Default Risk Research for Unbanked Population

Liu Cuiyi, Qiu Yang, Xu Shifeng Singapore Management University

ABSTRACT

The technical report explores significant factors that affect default risk on home credit, mainly containing three distinct parts: Exploratory Data Analysis (EDA), Cluster Analysis and Predictive Modelling.

In EDA, an overview about current status and several measures which can differentiate the default group are provided. Cluster analysis filters the group with 100 percent default rate and concludes its main features. Most importantly, we built several predictive models using Decision Tree method as well as Logistic Regression and selected the optimal one with the lowest Average Squared Error (ASE) at 0.07 and high accuracy rate at 91%. Besides, the most significant variables are selected and ranked from models to evaluate one's default risk.

From the perspective of management, we offered several recommendations on specific groups and useful factors.

INTRODUCTION

In the present years, credit loan becomes one of the most important fundraising approaches for individuals or companies. With the tendency, the term of credit risk attracts more attention to the financial industry. How to make a reliable prediction of clients' repayment abilities turns into a significantly valuable research project.

Embarking from the motivation, we collected data of loan customers' personal information, credit records and whether they repaid their loan (if a loan customer did not complete his repayment, we call it as credit risk) from Home Credit, an international consumer finance provider, aiming at lending money to underserved customers with little or no credit history. In total, 24825(8%) of 307511 customers of Home Credit defaulted their loan, causing loss up to $657,409,294.80. And our aim is to differentiate those high default risk population by primary factors and predict with considerable accuracy.

RESEARCH AND OBJECTIVES

This paper target at unbanked population and help Home Credit to figure out the principal factors affecting repayment. To a further step, providing a regression model predicting repayment as well as specific solutions is also included in our goal.

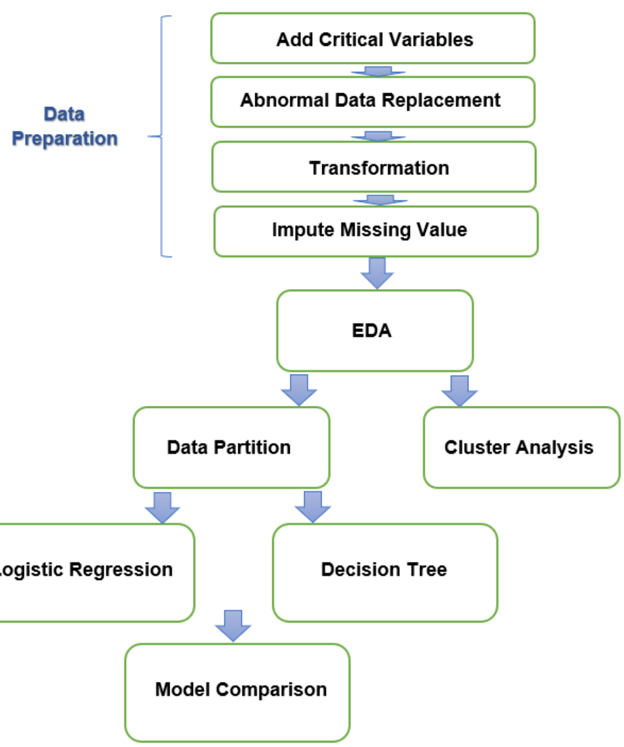


Figure 1 Process Flow of the Project

DATA PREPARTATION

Data Cleaning

After exploring the relationship between default risk and other variables, we select useful predictors and preprocess the data using JMP software. To reduce pairwise correlation, we combine some variables by computing ratio or statistics, like dividing credit by income to get the percentage. By checking variables' distributions, we recognize unusual values as missing values and delete outliers. In total, there are 30 predictor variables in use to build the predictive model. The following content is steps in detail.

File: application_train, POS_CASH_balance, credit_card_balance, previous_application, installments_payments
Add supplementary variables to reduce pairwise correlation, like combine some variables by computing ratio or statistics:

- Credit Income Percent = Amount of credit / Amount of income
- Annuity Income Percent = Amount of annuity / Amount of income
- Credit Term = Amount of credit / Amount of annuity
- Days Employed Birth Ratio = Days Employed / Days Birth
- AMT Payment / Regularity = $If\begin{pmatrix} \frac{\text{AMT\_PAYMENT\_TOTAL\_CURRENT}}{\text{AMT\_INST\_MIN\_REGULARITY}} \geq 1 \Rightarrow 1 \\ else \quad \Rightarrow \frac{\text{AMT\_PAYMENT\_TOTAL\_CURRENT}}{\text{AMT\_INST\_MIN\_REGULARITY}} \end{pmatrix}$
- AMT Credit / Application = AMT_CREDIT / AMT_APPLICATION
- Days Overdue = DAYS_ENTRY_PAYMENT - DAYS_INSTALMENT
- AMT Payment / Installment = AMT_PAYMENT / AMT_INSTALLMENT

For the "Days Employed" variable, replace unusual values with missing values. For the "AMT INCOME TOTAL" variable, recognize values three times the range between 10% quantile and 90% quantile as outliers, delete these outliers to make its distribution approximate to normal distribution.

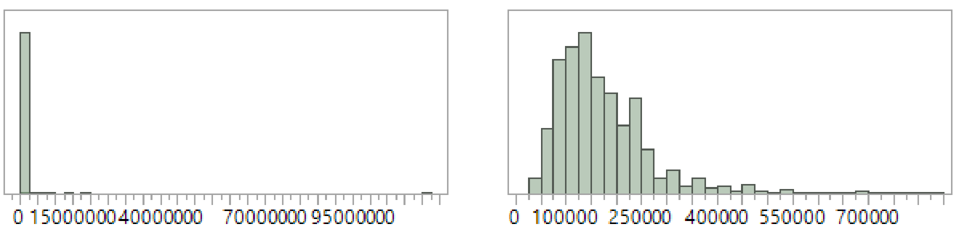


Figure 2 Distribution of AMT_INCOME_TOTAL

File: bureau
Add supplementary variable:

- Active Loan = $If\begin{pmatrix} \text{DAYS\_CREDIT\_ENDDATE} > 0 \Rightarrow 1 \\ else \quad \Rightarrow 0 \end{pmatrix}$

File: Model_Predict_Credit_Default_Risk
Combine all the useful predictors and target variable into one single file using "Table Summary" and "Table Update" tools in JMP.

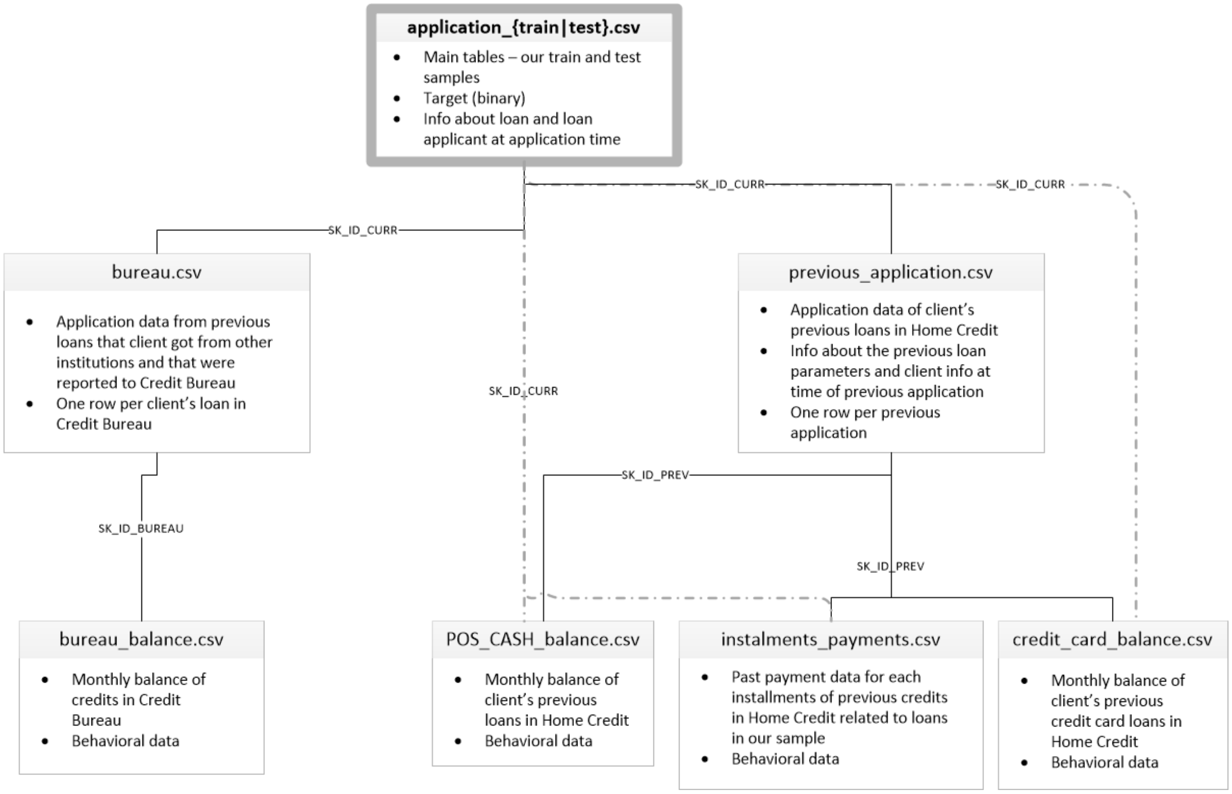


Figure 3 Table Structure

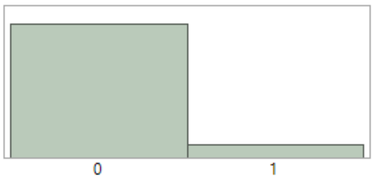
DATA DESCRIPTION

Variable Name	Data Type	Description	Data
SK_ID_CURR	ID	ID of loan in our sample	Original
TARGET	Binary	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)	Original
AMT_INCOME_TOTAL	Interval	Income of the client	Original
AMT_CREDIT	Interval	Credit amount of the loan	Original
AMT_ANNUITY	Interval	Loan annuity	Original
CREDIT_INCOME_PERCENT	Interval	= Amount of Credit / Amount of Income	Derived
ANNUITY_INCOME_PERCENT	Interval	= Amount of Annuity / Amount of Income	Derived
CREDIT_TERM	Interval	= Amount of Credit / Amount of Annuity	Derived
CODE_GENDER	Nominal	Gender of the client	Original
NAME_INCOME_TYPE	Nominal	Clients income type (businessman, working, maternity leave, etc)	Original
NAME_EDUCATION_TYPE	Nominal	Level of highest education the client achieved	Original
OCCUPATION_TYPE	Nominal	What kind of occupation does the client have	Original
YEARS_BIRTH	Interval	Client's age in years at the time of application	Original
DAYS_EMPLOYED	Interval	How many days before the application the	Original

		person started current employment	
DAYS_EMPLOYED_BIRTH_PERCENT	Interval	= Days Employed / Days Birth	Derived
DAYS_REGISTRATION	Interval	How many days before the application did client change his registration	Original
DAYS_ID_PUBLISH	Interval	How many days before the application did client change the identity document with which he applied for the loan	Original
Sum (Active_Loan)	Interval	Count of active loan need to pay at present	Derived
Max (AMT_CREDIT_SUM)	Interval	Current credit amount for the Credit Bureau credit	Original
Max (AMT_CREDIT_SUM_LIMIT)	Interval	Current credit limit of credit card reported in Credit Bureau	Original
Max (DAYS_CREDIT_ENDDATE)	Interval	Remaining duration of CB credit (in days) at the time of application in Home Credit	Original
Previous_Loans_Count	Interval	Count of loans received prior to the current loan	Derived
PL_Max(CNT_INSTALMENT_FUTURE)	Interval	Installments left to pay on the previous cred-it	Original
PL_Max(Days Overdue)	Interval	Maximum days overdue of previous loans	Derived
CC_Mean(AMT_BALANCE)	Interval	Balance during the month of previous credit	Original
CC_Mean(AMT_CREDIT_LIMIT)	Interval	Credit card limit during the month of the previous credit	Original
CC_Mean(AMT_INST_MIN)	Interval	Minimal installment for this month of the previous credit	Original
Previous_Application_Count	Interval	Count of previous loan applications	Derived
PA_Mean(AMT_ANNUITY)	Interval	Annuity of previous application	Original
PA_Mean(RATE_DOWN_PAYMENT)	Interval	Down payment rate normalized on previous credit	Original
PA_Mean(Percent_Credit_Application)	Interval	Average ratio of the approved amount of credit to the applied amount of credit of previous loan applications	Derived
INST_Max(Days_Overdue)	Interval	Maximum days overdue of previous install-ments	Derived
INST_Min(Percentage_of_Paid)	Interval	Minimum ratio of the amount of payment to installment regularity	Derived

EXPLORATARY DATA ANALYSIS

Distribution of Target: People who have payment difficulties occupies 8% of the total amount of population with higher income compared to people who don't have payment difficulties.



Frequencies

Level	Count	Prob
0	282231	0.91923
1	24800	0.08077
Total	307031	1.00000

N Missing 0
2 Levels

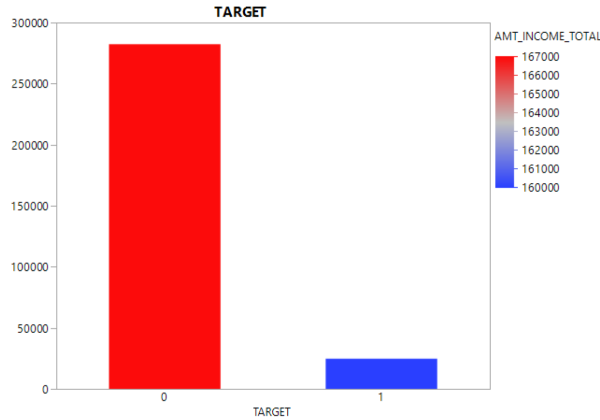


Figure 4

Target versus Income Type: First set Target as continuous variable to see the average default risk rates of different population, we see the unemployed have higher default risk than working people, while businessmen and students have no default risk.

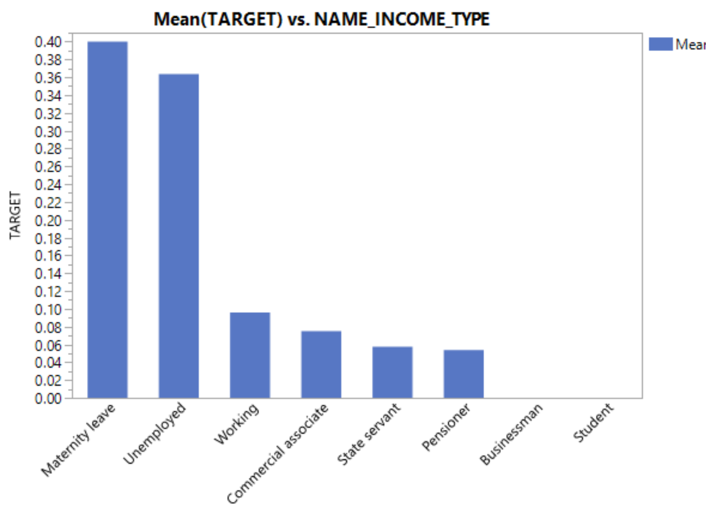


Figure 5

Target versus Education Type: Education levels are negatively correlated with default risk rates and positively correlated with the amounts of income.

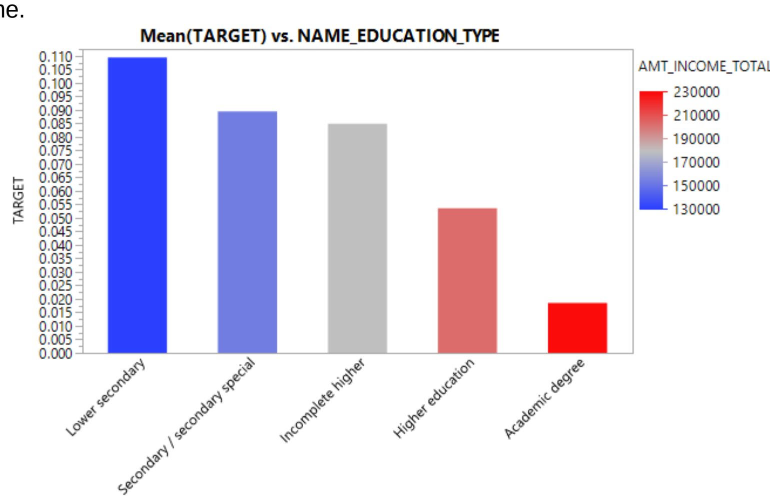


Figure 6

Target versus Occupation Type: Higher skilled workforce has lower default risk, higher income and higher education level compared to the lower skilled.

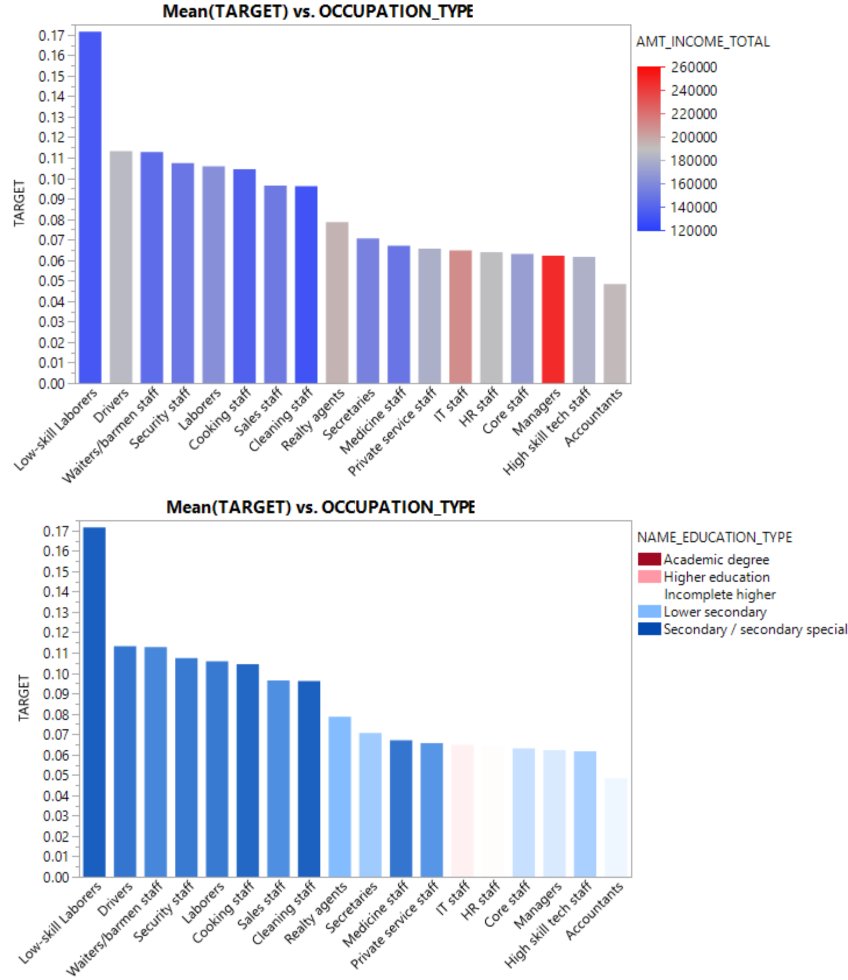


Figure 7

Target versus Gender: More women received loans than man did while women have average lower default risk rate.

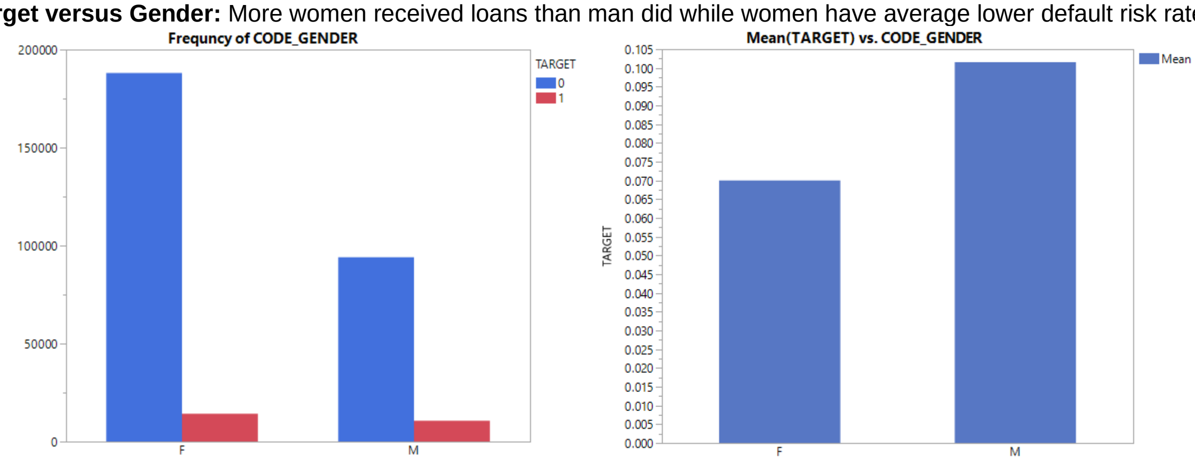


Figure 8

Descriptive analysis of the population who have default risk (Target=1):

In total, 24825(8%) of 307511 customers made late payment on the current loan, may cause loss up to $657,409,294.80. The average age of who has payment difficulty is 41 years old, with a yearly income of $165,611.76 on average. Of them, 43% are male while 57% are female. On average, they spread loans over 20 terms, repaying $26,481.74 as annuity. The ratio of working days over born days is 12.7% versus 16.00% of non-default group.

Figure 9

Pairwise correlation between predictors: The color map on correlations shows that most pairs of variables have low correlation.

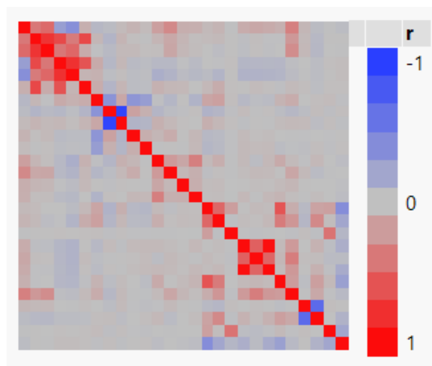


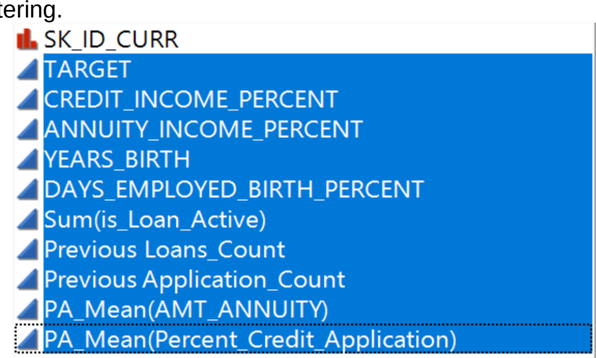
Figure 10 Color Map on Correlations

CLUSTER ANALYSIS

In this part, we will do the clustering analysis for the dataset. We use JMP Pro to do the cluster of the dataset. We make the different customer ID as the observations and keep all the features of customers as variables. There are 307511 observations in the dataset. Because of too large size of the observation, we must to apply the K-Mean clustering in this dataset.

Variable Selection

The K-Mean clustering can only deal with the interval data, so we are supposed to change the binary variable to interval variable and exclude the nominal variable. After that, we find the variables that have too much missing variable, and we exclude these variables. In order to increase the quality of clustering, we plan to choose Target and other 9 variables to do the K-Mean clustering.



K-Mean Clustering

After we choose 10 significant variables, we begin to do the clustering. We range the number of clusters from 3 to 8. From the result, we find that the optimal number of clustering equals to 8.

Method	NCluster	CCC	Best
K Means Cluster	3	-399.83	
K Means Cluster	4	-508.33	
K Means Cluster	5	-376.99	
K Means Cluster	6	-200.51	
K Means Cluster	7	-127.05	
K Means Cluster	8	-54.499	Optimal CCC

Figure 11

In the 8 cluster K-Mean Clustering, the figures show that the cluster 2 and cluster 5 has small sample size, so these clusters are not significant. In the Target column, because the default people are labeled 1, and the no default people are labeled 0. So, we can define the mean of Target as the default rate. It is indicated in the figure that the cluster 1 has the 100 percent default rate. In the meantime, the default rates of the other clusters, except cluster 2 and 5, are all less than overall default rate. In conclusion, the default population was well distinguished as cluster 1. It is good value to analyze the different between the cluster 1 and other clusters.

Cluster Summary		
Cluster	Count	TARGET
1	16193	1
2	84	0.23809524
3	23557	0.00233476
4	27478	0.01171847
5	3	0.33333333
6	33080	0.00535067
7	24658	0.00060832
8	75512	0

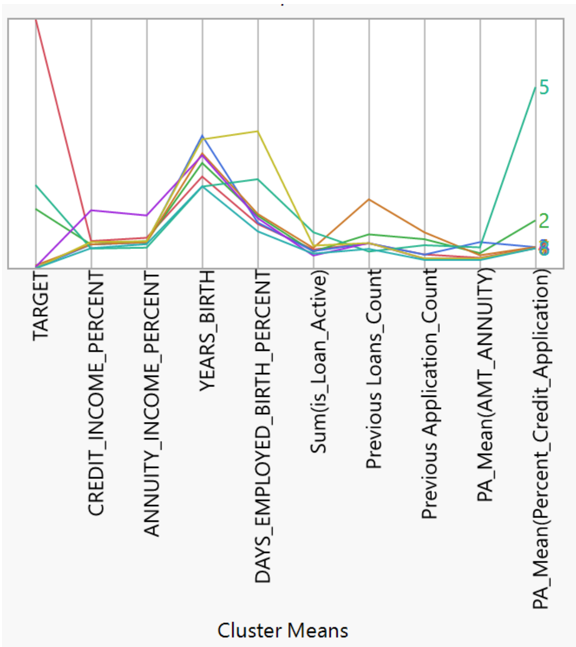


Figure 12

In order to find the features of the cluster 1, we use the Graph Builder to create a Parallel graph about the K-Mean clustering. And we focus on the cluster 1.

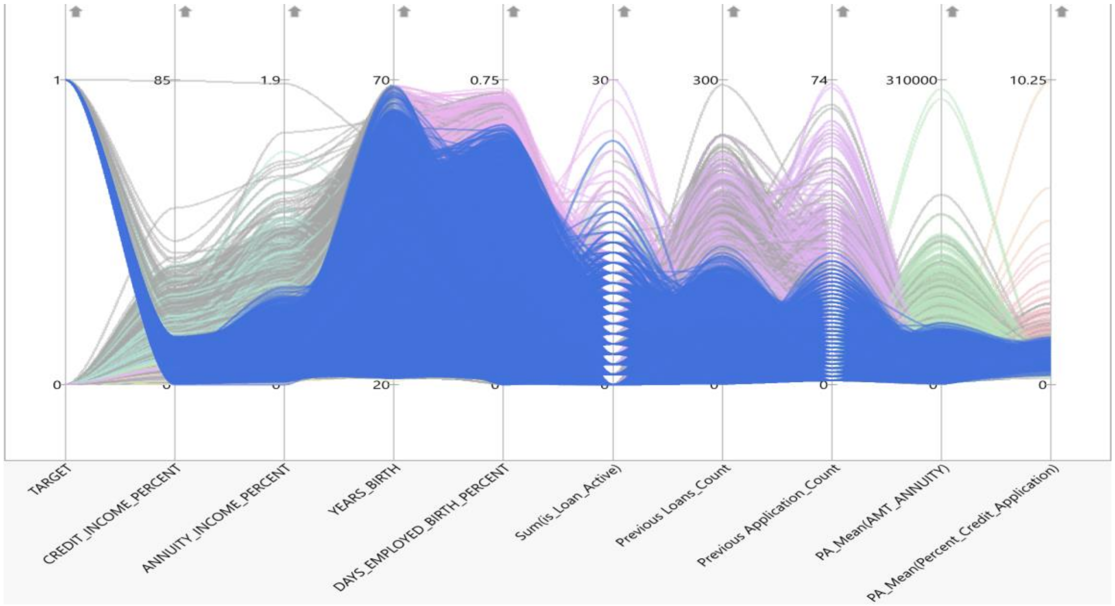


Figure 13

It is demonstrated in the graph, the cluster 1 has relatively low level of Previous_Loans_Count, Previous_Application_Count, PA_Mean(AMT_Annuity) and PA_Mean(Percent_Credit_Application). In four dimensions, the sum credit term of previous loans, the number of previous loans applications, the average previous loans' annuity and the ratio of issued credit to applicated credit, cluster1 has special features. And when these 4 variables of a customer above a special level, the default rate of this people will be very low.

In conclusion of the clustering analysis, we think the pre-loans/pre-application related experience is an important index to predict the default rate.

PREDICTIVE MODELS

To predict a binary target variable, client with payment difficulties or not, we use Decision Tree model and Logistic Regression model, which are insensitive to outliers. We use SAS EM to build the models, the steps are as follow:

- **Data Import and Edit:** Import File "Model_Predict_Credit_Default_Risk" to the diagram. Set "SK_ID_CURR" as ID variable and "TARGET" as Target variable; set the left variables as input variables. Set income type, education type and occupation type variable as nominal variable; set TARGET as binary variable and the left as interval variables.
- **Data Partition:** Split the data into 4:3:3, where 40% for train, 30% for validation, 30% for test.
- **Logistic Regression:** Before doing Logistic Regression, create dummy indicators for nominal variables and log 10 transform interval variables to stabilize variance and improve model fits. Since regression ignore altogether observations that contain missing values, which reduces the size of the data set and weak the predictive power of the model, we impute missing value using "tree surrogate" and "median" impute method for class and interval variables respectively. The reason of using the median impute method is that it is less sensitive to extreme values than the mean and is therefore more accurate to be used for missing values. We set Logistic Regression model based on different variable selection methods, including "backward", "forward" and "stepwise".

Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Backward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...

Figure 14 Regression Node

- **Decision Tree:** We set Decision Tree model based on different maximum branch: 2, 3 and 4 branches. We choose "LARGEST" to build the full tree. Set maximum depth as 50 and minimum split size as 25 observations. We use Average Square Error (ASE) indicator to measure the goodness of the model because the higher the value of target variable, the higher the default risk.

Splitting Rule	
Interval Target Criterio	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.1
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	50
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rule	0
Split Size	25
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Largest
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25

Figure 15 Decision Tree Node

- **Model Comparation:** We assess and compare models by ASE measure.

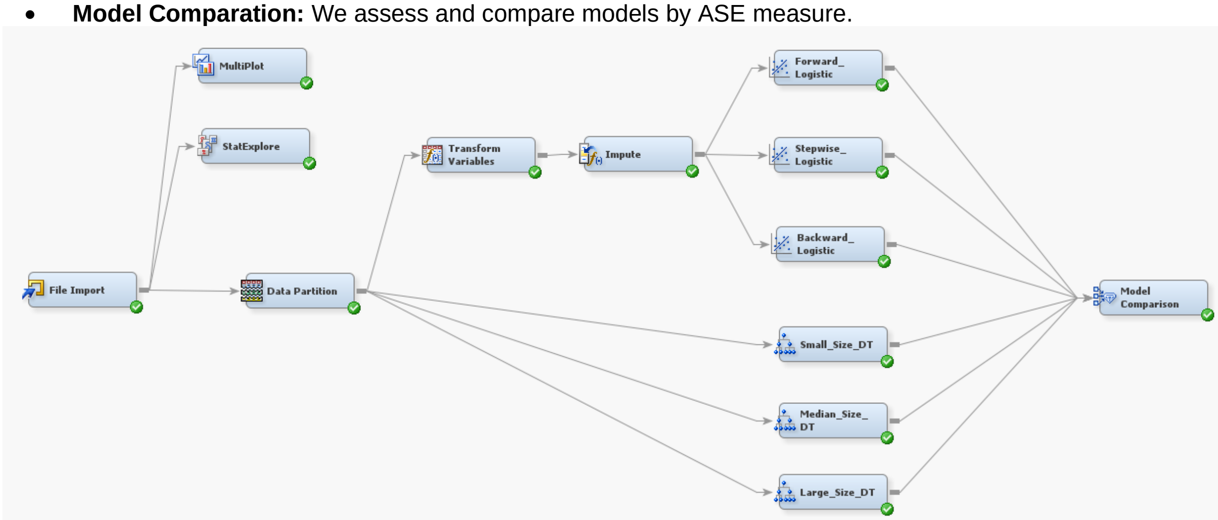


Figure 16 Process Flow in SAS EM

Result of Logistic Regression Model

Of the three logistic regression models, the model using "backward" selection method has the lowest ASE, which means it is the best fit model. From the Iteration Plot, the model get the best result at the 12^{th} step.

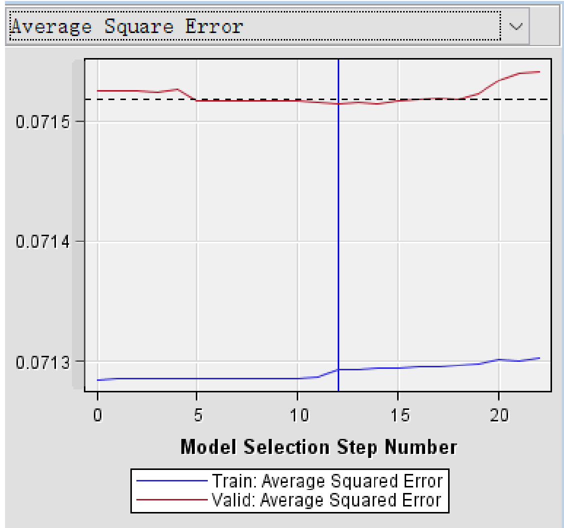


Figure 17 Iteration Plot

From the output of the model results, 51 variables (include dummy variables) are selected into the final model. The overall Chi Square statistics is less than 0.0001, which indicates that the model is valid and fits well.

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
68930.661	64539.586	4391.0750	51	<.0001

Figure 18

Result of Decision Tree Model

Of the three Decision Tree models, the model with 4 maximum branches has the lowest ASE. From the results of the model, the fit statistics tell us the misclassification rate for validation data is around 8.28% and the Average Square Error is 0.0731. These two statistics are similar between train and validation data, indicating consistent and valid result.

Target	Target Label	Fit Statistics	Statistics Label	Train ▲	Validation	Test
TARGET	TARGET	_ASE_	Average Squared Error	0.069741	0.073054	0.073098
TARGET	TARGET	_MISC_	Misclassification Rate	0.080262	0.08275	0.08252
TARGET	TARGET	_RASE_	Root Average Squared Error	0.264086	0.270284	0.270367
TARGET	TARGET	_MAX_	Maximum Absolute Error	0.991826	1	1
TARGET	TARGET	_SSE_	Sum of Squared Errors	17129.98	13457.78	13466.28
TARGET	TARGET	_NOBS_	Sum of Frequencies	122811	92109	92111
TARGET	TARGET	_DFT_	Total Degrees of Freedom	122811	.	.
TARGET	TARGET	_DIV_	Divisor for ASE	245622	184218	184222

Figure 19 Fit Statistics

From Variable Importance table, the most significant variable is “credit term”, followed by “days employed” and “maximum days of credit end-date” and so on. A higher importance value means a greater importance of the variable in predicting the default risk rate. The most common splitting rule is Credit Term, which implies that Credit Term is a significant characteristic to distinguish between different groups of clients.

Variable Name	Number of Splitting Rules	Importance ▼	Validation Importance
CREDIT_TERM	30	1.0000	1.0000
DAYS_EMPLOYED	8	0.3765	0.3803
Max_DAYS_CREDIT_ENDDATE_	8	0.2482	0.0376
VAR30	7	0.2428	0.1595
Sum_Active_Loan_	9	0.2393	0.1114
INST_Max_Days_Overdue_	9	0.2360	0.0000
NAME_EDUCATION_TYPE	5	0.2360	0.1736
CC_Mean_AMT_BALANCE_	3	0.2185	0.1479
INST_Min_AMT_Payment___Installme	6	0.2114	0.0000
YEARS_BIRTH	7	0.2048	0.0996
CREDIT_INCOME_PERCENT	5	0.1978	0.0405
DAYS_EMPLOYED_BIRTH_RATIO	4	0.1828	0.0164
PA_Mean_AMT_ANNUITY_	5	0.1740	0.0000
CC_Mean_AMT_INST_MIN_	6	0.1721	0.0000
AMT_CREDIT	4	0.1648	0.0967
DAYS_ID_PUBLISH	5	0.1588	0.0340
Previous_Loans_Count	3	0.1554	0.0934
PL_Max_Days_Overdue_	5	0.1429	0.0000
Previous_Application_Count	3	0.1400	0.0828
PA_Mean_RATE_DOWN_PAYMENT_	4	0.1323	0.0000
ANNUITY_INCOME_PERCENT	2	0.1322	0.0000
Max_AMT_CREDIT_SUM_LIMIT_	2	0.1267	0.0885
OCCUPATION_TYPE	1	0.1204	0.0000
CC_Mean_AMT_CREDIT_LIMIT_	1	0.1026	0.0000
Max_AMT_CREDIT_SUM_	3	0.1024	0.0000
AMT_ANNUITY	1	0.0844	0.0000
DAYS_REGISTRATION	2	0.0709	0.0232
PL_Max_CNT_INSTALMENT_FUTURE_	2	0.0592	0.0000
AMT_INCOME_TOTAL	1	0.0574	0.0000
NAME_INCOME_TYPE	0	0.0000	0.0000

Figure 20 Variable Importance Table

The plot shows ten variables with the lowest p-value which indicates the highest predictive power of the variables. Red represents positive relationship with the target variable while blue represents negative relationship. In other words, the larger the values of the blue variables, the lower the default risk rate. In conclusion, people with higher default risk tend to have the following characteristics: lower age, smaller amount of down-payment, lower rate of payment by installment, more active loans, more loan applications and more days overdue.

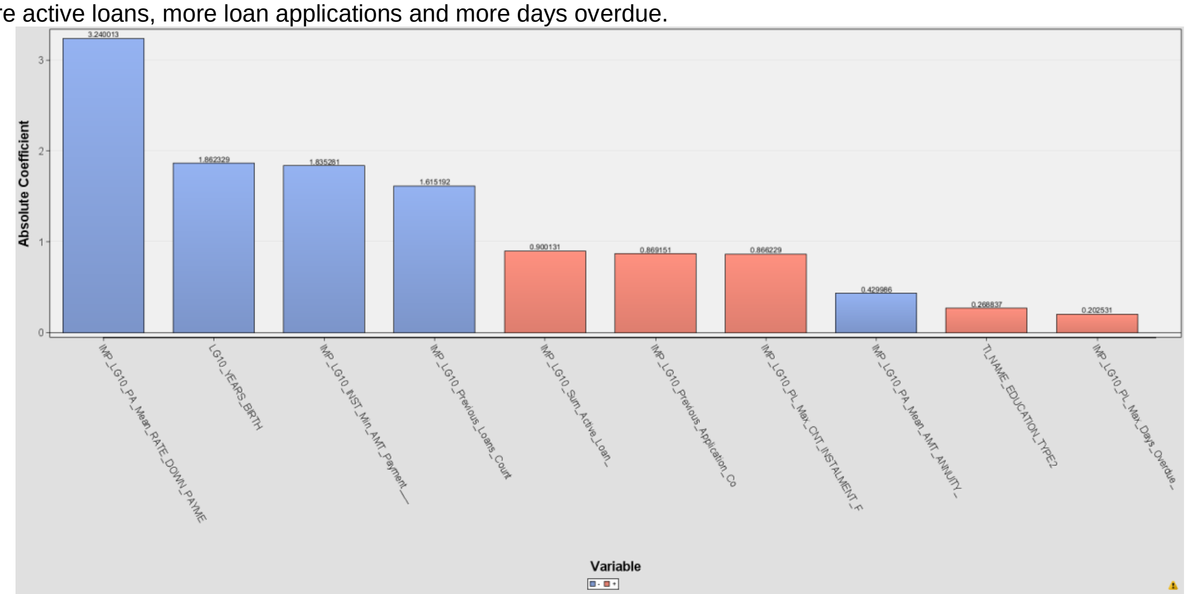


Figure 21 Absolute Coefficient Plot

Model Comparison between Logistic Regression and Decision Tree

Each model has low ASE at around 0.07 and misclassification at around 0.08. All the models show high Accuracy rate at above 91%, which indicates that the model is estimated to give an accurate prediction 91% of the time. Overall, the Decision Tree models have lower ASE and misclassification rate compared to logistic regression models, which means greater fitness of the Decision Tree models.

Fit Statistics
Model Selection based on Train: Average Squared Error (_ASE_)

Selected Model	Model Node	Model Description	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error	Valid: Misclassification Rate
Y	Tree	Large_Size_DT	0.069741	0.080262	0.073054	0.082750
	Tree3	Median_Size_DT	0.070187	0.080245	0.073503	0.081990
	Tree2	Small_Size_DT	0.071079	0.080644	0.072986	0.081610
	Reg2	Backward_Logistic	0.071191	0.080815	0.071356	0.080796
	Reg	Stepwise_Logistic	0.071262	0.080799	0.071411	0.080806
	Reg3	Forward_Logistic	0.071262	0.080799	0.071411	0.080806

Figure 22 Fit Statistics of Models

MODEL	ACCURACY
LARGE SIZE DECISION TREE	91.97%
MEDIAN SIZE DECISION TREE	91.98%
SMALL SIZE DECISION TREE	91.94%
BACKWARD LOGISTIC	91.92%
STEPWISE LOGISTIC	91.92%
FORWARD LOGISTIC	91.92%

Figure 23 Model Accuracy

The large size Decision Tree has the largest cumulative lift and larger ROC. From ROC curve, a plot of the true positive rate (y-axis) versus the false positive rate (x-axis), the further it is from the diagonal line, the better the model is for discriminating between negatives and positives.

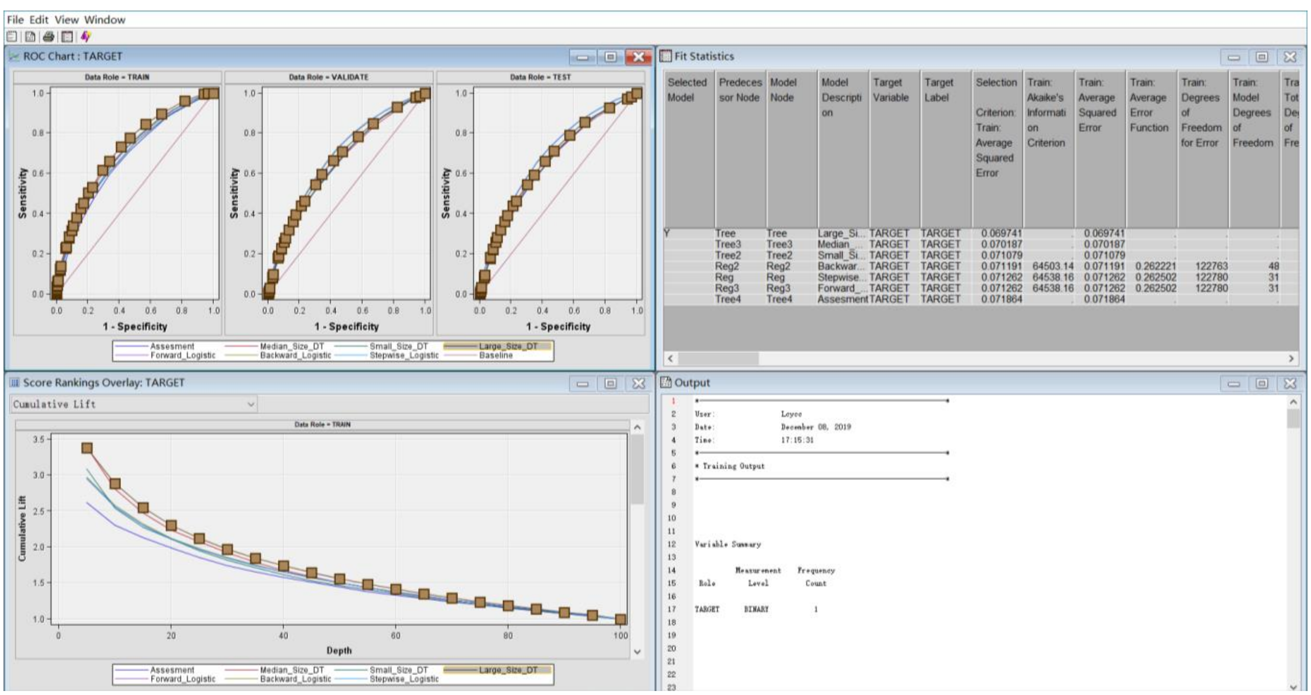


Figure 24

FUTURE WORK

This paper aims to build a predictive model based on alternative data, there are several future works to improve on the model. Through predictive modelling, most of variables are founded highly significant with target variable, however, contributes quite a few R-Square (a statistic indicator represents the fitness of regression) to the model. One of the reasons attributes to our ineffective features building. Advanced feature engineering technique would be implemented such as PCA and boosting in the future work.

CONCLUSION

According to the result of the models, clients who need to pay higher annuities are more likely to make late payments or default, so the company shall appropriately extend the credit term for high-risk groups with a low annuity. The high-default-risk population usually have the following set of characteristics: under three years working experience, high number of overdue days of previous loans, low age and high number of active loans. Specifically, we selected the highest risk (45%) population, whose average age below 37, annuity/income ratio greater than or equal to 0.071 and high school education applying largest tree method, which is supposed to be paid close attention to.

For these clients, companies shall reduce credit amount, increase down payment rate or collateral, reduce annuity and extend credit term to reduce payment difficulties and as a result, reduce the number of late payments and default risk.

REFERENCE

[1] Home Credit Group (2018), Home Credit Default Risk, from
https://www.kaggle.com/c/home-credit-default-risk/overview/description

[2] Will Koehrsen (2018), Start Here: A Gentle Introduction, from

https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction/data#Introduction:-Home-Credit-Default-Risk-Competition

[3] Will Koehrsen (2018), Introduction to Manual Feature Engineering, from

https://www.kaggle.com/willkoehrsen/introduction-to-manual-feature-engineering/notebook#Results

[4] Lathwal (2017), Home Credit: Complete EDA + Feature Importance, from

https://www.kaggle.com/codename007/home-credit-complete-eda-feature-importance#5.-Data-Exploration

[5] The Economist Intelligence Unit（2019）, Singapore: risk assessment, from

http://country.eiu.com.libproxy.smu.edu.sg/article.aspx?articleid=1908497574&Country=Singapore&topic=Risk&subtopic=Credit+risk&subsubtopic=Overview

ACKNOWLEDGEMENTS

We would like to extend our sincere gratitude to Prof. Kam Tin Seong, Associate Professor of Information Systems (Practice) at the Singapore Management University (SMU) for offering constant encouragement and support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Liu Cuiyi
Enterprise: Singapore Management University
E-mail: cuiyi.liu.2019@mitb.smu.edu.sg

Name: Qiu Yang
Enterprise: Singapore Management University
E-mail: yang.qiu.2019@mitb.smu.edu.sg

Name: Xu Shifeng
Enterprise: Singapore Management University
E-mail: sfxu.2019@mitb.smu.edu.sg