Technical Documentation

Credit Risk Modeling:
Application Credit Score

A Protocol on Developing and Implementing an Application Credit Scorecard

Contact info:

Y. Staeva

E-Mail: julianastaeva@gmail.com

List of Contents

1. Introduction…….1

2. Model development

    2.1 Data Preparation……..2

    2.2 Data Modeling…….4

    2.3 Model Validation…….6

3. Application Score Card

    3.1 Classing of predictive variables and Score points…….10

    3.2 Automated Decision Threshold……..13

1. Introduction

Credit Score is a mathematical model used by financial institutions to determine whether a client is creditworthy or not. Application scorecards are made for the purpose of calculating the risk of a client not repaying their loan to a financial institution. Application scorecards are created specifically for clients who are applying for a loan. The aim of credit risk modeling is to accelerate the process of estimating a client's creditworthiness, and whether that client will be approved or denied a loan.

The Credit Score model in this project is developmed by using an example data set with client records within a bank. All data is provided by Experian.

In terms of technical resources, all steps of this project are performed in Google Colaboratory software environment. The programming language chosen for this project is Python 3.8.

2. Model development

The model development consists of four steps:
Data cleansing, Data preparation, Data modeling and Model Validation.
In this project the provided data with client records was already cleansed from empty/false/duplicated values, therefore the data cleansing step is intentionally skipped. The other steps of the model development process are described in detail in the next paragraphs.

2.1 Data Preparation

This model was constructed by using an example set of bank client data. The set contained c.a twenty-four thousand client records with various data, that can be divided into three categories:

- Data, describing the financial state of the client – Gross Annual Income, Time at work, Number of dependents, Residential status etc.
- Data, describing the client's relationship with the financial institution – Loan Amount, Loan Payment Frequency, Loan Payment method etc.
- Data, describing the client's creditworthiness, gathered by other financial institutions – Bureau Score, SP ER Reference, SP Number of searches etc.

During the data preparation stage an analysis of all variables in the data frame is performed in order to select the predictive features for the model. The first step of the model development is to select a target variable. This is the variable, that is to be predicted. In this project, the target variable is calculated by using the Current delinquency status and the Final decision parameters.
The target variable is calculated based on the following logic:

Current Delinquency Status	Final decision	Target
0	Accept	Good
1	Accept	Good
2	Accept	Indeterminate
3+	Accept	Bad
N/A	Accept	NTU
N/A	Decline	Rejects

Clients who are creditworthy, but haven't applied for a credit, are marked as NTU's. Clients, who are not creditworthy, are marked as Rejects. To predict the target variable, I use only data of clients that are marked as Good or Bad. The target variable for the model is converted into numeric metric – 1 for Good and 0 for Bad.

The next step of model development is the selection of predictive variables. All variables in the given data set were classified and their influence on the target variable was analyzed by means of probability of the client repaying their loan. For metric variables, such as Gross Annual Income or Time in Employment, the correlation coefficient with the target variable was measured. After a possible set of predictive variables was found, all variables were tested for correlation and dependence among each other. The final set of predictive variables included overall eight features, which are listed below:

Predictive Feature	**Information Value**
Bureau score	0,5127
Cheque card flag	0,2268
Occupation code	0,0570
Residential status	0,1365
Marital status	0,0534
Loan payment frequency	0,0909
Loan payment method	0,0717
Insurance required	0,0777

2.2 Data Modeling

The model is created by using multivariate linear regression. This modeling method is suitable for application and behavioral scoring. The model consists of one dependent variable and multiple independent ones. In this case, the dependent variable is the one that is to be predicted – the Target variable - and the independent variables are in the predictive set.

There are overall eight predictive variables, that will be used for data modeling. Seven of them are nominal, and one is metric. The metric variable was coarse classified, in order to avoid clustering in the final outcome of the model. Some of the predictive features have greater influence than others, however features with less Information value were intentionally included in the model. This is to ensure that no more than 5% of the population should become the same Credit score, otherwise the model does not differentiate good and bad clients well enough.

The model was constructed by using dummy variables for each variable class. Note, that the number of dummy variables created for each parameter should be N-1 , whereas N is the number of classes of the predictive variable. In this case, the dummy variables were generated automatically by the software environment.

The next step in data modeling is to divide the original data frame with client records into a train and test set. In this project the train : test proportion is 8 : 2, meaning that the training set included 80% of the original data frame and the test set contained 20% of the data. The original data frame is randomized before the split, so that the train and test sets would contain a balanced population of good and bad clients.

The generated model is presented in Fig. 1.

The Generalized Linear Model result show that all the selected predictive features are statistically significant, having a P value below 0.05. The first column with numerical data – coef – contains the multivariate regression coefficients for each variable. These are the coefficients, which will be used for the creation of the Score Card. There are no inversions in the coefficient results, meaning that each variable class has a monotonic coefficient score and can be used for the linear model.

For example, there is a monotonic dependency between the Bureau Score and the target variable – the higher the Bureau Score is, the higher the probability of a client returning their loan to the bank, therefore a higher regression coefficient is assigned with each class.

Generalized Linear Model Regression Results

Dep. Variable:	Num_Target	No. Observations:	14374
Model:	GLM	Df Residuals:	14361
Model Family:	Binomial	Df Model:	12
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3195.1
Date:	Fri, 18 Jun 2021	Deviance:	6390.3
Time:	07:27:03	Pearson chi2:	1.43e+04
No. Iterations:	100		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.6122	0.013	46.855	0.000	0.587	0.638
Bureau_Score_[636.0, 830.0)	-1.1097	0.097	-11.452	0.000	-1.300	-0.920
Bureau_Score_[830.0, 935.0)	-0.4672	0.073	-6.384	0.000	-0.611	-0.324
Bureau_Score_[935.0, 970.0)	0.2633	0.097	2.714	0.007	0.073	0.454
Bureau_Score_[970.0, 995.0)	0.5034	0.112	4.483	0.000	0.283	0.723
Bureau_Score_[995.0, 1020.0)	0.6701	0.140	4.780	0.000	0.395	0.945
Bureau_Score_[1020.0, inf)	0.7523	0.203	3.703	0.000	0.354	1.151
Cheque_Card_Flag_N	0.1170	0.041	2.878	0.004	0.037	0.197
Cheque_Card_Flag_Y	0.4952	0.039	12.857	0.000	0.420	0.571
Occupation_code_P|B|O	0.2914	0.039	7.402	0.000	0.214	0.369
Occupation_code_M	0.3208	0.040	7.971	0.000	0.242	0.400
Residential_Status_H	0.3385	0.044	7.643	0.000	0.252	0.425
Residential_Status_L|O|T	0.2737	0.047	5.862	0.000	0.182	0.365
Loan_Payment_Frequency_F|W|X	0.2970	0.043	6.900	0.000	0.213	0.381
Loan_Payment_Frequency_M	0.3152	0.044	7.147	0.000	0.229	0.402
Loan_Payment_Method_B	0.4676	0.038	12.202	0.000	0.392	0.543
Loan_Payment_Method_Q|S|X	0.1446	0.042	3.449	0.001	0.062	0.227
Insurance_Required_N	0.3679	0.037	9.910	0.000	0.295	0.441
Insurance_Required_Y|Nan	0.2443	0.039	6.217	0.000	0.167	0.321
Marital_Status_M	0.3211	0.042	7.557	0.000	0.238	0.404
Marital_Status_Other	0.2911	0.043	6.761	0.000	0.207	0.375

FIGURE 1. Generalized Linear Model – Results

The coefficients need to be logical for experts in the field of finances. There should be no inversions or contradictory scoring. For example , it is not logical to expect unemployed clients to be more likely of returning a loan, therefore they are assigned less score points, than clients, who are employed.

The regression coefficients are used in the score card in accordance to the following formula:

*Score Card points per variable class = Regression coefficient * 100.0*

All coefficient are multiplied by one hundred and rounded for convenience, when assigning points to the clients in the Score card. Further details about the final Credit Score and the predictive variables are given, when describing the Credit Score Card in section 3.

2.3 Model Validation

The purpose of model validation is to assure that the model can differentiate good and bad clients of different data samples.

To verify that the model calculates the client credit score correctly and that credit score is evenly distributed among the clients list, a Score distribution table was made. The overall score results of the original data frame was divided into 20 classes, with each class containing a score interval. The percentage of good, bad and total clients within that score class is also given in the score distribution report (see Fig. 2).

Target Score	Bad	Good	Bad_Rate	Good_Bad_Odds	Total_%
(116.0, 154.0]	133	569	18.95	4.28	4.89
(154.0, 193.0]	127	721	14.98	5.68	5.91
(193.0, 210.0]	78	517	13.11	6.63	4.15
(210.0, 218.0]	73	657	10.00	9.00	5.09
(218.0, 227.0]	79	625	11.22	7.91	4.91
(227.0, 237.0]	66	706	8.55	10.70	5.38
(237.0, 255.0]	48	676	6.63	14.08	5.05
(255.0, 266.0]	49	683	6.69	13.94	5.10
(266.0, 275.0]	35	605	5.47	17.29	4.46
(275.0, 304.0]	42	769	5.18	18.31	5.65
(304.0, 318.0]	24	611	3.78	25.46	4.43
(318.0, 335.0]	30	797	3.63	26.57	5.76
(335.0, 340.0]	32	623	4.89	19.47	4.56
(340.0, 354.0]	17	653	2.54	38.41	4.67
(354.0, 362.0]	26	927	2.73	35.65	6.64
(362.0, 371.0]	9	550	1.61	61.11	3.90
(371.0, 385.0]	18	755	2.33	41.94	5.39
(385.0, 388.0]	14	784	1.75	56.00	5.56
(388.0, 396.0]	25	1096	2.23	43.84	7.81
(396.0, 399.0]	0	101	0.00	inf	0.70

FIGURE 2. Score Distribution Report

The application model for this project produces a fair credit score distribution among the client data, as there are no clusters in a single interval present.
To validate the model, the Kolmogorov – Smirnov statistical test is performed.
The KS test compares the cumulative score distribution of the model in different data sets. A valid KS test implies that the model is expected to perform well on new data sets. If there is a significant difference in the distribution of multiple data sets, the test will fail.
To execute the test, I split the original data frame again into a development and validation set. The proportion is: development : validation = 8 : 2, meaning that 80% of the original data frame is included in the development set and 20% is used for the validation set. Afterwards the cumulative amount of good, bad and total clients is calculated for each set. There are a total of nine cases to be validated:

Distribution of Bad clients within:
development set and validation set
development set and original set
validation set and original set
Distribution of Good clients within:
development set and validation set
development set and original set
validation set and original set
Distribution of all clients within:
development set and validation set
development set and original set
validation set and original set

After performing the validation test on all sample sets using the generated model, all distribution comparisons were valid.

The second metric, that is used to validate the model performance, is the Gini Coefficient. The Gini coefficient shows how well does the model discriminate good and bad clients. A perfect model would produce a Gini coefficient equal to 1, meaning that it creates a flawless separation of good and bad clients, without false positive and false negative outcomes. Since a perfect model is not realistic, a good model should aim for a Gini coefficient of 0.5 to 0.6.

For this model, I have used the cumulative distribution of good and bad clients in the original data set. The Gini coefficient is calculated by using Brown's formula:

$$\text{Gini coefficient} = 1 - \Sigma[G_{(i)} + G_{(i-1)}]*[B_{(i)} - B_{(i-1)}]$$

whereas:

G(i) is the cumulative proportion of good clients in the i-th interval

B(i) is the cumulative proportion of bad clients in the i-th interval

The sum is from $i = 2$ to n , where n is the number of intervals

The generated model has a Gini coefficient of 0.4282

3. Application Score Card

The result of the generated linear model is an Application Score Card. A graphical representation of the Score Card is given in Fig. 3.

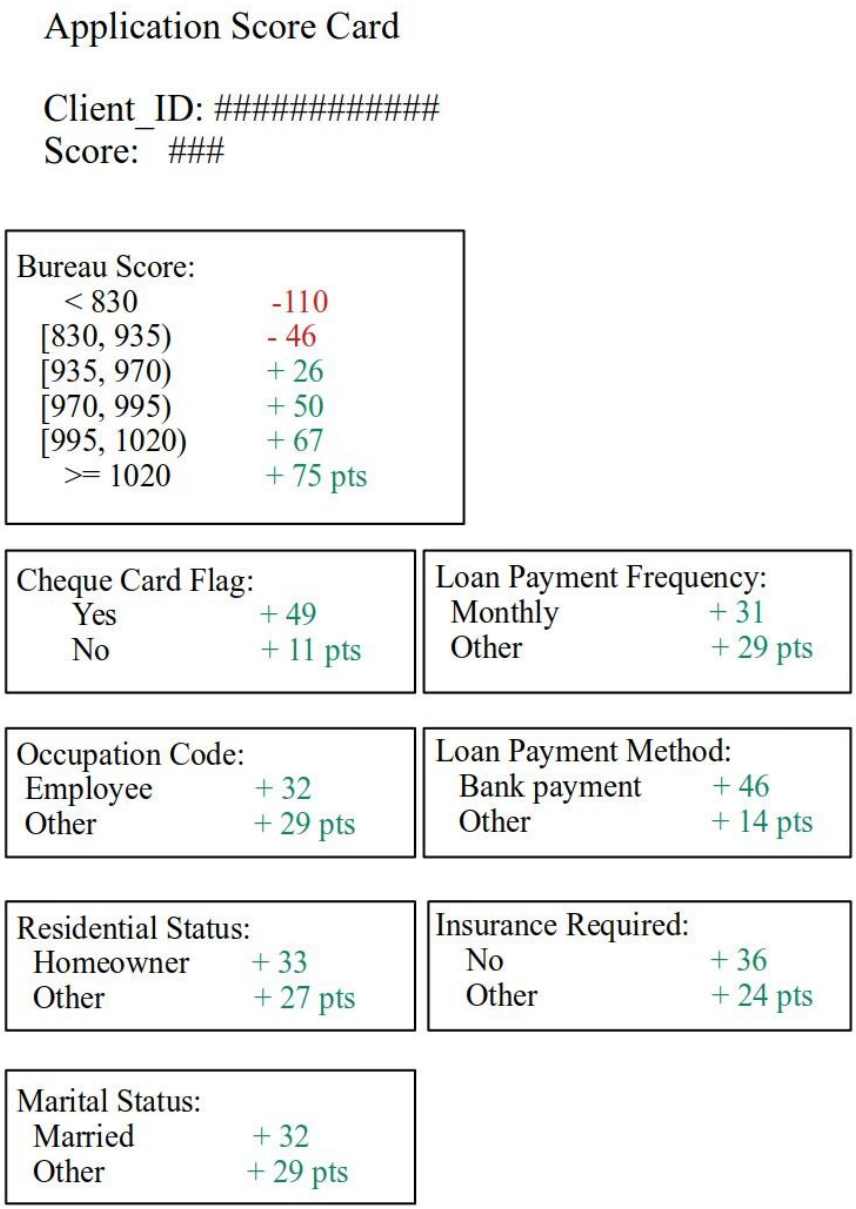

FIGURE 3. Application Score Card

3.1 Classing of predictive variables and Score points

The predictive variables and their class distribution, as well as scoring, are described in the next paragraph.

- **Bureau Score**

  Classing and score points, assigned to the client in accordance to the class:

< 830	- 110
[830, 935)	- 46
[935, 970)	+ 26
[970, 995)	+ 50
[995, 1020)	+ 67
>= 1020	+ 75 pts

- **Cheque Card Flag**

  Classing and score points, assigned to the client in accordance to the class:

Yes	+ 49
No / NA	+ 11 pts

  Possible Cheque card flag status is:

  Yes, No or not available

- **Occupation Code**

  Classing and score points, assigned to the client in accordance to the class:

Employee	+ 32
Other	+ 29 pts

  Possible Occupation status is:

  Employee, Self-employed, Pensioner, Other

- **Residential Status**

  Classing and score points, assigned to the client in accordance to the class:

Homeowner	+ 33
Other	+ 27 pts

  Possible Residential status is:

  Homeowner, Tenant, Living with parents, Other

- **Marital Status**

  Classing and score points, assigned to the client in accordance to the class:

Married	+ 32
Other	+ 29 pts

  Possible Marital status is:

  Married, Divorced, Single, Widow, not given

- **Loan Payment Frequency**

  Classing and score points, assigned to the client in accordance to the class:

Monthly	+ 31
Other	+ 29 pts

  Possible Loan payment frequency status is:

  Monthly, Weekly, Two times per month, not given or not available

- **Loan Payment Method**

  Classing and score points, assigned to the client in accordance to the class:

Bank Payment	+ 46
Other	+ 14 pts

  Possible Loan payment method status is:

  Bank payment, Cheque, Standing order, not given or not available

- **Insurance Required**

  Classing and score points, assigned to the client in accordance to the class:

No	+ 36
Other	+ 24 pts

  Possible Insurance required status is:

  No, Yes or not available

Each client has an intercept of 61 score points. The total credit score is calculated by using the following formula:

$$Credit\ Score = intercept + \Sigma\ scorecard\ points$$

The descriptive statistics of the client score for the original data frame of c.a 17 000 records are:

Minimal score: 114 pts

Maximal score: 395 pts

Mean value: 273 pts

Standard deviation: 82 pts

Quartiles:

25 % : 213 pts

50 % : 271 pts

75 % : 352 pts

3.2 Automated Decision Threshold

Taking into account the descriptive statistics of the overall client scoring, as well as the score distribution report, a proposed threshold for an automated decision making is:

Clients with an Application score below 227 points could be rejected.

Clients with an Application score in the interval: (227, 340] could be revised from an employee.

Clients with an Application score above 340 points could be approved.

Argumentation:

According to the score distribution report (see Fig. 4) the probability of a client with a credit score below 227 points being marked as a Good client is less than 10 %. Furthermore, the cumulative distribution table of bad clients (see Fig. 5) shows, that 52 % of the Bad clients have a credit score lesser or equal to 227 points. For this reason, the proposed threshold marks clients with a credit score less than 227 points as Rejected.

Similary, the score distribution report indicates, that clients with credit score above 340 points have a high probability – above 35 % - of being marked as Good clients. According to the cumulative distribution table of good clients (see Fig. 6) 36 % of all good clients have a credit score greater than 340 points.

All figures are shown in the following pages.

Figures:

Target Score	Bad	Good	Bad_Rate	Good_Bad_Odds	Total_%
(116.0, 154.0]	133	569	18.95	4.28	4.89
(154.0, 193.0]	127	721	14.98	5.68	5.91
(193.0, 210.0]	78	517	13.11	6.63	4.15
(210.0, 218.0]	73	657	10.00	9.00	5.09
(218.0, 227.0]	79	625	11.22	7.91	4.91
(227.0, 237.0]	66	706	8.55	10.70	5.38
(237.0, 255.0]	48	676	6.63	14.08	5.05
(255.0, 266.0]	49	683	6.69	13.94	5.10
(266.0, 275.0]	35	605	5.47	17.29	4.46
(275.0, 304.0]	42	769	5.18	18.31	5.65
(304.0, 318.0]	24	611	3.78	25.46	4.43
(318.0, 335.0]	30	797	3.63	26.57	5.76
(335.0, 340.0]	32	623	4.89	19.47	4.56
(340.0, 354.0]	17	653	2.54	38.41	4.67
(354.0, 362.0]	26	927	2.73	35.65	6.64
(362.0, 371.0]	9	550	1.61	61.11	3.90
(371.0, 385.0]	18	755	2.33	41.94	5.39
(385.0, 388.0]	14	784	1.75	56.00	5.56
(388.0, 396.0]	25	1096	2.23	43.84	7.81
(396.0, 399.0]	0	101	0.00	inf	0.70

FIGURE 4. Score Distribution Report

Cumulative distribution tables:

```
Score
(116.0, 154.0]     14.93
(154.0, 193.0]     28.00
(193.0, 210.0]     36.27
(210.0, 218.0]     44.44
(218.0, 227.0]     52.62
(227.0, 237.0]     59.64
(237.0, 255.0]     64.98
(255.0, 266.0]     70.22
(266.0, 275.0]     73.87
(275.0, 304.0]     78.49
(304.0, 318.0]     81.07
(318.0, 335.0]     84.71
(335.0, 340.0]     88.00
(340.0, 354.0]     89.87
(354.0, 362.0]     92.80
(362.0, 371.0]     93.69
(371.0, 385.0]     95.82
(385.0, 388.0]     97.51
(388.0, 396.0]    100.00
(396.0, 399.0]    100.00
Name: Cum_%_Bad_All, dtype: float64
```

FIGURE 5. Cumulative distribution table of bad clients

```
Score
(116.0, 154.0]      4.25
(154.0, 193.0]      9.75
(193.0, 210.0]     13.62
(210.0, 218.0]     18.48
(218.0, 227.0]     23.25
(227.0, 237.0]     28.52
(237.0, 255.0]     33.55
(255.0, 266.0]     38.89
(266.0, 275.0]     43.33
(275.0, 304.0]     49.13
(304.0, 318.0]     53.59
(318.0, 335.0]     59.44
(335.0, 340.0]     64.06
(340.0, 354.0]     68.99
(354.0, 362.0]     75.65
(362.0, 371.0]     79.65
(371.0, 385.0]     85.24
(385.0, 388.0]     91.13
(388.0, 396.0]     99.26
(396.0, 399.0]    100.00
Name: Cum_%_Good_All, dtype: float64
```

FIGURE 6. Cumulative distribution table of good clients