FOUNDATION FOR ORGANISATIONAL RESEARCH AND EDUCATION

NEW DELHI

Academic Session 2023-2025

Customer Classification (Loan Data) on the basis of Cluster data by using cross validation and Ensemble Learning

Machine Learning for Managers

FMG 32 Section A

Submitted to:

Prof. Amarnath Mitra

Submitted by:

321035 - Prityush Agarwal

Table of Contents

1. Project Objectives

→ The first objective is to classify the consumer (loan) data of the bank into segments or clusters using cross-validation.

→ The second objective is to classify the consumer (loan) data of the bank into segments or clusters using ensemble methods.

→ The third objective is to determine the appropriate classification model.

→ The fourth objective is to identify significant variables or features and their thresholds for classification.

2. Data Description

2.1. Data Source, Size and Shape

2.1.1. Link of the data: https://www.kaggle.com/datasets/mrferozi/loan-data-for-dummy-bank

2.1.2. The size of data is 30 MB.

2.1.3. Dimension of Data

- Number of Variables: The number of variables in the csv file is 30.
- Number of records: The number of records in the csv file is 1,06,485 (excluding naming column).

2.2. Description of Variables

2.2.1 Index variables: id – gives the loan a unique identification (year, issue_d and final_d won't be used for evaluation purpose another variable term is being used to gauge how much time it took to repay the loan).

2.2.2. Variables having categorical or non-categorical variables

2.2.2.1 Variables or Features having Nominal Categories:

→ cluster: This is the outcome variable. The results of the outcome variable I got from the previous project where we did unsupervised learning using K-means clustering.
→home_ownership - home ownership status provided by the borrower during registration
→term – Term of the loan
→application_type – Explains the status whether the account is individual or joint
→purpose – This variable tells the purpose why the loan was taken
→loan_condition – This variable tells the status of the loan whether the loan is good or bad
→region – The region the loan was taken from

2.2.2.2 Variables or Features having Ordinal Categories:

→income_category – This variable tells the bracket under which the person earns
→interest_payments – This variable tells whether the interest payments on the loan is low or high
→grade – This variable tells the assigned grade of the loan

2.2.2.3. Non-Categorical Variables:

→emp_length_int – The number of years the person is employed
→annual_inc – This variable tells the annual income the person earns
→loan_amount – The variable tells the amount of loan that has been taken by the person

→interest_rate – The variable tells the interest rate at which the loan needs to be paid
→dti - A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower's self-reported monthly income.
→total_pymnt – This variable explains the total payment done against the loan
→total_rec_prncp – This variable explains the total received principal gotten from the loan
→recoveries – This variable explains the recoveries made from the bad loan
→installment – This variable explains the instalment made against the loan

2.3. Descriptive Statistics

2.3.1. Descriptive Statistics of Outcome Categorical Variables

It provides the statistics of cluster variable (categorical variable) by giving frequency as well as relative frequency (in %).

Row ID	count	Relative Frequency (in %)
cluster_0	77618	72.891
cluster_1	1183	1.111
cluster_2	27684	25.998

2.3.2. Descriptive Statistics of Input Categorical Variables

2.3.2.1.It provides the statistics of input variable (categorical variable) by giving frequency (count) as well as relative frequency (in %).

home_ownership

Row ID	count	Relative Frequency (in %)
MORTGAGE	53079	49.846
NONE	5	0.005
OTHER	24	0.023
OWN	10502	9.862
RENT	42875	40.264

Term

Row ID	count	Relative Frequency in %
36 months	74473	69.938
60 months	32012	30.062

application_type

Row ID	count	Relative Frequency in %
INDIVIDUAL	106427	99.946
JOINT	58	0.054

Purpose

Row ID	count	Relative Frequency in %
car	1043	0.979
credit_card	24716	23.211
debt_consolidation	63092	59.25
educational	50	0.047
home_improvement	6193	5.816
house	433	0.407
major_purchase	2023	1.9
medical	1007	0.946
moving	661	0.621
other	5164	4.85
renewable_energy	67	0.063
small_business	1188	1.116
vacation	585	0.549
wedding	263	0.247

loan_condition

Row ID	count	Relative Frequency in %
Bad Loan	8084	7.592
Good Loan	98401	92.408

income_category

Row ID	I count	D Relative Frequency in %
High	2020	1.897
Low	87400	82.077
Medium	17065	16.026

interest_payments

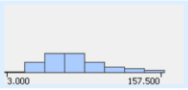
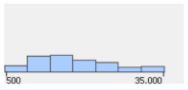
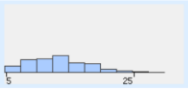
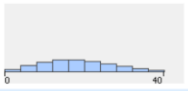
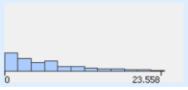
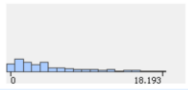
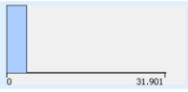
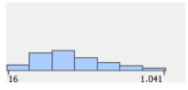
Row ID	I count	D Relative Frequency in %
High	50610	47.528
Low	55875	52.472

Grade

Row ID	I count	D Relative Frequency in %
A	17753	16.672
B	30642	28.776
C	29537	27.738
D	16703	15.686
E	8487	7.97
F	2729	2.563
G	634	0.595

2.3.3. Descriptive Statistics: Non-Categorical Variables

2.3.3.1. Measures of Central Tendency and Dispersion

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
emp_length_int	0.5	6.0538	?	10	3.5077	-0.2089	-1.4608	0	0	0	
annual_inc	3,000	67,641.7902	?	157,500	29,244.9179	0.7392	0.1725	0	0	0	
loan_amount	500	14,786.9029	?	35,000	8,427.6836	0.6771	-0.2624	0	0	0	
interest_rate	5.32	13.1438	?	25.28	4.2443	0.3358	-0.3667	0	0	0	
dti	0.0	18.1547	?	39.99	8.3112	0.2209	-0.5497	0	0	0	
total_pymnt	0.0	6,247.0389	?	23,558.04	5,463.9378	1.1313	0.6196	0	0	0	
total_rec_prncp	0.0	4,447.5	?	18,192.7	4,191.2341	1.2866	1.0144	0	0	0	
recoveries	0.0	45.972	?	31,900.52	415.898	19.6333	766.3308	0	0	0	
installment	15.67	418.0371	?	1,040.76	212.94	0.6264	-0.1607	0	0	0	

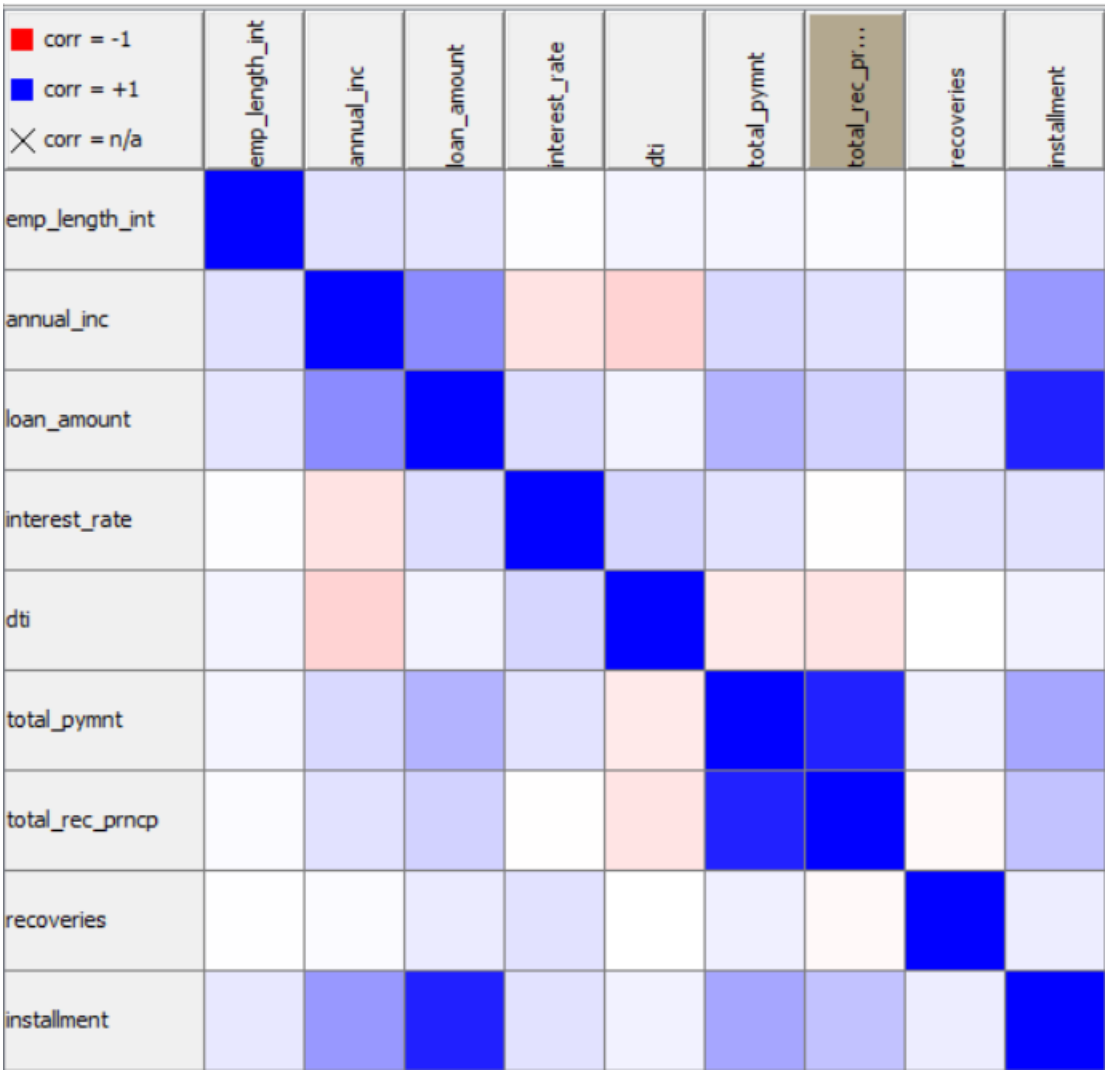
Name ↓	Type	# Missi...	Minimum	Maximum	25% Quantile	50% Quantile (Median)	75% Quantile	Mean	Mean Absolut...	Standard Devi...	Variance	Skewness	Kurtosis
emp_length_int	Number (double)	0	0.5	10	3	6.05	10	6.054	3.115	3.508	12.304	-0.209	10.711
annual_inc	Number (double)	0	3,000	157,500	45,000	65,000	85,000	67,641.79	23,007.866	29,244.918	855,265,222.1...	0.739	-1.265
loan_amount	Number (double)	0	500	35,000	8,000	13,000	20,000	14,786.903	6,867.874	8,427.684	71,025,850.905	0.677	1.924
interest_rate	Number (double)	0	5.32	25.28	9.99	12.99	15.88	13.144	3.417	4.244	18.014	0.336	2.689
dti	Number (double)	0	0	39.99	11.9	17.71	23.97	18.155	6.819	8.311	69.076	0.221	4.031
total_pymnt	Number (double)	0	0	23,558.04	1,918.7	4,868.8	8,905.36	6,247.039	4,282.876	5,463.938	29,854,616.597	1.131	-4.543
total_rec_prncp	Number (double)	0	0	18,192.7	1,203.92	3,200	6,140.795	4,447.5	3,235.237	4,191.234	17,566,443.385	1.287	-7.438
recoveries	Number (double)	0	0	31,900.52	0	0	0	45.972	89.481	415.898	172,971.14	19.633	-5,618.971
installment	Number (double)	0	15.67	1,040.76	261.65	382.87	547.84	418.037	171.352	212.94	45,343.458	0.626	1.178

2.3.3.2.Correlation Statistics (using Test of Correlation)

Row ID	First column name	Second column name	Correlation value	p value	Degrees of freedom
Row0	emp_length_int	annual_inc	0.11911012653120...	0.0	106483
Row1	emp_length_int	loan_amount	0.10171130496082...	0.0	106483
Row2	emp_length_int	interest_rate	0.00675005509605...	0.027617319551575648	106483
Row3	emp_length_int	dti	0.0445547463056736	0.0	106483
Row4	emp_length_int	total_pymnt	0.03741256187103...	0.0	106483
Row5	emp_length_int	total_rec_prncp	0.01743498543779...	1.27220303181019 29E-8	106483
Row6	emp_length_int	recoveries	0.00435276338086...	0.1554954763978429	106483
Row7	emp_length_int	installment	0.08826361219589...	0.0	106483
Row8	annual_inc	loan_amount	0.45639069673328...	0.0	106483
Row9	annual_inc	interest_rate	-0.10890036507563...	3.130798239529113E-278	106483
Row10	annual_inc	dti	-0.1734362843491869	0.0	106483
Row11	annual_inc	total_pymnt	0.14931826101783...	0.0	106483
Row12	annual_inc	total_rec_prncp	0.11262203460731...	0.0	106483
Row13	annual_inc	recoveries	0.01508971149248...	8.466150787089788E-7	106483
Row14	annual_inc	installment	0.4045213669226771	0.0	106483
Row15	loan_amount	interest_rate	0.13267165502297...	0.0	106483
Row16	loan_amount	dti	0.04671651876506...	0.0	106483
Row17	loan_amount	total_pymnt	0.29787380746640...	0.0	106483
Row18	loan_amount	total_rec_prncp	0.1768913105828052	0.0	106483
Row19	loan_amount	recoveries	0.07681736222226...	0.0	106483
Row20	loan_amount	installment	0.8727387314553934	0.0	106483
Row21	interest_rate	dti	0.1625930381940522	0.0	106483
Row22	interest_rate	total_pymnt	0.10899088797054...	0.0	106483
Row23	interest_rate	total_rec_prncp	-0.00264481012310...	0.3881117024827133	106483
Row24	interest_rate	recoveries	0.11364039053627...	0.0	106483
Row25	interest_rate	installment	0.11244970536063...	0.0	106483
Row26	dti	total_pymnt	-0.08140112488851...	5.679030877555581E-156	106483
Row27	dti	total_rec_prncp	-0.10606357586467...	5.918196961921463E-264	106483
Row28	dti	recoveries	5.401896161005 56...	0.8600796044771082	106483
Row29	dti	installment	0.04974955052132...	0.0	106483
Row30	total_pymnt	total_rec_prncp	0.8707573781921869	0.0	106483
Row31	total_pymnt	recoveries	0.06023104760486...	0.0	106483
Row32	total_pymnt	installment	0.3504988922257118	0.0	106483
Row33	total_rec_prncp	recoveries	-0.02247225026477...	2.2339385071952356E-13	106483
Row34	total_rec_prncp	installment	0.23863177319903...	0.0	106483
Row35	recoveries	installment	0.07088416491241...	0.0	106483

The variables are correlated if the value of p is less than 0.05. The variables that are not correlated are emp_lenght_int and recoveries, interest_rate and total_rec_prncp, and dti and recoveries because the p-value is less than 0.05.

Row ID	emp_length_int	annual_inc	loan_amount	interest_rate	dti	total_pymnt	total_rec_prncp	recoveries	installment
emp_length_int	1.0	0.11911012653120064	0.10171130496082201	0.006750055096057345	0.0445547463056736	0.037412561871036316	0.017434985437797604	0.004352763380868477	0.08826361219589225
annual_inc	0.11911012653120064	1.0	0.45639069673328797	-0.10890036507563082	-0.1734362843491869	0.14931826101783496	0.11262203460731779	0.01508971149248958	0.4045213669226771
loan_amount	0.10171130496082201	0.45639069673328797	1.0	0.13267165502297726	0.04671651876506626	0.29787380746640885	0.1768913105828052	0.07681736222226498	0.8727387314553934
interest_rate	0.006750055096057345	-0.10890036507563082	0.13267165502297726	1.0	0.1625930381940522	0.10899088797054506	-0.002644810231001395	0.11364039053627975	0.11244970536063416
dti	0.0445547463056736	-0.1734362843491869	0.04671651876506626	0.1625930381940522	1.0	-0.08140112488851041	-0.10606357586467281	5.401896161005566E-4	0.04974955052132508
total_pymnt	0.037412561871036316	0.14931826101783496	0.29787380746640885	0.10899088797054506	-0.08140112488851041	1.0	0.8707573781921869	0.06023104760486053	0.3504988922257118
total_rec_prncp	0.017434985437797604	0.11262203460731779	0.1768913105828052	-0.002644810231001395	-0.10606357586467281	0.8707573781921869	1.0	-0.022472250264771516	0.23863177319903192
recoveries	0.004352763380868477	0.01508971149248958	0.07681736222226498	0.11364039053627975	5.401896161005566E-4	0.06023104760486053	-0.022472250264771516	1.0	0.07088416491241371
installment	0.08826361219589225	0.4045213669226771	0.8727387314553934	0.11244970536063416	0.04974955052132508	0.3504988922257118	0.23863177319903192	0.07088416491241371	1.0



3. Analysis of Data

3.1. Data Pre-Processing

3.1.1. Missing Data Statistics and Treatment

3.1.1.1. Missing Data Statistics: 0

3.1.1.2. Missing Data Treatment: 0

3.1.1.2.1. Removal of Records with More Than 50% Missing Data: None

3.1.1.3. Missing Data Statistics of categorical Variables: 0

3.1.1.3.1. Missing Data Treatment: Categorical Variables or Features: 0

3.1.1.3.1.1. Removal of Variables or Features with More Than 50% Missing Data: None

3.1.1.4. Missing Data Statistics of non-categorical Variables: 0

3.1.1.4.1. Missing Data Treatment of non-categorical Variables: 0

3.1.1.4.1.1. Removal of Variables or Features with More Than 50% Missing Data: None

3.1.2. Numerical Encoding of Categorical Variables

In this case, category to number node will be used to encode the categorical variables.

home_ownership
mortagage - 3, none - 5, other - 4, own - 2, rent – 1

Term
36 months – 1, 60 months – 2

application_type
Individual - 1, Joint - 2

Purpose
Credit card – 1, car – 2, small business – 3, other – 4, wedding – 5, debt consolidation – 6, home improvement – 7, major purchase – 8, medical – 9, moving – 10, vacation – 11, house – 12, renewable energy – 13, educational - 14

loan_condition
Good Loan – 1, Bad Loan - 2

Region
Munster – 1, Leinster – 2, Cannught – 3, Ulster – 4, Northern-Irl - 5

income_category
Low – 1, Medium – 2, High – 3

interest_payments
Low – 1, High – 2

Grade
B – 1, C – 2, A – 3, E – 4, F – 5, D – 6, G – 7

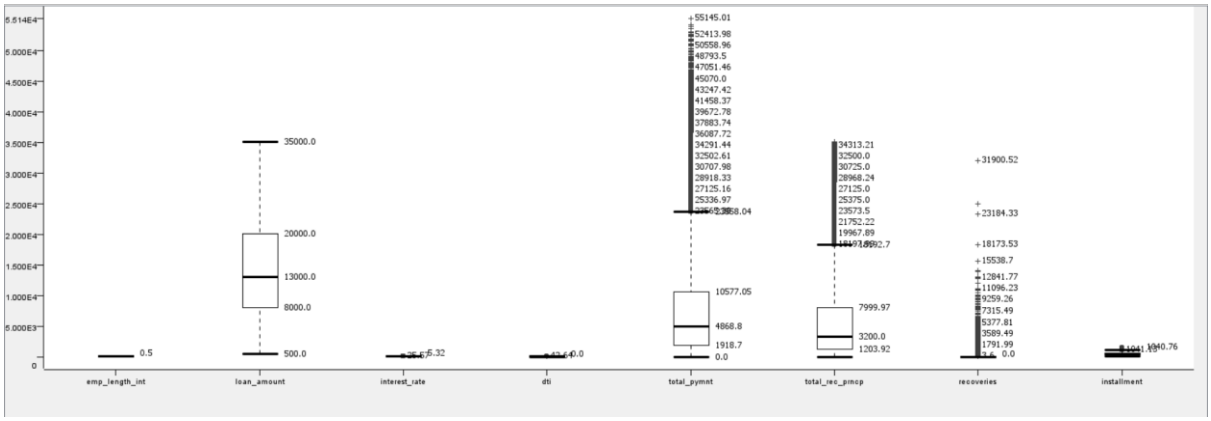
3.1.3. Outlier Statistics and Treatment

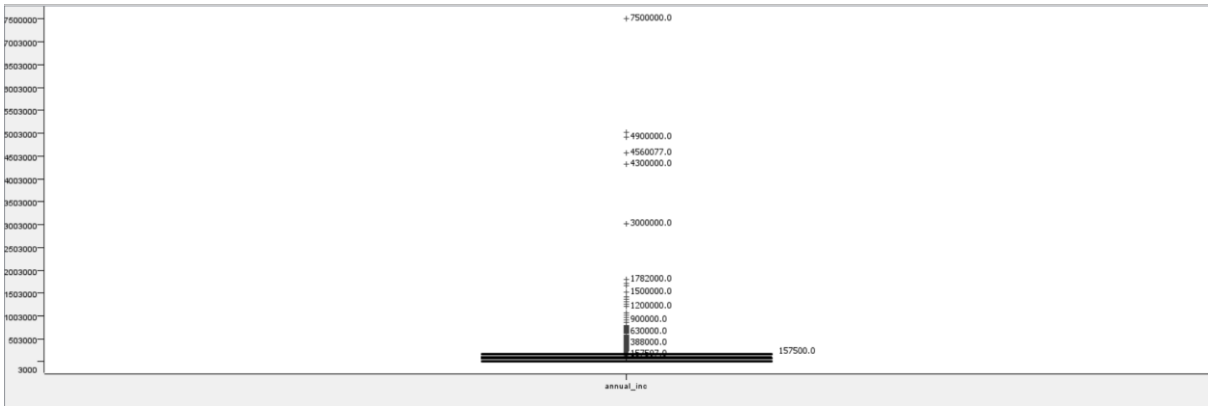
3.1.3.1.Outlier Statistics: Non-Categorical Variables

Row ID	emp_length_int	annual_inc	loan_amount	interest_rate	dti	total_pymnt	total_rec_prncp	recoveries	installment
Minimum	0.5	3,000	500	5.32	0	0	0	0	15.67
Smallest	0.5	3,000	500	5.32	0	0	0	0	15.67
Lower Quartile	3	45,000	8,000	9.99	11.9	1,918.7	1,203.92	0	261.65
Median	6.05	65,000	13,000	12.99	17.71	4,868.8	3,200	0	382.87
Upper Quartile	10	90,000	20,000	16.2	23.98	10,577.05	7,999.97	0	573.35
Largest	10	157,500	35,000	25.28	42.1	23,558.04	18,192.7	0	1,040.76
Maximum	10	7,500,000	35,000	28.99	104	55,145.01	35,000.03	31,900.52	1,445.46

3.1.3.2.Normalization using Min-Max Scaler

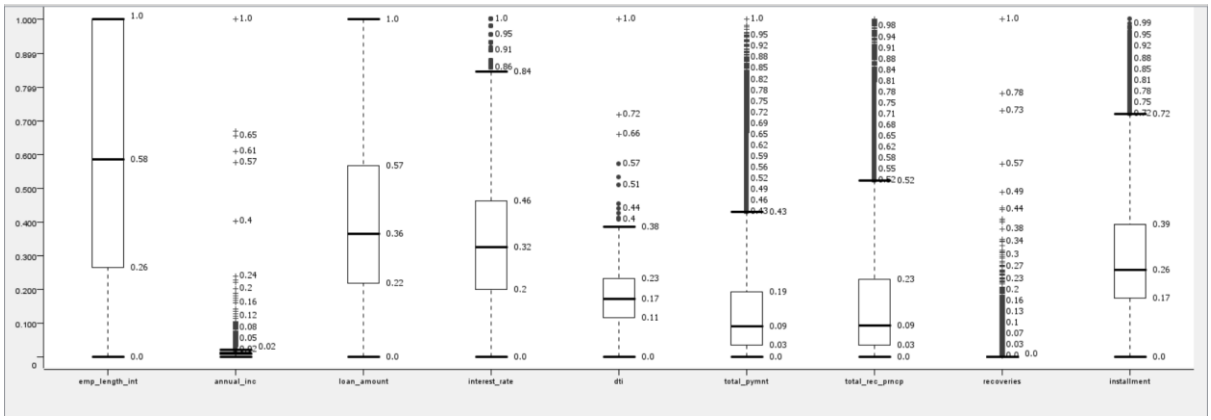
Before Normalization



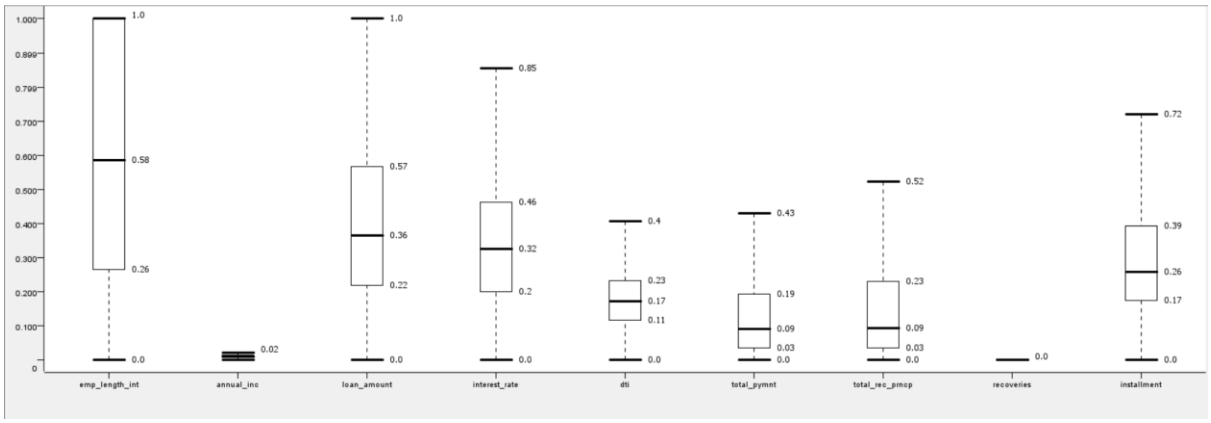


After Normalization

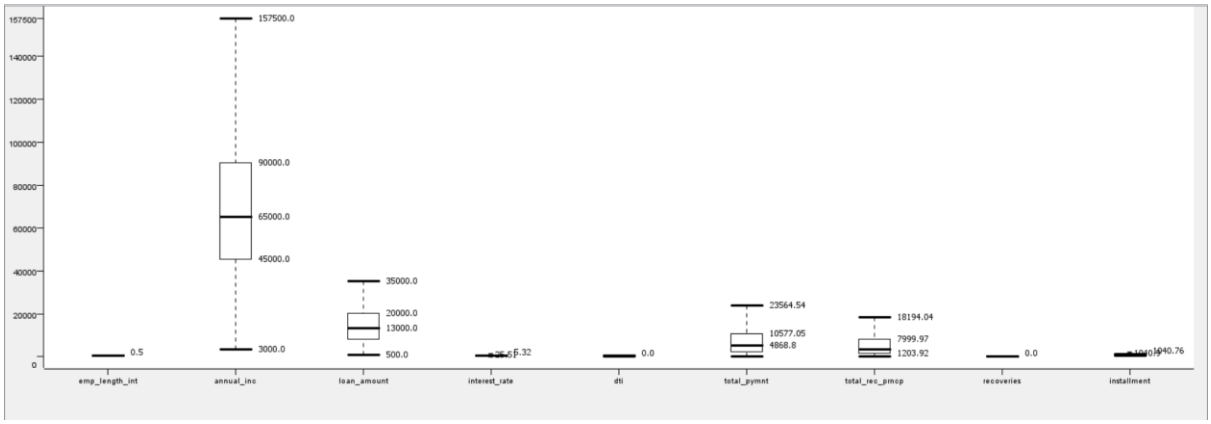
Min-Max Scaler Normalization (between 0 and 1) for variables: annual_inc, interest_rate, dti, total_pymnt, total_rec_prncp, recoveries, installment



Using numeric outliers' node to remove the outliers.



De-normalizing the data



3.1.4. Data Bifurcation

The bifurcation schema used is stratified sampling on the basis of outcome variable cluster variable with 80% (training data) and 20% (testing data).

3.2. Data Analysis

3.2.1. Cross-Validation using Decision Tree

Cross-validation using a decision tree involves splitting the dataset into k subsets, training the decision tree on k-1 subsets and validating on the remaining subset by repeating this process k times and averaging the results to assess the model's performance and generalization ability.

3.2.2. Cross-Validation using Other Methods

3.2.2.1. Logistic Regression

Cross-validation with logistic regression involves partitioning the dataset into training and validation sets, fitting the logistic regression model on the training data and evaluating its performance on the validation set. This process is repeated multiple times with different partitions to estimate the model's generalization performance and minimize overfitting.

3.2.2.2. K-Nearest Neighbours

Cross-validation with KNN entails splitting the dataset into training and validation sets, then iterating through different values of k (number of nearest neighbours) to find the optimal k value that minimizes error on the validation set. This process helps assess the KNN model's performance and its ability to generalize to new data.

3.2.3. Ensemble Method using Random Forest

Random forest is an ensemble learning method where multiple decision trees are trained on random subsets of the data and features. During prediction, each tree votes on the outcome and the final prediction is determined by the majority vote. This approach improves prediction accuracy and reduces overfitting compared to individual decision trees.

3.2.4. Ensemble Method using XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm that uses a gradient boosting framework. It sequentially builds multiple decision trees, each correcting the errors of the previous one. XGBoost incorporates regularization techniques to prevent overfitting and is known for its efficiency and effectiveness in various machine learning tasks.

3.2.1.1. Model Performance Evaluation of Cross-Validation using Decision Tree

Without pruning

Row ID	cluster_0	cluster_2	cluster_1
cluster_0	77540	75	0
cluster_2	81	27130	457
cluster_1	0	470	696

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_0	77540	81	28753	75	0.999	0.999	0.999	0.997	0.999	?	?
cluster_2	27130	545	78236	538	0.981	0.98	0.981	0.993	0.98	?	?
cluster_1	696	457	104826	470	0.597	0.604	0.597	0.996	0.6	?	?
Overall	?	?	?	?	?	?	?	?	?	0.99	0.975

With pruning

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_0	77557	98	28769	61	0.999	0.999	0.999	0.997	0.999	?	?
cluster_2	27028	318	78483	656	0.976	0.988	0.976	0.996	0.982	?	?
cluster_1	925	559	104743	258	0.782	0.623	0.782	0.995	0.694	?	?
Overall	?	?	?	?	?	?	?	?	?	0.991	0.977

Row ID	cluster_0	cluster_2	cluster_1
cluster_0	77557	61	0
cluster_2	97	27028	559
cluster_1	1	257	925

Cluster 0

- This cluster shows very high-performance metrics with high recall, precision, sensitivity and specificity indicating robust predictive power.
- The model correctly identifies a vast majority of cases evidenced by the high true positive count.
- There are very few false positives and false negatives, indicating minimal misclassification.

Cluster 1

- This cluster exhibits lower performance metrics compared to cluster 0 and cluster 2 with lower recall, precision and F-measure.
- The model correctly identifies a substantial portion of cases in this cluster but there are notable false positives and false negatives indicating some misclassification.
- Precision is relatively lower in this cluster compared to the others, suggesting a higher rate of false positives.

Cluster 2

- This cluster also exhibits strong performance metrics with high recall, precision, sensitivity and specificity though slightly lower than cluster 0.
- The model correctly identifies the vast majority of cases in this cluster with a high true positive count.
- There is a relatively low number of false positives and false negatives indicating good classification accuracy.

Comparative analysis of decision tree with and without pruning

- The performance metrics for both decision trees with and without pruning are identical across all clusters and overall accuracy is different where decision tree with pruning gives the higher accuracy.
- Both models demonstrate high recall, precision, sensitivity, specificity and F-measure for cluster 0 and cluster 2.
- For cluster 1, both models show lower performance metrics compared to cluster 0 and cluster 2, with lower recall, precision and F-measure suggesting a greater degree of misclassification or uncertainty in predicting loan condition for this cluster.
- Overall, both decision tree models achieve high accuracy and Cohen's Kappa coefficient indicating strong predictive performance across clusters.

3.2.2.1.Model Performance Evaluation of Cross-Validation using Other Methods

3.2.2.1.1. Logistic Regression

Row ID	Logit	Variable	Coeff.	Std. Err.	z-score	P>|z|
Row1	cluster_0	home_ownership=MORTGAGE	11.325	945,077.669	0	1
Row2	cluster_0	home_ownership=NONE	6.956	110,278,11...	0	1
Row3	cluster_0	home_ownership=OTHER	15.652	97,277,945.94	0	1
Row4	cluster_0	home_ownership=OWN	11.243	945,139.061	0	1
Row5	cluster_0	home_ownership=RENT	11.492	945,175.141	0	1
Row6	cluster_0	income_category=Low	102.304	70,921.478	0.001	0.999
Row7	cluster_0	income_category=Medium	44.386	120,190.177	0	1
Row8	cluster_0	term=60 months	0.318	0.233	1.363	0.173
Row9	cluster_0	application_type=JOINT	-0.418	149,278,78...	-0	1
Row10	cluster_0	purpose=credit_card	0.072	1.028	0.07	0.944
Row11	cluster_0	purpose=debt_consolidation	-0.235	1.018	-0.23	0.818
Row12	cluster_0	purpose=educational	1.987	63,349,601....	0	1
Row13	cluster_0	purpose=home_improvement	-0.306	1.053	-0.291	0.771
Row14	cluster_0	purpose=house	0.833	1.868	0.446	0.656
Row15	cluster_0	purpose=major_purchase	-1.439	1.187	-1.212	0.225
Row16	cluster_0	purpose=medical	-1.675	1.348	-1.243	0.214
Row17	cluster_0	purpose=moving	0.606	1.401	0.432	0.666
Row18	cluster_0	purpose=other	-0.546	1.08	-0.506	0.613
Row19	cluster_0	purpose=renewable_energy	2.034	79,123,021....	0	1
Row20	cluster_0	purpose=small_business	-0.646	1.202	-0.538	0.591
Row21	cluster_0	purpose=vacation	0.414	1.572	0.263	0.792
Row22	cluster_0	purpose=wedding	-23.581	30,340.643	-0.001	0.999
Row23	cluster_0	interest_payments=Low	-0.03	0.376	-0.08	0.936
Row24	cluster_0	loan_condition=Good Loan	-0.582	0.406	-1.432	0.152
Row25	cluster_0	grade=B	1.17	0.356	3.289	0.001
Row26	cluster_0	grade=C	2.432	0.53	4.587	0
Row27	cluster_0	grade=D	3.313	0.74	4.479	0
Row28	cluster_0	grade=E	4.865	0.926	5.253	0
Row29	cluster_0	grade=F	4.387	1.142	3.84	0
Row30	cluster_0	grade=G	2.832	1.044	2.714	0.007
Row31	cluster_0	region=cannught	0.4	0.279	1.43	0.153
Row32	cluster_0	region=leinster	0.562	0.25	2.244	0.025
Row33	cluster_0	region=munster	0.745	0.301	2.478	0.013
Row34	cluster_0	region=ulster	0.005	0.251	0.018	0.985
Row35	cluster_0	emp_length_int	-0.047	0.243	-0.192	0.848
Row36	cluster_0	annual_inc	-218.757	168,765.335	-0.001	0.999
Row37	cluster_0	loan_amount	-14.401	0.756	-19.045	0
Row38	cluster_0	interest_rate	-7.179	1.368	-5.246	0
Row39	cluster_0	dti	3.455	0.495	6.984	0
Row40	cluster_0	total_pymnt	-0.482	0.678	-0.71	0.477
Row41	cluster_0	total_rec_prncp	-3.654	0.723	-5.056	0
Row42	cluster_0	recoveries	3.455	6.415	0.539	0.59
Row43	cluster_0	installment	-0.632	0.643	-0.982	0.326
Row44	cluster_0	Constant	56.705	803,474.946	0	1
Row45	cluster_2	home_ownership=MORTGAGE	-17.55	470,482.832	-0	1

Row46	cluster_2	home_ownership=NONE	-6.952	110,279,12...	-0	1
Row47	cluster_2	home_ownership=OTHER	-15.655	97,316,168....	-0	1
Row48	cluster_2	home_ownership=OWN	-17.906	470,482.832	-0	1
Row49	cluster_2	home_ownership=RENT	-17.652	470,482.832	-0	1
Row50	cluster_2	income_category=Low	13.832	38,449.762	0	1
Row51	cluster_2	income_category=Medium	46.404	72,025.733	0.001	0.999
Row52	cluster_2	term=60 months	-0.214	0.131	-1.637	0.102
Row53	cluster_2	application_type=JOINT	0.418	149,278,78...	0	1
Row54	cluster_2	purpose=credit_card	-0.528	0.637	-0.83	0.406
Row55	cluster_2	purpose=debt_consolidation	-0.42	0.63	-0.666	0.506
Row56	cluster_2	purpose=educational	-1.984	63,349,601....	-0	1
Row57	cluster_2	purpose=home_improvement	-0.395	0.64	-0.616	0.538
Row58	cluster_2	purpose=house	-0.056	0.886	-0.063	0.95
Row59	cluster_2	purpose=major_purchase	-0.979	0.723	-1.354	0.176
Row60	cluster_2	purpose=medical	-1.26	0.81	-1.555	0.12
Row61	cluster_2	purpose=moving	-0.516	0.837	-0.616	0.538
Row62	cluster_2	purpose=other	-0.763	0.661	-1.154	0.249
Row63	cluster_2	purpose=renewable_energy	-2.034	79,123,022....	-0	1
Row64	cluster_2	purpose=small_business	-0.301	0.692	-0.435	0.663
Row65	cluster_2	purpose=vacation	-0.399	1.055	-0.379	0.705
Row66	cluster_2	purpose=wedding	-22.42	30,340.643	-0.001	0.999
Row67	cluster_2	interest_payments=Low	-0.148	0.228	-0.651	0.515
Row68	cluster_2	loan_condition=Good Loan	0.074	0.246	0.301	0.763
Row69	cluster_2	grade=B	0.093	0.195	0.478	0.632
Row70	cluster_2	grade=C	0.225	0.302	0.746	0.455
Row71	cluster_2	grade=D	-0.01	0.422	-0.024	0.981
Row72	cluster_2	grade=E	0.204	0.521	0.391	0.696
Row73	cluster_2	grade=F	0.03	0.666	0.045	0.964
Row74	cluster_2	grade=G	0.329	0.633	0.519	0.604
Row75	cluster_2	region=cannught	0.045	0.172	0.26	0.795
Row76	cluster_2	region=leinster	0.235	0.142	1.656	0.098
Row77	cluster_2	region=munster	0.38	0.167	2.27	0.023
Row78	cluster_2	region=ulster	0.339	0.139	2.431	0.015
Row79	cluster_2	emp_length_int	-0.184	0.137	-1.34	0.18
Row80	cluster_2	annual_inc	223.279	168,765.334	0.001	0.999
Row81	cluster_2	loan_amount	-1.182	0.236	-5.016	0
Row82	cluster_2	interest_rate	-0.029	0.783	-0.037	0.971
Row83	cluster_2	dti	2.779	0.326	8.52	0
Row84	cluster_2	total_pymnt	0.31	0.419	0.74	0.459
Row85	cluster_2	total_rec_prncp	-0.658	0.445	-1.478	0.139
Row86	cluster_2	recoveries	-1.519	3.699	-0.411	0.681
Row87	cluster_2	installment	0.313	0.277	1.129	0.259
Row88	cluster_2	Constant	-75.848	111,026.048	-0.001	0.999

Cluster 0

- **Variables with Positive Impact**
  - Home ownership status (except NONE) positively influences the prediction of loan condition.
  - Higher grades (B, C, D, E, F, G) positively impact the prediction of loan condition.
  - Higher debt-to-income ratio (dti) positively influences the prediction of loan condition.
- **Variables with Negative Impact**
  - Lower loan amounts negatively impact the prediction of loan condition.
- **Insights**
  - Home ownership, grade and debt-to-income ratio are significant predictors of loan condition in Cluster 0.
  - Borrowers with higher grades, higher debt-to-income ratios and specific home ownership statuses are more likely to be classified as having a good loan condition in this cluster.

Cluster 1

- **Variables with Positive Impact**
  - Higher loan amounts positively influence the prediction of loan condition.
  - Higher interest rates positively influence the prediction of loan condition.
- **Variables with Negative Impact**
  - Home ownership status (except NONE) negatively influences the prediction of loan condition.
  - Lower grades (B, C, D, E, F, G) negatively impact the prediction of loan condition.
  - Higher debt-to-income ratio (dti) negatively influences the prediction of loan condition.
- **Insights**
  - Home ownership, grade, loan amount, interest rate, and debt-to-income ratio are significant predictors of loan condition in cluster 1.
  - Borrowers with specific home ownership statuses, lower grades, lower loan amounts, lower interest rates and lower debt-to-income ratios are more likely to be classified as having a good loan condition in this cluster.

Cluster 2

- **Variables with Positive Impact**
  - Home ownership status (except NONE) positively influences the prediction of loan condition.
  - Higher grades (B, C, G) positively impact the prediction of loan condition.
  - Higher annual income positively influences the prediction of loan condition.
  - Lower debt-to-income ratio (dti) positively influences the prediction of loan condition.
- **Variables with Negative Impact**
  - Lower loan amounts negatively impact the prediction of loan condition.
- **Insights**
  - Home ownership, grade annual income and debt-to-income ratio are significant predictors of loan condition in cluster 2.
  - Borrowers with specific home ownership statuses, higher grades, higher annual incomes and lower debt-to-income ratios are more likely to be classified as having a good loan condition in this cluster.

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_0	77462	139	28728	156	0.998	0.998	0.998	0.995	0.998	?	?
cluster_2	26972	390	78411	712	0.974	0.986	0.974	0.995	0.98	?	?
cluster_1	949	573	104729	234	0.802	0.624	0.802	0.995	0.702	?	?
Overall	?	?	?	?	?	?	?	?	?	0.99	0.974

Row ID	cluster_0	cluster_2	cluster_1
cluster_0	77462	156	0
cluster_2	139	26972	573
cluster_1	0	234	949

Cluster 0

- Cluster 0 exhibits extremely high-performance metrics with almost perfect recall, precision, sensitivity and specificity.
- The model effectively identifies true positives while minimizing false positives and false negatives indicating robust predictive power.
- Borrowers in this cluster are likely to have characteristics that make them highly reliable for loan repayment, resulting in minimal misclassifications.

Cluster 1

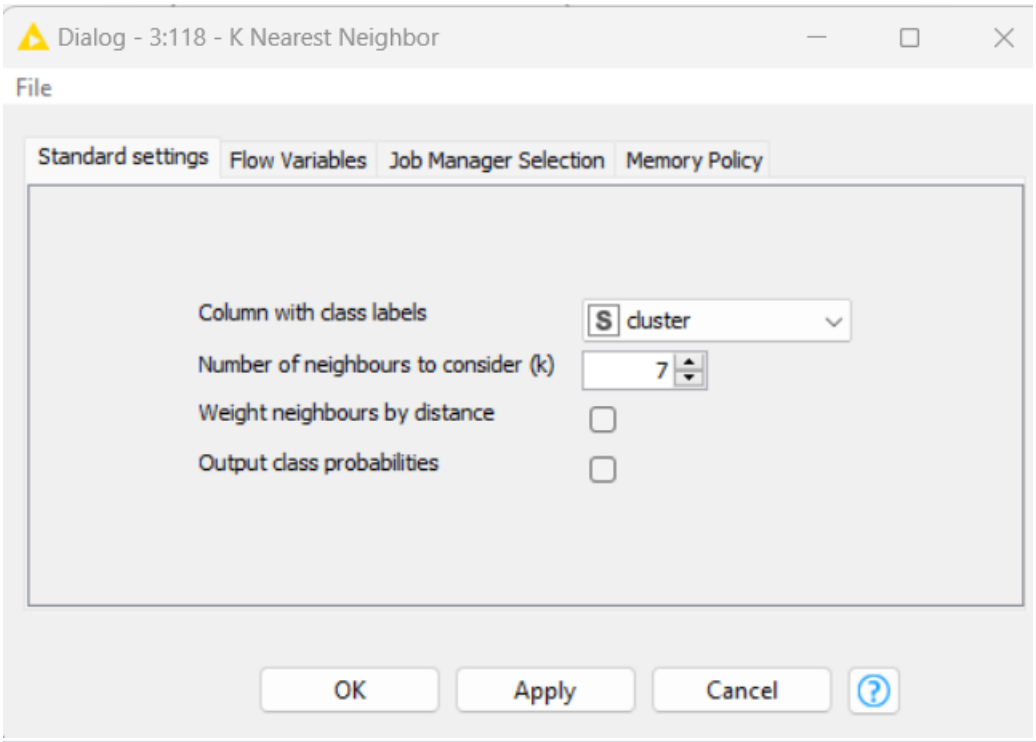
- Cluster 1 exhibits lower performance metrics compared to cluster 0 and cluster 2 with moderate recall, precision and F-measure.
- The model correctly identifies a significant portion of true positives but has a higher rate of false positives and false negatives compared to cluster 0 and cluster 2.
- Borrowers in this cluster may have characteristics associated with higher risk or variability in loan repayment behaviour leading to less reliable predictions compared to other clusters.

Cluster 2

- Cluster 2 demonstrates high performance metrics with strong recall, precision, sensitivity and specificity.
- The model effectively identifies true positives while maintaining a low false positive rate, suggesting reliable predictions for loan condition in this cluster.
- Borrowers in this cluster are likely to have characteristics associated with lower risk, contributing to the model's high accuracy.

3.2.2.1.2. K-Nearest Neighbours

K=7



Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_0	71232	26401	2466	6386	0.918	0.73	0.918	0.085	0.813	?	?
cluster_2	2354	6498	72303	25330	0.085	0.266	0.085	0.918	0.129	?	?
cluster_1	0	0	105302	1183	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.691	0.004

Row ID	cluster_0	cluster_2	cluster_1
cluster_0	71232	6386	0
cluster_2	25330	2354	0
cluster_1	1071	112	0

K=9

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_0	72951	27045	1822	4667	0.94	0.73	0.94	0.063	0.821	?	?
cluster_2	1730	4759	74042	25954	0.062	0.267	0.062	0.94	0.101	?	?
cluster_1	0	0	105302	1183	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.701	0.003

Row ID	cluster_0	cluster_2	cluster_1
cluster_0	72951	4667	0
cluster_2	25954	1730	0
cluster_1	1091	92	0

K=11

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_0	74203	27537	1330	3415	0.956	0.729	0.956	0.046	0.827	?	?
cluster_2	1276	3469	75332	26408	0.046	0.269	0.046	0.956	0.079	?	?
cluster_1	0	0	105302	1183	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.709	0.003

Row ID	cluster_0	cluster_2	cluster_1
cluster_0	74203	3415	0
cluster_2	26408	1276	0
cluster_1	1129	54	0

K=13

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_0	75116	27908	959	2502	0.968	0.729	0.968	0.033	0.832	?	?
cluster_2	920	2541	76260	26764	0.033	0.266	0.033	0.968	0.059	?	?
cluster_1	0	0	105302	1183	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.714	0.001

Row ID	cluster_0	cluster_2	cluster_1
cluster_0	75116	2502	0
cluster_2	26764	920	0
cluster_1	1144	39	0

\

K=15

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_0	75742	28146	721	1876	0.976	0.729	0.976	0.025	0.835	?	?
cluster_2	702	1895	76906	26982	0.025	0.27	0.025	0.976	0.046	?	?
cluster_1	0	0	105302	1183	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.718	0.001

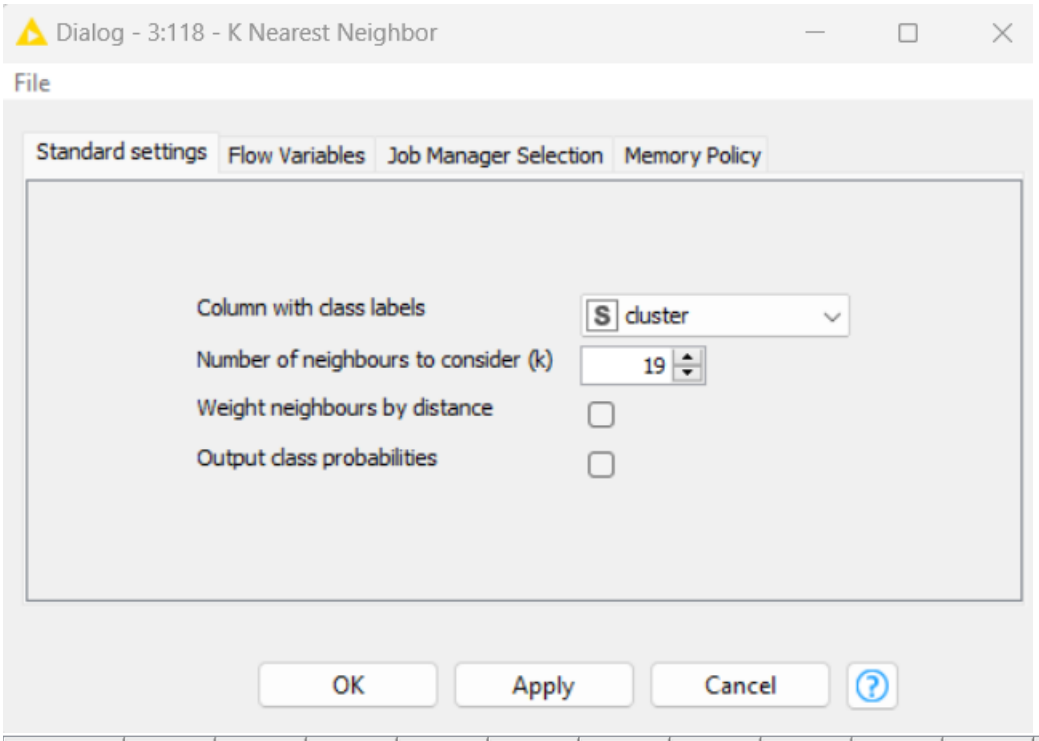
Row ID	cluster_0	cluster_2	cluster_1
cluster_0	75742	1876	0
cluster_2	26982	702	0
cluster_1	1164	19	0

K=17

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_0	76212	28317	550	1406	0.982	0.729	0.982	0.019	0.837	?	?
cluster_2	529	1427	77374	27155	0.019	0.27	0.019	0.982	0.036	?	?
cluster_1	0	0	105302	1183	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.721	0.001

Row ID	cluster_0	cluster_2	cluster_1
cluster_0	76212	1406	0
cluster_2	27155	529	0
cluster_1	1162	21	0

K=19



Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_0	76594	28448	419	1024	0.987	0.729	0.987	0.015	0.839	?	?
cluster_2	408	1035	77766	27276	0.015	0.283	0.015	0.987	0.028	?	?
cluster_1	0	0	105302	1183	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.723	0.002

Row ID	cluster_0	cluster_2	cluster_1
cluster_0	76594	1024	0
cluster_2	27276	408	0
cluster_1	1172	11	0

In KNN, the number of neighbours to be considered are from k=7 to 19. From the images, it is seen that as the number of k increases the accuracy also increases. For k=19, as the accuracy is the highest from all the other k's, this cluster will be considered.

Cluster 0

- Cluster 0 exhibits a high recall rate indicating that the majority of true positives are correctly identified.
- Precision is moderate, suggesting that while many identified instances are correct (true positives) there is also a significant number of false positives.
- Specificity is extremely low, indicating that the model has a high rate of falsely classifying negative instances as positive.
- The F-measure, which balances precision and recall, is relatively high indicating reasonable overall model performance.

Cluster 1

- Cluster 1 does not have any true positives predicted by the model, indicating that it may not have been effectively classified.
- The model seems to have classified all instances in this cluster as negative.
- Specificity is 100%, indicating that all negative instances were correctly classified as negative.

Cluster 2

- Cluster 2 exhibits a very low recall rate indicating that only a small portion of true positives are correctly identified.
- Precision is low, indicating that while some identified instances are correct (true positives), there is a significant number of false positives.
- Specificity is high, suggesting that the model has a low rate of falsely classifying negative instances as positive.
- The F-measure is very low, indicating poor overall model performance.

3.2.3.1. Model Performance Evaluation of Random Forest

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_0	15509	28	5746	14	0.999	0.998	0.999	0.995	0.999	?	?
cluster_2	5438	112	15648	99	0.982	0.98	0.982	0.993	0.981	?	?
cluster_1	139	71	20989	98	0.586	0.662	0.586	0.997	0.622	?	?
Overall	?	?	?	?	?	?	?	?	?	0.99	0.975

Row ID	cluster_0	cluster_2	cluster_1
cluster_0	15509	14	0
cluster_2	28	5438	71
cluster_1	0	98	139

Cluster 0

- Cluster 0 exhibits extremely high-performance metrics, with almost perfect recall, precision, sensitivity and specificity.
- The model effectively identifies true positives while minimizing false positives and false negatives, indicating robust predictive power.
- Borrowers in this cluster are likely to have characteristics that make them highly reliable for loan repayment, resulting in minimal misclassifications.

Cluster 1

- Cluster 1 exhibits lower performance metrics compared to cluster 0 and cluster 2, with moderate recall, precision, and F-measure.
- The model correctly identifies a significant portion of true positives but has a higher rate of false positives and false negatives compared to cluster 0 and cluster 2.
- Borrowers in this cluster may have characteristics associated with higher risk or variability in loan repayment behaviour, leading to less reliable predictions compared to other clusters.

Cluster 2

- Cluster 2 demonstrates high performance metrics, with strong recall, precision, sensitivity, and specificity.
- The model effectively identifies true positives while maintaining a low false positive rate, suggesting reliable predictions for loan condition in this cluster.
- Borrowers in this cluster are likely to have characteristics associated with lower risk, contributing to the model's high accuracy.

3.2.3.2. Model Performance Evaluation of XGBoost

Row ID	S Feature Name	D Weight	D Gain	D Cover	D Total Gain	D Total Cover
Row0	home_ownership=RENT	83	0.399	41.645	33.142	3,456.494
Row1	home_ownership=OWN	39	0.235	59.073	9.162	2,303.84
Row2	home_ownership=MO...	79	0.297	77.243	23.427	6,102.233
Row3	home_ownership=OT...	?	?	?	?	?
Row4	home_ownership=NONE	?	?	?	?	?
Row5	home_ownership=ANY	?	?	?	?	?
Row6	income_category=Low	118	282.615	3,269.337	33,348.619	385,781.766
Row7	income_category=Me...	22	271.116	2,369.045	5,964.544	52,118.997
Row8	income_category=High	40	158.954	5,236.025	6,358.173	209,441.018
Row9	term=36 months	67	0.35	82.638	23.418	5,536.768
Row10	term=60 months	?	?	?	?	?
Row11	application_type=IND...	?	?	?	?	?
Row12	application_type=JOINT	?	?	?	?	?
Row13	purpose=credit_card	70	0.361	35.689	25.237	2,498.241
Row14	purpose=car	5	0.288	92.605	1.441	463.024
Row15	purpose=small_business	12	0.405	93.907	4.858	1,126.888
Row16	purpose=other	42	0.328	80.147	13.795	3,366.171
Row17	purpose=wedding	?	?	?	?	?
Row18	purpose=debt_consoli...	69	0.244	84.882	16.833	5,856.878
Row19	purpose=home_impro...	78	0.269	49.186	20.97	3,836.504
Row20	purpose=major_purch...	5	0.173	48.84	0.867	244.199
Row21	purpose=medical	11	0.214	578.037	2.35	6,358.405
Row22	purpose=moving	3	0.439	183.44	1.317	550.32
Row23	purpose=vacation	?	?	?	?	?
Row24	purpose=house	3	1.301	119.199	3.902	357.596
Row25	purpose=renewable_...	?	?	?	?	?
Row26	purpose=educational	?	?	?	?	?
Row27	interest_payments=Low	13	0.311	74.778	4.038	972.112
Row28	interest_payments=High	?	?	?	?	?
Row29	loan_condition=Good ...	38	0.366	86.86	13.917	3,300.671
Row30	loan_condition=Bad L...	?	?	?	?	?
Row31	grade=B	61	0.297	65.546	18.147	3,998.326
Row32	grade=C	60	0.192	23.894	11.544	1,433.62
Row33	grade=A	18	0.247	11.527	4.448	207.489
Row34	grade=E	22	0.195	14.104	4.287	310.29
Row35	grade=F	4	0.47	285.769	1.882	1,143.075
Row36	grade=D	71	0.291	36.292	20.665	2,576.741
Row37	grade=G	?	?	?	?	?
Row38	emp_length_int	495	0.322	44.072	159.452	21,815.416
Row39	annual_inc	661	235.429	1,965.817	155,618.564	1,299,404.957
Row40	loan_amount	795	2.234	148.176	1,776.077	117,799.876
Row41	interest_rate	780	0.336	90.848	261.811	70,861.508
Row42	dti	1,343	0.434	139.178	582.82	186,915.446
Row43	total_pymnt	1,084	0.957	85.399	1,036.953	92,572.892
Row44	total_rec_prncp	914	1.04	125.538	950.93	114,742.144
Row45	recoveries	16	0.396	394.819	6.329	6,317.097
Row46	installment	860	1.003	135.764	862.912	116,756.687

1. **Home Ownership and Income**

   Factors related to home ownership status and income categories have significant importance in predicting loan conditions. Renting status, owning a home and income levels (low, medium, high) play crucial roles in assessing borrower risk.

2. **Loan Term and Purpose**

   Loan term duration and the purpose of the loan are important factors. Shorter loan terms and specific loan purposes such as debt consolidation, credit card usage and small business ventures carry weight in determining loan conditions.

3. **Interest Payments and Loan Grade**

   The type of interest payments (low or high) and the assigned loan grade (A, B, C, etc.) significantly influence loan conditions. Lower interest payments and higher loan grades typically indicate lower-risk borrowers.

4. **Employment Length and Annual Income**

   Borrowers' employment length and annual income are essential factors in assessing loan conditions. Longer employment length and higher annual income tend to correlate with lower risk.

5. **Financial Metrics**

   Financial metrics such as loan amount, interest rate, debt-to-income ratio, total payments, and instalment payments are critical in determining loan conditions. Higher loan amounts, interest rates, and debt-to-income ratios may signal increased borrower risk.

6. **Specific Loan Purposes**

   Certain loan purposes, such as debt consolidation, small business ventures, and credit card usage, carry more weight in predicting loan conditions. These purposes may indicate varying levels of financial stability or risk for borrowers.

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_0	15509	10	5764	14	0.999	0.999	0.999	0.998	0.999	?	?
cluster_2	5450	91	15669	87	0.984	0.984	0.984	0.994	0.984	?	?
cluster_1	160	77	20983	77	0.675	0.675	0.675	0.996	0.675	?	?
Overall	?	?	?	?	?	?	?	?	?	0.992	0.979

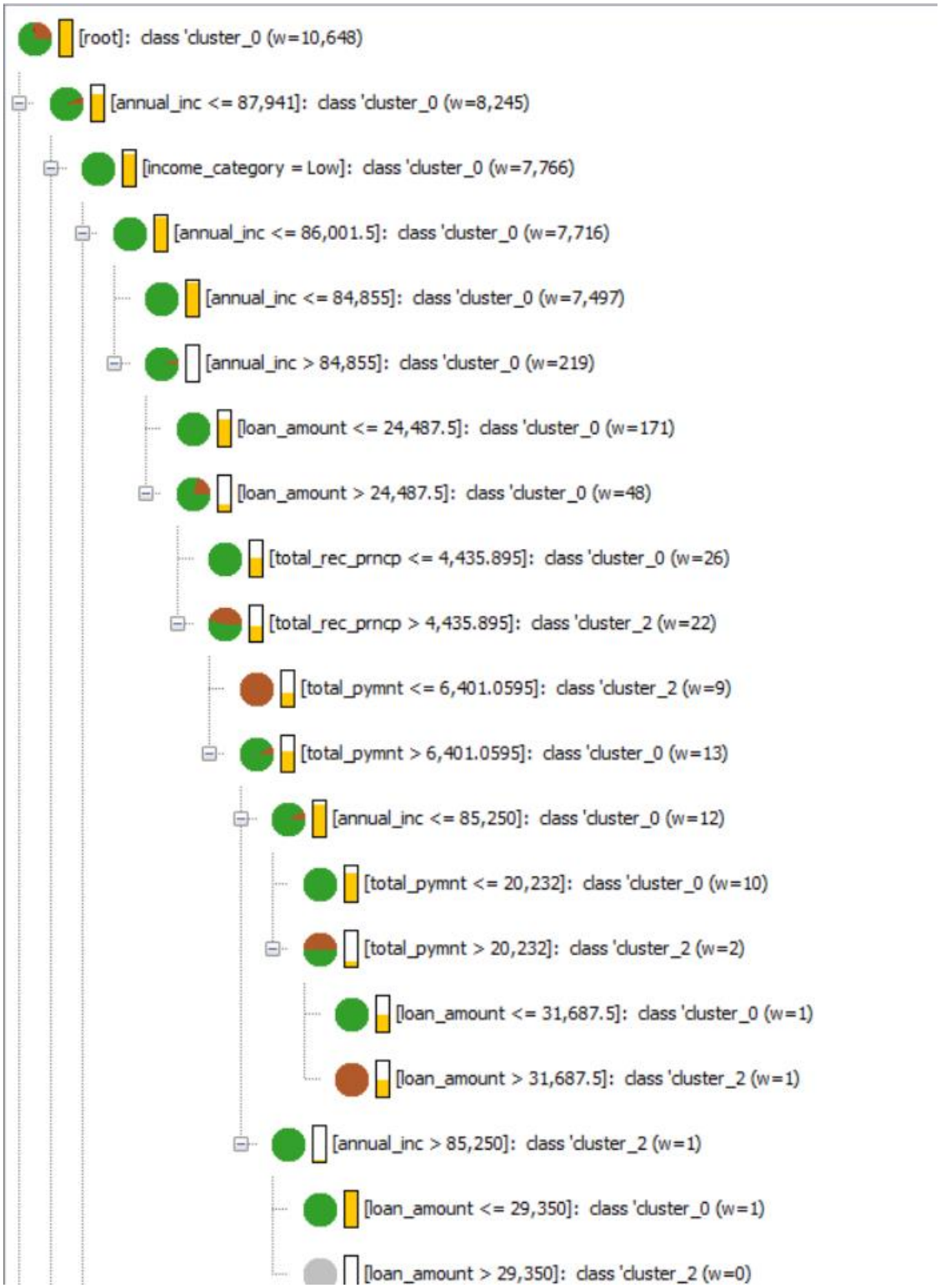
Row ID	cluster_0	cluster_2	cluster_1
cluster_0	15509	14	0
cluster_2	10	5450	77
cluster_1	0	77	160

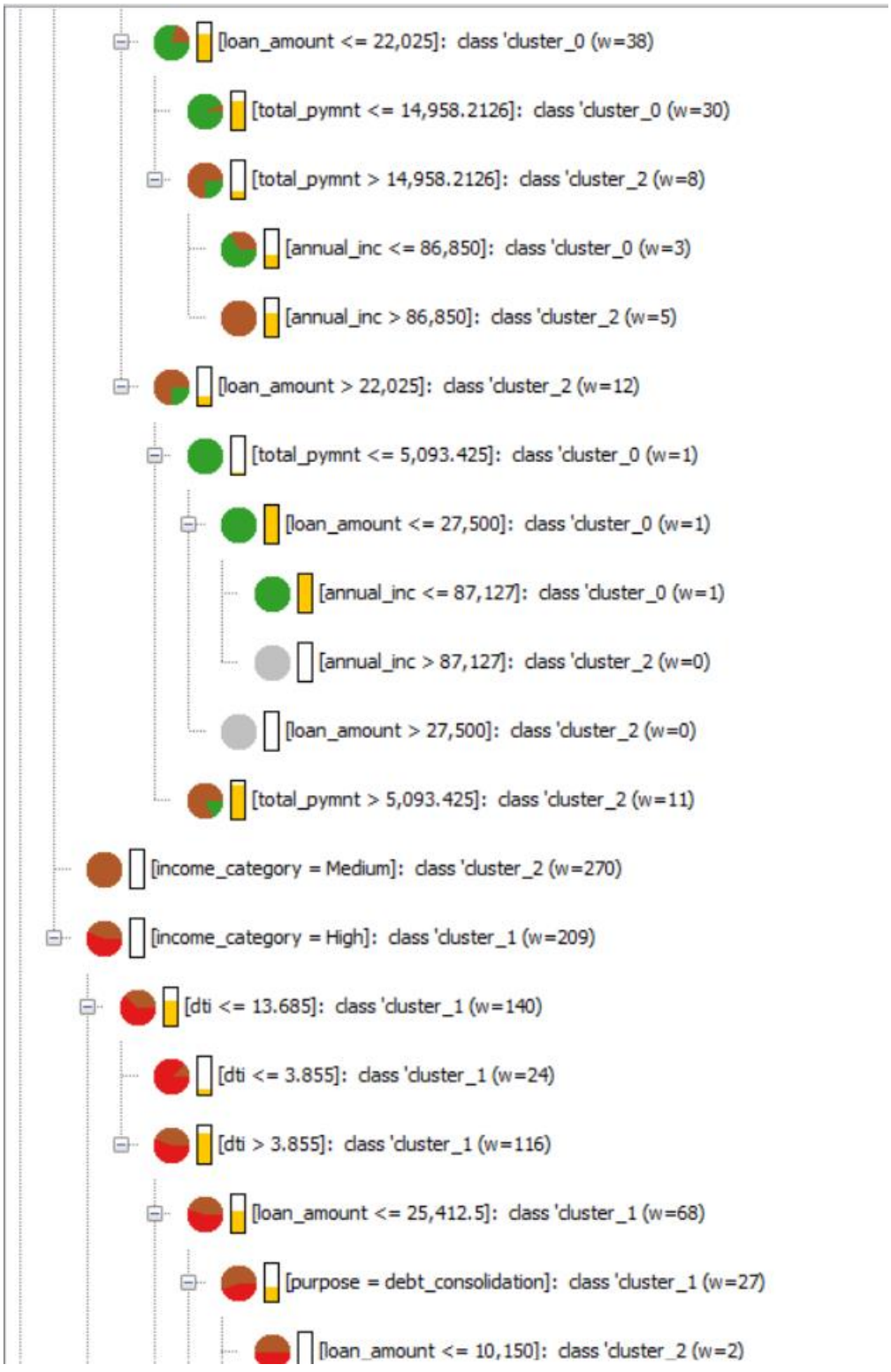
1. **Cluster 0**
    - This cluster has a high number of true positives (15509) and a very low false positive rate (10), indicating that the model is very good at correctly identifying positive cases while minimizing false alarms.
    - The recall, precision, sensitivity and F-measure are all very high, indicating excellent performance in correctly identifying positive cases and minimizing false positives.
    - The specificity is also high at 99.83%, indicating a low false positive rate.
2. **Cluster 1**
    - This cluster has a lower number of true positives (160) compared to the other clusters, and a higher false positive rate (77), indicating that the model's performance in identifying positive cases is not as strong in this cluster.
    - The recall, precision, sensitivity and F-measure are all moderate, indicating that the model's performance in correctly identifying positive cases and minimizing false positives is average compared to the other clusters.
    - The specificity is still high at 99.63% indicating a relatively low false positive rate.
3. **Cluster 2**
    - Similar to Cluster 0, this cluster has high values for true positives (5450) and a relatively low false positive rate (91), indicating good performance in correctly identifying positive cases while minimizing false alarms.
    - The recall, precision, sensitivity and F-measure are all high, suggesting that the model performs well in correctly identifying positive cases and minimizing false positives.
    - The specificity is also high at 99.42% indicating a low false positive rate.

3.3. Variable or Feature Analysis for Decision Tree

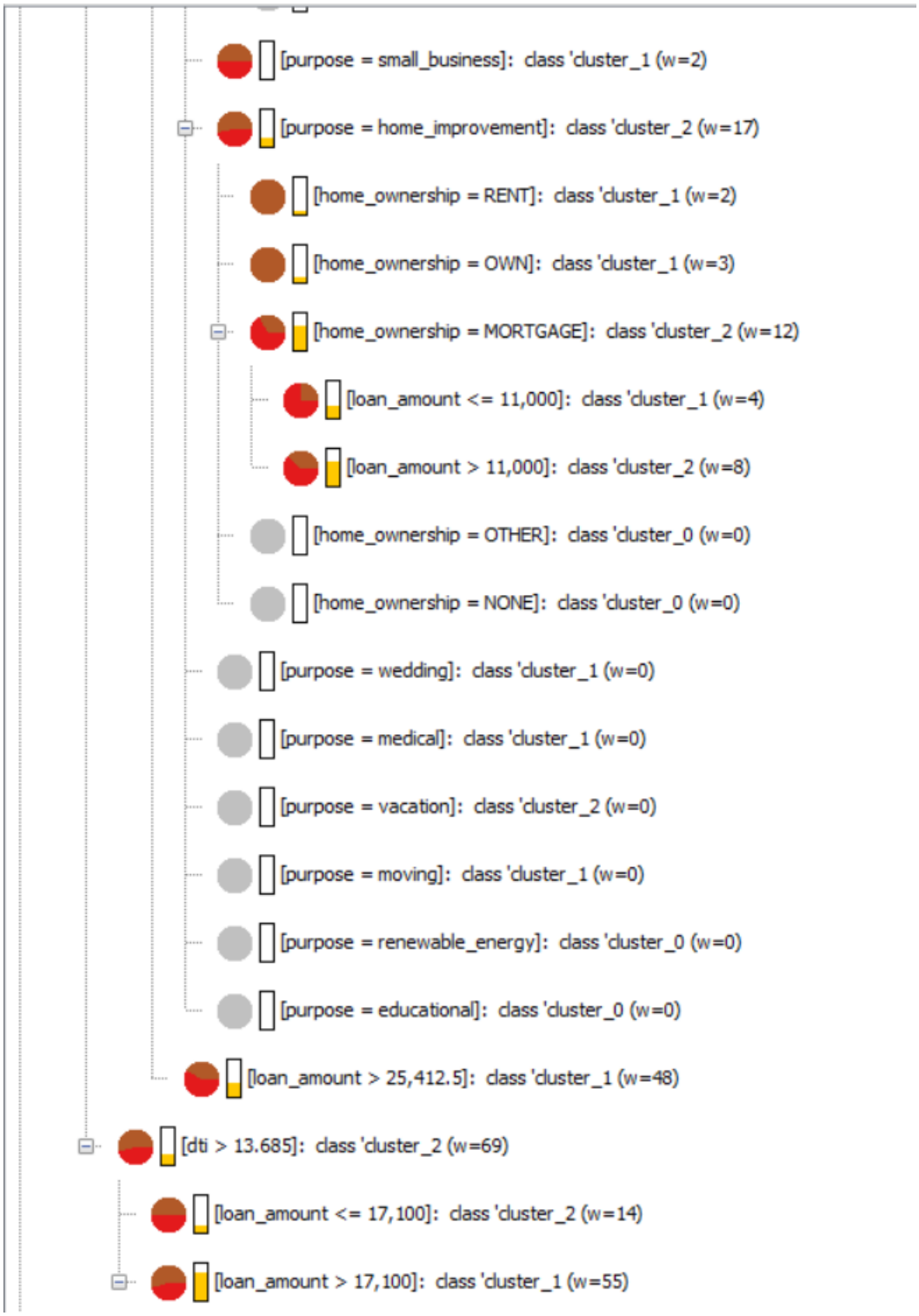
3.3.1. List of Relevant or Important Variables

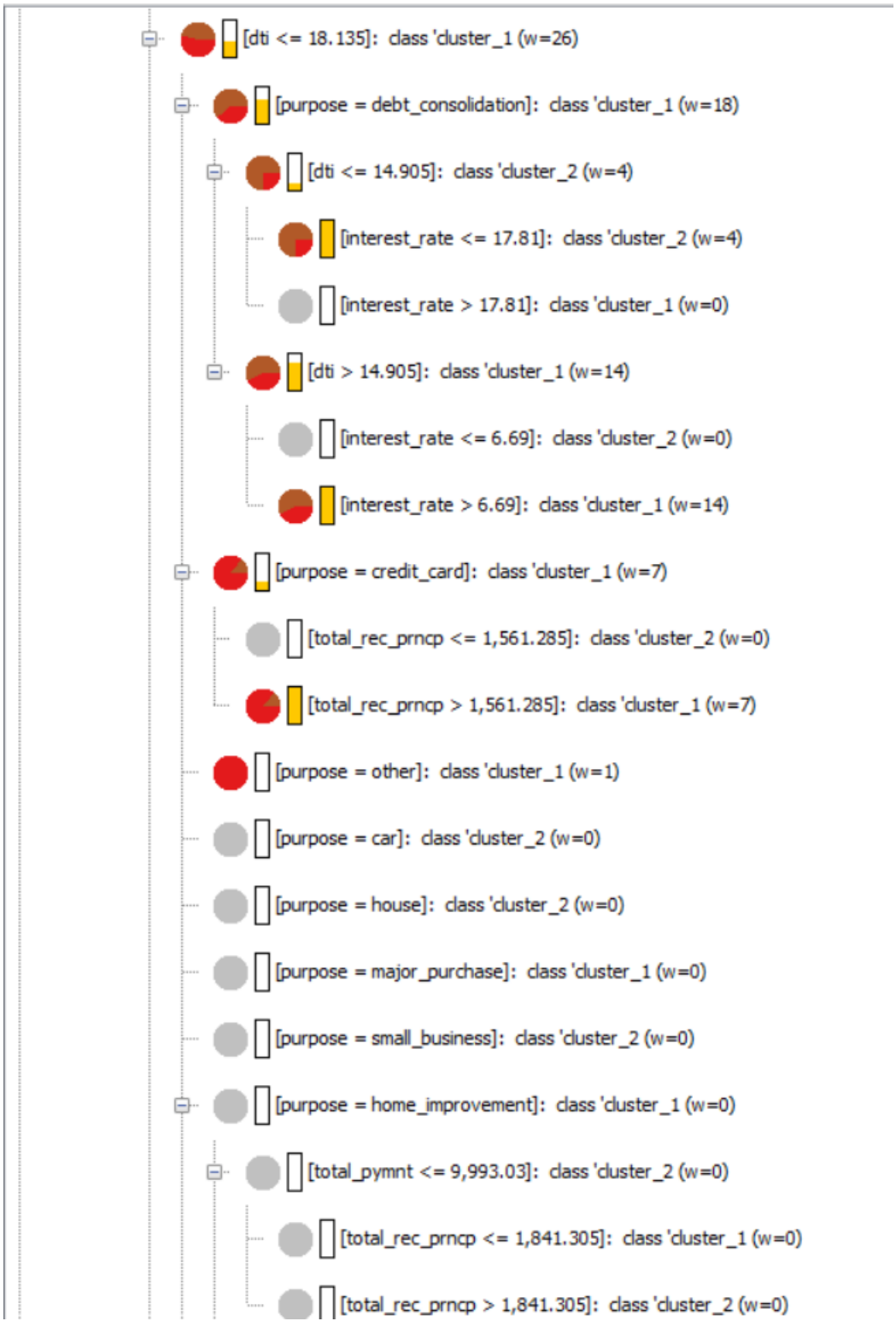
This image describes the variables that were important and contributed in the cross validation using decision tree to predict which cluster the record belonged to as well as the threshold onto which decision were made.

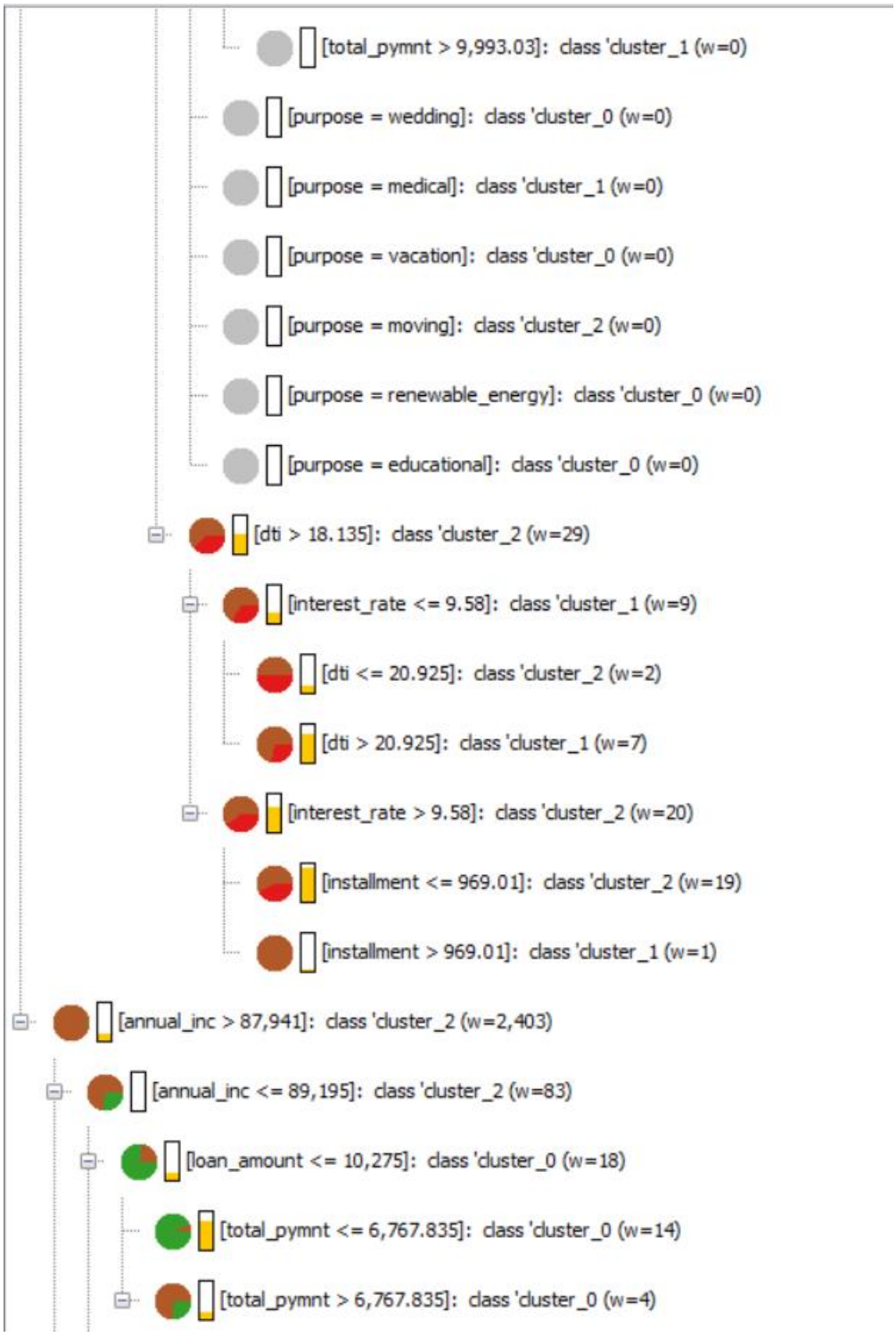


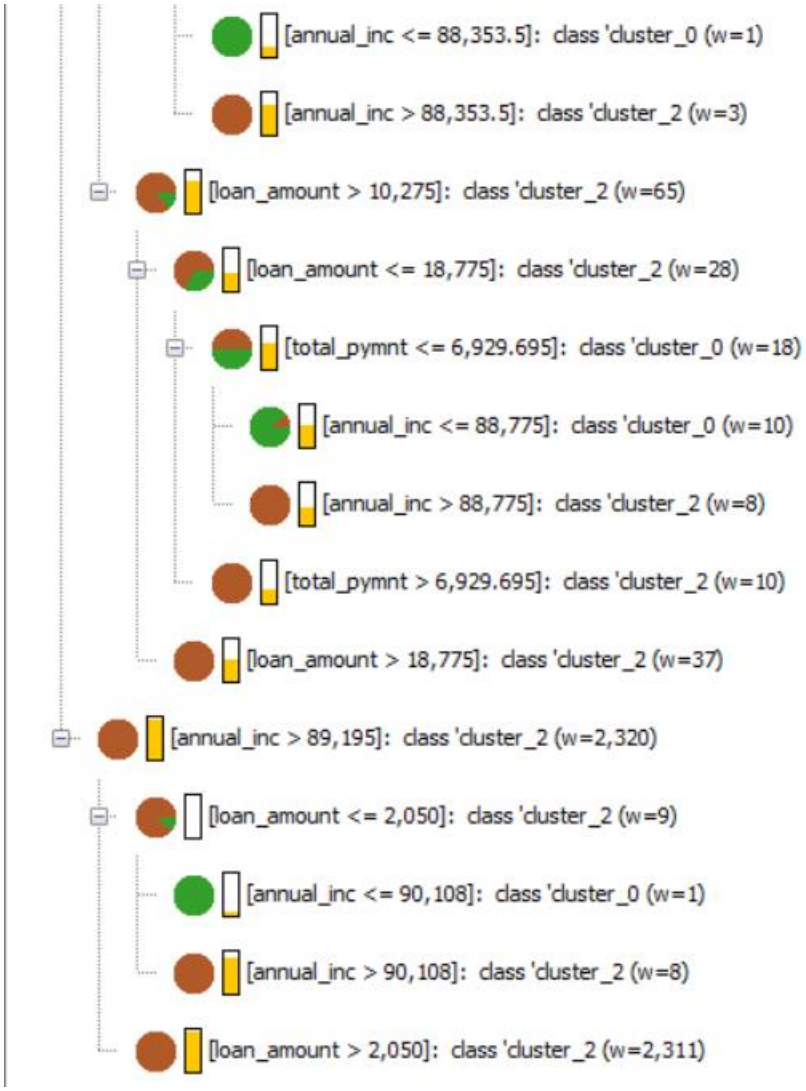

[loan_amount <= 22,025]: class 'cluster_0 (w=38)
[total_pymnt <= 14,958.2126]: class 'cluster_0 (w=30)
[total_pymnt > 14,958.2126]: class 'cluster_2 (w=8)
[annual_inc <= 86,850]: class 'cluster_0 (w=3)
[annual_inc > 86,850]: class 'cluster_2 (w=5)
[loan_amount > 22,025]: class 'cluster_2 (w=12)
[total_pymnt <= 5,093.425]: class 'cluster_0 (w=1)
[loan_amount <= 27,500]: class 'cluster_0 (w=1)
[annual_inc <= 87,127]: class 'cluster_0 (w=1)
[annual_inc > 87,127]: class 'cluster_2 (w=0)
[loan_amount > 27,500]: class 'cluster_2 (w=0)
[total_pymnt > 5,093.425]: class 'cluster_2 (w=11)
[income_category = Medium]: class 'cluster_2 (w=270)
[income_category = High]: class 'cluster_1 (w=209)
[dti <= 13.685]: class 'cluster_1 (w=140)
[dti <= 3.855]: class 'cluster_1 (w=24)
[dti > 3.855]: class 'cluster_1 (w=116)
[loan_amount <= 25,412.5]: class 'cluster_1 (w=68)
[purpose = debt_consolidation]: class 'cluster_1 (w=27)
[loan_amount <= 10,150]: class 'cluster_2 (w=2)


[loan_amount > 10,150]: class 'cluster_1 (w=25)
[total_pymnt <= 21,194.83]: class 'cluster_1 (w=23)
[total_pymnt > 21,194.83]: class 'cluster_2 (w=2)
[purpose = credit_card]: class 'cluster_1 (w=14)
[emp_length_int <= 3.5]: class 'cluster_2 (w=5)
[total_rec_prncp <= 6,135.025]: class 'cluster_2 (w=2)
[total_rec_prncp > 6,135.025]: class 'cluster_1 (w=3)
[emp_length_int > 3.5]: class 'cluster_1 (w=9)
[emp_length_int <= 6.025]: class 'cluster_1 (w=2)
[emp_length_int > 6.025]: class 'cluster_1 (w=7)
[interest_rate <= 11.76]: class 'cluster_1 (w=6)
[interest_rate > 11.76]: class 'cluster_2 (w=1)
[purpose = other]: class 'cluster_1 (w=4)
[dti <= 11.2]: class 'cluster_1 (w=4)
[dti > 11.2]: class 'cluster_2 (w=0)
[purpose = car]: class 'cluster_2 (w=1)
[purpose = house]: class 'cluster_2 (w=1)
[purpose = major_purchase]: class 'cluster_1 (w=2)
[dti <= 9.87]: class 'cluster_1 (w=2)
[dti > 9.87]: class 'cluster_2 (w=0)


[purpose = small_business]: class 'cluster_1 (w=2)
[purpose = home_improvement]: class 'cluster_2 (w=17)
[home_ownership = RENT]: class 'cluster_1 (w=2)
[home_ownership = OWN]: class 'cluster_1 (w=3)
[home_ownership = MORTGAGE]: class 'cluster_2 (w=12)
[loan_amount <= 11,000]: class 'cluster_1 (w=4)
[loan_amount > 11,000]: class 'cluster_2 (w=8)
[home_ownership = OTHER]: class 'cluster_0 (w=0)
[home_ownership = NONE]: class 'cluster_0 (w=0)
[purpose = wedding]: class 'cluster_1 (w=0)
[purpose = medical]: class 'cluster_1 (w=0)
[purpose = vacation]: class 'cluster_2 (w=0)
[purpose = moving]: class 'cluster_1 (w=0)
[purpose = renewable_energy]: class 'cluster_0 (w=0)
[purpose = educational]: class 'cluster_0 (w=0)
[loan_amount > 25,412.5]: class 'cluster_1 (w=48)
[dti > 13.685]: class 'cluster_2 (w=69)
[loan_amount <= 17,100]: class 'cluster_2 (w=14)
[loan_amount > 17,100]: class 'cluster_1 (w=55)


[dti <= 18.135]: class 'cluster_1 (w=26)
[purpose = debt_consolidation]: class 'cluster_1 (w=18)
[dti <= 14.905]: class 'cluster_2 (w=4)
[interest_rate <= 17.81]: class 'cluster_2 (w=4)
[interest_rate > 17.81]: class 'cluster_1 (w=0)
[dti > 14.905]: class 'cluster_1 (w=14)
[interest_rate <= 6.69]: class 'cluster_2 (w=0)
[interest_rate > 6.69]: class 'cluster_1 (w=14)
[purpose = credit_card]: class 'cluster_1 (w=7)
[total_rec_prncp <= 1,561.285]: class 'cluster_2 (w=0)
[total_rec_prncp > 1,561.285]: class 'cluster_1 (w=7)
[purpose = other]: class 'cluster_1 (w=1)
[purpose = car]: class 'cluster_2 (w=0)
[purpose = house]: class 'cluster_2 (w=0)
[purpose = major_purchase]: class 'cluster_1 (w=0)
[purpose = small_business]: class 'cluster_2 (w=0)
[purpose = home_improvement]: class 'cluster_1 (w=0)
[total_pymnt <= 9,993.03]: class 'cluster_2 (w=0)
[total_rec_prncp <= 1,841.305]: class 'cluster_1 (w=0)
[total_rec_prncp > 1,841.305]: class 'cluster_2 (w=0)


[total_pymnt > 9,993.03]: class 'cluster_1 (w=0)
[purpose = wedding]: class 'cluster_0 (w=0)
[purpose = medical]: class 'cluster_1 (w=0)
[purpose = vacation]: class 'cluster_0 (w=0)
[purpose = moving]: class 'cluster_2 (w=0)
[purpose = renewable_energy]: class 'cluster_0 (w=0)
[purpose = educational]: class 'cluster_0 (w=0)
[dti > 18.135]: class 'cluster_2 (w=29)
[interest_rate <= 9.58]: class 'cluster_1 (w=9)
[dti <= 20.925]: class 'cluster_2 (w=2)
[dti > 20.925]: class 'cluster_1 (w=7)
[interest_rate > 9.58]: class 'cluster_2 (w=20)
[installment <= 969.01]: class 'cluster_2 (w=19)
[installment > 969.01]: class 'cluster_1 (w=1)
[annual_inc > 87,941]: class 'cluster_2 (w=2,403)
[annual_inc <= 89,195]: class 'cluster_2 (w=83)
[loan_amount <= 10,275]: class 'cluster_0 (w=18)
[total_pymnt <= 6,767.835]: class 'cluster_0 (w=14)
[total_pymnt > 6,767.835]: class 'cluster_0 (w=4)



In the decision tree analysis, we see that these were the important variables that contributed in the cross validation using decision tree which are: -

annual_inc (annual income), income_category, loan_amount, total_rec_prncp (total received principal), total_payment, dti, purpose, emp_length_int, interest_rate, home_ownership

3.3.2. List of Non-Relevant or Unimportant Variables

Recoveries, term, application_type, loan_condition, region, interest_payments, grade, installment

3.4. Variable or Feature Analysis for Random Forest and XGBoost

3.4.1. Variables or Features that are important

From the tree view, the features that were shown in the tree view were the important features that determined the results which are: -

home_ownership, income_category, interest_rate, total_payment, loan_amount, annual_inc, emp_length_int, total_rec_prncp, installment, term, recoveries, loan_condition, Grade, dti, purpose

3.4.2. Variables or Features that are non-relevant

These variables or features were not important as it these variables were not a part of the tree view.

application_type, loan_condition, region, interest_payments

3.5. Variable or Feature Analysis for Cross Validation using Logistic Regression and K-Nearest Neighbour

3.5.1. Variables or Features that are important

Grade, dti, loan_amount, interest_rate, total_rec_prncp

These variables had $p < 0.05$ which shows its significance in the logistic regression equation i.e. the impact of these variables is more in the classification of customers.

3.5.2. Variables or Features that are non-relevant

annual_inc (annual income), income_category, total_payment, purpose, emp_length_int, home_ownership, recoveries, term, application_type, loan_condition, region, interest_payments, grade, installment

Some of the variables had higher coefficients that should have impacted the regression equation but they have less significance due the p value being greater than 0.05.

4. Results and Observations

4.1. Comparing Supervised Learning models: Cross Validation using Decision Tree VS Cross Validation using Logistic Regression, KNN

Cross validation using Decision Tree

No pruning

cluster \ Prediction (cluster)	cluster_0	cluster_2	cluster_1
cluster_0	77540	75	0
cluster_2	81	27130	457
cluster_1	0	470	696

Correct classified: 105,366 Wrong classified: 1,083

Accuracy: 98.983% Error: 1.017%

Cohen's kappa (κ): 0.975%

Pruning

cluster \ Prediction (cluster)	cluster_0	cluster_2	cluster_1
cluster_0	77557	61	0
cluster_2	97	27028	559
cluster_1	1	257	925

Correct classified: 105,510 Wrong classified: 975

Accuracy: 99.084% Error: 0.916%

Cohen's kappa (κ): 0.977%

Cross validation using other methods

Logistic regression

cluster \ Prediction (cluster)	cluster_0	cluster_2	cluster_1
cluster_0	77465	153	0
cluster_2	140	26980	564
cluster_1	0	209	974

Correct classified: 105,419 Wrong classified: 1,066

Accuracy: 98.999% Error: 1.001%

Cohen's kappa (κ): 0.975%

KNN

K=7

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	71167	6448	3
cluster_2	25304	2379	1
cluster_1	1073	110	0

Correct classified: 73,546 Wrong classified: 32,939

Accuracy: 69.067% Error: 30.933%

Cohen's kappa (κ): 0.004%

K=9

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	72875	4743	0
cluster_2	26033	1651	0
cluster_1	1101	82	0

Correct classified: 74,526 Wrong classified: 31,959

Accuracy: 69.987% Error: 30.013%

Cohen's kappa (κ): -0.002%

K=11

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	74233	3385	0
cluster_2	26395	1289	0
cluster_1	1126	57	0

Correct classified: 75,522 Wrong classified: 30,963

Accuracy: 70.923% Error: 29.077%

Cohen's kappa (κ): 0.004%

K=13

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	75088	2530	0
cluster_2	26769	915	0
cluster_1	1144	39	0

Correct classified: 76,003 Wrong classified: 30,482

Accuracy: 71.374% Error: 28.626%

Cohen's kappa (κ): 0.001%

K=15

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	75742	1876	0
cluster_2	26982	702	0
cluster_1	1164	19	0

Correct classified: 76,444 Wrong classified: 30,041

Accuracy: 71.789% Error: 28.211%

Cohen's kappa (κ): 0.001%

K=17

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	76212	1406	0
cluster_2	27155	529	0
cluster_1	1162	21	0

Correct classified: 76,741 Wrong classified: 29,744

Accuracy: 72.067% Error: 27.933%

Cohen's kappa (κ): 0.001%

K=19

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	76555	1063	0
cluster_2	27265	419	0
cluster_1	1166	17	0

Correct classified: 76,974 Wrong classified: 29,511

Accuracy: 72.286% Error: 27.714%

Cohen's kappa (κ): 0.002%

Random Forest

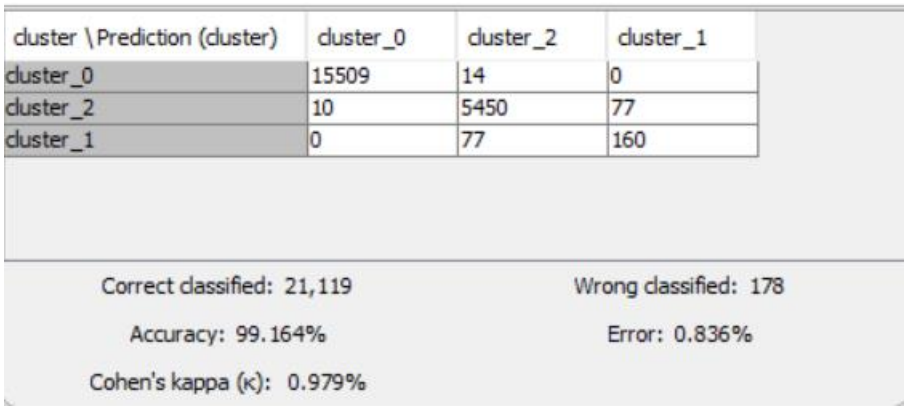
cluster \ Prediction (cluster)	cluster_0	cluster_2	cluster_1
cluster_0	15509	14	0
cluster_2	28	5438	71
cluster_1	0	98	139

Correct classified: 21,086 Wrong classified: 211

Accuracy: 99.009% Error: 0.991%

Cohen's kappa (κ): 0.975%

XGBoost



cluster \Prediction (cluster)	cluster_0	cluster_2	cluster_1
cluster_0	15509	14	0
cluster_2	10	5450	77
cluster_1	0	77	160

Correct classified: 21,119 Wrong classified: 178

Accuracy: 99.164% Error: 0.836%

Cohen's kappa (κ): 0.979%

	Cross Validation				**Ensemble Learning**	
Metrics	**Decision Tree (no pruning)**	**Decision Tree (pruning)**	**Logistic Regression**	**KNN**	**Random Forest**	**XGBoost**
Accuracy (in %)	98.983	99.084	98.999	72.286	99.009	99.164
Error (in %)	1.017	0.916	1.001	27.714	0.991	0.836
Cohen's Kappa (in %)	0.975	0.977	0.975	0.002	0.975	0.979
Correctly classified	105366	105510	105419	76974	21086	21119
Wrongly Classified	1083	975	1066	29511	211	178

- **Cross validation using Decision Trees**: Both with and without pruning show high accuracy and Cohen's Kappa scores indicating good performance. Pruning helps slightly improve accuracy and reduce misclassification.
- **Cross validation using Logistic Regression**: This algorithm Shows high accuracy and Cohen's Kappa score similar to decision trees, indicating robustness and effectiveness for the given dataset.
- **Cross validation using KNN**: Performs significantly lower compared to other models, with the lowest accuracy and Cohen's Kappa score. This suggests that KNN might not be suitable for this dataset or may require further tuning of hyperparameters.
- **Random Forest and XGBoost (Ensemble learning):** Both ensemble methods perform exceptionally well with high accuracy and Cohen's Kappa scores. XGBoost outperforms Random Forest slightly in terms of accuracy and Cohen's Kappa, indicating its superior predictive power for this dataset.

For this dataset, ensemble learning methods like Random Forest and XGBoost along with Decision Trees with pruning, seem to be the most effective models in terms of accuracy and robustness. Logistic Regression also performs well and provides interpretable results which can be advantageous in certain scenarios. However, KNN appears to be less suitable due to it less accuracy.

4.1. Variable or Feature Analysis

Variables Important for Random Forest and XGBoost

1. **home_ownership**: This variable represents the type of home ownership (e.g., rent, own, mortgage). It could be important because individuals who own their homes outright or have a mortgage may be seen as more financially stable compared to those who rent.
2. **income_category**: This variable likely represents different income levels or categories of customers. Income is a crucial factor in determining loan risk and affluence. Higher income individuals are generally considered less risky borrowers and may be categorized as more affluent.
3. **interest_rate**: The interest rate on a loan is a key determinant of the cost of borrowing and loan risk. Higher interest rates may indicate riskier loans or borrowers with lower creditworthiness.
4. **total_payment**: This variable likely represents the total amount paid by the borrower over the life of the loan. It reflects the borrower's ability to make payments and manage their debt obligations.
5. **loan_amount**: The amount of the loan is another important factor in assessing loan risk. Higher loan amounts may indicate larger financial commitments and potentially higher risk.
6. **annual_inc**: Annual income is a fundamental factor in assessing a borrower's ability to repay a loan. Higher income individuals are generally more likely to be able to meet their loan obligations.
7. **emp_length_int**: This variable represents the length of employment in years. Stable employment history can indicate financial stability and ability to repay loans.
8. **total_rec_prncp**: This variable likely represents the total amount of principal received by the borrower. It reflects the borrower's repayment behaviour and ability to manage their debt.
9. **installment**: The instalment amount represents the monthly payment on the loan. It is an important factor in assessing the borrower's ability to make regular payments.
10. **term**: The loan term (e.g., 36 months, 60 months) is important as it determines the duration of the loan and the repayment period.
11. **recoveries**: This variable likely represents the amount recovered by the lender after a loan default. It reflects the effectiveness of the lender's recovery process and the overall risk associated with lending.

12. **loan_condition**: This variable likely represents the loan condition (e.g., good loan, bad loan) and is directly related to loan risk assessment.
13. **Grade**: This variable likely represents the loan grade assigned by the lender based on the borrower's creditworthiness. Higher grade loans are considered lower risk.
14. **dti**: Debt-to-income ratio is a crucial factor in assessing a borrower's financial health and ability to manage additional debt obligations.
15. **purpose**: The purpose of the loan (e.g., debt consolidation, home improvement) can provide insights into the borrower's financial behaviour and intentions.

Variables not important for Random Forest and XGBoost

application_type, loan_condition, region, interest_payments

Variables important for Cross Validation using Decision Tree

annual_inc (annual income), income_category, loan_amount, total_rec_prncp (total received principal), total_payment, dti, purpose, emp_length_int, interest_rate, home_ownership

Variables not important for Cross Validation using Decision Tree

Recoveries, term, application_type, loan_condition, region, interest_payments, grade, installment

Variables or Features that are important for Cross Validation using Logistic Regression and K-Nearest Neighbour

Grade, dti, loan_amount, interest_rate, total_rec_prncp

These variables had $p < 0.05$ which shows its significance in the logistic regression equation i.e. the impact of these variables is more in the classification of customers.

Variables or Features that are non-relevant Cross Validation using Logistic Regression and K-Nearest Neighbour

annual_inc (annual income), income_category, total_payment, purpose, emp_length_int, home_ownership, recoveries, term, application_type, loan_condition, region, interest_payments, grade, installment

5. Managerial Insights

	Cross Validation				Ensemble Learning	
Metrics	**Decision Tree (no pruning)**	**Decision Tree (pruning)**	**Logistic Regression**	**KNN**	**Random Forest**	**XGBoost**
Accuracy (in %)	98.983	99.084	98.999	72.286	99.009	99.164

XGBoost Ensemble learning has the highest accuracy followed closely by Decision tree (with pruning). KNN has the lowest of accuracy when compared to all the models, thus it won't be preferred when classifying bank customers on the basis of affluency and loan risk.

Managerial insights according to the appropriate model (XGBoost ensemble learning)

1. **Credit Risk Assessment**
    - XGBoost can help banks assess the credit risk associated with individual customers by analyzing their financial profiles, transaction history, credit scores and other relevant data.
    - By leveraging XGBoost, banks can classify customers into different risk categories (e.g., low risk, moderate risk, high risk) with high accuracy, enabling them to make informed decisions about lending and credit approvals.

2. **Customer Segmentation**
    - XGBoost can aid banks in segmenting their customer base into distinct groups based on various attributes such as income level, spending behaviour, banking preferences and risk profile.
    - By utilizing XGBoost for customer segmentation, banks can tailor their marketing strategies, product offerings and customer service initiatives to meet the specific needs and preferences of different customer segments effectively.

3. **Cross-Selling and Upselling**
    - XGBoost can help banks identify opportunities for cross-selling and upselling additional financial products and services to existing customers.
    - By leveraging XGBoost, banks can analyse customer data to identify relevant product recommendations and personalize marketing campaigns, thereby increasing customer engagement, loyalty and revenue generation.

4. **Loan Default Prediction**

- XGBoost can be utilized to predict the likelihood of loan defaults by assessing various factors such as credit history, income stability, debt-to-income ratio and employment status.
- Banks can use XGBoost to classify loan applications as either low-risk or high-risk, enabling them to make informed decisions about loan approvals, interest rates and repayment terms to minimize default risks and optimize loan portfolio performance.