City University of Hong Kong MS4224 Enterprise Data Mining 2019/20 Semester A Project Report

Quantitative Analysis on Home Credit's Operational Strategy

Team: LamboKiwi

Chow Sai Yat Morris 54796760

Table of contents

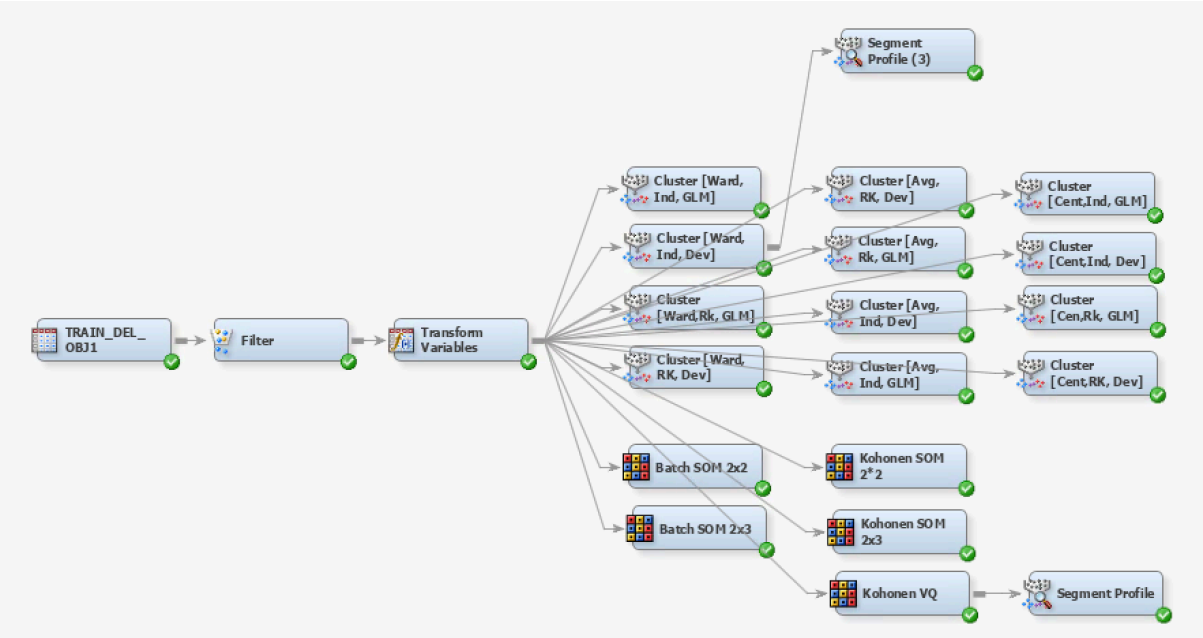
Company background & project objectives:

The Home Credit Group focuses on providing loan products to the unbanked population. The company relies on data to predict their client's repayment abilities to minimize risks. Currently, the company offers a limited number of loan types compared to the market. Therefore, we would like to use data mining approaches to discover potential business

opportunities for the company, as well as investigate in the customer behaviours to adjust the operational and strategic plans.

Objective 1: Customer Segmentation

The purpose is to improve customer intimacy by proposing business strategies to different customer segments based on their characteristics.



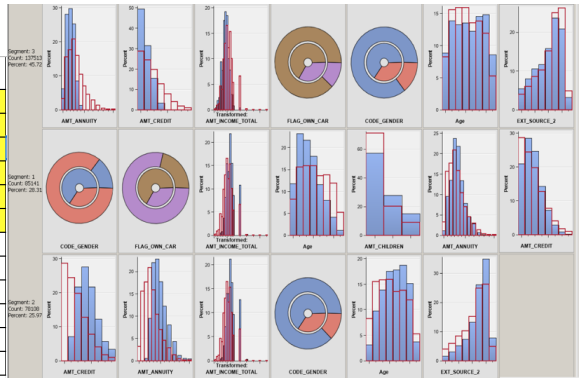
Methodology:

- Clustering analysis and Self-organizing map are used as they help to segmentize the customers according to their demographic information (e.g. age, gender, family size), credit score, as well as income level.
- To select the best model, the distributions of the variables in segments and statistics (CCC, Pseudo F, R-square, Within STD) are compared for model selection.

Clustering analysis results:

Average, ward and centroid methods are used to discover the number of clusters. Since the distributions of our variables are biased, we used internal standardization. Along with the ordinal (Index, Rank) and nominal (GLM, Deviation) encoding approaches, a total of 12 models were created. Yet half of these models obtained the same and comparatively better statistic results of the selection criteria. Therefore, we randomly selected the model with “Average Standardization Index Deviation” that generated 3 segments for later comparison.

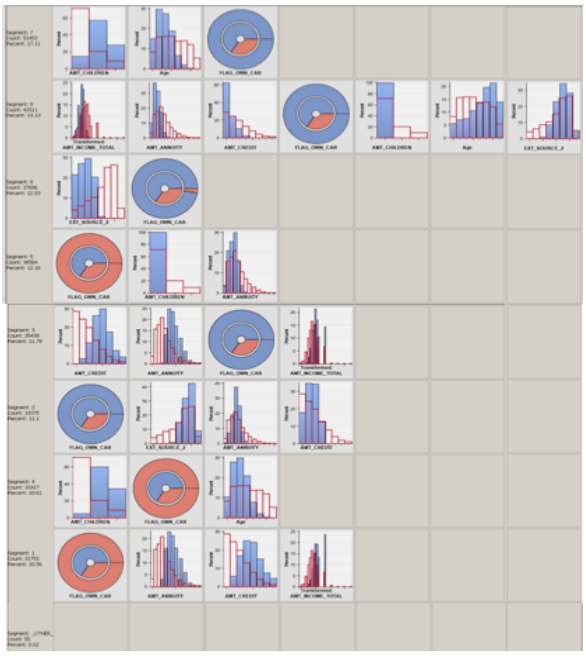
Cluster Method	Internal Standardization	Ordinal Encoding	Nominal Encoding	No. of Cluster (K)	CCC	PSF	RSQ	Within-STD
Average	Standardization	Index	Deviation	3	637.0984	57754.59	0.277487	0.803726
Average	Standardization	Rank	Deviation	3	637.0984	57754.59	0.277487	0.803726
Centroid	Standardization	Index	Deviation	3	637.0984	57754.59	0.277487	0.803726
Centroid	Standardization	Rank	Deviation	3	637.0984	57754.59	0.277487	0.803726
Ward	Standardization	Index	Deviation	3	637.0984	57754.59	0.277487	0.803726
Ward	Standardization	Rank	Deviation	3	637.0984	57754.59	0.277487	0.803726
Average	Standardization	Index	GLM	5	428.4803	42384.77	0.360494	0.685056
Average	Standardization	Rank	GLM	5	428.4803	42384.77	0.279012	0.685056
Centroid	Standardization	Index	GLM	5	428.4803	42384.77	0.360494	0.685056
Centroid	Standardization	Rank	GLM	5	428.4803	42384.77	0.360494	0.685056
Ward	Standardization	Index	GLM	5	428.4803	42384.77	0.360494	0.685056
Ward	Standardization	Rank	GLM	5	428.4803	42384.77	0.360494	0.685056



Self-organizing map results:

Batch SOM, Kohonen SOM and Kohonen VQ are used to form different sizes of matrices. However, batch SOM gave negative CCC value when forming 2x2 and 2x3 matrices, and we would not consider model with CCC less than 3. Among these obtained results, Kohonen VQ with 10 clusters generated was chosen to do cluster analysis as it has higher CCC and R-square.

SOM Method	Internal Standardization	No. of Cluster	CCC	PSF	RSQ	Within-STD
Kohonen VQ	Standardization	10	594.647	33748.22	0.502467	0.602957
Kohonen SOM	Standardization	2x3	480.6263	42157.81	0.412063	0.657167
Kohonen SOM	Standardization	2x2	397.3765	43821.32	0.304158	0.714542
Batch SOM	Standardization	2x2	-112.447	26061.84	0.206325	0.755468
Batch SOM	Standardization	2x3	-411.461	18229.4	0.232575	0.827179



Interpretation & Recommendations:

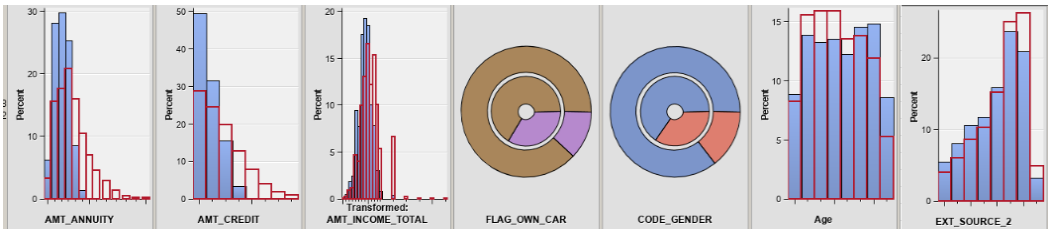
Model (with 3 segments) from clustering analysis was selected as the best model as it has more significant differences in characteristics in each generated segment and

relatively better statistical results.

Method	Internal Standardization	No. of Cluster (K)	CCC	PSF	RSQ	Within-STD
Average	Standardization	3	637.0984	57754.59	0.277487	0.803726
Kohonen VQ	Standardization	10	594.647	33748.22	0.502467	0.602957

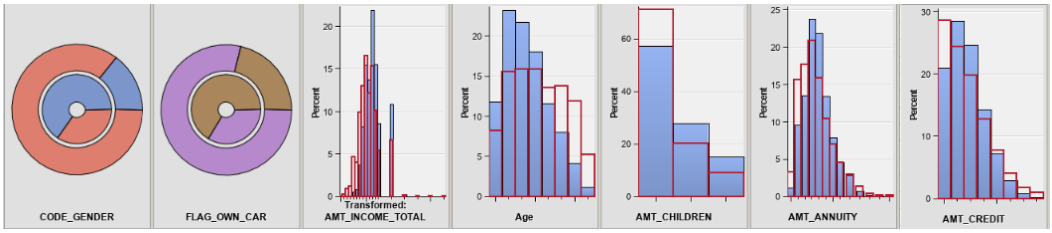
Recommendations for segment 3:

Segment 3 features car owners and customers with a low credit score. So we see there is a potential for our company to introduce car loans business. For customers with a low credit score, we suggest setting a higher interest rate for their cash loan applications to minimize the default risk.



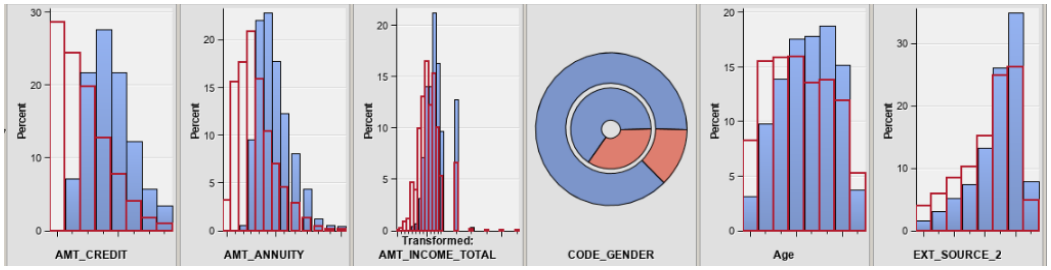
Recommendations for segment 1:

Segment 1 features young adults with family. As this group of people are most likely working-class and have children, there are potential for tax loan and mortgage businesses.



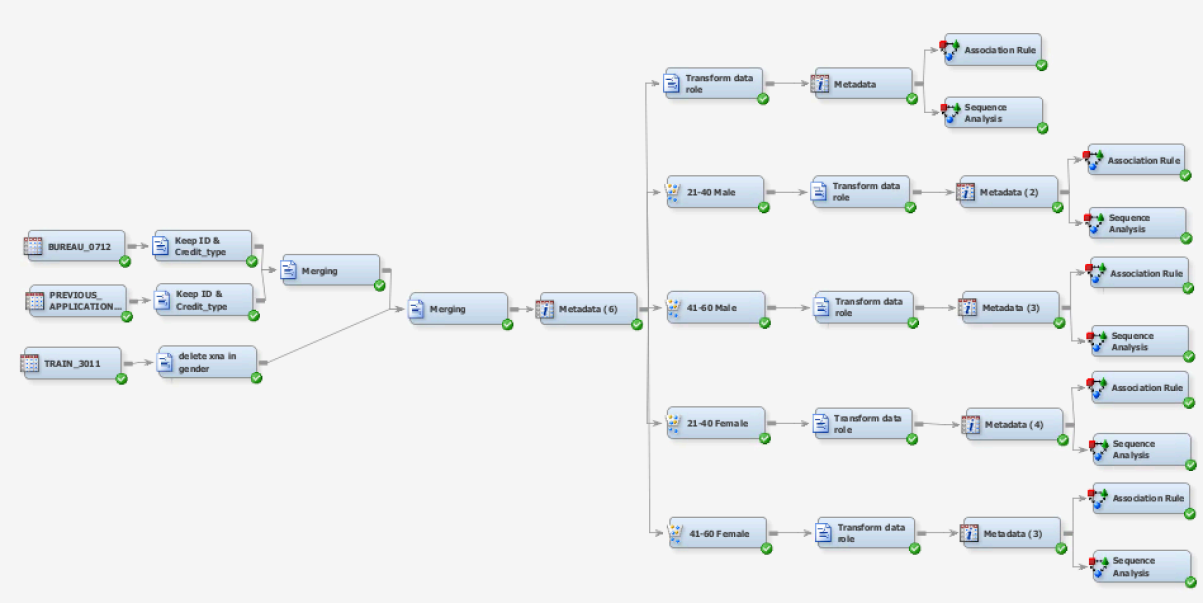
Recommendations for segment 2:

Customers in segment 2 have a high credit score and income level. They have large business capability for our company and therefore we suggest putting more resources or provide premium services to build long-term relationships with them.



Objective 2: Identify loan application patterns

Through identifying the loan application pattern of our customers in our company as well as other institutes, it is aimed to adjust the marketing strategies and discover opportunities for new loan products by finding out popular loan products.



Methodology:

- Association rule for all customer
- Sequence Analysis for all customer
- Association rule and Sequence Analysis with different customer segment

Results for Association Rule:

We select the rules with lift greater than 1 as a significant rule. Confidence percentage was then be used as the selection criteria.

Results are similar before and after customer segmentation. It is found that there are 3 rules with confidence greater than 85%.

{Cash loans ⇒ Consumer loans}

{Mortgage ⇒ Consumer loans}

{Consumer Credit ⇒ Consumer loans}

All customers

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift ▼	Transaction Count	Rule
2	27.62	59.30	6.21	2.15	19101	Mortgage ==> Credit card
2	10.47	22.49	6.21	2.15	19101	Credit card ==> Mortgage
2	27.62	57.12	11.28	2.07	34691	Car loan ==> Credit card
2	19.75	40.84	11.28	2.07	34691	Credit card ==> Car loan
2	27.62	52.80	5.03	1.91	15467	Consumer credit ==> Credit card
2	9.53	18.21	5.03	1.91	15467	Credit card ==> Consumer credit
2	52.58	53.54	45.82	1.02	140885	Consumer loans ==> Cash loans
2	85.56	87.14	45.82	1.02	140885	Cash loans ==> Consumer loans
2	85.56	85.83	8.99	1.00	27644	Mortgage ==> Consumer loans
2	10.47	10.51	8.99	1.00	27644	Consumer loans ==> Mortgage
2	85.56	85.30	8.13	1.00	24989	Consumer credit ==> Consumer loans

Male aged 21-40

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift ▼	Transaction Count	Rule
2	29.91	59.15	13.06	1.98	6951.0	Mortgage ==> Credit card
2	22.08	43.66	13.06	1.98	6951.0	Credit card ==> Mortgage
2	44.09	50.66	16.86	1.15	8972.0	Cash loans ==> Revolving loans
2	84.48	87.61	29.15	1.04	15515	Cash loans ==> Consumer loans
2	84.48	85.16	18.80	1.01	10008	Mortgage ==> Consumer loans

Female aged 21-40

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift ▼	Transaction Count	Rule
2	27.90	59.40	5.62	2.13	10995	Mortgage ==> Credit card
2	9.46	20.15	5.62	2.13	10995	Credit card ==> Mortgage
2	20.76	42.68	11.91	2.06	23294	Credit card ==> Car loan
2	27.90	57.36	11.91	2.06	23294	Car loan ==> Credit card
2	10.52	19.77	5.52	1.88	10790	Credit card ==> Consumer credit
2	27.90	52.43	5.52	1.88	10790	Consumer credit ==> Credit card
2	86.36	87.72	48.39	1.02	94675	Cash loans ==> Consumer loans

Male aged 41-60

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift ▼	Transaction Count	Rule
2	11.94	26.19	7.09	2.19	7449.0	Credit card ==> Mortgage
2	27.08	59.36	7.09	2.19	7449.0	Mortgage ==> Credit card
2	18.24	38.02	10.30	2.08	10816	Credit card ==> Car loan
2	27.08	56.45	10.30	2.08	10816	Car loan ==> Credit card
2	7.85	15.56	4.21	1.98	4425.0	Credit card ==> Consumer credit
2	27.08	53.66	4.21	1.98	4425.0	Consumer credit ==> Credit card
2	48.08	49.10	41.25	1.02	43335	Consumer loans ==> Cash loans
2	84.01	85.79	41.25	1.02	43335	Cash loans ==> Consumer loans

Female aged 41-60

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift ▼	Transaction Count	Rule
2	27.90	59.27	5.76	2.12	11652	Mortgage ==> Credit card
2	9.71	20.63	5.76	2.12	11652	Credit card ==> Mortgage
2	20.53	42.26	11.79	2.06	23875	Credit card ==> Car loan
2	27.90	57.44	11.79	2.06	23875	Car loan ==> Credit card
2	10.40	19.55	5.45	1.88	11042	Credit card ==> Consumer credit
2	27.90	52.46	5.45	1.88	11042	Consumer credit ==> Credit card
2	86.37	87.76	48.19	1.02	97550	Cash loans ==> Consumer loans

Results for Sequence Analysis:

We select the rules with lift greater than 1 as a significant rule. Confidence percentage was then be used as the selection criteria.

Repeat purchase on same loan type were found for all customers. Customer segmented by sex and age found to have slightly different preference on loan purchase.

Below lists the preference of customers in different segments
Male Aged 21-40: Mortgage
Female Aged 21-40: Car Loan
Male Aged 41-60: Car Loan
Female Aged 41-60: Car Loan
Young Adult: {Credit Card/Revolving loans ⇒ Consumer Loans}
Elder Adult: Cash Loan

According to the Rule Matrix map, it is found that no matter which type of loan the customer applied, they will apply consumer loan later.

All Customers

Chain Length	Transaction Count	Support(%) ▼	Confidence(%)	PseudoLift	Rule
2	111316	42.25	49.05	0.57	Consumer loans ==> Consumer loans
2	84825	32.19	60.23	1.13	Cash loans ==> Cash loans
2	82144	31.18	36.20	0.68	Consumer loans ==> Cash loans
2	82114	31.16	58.31	0.68	Cash loans ==> Consumer loans
3	53134	20.17	47.73	0.55	Consumer loans ==> Consumer loans ==> Consumer loans
3	51910	19.70	61.20	1.14	Cash loans ==> Cash loans ==> Cash loans
2	50337	19.10	59.27	0.69	Credit card ==> Consumer loans
2	50201	19.05	22.12	0.69	Consumer loans ==> Credit card
2	42875	16.27	70.60	3.06	Car loan ==> Car loan

Male aged 21-40

Chain Length	Transaction Count	Support(%) ▼	Confidence(%)	PseudoLift	Rule
2	21293	47.41	55.61	0.65	Consumer loans ==> Consumer loans
3	11094	24.70	52.10	0.61	Consumer loans ==> Consumer loans ==> Consumer loans
2	9888	22.02	50.33	0.59	Revolving loans ==> Consumer loans
2	9830	21.89	61.75	0.72	Credit card ==> Consumer loans
2	9794	21.81	25.58	0.58	Consumer loans ==> Revolving loans
2	9688	21.57	25.30	0.71	Consumer loans ==> Credit card
2	8938	19.90	56.15	1.58	Credit card ==> Credit card
2	8741	19.46	57.00	0.67	Cash loans ==> Consumer loans
2	8734	19.45	22.81	0.67	Consumer loans ==> Cash loans
2	8414	18.74	71.60	2.74	Mortgage ==> Mortgage
2	8123	18.09	69.12	0.81	Mortgage ==> Consumer loans

Female aged 21-40

Chain Length	Transaction Count	Support(%) ▼	Confidence(%)	PseudoLift	Rule
2	71113	42.08	48.45	0.56	Consumer loans ==> Consumer loans
2	59023	34.93	62.26	1.11	Cash loans ==> Cash loans
2	56447	33.40	38.46	0.69	Consumer loans ==> Cash loans
2	56390	33.37	59.48	0.68	Cash loans ==> Consumer loans
3	37015	21.90	62.71	1.12	Cash loans ==> Cash loans ==> Cash loans
3	33453	19.79	47.04	0.54	Consumer loans ==> Consumer loans ==> Consumer loans
2	32466	19.21	22.12	0.68	Consumer loans ==> Credit card
2	32413	19.18	59.39	0.68	Credit card ==> Consumer loans
3	30185	17.86	51.14	0.59	Cash loans ==> Cash loans ==> Consumer loans
3	30121	17.82	53.42	0.95	Cash loans ==> Consumer loans ==> Cash loans
3	30107	17.81	53.34	0.95	Consumer loans ==> Cash loans ==> Cash loans
2	29927	17.71	73.69	3.07	Car loan ==> Car loan

Male aged 41-60

Chain Length	Transaction Count	Support(%) ▼	Confidence(%)	PseudoLift	Rule
2	37692	42.52	50.22	0.59	Consumer loans ==> Consumer loans
2	24109	27.20	55.85	1.15	Cash loans ==> Cash loans
2	24042	27.12	55.69	0.66	Cash loans ==> Consumer loans
2	23983	27.05	31.95	0.66	Consumer loans ==> Cash loans
3	18506	20.87	49.10	0.58	Consumer loans ==> Consumer loans ==> Consumer loans
2	16791	18.94	59.03	0.70	Credit card ==> Consumer loans
2	16600	18.72	22.12	0.69	Consumer loans ==> Credit card
2	14365	16.20	50.50	1.57	Credit card ==> Credit card
3	13873	15.65	57.54	1.18	Cash loans ==> Cash loans ==> Cash loans
2	12296	13.87	64.17	2.97	Car loan ==> Car loan
2	12120	13.67	16.15	0.75	Consumer loans ==> Car loan
2	12024	13.56	62.75	0.74	Car loan ==> Consumer loans

Female aged 41-60

Chain Length	Transaction Count	Support(%) ▼	Confidence(%)	Pseudo Lift	Rule
2	73624	42.11	48.48	0.56	Consumer loans ==> Consumer loans
2	60716	34.73	62.17	1.11	Cash loans ==> Cash loans
2	58161	33.27	38.30	0.69	Consumer loans ==> Cash loans
2	58072	33.22	59.46	0.68	Cash loans ==> Consumer loans
3	38037	21.76	62.65	1.12	Cash loans ==> Cash loans ==> Cash loans
3	34628	19.81	47.03	0.54	Consumer loans ==> Consumer loans ==> Consumer loans
2	33601	19.22	22.12	0.68	Consumer loans ==> Credit card
2	33546	19.19	59.39	0.68	Credit card ==> Consumer loans
3	31057	17.76	51.15	0.59	Cash loans ==> Cash loans ==> Consumer loans
3	31008	17.74	53.31	0.95	Consumer loans ==> Cash loans ==> Cash loans
3	30970	17.71	53.33	0.95	Cash loans ==> Consumer loans ==> Cash loans
2	30579	17.49	73.56	3.09	Car loan ==> Car loan

Interpretation & Recommendations:

From the results, it is known that Mortgage, Credit card and Car loan are popular products in other institutes. Instead of categorising the lons in three big group, which is Consumer loan, Cash loan and Revolving loan, it is recommended that HomeCredit should diversify its product to satisfy different customers' need.

As repeat applications on same loan type was found, it is suggested that HomeCredit should offer lower interest rate for repeat application to encourage repeat loan applications. Moreover, as most customer will apply consumer loans later, cross selling consumer loans with other product is suggested.

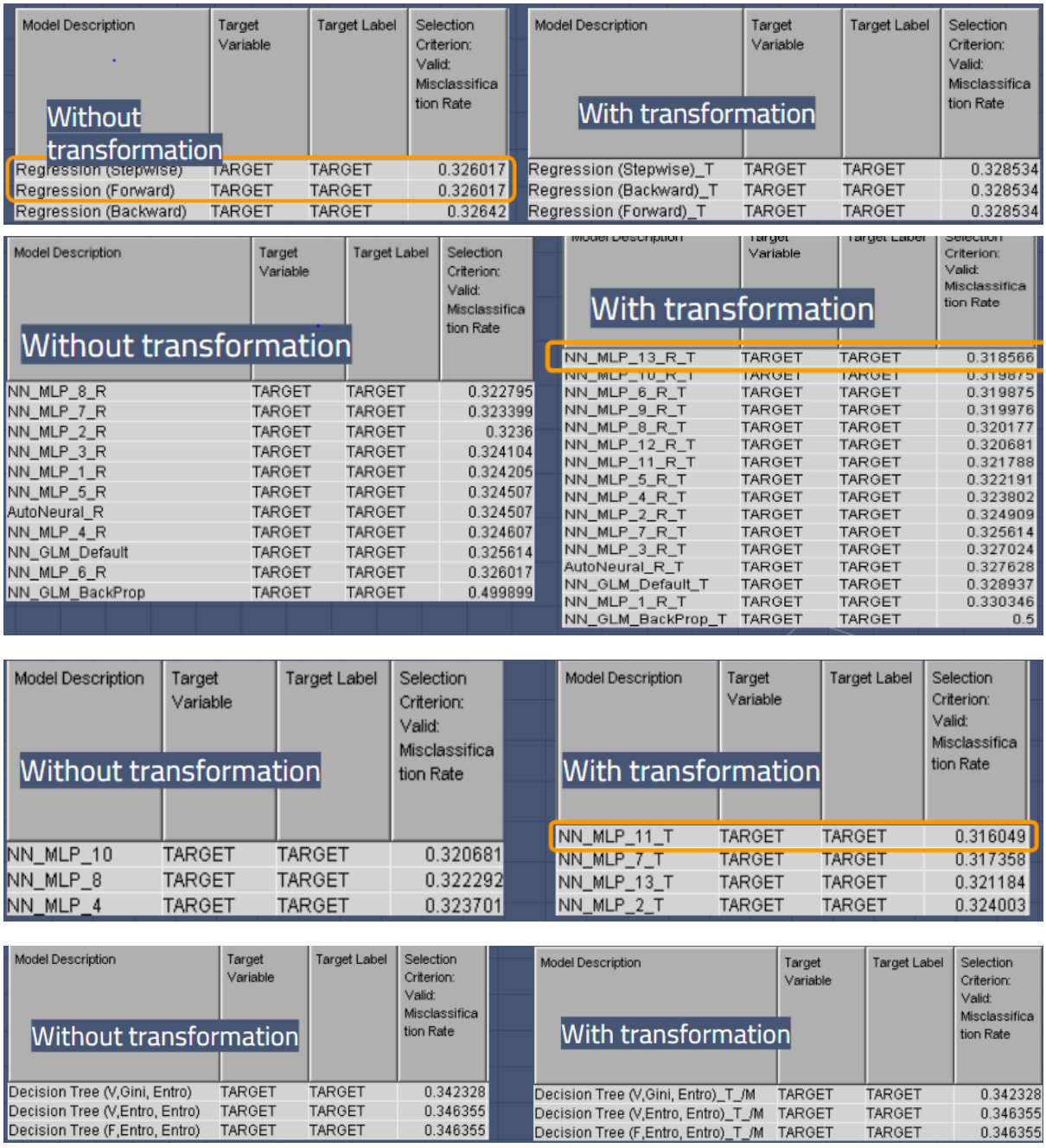
Last but not least, as customer in different segments having different preference, HomeCredit can send loan information which suits their preference to customer according to their demographic information. For example, send mortgage information to young male while sending car loan information to young female.

Objective 3: Identify business risk with loan overdue prediction

Many people struggle to get loans owing to insufficient or non-existent credit histories. Therefore we strive to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience.

Thus, in Objective 3, we focused on the prediction of the occurrence of customer's overdue in order to ensure that clients capable of repayment are not rejected and that the loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful. In model selection, we make use of a variety of models, including logistic regression, neural network, and decision tree to predict clients' repayment abilities and make use of the prediction results to allocate resources to customers/segments with lower default risk.

The model detail that we used in the predictions are, logistic regression since the type of the target is a binary variable. In the decision tree model, different model criteria for the decision tree as in ProbChisq, Entropy, Gini, Variance, and ProbF are chosen for each type of target. Moreover, the auto neural network and network by using GLM and MLP are applied in the neural network model. The result of the models are as following;



Without transformation

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Regression (Stepwise)	TARGET	TARGET	0.326017
Regression (Forward)	TARGET	TARGET	0.326017
Regression (Backward)	TARGET	TARGET	0.32642

With transformation

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Regression (Stepwise)_T	TARGET	TARGET	0.328534
Regression (Backward)_T	TARGET	TARGET	0.328534
Regression (Forward)_T	TARGET	TARGET	0.328534

Without transformation

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
NN_MLP_8_R	TARGET	TARGET	0.322795
NN_MLP_7_R	TARGET	TARGET	0.323399
NN_MLP_2_R	TARGET	TARGET	0.3236
NN_MLP_3_R	TARGET	TARGET	0.324104
NN_MLP_1_R	TARGET	TARGET	0.324205
NN_MLP_5_R	TARGET	TARGET	0.324507
AutoNeural_R	TARGET	TARGET	0.324507
NN_MLP_4_R	TARGET	TARGET	0.324607
NN_GLM_Default	TARGET	TARGET	0.325614
NN_MLP_6_R	TARGET	TARGET	0.326017
NN_GLM_BackProp	TARGET	TARGET	0.499899

With transformation

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
NN_MLP_13_R_T	TARGET	TARGET	0.318566
NN_MLP_10_R_T	TARGET	TARGET	0.319875
NN_MLP_6_R_T	TARGET	TARGET	0.319875
NN_MLP_9_R_T	TARGET	TARGET	0.319976
NN_MLP_8_R_T	TARGET	TARGET	0.320177
NN_MLP_12_R_T	TARGET	TARGET	0.320681
NN_MLP_11_R_T	TARGET	TARGET	0.321788
NN_MLP_5_R_T	TARGET	TARGET	0.322191
NN_MLP_4_R_T	TARGET	TARGET	0.323802
NN_MLP_2_R_T	TARGET	TARGET	0.324909
NN_MLP_7_R_T	TARGET	TARGET	0.325614
NN_MLP_3_R_T	TARGET	TARGET	0.327024
AutoNeural_R_T	TARGET	TARGET	0.327628
NN_GLM_Default_T	TARGET	TARGET	0.328937
NN_MLP_1_R_T	TARGET	TARGET	0.330346
NN_GLM_BackProp_T	TARGET	TARGET	0.5

Without transformation

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
NN_MLP_10	TARGET	TARGET	0.320681
NN_MLP_8	TARGET	TARGET	0.322292
NN_MLP_4	TARGET	TARGET	0.323701

With transformation

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
NN_MLP_11_T	TARGET	TARGET	0.316049
NN_MLP_7_T	TARGET	TARGET	0.317358
NN_MLP_13_T	TARGET	TARGET	0.321184
NN_MLP_2_T	TARGET	TARGET	0.324003

Without transformation

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Decision Tree (V,Gini, Entro)	TARGET	TARGET	0.342328
Decision Tree (V,Entro, Entro)	TARGET	TARGET	0.346355
Decision Tree (F,Entro, Entro)	TARGET	TARGET	0.346355

With transformation

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Decision Tree (V,Gini, Entro)_T_/M	TARGET	TARGET	0.342328
Decision Tree (V,Entro, Entro)_T_/M	TARGET	TARGET	0.346355
Decision Tree (F,Entro, Entro)_T_/M	TARGET	TARGET	0.346355

Results and Interpretation:

With the overall model comparison, the best model is the neural network model by using the transformed variable with 11 hidden units and MLP network which obtains the lowest misclassification rate among the models. (0.316049)

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Misclassification Rate
NN_MLP_11_T	TARGET	TARGET	0.316049	0.322297
NN_MLP_10	TARGET	TARGET	0.320681	0.32129

By using the decision tree method to interpret the result of the best model, in terms of the variable importance, we discovered that the most important feature of a client that will gradually affect the probability to overdue are EXT_SOURCE_3(TU Score), EXT_SOURCE_2(FICO Score), education type, gender and days employed.

We may use these feature importances as a method of dimensionality reduction in future work in order to reduce the multicollinearity and the number of the variables that obstruct the pattern finding.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance
IMP_EXT_SOURCE_3	Imputed: EXT_SOURCE_3	8	1.0000	1.0000
LOG_IMP_EXT_SOURCE_2	Transformed: Imputed: EXT_SOURCE_2	5	0.7372	0.7082
NAME_EDUCATION_TYPE	NAME_EDUCATION_TYPE	3	0.2847	0.2681
CODE_GENDER	CODE_GENDER	3	0.2816	0.2652
LOG_IMP_REP_DAYS_EMPLOYED	Transformed: Imputed: Replacement: DAYS_EMPLOYED	2	0.1848	0.2026
Age	Age	2	0.1638	0.1872
FLAG_OWN_CAR	FLAG_OWN_CAR	2	0.1532	0.1286
TI_REGION_RATING_CLIENT_W_C1	REGION_RATING_CLIENT_W_CITY:1	1	0.0928	0.1110
LOG_IMP_payToincome	Transformed: Imputed payToincome	1	0.0863	0.0863
TI_REGION_RATING_CLIENT_W_C3	REGION_RATING_CLIENT_W_CITY:3	1	0.0859	0.0851
NAME_CONTRACT_TYPE	NAME_CONTRACT_TYPE	1	0.0498	0.0461

While looking at the odds ratio estimates, we can discover that males, people who own a cash loan and a lower education level holder will have a higher chance of default. It gave us some insights that tightening cash loan policies can reduce the overdue probability and we should allocate more resources in revolving loans since there is more customer with good quality. We can also connect the model to the scoring engine to evaluate the customer status for risk management.

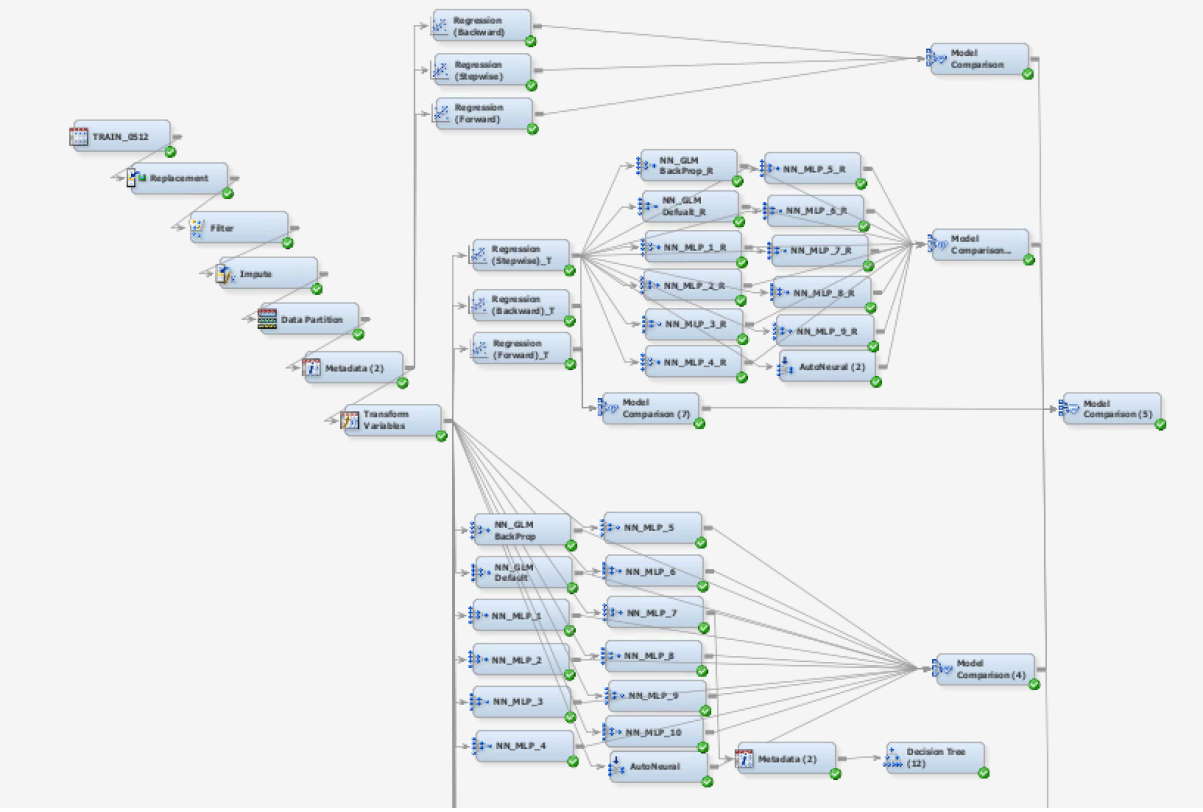
Variables	Prob(overdue)
Gender	**M** > F
Contract type	**Cash loan** > Revolving loan
Education type	**Lower secondary** > Secondary > Incomplete higher > Higher education > Degree

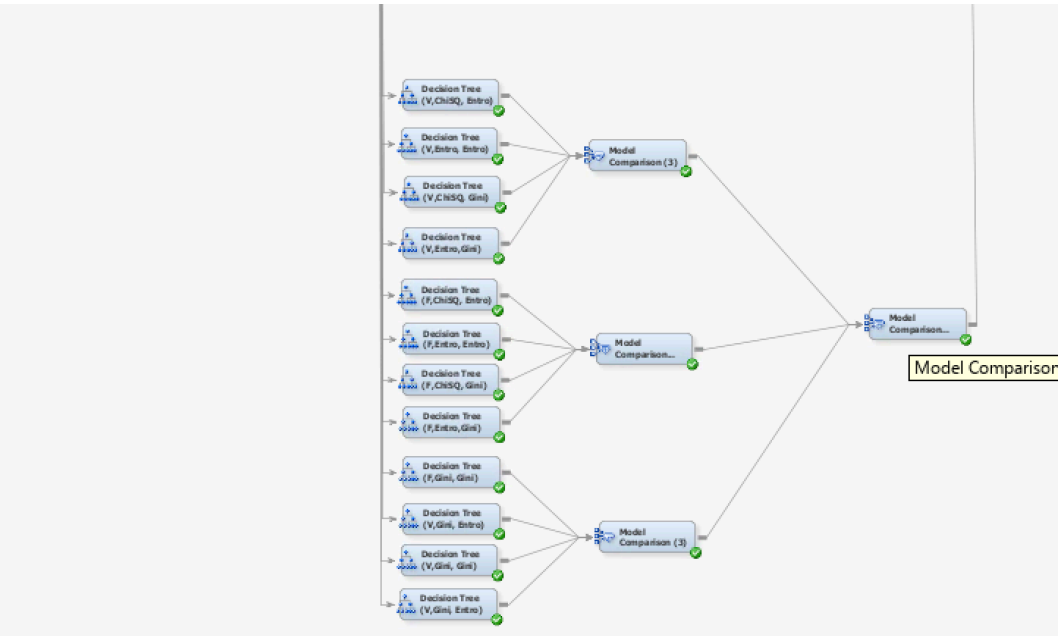
Odds Ratio Estimates

Effect		Point Estimate
CODE_GENDER	F vs M	0.674
FLAG_OWN_CAR	0 vs 1	1.316
IMP_EXT_SOURCE_2		0.123
IMP_EXT_SOURCE_3		0.059
IMP_REP_DAYS_EMPLOYED		1.000
IMP_payToincome		2.924
NAME_CONTRACT_TYPE	Cash loans vs Revolving loans	1.517
NAME_EDUCATION_TYPE	Academic degree vs Secondary / secondary special	0.214
NAME_EDUCATION_TYPE	Higher education vs Secondary / secondary special	0.677
NAME_EDUCATION_TYPE	Incomplete higher vs Secondary / secondary special	0.803
NAME_EDUCATION_TYPE	Lower secondary vs Secondary / secondary special	1.184
NAME_FAMILY_STATUS	Civil marriage vs Widow	1.179
NAME_FAMILY_STATUS	Married vs Widow	0.935
NAME_FAMILY_STATUS	Separated vs Widow	1.092
NAME_FAMILY_STATUS	Single / not married vs Widow	1.080
NAME_INCOME_TYPE	Commercial associate vs Working	0.874
NAME_INCOME_TYPE	Maternity leave vs Working	109.119
NAME_INCOME_TYPE	Pensioner vs Working	0.658
NAME_INCOME_TYPE	State servant vs Working	0.779
NAME_INCOME_TYPE	Student vs Working	0.008
NAME_INCOME_TYPE	Unemployed vs Working	5.265
REGION_RATING_CLIENT_W_CITY	1 vs 3	0.675
REGION_RATING_CLIENT_W_CITY	2 vs 3	0.833

Objective 4: Predict client's credit limit

The objective is to predict the credit amount for Home Credit applicants in order to provide a systemic lending mechanism for minimising the risk of default payment.





Methodology:

I. Linear Regression

The regression models are built with stepwise, forward and backward variable selection using without transformed and transformed variables respectively. In the linear regression model comparison, the stepwise regression using transformed variables is selected as it has a smallest ASE of 0.159101 and all parameters could make significant predictions on the credit limit.

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error
Without transformation				With transformation			
Regression (Stepwise)	AMT_CRED...	AMT_CRED...	3.36E10	Regression (Stepwise)_T	LOG_AMT_...	Transforme...	0.159101
Regression (Backward)	AMT_CRED...	AMT_CRED...	3.36E10	Regression (Backward)_T	LOG_AMT_...	Transforme...	0.159101
Regression (Forward)	AMT_CRED...	AMT_CRED...	3.36E10	Regression (Forward)_T	LOG_AMT_...	Transforme...	0.159101

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	22	2688.741474	122.215522	762.05	<.0001
Error	23397	3752.354699	0.160378		
Corrected Total	23419	6441.096174			

Significant model

Model Fit Statistics

R-Square	0.4174	Adj R-Sq	0.4169
AIC	-42840.8631	BIC	-42838.8149
SBC	-42655.4522	C(p)	21.4828

Type 3 Analysis of Effects

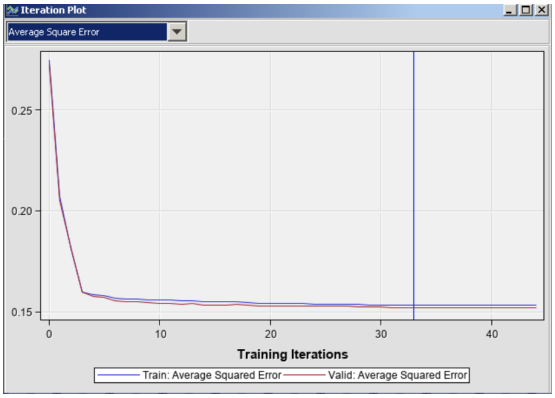
Effect	DF	Sum of Squares	F Value	Pr > F
CODE_GENDER	1	33.1324	206.59	<.0001
LOG_AMT_INCOME_TOTAL	1	1227.0419	7650.96	<.0001
LOG_Age	1	115.7635	721.82	<.0001
LOG_IMP_EXT_SOURCE_2	1	5.7360	35.77	<.0001
LOG_IMP_EXT_SOURCE_3	1	21.0077	130.99	<.0001
LOG_IMP_REP_DAYS_EMPLOYED	1	5.5114	34.36	<.0001
M_EXT_SOURCE_2	1	1.8125	11.30	0.0008
M_EXT_SOURCE_3	1	1.6656	10.39	0.0013
M_REP_DAYS_EMPLOYED	1	0.7986	4.98	0.0257
REGION_RATING_CLIENT_W_CITY	2	24.9311	77.73	<.0001
TG_AMT_CHILDREN	2	2.7980	8.72	0.0002
TG_AMT_FAM_MEMBERS	3	12.6684	26.33	<.0001
TG_NAME_EDUCATION_TYPE	2	70.7960	220.72	<.0001
TG_NAME_HOUSING_TYPE	1	0.6709	4.18	0.0408
TG_NAME_INCOME_TYPE	3	3.6168	7.52	<.0001

II. Neural Network After Regression

Out of 12 neural network after regression models, we found out that neural network using multilayer perceptron with 9 hidden units obtains smallest ASE (0.15205) which is slightly less than stepwise linear regression previously.

With transformation

Model Description	Target	Target Label	Selection Criterion: Valid: Average Squared Error
NN_MLP_4_R	LOG_AMT_...	Transforme...	0.15205
NN_MLP_7_R	LOG_AMT_...	Transforme...	0.15286
AutoNeural (2)	LOG_AMT_...	Transforme...	0.152974
NN_MLP_3_R	LOG_AMT_...	Transforme...	0.153023
NN_MLP_8_R	LOG_AMT_...	Transforme...	0.153267
NN_MLP_6_R	LOG_AMT_...	Transforme...	0.153327
NN_MLP_5_R	LOG_AMT_...	Transforme...	0.153783
NN_MLP_2_R	LOG_AMT_...	Transforme...	0.153866
NN_MLP_9_R	LOG_AMT_...	Transforme...	0.154301
NN_MLP_1_R	LOG_AMT_...	Transforme...	0.155196
NN_GLM Defualt_R	LOG_AMT_...	Transforme...	0.160687
NN_GLM BackProp_R	LOG_AMT_...	Transforme...	0.272766

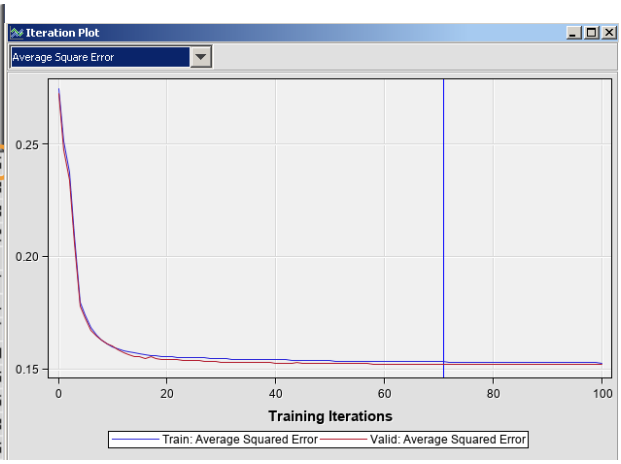


III. Neural Network

We select the best neural network from 13 models using generalized linear GLM and multilayer perceptron MLP methods. In the MLP, the selected variable in the best linear regression model will be the input variables of the neural network, and we try to use different hidden units (1-10) in order to come with the smallest ASE. We find out the MLP neural network with 7 hidden units which variables are selected from the stepwise regression has the smallest ASE with 0.151835.

With transformation

Model Description	Target	Target Label	Selection Criterion: Valid: Average Squared Error
NN_MLP_7	LOG_AMT_...	Transforme...	0.151835
NN_MLP_8	LOG_AMT_...	Transforme...	0.151858
NN_MLP_3	LOG_AMT_...	Transforme...	0.152673
NN_MLP_2	LOG_AMT_...	Transforme...	0.153122
NN_MLP_10	LOG_AMT_...	Transforme...	0.15321
NN_MLP_6	LOG_AMT_...	Transforme...	0.153247
NN_MLP_5	LOG_AMT_...	Transforme...	0.153304
NN_MLP_4	LOG_AMT_...	Transforme...	0.153477
NN_MLP_9	LOG_AMT_...	Transforme...	0.154179
AutoNeural	LOG_AMT_...	Transforme...	0.155206
NN_MLP_1	LOG_AMT_...	Transforme...	0.155206
NN_GLM Default	LOG_AMT_...	Transforme...	0.160688
NN_GLM BackProp	LOG_AMT_...	Transforme...	0.272766



IV. Regression Tree

Out of the 12 regression tree models, the best regression tree models are selected using F-test, Chi Square and Entropy as well as F-test, Gini, Gini combinations. It shows a consistent performance in training and validation data. It minimizes the validation average error to 0.165452.

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error
With transformation			
Decision Tree (F,ChiSQ, Entro)	LOG_AMT_...	Transforme...	0.165452
Decision Tree (F,Gini, Gini)	LOG_AMT_...	Transforme...	0.165452
Decision Tree (V,ChiSQ, Entro)	LOG_AMT_...	Transforme...	0.165573

Model Selection:

Compared with all the methods we have used, we find that MLP neural network with 7 hidden units which variables are selected from the stepwise regression has the smallest ASE with 0.151835. Through the model, we highly recommend Home Credit staff should seriously examine important documents such as income statement, age and education level during the approval process.

Model Description	Target	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Average Squared Error
With transformation				
NN_MLP_7	LOG_AMT_...	Transforme...	0.151835	0.153037
NN_MLP_4_R	LOG_AMT_...	Transforme...	0.15205	0.153219
Regression (Stepwise)_T	LOG_AMT_...	Transforme...	0.159101	0.16022
Decision Tree (F,ChiSQ, Entro)	LOG_AMT_...	Transforme...	0.165452	0.164568

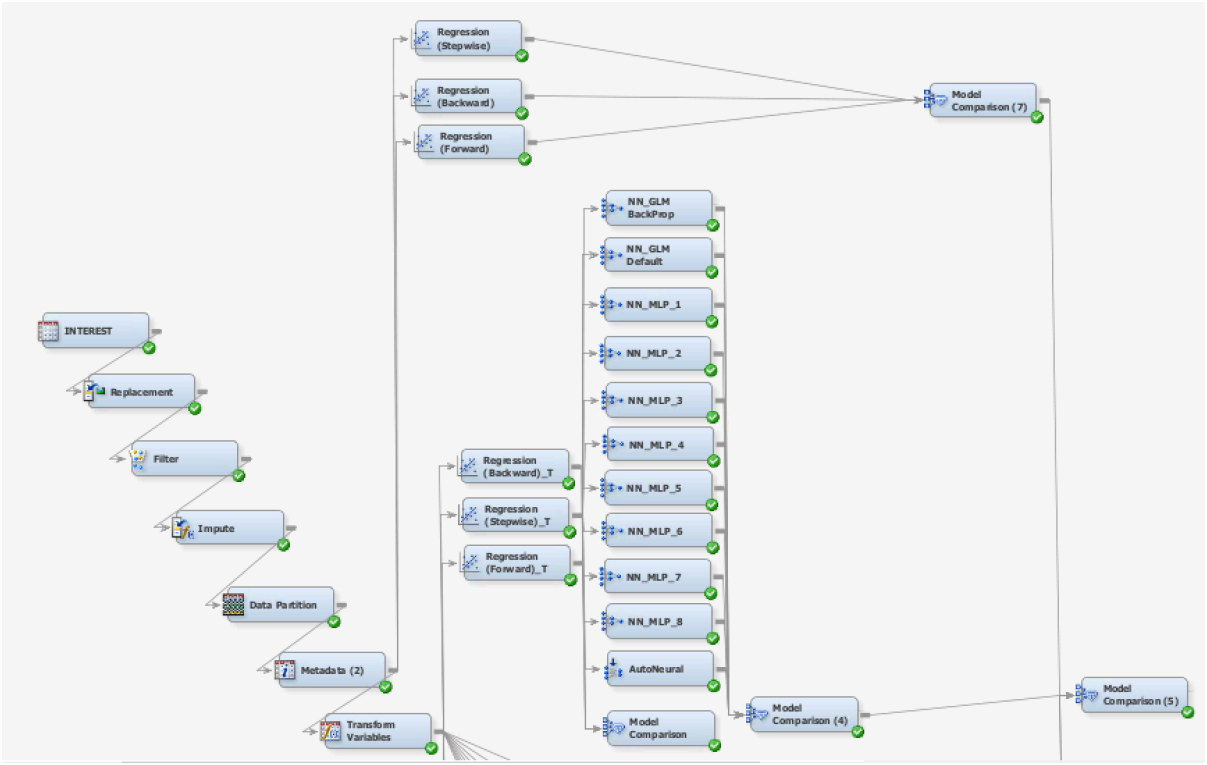
Variable Importance

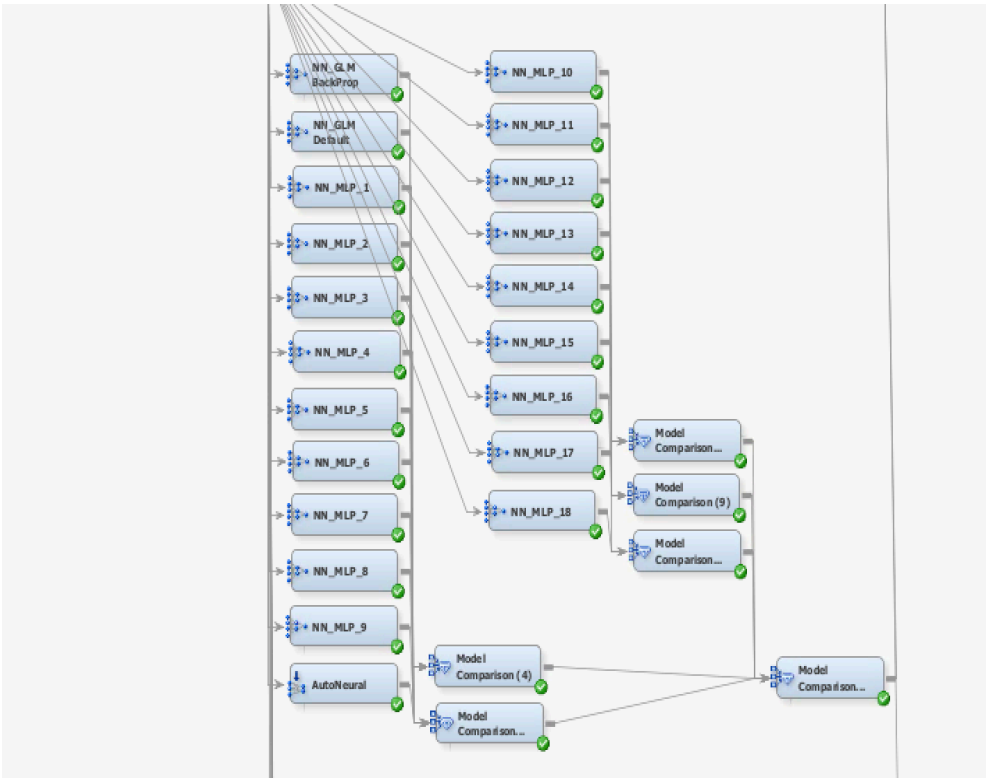
More important variables

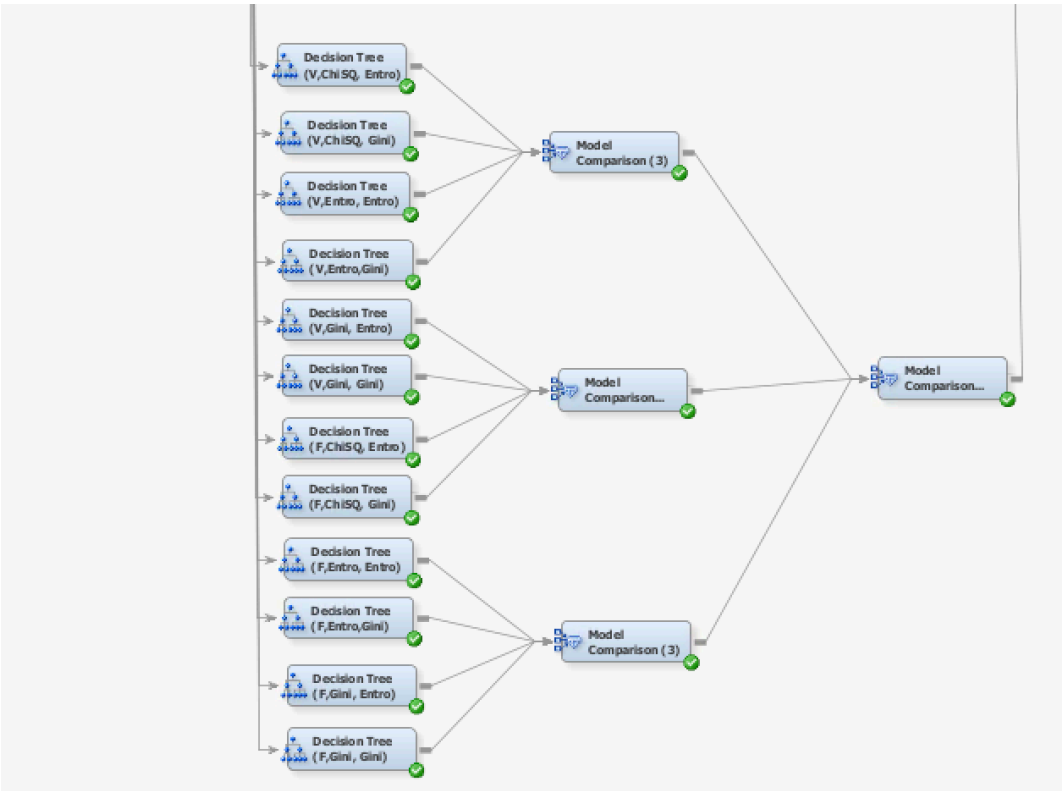
Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LOG_AMT_INCOME_TOTAL	Transformed: AMT_INCOME_TOTAL	17	1.0000	1.0000	1.0000
LOG_Age	Transformed: Age	19	0.3530	0.3505	0.9929
TG_NAME_EDUCATION_TYPE	Transformed: NAME_EDUCATION_TYPE	8	0.2078	0.2114	1.0176
LOG_IMP_EXT_SOURCE_3	Transformed: Imputed: EXT_SOURCE_3	6	0.0962	0.0814	0.8459
CODE_GENDER	CODE_GENDER	6	0.0788	0.0699	0.8873
TG_AMT_CHILDREN	Transformed: AMT_CHILDREN	1	0.0267	0.0155	0.5805
REGION_RATING_CLIENT	REGION_RATING_CLIENT	1	0.0186	0.0200	1.0728
LOG_IMP_EXT_SOURCE_2	Transformed: Imputed: EXT_SOURCE_2	1	0.0176	0.0071	0.4040
LOG_IMP_REP_DAYS_EMPLOYED	Transformed: Imputed: Replacement: DAYS_EMPLOYED	1	0.0077	0.0057	0.7330

Objective 5: Predict interest rate level

To maximize the profit of Home Credit, we tried to build and choose the best model in order to predict interest rate level for new entry customers.


Regression (Stepwise)
Regression (Backward)
Regression (Forward)
Model Comparison (7)
NN_GLM BackProp
NN_GLM Default
NN_MLP_1
NN_MLP_2
NN_MLP_3
NN_MLP_4
NN_MLP_5
NN_MLP_6
NN_MLP_7
NN_MLP_8
AutoNeural
Model Comparison
INTEREST
Replacement
Filter
Impute
Data Partition
Metadata (2)
Transform Variables
Regression (Backward)_T
Regression (Stepwise)_T
Regression (Forward)_T
Model Comparison (4)
Model Comparison (5)


NN_GLM BackProp
NN_GLM Default
NN_MLP_1
NN_MLP_2
NN_MLP_3
NN_MLP_4
NN_MLP_5
NN_MLP_6
NN_MLP_7
NN_MLP_8
NN_MLP_9
AutoNeural
NN_MLP_10
NN_MLP_11
NN_MLP_12
NN_MLP_13
NN_MLP_14
NN_MLP_15
NN_MLP_16
NN_MLP_17
NN_MLP_18
Model Comparison (4)
Model Comparison...
Model Comparison (9)
Model Comparison...
Model Comparison...
Model Comparison...



Methodology

We have used 3 major ones which are linear regression, decision tree and neural network.

As our target is the interest rate level, we chose linear regression instead of logistic and tried three variable selection methods which are Stepwise, Backward and Forward. Then, we added neural network after regression in order to enhance the accuracy of our model.

In the decision tree model, we used regression tree and applied different model criterions which are ProbChisq, Entropy, Gini, Variance, and ProbF to Interval, Nominal and Ordinal target respectively.

For neural network, GLM, MLP and autoneural are both applied.

The results for each method are as follows:

Linear Regression:

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error
Regression (Backward)	RATE_INTE...		0.005952
Regression (Stepwise)	RATE_INTE...		0.00603
Regression (Forward)	RATE_INTE...		0.00603

Without transformation

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error
Regression (Stepwise)_T	LOG_RATE...	Transforme...	0.002453
Regression (Forward)_T	LOG_RATE...	Transforme...	0.002453
Regression (Backward)_T	LOG_RATE...	Transforme...	0.002745

With transformation

Neural Network after Regression:

Model Description	Target	Target Label	Selection Criterion: Valid: Average Squared Error
NN_MLP_6	LOG_RATE...	Transforme...	0.002286
NN_MLP_7	LOG_RATE...	Transforme...	0.002328
NN_MLP_5	LOG_RATE...	Transforme...	0.002348
NN_MLP_8	LOG_RATE...	Transforme...	0.002368
NN_MLP_4	LOG_RATE...	Transforme...	0.002409
NN_MLP_3	LOG_RATE...	Transforme...	0.002416
NN_GLM Default	LOG_RATE...	Transforme...	0.002491
AutoNeural	LOG_RATE...	Transforme...	0.002713
NN_MLP_2	LOG_RATE...	Transforme...	0.00272
NN_MLP_1	LOG_RATE...	Transforme...	0.002742
NN_GLM BackProp	LOG_RATE...	Transforme...	0.00344

With transformation

Regression Tree:

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error
Decision Tree (V,ChiSQ, Entro)	LOG_RATE...	Transforme...	0.001745
Decision Tree (V,Gini, Entro)	LOG_RATE...	Transforme...	0.002498
Decision Tree (F,Entro, Entro)	LOG_RATE...	Transforme...	0.002498

With transformation

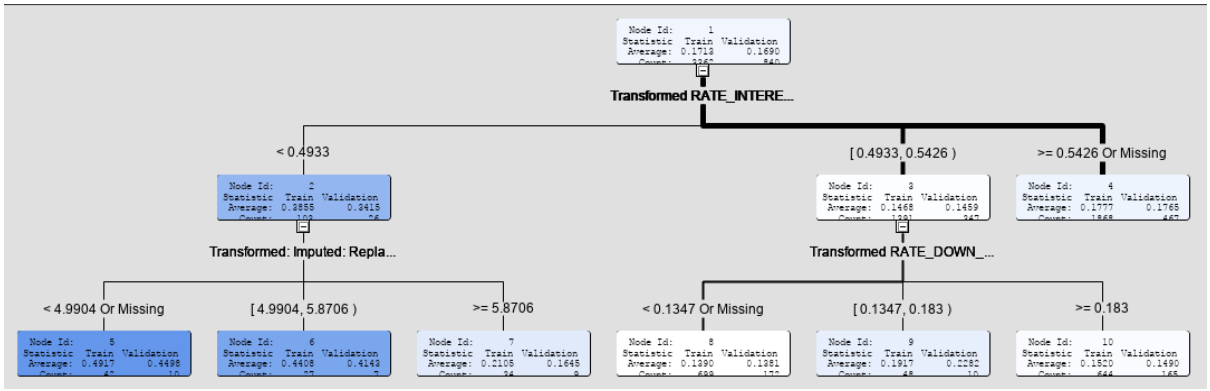
Neural Network:

Model Description	Target	Target Label	Selection Criterion: Valid: Average Squared Error
NN_MLP_2	LOG_RATE...	Transforme...	0.001838
NN_MLP_9	LOG_RATE...	Transforme...	0.002021
NN_MLP_17	LOG_RATE...	Transforme...	0.002066
NN_MLP_10	LOG_RATE...	Transforme...	0.002113
NN_MLP_16	LOG_RATE...	Transforme...	0.002569

With transformation

Interpretation and Recommendations:

In overall model comparison, Regression Tree using Variance, ProbChisq and Entropy as model criterion for Interval, Nominal and Ordinal target respectively, performed the best among all the models.



Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LOG_DAYS_DECISION	Transformed DAYS_DECISI...	0	3	1.0000	1.0000	1.0000
LOG_IMP_REP_DAYS_FIRST_DUE	Transformed: Imputed: Repl...	0	3	0.9991	0.9992	1.0001
LOG_IMP_REP_DAYS_TERMINATION	Transformed: Imputed: Repl...	0	3	0.9690	0.9646	0.9955
LOG_RATE_INTEREST_PRIVILEGED	Transformed RATE_INTERE...	1	1	0.9685	0.9145	0.9442
LOG_RATE_DOWN_PAYMENT	Transformed RATE_DOWN_...	1	1	0.7789	0.7437	0.9549
NAME_CLIENT_TYPE		0	1	0.7556	0.6998	0.9262
LOG_IMP_REP_DAYS_LAST_DUE_1ST_VE	Transformed: Imputed: Repl...	1	1	0.5264	0.6412	1.2181
LOG_PAYTOINC	Transformed PAYTOINC	0	1	0.3579	0.4291	1.1988
LOG_SELLERPLACE_AREA	Transformed SELLERPLAC...	0	1	0.3509	0.4206	1.1988

Important factors

According to the output obtained, we recommend that apart from studying and analyzing customers' demographic information, Home Credit could also focus on the above variables, for instance, the number of days used to decision whether or not to lend money to this customer, number of days past due from the previous loan, Pay-to_Income ratio as they imposed significant impact to the model, so that Home Credit could maximize profit with the most suitable interest rate level assigned to each customer.