ASSIGNMNET 2

DATA MINING

MUHAMMAD IZZUDDIN BIN AHAMAD SFHAFI

WQD170041

TITLE: DATA ANALYTIC AND MODELLING ON CENSUS INCOME DATASET USING ENTERPRISE MINER

Contents

1. Introduction

The Census Income dataset is derived from UCI dataset archieve. The challenge from this data set is to determine the binary outcome of if a person is having an annual income more 50,000 USD or less than 50,000 USD based on the 14 variables and 32562 rows of dataset.

URL link of the Census Income Data Set: https://archive.ics.uci.edu/ml/datasets/census+income

1.1 Objective

To find best predicting model between Decision Tree and Logistic that predict if a person has annual income above $50,000 USD or below $50,000 based on the given variable

The variables consist of:

- Age = the age of the subject, interval
- Workclass = Occupation of the subject, nominal (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked)
- Fnlwgt = ID of the subjects, ID
- Education = Level of education of the subject, ordinal (Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool)
- Education-num: Years of education that the subject has, continuous value
- Marital status: Married status of the subject, nominal (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse)
- Occupation: The job of the subject (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces)
- Relationship: The relationship status of the subject of the subject (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried)
- Race: the ethnicity of the subject, nominal (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black)
- Sex: The Gender of the subject, binary (Female or Male)
- Target: above 50,000 USD annual income or below 50,000 USD annual income, binary

2. Uploading Data into Enterprise Miner

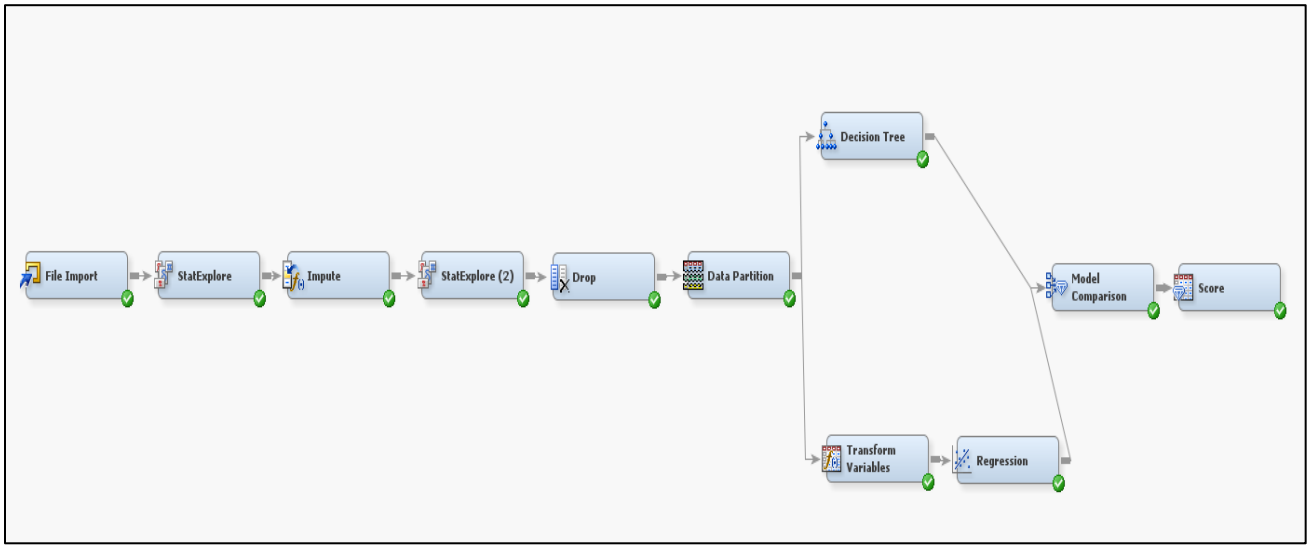


Figure 1: Process Flow for the modelling

2.1 Select Data Roles and Level

Above is the process flow for the data analytic and modelling process. The dataset is in CSV form. The data for each variable were selected based on criteria in the diagram below

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	**Input**	**Interval**	**No**		**No**	.	.
Target	**Target**	**Binary**	**No**		**No**	.	.
Workclass	**Input**	**Nominal**	**No**		**No**	.	.
capital_gain	**Input**	**Interval**	**No**		**No**	.	.
capital_loss	**Input**	**Interval**	**No**		**No**	.	.
country	**Input**	**Nominal**	**No**		**No**	.	.
education	**Input**	**Ordinal**	**No**		**No**	.	.
education_num	**Input**	**Interval**	**No**		**No**	.	.
fnlwgt	**ID**	**Interval**	**No**		**No**	.	.
hour_per_week	**Input**	**Interval**	**No**		**No**	.	.
mariatl_status	**Input**	**Nominal**	**No**		**No**	.	.
occupation	**Input**	**Nominal**	**No**		**No**	.	.
race	**Input**	**Binary**	**No**		**No**	.	.
relationship	**Input**	**Nominal**	**No**		**No**	.	.
sex	**Input**	**Binary**	**No**		**No**	.	.

Figure 2: The roles and data format for each variable

The fnlwgt variable was selected as ID while the Target is binary output which consist if the annual income was above or below 50,000 USD annually. Education was selected as ordinal as we hypothesize that different stages of education has different ranking (school, high school, college, etc). Workclass, Country, Marital status, occupation and relationship variables were selected as ordinal level to represent categorical nature of the variable and the remaining variables were selected as nominal

3. Initial Data Exploration

The initial data exploration was conducted to have initial understanding of the dataset. Using Explore Plot on the import node. We will explore the correlations between variables to have better understanding on the dataset.

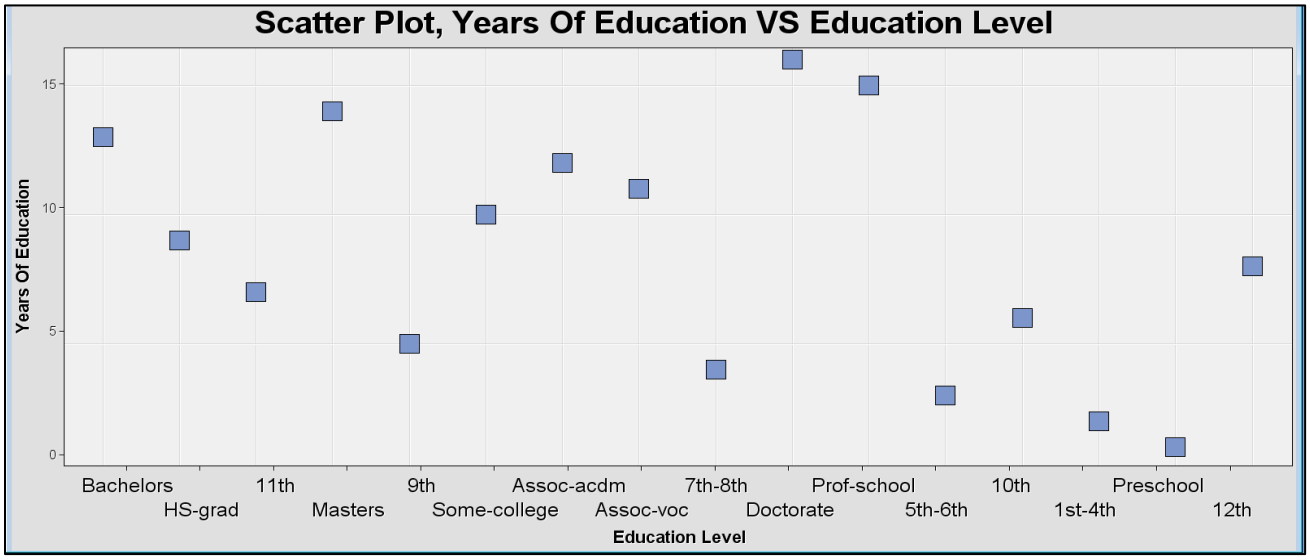


Figure 3: Scatter Plot of Years of Education Vs Education Level

The scatter plot showed that the year of education was tallied with education level. The highest education level was Doctorate, followed by Prof-school, master and then bachelor was in accordance with the highest number of years of education. This showed positive correlation between these two variables.

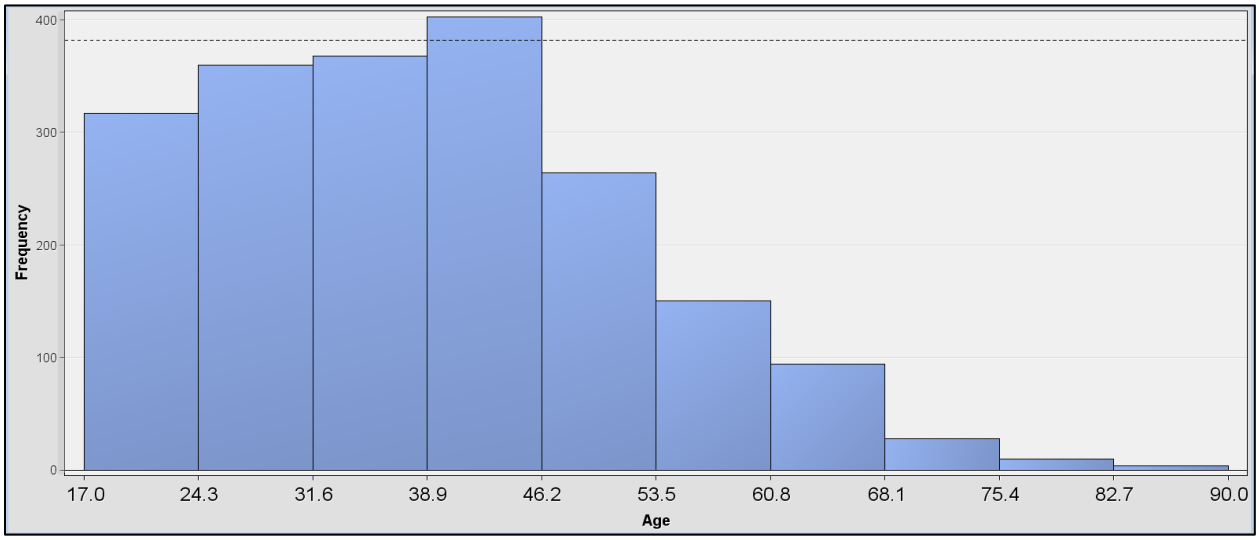


Figure 4: Histogram of Age

Based on our findings on histogram of Age. The youngest age for the subject was 17 years old while the highest age was around 90 years old. Most of the age population fall between 38.9-46.2 years old, followed by 31.6-38.9, 24.3-31.6 and 17-24.3 years old. All of these age bracket has population more than 300. The lowest population of age bracket was between 82.7-90

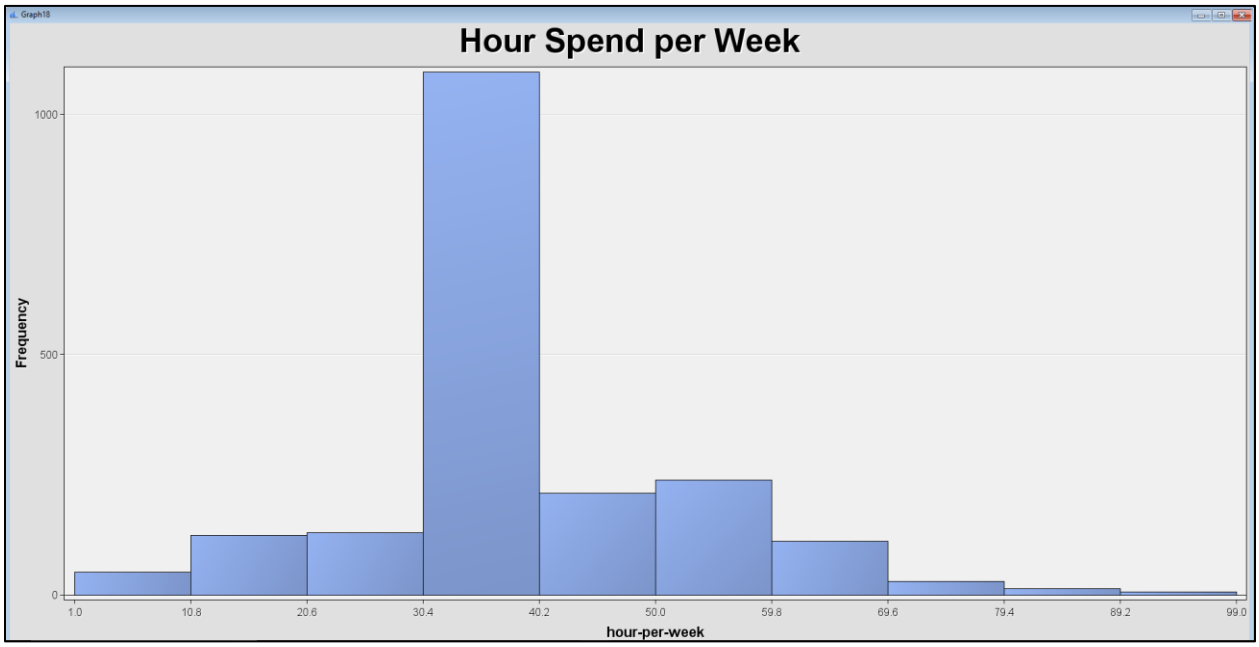


Figure 5: Hour Spend Per Week Bar Chart

Based on the Hour Spend per week, most of the subjects spend around 30-40 hours per week working which is consider common working hour. And this population of working hour made up more than 50% of the whole working hour population.

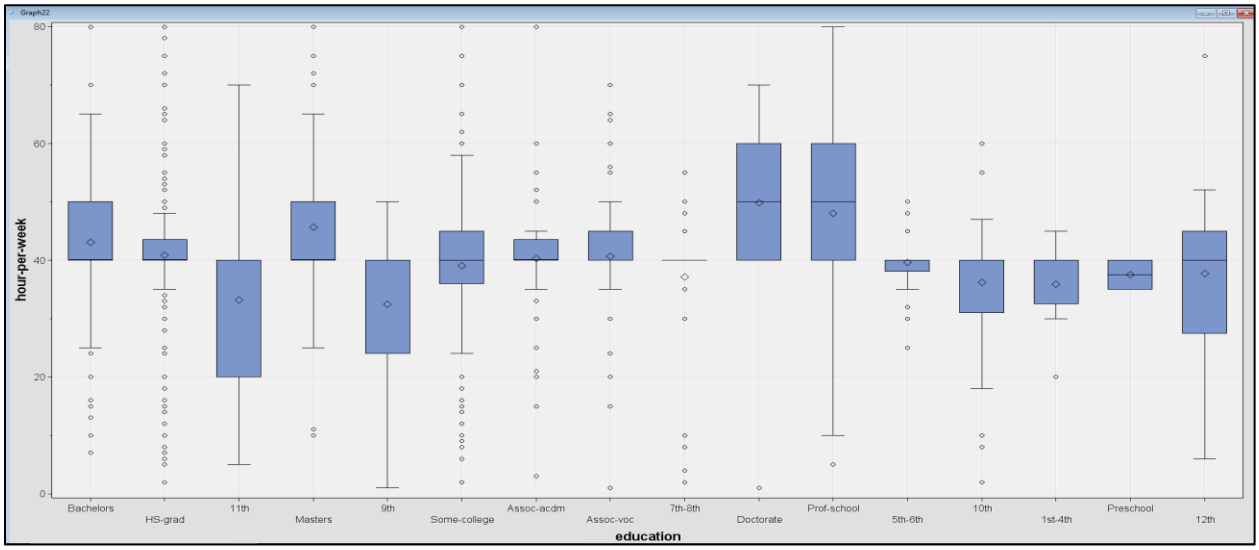


Figure 6: Boxplot Total Hour Per week Vs Education

We use boxplot to investigate the relationship on how many hours per week spend base on level of education. From the plot, it showed that the higher educated worker (bachelor, master, prof-school and Doctorate) tend to spend more than 40 hours per week working compare to other type of education level. This further support our initial assumption which showed higher educated subject tend to make 50, 000 USD annually due to their longer working hour.

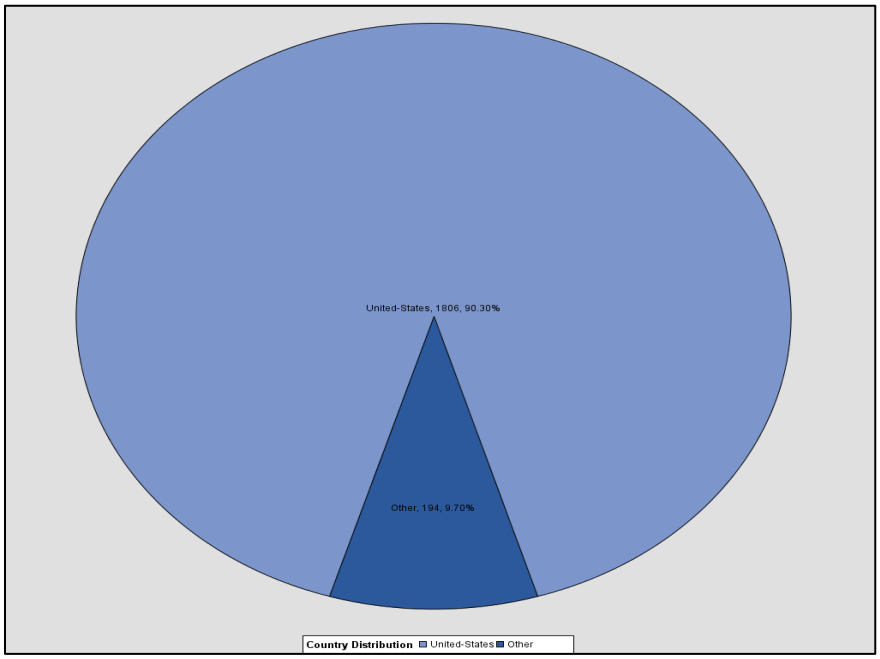


Figure 7: Pie Chart for country distribution

Based on the pie chart, it showed that the more than 90% of the subjects originated from USA.

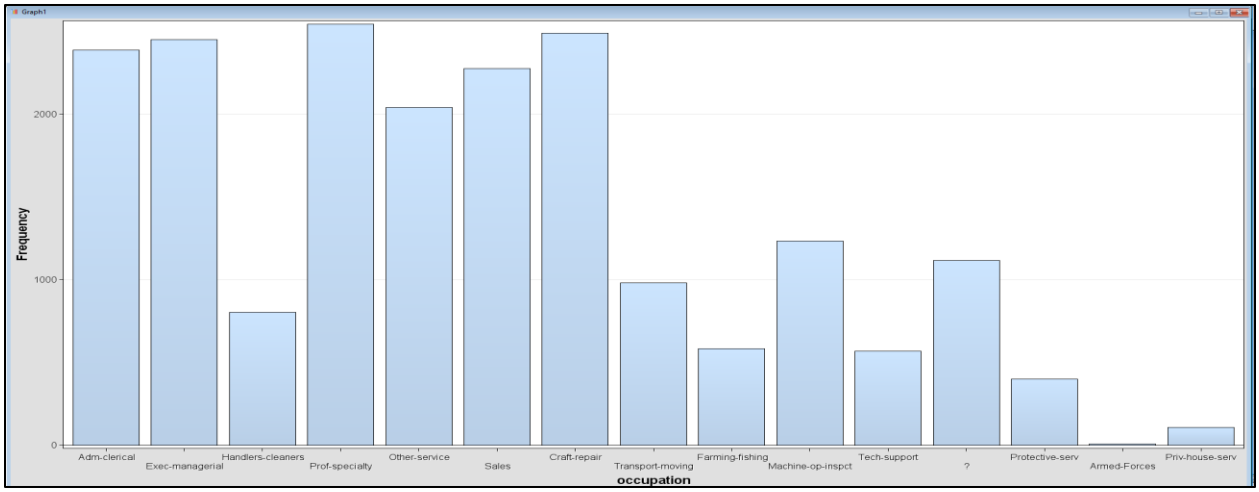


Figure 8: bar Chart for Occupation Type

Based on type of Occupation bar chart. We found out that Prof-speciality made the highest population of profession while the lowest one was the armed-forces.

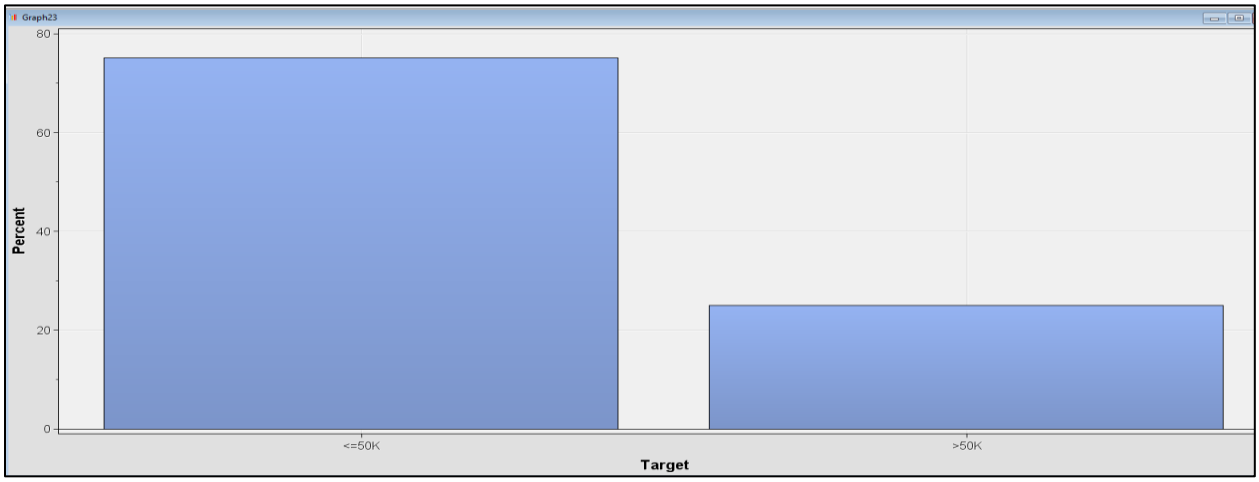


Figure 9: Percentage Bar Chart for target Varaible

Based on the percentage bar chart for our target variable, we found that more than 75% of the population made income below 50,000 annually and around 25% made annual income more than 50,000 USD.

4. Correlation of Variable

2. Stat Explore

StatExplore node was used to generate Descriptive Statistic. Chi Square Plot and Variable Worth plot are generated to determine the most important variable the target (income above >50k USD or <50k USD)

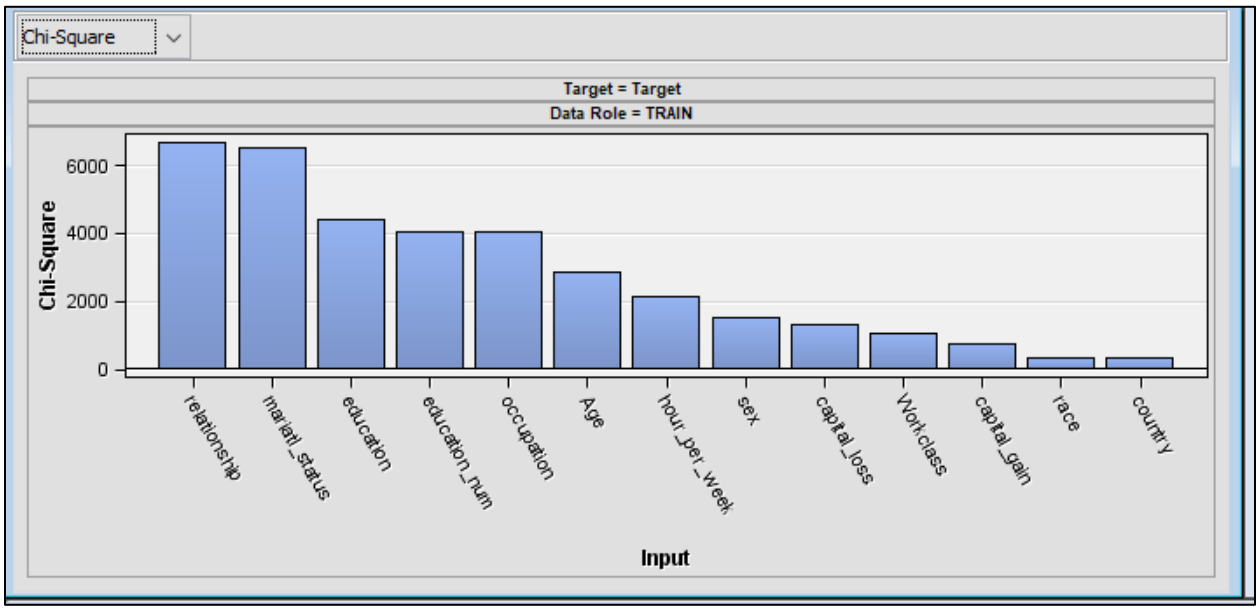


Figure 10: Chi Square for variables

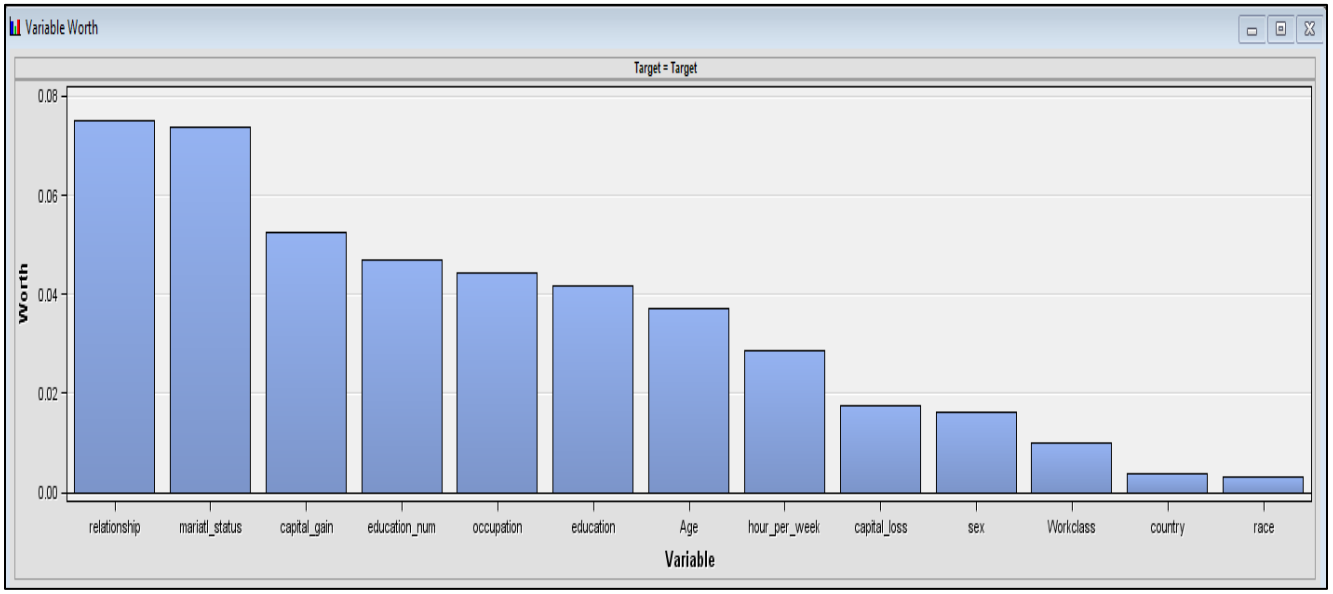


Figure 11: Variable Worth for the variable

The findings of both Chi Square and Variable Worth lead to interesting discovery, our initial assumptions were that education and occupation will become the most important factors of having annual income above or below 50k USD. But it turns out people the most important variable was relationship (rank 1) followed by marital status (rank 2). Race and country were the weakest variable that influence our target.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Workclass	INPUT	10	1	Private	69.70	Self-emp-not-inc	7.80
TRAIN	country	INPUT	43	1	United-States	89.58	Mexico	1.97
TRAIN	education	INPUT	17	2	HS-grad	32.25	Some-college	22.39
TRAIN	mariatl_status	INPUT	8	1	Married-civ-spouse	45.99	Never-married	32.81
TRAIN	occupation	INPUT	15	0	Prof-specialty	12.71	Craft-repair	12.59
TRAIN	race	INPUT	6	1	White	85.42	Black	9.59
TRAIN	relationship	INPUT	7	2	Husband	40.51	Not-in-family	25.50
TRAIN	sex	INPUT	3	1	Male	66.92	Female	33.08
TRAIN	Target	TARGET	2	0	<=50K	75.92	>50K	24.08

Figure 12: Variable Summary Statistic

Based on Variable Summary Statistics, our further investigation showed that the mode percentage for country (United States) and race (white) contribute to the highest percentage of the mode population (89.58% and 85,42%) which is believed to be the main reason to the lowest rank of importance for both Chi-Square and Variable Worth plot.

5. Missing Value

Data Role	Variable Name	Role	Number of Levels	Missing
TRAIN	Workclass	INPUT	10	1
TRAIN	country	INPUT	43	1
TRAIN	education	INPUT	17	2
TRAIN	mariatl_status	INPUT	8	1
TRAIN	occupation	INPUT	15	0
TRAIN	race	INPUT	6	1
TRAIN	relationship	INPUT	7	2
TRAIN	sex	INPUT	3	1
TRAIN	Target	TARGET	2	0

Figure 13: Missing Value for each variables

From StatExplore result, it was found out that the missing value were so minor. Nevertheless, instead of removing the missing value, imputation method was chosen to fill up all the missing values

6. Imputation

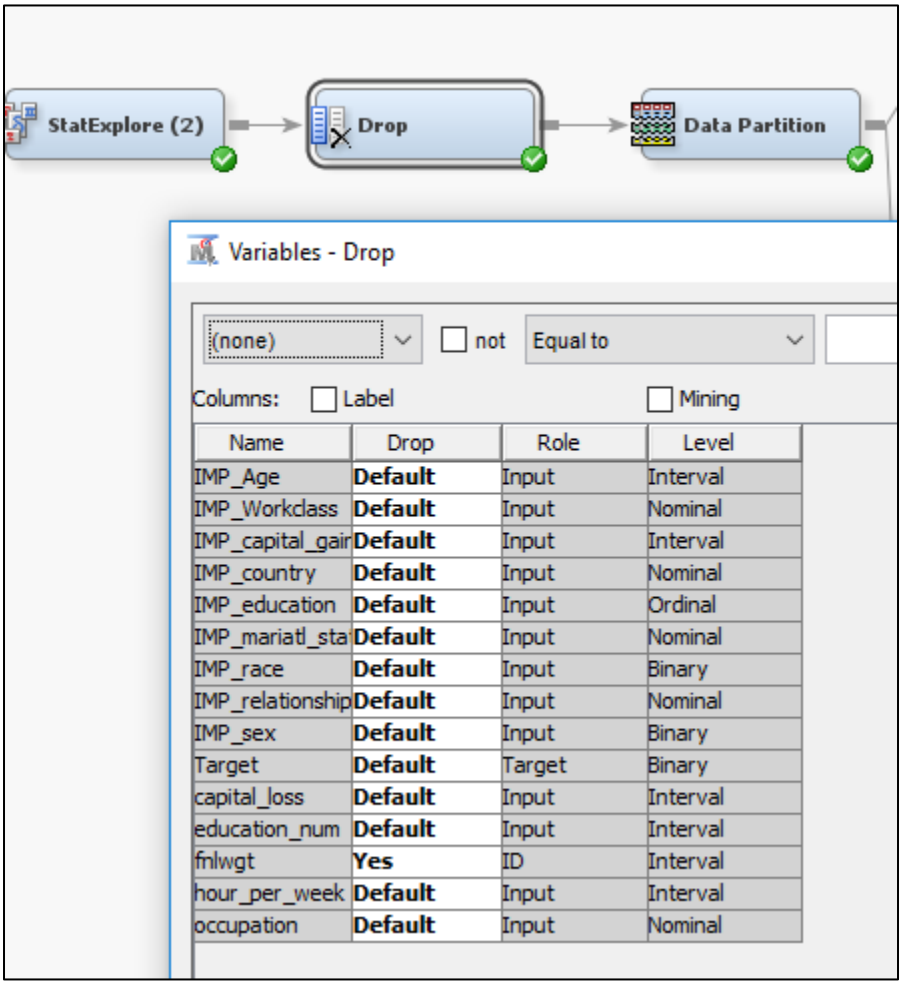
Under Imputation, Distribution method was selected for all the missing variable. The main reason that for this selection due to the very small portion of the data set (less than 0.1%) were missing and secondly due to the some of the missing value came from categorical variable (which statistic method such as average, mean median are not feasible).

40	Variable Name	Method	Imputed Variable	Value	Role	Level	Label	for TRAIN
41								
42	Age	DISTRIBUTION	IMP_Age	.	INPUT	INTERVAL		11
43	Workclass	DISTRIBUTION	IMP_Workclass	.	INPUT	NOMINAL		1
44	capital_gain	DISTRIBUTION	IMP_capital_gain	.	INPUT	INTERVAL	capital-gain	1
45	country	DISTRIBUTION	IMP_country	.	INPUT	NOMINAL		1
46	education	DISTRIBUTION	IMP_education	.	INPUT	ORDINAL		2
47	mariatl_status	DISTRIBUTION	IMP_mariatl_status	.	INPUT	NOMINAL	mariatl-status	1
48	race	DISTRIBUTION	IMP_race	.	INPUT	BINARY		1
49	relationship	DISTRIBUTION	IMP_relationship	.	INPUT	NOMINAL		2
50	sex	DISTRIBUTION	IMP_sex	.	INPUT	BINARY		1

Figure 14:After result of imputation using Distribution Method

7. Drop

Fnlweight variable consist of more than 21000 unique values and did not contribute to any significant under Chi-square and Variable Worth plot. The fnlwght variable was drop under drop node to reduce the complexity of the final modelling



Name	Drop	Role	Level
IMP_Age	**Default**	Input	Interval
IMP_Workclass	**Default**	Input	Nominal
IMP_capital_gair	**Default**	Input	Interval
IMP_country	**Default**	Input	Nominal
IMP_education	**Default**	Input	Ordinal
IMP_mariatl_sta	**Default**	Input	Nominal
IMP_race	**Default**	Input	Binary
IMP_relationship	**Default**	Input	Nominal
IMP_sex	**Default**	Input	Binary
Target	**Default**	Target	Binary
capital_loss	**Default**	Input	Interval
education_num	**Default**	Input	Interval
fnlwgt	**Yes**	ID	Interval
hour_per_week	**Default**	Input	Interval
occupation	**Default**	Input	Nominal

Figure 15: Dropping fnlwgt variable under Drop node

8. Data Partition

The dataset was split into 40% training set, 30% validate set and 20% test set under the partition node

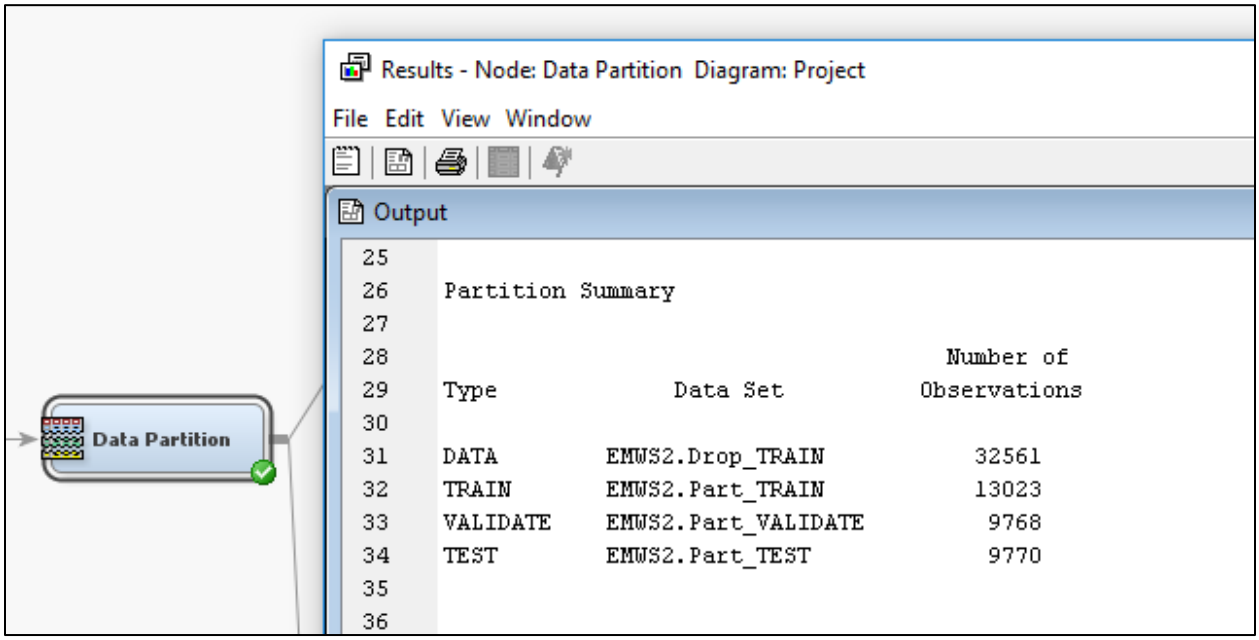


Figure 16: Data partition

9. Modelling

One of the main objective for this research is to compare the performance model between Logistic Regression and Decision Tree. After Data Partition node, the connecting line was split into 2 models which are Logistic Regression and Decision Tree

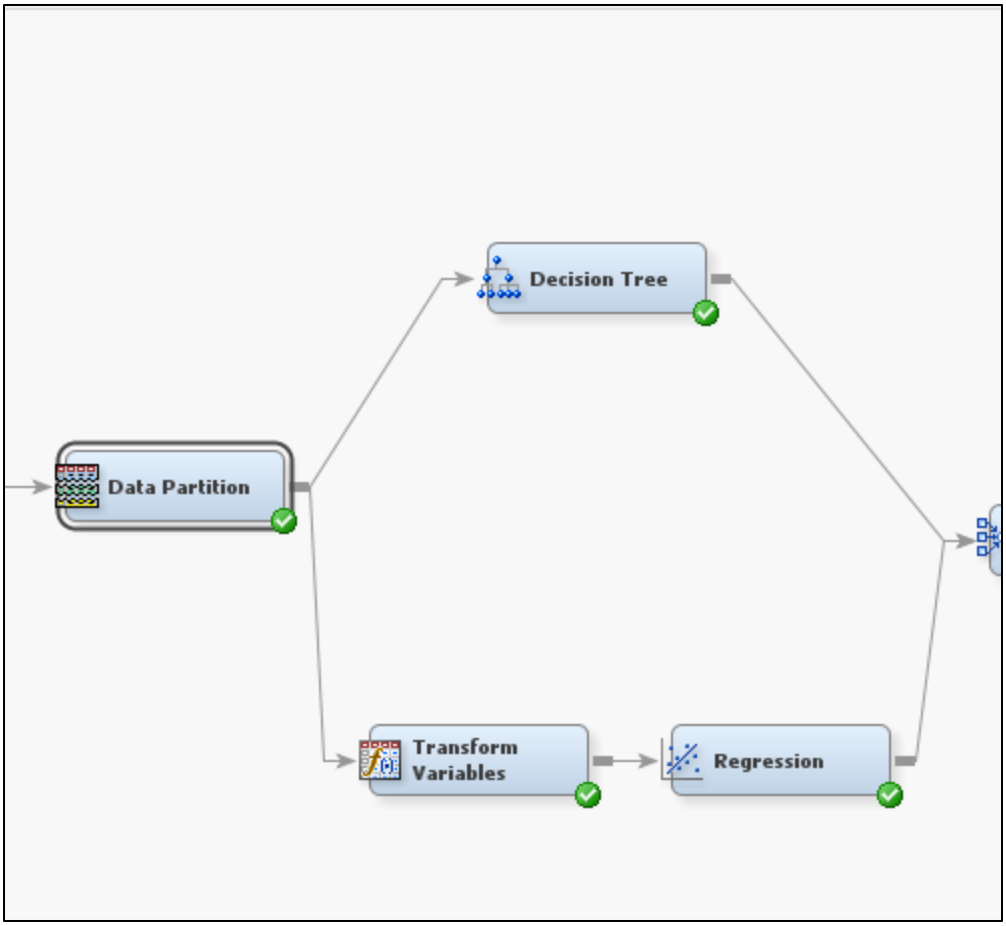


Figure 17: Generate Decision Tree And Logistic Regression Node

9.1 Decision Tree

The default setting for Decision Tree was used, All the parameters required for the decision tree were set as below

Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25

Figure 18: Decision Tree Parameter

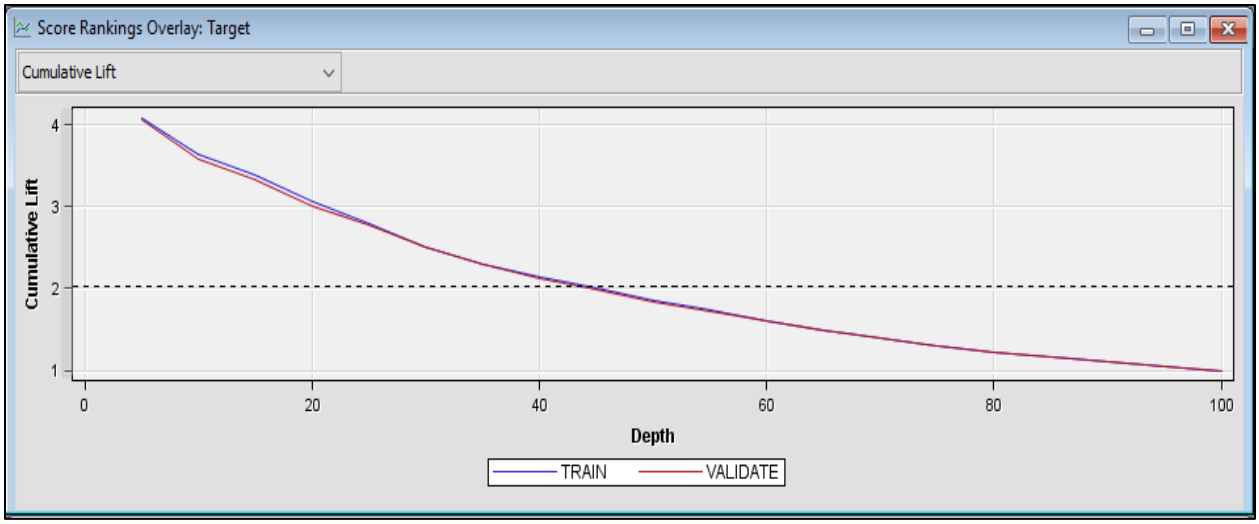


Figure 19: Score Ranking Overlay: Target

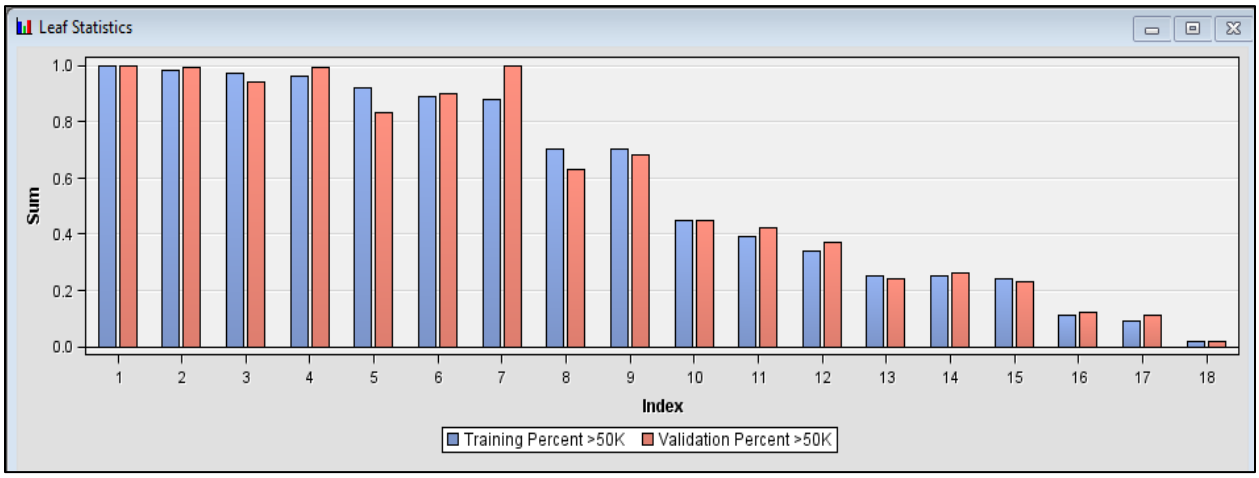


Figure 20: Leaf Statistic

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Target		_NOBS_	Sum of Frequencies	13023	9768	9770
Target		_MISC_	Misclassification Rate	0.142901	0.147011	0.146059
Target		_MAX_	Maximum Absolute Err...	0.996528	0.996528	0.996528
Target		_SSE_	Sum of Squared Errors	2661.742	2048.065	2024.203
Target		_ASE_	Average Squared Error	0.102194	0.104835	0.103593
Target		_RASE_	Root Average Squared...	0.319678	0.323783	0.321858
Target		_DIV_	Divisor for ASE	26046	19536	19540
Target		_DFT_	Total Degrees of Free...	13023	.	.

Figure 21: Fit Statistics

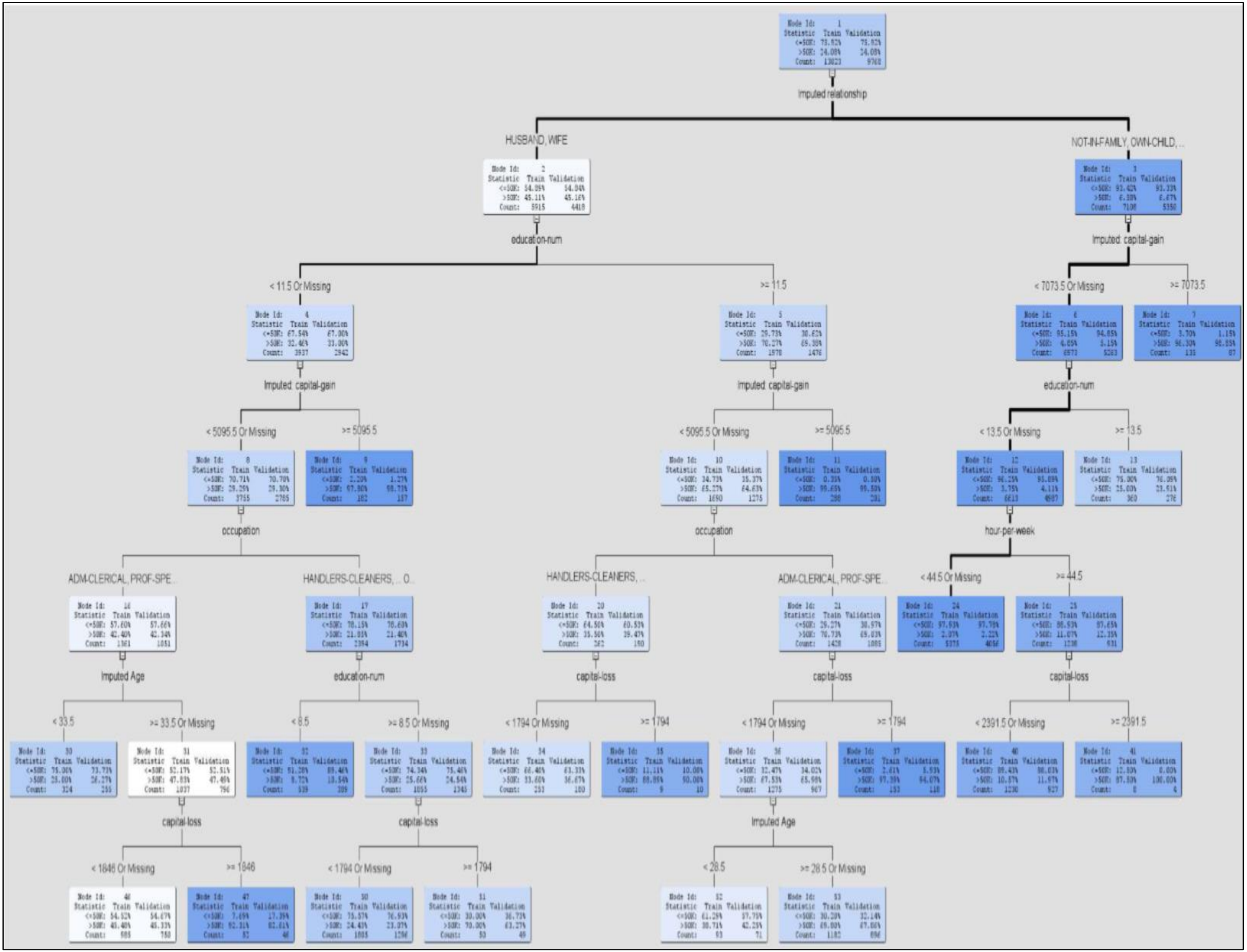
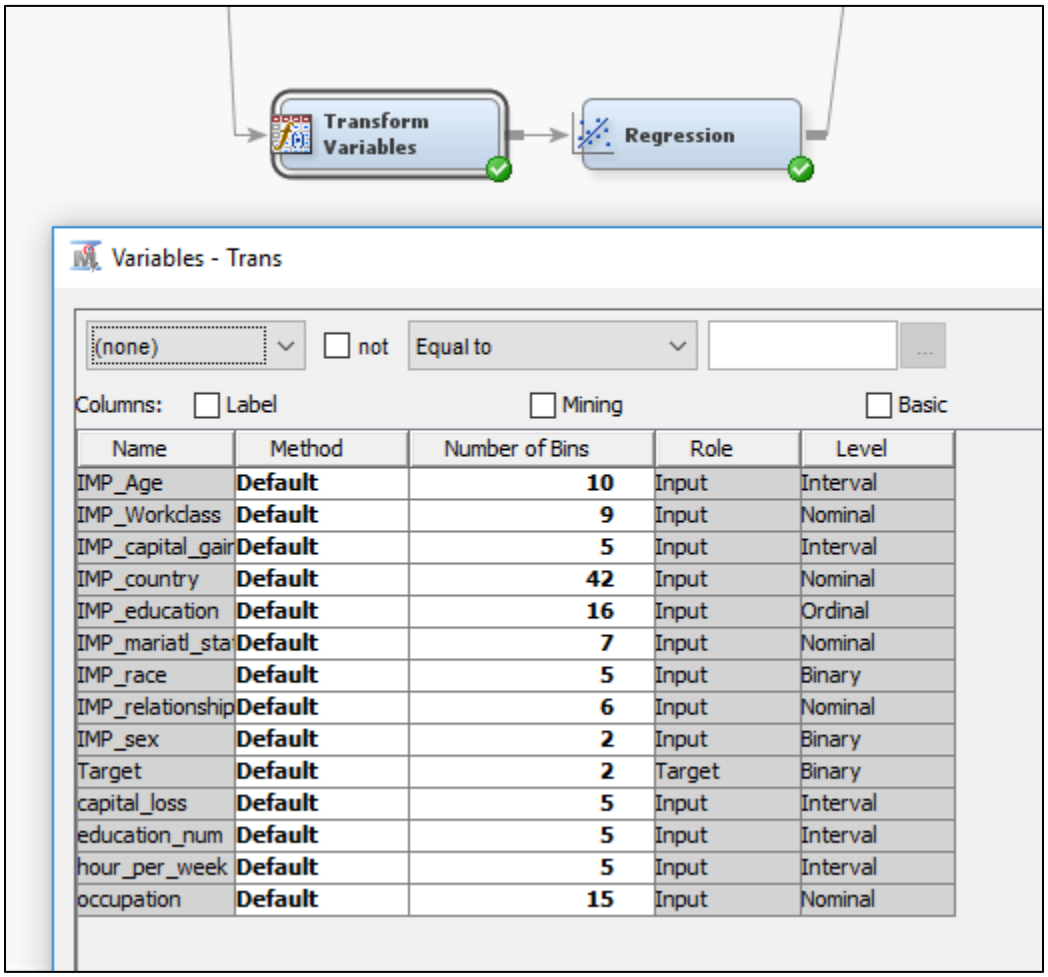


Figure 22: Decision Tree Split Criteria

9.2 Logistic Regression

For logistic regression, a transform node was introduced before the logistic regression node. The transform node was used to stabilize variance, remove nonlinearity, improve additivity, and counter non-normality.



Name	Method	Number of Bins	Role	Level
IMP_Age	Default	10	Input	Interval
IMP_Workclass	Default	9	Input	Nominal
IMP_capital_gair	Default	5	Input	Interval
IMP_country	Default	42	Input	Nominal
IMP_education	Default	16	Input	Ordinal
IMP_mariatl_sta	Default	7	Input	Nominal
IMP_race	Default	5	Input	Binary
IMP_relationship	Default	6	Input	Nominal
IMP_sex	Default	2	Input	Binary
Target	Default	2	Target	Binary
capital_loss	Default	5	Input	Interval
education_num	Default	5	Input	Interval
hour_per_week	Default	5	Input	Interval
occupation	Default	15	Input	Nominal

Figure 23: Transform variable criteria

For the nominal value, the numbers of bins were set based on the total number of nominal value for each variables. For interval variables, the Age variable was divided into 10 buckets while the remaining interval variables (capital gain, capital loss, education num and hour per week) were divide into 5 buckets.

For Regression Node, other than Regression type = logistic, all other parameter will use the default setting.

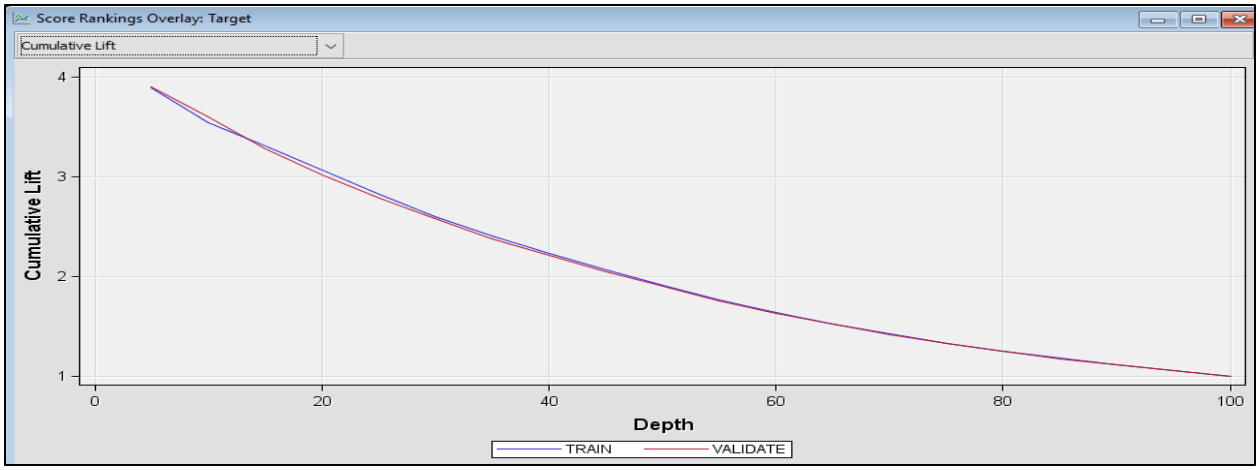


Figure 24: Score Ranking Overlay:Target

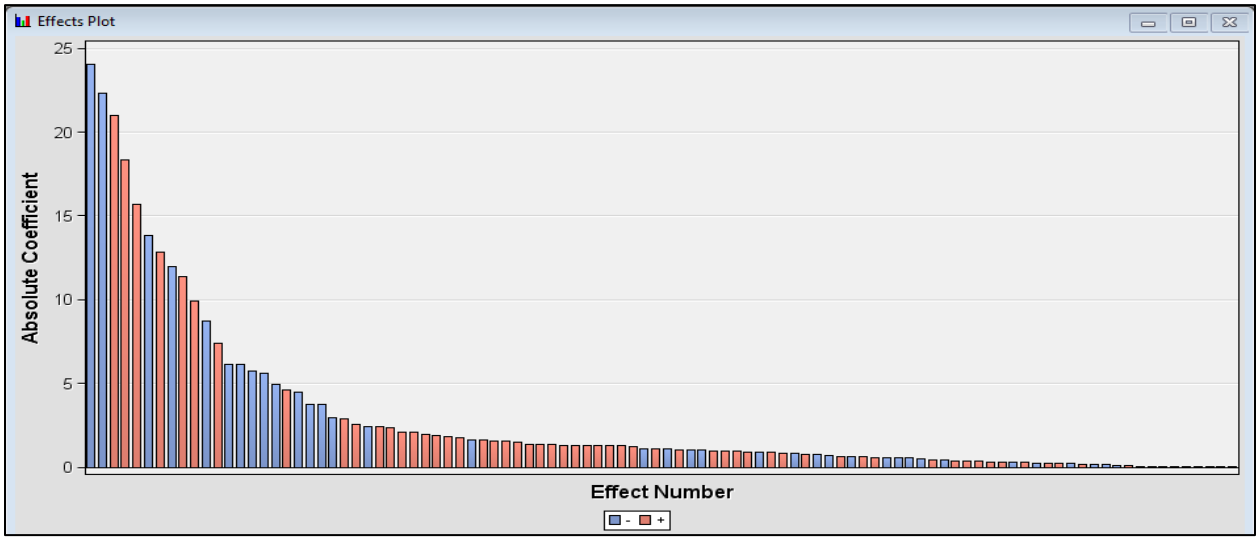


Figure 25: Effect Plot

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Target		_AIC_	Akaike's Information C...	8425.033	.	.
Target		_ASE_	Average Squared Error	0.101303	0.103135	0.101545
Target		_AVERR_	Average Error Function	0.315789	0.323838	0.322088
Target		_DFE_	Degrees of Freedom f...	12923	.	.
Target		_DFM_	Model Degrees of Fre...	100	.	.
Target		_DFT_	Total Degrees of Free...	13023	.	.
Target		_DIV_	Divisor for ASE	26046	19536	19540
Target		_ERR_	Error Function	8225.033	6326.49	6293.598
Target		_FPE_	Final Prediction Error	0.10287	.	.
Target		_MAX_	Maximum Absolute Err...	0.999892	0.999901	0.999997
Target		_MSE_	Mean Square Error	0.102086	0.103135	0.101545
Target		_NOBS_	Sum of Frequencies	13023	9768	9770
Target		_NW_	Number of Estimate ...	100	.	.
Target		_RASE_	Root Average Sum of ...	0.318281	0.321146	0.318662
Target		_RFPE_	Root Final Prediction ...	0.320734	.	.
Target		_RMSE_	Root Mean Squared E...	0.31951	0.321146	0.318662
Target		_SBC_	Schwarz's Bayesian C...	9172.48	.	.
Target		_SSE_	Sum of Squared Errors	2638.526	2014.839	1984.195
Target		_SUMW_	Sum of Case Weights ...	26046	19536	19540
Target		_MISC_	Misclassification Rate	0.145281	0.150184	0.148516

Figure 26: Fit Statistic

10. Model Performance Node And Score Node

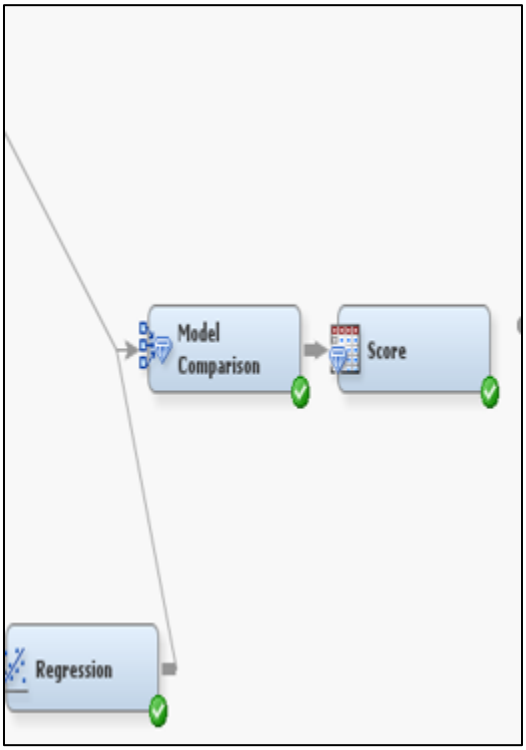


Figure 27:Model Comparison Node And Score Node

We compare the performance for both Decision Tree and Logistic Regression using Model Comparison node.

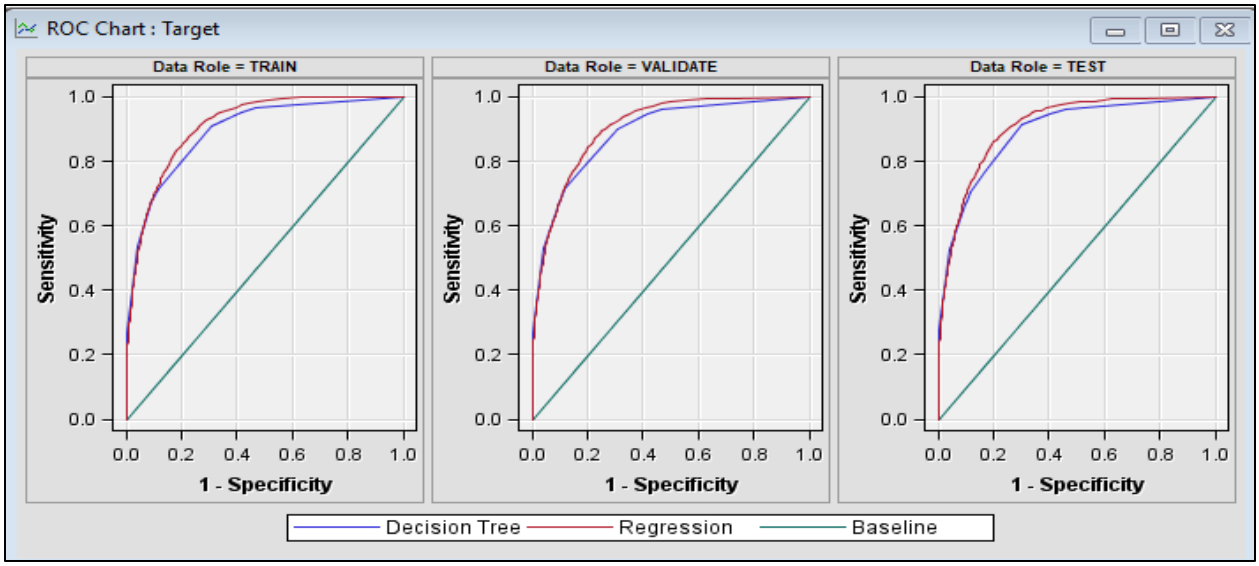


Figure 28: ROC Chart Decision Tree Vs Logistic Regression

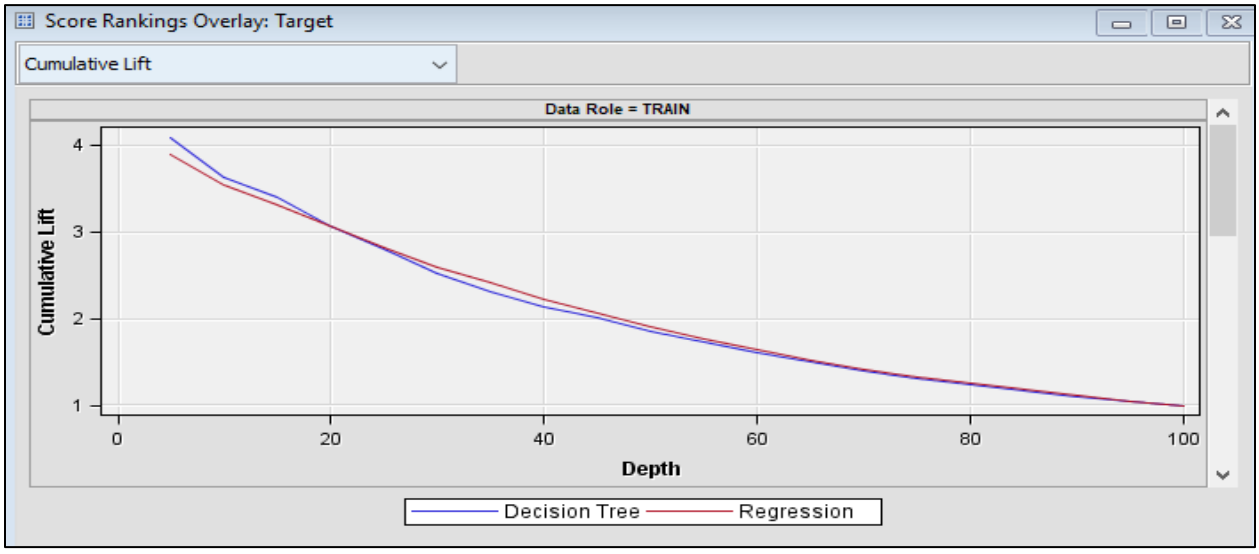


Figure 29: Score Ranking Overlay: Target

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error
Decision Tr...	Target		0.147011	.	0.102194	.	.	.	13023	26046	.	.	0.996528
Regression	Target		0.150184	8425.033	0.101303	0.315789	12923	100	13023	26046	8225.033	0.10287	0.999892

Figure 30: Fit Statistic Decision Tree Vs Regression Tree

10.1 Model Comparison

Based on the Performance of the ROC curve for both model, the accuracy of both model is around 0.90 which is consider good (above 0.90 is consider excellent) with Logistic Regression perform slightly better.

10.2 Fit Statistic

The Valid. Misclassification rate for Decision Tree was 0.147 while for the Logistic Regression was 0.150. We can say that the Decision tree perform 0.3% better than logistic regression under this matrix.

Fit Statistics Table
Target: Target

Data Role=Train

Statistics	Tree	Reg
Train: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.24	0.28
Train: Kolmogorov-Smirnov Statistic	0.60	0.65
Train: Akaike's Information Criterion	.	8425.03
Train: Average Squared Error	0.10	0.10
Train: Roc Index	0.89	0.91
Train: Average Error Function	.	0.32
Train: Cumulative Percent Captured Response	36.33	35.43
Train: Percent Captured Response	15.86	15.94
Selection Criterion: Valid: Misclassification Rate	0.15	0.15
Train: Degrees of Freedom for Error	.	12923.00
Train: Model Degrees of Freedom	.	100.00
Train: Total Degrees of Freedom	13023.00	13023.00
Train: Divisor for ASE	26046.00	26046.00
Train: Error Function	.	8225.03
Train: Final Prediction Error	.	0.10
Train: Gain	263.11	254.08
Train: Gini Coefficient	0.78	0.82
Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.60	0.65
Train: Kolmogorov-Smirnov Probability Cutoff	0.11	0.22
Train: Cumulative Lift	3.63	3.54
Train: Lift	3.17	3.19
Train: Maximum Absolute Error	1.00	1.00
Train: Misclassification Rate	0.14	0.15
Train: Mean Square Error	.	0.10
Train: Sum of Frequencies	13023.00	13023.00
Train: Number of Estimate Weights	.	100.00
Train: Root Average Sum of Squares	0.32	0.32
Train: Cumulative Percent Response	87.44	85.26
Train: Percent Response	76.41	76.80
Train: Root Final Prediction Error	.	0.32
Train: Root Mean Squared Error	.	0.32
Train: Schwarz's Bayesian Criterion	.	9172.48
Train: Sum of Squared Errors	2661.74	2638.53
Train: Sum of Case Weights Times Freq	.	26046.00

Figure 31: Fit Statistic Table for train data set

Data Role=Valid

Statistics	Tree	Reg
Valid: Kolmogorov-Smirnov Statistic	0.60	0.64
Valid: Average Squared Error	0.10	0.10
Valid: Roc Index	0.89	0.90
Valid: Average Error Function	.	0.32
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.29	0.22
Valid: Cumulative Percent Captured Response	35.76	36.01
Valid: Percent Captured Response	15.40	16.50
Valid: Divisor for VASE	19536.00	19536.00
Valid: Error Function	.	6326.49
Valid: Gain	257.48	260.05
Valid: Gini Coefficient	0.77	0.81
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.60	0.64
Valid: Kolmogorov-Smirnov Probability Cutoff	0.25	0.21
Valid: Cumulative Lift	3.57	3.60
Valid: Lift	3.08	3.30
Valid: Maximum Absolute Error	1.00	1.00
Valid: Misclassification Rate	0.15	0.15
Valid: Mean Square Error	.	0.10
Valid: Sum of Frequencies	9768.00	9768.00
Valid: Root Average Squared Error	0.32	0.32
Valid: Cumulative Percent Response	86.08	86.69
Valid: Percent Response	74.20	79.51
Valid: Root Mean Square Error	.	0.32
Valid: Sum of Square Errors	2048.06	2014.84
Valid: Sum of Case Weights Times Freq	.	19536.00

Figure 32: Fit Statistic Table for valid data set

Data Role=Test

Statistics	Tree	Reg
Test: Kolmogorov-Smirnov Statistic	0.61	0.66
Test: Average Squared Error	0.10	0.10
Test: Roc Index	0.89	0.91
Test: Average Error Function	.	0.32
Test: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.18	0.28
Test: Cumulative Percent Captured Response	36.09	36.17
Test: Percent Captured Response	15.61	16.53
Test: Divisor for TASE	19540.00	19540.00
Test: Error Function	.	6293.60
Test: Gain	260.94	261.67
Test: Gini Coefficient	0.78	0.81
Test: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.61	0.65
Test: Kolmogorov-Smirnov Probability Cutoff	0.11	0.23
Test: Cumulative Lift	3.61	3.62
Test: Lift	3.12	3.31
Test: Maximum Absolute Error	1.00	1.00
Test: Misclassification Rate	0.15	0.15
Test: Lower 95% Conf. Limit for TMISC	.	0.14
Test: Upper 95% Conf. Limit for TMISC	.	0.16
Test: Mean Square Error	.	0.10
Test: Sum of Frequencies	9770.00	9770.00
Test: Root Average Squared Error	0.32	0.32
Test: Cumulative Percent Response	86.93	87.10
Test: Percent Response	75.26	79.71
Test: Root Mean Square Error	.	0.32
Test: Sum of Square Errors	2024.20	1984.20
Test: Sum of Case Weights Times Freq	19540.00	19540.00

Figure 33:Fit Statistic Table for test data set

Event Classification Table

Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Reg	Regression	TRAIN	Target		1219	9214	673	1917
Reg	Regression	VALIDATE	Target		920	6869	547	1432
Tree	Decision Tree	TRAIN	Target		1469	9495	392	1667
Tree	Decision Tree	VALIDATE	Target		1110	7090	326	1242

Figure 34: Classification Table Decision Tree Vs Logistic Regression

Score Node

Score Input Variables

Variable Name	Role	Creator	Comment	Label	Variable Hidden	Used in Score Code
Age	INPUT				Y	Y
Target	TARGET				N	N
capital_gain	INPUT			capital-gain	Y	Y
capital_loss	INPUT			capital-loss	N	Y
education_num	INPUT			education-num	N	Y
hour_per_week	INPUT			hour-per-week	N	Y
occupation	INPUT				N	Y
relationship	INPUT				Y	Y

Score Output Variables

Variable Name	Function	Creator	Label
EM_CLASSIFICATION	CLASSIFICATION	Score	Prediction for Target
EM_EVENTPROBABILITY	PREDICT	Score	Probability for level >50K of Target
EM_PROBABILITY	PREDICT	Score	Probability of Classification
EM_SEGMENT	TRANSFORM	Score	Node
IMP_Age	TRANSFORM	Impt	Imputed Age
IMP_capital_gain	TRANSFORM	Impt	Imputed: capital-gain
IMP_relationship	TRANSFORM	Impt	Imputed relationship
I_Target	CLASSIFICATION	Tree	Into: Target
P_Target_50K	PREDICT	Tree	Predicted: Target=>50K
P_Target__50K	PREDICT	Tree	Predicted: Target=<=50K
Q_Target_50K	TRANSFORM	Tree	Unadjusted P: Target=>50K
Q_Target__50K	TRANSFORM	Tree	Unadjusted P: Target=<=50K
U_Target	CLASSIFICATION	Tree	Unnormalized Into: Target
V_Target_50K	PREDICT	Tree	Validated: Target=>50K
V_Target__50K	PREDICT	Tree	Validated: Target=<=50K
NODE	TRANSFORM	Tree	Node
WARN	ASSESS	Tree	Warnings

Figure 35: Score Input Node Summary

Class Variable Summary Statistics

Data Role=TEST Output Type=CLASSIFICATION

Variable	Numeric Value	Formatted Value	Frequency Count	Percent
I_Target	.	<=50K	8234	84.2784
I_Target	.	>50K	1536	15.7216

Data Role=TEST Output Type=SEGMENT

Variable	Numeric Value	Formatted Value	Frequency Count	Percent
NODE	7	7	96	0.9826
NODE	9	9	149	1.5251
NODE	11	11	223	2.2825
NODE	13	13	286	2.9273
NODE	24	24	4102	41.9857
NODE	30	30	248	2.5384
NODE	32	32	437	4.4729
NODE	34	34	166	1.6991
NODE	35	35	14	0.1433
NODE	37	37	119	1.2180
NODE	40	40	853	8.7308
NODE	41	41	5	0.0512
NODE	46	46	751	7.6868
NODE	47	47	40	0.4094
NODE	50	50	1324	13.5517
NODE	51	51	45	0.4606
NODE	52	52	67	0.6858
NODE	53	53	845	8.6489

Data Role=TRAIN Output Type=CLASSIFICATION

Variable	Numeric Value	Formatted Value	Frequency Count	Percent
I_Target	.	<=50K	10964	84.1895
I_Target	.	>50K	2059	15.8105

Figure 36: Class Variable Summary Statistics

Interval Variable Summary Statistics

Variable Name=P_Target_50K

Statistics	Label	TRAIN	VALIDATE	TEST
MEAN	Mean	0.24	0.24	0.24
STD	Standard Deviation	0.28	0.29	0.29
N	Non Missing	13023.00	9768.00	9770.00
MIN	Minimum	0.02	0.02	0.02
P25	25th Percentile	0.02	0.02	0.02
MEDIAN	Median	0.11	0.11	0.11
P75	75th Percentile	0.34	0.34	0.34
MAX	Maximum	1.00	1.00	1.00

Variable Name=P_Target__50K

Statistics	Label	TRAIN	VALIDATE	TEST
MEAN	Mean	0.76	0.76	0.76
STD	Standard Deviation	0.28	0.29	0.29
N	Non Missing	13023.00	9768.00	9770.00
MIN	Minimum	0.00	0.00	0.00
P25	25th Percentile	0.66	0.66	0.66
MEDIAN	Median	0.89	0.89	0.89
P75	75th Percentile	0.98	0.98	0.98
MAX	Maximum	0.98	0.98	0.98

Figure 37: Interval Variable Summary Statistic

11. Performance Assessment Validation Set and Test Set under Fit Statistic

We selected several matrices from Fit Statistic Table to assess the performance of our model

	Performance Direction	**Decision Tree**	**Logistic Regression**
Validate Set			
• Kolmogorov-Smirnov Statistic	Largest	0.60	0.64
• ROC Index	Largest	0.89	0.90
• Gini Coefficient	Largest	0.77	0.81
• Average Square Error	Smallest	0.10	0.10
Test Set			
• Kolmogorov-Smirnov Statistic	Largest	0.61	0.66
• ROC Index	Largest	0.89	0.91
• Gini Coefficient	Largest	0.78	0.81
• Average Square Error	Smallest	0.10	0.10

Table 1: Fit Statistic Table Performance Measurement

Based on the selected matrices, we can concluded Logistic Regression has slight better performance than the Decision Tree.

12. Conclusion

Based on the performance of our Fit Statistic Matrix, we concluded that both of our model (decision tree and Logistic Regression) display a good performance with Logistic Regression perform slightly better than our logistic regression.