WQD7005
Data Mining Group Assignment and Project

Instructor: Prof Dr. Teh Ying Wah
Project Title: Credit Card Approval Prediction

Group member:

Lee Ying Qiu	17108552
Clement Lee	S2128268
Ng Sin Yu	S2111068
Syalin Dania	S2018106

Introduction

- Commercial banks receive numerous amount of credit card application
- How they approve/reject? Credit record is one of the important factors in determining applicant's eligibility
- A good credit record is often associated with lower risk and vice versa
- Manually assess the credit record? Infeasible
- Or we can automate the assessment process?

LEE YING QIU 17108552

Analysis Goal

- A bank seek to automate the credit assessment process to increase the processing volumes by predicting the credit risk of an applicant when applying for credit card
- Objectives:

a) To understand the determinants in determining the credit risk of an applicant (Group Assignment)

b) To explore the sample dataset to find patterns between the variables before developing model to predict the credit risk of applicants (Group Assignment)

c) To perform data cleansing to remove noisy data (Group Project)

d) To implement various data mining techniques (e.g., HP Decision Tree, HP Random Forest, HP SVM, and HP Neural Network) and create predictive models (Group Project)

e) To examine the accuracy and reliability of the candidate predictive models (Group Project)

LEE YING QIU 17108552

Analysis Data and Data Pre-processing

- 2 datasets: application_record.csv & credit_record.csv selected from Kaggle
- application_record.csv: describes applicant's personal information and basic financial status
- credit_record.csv: credit record of the applicant engaged with the bank
- 2 datasets are merged by the common key: ID (using SAS Studio)
- New attributes: TARGET, ACCOUNT_LENGTH, AGE, YEARS_EMPLOYED

LEE YING QIU 17108552

Accessing and Assaying Prepared Data

- Data Source Wizard (SAS e-miner) auto assign the Role and Level for each variable
- However, some of the assigned measurement level should be reassigned

Name	Role	Level
ACCOUNT_LENGTH	Input	Interval
AGE	Input	Interval
AMT_INCOME_TOTAL	Input	Interval
CNT_CHILDREN	Input	Interval
CNT_FAM_MEMBERS	Input	Interval
EDUCATION_TYPE	Input	Interval
EMAIL	Input	Interval
FAMILY_STATUS	Input	Interval
GENDER	Input	Interval
HOUSING_TYPE	Input	Interval
ID	ID	Nominal
INCOME_TYPE	Input	Interval
OCCUPATION_TYPE	Input	Nominal
OWN_CAR	Input	Interval
OWN_MOBIL	Input	Interval
OWN_REALTY	Input	Interval
PHONE	Input	Interval
TARGET	Target	Interval
WORK_PHONE	Input	Interval
YEARS_EMPLOYED	Input	Interval

Name	Role	Level
ACCOUNT_LENGTH	Input	Interval
AGE	Input	Interval
AMT_INCOME_TOTAL	Input	Interval
CNT_CHILDREN	Input	Interval
CNT_FAM_MEMBERS	Input	Interval
EDUCATION_TYPE	Input	Nominal
EMAIL	Input	Binary
FAMILY_STATUS	Input	Nominal
GENDER	Input	Binary
HOUSING_TYPE	Input	Nominal
ID	ID	Nominal
INCOME_TYPE	Input	Nominal
OCCUPATION_TYPE	Input	Nominal
OWN_CAR	Input	Binary
OWN_MOBIL	Input	Binary
OWN_REALTY	Input	Binary
PHONE	Input	Binary
TARGET	Target	Binary
WORK_PHONE	Input	Interval
YEARS_EMPLOYED	Input	Interval

VARIABLES SUCH AS *EMAIL, GENDER, OWN_CAR, OWN_MOBIL, OWN_REALTY, PHONE* AND *TARGET* WITH ONLY TWO POSSIBLE LEVELS: 0 AND 1 WILL BE RE-ASSIGNED TO BINARY CLASS WHILE VARIABLES WITH MORE THAN 2 LEVELS WILL BE RE-ASSIGNED TO NOMINAL CLASS.

LEE YING QIU 17108552

Methodology

- SAS 'SEMMA'
- Focus on 1st 2 Letters 'S', 'E' for this Group Assignment

- SAMPLE - Data sampled not too large , but large enough to contain sufficient features, variables and observations for analysis
- EXPLORE – Univariate, Bivariate and Multivariate Analysis

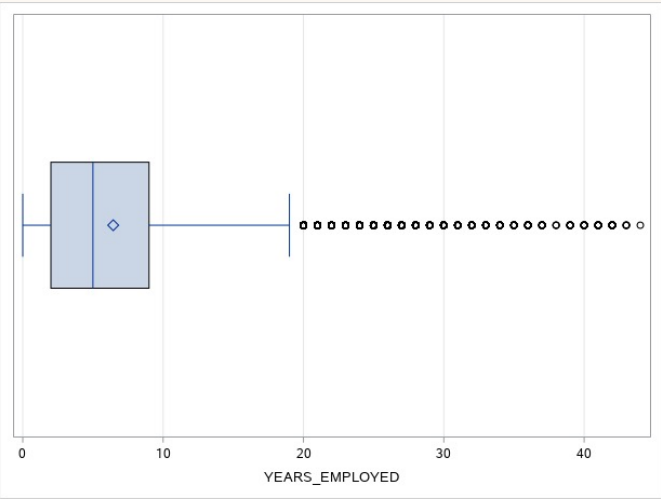
LEE YING QIU 17108552

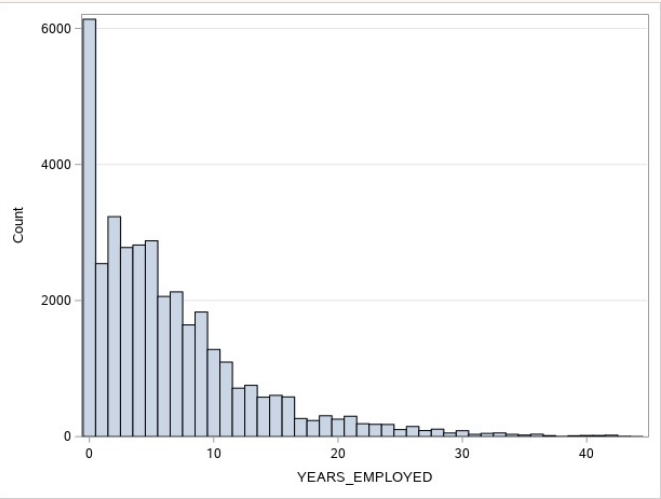
Univariate Analysis

- Analysis of a single variable one at a time
- Boxplot and Histogram plotted for Interval Variables
- Bar Chart plotted for Binary and Nominal Variables
- Check for completeness, consistency and noise
- Pie chart for TARGET Variable to visualize proportion on High Risk vs Low Risk profiles

CLEMENT LEE S2128268

Univariate Analysis : Interval Variables

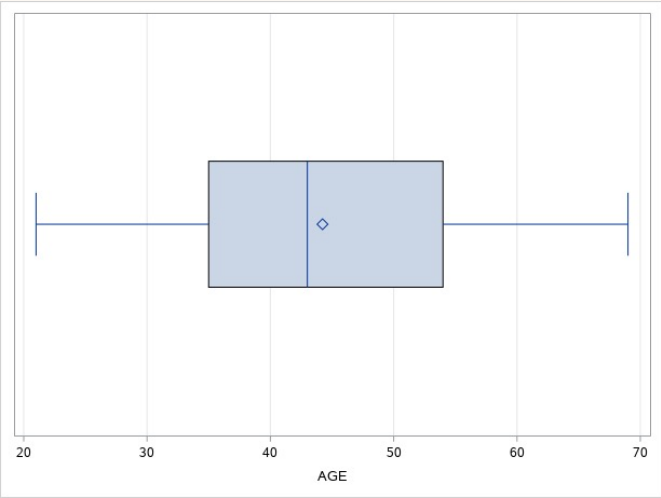


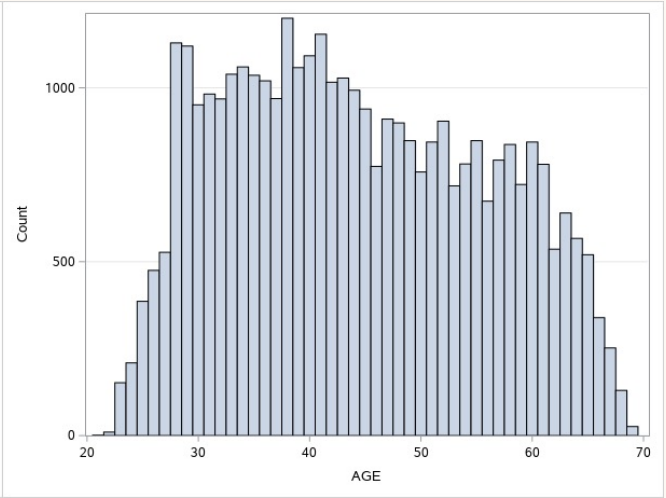


Boxplot and Histogram plotted to visualize distribution and identify outliers

CLEMENT LEE S2128268

Univariate Analysis : Interval Variables

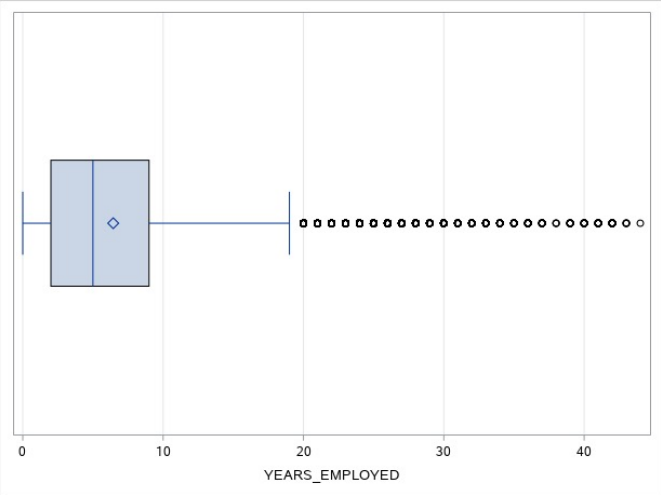


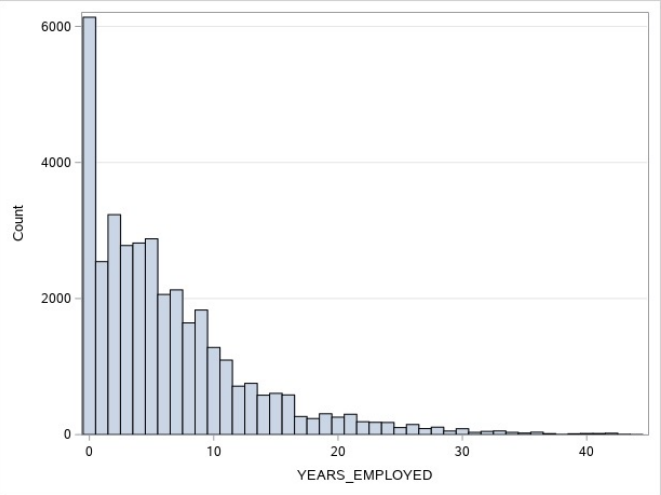


Boxplot and Histogram of AGE Variable

CLEMENT LEE S2128268

Univariate Analysis : Interval Variables



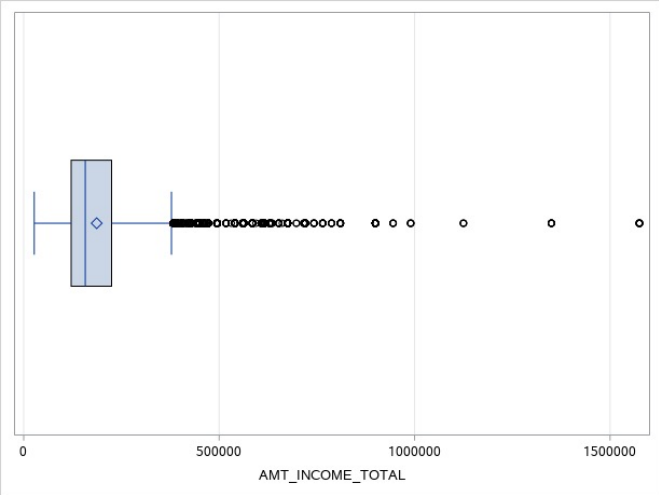


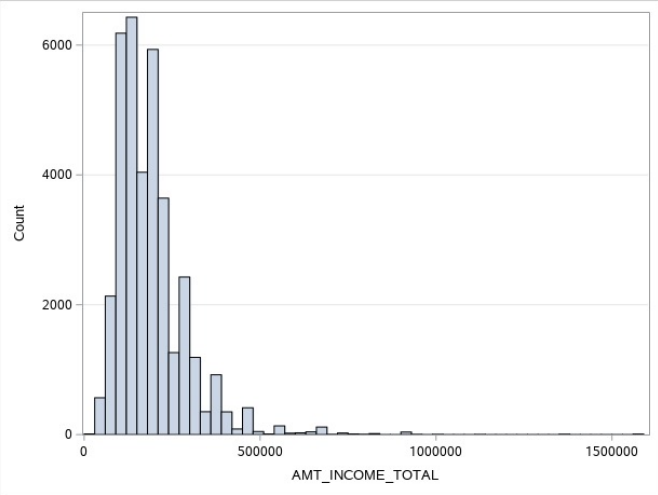
Boxplot and Histogram of YEARS_EMPLOYED Variable

- Outliers not to be excluded
- Ages up to 69

CLEMENT LEE S2128268

Univariate Analysis : Interval Variables



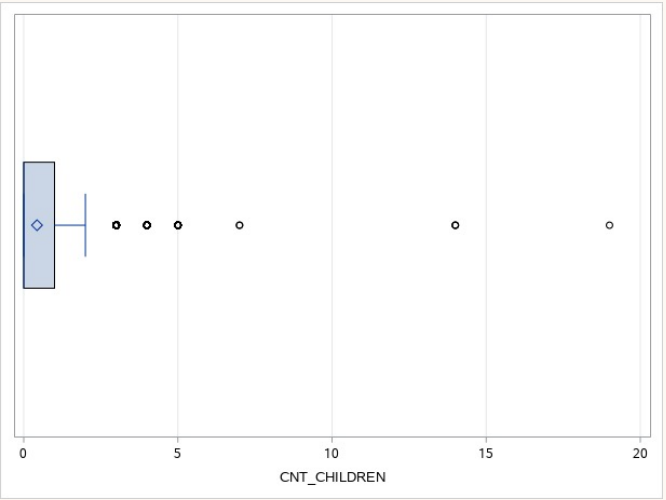


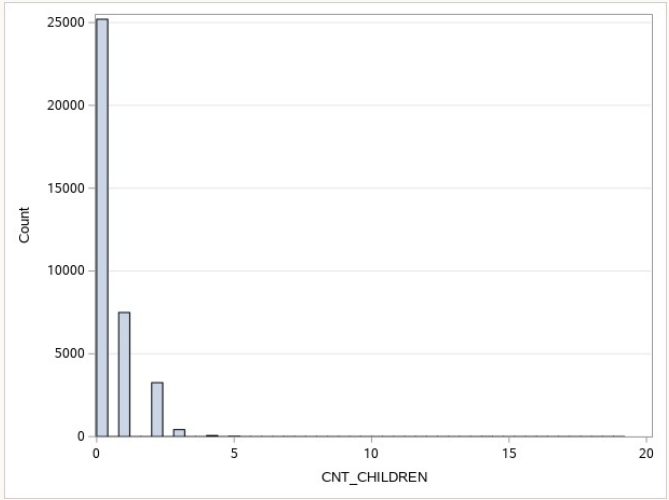
Boxplot and Histogram of AMT_INCOME_TOTAL Variable

- Outliers not to be excluded
- There are persons with high number of years employed

CLEMENT LEE S2128268

Univariate Analysis : Interval Variables



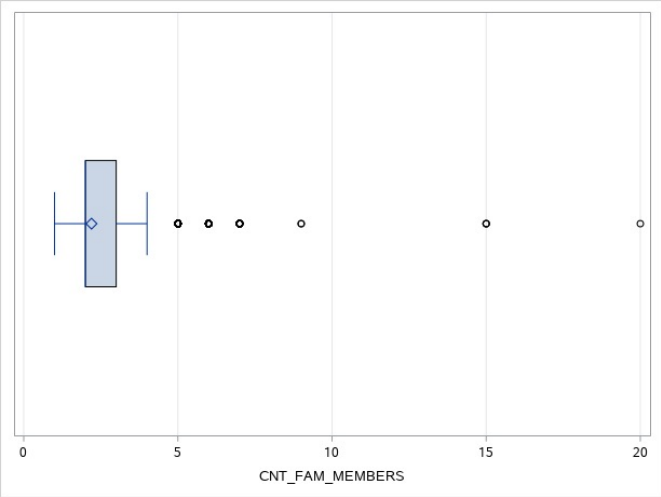


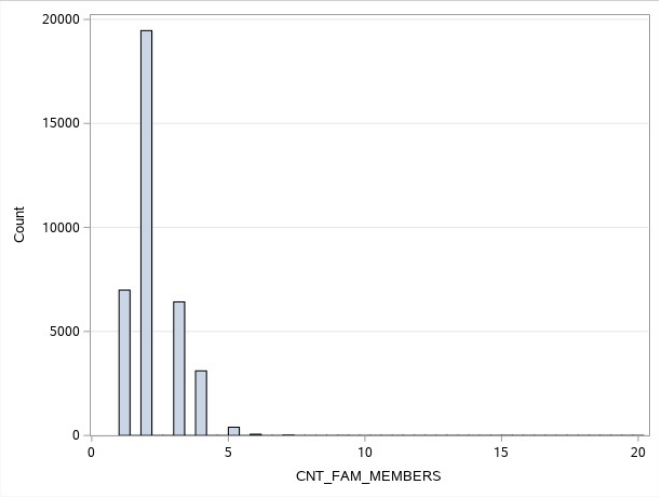
Boxplot and Histogram of CNT_CHILDREN Variable

- Outlier at 19 to exclude
- Possible error

CLEMENT LEE S2128268

Univariate Analysis : Interval Variables



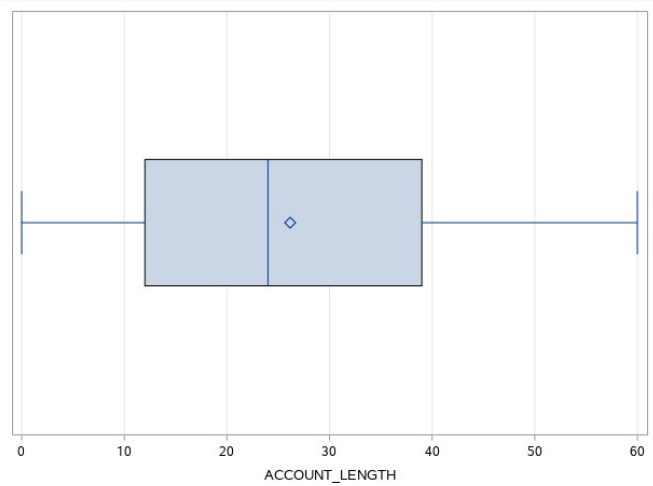


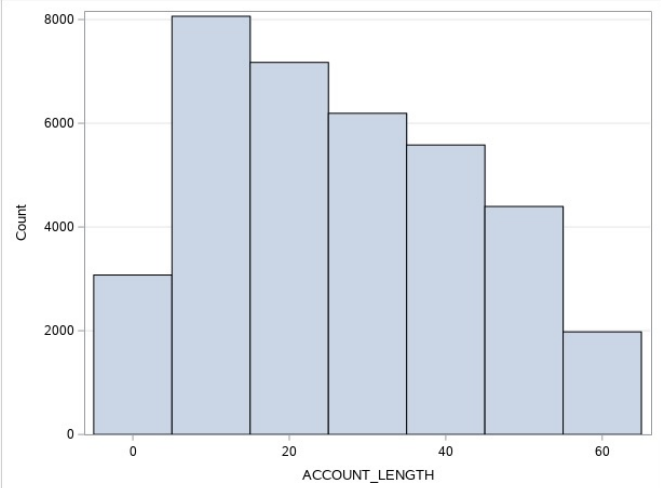
Boxplot and Histogram of CNT_FAM_MEMBERS Variable

- Outlier at 20 to exclude
- Possible error

CLEMENT LEE S2128268

Univariate Analysis : Interval Variables





Boxplot and Histogram of ACCOUNT_LENGTH Variable

CLEMENT LEE S2128268

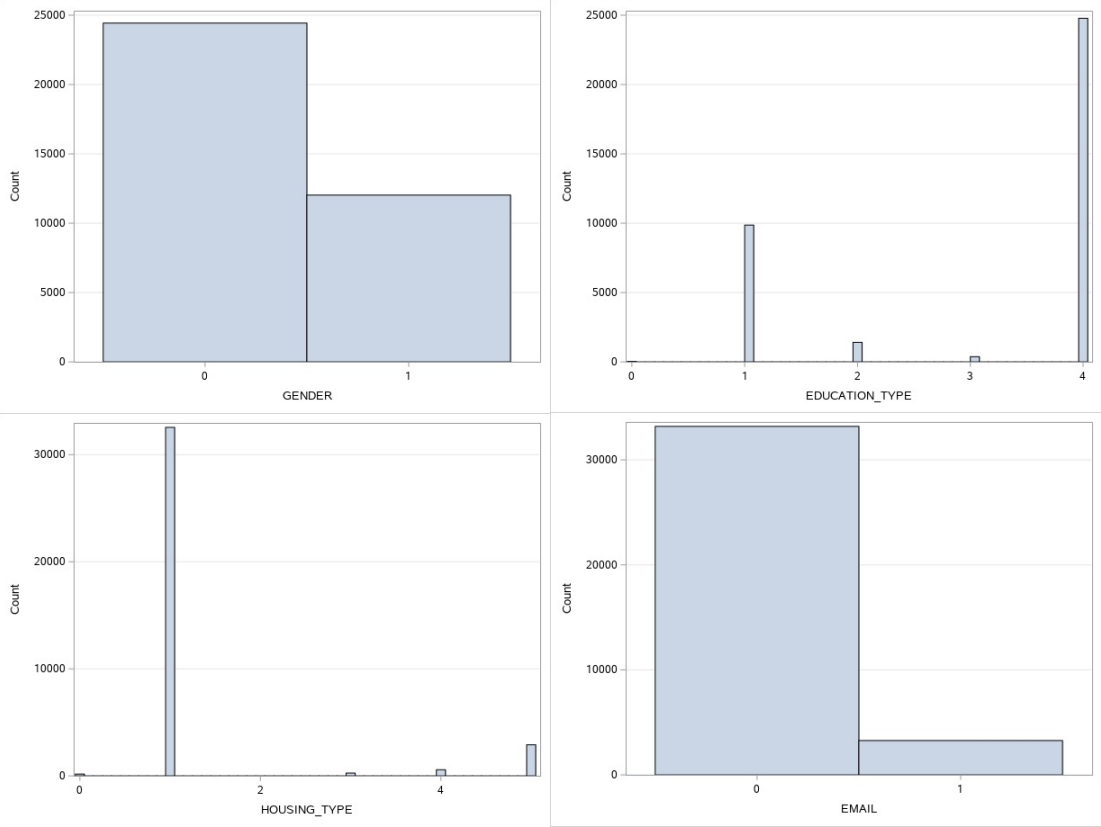
Univariate Analysis : Interval Variables

CLEMENT LEE S2128268

Descriptive Statistics for Numeric Variables

Variable	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
AGE	36457	0	21.0000000	44.2313685	43.0000000	69.0000000	11.5041274
YEARS_EMPLOYED	36457	0	0	6.4444963	5.0000000	44.0000000	6.5706770
AMT_INCOME_TOTAL	36457	0	27000.00	186685.74	157500.00	1575000.00	101789.23
CNT_CHILDREN	36457	0	0	0.4303152	0	19.0000000	0.7423669
CNT_FAM_MEMBERS	36457	0	1.0000000	2.1984530	2.0000000	20.0000000	0.9116861
ACCOUNT_LENGTH	36457	0	0	26.1641934	24.0000000	60.0000000	16.5018545

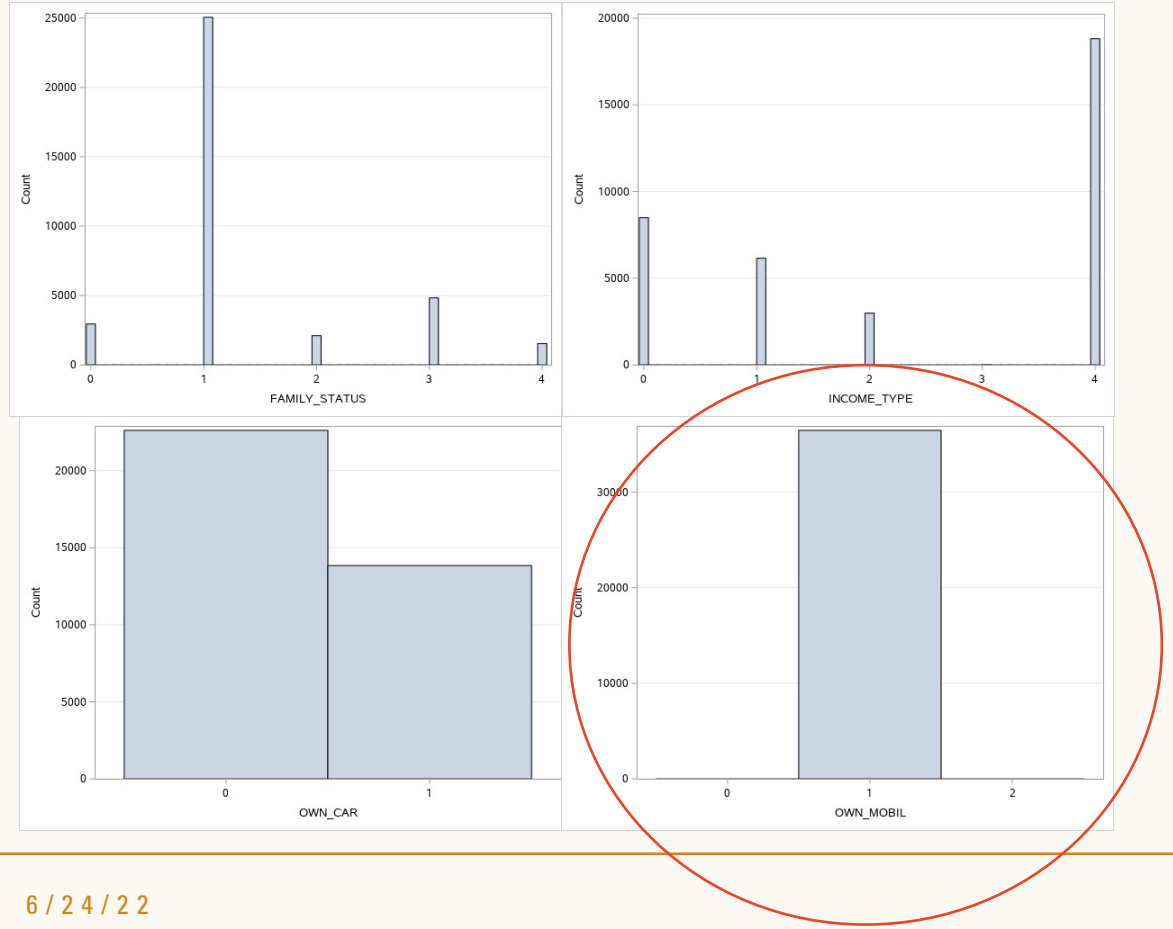
Univariate Analysis : Binary and Nominal Variables



Bar Chart plotted to check inconsistencies and noise

CLEMENT LEE S2128268

Univariate Analysis : Binary and Nominal Variables

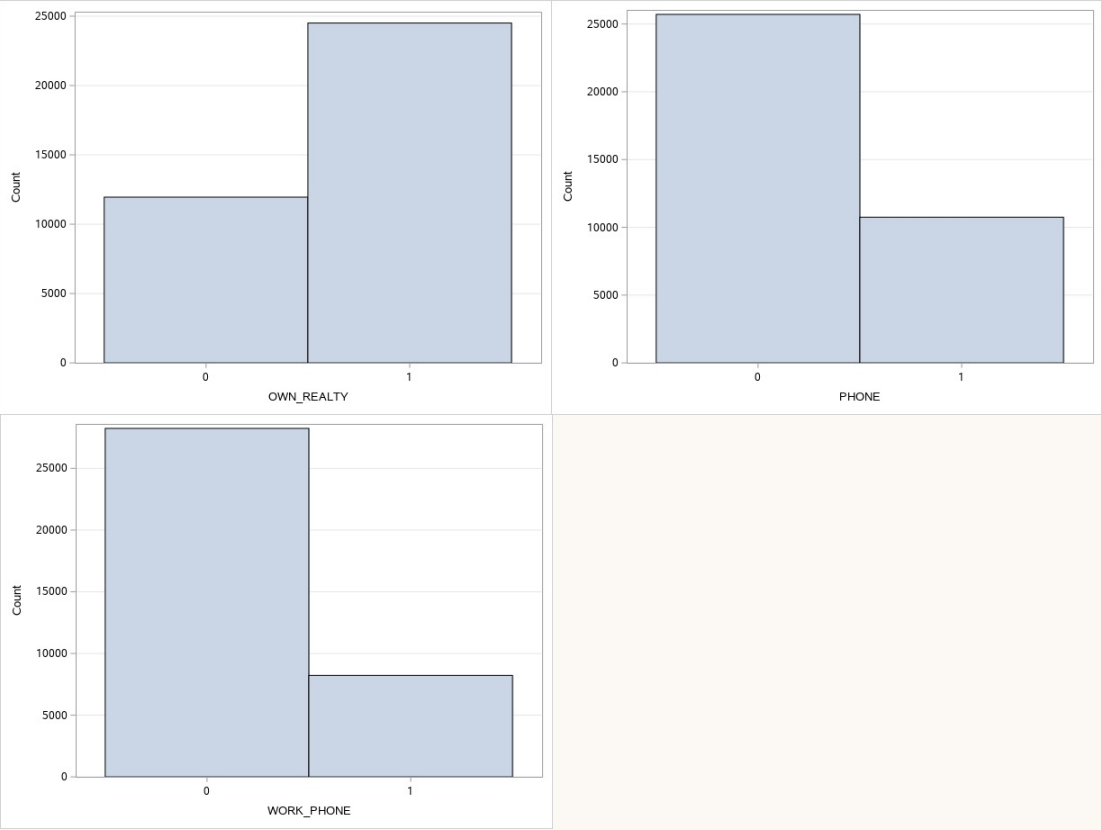


Bar Chart plotted to check inconsistencies and noise

- OWN_MOBIL is constant variable and to be excluded from further analysis

CLEMENT LEE S2128268

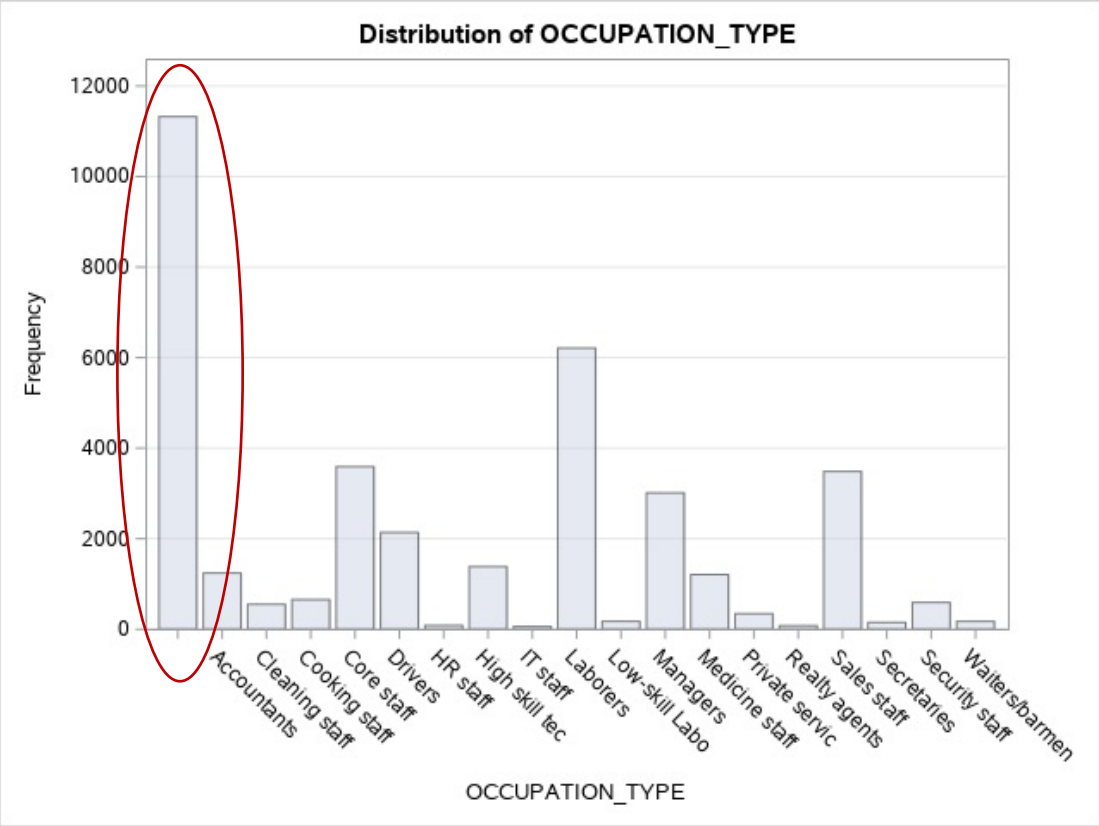
Univariate Analysis : Binary and Nominal Variables



Bar Chart plotted to check inconsistencies and noise

CLEMENT LEE S2128268

Univariate Analysis : Binary and Nominal Variables

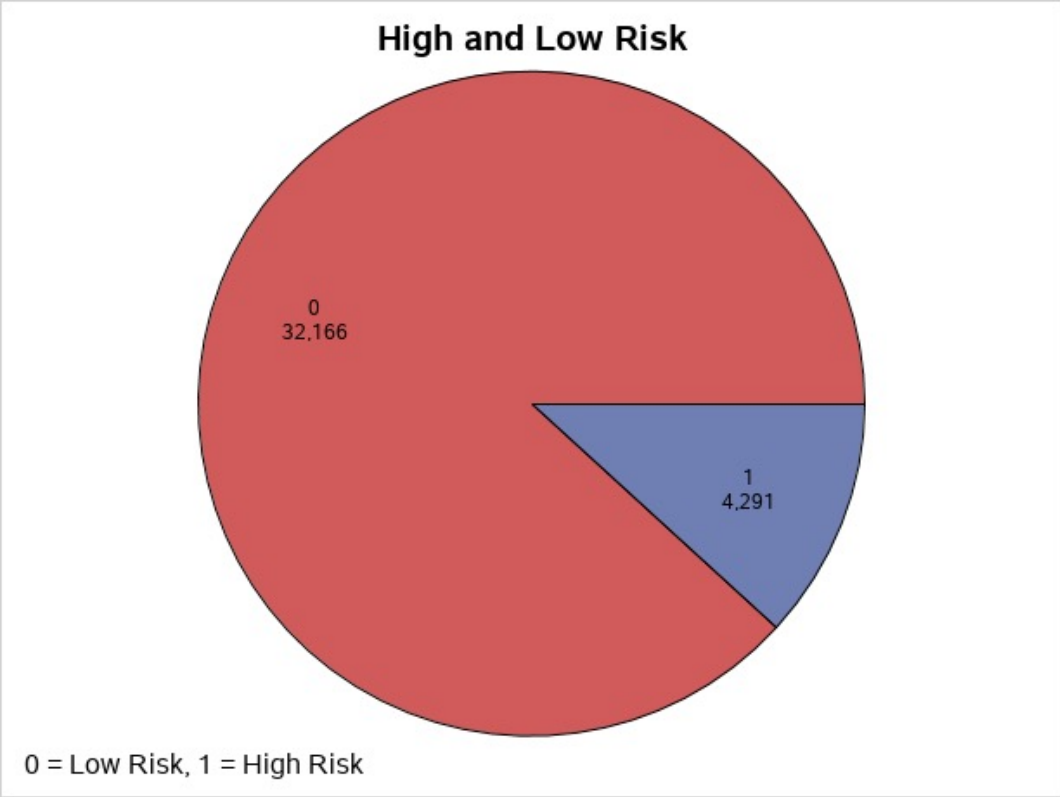


Bar Chart plotted to check inconsistencies and noise

- High number of missing values in OCCUPATION_TYPE

CLEMENT LEE S2128268

Univariate Analysis : Binary and Nominal Variables



Pie Chart plotted to visualize proportion of High Risk vs Low Risk profiles

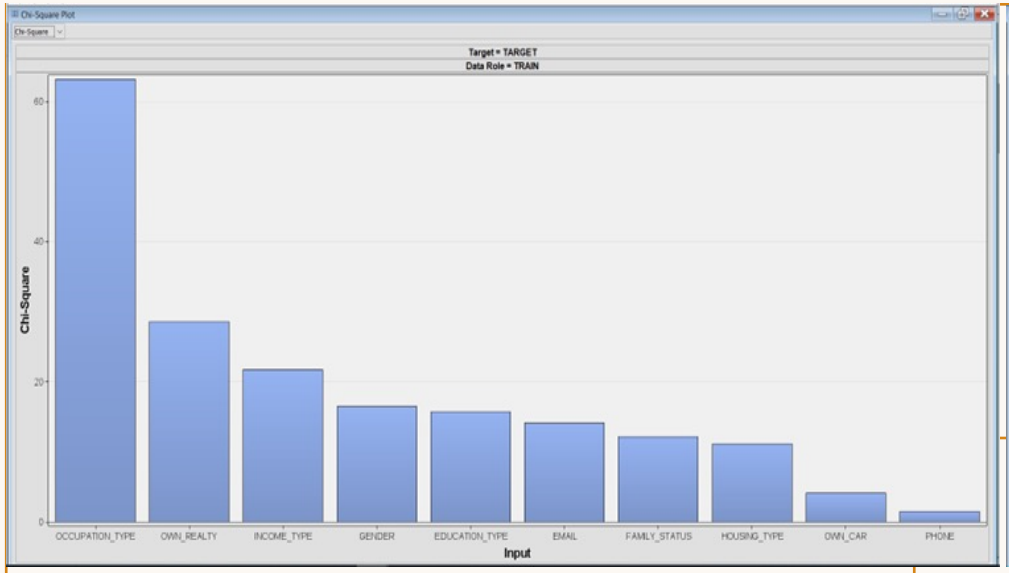
CLEMENT LEE S2128268

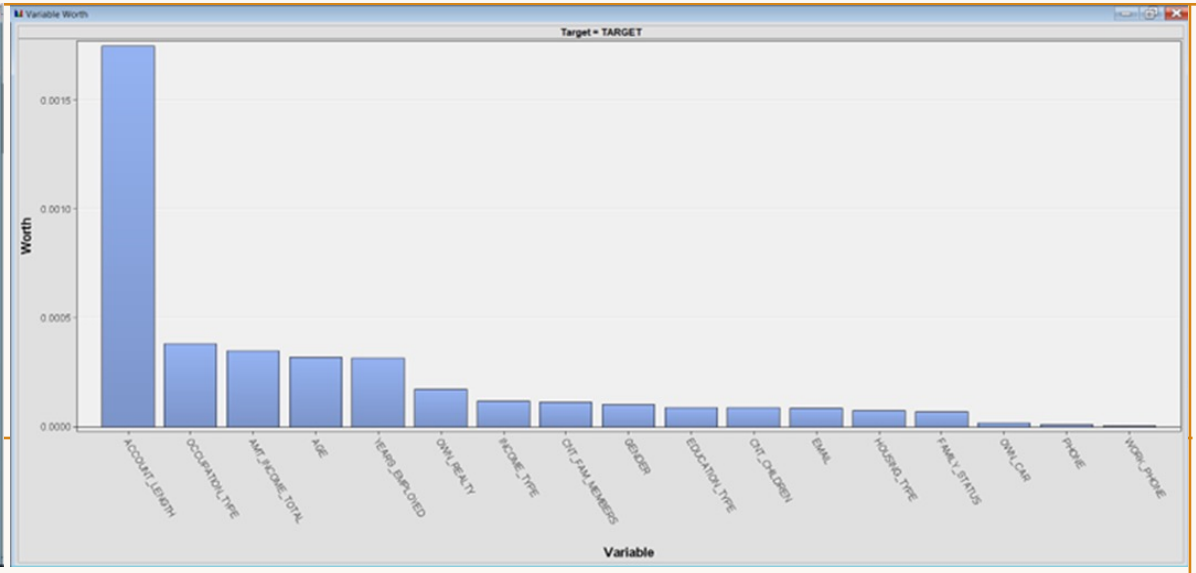
Univariate Analysis

- Dataset re-checked for missing values

ID	GENDER	OWN_CAR	OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	INCOME_TYPE	FAMILY_STATUS	HOUSING_TYPE	EDUCATION_TYPE	OWN_MOBIL
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing

WORK_PHONE	PHONE	EMAIL	OCCUPATION_TYPE	CNT_FAM_MEMBERS	AGE	YEARS_EMPLOYED	TARGET	ACCOUNT_LENGTH	Frequency	Percent
Non-missing	Non-missing	Non-missing		Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	11323	31.0585
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	25134	68.9415

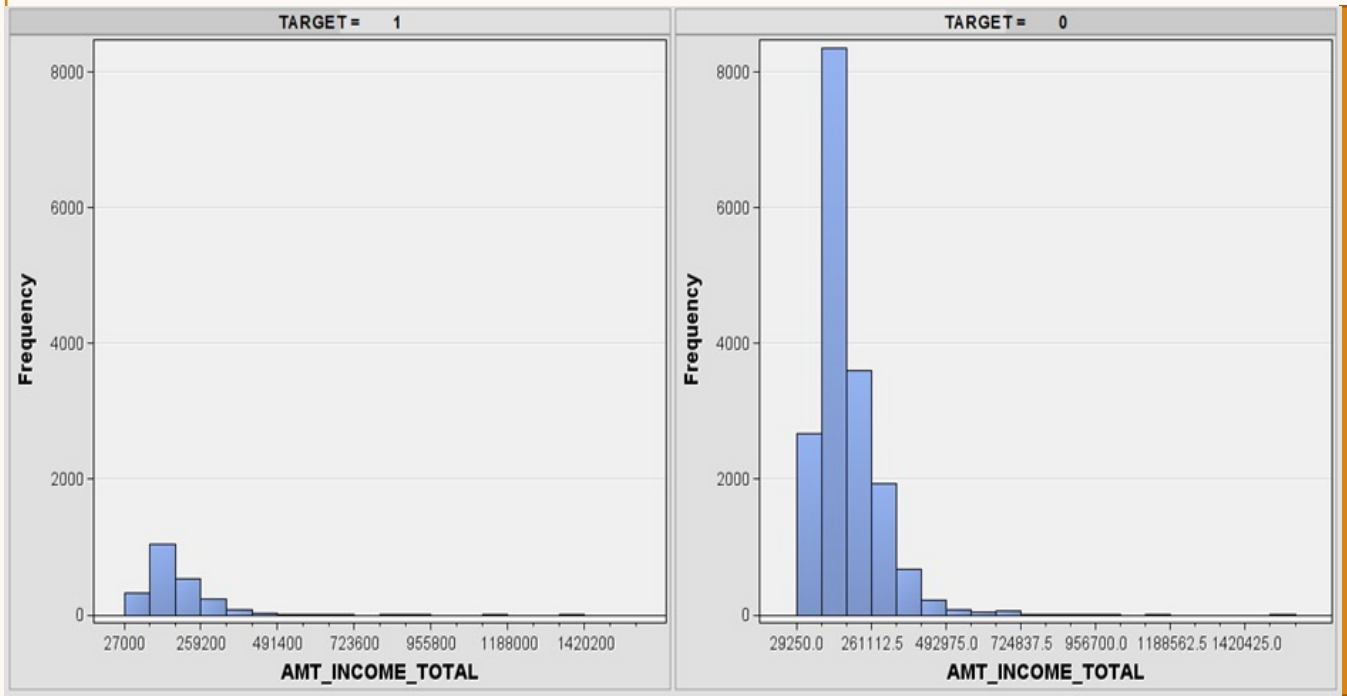




Bivariate Analysis

- Usually conducted to identify if a statistical association exists in between two variable
- Worth Analyse bar chart & Chi Square Test
- Target Variable -> Target (Risk of Profile of a Customer)

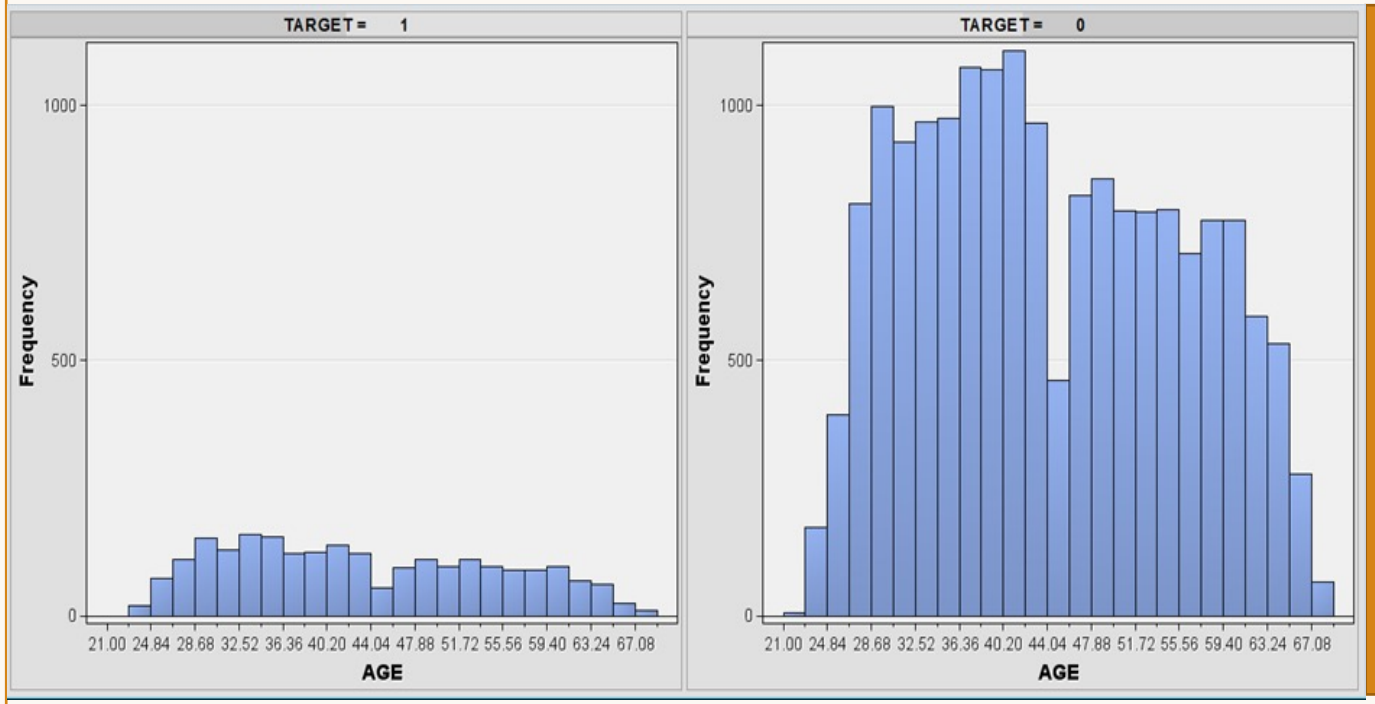
Total Annual Income VS Risk of Profile



- Both Target Class are Right Skewed Histogram
- Highest Income Range is from 104k to 181k. Most of the customer in both class is having salary lower than average.
- Majority of customer with high risk profile (target = 1) have lower annual income compared to low risk profile applicant.

NG SIN YU S211068

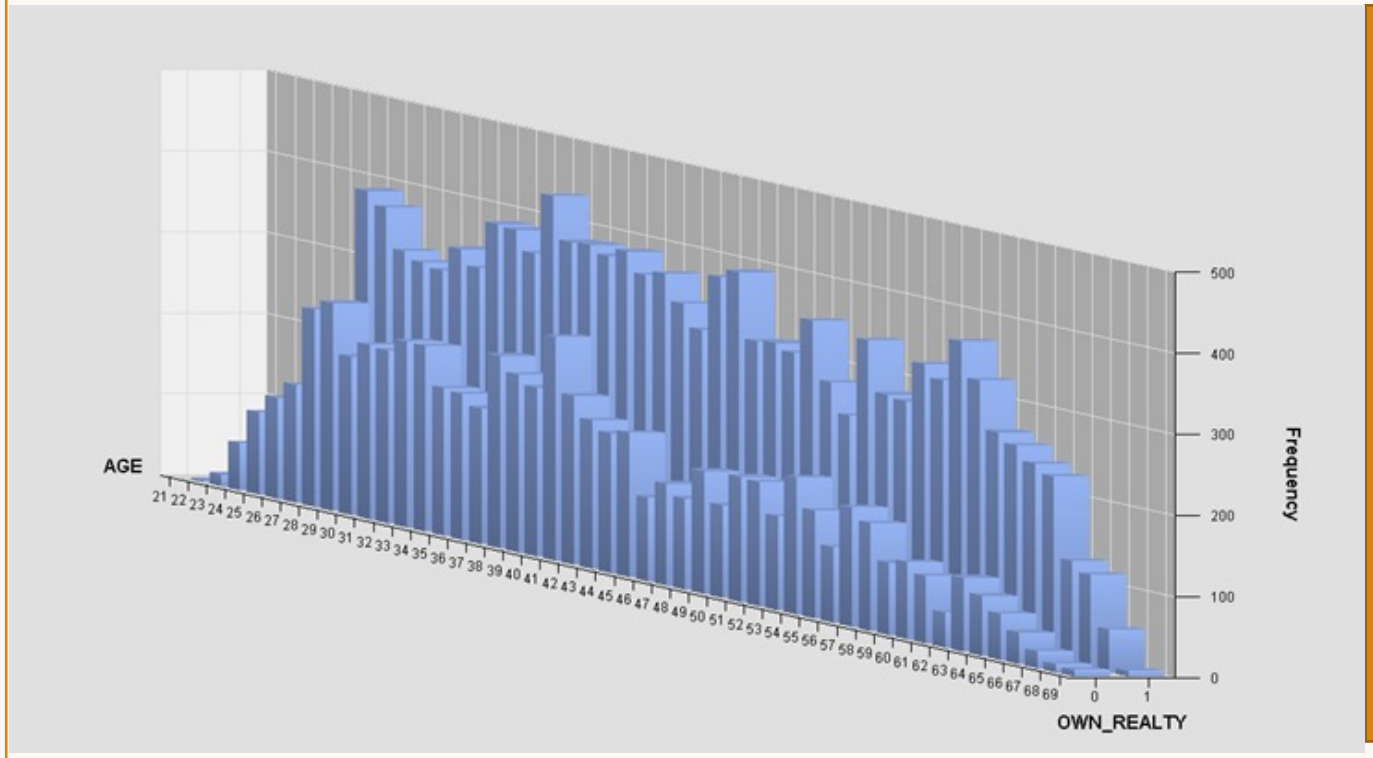
Age VS Risk of Profile



- Most of the applicant that is above age 36 is having a low risk profile, this is the age which most of the people reach financial stability .
- However, as they grew old, the credit risk increase as stated in risk modelling research.

NG SIN YU S211068

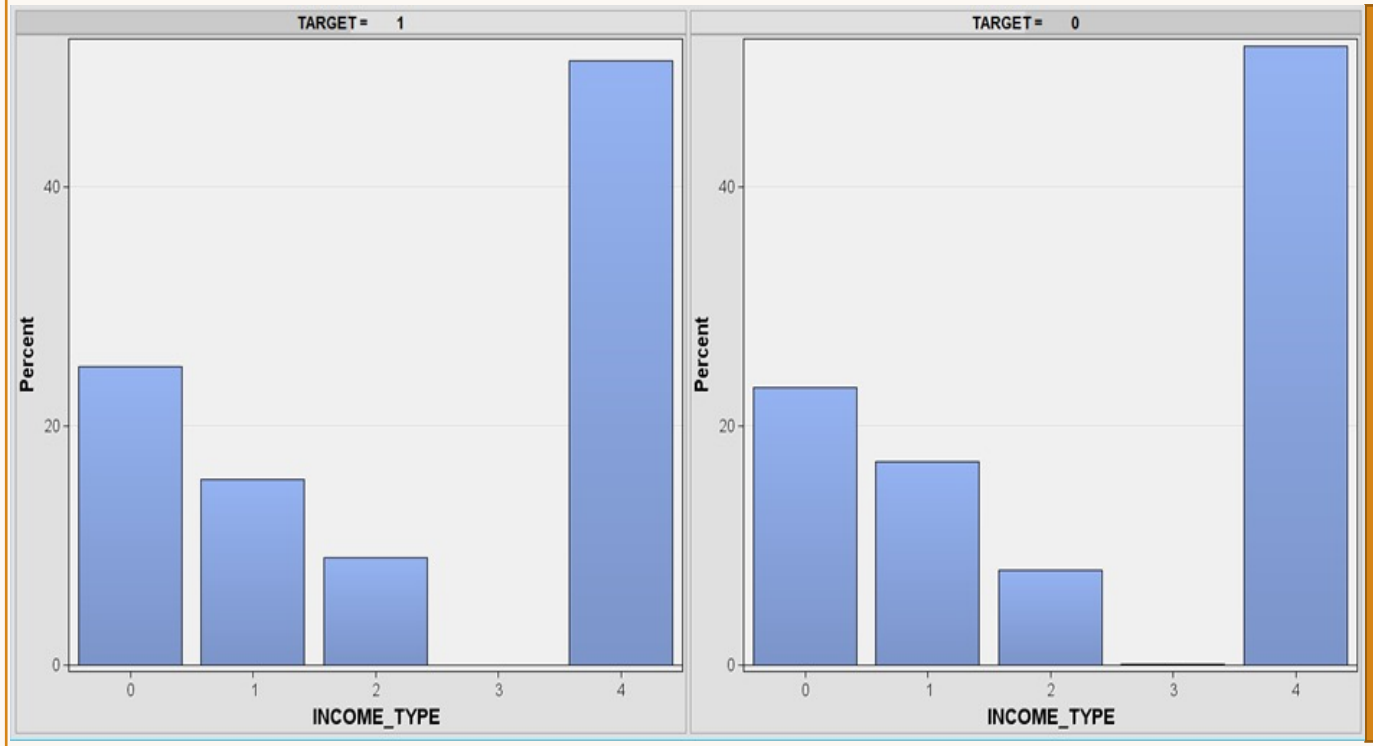
Age VS Own Realty



- Compared to early 20 or 30's years old, the proportion of applicants that owns a property spike in the age of late 30's to 40's.
- This might be because after years of working, most of the people started to get married and have their own families.

NG SIN YU S211068

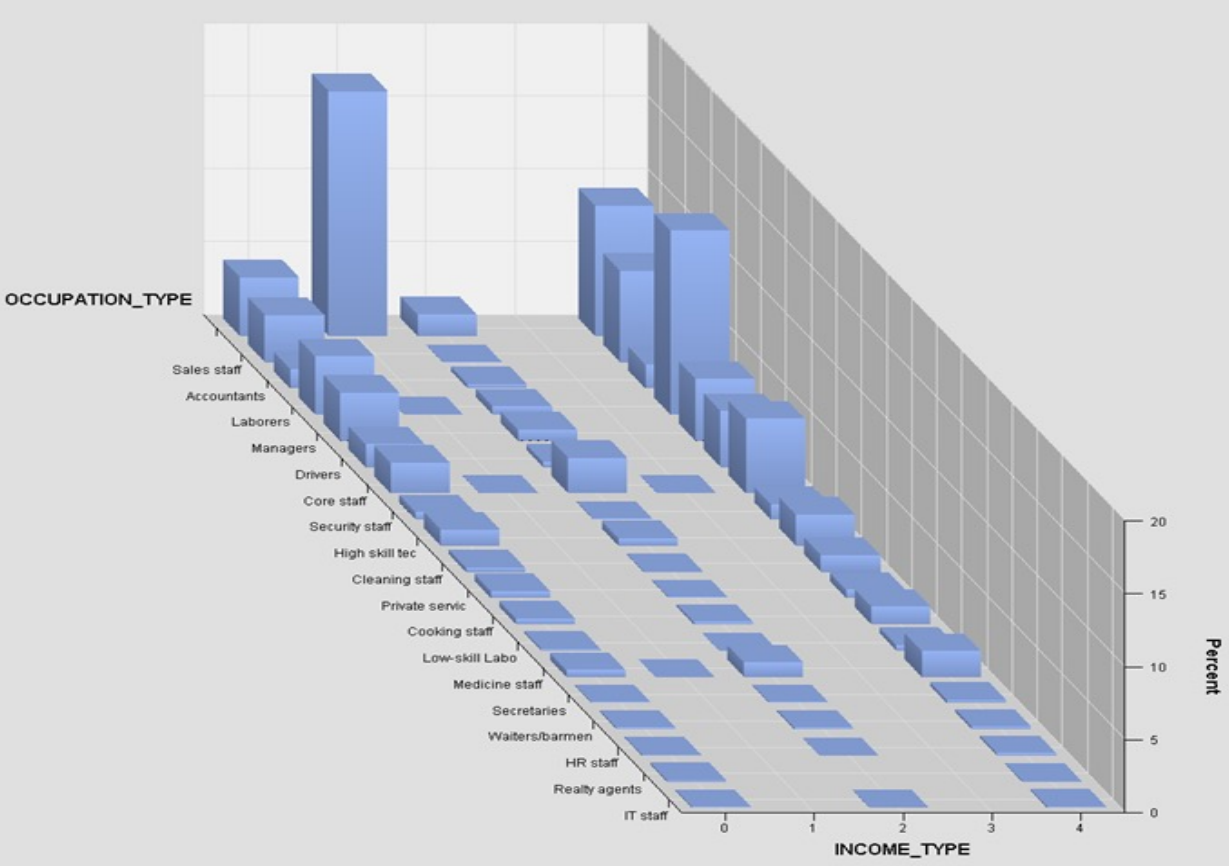
Income Type VS Risk of Profile



- Income type has been coded into 0, 1, 2, 3 and 4 which represent commercial, pensioner, state servant, student and working.
- Most likely a student will not apply for a credit card as they're potentially not financially independent.

NG SIN YU S211068

Income Type VS Occupation Type



- As mentioned previously, we grouped occupation type with null value as unknown occupation type.
- However, from income type, we can notice that those without occupation type are mostly come from the group of people that had already retired, thus they don't have ongoing job.

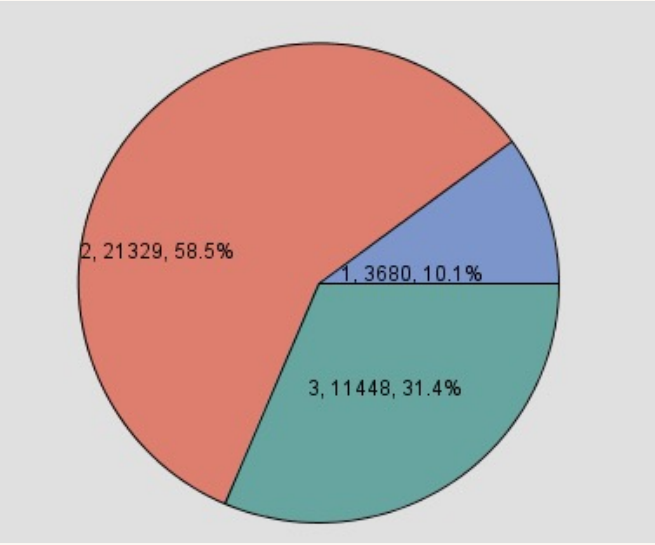
NG SIN YU S211068

Multivariate Analysis

SYALIN DANIA S2018106

Cluster Analysis

Variables : AMT_INCOME_TOTAL, OWN_REALTY, YEARS_EMPLOYED



Segment ID	Root-Mean Square Standard Deviation	Nearest Cluster
1	0.973861	2
2	0.551933	3
3	0.599322	2

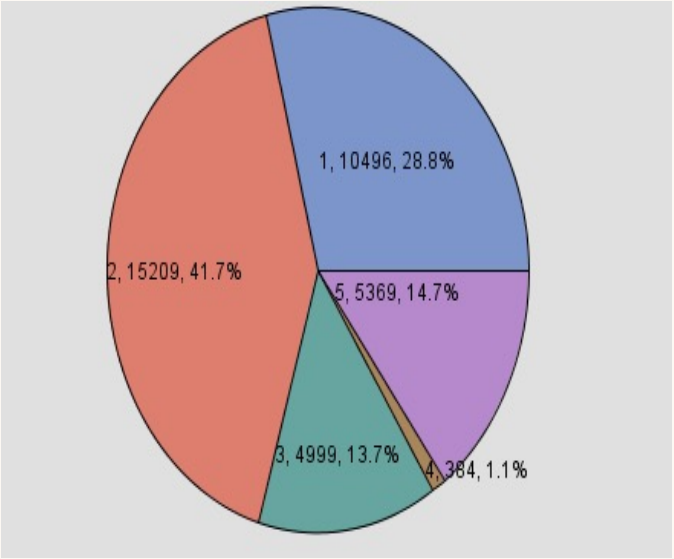
•The Root-Mean-Square-Deviation (RMSD) from our analysis shows that it is non-negative, and a value of 0 would indicate a perfect fit to the data. In general, a lower RMSD is better than a higher one. In this case, Cluster 2 is the best cluster followed by Cluster 3 and Cluster 1. The statistics also indicates that Cluster 2 and Cluster 3 are close to each other.

Multivariate Analysis

SYALIN DANIA S2018106

Cluster Analysis

Variables : ACCOUNT_LENGTH, AGE, AMT_INCOME_TOTAL, YEARS_EMPLOYED



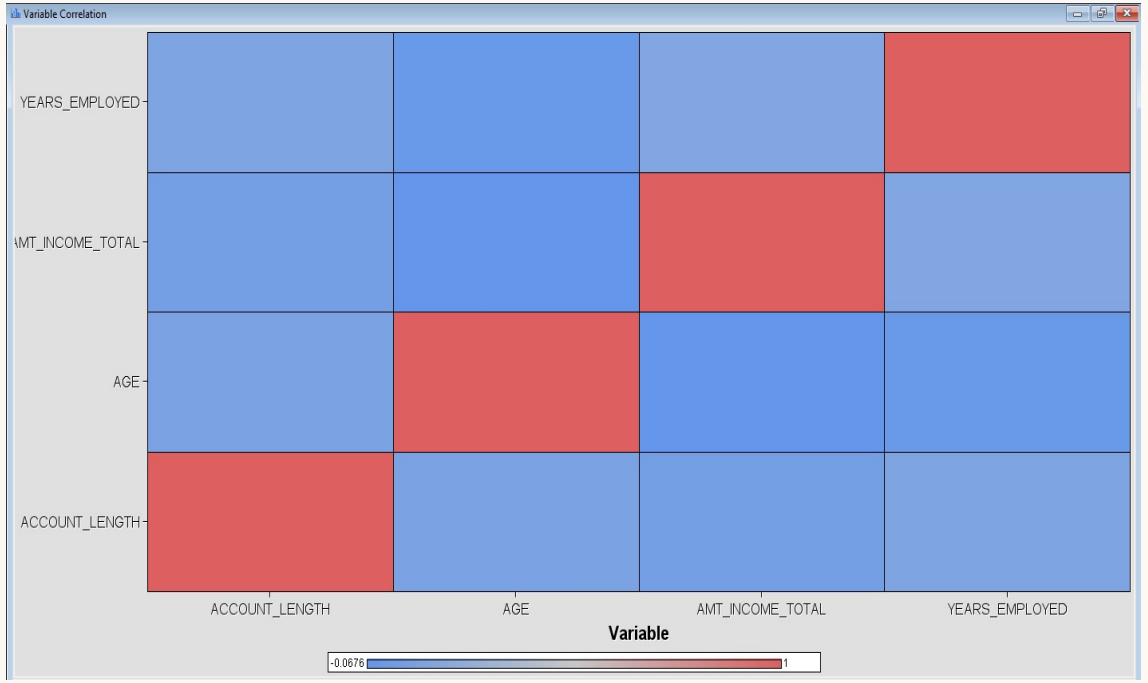
Segment ID	Root-Mean Square Standard Deviation	Nearest Cluster
1	0.673124	2
2	0.642216	3
3	0.771291	2
4	1.182269	3
5	0.842269	2

•The Root-Mean-Square-Deviation (RMSD) from our analysis, as compared to previous observation, the values for RMSD in these clusters are higher as compared, with Cluster 2 with the lowest value of RMSD, followed by Cluster 4 with highest value of RMSD.

Multivariate Analysis

Correlation Matrix

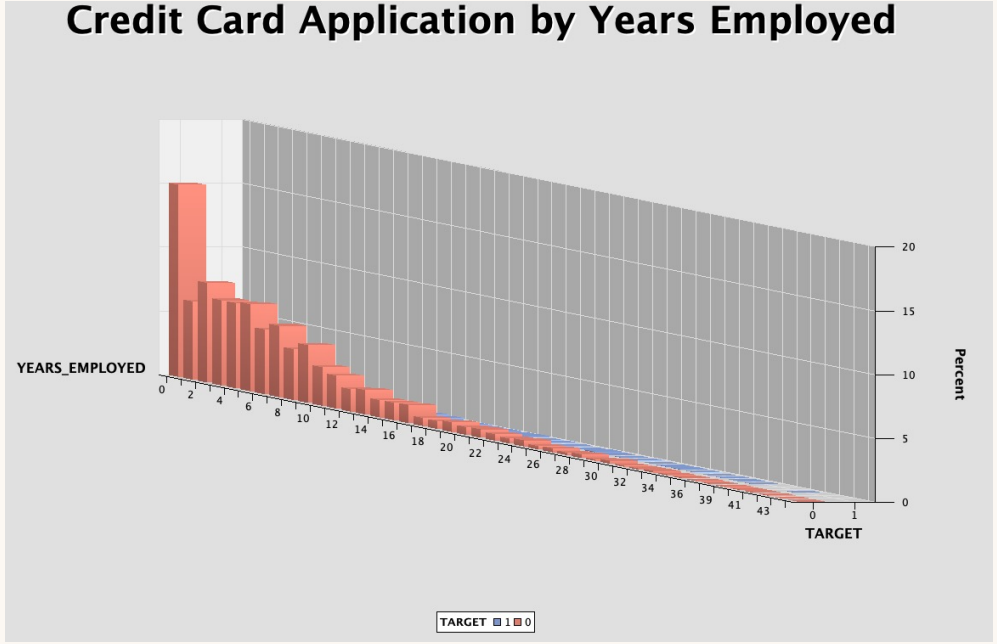
From the correlation matrix here, it summarised how variables like ACCOUNT_LENGTH, AGE, AMT_INCOME_TOTAL, and YEARS_EMPLOYED are correlated with each other. The red colour indicated that the two variables are highly correlated while the lighter colour means that the variables are less correlated.



SYALIN DANIA S2018106

Pattern Discovery with SAS E-miner

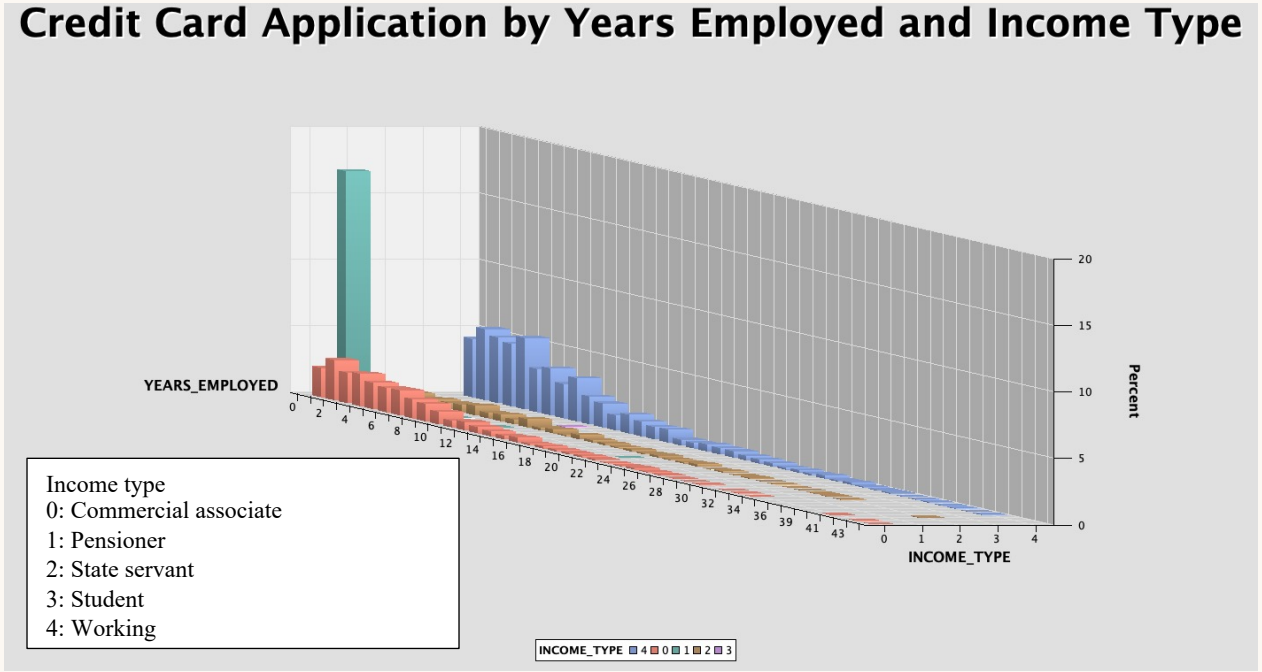
- Credit Card Application by Number of Years Employed



There is a huge spike of 0 years employment for low-risk applicants which accounted for 15%. Does this scenario infer that most of the applicants are recent graduates who entering the workforce?

LEE YING QIU 17108552

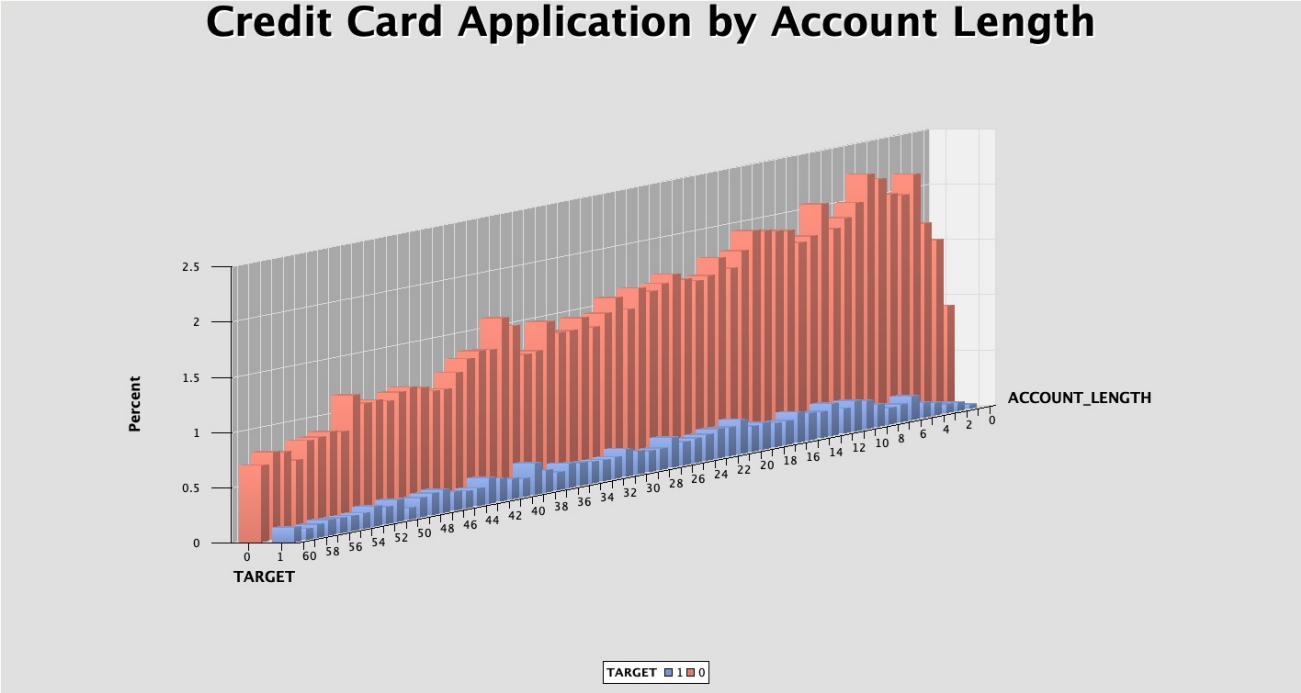
Pattern Discovery with SAS E-miner



When INCOME_TYPE is included, the source of huge spike comes from income category where the applicants receive income in pension form. They can use retirement savings as their source of annual income to apply for credit card.

LEE YING QIU 17108552

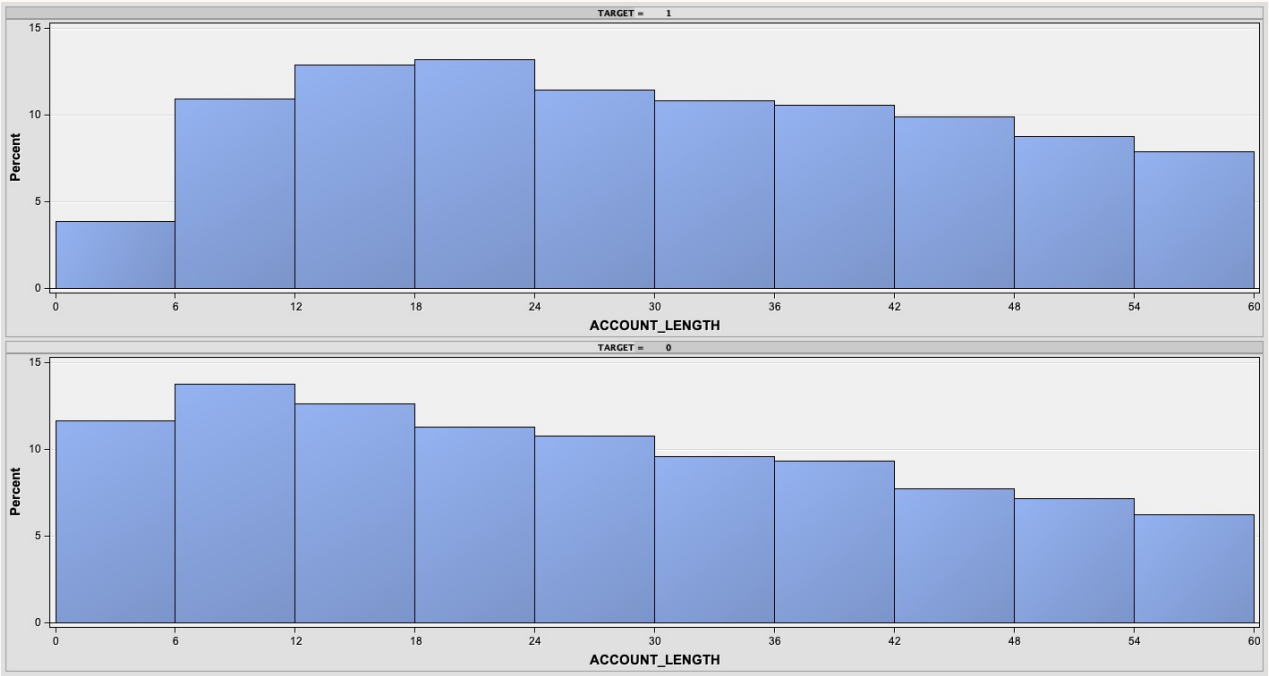
Pattern Discovery with SAS E-miner



A long-standing account with good track record (i.e., on time payments) will get approval easier than the account with miss or late payments. Most of the accounts are newly opened and not more than 2 years or 24 months.

LEE YING QIU 17108552

Pattern Discovery with SAS E-miner



For low-risk applicants, about 50% of their accounts are not more than 2 years. A possible reason for this phenomenon is that many of the applicants are young working adults who demonstrate impulse buying behavior and the desire for social status. In fact, they are the prime target of credit card issuer.

LEE YING QIU 17108552

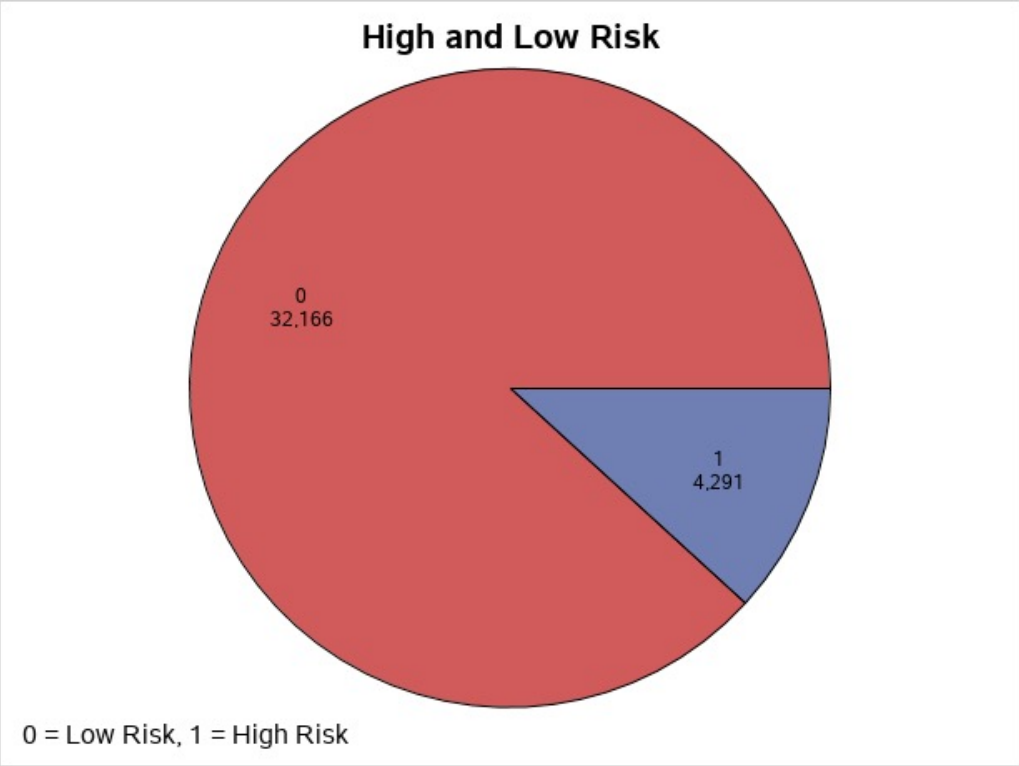
Summary

Variable Name	Incomplete	Noisy	Inconsistent	Intentional
ACCOUNT_LENGTH				
AGE				
AMT_INCOME_TOTAL		YES		
CNT_CHILDREN		YES		YES
CNT_FAM_MEMBERS		YES		
EDUCATION_TYPE				
EMAIL				
FAMILY_STATUS				
GENDER				
HOUSING_TYPE				
ID				
INCOME_TYPE				
OCCUPATION_TYPE	YES			
OWN_CAR				
OWN_MOBIL				YES
OWN_REALTY				
PHONE				
TARGET				
WORK_PHONE				
YEARS_EMPLOYED		YES		

a) Both outliers from CNT_CHILDREN and CNT_FAM_MEMBERS will be removed.
b) OWN_MOBIL will be dropped and excluded from the analysis.
c) OCCUPATION_TYPE revealed a large numbers of missing values and these values will be replaced with constant.

LEE YING QIU 17108552

Summary



TARGET data is highly imbalanced in which low risk applicants accounted for almost 90% in the sample. Therefore, it is required to perform the oversampling to the rare event by putting a higher proportion of rare event observations (high risk applicants).

LEE YING QIU 17108552

Methodology (Cont'd)

- SAS 'SEMMA'
- Focus on 3rd, 4th and 5th letters 'M', 'M' and 'A' for this Group Assignment

- MODIFY – Perform oversampling of rare event sample and data cleansing to remove the noisy data
- MODEL – Perform predictive model training to identify if a customer is likely to be a high-risk profile or low-risk profile
- ASSESS– Assess and determine the best model to be used

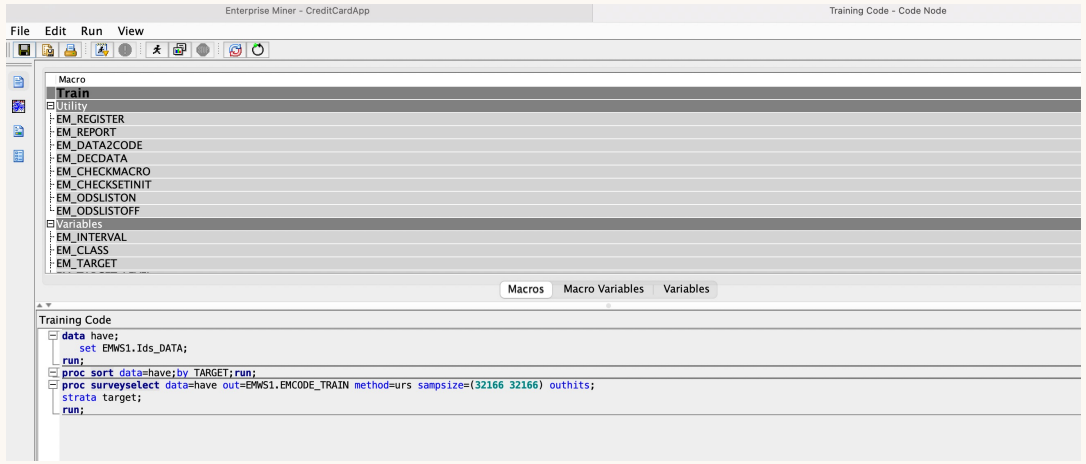
LEE YING QIU 17108552

Modification (Diagram)

Modification

OVERSAMPLING

- Oversampling involves choosing random samples from the minority class with replacement and adding additional copies of this example to the training set.
- We use the proc surveyselect in SAS Code to oversample high risk target and make it the same amount to low-risk target in a way to make our sample balanced.

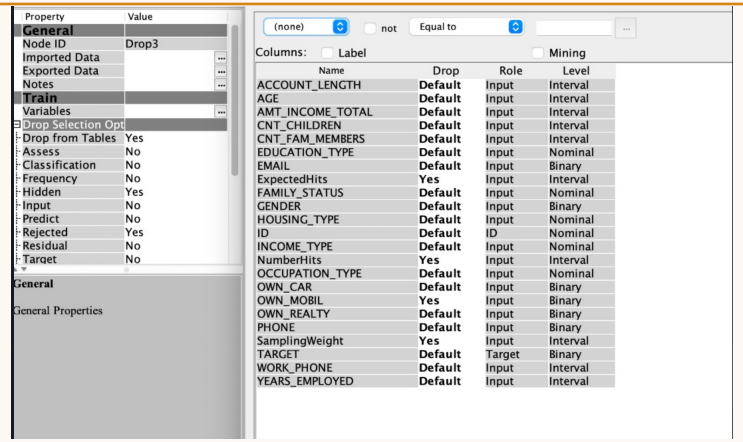


Port	Table	Role	Data Exists
TRAIN	EMWS1.EMCODE_TRAIN	Train	Yes
VALIDATE	EMWS1.EMCODE_VALIDATE	Validate	No
TEST	EMWS1.EMCODE_TEST	Test	No
SCORE	EMWS1.EMCODE_SCORE	Score	No
TRANSACTION	EMWS1.EMCODE_TRANSAC...	Transaction	No

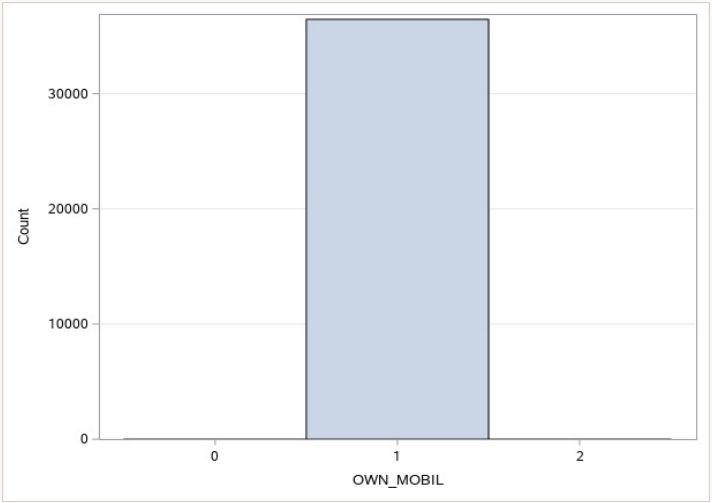
Modification

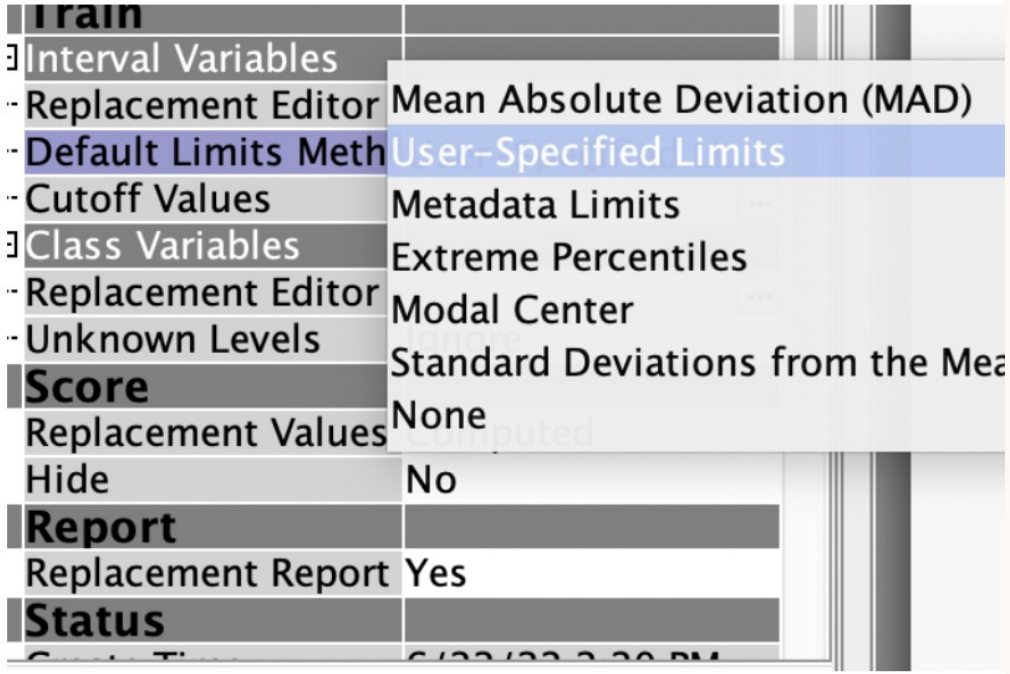
DROP

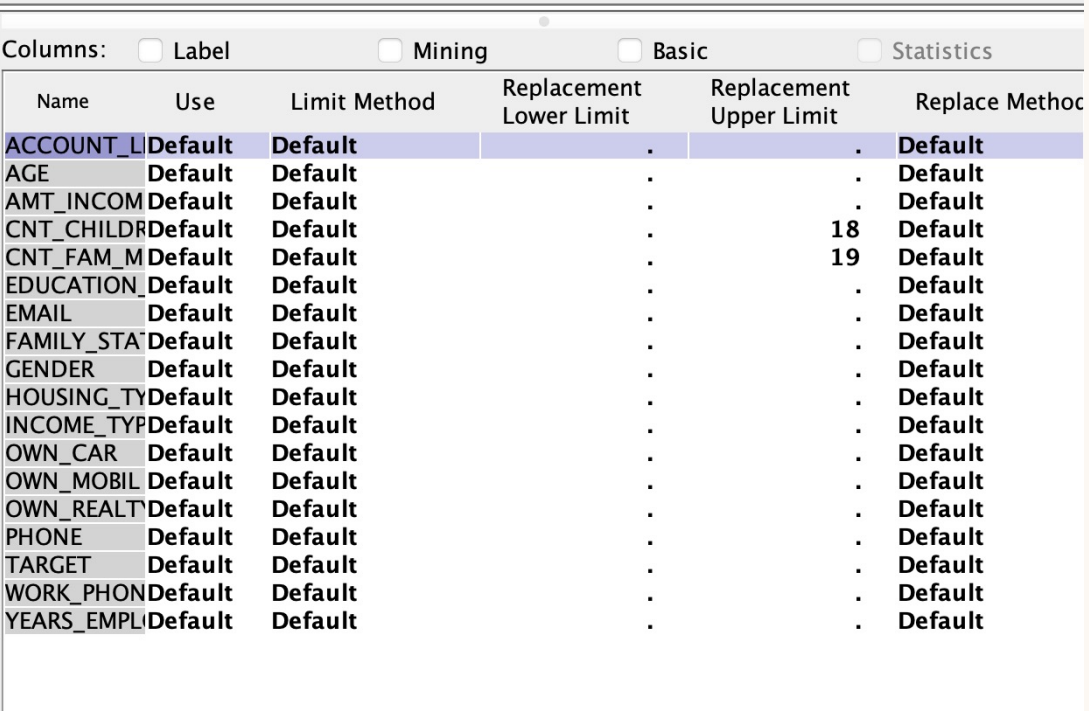
- In this node, we want to drop selected variables from our dataset, which OWN_MOBIL column because it is intentional, where all the values in OWN_MOBIL is zero, so we want to remove it.



- OWN_MOBIL is constant variable and to be excluded from further analysis







Columns: Label Mining Basic Statistics

Name	Use	Limit Method	Replacement Lower Limit	Replacement Upper Limit	Replace Method
ACCOUNT_L	Default	Default	.	.	Default
AGE	Default	Default	.	.	Default
AMT_INCOM	Default	Default	.	.	Default
CNT_CHILDR	Default	Default	.	18	Default
CNT_FAM_M	Default	Default	.	19	Default
EDUCATION_	Default	Default	.	.	Default
EMAIL	Default	Default	.	.	Default
FAMILY_STA	Default	Default	.	.	Default
GENDER	Default	Default	.	.	Default
HOUSING_TY	Default	Default	.	.	Default
INCOME_TYP	Default	Default	.	.	Default
OWN_CAR	Default	Default	.	.	Default
OWN_MOBIL	Default	Default	.	.	Default
OWN_REALT	Default	Default	.	.	Default
PHONE	Default	Default	.	.	Default
TARGET	Default	Default	.	.	Default
WORK_PHON	Default	Default	.	.	Default
YEARS_EMPL	Default	Default	.	.	Default

Modification

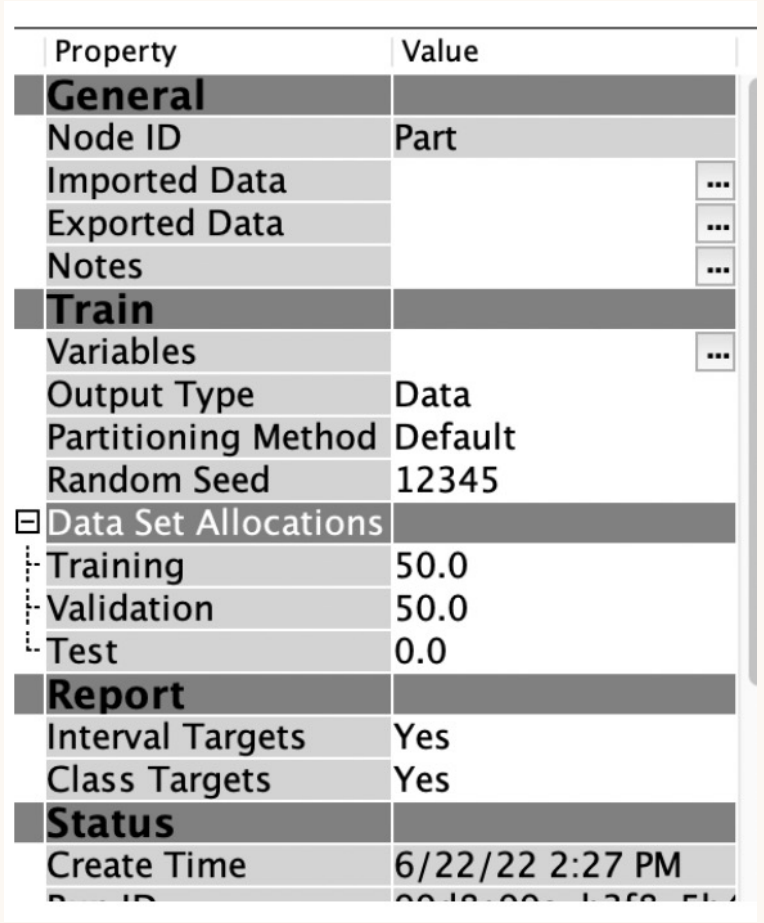
REPLACEMENT

- In the replacement node, we set the Default limits Method to User-specified Limits, and we set the limit of 18 to CNT_CHILDREN and 19 to CNT_FAM_MEMBERS.
- The purpose of this is , to remove the outlier that we have in our dataset which is 19 for CNT_CHILDREN and 20 for CNT_FAM_MEMBERS.

Modification

DATA PARTITION

- In data mining, data partitioning is the division of all available data into two or three non-overlapping sets: the training set, the validation set, and the test set. If the data set is very large, only a portion of it is usually chosen for partitioning.
- As for our dataset we split into two sets which are training set and validation set with ratio of 50/50.



Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	6/22/22 2:27 PM

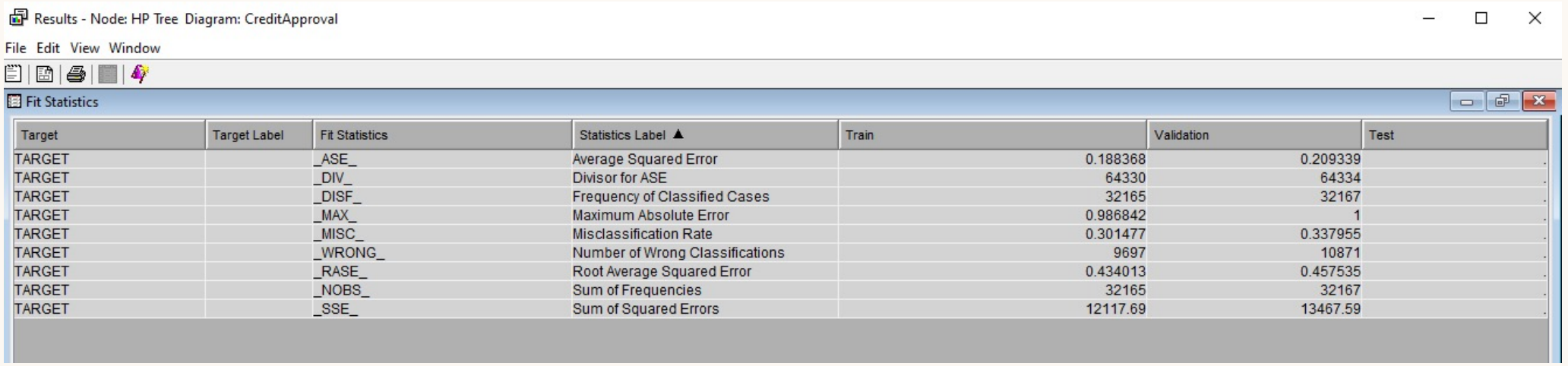
Modelling

- Predictive modelling is the process of using the known result to create, process and validate a model that can be used to forecast future outcomes.
- We will train 4 predictive models to predict if a customer is likely to be a high-risk profile or low-risk profile
- All the four model is trained using HP node – High performance node as our data set is huge, thus HP node enable use to derive more accurate insights, and train model more robustly.

NG SIN YU S2111068

HP Decision Tree

- Train decision tree with 10 –fold cross validation
- Accuracy: 67 %
- Misclassification Rate: 33 %



Target	Target Label	Fit Statistics	Statistics Label ▲	Train	Validation	Test
TARGET		_ASE_	Average Squared Error	0.188368	0.209339	.
TARGET		_DIV_	Divisor for ASE	64330	64334	.
TARGET		_DISF_	Frequency of Classified Cases	32165	32167	.
TARGET		_MAX_	Maximum Absolute Error	0.986842	1	.
TARGET		_MISC_	Misclassification Rate	0.301477	0.337955	.
TARGET		_WRONG_	Number of Wrong Classifications	9697	10871	.
TARGET		_RASE_	Root Average Squared Error	0.434013	0.457535	.
TARGET		_NOBS_	Sum of Frequencies	32165	32167	.
TARGET		_SSE_	Sum of Squared Errors	12117.69	13467.59	.

HP Decision Tree

Variable Importance

Variable Name	Label	Number of Splitting Rules	Sum of Square Errors	Importance	Validation Sum of Square Errors	Validation Importance
ACCOUNT_LENGTH		54	27.35406	1	24.1961	1
REP_OCCUPATION_TYPE	Replacement: OCCUPATION...	50	24.83139	0.907777	20.04276	0.828347
AGE		63	23.71375	0.866919	19.2117	0.794
YEARS_EMPLOYED		54	20.83719	0.761759	16.48257	0.681208
AMT_INCOME_TOTAL		46	19.90914	0.727831	15.88706	0.656596
FAMILY_STATUS		21	13.95429	0.510136	11.60441	0.479598
INCOME_TYPE		16	12.44869	0.455095	12.06768	0.498745
REP_CNT_CHILDREN	Replacement: CNT_CHILDR...	15	12.28127	0.448974	11.52437	0.47629
EDUCATION_TYPE		15	10.88926	0.398086	8.975743	0.370958
OWN_REALTY		10	9.905296	0.362114	9.516999	0.393328
PHONE		10	9.891501	0.36161	4.910588	0.20295
REP_CNT_FAM_MEMBERS	Replacement: CNT_FAM_ME...	12	8.889465	0.324978	6.700135	0.27691
OWN_CAR		5	8.862958	0.324009	8.286487	0.342472
HOUSING_TYPE		7	8.743655	0.319647	6.402373	0.264603
GENDER		6	8.252425	0.301689	6.776616	0.280071
WORK_PHONE		8	7.887591	0.288352	5.791795	0.239369
EMAIL		4	5.260016	0.192294	4.752157	0.196402

Important Variables Involves in Training Decision Tree

NG SIN YU S2111068

HP Random Forest

- Set model with maximum tree of 100: Optimum tree among misclassification rate and average square error
- Accuracy is 75 %, Misclassification = 25 %

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TARGET		_ASE_	Average Squared Error	0.203807	0.210619	.
TARGET		_DIV_	Divisor for ASE	64330	64334	.
TARGET		_MAX_	Maximum Absolute Error	0.785927	0.785927	.
TARGET		_NOBS_	Sum of Frequencies	32165	32167	.
TARGET		_RASE_	Root Average Squared Error	0.45145	0.458932	.
TARGET		_SSE_	Sum of Squared Errors	13110.92	13549.95	.
TARGET		_DISF_	Frequency of Classified Cases	32165	32167	.
TARGET		_MISC_	Misclassification Rate	0.211441	0.252433	.
TARGET		_WRONG_	Number of Wrong Classifications	6801	8120	.

NG SIN YU S2111068

HP Random Forest

Variable Importance

Variable Name	Number of Splitting Rules	Train: Gini Reduction ▼	Train: Margin Reduction	OOB: Gini Reduction	OOB: Margin Reduction	Valid: Gini Reduction
ACCOUNT_LENGTH	2003	0.013199	0.026399	0.010143	0.023013	0.010357
REP_OCCUPATION_TYPE	1954	0.008181	0.016362	0.003731	0.012501	0.004010
AGE	2128	0.006273	0.012545	0.003334	0.009333	0.003589
AMT_INCOME_TOTAL	1964	0.005535	0.011069	0.003076	0.008354	0.003082
YEARS_EMPLOYED	1908	0.005096	0.010192	0.002886	0.007734	0.002812
FAMILY_STATUS	1023	0.002718	0.005437	0.001256	0.004014	0.001443
INCOME_TYPE	1018	0.002632	0.005265	0.001345	0.003955	0.001538
EDUCATION_TYPE	948	0.002528	0.005057	0.001302	0.003814	0.001276
REP_CNT_FAM_MEMBERS	1175	0.002423	0.004847	0.001261	0.003568	0.001403
REP_CNT_CHILDREN	1014	0.002023	0.004046	0.001202	0.003139	0.001116
HOUSING_TYPE	802	0.001967	0.003935	0.000882	0.002789	0.000900
OWN_CAR	600	0.001673	0.003345	0.001092	0.002653	0.001189
OWN_REALTY	483	0.001452	0.002905	0.000987	0.002346	0.000704
PHONE	613	0.001334	0.002667	0.000712	0.001953	0.000656
WORK_PHONE	583	0.001184	0.002368	0.000514	0.001619	0.000553
GENDER	566	0.001117	0.002234	0.000617	0.001631	0.000626
EMAIL	415	0.000881	0.001763	0.000447	0.001267	0.000601

- All the interval variable has been automatically normalized and scale to 0-1 by log transformation in HP model

Important Variables Involves in Training Random Forest

NG SIN YU S2111068

HP SVM

Output Name	Formula	Power	Role	Output Level	Input Name	Label	Input Level
LG10_ACCOUN...	log10(ACCOUN...	0	INPUT	INTERVAL	ACCOUNT_LEN...	Transformed AC...	INTERVAL
LG10_AGE	log10(AGE+ 1)	0	INPUT	INTERVAL	AGE	Transformed AGE	INTERVAL
LG10_AMT_INC...	log10(AMT_INC...	0	INPUT	INTERVAL	AMT_INCOME_T...	Transformed AM...	INTERVAL
LG10_REP_CN...	log10(REP_CNT...	0	INPUT	INTERVAL	REP_CNT_CHIL...	Transformed Re...	INTERVAL
LG10_REP_CN...	log10(REP_CNT...	0	INPUT	INTERVAL	REP_CNT_FAM...	Transformed Re...	INTERVAL
LG10_YEARS_E...	log10(YEARS_E...	0	INPUT	INTERVAL	YEARS_EMPLO...	Transformed YE...	INTERVAL

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
TARGET		_ASE_	Average Squared Error	0.231133	0.233068
TARGET		_DIV_	Divisor for ASE	64330	64334
TARGET		_MAX_	Maximum Absolute Err...	0.86189	0.86189
TARGET		_NOBS_	Sum of Frequencies	32165	32167
TARGET		_RASE_	Root Average Squared...	0.480763	0.482771
TARGET		_SSE_	Sum of Squared Errors	14868.78	14994.21
TARGET		_DISF_	Frequency of Classifie...	32165	32167
TARGET		_MISC_	Misclassification Rate	0.343603	0.357696
TARGET		_WRONG_	Number of Wrong Cla...	11052	11506

- Performs feature scaling for all the interval variables
- Misclassification Rate is 35 %
- Accuracy of Model is 65 %

NG SIN YU 2111068

HP Neural Network

- Apply feature scaling -> log transformation for interval variable
- Apply feature selection -> select important variable based on sequential selection
- Years of Employed -> Rejected by variance Explained
- Number of Children -> Rejected by variance Explained

NG SIN YU S2111068

HP Neural Network

- Two Layers With Direct – 15 number of hidden neurons
- Misclassification and root average squared error is high

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TARGET		_ASE_	Average Squared ...	0.241415	0.24196	.
TARGET		_DIV_	Divisor for ASE	64330	64334	.
TARGET		_MAX_	Maximum Absolut...	0.817889	0.81815	.
TARGET		_NOBS_	Sum of Frequenci...	32165	32167	.
TARGET		_RASE_	Root Average Sq...	0.49134	0.491894	.
TARGET		_SSE_	Sum of Squared ...	15530.24	15566.24	.
TARGET		_DISF_	Frequency of Cla...	32165	32167	.
TARGET		_MISC_	Misclassification ...	0.43482	0.434234	.
TARGET		_WRONG_	Number of Wrong...	13986	13968	.

NG SIN YU 2111068

Assess

- Model Comparison Node Fit Statistics

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Average Squared Error	Train: Divisor for ASE	Train: Maximum Absolute Error	Train: Sum of Frequencies	Train: Root Average Squared Error	Train: Sum of Squared Errors	Train: Frequency of Classified Cases	Train: Misclassification Rate
Y	HPDMForest	HPDMForest	HP Forest	TARGET		0.252433	0.203807	64330	0.785927	32165	0.45145	13110.92	32165	0.211441
	HPTree	HPTree	HP Tree	TARGET		0.337955	0.188368	64330	0.986842	32165	0.434013	12117.69	32165	0.301477
	HPSVM	HPSVM	HP SVM	TARGET		0.357696	0.231133	64330	0.86189	32165	0.480763	14868.78	32165	0.343603
	HPNNA	HPNNA	HP Neural	TARGET		0.434234	0.241415	64330	0.817889	32165	0.49134	15530.24	32165	0.43482

Assess

- Model Comparison Node Fit Statistics from Output window

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	HPDMForest	HP Forest	0.25243	0.20381	0.21144	0.21062
	HPTree	HP Tree	0.33796	0.18837	0.30148	0.20934
	HPSVM	HP SVM	0.35770	0.23113	0.34360	0.23307
	HPNNA	HP Neural	0.43423	0.24142	0.43482	0.24196

Assess

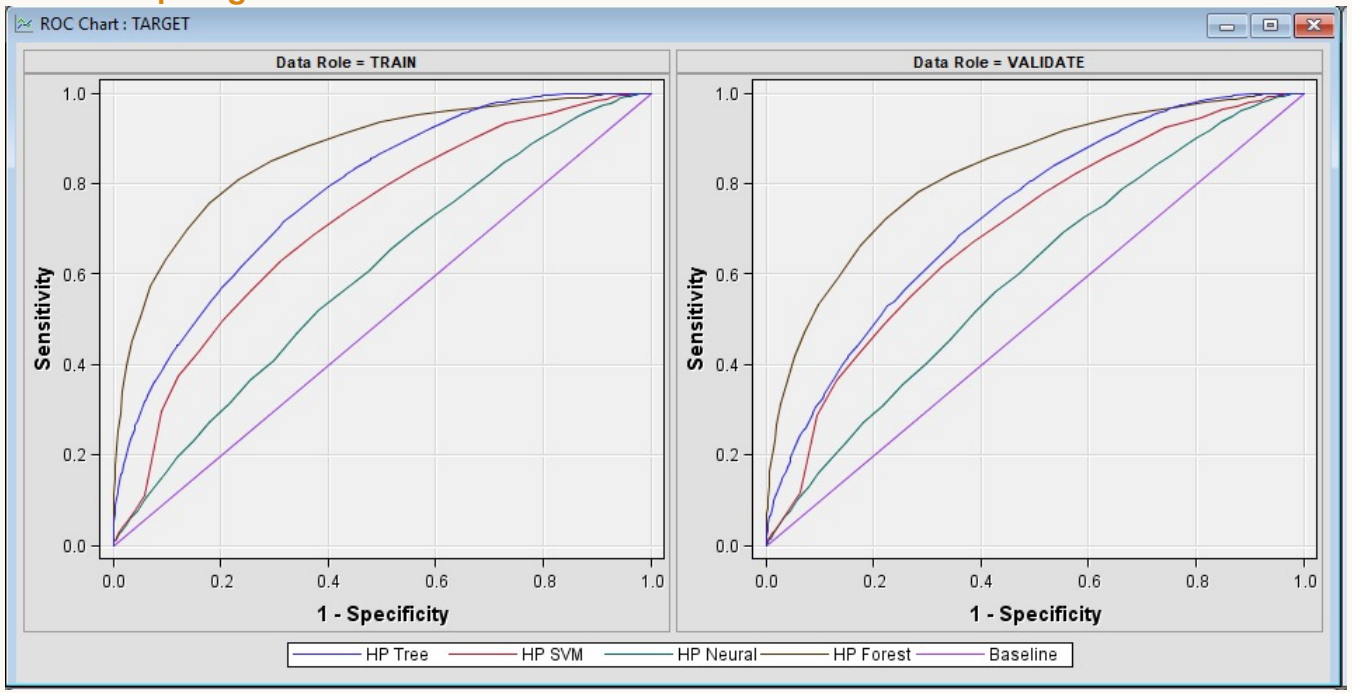
- Performance Matrix

Model	Accuracy	Precision	Recall	F-score	ROC area
Random Forest	75%	0.7326	0.7797	0.7554	0.845
Decision Tree	67%	0.6498	0.7027	0.6752	0.711
SVM	65%	0.6340	0.6730	0.6529	0.589
Neural Networks	57%	0.5613	0.6018	0.5808	0.59

CLEMENT LEE S2128268

Assess

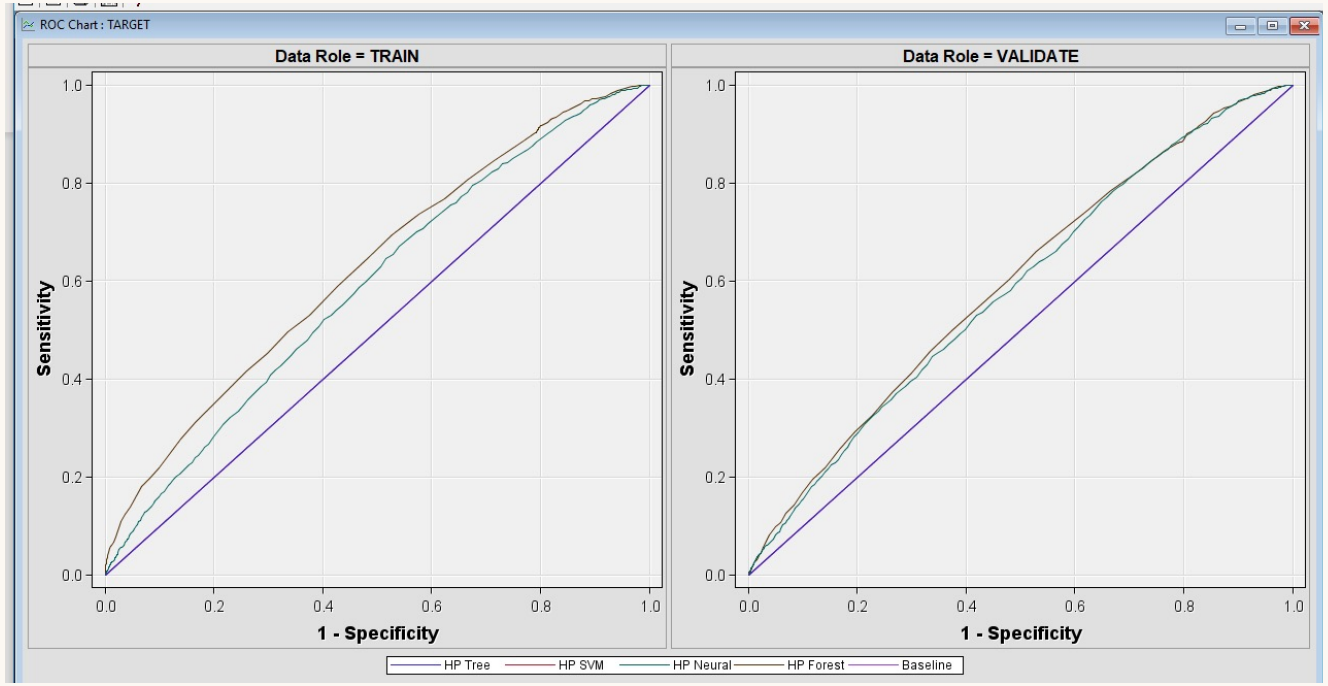
- ROC curves with oversampling



CLEMENT LEE S2128268

Assess

- ROC curve without Oversampling

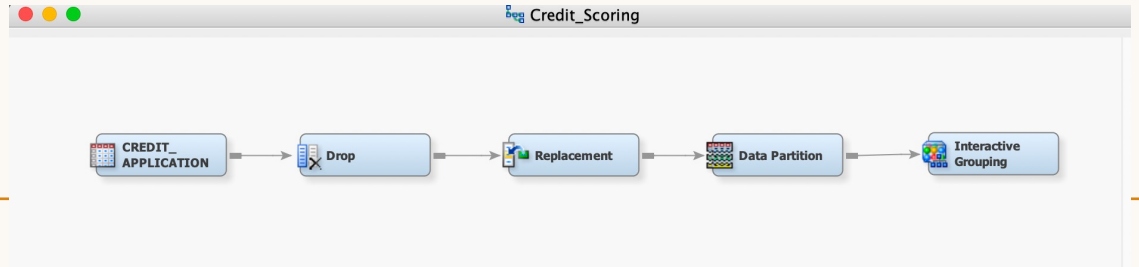


CLEMENT LEE S2128268

Credit Scoring in SAS E-Miner

LEE YING QIU 17108552

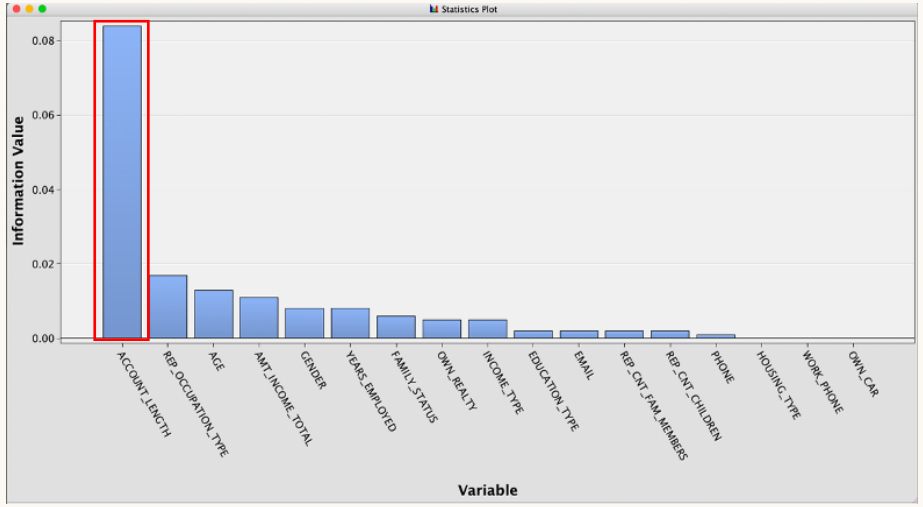
- Explored the credit scoring module in SAS E-Miner to predict the potential risk of applicants by building the scorecard model for the attributes (characteristics)
- Prior to the scorecard modelling, "Interactive Grouping" node is added after "Data Partition" node to perform grouping of attributes and initial screening of attributes
- Assess the strength of attributes individually using measurements such as Weight of Evidence Measure (WoE) and Information Value (IV)



Credit Scoring in SAS E-Miner

Output Variables

Variable	Information Value	Level for Interactive	Information Value Ordering	Gini Statistic	Calculated Role	New Role	Pre-Defined Grouping	Level
ACCOUNT_LENGTH	0.084	INTERVAL	1	12.019	Input	Default		INTERVAL
REP_OCCUPATION_TYPE	0.017	NOMINAL	2	7.118	Input	Default		NOMINAL
AGE	0.013	INTERVAL	3	5.895	Input	Default		INTERVAL
AMT_INCOME_TOTAL	0.011	INTERVAL	4	5.431	Input	Default		INTERVAL
GENDER	0.008	BINARY	5	4.326	Input	Default		BINARY
YEARS_EMPLOYED	0.008	INTERVAL	6	4.339	Input	Default		INTERVAL
FAMILY_STATUS	0.006	NOMINAL	7	3.51	Input	Default		NOMINAL
OWN_REALTY	0.005	BINARY	8	3.418	Input	Default		BINARY
INCOME_TYPE	0.005	NOMINAL	9	3.619	Input	Default		NOMINAL
EDUCATION_TYPE	0.002	NOMINAL	10	1.83	Input	Default		NOMINAL
EMAIL	0.002	BINARY	11	1.336	Input	Default		BINARY
REP_CNT_FAM_MEMBERS	0.002	INTERVAL	12	1.991	Input	Default		INTERVAL
REP_CNT_CHILDREN	0.002	INTERVAL	13	1.675	Input	Default		INTERVAL
PHONE	0.001	BINARY	14	1.158	Input	Default		BINARY
HOUSING_TYPE	0	NOMINAL	15	0.652	Input	Default		NOMINAL
WORK_PHONE	0	INTERVAL	16	0.567	Input	Default		INTERVAL
OWN_CAR	0	BINARY	17	0.487	Input	Default		BINARY

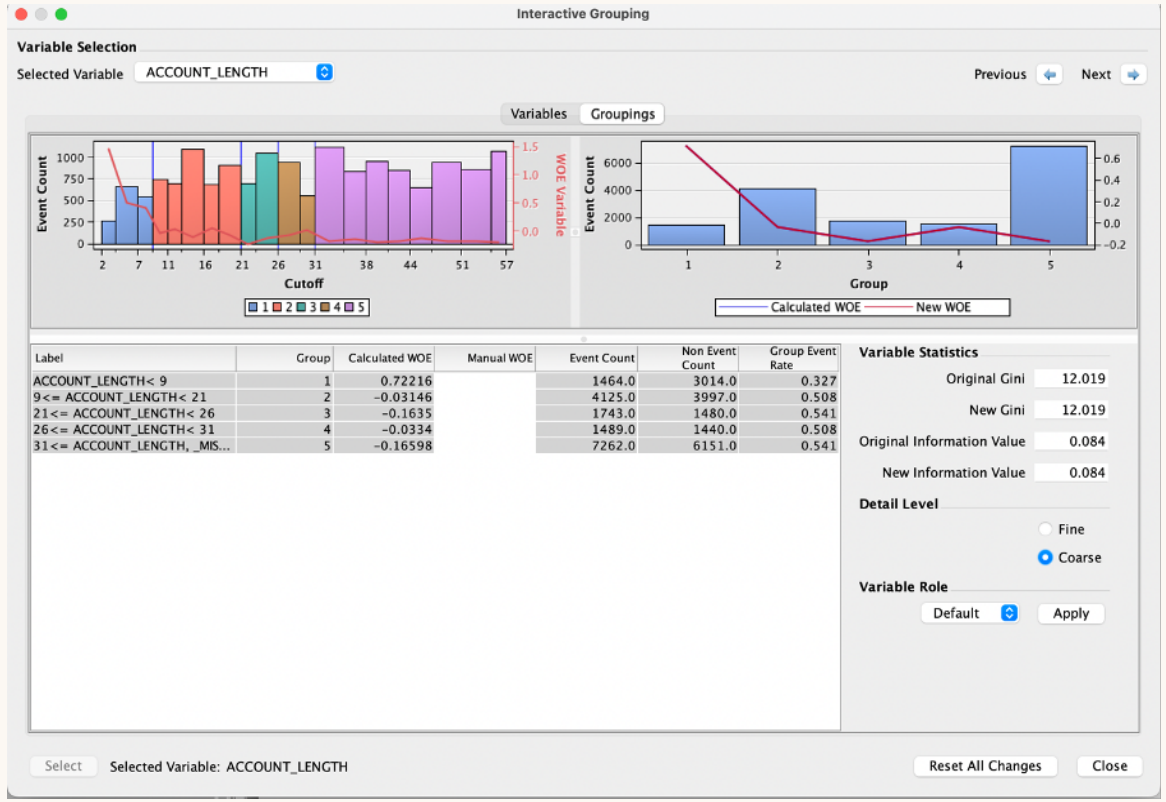


Information Value (IV)

Based on the output table and plot, ACCOUNT_LENGTH appeared to be the highest IV (0.084) and the most important candidate input to build scorecard

LEE YING QIU 17108552

Credit Scoring in SAS E-Miner

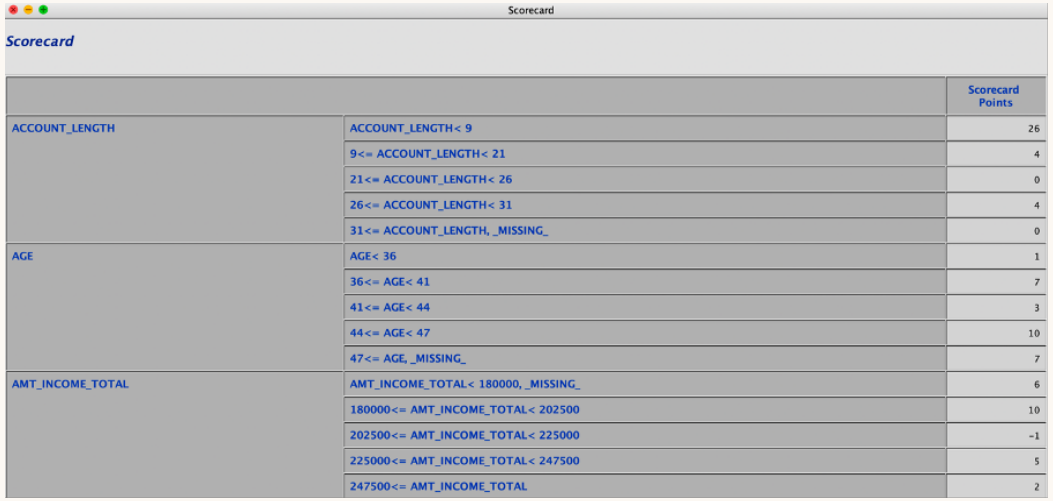


Weight of Evidence Measure (WoE)

Obtained the WoE for ACCOUNT_LENGTH. The event count refers to the number of high risk applicants whereas non-event count refers to the number of low risk applicants. For example, for group that the *ACCOUNT_LENGTH* less than 9 months, positive value of WOE is obtained where the proportion of low risk applicants is higher than high risk applicants. It implies that the applicant falls in that group has lower credit risk

LEE YING QIU 17108552

Credit Scoring in SAS E-Miner


Scorecard

Scorecard

		Scorecard Points
ACCOUNT_LENGTH	ACCOUNT_LENGTH< 9	26
	9<= ACCOUNT_LENGTH< 21	4
	21<= ACCOUNT_LENGTH< 26	0
	26<= ACCOUNT_LENGTH< 31	4
	31<= ACCOUNT_LENGTH, _MISSING_	0
AGE	AGE< 36	1
	36<= AGE< 41	7
	41<= AGE< 44	3
	44<= AGE< 47	10
	47<= AGE, _MISSING_	7
AMT_INCOME_TOTAL	AMT_INCOME_TOTAL< 180000, _MISSING_	6
	180000<= AMT_INCOME_TOTAL< 202500	10
	202500<= AMT_INCOME_TOTAL< 225000	-1
	225000<= AMT_INCOME_TOTAL< 247500	5
	247500<= AMT_INCOME_TOTAL	2

ACCOUNT_LENGTH	ACCOUNT_LENGTH < 9	26 pts
AGE	36 <= AGE < 41	7 pts
AMT_INCOME_TOTAL	180000 <= AMT_INCOME_TOTAL < 202500	10 pts
TOTAL POINTS		43 pts

Connect a "Scorecard" node to construct the scorecard points using the attributes. Based on the scorecard points, the credit scoring of a credit card applicant can be calculated. Assuming a new credit card applicant which possess characteristics as follows:

LEE YING QIU 17108552

Conclusion

Learning outcome :

a) Identified and sampled data.

b) Understand and explore the data by Univariate, Bivariate and Multivariate Analysis.

c) Utilize SAS Enterprise Miner and SAS Studio to clean the data.

d) Realize importance of data mining

e) Assessing the model which generate the classification results with the highest accuracy (lowest misclassification)

f) Explore the Credit Scoring module in SAS Enterprise Miner to assist in building scorecard for each profile