MMF2000 RETAIL CREDIT RISK MODELING ASSIGNMENT REPORT

YILING QI, TIANYI LONG, JUNHAO WANG, TIANRAN LI, CHANGJUN LIU

1. Introduction

Small businesses play an important role in economy and they have been receiving greater recognition as an essential driver of the economic growth. Retail credit risk is the risk of loss due to a consumer's failure or inability to repay on a credit product. A healthy economy is where the small businesses are healthy, prosperous, and sustainable.

1.1. **Purpose.**

The Small Business behavior score is an internal, proprietary credit report that supplement credit data on small businesses, covering deeper analysis on small business payment history, credit utilization over time and the amounts and types of products purchased, which helps to demonstrate the payment pattern of small businesses. Our model is developed with the purpose of using scorecard with increasing transparency to monitor the financial health of small businesses in order to identify changing risk to take corrective actions and improve portfolio profitability.

1.2. **Portfolio.**

According to the attached dataset, the Small Business portfolio consists of Widely held customer, Startup business, Term loan customer, Demand loan customer, OLL customer, Visa customer, and etc.

1.3. **Model Use.**

The Small Business model could be used within a Financial Institution to measure the credit score of large different variety of customers, which could indicate the predicted probability that the customer will exhibit a certain behavior. In this case, underwriters in financial institutions are able to review each specific case manually and integrate all data sources, and finally make financial decisions based on this credit score.

1.4. **Economic and Market Outlook.**

The Bank of Canada has repeatedly committed to keeping the policy interest rate very low for a long time, and until October 28th, 2020, it has kept interest rates at a steady 0.25%, due to the economic slump triggered by the COVID-19 pandemic. Unless we can see a sharp recovery in the economy and a huge spike in inflation, the rate is likely to remain at the effective lower bound. As the coronavirus disease (COVID-19) spreads globally, fear and uncertainty are rising, roiling financial markets and pushing the global economy towards recession, which impacts the revenues and earnings of many companies, influencing their ability to make interest payments on debt and, ultimately, repay the debt itself. Therefore, leading to increasing the default risk.

1.5. Model Development Process.

The model development process includes the data pre-processing, data cleaning, variable reduction using Weighted of Evidence (WOE), Information Value (IV), Variable cluster, feature engineering and model fitting using logistic regression model. After that, we created scorecard scaling and assessment using rank ordering and checked the population stability. We also created a benchmark model using Decision Tree to compare with the first model.

2. Data

2.1. Data Sources.

We have our data sources from both external data, for example, the Credit Bureau Information, and other internal data.

2.2. Time Frames.

According to our data dictionary, the observation period lasts 24 months that historically leading up to the observation point. Also, by the definition of our target variable, there is a 12-month performance period after observation point.

2.3. Target Variable Definition.

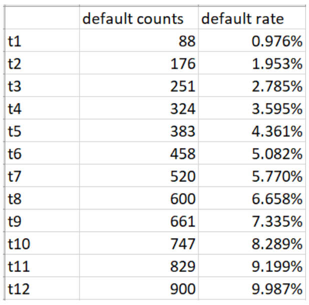
In our dataset, we have sets of variable ti represents Default Flag i month after observation point, where $i = 1, ..., 12$. We notice that whenever $ti = 1$, all following $t(i + 1)$ to $t12$ are 1, which intuitively makes sense since each ti indicates whether default or not within i month after observation point, and each $t(i+1)$ contains the case of ti, which means $t12 = 1$ if it defaults any month within this 12 months. Thus, we directly defined t12 as our target variable.

2.4. Population Exclusions.

The original number of the customer is 9028, and we have to exclude 16 Widely Held and deceased customers in total, due to the fact that they are no longer active and are not able to pay any more. Thus, keeping them in our dataset would not provide useful information. After removing these customers, we now have 9012 customers in our new dataset so that we still have enough data to analyze.

2.5. Modeling Population.

For total 9012 customers, there are 900 customers are in default by summing up along the column $t12$. In addition, the default rate is the total number of customers in default divided by total population, which is 9.987% in this case. The default counts and default rates for on all 12 observation points are as follows:



	default counts	default rate
t1	88	0.976%
t2	176	1.953%
t3	251	2.785%
t4	324	3.595%
t5	383	4.361%
t6	458	5.082%
t7	520	5.770%
t8	600	6.658%
t9	661	7.335%
t10	747	8.289%
t11	829	9.199%
t12	900	9.987%

2.6. **Explanatory Variables.**
We create two new explanatory variables by using Monthly Debit and Credit Transactions, they are credit_to_debit_curr and credit_to_debit_prev$_i$ respectively. And the definition of these two new explanatory variables shown as below:

$$\text{credit\_to\_debit\_curr} = \frac{\text{credit\_curr}}{\text{debit\_curr}}$$
$$\text{credit\_to\_debit\_prev\_i} = \frac{\text{credit\_prev\_i}}{\text{debit\_prev\_i}}$$

Therefore, we choose the ratio of debit amount and credit amount from previous 12 months to current month, and create 13 more explanatory variables in total. This ratio is a good way for bank to access client's credit worthiness, and it generally represents the amount of debit a client have to the amount a client owes to bank. we prefer to see a higher debit to credit ratio, which indicates lower risk of default.

2.7. **Segmentation.**
According to the number of unique value in each column, we decide to segment our population into 2 groups by using variable $max\_ks\_max\_dlq\_24mos$, which represents the maximum revolving delinquency of a customer in the last 24 months. And there are only two values for this variable, 7 or 0 respectively, which clearly separates our population into 2 segmentation.

2.8. **Sampling Methodology.**
In order to measure the model performance, we need to use historical sample data to train and validate our model, then using out-of-time sample to test our model. We observed from the variable $TIME\_KEY$ that all the observations can be categorized into four time points: $2014/1/1, 2014/04/01$, 2014/07/01 and 2014/10/01 in the chronological order. Therefore we created two samples with the first one including the data on and before 2014/07/01 and the second one on 2014/10/01 as the out-of-time validation sample. We always keep data not used for training nor validation to measure model performance, and ensure sampling are sufficient random and representative to keep unbiased.

3. Scorecard Development

3.1. **Modeling Considerations.**
Since the model we used in this project is the one that most frequently used and applied by the industry. Therefore, the modeling technique, such as using WOE and IV to do variable reduction, as well as using step-wise logistic regression to fit our prediction model on this project would be appropriate.

3.2. **Variable Reduction.**

3.2.1. *Pre-Screening.*
There are total 518 explanatory variables initially in our dataset. We separated all explanatory variables into numerical and categorical variables, then excluding numerical variables that has F-statistics smaller than 6 using F-test, as well as excluding categorical variables when p value greater than 0.05 using Chi-squared test. In addition, we removed columns with more than 80% missing value and those with 0 variance (i.e. constants), which cannot provide useful information

for us to make business decision. After that, we have 264 categorical features and 81 numerical features left. For Pre-Screening, we created correlation matrix and dropped those features may cause multicollinearity issues with correlation greater than 0.85 and smaller than 0.1. Therefore, we removed 178 variables, example excluded variables including but not limit to following: 'NFP_N', 'deceased', 'dda_sum_OL_Days', 'BSL_SB_CUST_Y', 'PROFTYPE_C', 'PDLCUST_N', 'HasAIRBProduct_Y', 'WIDELYHD_N', 'BSL_CUST_AIRB_Y', 'OPFBCUST_N'. Finally, we have 123 explanatory variables remaining.

3.2.2. *Univariate Screening.*
After Pre-screening, we filtered out that there are 41 variables that already have the WOE calculated from the dataset. So we firstly calculate the IV for those 41 variables, and then we calculated the WOE and IV for the remaining categorical variables and numerical variables (after binning) respectively. The calculation metrics for WOE and IV are showed as below. Since IV smaller 0.02 can be regarded as very weak predictive power and IV greater than 0.3 can be regarded as very strong predictive power, but if it is too strong, they may be in some way a deviation of a target variable, thus in this case, we removed 28 variables that has IV smaller than 0.02 or greater than 0.5, After that, there are 52 categorical variables and 9 numerical variables left, as well as 33 variables with WOE originally, total 94 variables remaining now. GRP_ALL8150 (IV=0.319694) is one of the significant variable, which indicates strong predictive power, therefore, as variables changes from one bin to another bin, it could show very different value of WOE and default rate that completely gives us very different information. On the contrary, 'GRP_TBSSC100' is one of the non-significant variables with very low IV (≤ 0.02), thus, when this variable changes, there would be only negligible change happens to WOE and default rate, since low IV shoes very weak predictive power to separate Non-default from Default accounts.

Weight of Evidence (**WOE**) is a measure of how much the evidence supports hypothesis. It tells the predictive power of an independent variable in relation to the dependent variable. Since it evolved from credit scoring world, it is generally described as a measure of the separation of good and bad customers. WOE_i for an attribute/group i is based on the probability density functions (f_G and f_B) for the two binary events and is defined as following:

$$f_G(i) = \frac{N_i^G}{N_{total}^G}$$
$$f_B(i) = \frac{N_i^B}{N_{total}^B}$$
$$WOE_i = ln(\frac{f_G(i)}{f_B(i)})$$

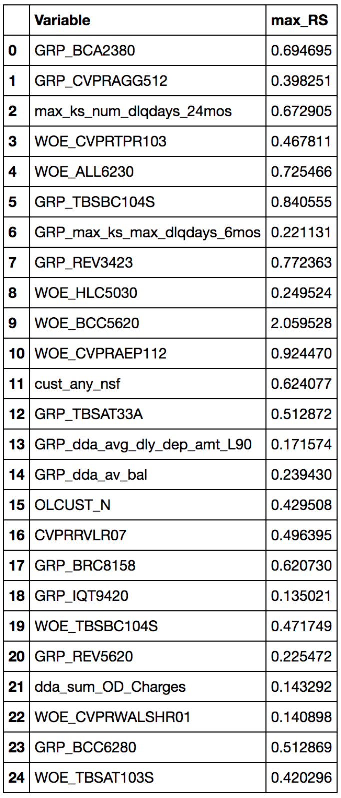
where

$$N^G = \text{the number of Non-default accounts}$$
$$N^B = \text{the number of Default accounts}$$

Information Value (**IV**) measures the predictive power of a single variable specifically, the ability to separate good accounts from bad accounts. It is one of the most useful technique to select important variables in a predictive model and it helps to rank variables on the basis of their importance. Higher IV equals higher predictive power. IV is the weighted sum of the WOE of the variable's attributes. Shown as following;

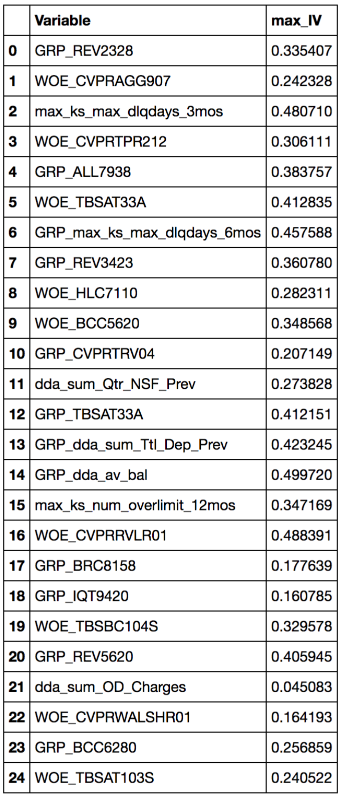
$$IV = \sum_{i=1}^{n} [f_G(i) - f_B(i)] \cdot WOE_i$$

3.2.3. *Multivariate Screening.*
We created total 25 clusters using VarClusHi, which is a Python module to perform variable clustering (varclus) with a hierarchical structure to perform a nice dimension reduction algorithm. And the algorithm is based on R^2 ratio $= \frac{1-RS_{own\ cluster}}{1-RS_{next\ closest\ cluster}}$ and then select one variable from each cluster which is having lowest $1-R^2$ ratio could best represent that clusters with little loss of information, also select one variable from each cluster which is having highest Information Value. Therefore, we finally selected 2 variables in each cluster. For example, cluster 0 includes variables such as 'GRP_REV2327', 'GRP_REV2328', 'GRP_REV2350', which involves the information of total number of revolving trades ever 30 or more days delinquent or derogatory in past months and cluster 1 includes 'WOE_CVPRAGG907', 'WOE_CVPRAGG519', which involves the information of the WOE of maximum aggregate bankcard balances over past months. Thus, after combining above metrics together, we have 37 variables (after dropping the duplicates, i.e a variable is the one with lowest $1 - R^2$ ratio and also the one with highest IV) remaining for us proceed to the next step.



	Variable	max_RS
0	GRP_BCA2380	0.694695
1	GRP_CVPRAGG512	0.398251
2	max_ks_num_dlqdays_24mos	0.672905
3	WOE_CVPRTPR103	0.467811
4	WOE_ALL6230	0.725466
5	GRP_TBSBC104S	0.840555
6	GRP_max_ks_max_dlqdays_6mos	0.221131
7	GRP_REV3423	0.772363
8	WOE_HLC5030	0.249524
9	WOE_BCC5620	2.059528
10	WOE_CVPRAEP112	0.924470
11	cust_any_nsf	0.624077
12	GRP_TBSAT33A	0.512872
13	GRP_dda_avg_dly_dep_amt_L90	0.171574
14	GRP_dda_av_bal	0.239430
15	OLCUST_N	0.429508
16	CVPRRVLR07	0.496395
17	GRP_BRC8158	0.620730
18	GRP_IQT9420	0.135021
19	WOE_TBSBC104S	0.471749
20	GRP_REV5620	0.225472
21	dda_sum_OD_Charges	0.143292
22	WOE_CVPRWALSHR01	0.140898
23	GRP_BCC6280	0.512869
24	WOE_TBSAT103S	0.420296

(a) Variables with lowest $1 - R^2$ ratio in each cluster



	Variable	max_IV
0	GRP_REV2328	0.335407
1	WOE_CVPRAGG907	0.242328
2	max_ks_max_dlqdays_3mos	0.480710
3	WOE_CVPRTPR212	0.306111
4	GRP_ALL7938	0.383757
5	WOE_TBSAT33A	0.412835
6	GRP_max_ks_max_dlqdays_6mos	0.457588
7	GRP_REV3423	0.360780
8	WOE_HLC7110	0.282311
9	WOE_BCC5620	0.348568
10	GRP_CVPRTRV04	0.207149
11	dda_sum_Qtr_NSF_Prev	0.273828
12	GRP_TBSAT33A	0.412151
13	GRP_dda_sum_Ttl_Dep_Prev	0.423245
14	GRP_dda_av_bal	0.499720
15	max_ks_num_overlimit_12mos	0.347169
16	WOE_CVPRRVLR01	0.488391
17	GRP_BRC8158	0.177639
18	GRP_IQT9420	0.160785
19	WOE_TBSBC104S	0.329578
20	GRP_REV5620	0.405945
21	dda_sum_OD_Charges	0.045083
22	WOE_CVPRWALSHR01	0.164193
23	GRP_BCC6280	0.256859
24	WOE_TBSAT103S	0.240522

(b) Variables with highest IV in each cluster

3.3. **Model Fitting.** We fitted the logistic regression model by using variables from previous step, and then selected variables using the stepwise (forward and backward selection) algorithm, which indicates that we have evidence to show those remaining variables are significant in our model. Logit transformation is the log of the odds used to linearize probability of the event occurring and limits the outcome between 0 and 1, which definitely matches our expectation of probability of default, and estimates reflect how likely (the odds) it is that observed outcome can be predicted based the knowing inputs of the explanatory variables.

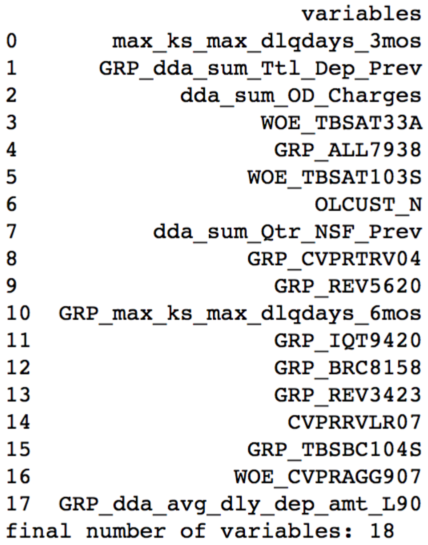
Logistic Regression Model:

$$Logit(p) = \beta_0 + \sum_{i=0}^{M} \beta_i x_i$$

where

$$\begin{aligned}
&p = \text{probability of default}\\
&\beta_0 = \text{intercept}\\
&\beta_i = \text{estimate of parameter } x_i\text{'s coefficient}\\
&x_i = \text{explanatory variables i}\\
&M = \text{number of explanatory variables}
\end{aligned}$$

The following are variables made into the final model



```
                        variables
0        max_ks_max_dlqdays_3mos
1       GRP_dda_sum_Ttl_Dep_Prev
2             dda_sum_OD_Charges
3                   WOE_TBSAT33A
4                    GRP_ALL7938
5                  WOE_TBSAT103S
6                       OLCUST_N
7            dda_sum_Qtr_NSF_Prev
8                  GRP_CVPRTRV04
9                    GRP_REV5620
10  GRP_max_ks_max_dlqdays_6mos
11                   GRP_IQT9420
12                   GRP_BRC8158
13                   GRP_REV3423
14                    CVPRRVLR07
15                 GRP_TBSBC104S
16                WOE_CVPRAGG907
17  GRP_dda_avg_dly_dep_amt_L90
final number of variables: 18
```

3.4. **Scorecard Scaling.**

We use most common scaling factor for credit scoring models, where Odds of 40:1 (non-defaulted: defaulted) at 740 points, with odds doubling every 20 points. The score scaling table is shown as following:

Score	Odds
680	5
700	10
720	20
740	40
760	80
780	160

And according to this table and the coefficients we get from model fitting step, we are able to derive that:

$$Factor = \frac{20}{ln(2)} = 28.8539$$
$$Offset = 740 - (28.8539 \times ln(40)) = 633.56$$
$$\alpha = -0.281056$$
$$n = 18$$

In general,

$$Score = \sum_{j,i=1}^{k,n} (-(WOE_{j,i} \times \beta_i + \frac{\alpha}{n}) \times Factor + \frac{Offset}{n})$$

where

β_i =regression coefficient for the explanatory variable i

WOE_j = weight of evidence for the group i for explanatory variable j

n = number of explanatory variables in the model

α =intercept

k =number of groups/bins in each explanatory variable

3.5. **Scorecard Assessment.**

3.5.1. *Rank-Ordering.*

Kolmogorov − Smirnov (KS) Statistic measures the maximum difference between two cumulative distributions (distribution of Not-Defaulted and distribution of Defaulted) ,

$$KS = sup|F_D(s) - F_{ND}(s)|$$

where

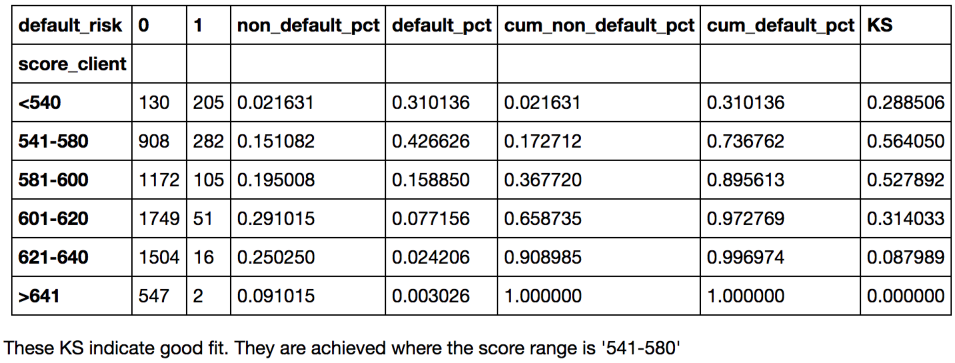
$F_D(s)$ =Cumulative distribution of Defaults by score

$F_{ND}(s)$ =Cumulative distribution of Non-Defaults by score

Result shown as below:

Sample	KS
1	50.83684980374411%
2	47.176801226366884%

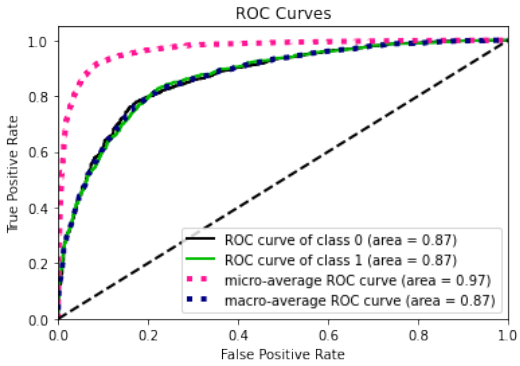
Also, we can see



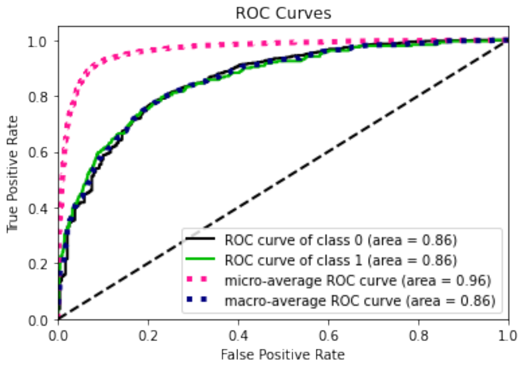
default_risk	0	1	non_default_pct	default_pct	cum_non_default_pct	cum_default_pct	KS
score_client							
<540	130	205	0.021631	0.310136	0.021631	0.310136	0.288506
541-580	908	282	0.151082	0.426626	0.172712	0.736762	0.564050
581-600	1172	105	0.195008	0.158850	0.367720	0.895613	0.527892
601-620	1749	51	0.291015	0.077156	0.658735	0.972769	0.314033
621-640	1504	16	0.250250	0.024206	0.908985	0.996974	0.087989
>641	547	2	0.091015	0.003026	1.000000	1.000000	0.000000

These KS indicate good fit. They are achieved where the score range is '541-580'

Receiver Operating Characteristic (ROC) Curve can be summarized by Area under the ROC Cruve (AUC). AUC represents an estimate of the probability that a randomly chosen instance of non defaulted account is correcly ranked higher than a randomly chosen instance of defaulted account. For random model AUC=50%. AUC is often used as a measure of quality of the classification models. Results shown as below:



(c) ROC Sample 1



(d) ROC Sample 2

Accuracy Ratio (AR) is the area between the Lorenz and Random Curve is the Gini Index.

$$AR = \frac{GINI}{perfect\ GINI}$$

where

AR for Perfect Model = 1

AR for Random Model = 0

AR = 2 × AUC - 1

Result shown as below:

Sample	AUC	AR
1	88.0297336%	76.05946720166337%
2	86.41391542%	72.82783083654141%

Both AUC and AR gives us desirable results, which indicates good model accurate and performance.

Lift Chart The lift value is the cumulative percentage of defaults per decile, divided by the overall population percentage of defaults. For Random Model, Lift = 1.0

Result shown as below:

Sample	Lift at 10%
1	5.099100222959626
2	4.730053999928477

3.5.2. *Population Stability.*
Assess whether the model development dataset is similar to population distributions through time.

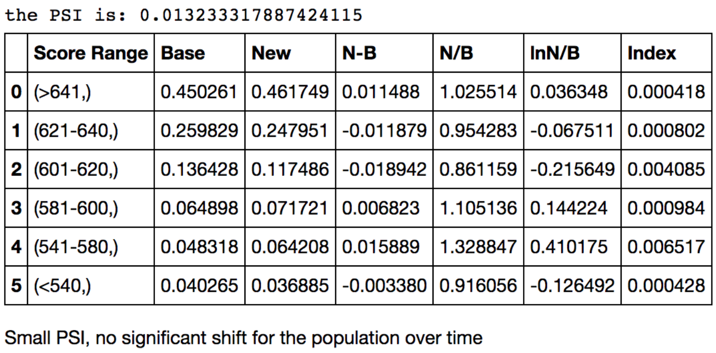
Popularity Stability Index (**PSI**) is a measure of how much the population has changed over a period of time, it quantifies population differences by measuring the shift between two sample distributions, and it also act as Industry standard indicator. Our PSI is 0.013233317887424115, which is smaller than 0.1, indicates there is no significant shift.

$$PSI = \sum_{i=1}^{k}(N_i - B_i) \cdot ln(\frac{N_i}{B_i})$$

where

$$N_i = \%\text{ of accounts in score band } i \text{ for the new population}$$
$$B_i = \%\text{ of accounts in score band } i \text{ for the base population}$$

Result shown as below:



```
the PSI is: 0.013233317887424115
```

	Score Range	Base	New	N-B	N/B	lnN/B	Index
0	(>641,)	0.450261	0.461749	0.011488	1.025514	0.036348	0.000418
1	(621-640,)	0.259829	0.247951	-0.011879	0.954283	-0.067511	0.000802
2	(601-620,)	0.136428	0.117486	-0.018942	0.861159	-0.215649	0.004085
3	(581-600,)	0.064898	0.071721	0.006823	1.105136	0.144224	0.000984
4	(541-580,)	0.048318	0.064208	0.015889	1.328847	0.410175	0.006517
5	(<540,)	0.040265	0.036885	-0.003380	0.916056	-0.126492	0.000428

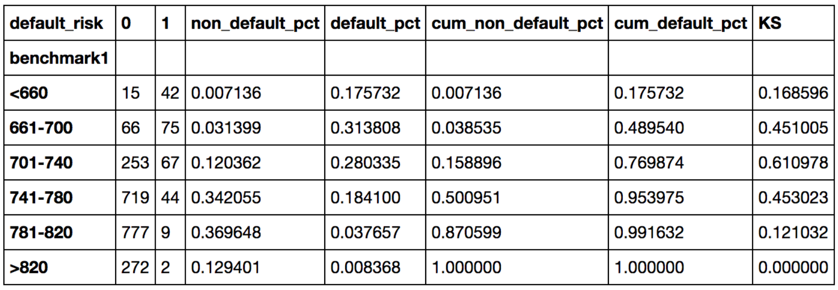
Small PSI, no significant shift for the population over time

3.5.3. *Scorecard Benchmarking.*

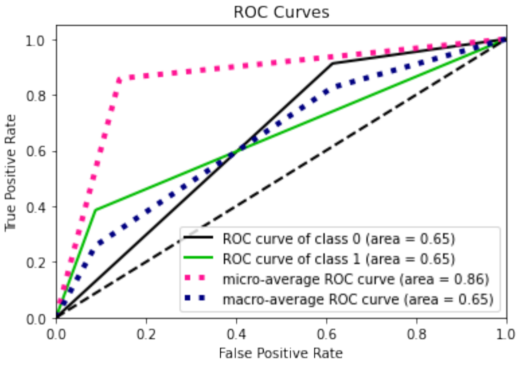
Compare scorecard to alternative scorecards, in this part, we generated the Decision Tree model as our Benchmark model to compare with the performance using AR, KS and lift at 10%. Result shown as below:

Sample	KS	AR	Lift at 10%
1	63.120240849215%	32.96422215940984%	5.356324863286965
2	61.09781877391127%	28.62757733580694%	5.483513929120624

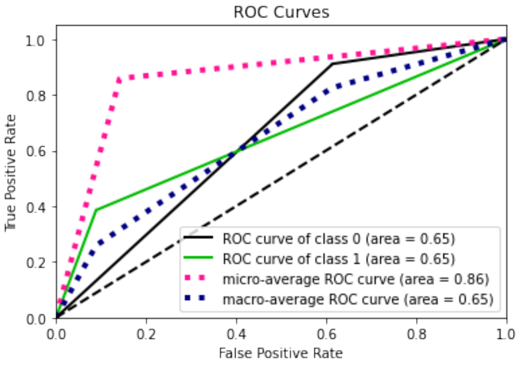
Also, we have the biggiest KS corresponds to the score range of '701-740',



default_risk	0	1	non_default_pct	default_pct	cum_non_default_pct	cum_default_pct	KS
benchmark1							
<660	15	42	0.007136	0.175732	0.007136	0.175732	0.168596
661-700	66	75	0.031399	0.313808	0.038535	0.489540	0.451005
701-740	253	67	0.120362	0.280335	0.158896	0.769874	0.610978
741-780	719	44	0.342055	0.184100	0.500951	0.953975	0.453023
781-820	777	9	0.369648	0.037657	0.870599	0.991632	0.121032
>820	272	2	0.129401	0.008368	1.000000	1.000000	0.000000



(e) ROC Benchmark Sample 1



(f) ROC Benchmark Sample 2

4. Model Limitations and Assumptions

4.1. **Model Limitation.**

Since we fitted the logistic regression in this project, and also applied some Machine Learning techniques to select features that we thought they are important and could best explain our target variable in order to improve the prediction accuracy. However, the criterion or the technique we applied to do feature selection and variable reduction may not be trivial, and different methods could present with different results. Also, we have to concern about issues such as separate steps for variable reduction and model fitting as well as distribution assumptions. In addition, there is also the issue we need to take into account is the trade-off between model bias and variance, since we always want to find the balance between them, but in reality, it could be hard, therefore, the over-fitting problem is also one of the limitation in our model.

4.2. **Model Assumption.**

1. Our target variable is binary, with only two state: Default or Not default.
2. Variables with smaller or equal to 15 unique values can be regard as numerical variable, and otherwise, is categorical variable.
3. Customer behavior, products, policies and procedures will not change through the whole process.
4. Historical experience is predictive of future behaviour.