Performance_Tag)
0
1


30DPD in 6 Months
count
1.00
0.75
0.50
0.25
0.00
69.6224%
12.7075%
7.7737%
3.6986%
1.3426%
0.4910%
0.1260%
0.0200%
2.0826%
0.8903%
0.6670%
0.3507%
0.1532%
0.0615%
0.0115%
0.0014%
0
2
4
6
30DPD-6M
factor(Performance_Tag)
0
1

- There are 0.0072% of defaulters whose payment goes beyond 5 times in 90 days in 12 months containing more defaulters as compared to 0 to 4 times late payment.

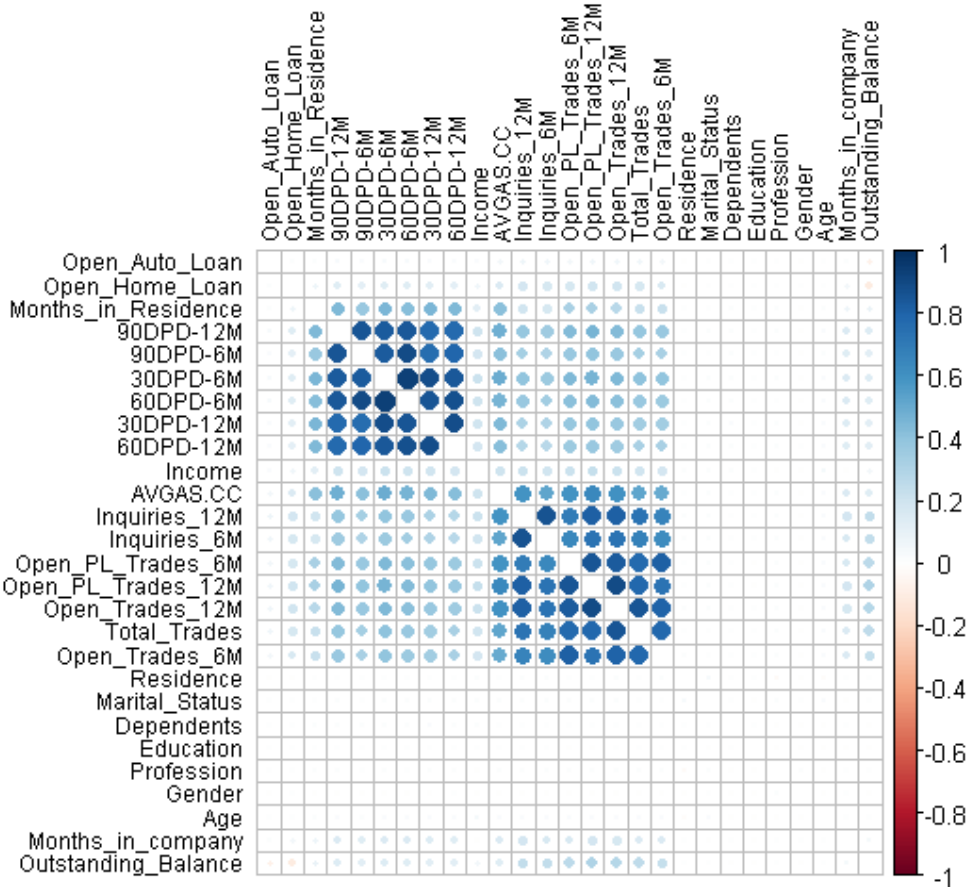






Correlation of data:

Finding the correlation between the features or predictors is important in the model because we can use the correlation to make the predictions. Correlation takes values between 1 to -1. According to the screenshot, Credit Bureau data's variables are highly correlated i.e. 90/60/30DPD in 12 Months, 90/60/30 DPD in 6 months, AVGAS_CC, Inquiries in 12/6 months, Open PL trades in 6/12 months, open trades in 6/12 months. The value of Dependents, Education, Profession, Gender, Age, Months in company and outstanding balance have less than -0.6 that are low correlated.



Treatment on outliers:

It's easy to detect the outliers in boxplot. Outliers generally lie outside 1.5*IQR where "Inter Quartile Range" is the difference between 75th and 25th quartiles.

Principle Component Analysis:

Data Scientist team analyses the data and correalation and they got to know that the data is wide. Its better to use the principle Component analysis to transform the data into new coordinate and remove the unnecessary variables. After checking the PCA on demographic data, then it gave the variance of 11.55% of first component. They obtain to find the component which gave the maximum variance to retain as much as information possible. The higher the variance, higher will be the information in those components.

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5     PC6     PC7
## Standard deviation     1.0749 1.0391 1.0186 1.0119 1.0033 0.99744 0.99484
## Proportion of Variance 0.1155 0.1080 0.1037 0.1024 0.1007 0.09949 0.09897
## Cumulative Proportion  0.1155 0.2235 0.3273 0.4296 0.5303 0.62981 0.72878
##                            PC8     PC9    PC10
## Standard deviation     0.97420 0.97100 0.90571
## Proportion of Variance 0.09491 0.09428 0.08203
## Cumulative Proportion  0.82369 0.91797 1.00000
```

Modelling on Raw data:

Linear Discriminant Analysis on demographic data:

To obtain more results, Data scientists performed the LDA modelling on the raw data by adding different variables each time.

By using Months_in_residence, Income, Months_in_company, Education variables with the performance tag, the model gave 95% accuracy.

```
## Call:
## lda(Performance_Tag1 ~ Months_in_residence + Income + Months_in_company +
##     Education, data = demcopy_train)
##
## Prior probabilities of groups:
##         0         1
## 0.9583381 0.0416619
##
## Group means:
##   Months_in_residence      Income Months_in_company    Education
## 0        -0.001159309  0.00782569       0.008438014 3.635722e-05
## 1         0.083846072 -0.18561785      -0.086951754 4.421433e-02
##
## Coefficients of linear discriminants:
##                            LD1
## Months_in_residence  0.2468127
## Income              -0.8425966
## Months_in_company   -0.4694156
```

According to the model, Linear Discriminant Analysis provide the Group mean, probabilities and coefficient for each individual parameter.

1. Probabilities: We can see that there are 95.83% and 4.16% of performance tag 0 and 1 in the Group respectively.
2. Group Mean: It shows the group center of gravity of all the variables.
3. Coefficients: It provide us the combination of all the variable that are required to form the LDA decision. For eg: LD1 = Months in residence*0.2468127 + Income*-0.8425966 +Months in company* -0.4694156+ Education*0.1885182

K-Fold Cross Validation:

K-fold cross validation is procedure to evaluate the performance of the algorithm on different dataset. In this technique, we reserve the sample of dataset and then train the model on remaining part of the dataset.
K-fold algorithm is as follow:
1. We have selected k fold number=10 and randomly split the data into k-subsets (here value=10)
2. Reserve one set and train the model on all other sets. In first iteration, the first fold is used to test the model and others are used to train the model. In second iteration, second fold is used to test the model and others are used to train the model.
3. After testing the model on reserve subset, we can record the prediction error. This process will repeat 10 times until all the folds have been used as testing set.

Data Scientist team used the same model with 4 predictors to check the validation. Kappa is generally used to measure the score of homogeneity or consensus exists in the ratings. It records the possibility of the agreement occurring by chance. From 0.81-1.0 gives the most perfect agreement. Kappa of our model is 0 that means agreement equal to chance. Model is giving the accuracy of 95%.

```
## Linear Discriminant Analysis
##
## 52398 samples
##     4 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (7 fold)
## Summary of sample sizes: 44912, 44912, 44912, 44914, 44912, 44913, ..
## Resampling results:
##
##   Accuracy   Kappa
##   0.9583381  0
```

Linear Discriminant Analysis on whole data

They again run the model using 8 predictors including Open_Trades_12M, Open_PL_Trades_12M, Inquiries_12M, 30DPD-6M, Open_`90DPD-12M, 60DPD-6M and Inquiries_6M. The probability of Performance tag 0 and 1 have been changed slightly. Performance tag 0 is 95.73% and 1 is 42.67%.

```
## lda(Performance_Tag ~ Open_Trades_12M + Open_PL_Trades_12M +
##     Inquiries_12M + `30DPD-6M` + Open_PL_Trades_6M + `90DPD-12M` +
##     `60DPD-6M` + Inquiries_6M, data = demcredit_train)
##
## Prior probabilities of groups:
##          0          1
## 0.95732662 0.04267338
##
## Group means:
##   Open_Trades_12M Open_PL_Trades_12M Inquiries_12M  `30DPD-6M`
## 0     -0.01435739         -0.0174909   -0.01227104 -0.02325777
## 1      0.28520593          0.3773948    0.26894013  0.48466772
##   Open_PL_Trades_6M `90DPD-12M`  `60DPD-6M` Inquiries_6M
## 0       -0.01647643 -0.02045018 -0.02225001 -0.009559173
## 1        0.33987362  0.45482147  0.45671140  0.229374437
##
## Coefficients of linear discriminants:
##                            LD1
## Open_Trades_12M    -0.6737568
## Open_PL_Trades_12M  0.7666290
## Inquiries_12M       0.3656019
## `30DPD-6M`          0.7080951
## Open_PL_Trades_6M   0.1898576
## `90DPD-12M`         0.2254565
## `60DPD-6M`         -0.2189451
## Inquiries_6M       -0.1835326
```

K fold Cross Validation:

```
## Linear Discriminant Analysis
##
## 52398 samples
##     8 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (7 fold)
## Summary of sample sizes: 44913, 44913, 44912, 44912, 44913, 44913, .
## Resampling results:
##
##   Accuracy   Kappa
##   0.9573266  0
```

LDA model 2 gave the accuracy of 95.73% and Kappa values are again zero. When they analyse the results of both the models, they know that it is still difficult to predict the good and bad customers and checking the model accuracy. So, They move ahead with Logistic regression.

Logistic Regression demographic data

```
##
## Call:
## glm(formula = Performance_Tag ~ Months_in_residence + Income +
##     Months_in_company + Education, family = binomial, data = demcopy_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4820  -0.3118  -0.2859  -0.2610   2.7616
##
## Coefficients:
##                     Estimate Std. Error  z value Pr(>|z|)
## (Intercept)         -3.16129    0.02240 -141.126  < 2e-16 ***
## Months_in_residence  0.05541    0.02145    2.583  0.00979 **
## Income              -0.20338    0.02267   -8.971  < 2e-16 ***
## Months_in_company   -0.11141    0.02223   -5.011 5.42e-07 ***
## Education            0.03932    0.01924    2.043  0.04102 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

There are 4 predictors in logistic regression that shows logistic equation like this:

logit(p) = log(p/(1-p))= β0 + β1*x1 + … + βk*xk

Each estimated coefficient is the expected change in the log odds of being in an honors class for a unit increase in the corresponding predictor variable holding the other predictor variables constant at certain value. Each exponentiated coefficient is the ratio of two odds, or the change in odds in the multiplicative scale for a unit increase in the corresponding predictor variable holding other variables at certain value. Here is the equation:

logit(p) = log(p/(1-p))= β0 + β1*Months in residence + β2*Income + β3*Months in company + β4*Education.

The coefficient of Months in residence says holding Income, Months in Company and Education at a fixed value. the odds of getting into an honors class for Months in residence(Months in residence = 1)over the odds of getting into an honors class for males (Months in residence = 0) is exp(.05541)=1.05

K fold cross validation

By performing the k fold cross validation using the same model gave the accuracy of 95.83%.

```
##
## No pre-processing
## Resampling: Cross-Validated (7 fold)
## Summary of sample sizes: 44912, 44912, 44912, 44914, 44912, 44913, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.9583381  0
```

Logistic regression whole data

They again run the model using 8 predictors including Open_Trades_12M, Open_PL_Trades_12M, Inquiries_12M, 30DPD-6M, Open_PL_Trades_6M and 90DPD-12M.

```
## Call:
## glm(formula = Performance_Tag ~ Open_Trades_12M + Open_PL_Trades_12M
##     Inquiries_12M + `30DPD-6M` + Open_PL_Trades_6M + `90DPD-12M`,
##     family = binomial, data = demcredit_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8420  -0.3245  -0.2522  -0.2215   2.8131
##
## Coefficients:
##                    Estimate Std. Error  z value Pr(>|z|)
## (Intercept)        -3.23426    0.02394 -135.080  < 2e-16 ***
## Open_Trades_12M    -0.41131    0.07531   -5.461 4.73e-08 ***
## Open_PL_Trades_12M  0.46473    0.06506    7.143 9.12e-13 ***
## Inquiries_12M       0.14072    0.03281    4.288 1.80e-05 ***
## `30DPD-6M`          0.19535    0.03011    6.489 8.66e-11 ***
## Open_PL_Trades_6M   0.11569    0.04853    2.384  0.01714 *
## `90DPD-12M`         0.09741    0.03114    3.129  0.00176 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

K fold cross validation

Using 6 predictors, Accuracy of model is 95.73.

```
## Generalized Linear Model
##
## 52398 samples
##     6 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (7 fold)
## Summary of sample sizes: 44913, 44913, 44912, 44912, 44913, 44913, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.9573266  0
```

Knn model

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique we generally look at 3 important aspects:

1. Ease to interpret output
2. Calculation time
3. Predictive Power

Implementation a KNN model by following the below steps:

1. Load the data
2. Initialise the value of k
3. For getting the predicted class, iterate from 1 to total number of training data points

    1. Calculate the distance between test data and each row of training data.
    2. Sort the calculated distances in ascending order based on distance values
    3. Get top k rows from the sorted array
    4. Get the most frequent class of these rows
    5. Return the predicted class

In this model, Using all the predictors, if k=5 then the accuracy is 95.62, If k=7 then the accuracy of model is 93.72 and k=9, accuracy of model is 95.73.

```
## k-Nearest Neighbors
##
## 52398 samples
##    27 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (7 fold)
## Summary of sample sizes: 44913, 44913, 44912, 44912, 44913, 44913, ...
## Resampling results across tuning parameters:
##
##   k  Accuracy   Kappa
##   5  0.9562960   1.190488e-03
##   7  0.9570976   3.613470e-04
##   9  0.9573075  -3.806243e-05
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 9.
```

Conclusion:

Logistic regression model gave the accuracy of 95.83 which is chosen as best model for this analysis.

Standard error of the model is low that means our values are significant.