Data: Holiday Package Prediction

07-15-2022

—

Group 3 - BA706 - Applied Analytic Modeling

Olena Dovzhenko

Zhiguang Guan

Risikat Hameed

Daisy Johanna Uy

Table of Contents

Description of the Problem

We need to analyze the customers' data to build a model to predict the potential customer who is going to purchase the newly introduced travel package. In the past, marketing costs were high because customers were contacted at random without looking at the available information. The company now wants to harness the available data of existing and potential customers to make the marketing expenditure more efficient.

File Import

Because our data is in a comma-separated values (CSV) file, we first imported the file into SAS Enterprise Miner. We created the diagram and added an Import node ('Project').

We imported the CSV file (under Train section of the Property panel on the right), then saved it as a SAS dataset with the Export node ('Save Data'). This allows us to easily connect the Export Nodel

Property	Value
General	
Node ID	FIMPORT
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Import File	H:\BA706\Travel.csv
Maximum rows to import	
Maximum columns to impor	10000
Delimiter	,
Name Row	Yes
Number of rows to skip	0
Guessing Rows	500
File Location	Local
File Type	csv
Advanced Advisor	No
Rerun	No
Score	
Role	Train
Report	
Summarize	No
Status	
Create Time	14/12/21 9:21 AM
Run ID	8a8d6297-66ec-4b30-b133-
Last Error	
Last Status	Complete
Last Run Time	14/12/21 9:28 AM
Run Duration	0 Hr. 0 Min. 4.07 Sec.
Grid Host	
User-Added Node	No

At this point, you can start your analysis of the data by dragging and dropping additional nodes. However, it is advisable to save your file as a SAS dataset in order to connect using the Data Sources in the Task Tree.

The target variable for our Holiday Package Prediction dataset is "ProdTaken". It is a binary variable, which identifies whether the customer has purchased the product: 0 - no; 1 - yes.

There are 19 input variables in the original (raw) dataset provided. We did not see signs of potential data leakage or duplication between the data.

Data Wrangling

After analysing the data, we rejected 5 variables:

1. CityTier - indicates the level of the destination city's development.
   Irrelevant, gives no valuable information to the model.
2. DurationOfPitch - duration of the marketing pitch.
   Redundant, gives no valuable information to the model.

3. OwnCar - identifies whether the customer has a car.
   Redundant given the presence of the Monthly Income variable, gives no valuable information to the model. (Note: We did try to run the decision tree model with this variable during our experimentation, but it was not an important variable. This confirmed its redundancy.)

4. PitchSatisfactionScore - Customer's satisfaction with marketing pitch. Can be from 1 to 5.
   Highly correlated with the Target variable.

5. ProductPitched - Advertised product during the marketing pitch.
   Gives no valuable information to the model.

Among the 14 variables left, 6 are Categorical (Nominal) variables:

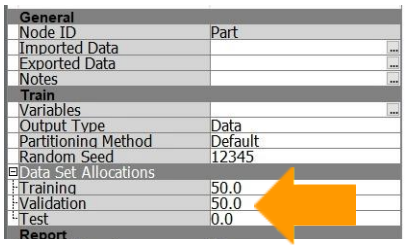
1. Designation:
   a. AVP (Assistant Vice President)
   b. Executive
   c. Manager
   d. Senior Manager
   e. VP (Vice President)
2. Gender:
   a. Female
   b. Male
   c. Undisclosed (Noted as Fe_Male)
3. MaritalStatus:
   a. Divorced
   b. Married
   c. Single
   d. Unmarried
4. Occupation:
   a. Freelance
   b. Large Business
   c. Salaried
   d. Small Business
5. ProductPitched
   a. Basic
   b. Deluxe
   c. King
   d. Standard
   e. Super Deluxe
6. TypeofContact
   a. Company invited
   b. Self Enquiry

We have collapsed the MaritalStatus variable with a Replacement node as categories “Unmarried” and “Single” are equal (settings seen below).

ProdTaken	_UNKNOWN_	_DEFAULT_	.	N	
REP_MaritalStatus	Married		1148	C	Married
REP_MaritalStatus	Single		811	C	Single
REP_MaritalStatus	Divorced		484	C	Divorced
REP_MaritalStatus	_UNKNOWN_	_DEFAULT_	.	C	

We then explored the source data, noting that it would be useful if the company could provide the information whether these trips are domestic or international. We assume that would highly increase the accuracy of the model as it will identify whether the customer should have a passport in order to enjoy the trip package.

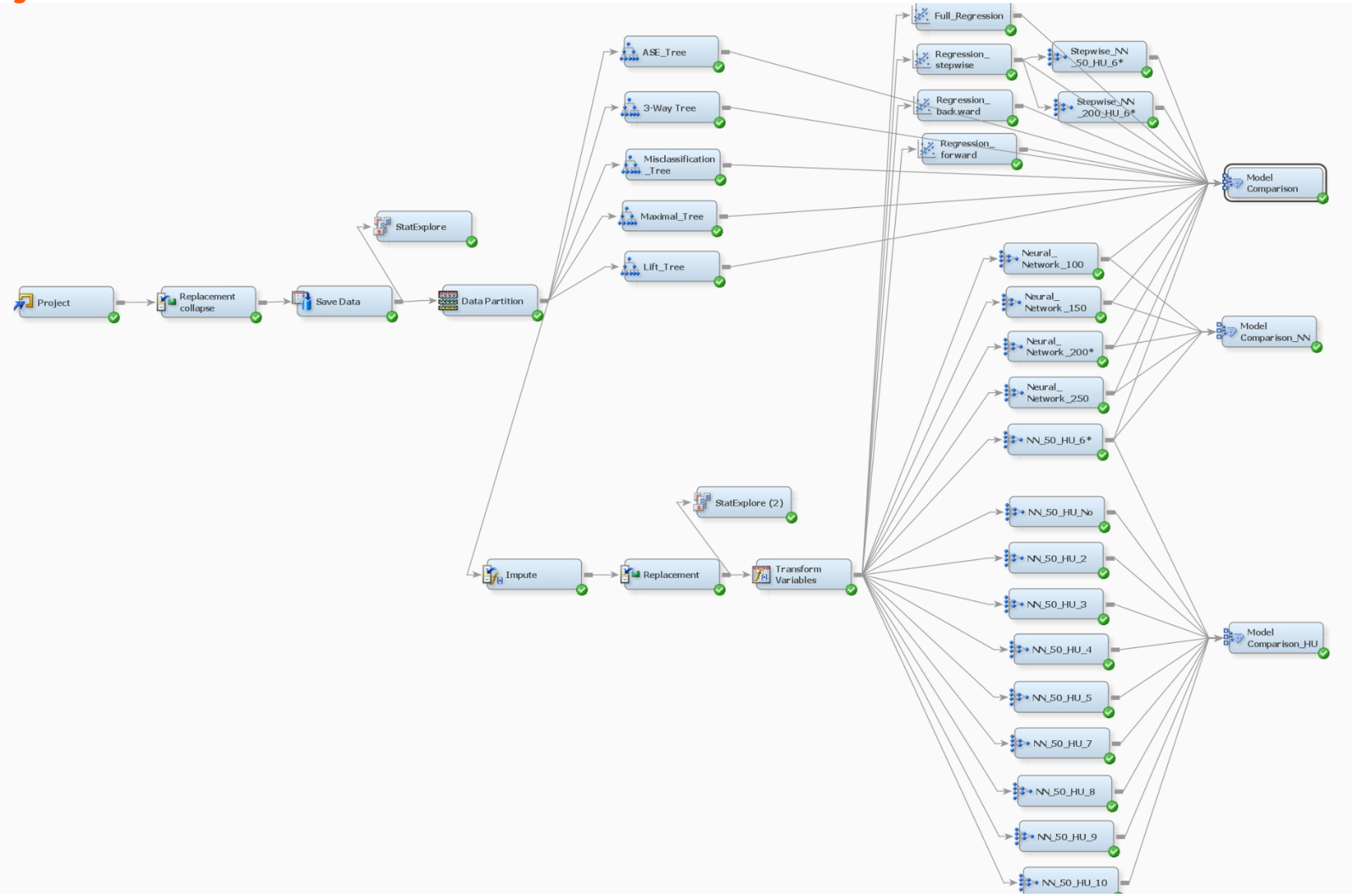
We also noted that we have 4,888 observations in the data set, along with missing values. We partitioned the data with a Data Partition node, and assigned 50% of the data for training and the other 50% for validation.



Data dictionary

#	Name	Model Role	Measurement Level	Description
1	Age	Input	Interval	Age of the Customer
2	CityTier	Rejected	Interval	Level of destination city development. Rejected
3	CustomerID	ID	Interval	Unique Customer ID
4	Designation	Input	Nominal	Customer's job title
5	DurationOfPitch	Rejected	Interval	Duration of the pitch. Rejected
6	Gender	Input	Nominal	Gender of the Customer
7	MaritalStatus	Input	Nominal	Marital status of the customer
8	MonthlyIncome	Input	Interval	Customer's monthly income
9	NumberOfChildrenVisiting	Input	Interval	Number of children who are supposed to join the trip
10	NumberOfFollowups	Input	Interval	Number of outreach interactions after initial pitch
11	NumberOfPersonVisiting	Input	Interval	Number of participants in the trip
12	NumberOfTrips	Input	Interval	Number of trips taken
13	Occupation	Input	Nominal	Type of Customer's employment
14	OwnCar	Rejected	Binary	Identifies whether the Customer has a car
15	Passport	Input	Binary	Identifies whether Customer has a passport when pitched
16	PitchSatisfactionScore	Rejected	Interval	Customer's satisfaction with marketing pitch. Can be from 1 to 5
17	PreferredPropertyStar	Input	Interval	Preferred property class. Can be from 1 to 5 (stars)
18	ProdTaken	Target	Binary	Identifies whether the customer has purchased the product. Values can be 0 or 1
19	ProductPitched	Rejected	Nominal	Advertised product during the marketing pitch
20	TypeofContact	Input	Nominal	How was customer interaction initiated?

Diagram


Project
Replacement collapse
Save Data
StatExplore
Data Partition
ASE_Tree
3-Way Tree
Misclassification _Tree
Maximal_Tree
Lift_Tree
Impute
Replacement
StatExplore (2)
Transform Variables
Full_Regression
Regression_ stepwise
Regression_ backward
Regression_ forward
Stepwise_NN _50_HU_6*
Stepwise_NN _200_HU_6*
Model Comparison
Neural_ Network_100
Neural_ Network_150
Neural_ Network_200*
Neural_ Network_250
NN_50_HU_6*
Model Comparison_NN
NN_50_HU_No
NN_50_HU_2
NN_50_HU_3
NN_50_HU_4
NN_50_HU_5
NN_50_HU_7
NN_50_HU_8
NN_50_HU_9
NN_50_HU_10
Model Comparison_HU

Decision Tree

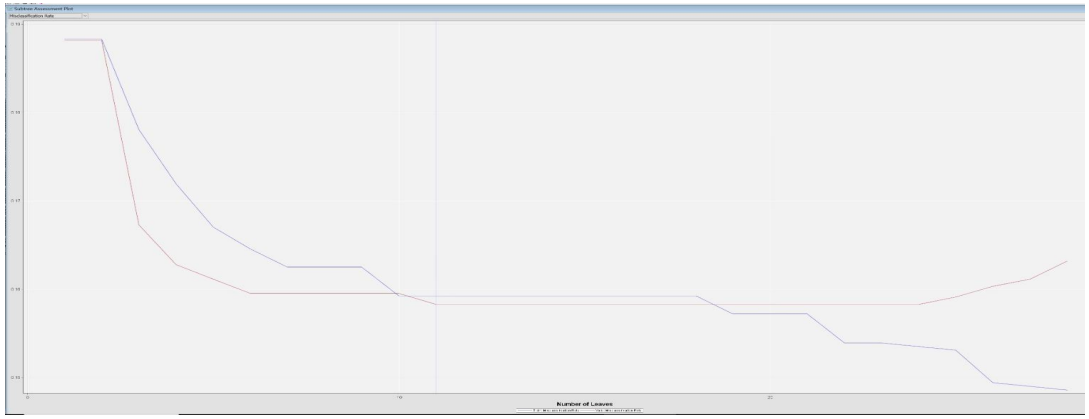
We ran decision tree models (returning trees with the assessment measures below). We froze our models and chose to disable training of nodes.

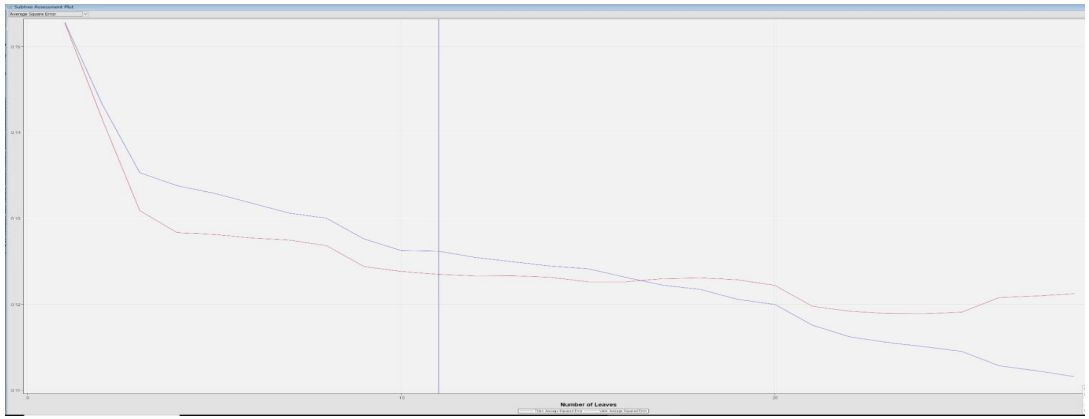
- Maximal Tree (largest average profit and smallest average loss if a profit or loss matrix is defined)
- Probability Tree (smallest average square error)
    - Maximum branch: 2
    - Maximum branch: 3 (referred to as 3-way tree)
- Misclassification Tree (lowest misclassification rate)
- Lift Tree (prediction of the top n% of the ranked observations)

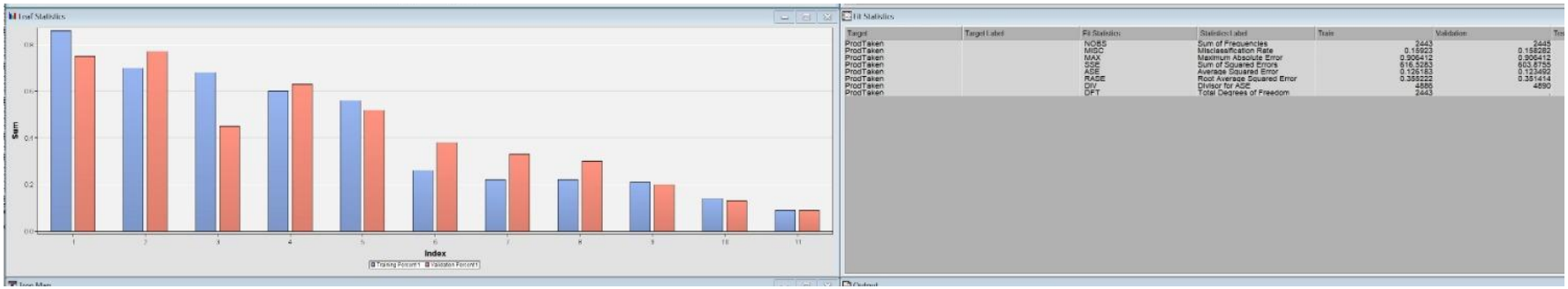
Comparison of Decision Tree Models

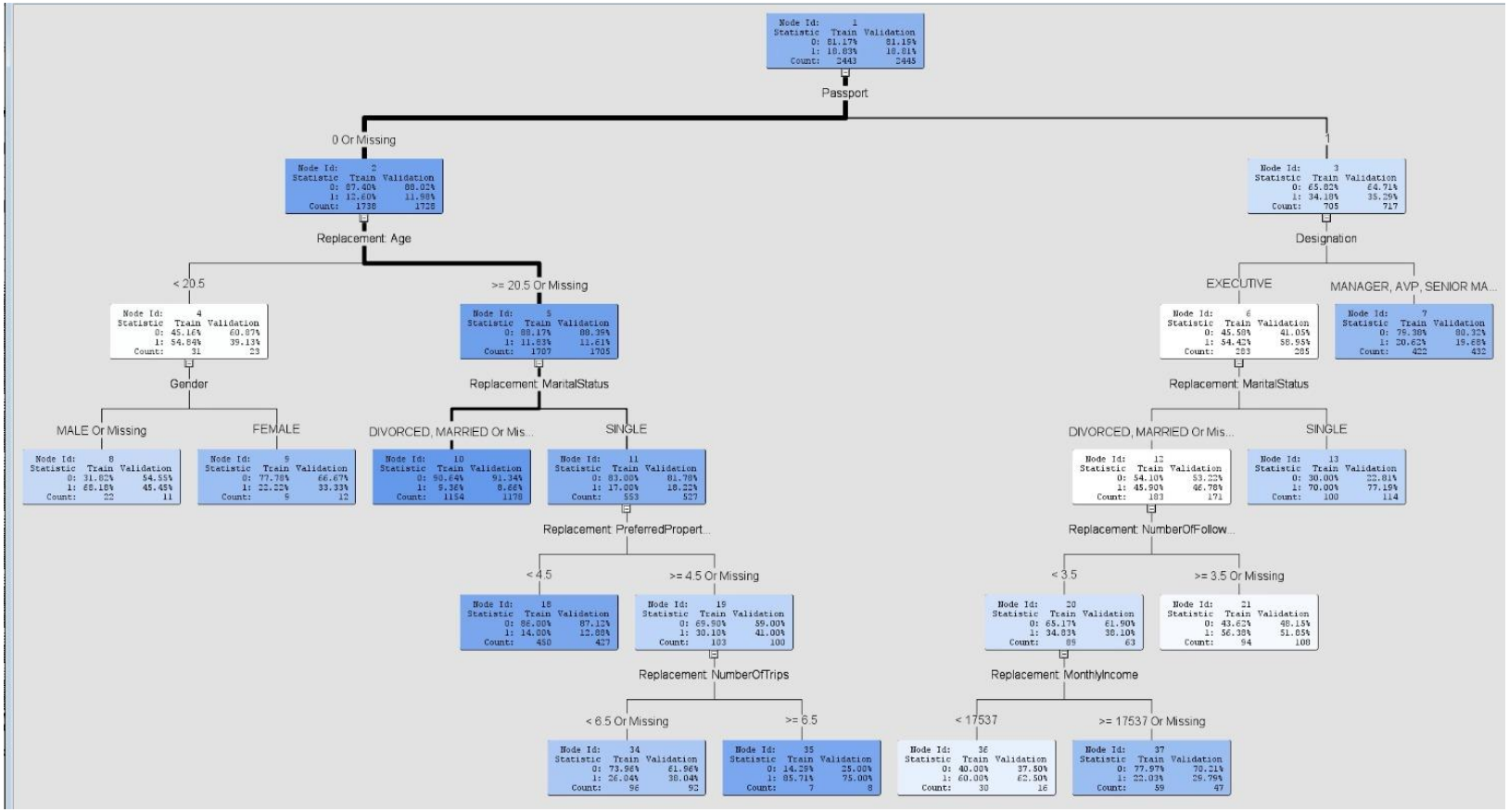
Maximal Tree

Use Frozen Tree	Yes
Use Multiple Targets	No
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 Rule	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






Leaf Statistics
Sum
Index
Fit Statistics
Target
Target Label
Fit Statistics
Statistics Label
Train
Validation
ProdTaken
NOBS
Sum of Frequencies
2443
2445
ProdTaken
MISC
Misclassification Rate
0.158282
ProdTaken
MAX
Maximum Absolute Error
0.906412
0.906412
ProdTaken
SSE
Sum of Squared Errors
616.5263
603.8755
ProdTaken
ASE
Average Squared Error
0.126183
0.123492
ProdTaken
RASE
Root Average Squared Error
0.355222
0.351414
ProdTaken
DIV
Divisor for ASE
4886
4890
ProdTaken
DFT
Total Degrees of Freedom
2443

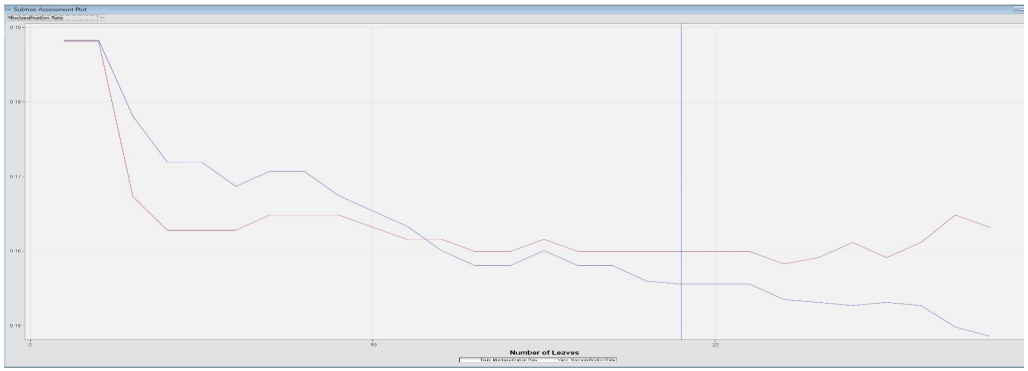

Node Id: 1
Statistic Train Validation
0: 81.17% 81.15%
1: 18.83% 18.81%
Count: 2443 2445
Passport
0 Or Missing
1
Node Id: 2
Statistic Train Validation
0: 87.40% 88.02%
1: 12.60% 11.98%
Count: 1738 1728
Node Id: 3
Statistic Train Validation
0: 65.82% 64.71%
1: 34.18% 35.29%
Count: 705 717
Replacement: Age
Designation
< 20.5
>= 20.5 Or Missing
EXECUTIVE
MANAGER, AVP, SENIOR MA...
Node Id: 4
Statistic Train Validation
0: 45.16% 60.87%
1: 54.84% 39.13%
Count: 31 23
Node Id: 5
Statistic Train Validation
0: 88.17% 88.39%
1: 11.83% 11.61%
Count: 1707 1705
Node Id: 6
Statistic Train Validation
0: 45.58% 41.05%
1: 54.42% 58.95%
Count: 283 285
Node Id: 7
Statistic Train Validation
0: 76.30% 80.32%
1: 30.62% 19.68%
Count: 422 432
Gender
Replacement: MaritalStatus
Replacement: MaritalStatus
MALE Or Missing
FEMALE
DIVORCED, MARRIED Or Mis...
SINGLE
DIVORCED, MARRIED Or Mis...
SINGLE
Node Id: 8
Statistic Train Validation
0: 31.82% 54.55%
1: 68.18% 45.45%
Count: 22 11
Node Id: 9
Statistic Train Validation
0: 77.78% 66.67%
1: 22.22% 33.33%
Count: 9 12
Node Id: 10
Statistic Train Validation
0: 90.64% 91.34%
1: 9.36% 8.66%
Count: 1154 1178
Node Id: 11
Statistic Train Validation
0: 83.00% 81.78%
1: 17.00% 18.22%
Count: 553 527
Node Id: 12
Statistic Train Validation
0: 54.10% 53.22%
1: 45.90% 46.78%
Count: 183 171
Node Id: 13
Statistic Train Validation
0: 30.00% 22.81%
1: 70.00% 77.19%
Count: 100 114
Replacement: PreferredPropert...
Replacement: NumberOfFollow...
< 4.5
>= 4.5 Or Missing
< 3.5
>= 3.5 Or Missing
Node Id: 18
Statistic Train Validation
0: 86.00% 87.12%
1: 14.00% 12.88%
Count: 450 427
Node Id: 19
Statistic Train Validation
0: 69.90% 59.00%
1: 30.10% 41.00%
Count: 103 100
Node Id: 20
Statistic Train Validation
0: 65.17% 61.90%
1: 34.83% 38.10%
Count: 89 63
Node Id: 21
Statistic Train Validation
0: 43.62% 48.15%
1: 56.38% 51.85%
Count: 94 108
Replacement: NumberOfTrips
Replacement: MonthlyIncome
< 6.5 Or Missing
>= 6.5
< 17537
>= 17537 Or Missing
Node Id: 34
Statistic Train Validation
0: 73.96% 61.96%
1: 26.04% 38.04%
Count: 96 92
Node Id: 35
Statistic Train Validation
0: 14.29% 25.00%
1: 85.71% 75.00%
Count: 7 8
Node Id: 36
Statistic Train Validation
0: 40.00% 37.50%
1: 60.00% 62.50%
Count: 30 16
Node Id: 37
Statistic Train Validation
0: 77.97% 70.21%
1: 22.03% 29.79%
Count: 59 47

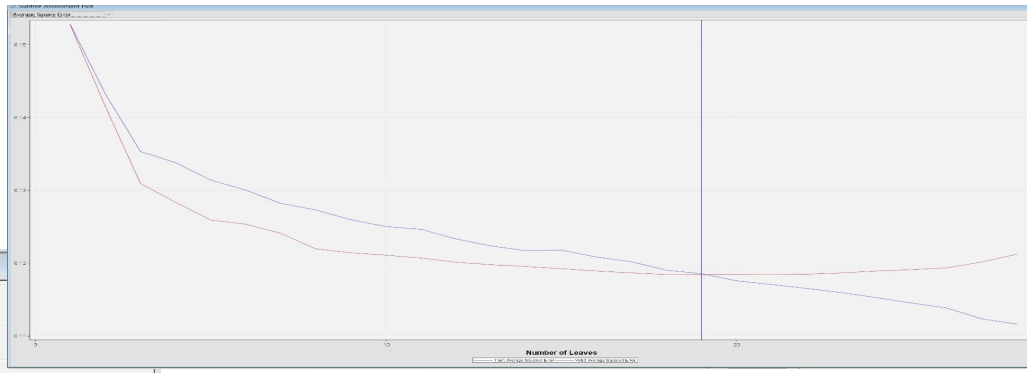
- The optimal number of leaves is 11.
- The variable used for the first split was Passport. The competing splits that were used in the second split were Age and Designation. Other important variables used in the third split were Gender and Marital Status. As can be seen in the tree, variables like Preferred Property Score and Number of Follow-Ups were included but less important.
- Valid average square error (ASE) = 0.123492

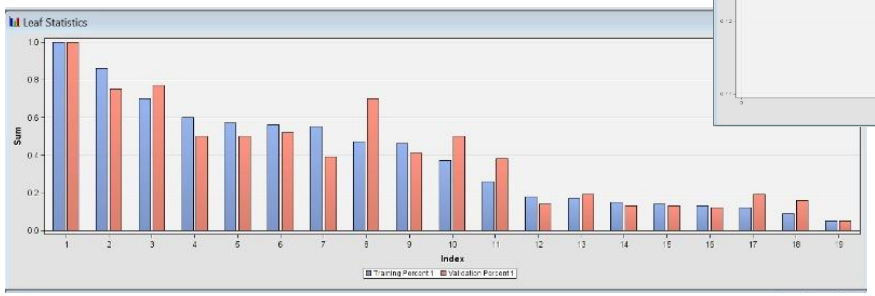
Valid misclassification rate = 0.158282

Probability Tree

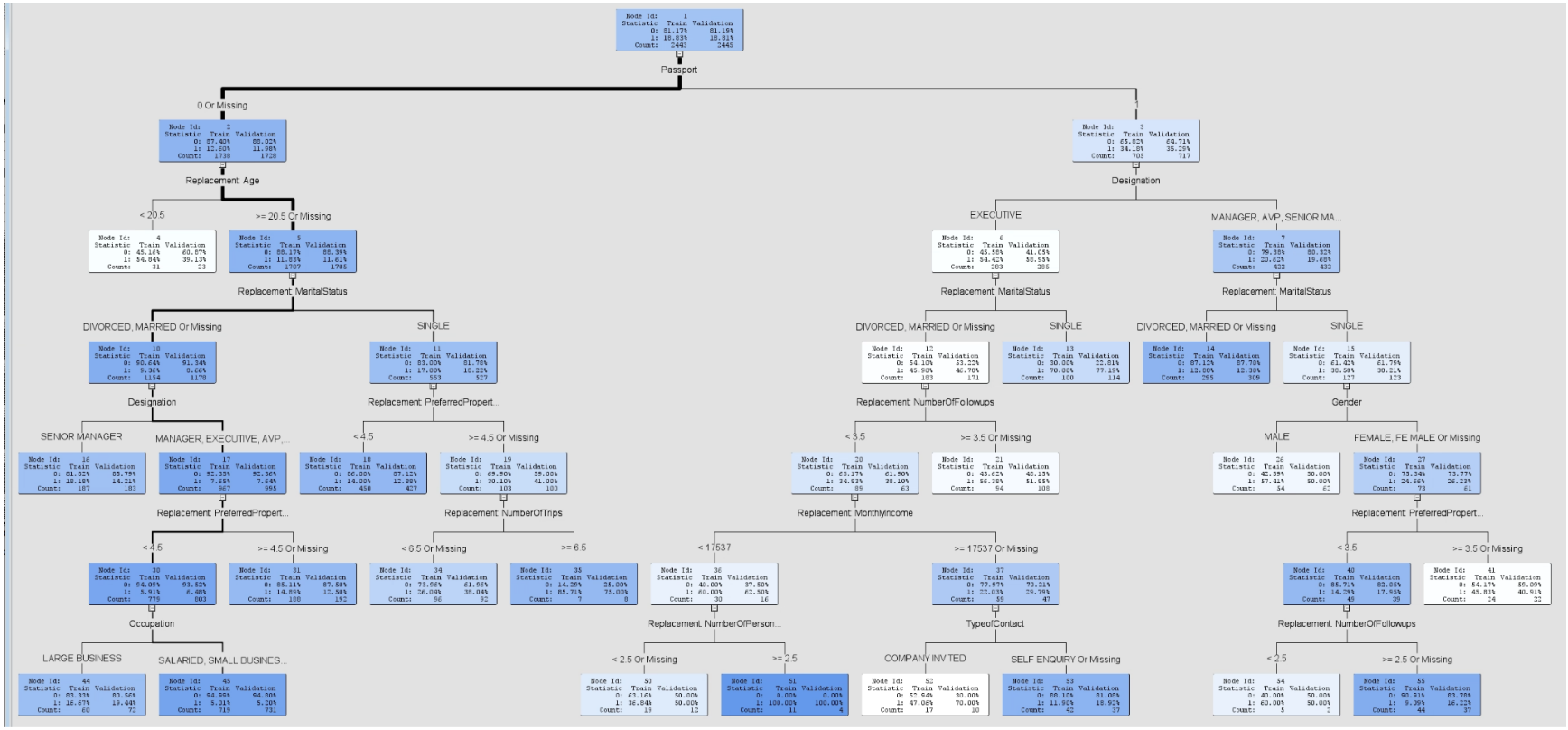
Use Frozen Tree	Yes
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rule	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	Average Square Error
Assessment Fraction	0.25







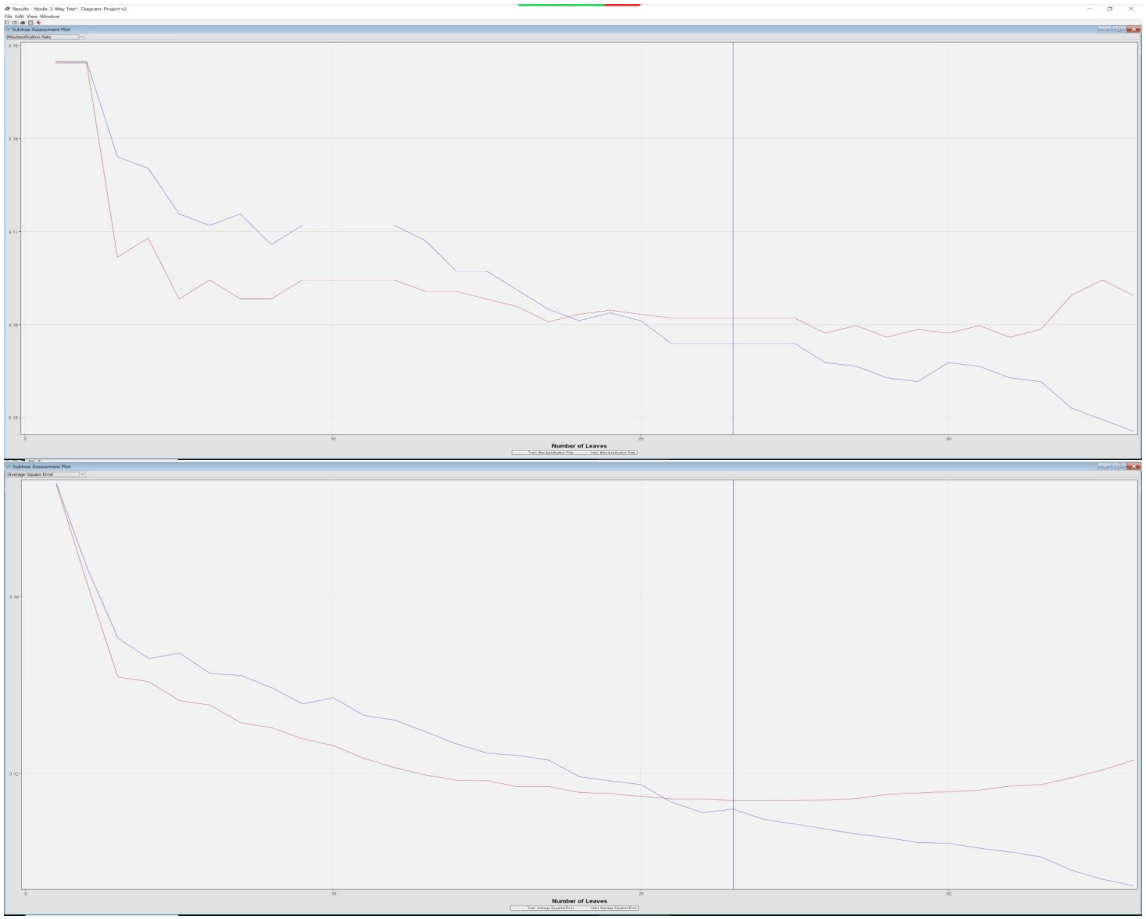
Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	2443	2445
MISC	Misclassification Rate	0.155546	0.159918
MAX	Maximum Absolute Error	0.94993	0.94993
SSE	Sum of Squared Errors	579.3291	579.0059
ASE	Average Squared Error	0.118569	0.118406
RASE	Root Average Squared...	0.344339	0.344102
DIV	Divisor for ASE	4886	4890
DFT	Total Degrees of Free...	2443	

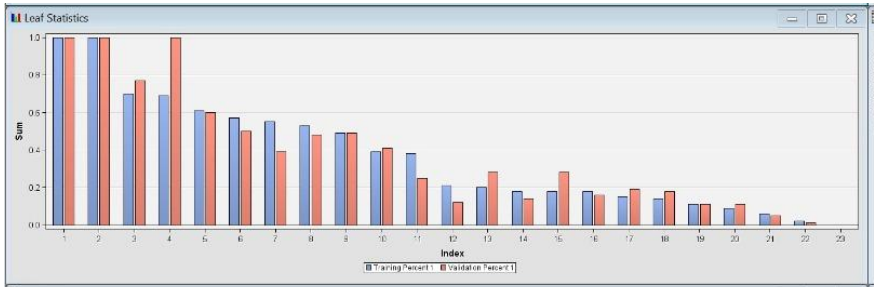


- The optimal number of leaves is 19.
- The variable used for the first split was Passport. The competing splits used for the second split were Age and Designation. For the third split, Marital Status was the variable used. Other important variables are Gender, Preferred Property Score and Number of Follow-ups. As can be seen in the tree, variables like Monthly Income were included but less important.
- Valid average square error (ASE) = 0.118406
- Valid misclassification rate = 0.159918

3-Way Tree

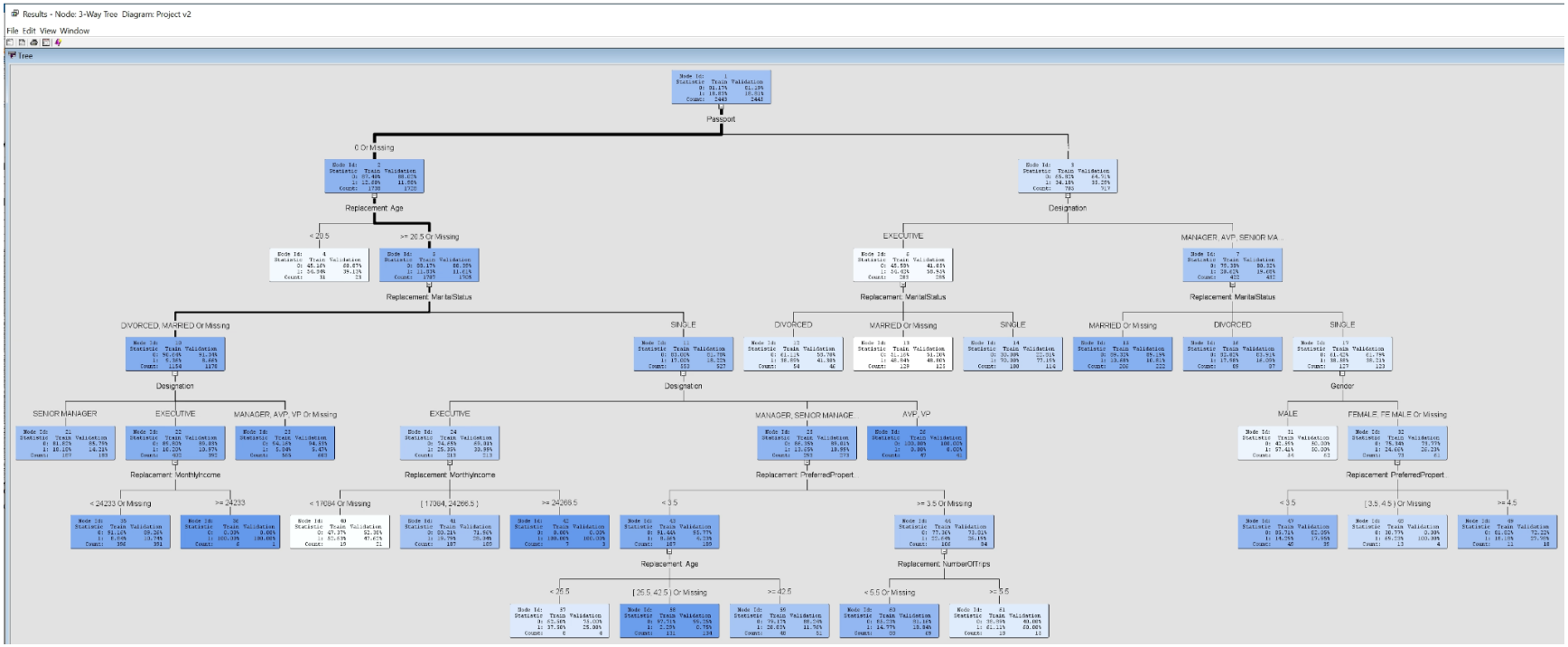
Use Frozen Tree	Yes
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	3
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rule	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	Average Square Error
Assessment Fraction	0.25





Fit Statistics

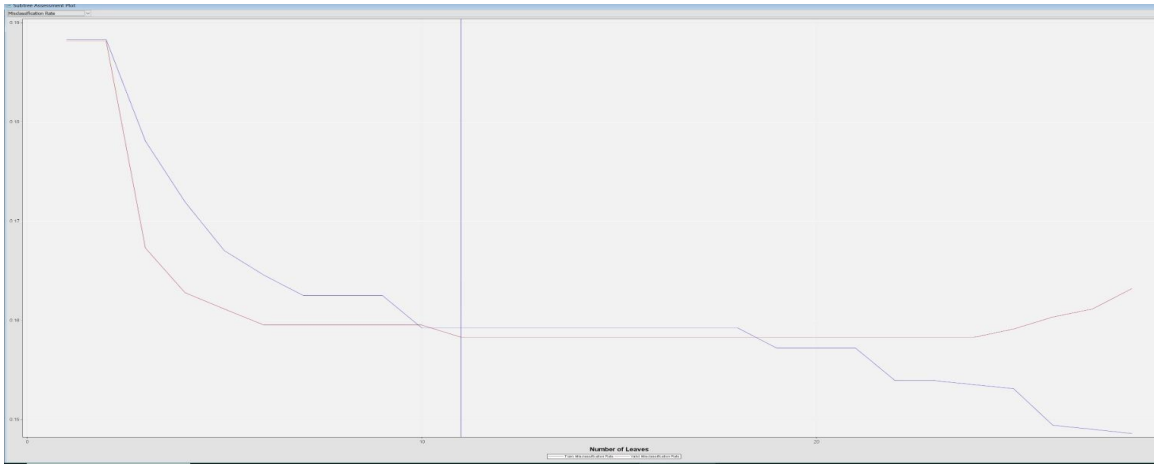
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
ProdTaken		NOBS	Sum of Frequencies	2443	2445
ProdTaken		MISC	Misclassification Rate	0.158002	0.160736
ProdTaken		MAX	Maximum Absolute Error	0.977099	0.977099
ProdTaken		SSE	Sum of Squared Errors	566.8406	572.1074
ProdTaken		ASE	Average Squared Error	0.116013	0.116995
ProdTaken		RASE	Root Average Squared...	0.340607	0.342046
ProdTaken		DIV	Divisor for ASE	4886	4890
ProdTaken		DFT	Total Degrees of Free...	2443	

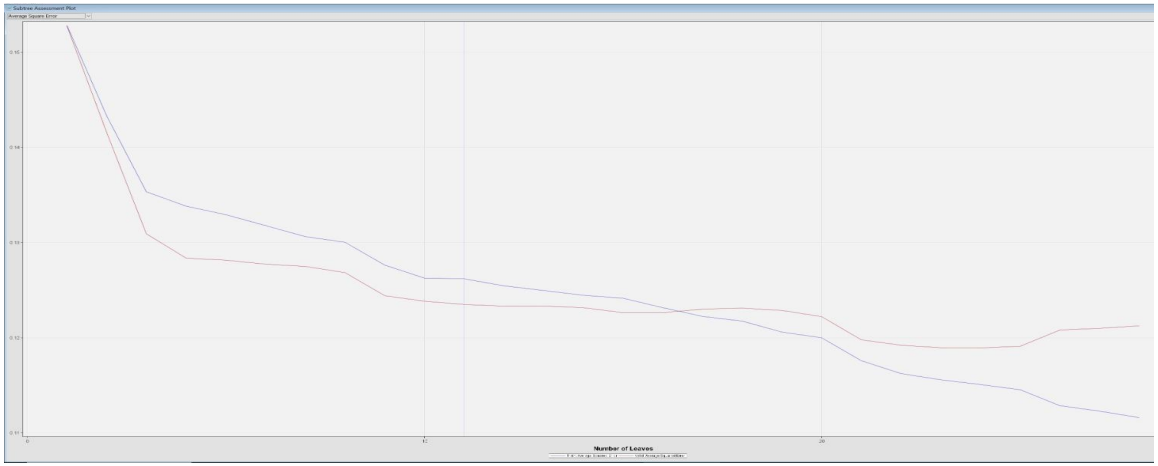


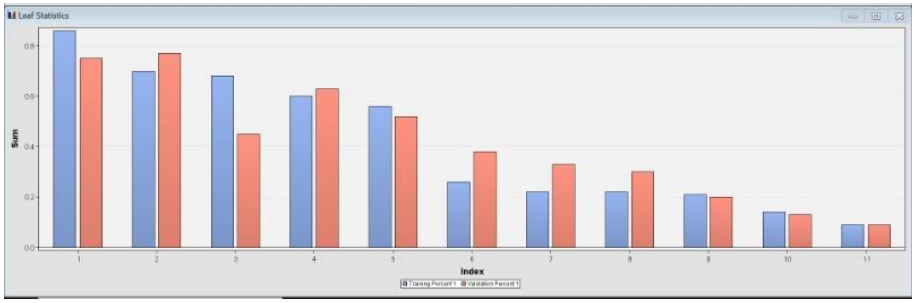
- The optimal number of leaves is 23.
- The variable used for the first split was Passport. The competing splits used in the second split were Age and Designation. Another important variable used in the third split is Marital Status. As can be seen in the tree, variables (Monthly Income, Preferred Property Score and Number of Trips) were included but less important.
- Valid average square error (ASE) = 0.116995
- Valid misclassification rate = 0.160736

Misclassification Tree

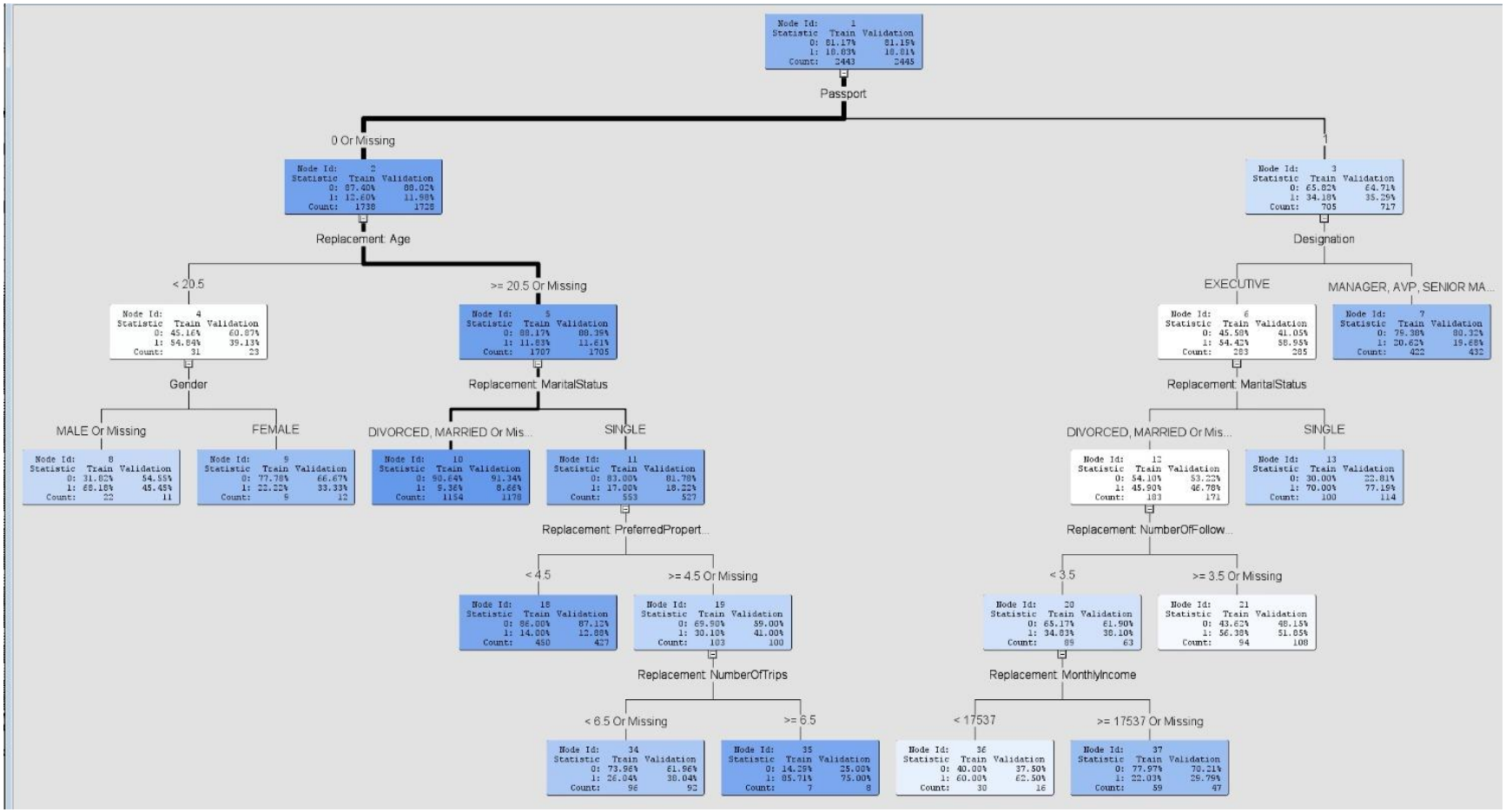
Use Frozen Tree	Yes
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rule	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	Misclassification
Assessment Fraction	0.25







Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	2443	2445
MISC	Misclassification Rate	0.15923	0.158282
MAX	Maximum Absolute Error	0.906412	0.906412
SSE	Sum of Squared Errors	616.5283	603.8766
ASE	Average Squared Error	0.126183	0.123492
RASE	Root Average Squared Error	0.355222	0.351414
DIV	Divisor for ASE	4886	4890
DFT	Total Degrees of Freedom	2443	

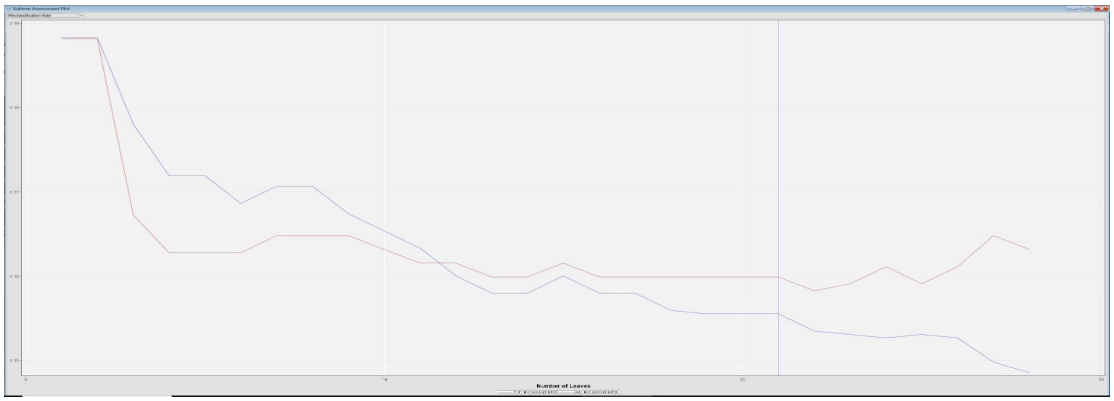


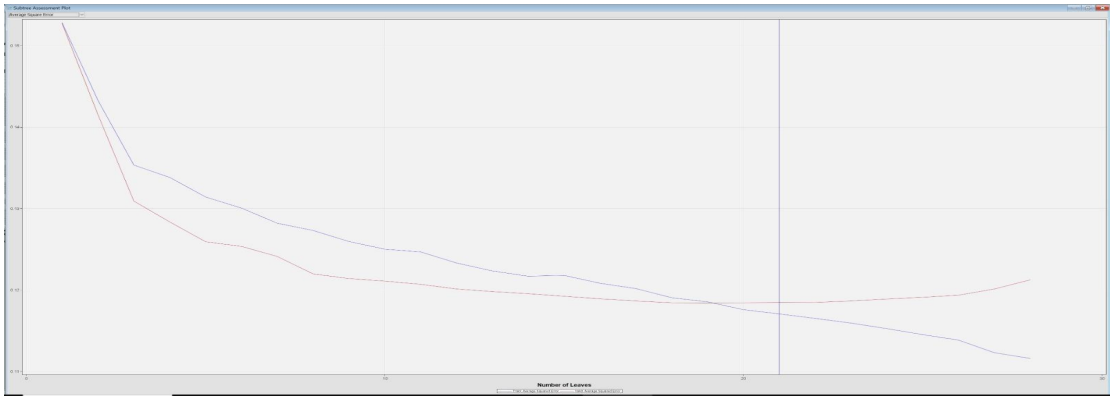
- The optimal number of leaves is 11.
- The variable used for the first split was Passport. The variables used for competing splits were Age and Designation. Other important variables used in the third split were Marital Status and Gender. As can be seen in the tree, variables like Preferred Property Score and Number of Follow-Ups were included but less important.
- Valid average square error (ASE) = 0.123492

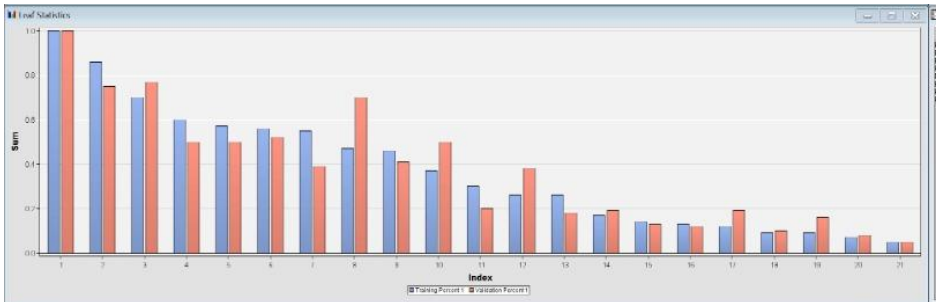
Valid misclassification rate = 0.158282

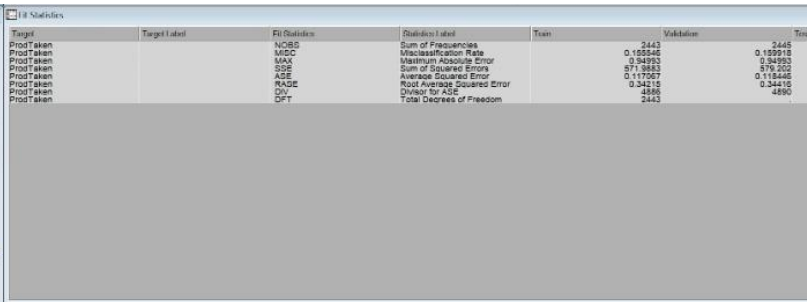
Lift Tree

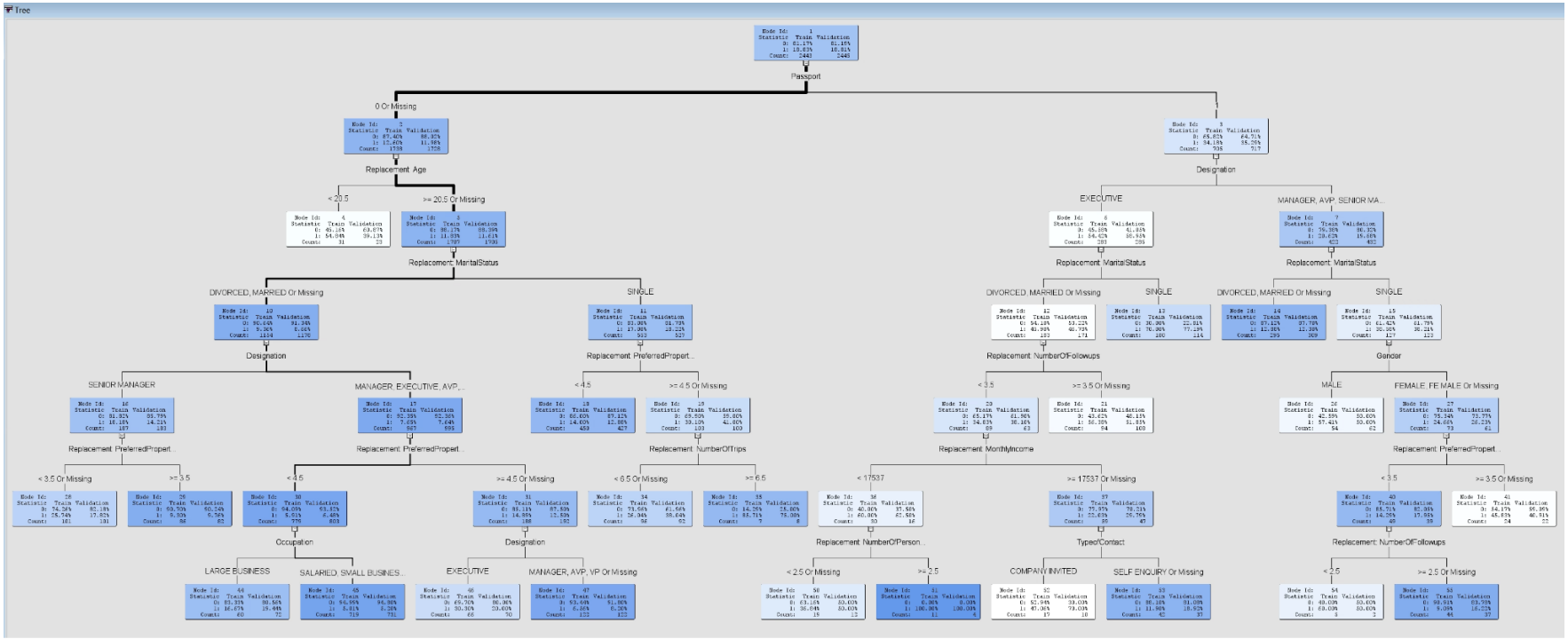
Use Frozen Tree	Yes
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf 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	Lift
Assessment Fraction	0.25
Cross Validation	











- The optimal number of leaves is 21.
- The variable used for the first split was Passport. The competing splits used in the second split were Age and Designation. Another important variable used in the third split was Marital Status. As can be seen in the tree, variables like Monthly Income, Gender, Preferred Property Score and Number of Follow-ups were included but less important.
- Valid average square error (ASE) = 0.118446
- Valid misclassification rate = 0.159918

The 3-way Tree is the best decision tree, based on its lowest valid ASE and valid misclassification rate.

Logistic Regression

Data Massaging

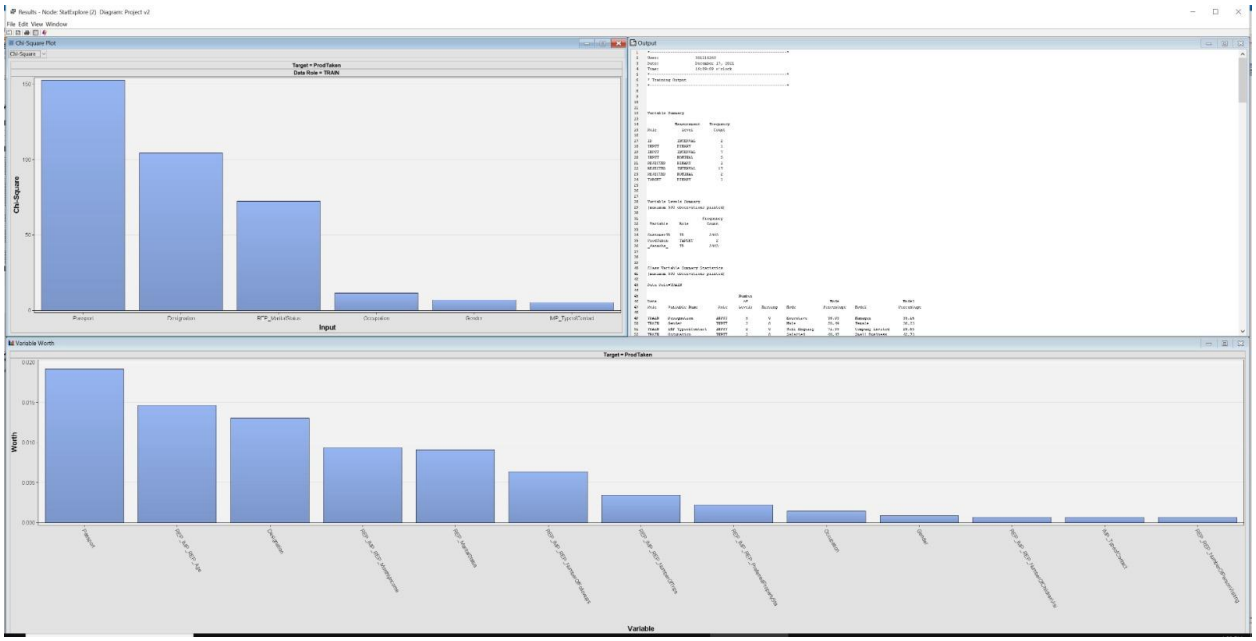
Train	
Variables	
Nonmissing Variabl	No
Missing Cutoff	50.0
Class Variables	
Default Input Metho	Count
Default Target Meth	None
Normalize Values	Yes
Interval Variables	
Default Input Metho	Mean
Default Target Meth	None
Default Constant Va	
Default Character Va	U
Default Number Var	
Method Options	
Random Seed	12345
Tuning Parameters	
Tree Imputation	
Score	
Hide Original Variat	Yes
Indicator Variables	
Type	None
Source	Imputed Variables
Role	Rejected

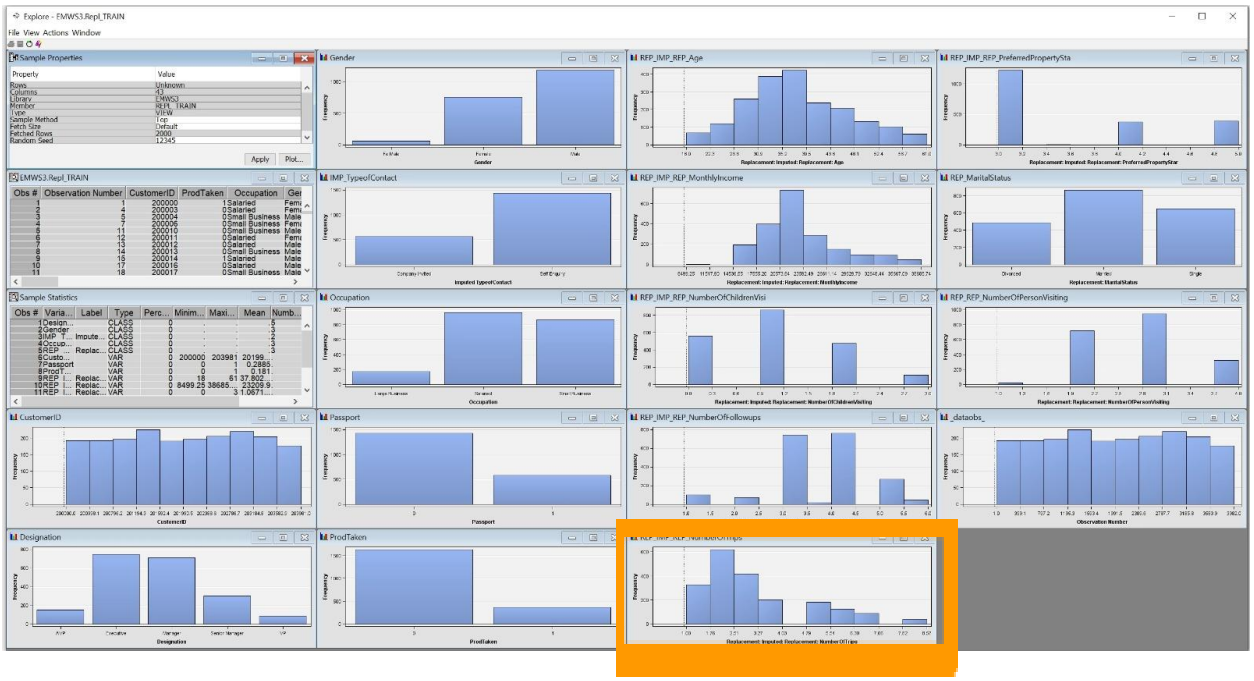
To prepare for linear regression, we used an Impute node (left) for the missing data. We then used a Replacement node (right) with default limits method set to a value of standard deviation from the mean. This was done to cap and floor outliers, reducing the variables that we need to transform with logarithms later. Less use of the logarithm would mean easier reporting of the data to audiences who are unfamiliar with data manipulation methods.

Train	
Interval Variables	
Replacement Editor	
Default Limits Method	Standard Deviations from th
Cutoff Values	
Class Variables	
Replacement Editor	
Unknown Levels	Ignore
Score	
Replacement Values	Computed
Hide	No
Report	
Replacement Report	Yes

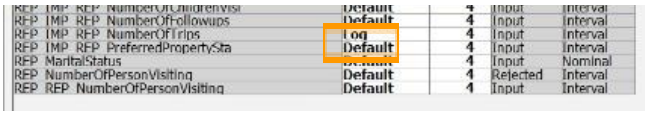
Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
IMP_REP_Age	37.797762478	INPUT	INTERVAL	Replacement: Age	119
IMP_REP_MonthlyIncome	23592.493172	INPUT	INTERVAL	Replacement: MonthlyIncome	130
IMP_REP_NumberOfChildrenVisiting	1.1806879403	INPUT	INTERVAL	Replacement: NumberOfChildrenVisiting	30
IMP_REP_NumberOfFollowups	3.7058580858	INPUT	INTERVAL	Replacement: NumberOfFollowups	19
IMP_REP_NumberOfTrips	3.2263083273	INPUT	INTERVAL	Replacement: NumberOfTrips	73
IMP_REP_PreferredPropertyStar	3.5736019737	INPUT	INTERVAL	Replacement: PreferredPropertyStar	11
IMP_TypeofContact	Self Enquiry	INPUT	NOMINAL		12

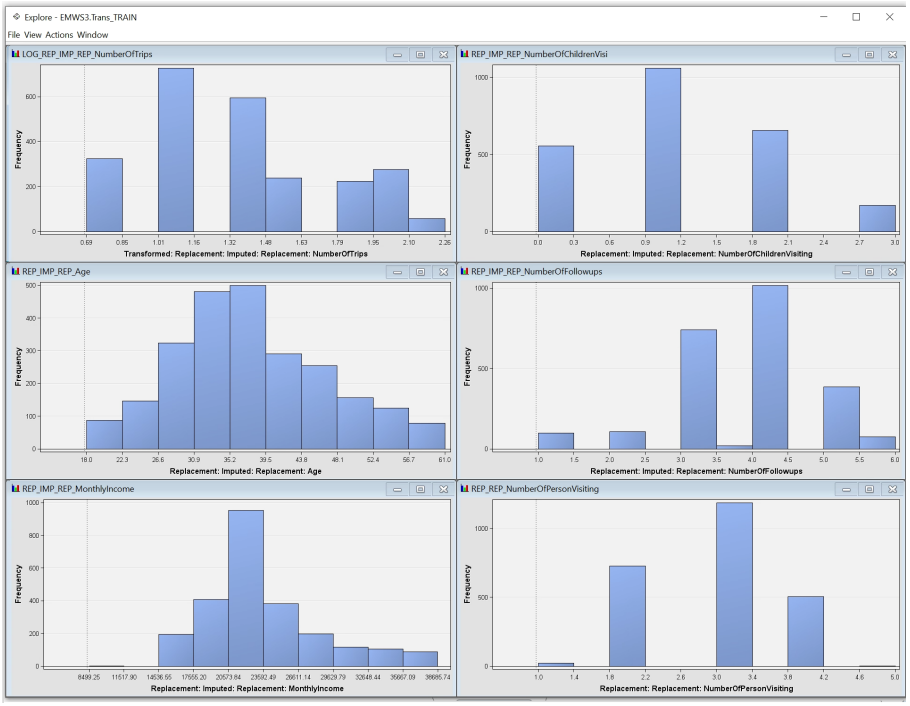
The data inputs (the right side of the data explored below) are less skewed than before.





However, we noticed the Number of Trips (at the lower right of the variables explored in the image above) still had a skewed distribution. So we decided to apply a Transform Node, changing the method of Number of Trips to use a logarithm.





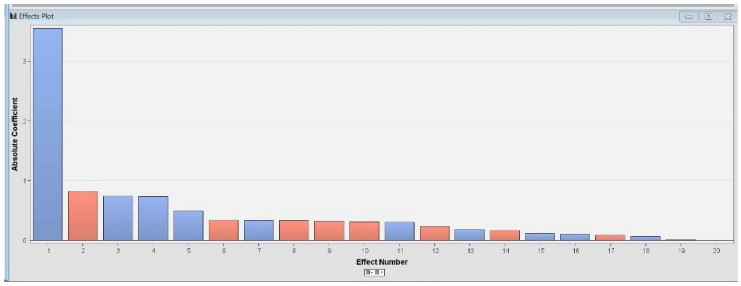
The resulting distribution for Number of Trips was less skewed (see screenshot on the left).

We considered converting some variables into dummy variables, but chose not to. We observed the intervals for our variables have only one value for each interval, so we assumed that change would not influence the results.

Comparison of Regressions

We chose to do full, backward, forward and stepwise logical regressions. Our variables did not seem to interact with each other, so we chose not to do polynomial regression.

It is interesting to note our effects plots for our regressions. In order of importance, Designation (Executive), Designation (AVP) and Passport had an appreciable effect on the odds of customers buying our travel package.



Full Regression

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
2363.546	1961.374	402.1724	19	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Designation	4	44.4639	<.0001
Gender	2	11.6772	0.0029
IMP_TypeofContact	1	7.3023	0.0069
LOG_REP_IMP_REP_NumberOfTrips	1	4.8442	0.0277
Occupation	2	4.7016	0.0953
Passport	1	155.8074	<.0001
REP_IMP_REP_Age	1	6.4904	0.0108
REP_IMP_REP_MonthlyIncome	1	0.0000	0.9997
REP_IMP_REP_NumberOfChildrenVisi	1	1.5930	0.2069
REP_IMP_REP_NumberOfFollowups	1	23.3390	<.0001
REP_IMP_REP_PreferredPropertySta	1	21.0444	<.0001
REP_MaritalStatus	2	63.1689	<.0001
REP_REP_NumberOfPersonVisiting	1	0.4535	0.5007

Odds Ratio Estimates

Effect		Point Estimate
Designation	AVP vs VP	0.648
Designation	Executive vs VP	3.084
Designation	Manager vs VP	1.212
Designation	Senior Manager vs VP	1.911
Gender	Fe Male vs Male	0.408
Gender	Female vs Male	0.733
IMP_TypeofContact	Company Invited vs Self Enquiry	1.399
LOG_REP_IMP_REP_NumberOfTrips		1.392
Occupation	Large Business vs Small Business	1.338
Occupation	Salaried vs Small Business	0.884
Passport	0 vs 1	0.228
REP_IMP_REP_Age		0.981
REP_IMP_REP_MonthlyIncome		1.000
REP_IMP_REP_NumberOfChildrenVisi		0.897
REP_IMP_REP_NumberOfFollowups		1.365
REP_IMP_REP_PreferredPropertySta		1.381
REP_MaritalStatus	Divorced vs Single	0.375
REP_MaritalStatus	Married vs Single	0.385
REP_REP_NumberOfPersonVisiting		0.932

- Valid ASE = 0.117825
- Valid misclassification rate = 0.155828

Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	2001.374	.
ASE	Average Squared Error	0.123312	0.117825
AVERR	Average Error Function	0.401427	0.384287
DFE	Degrees of Freedom for Error	2423	.
DFM	Model Degrees of Freedom	20	.
DFT	Total Degrees of Freedom	2443	.
DIV	Divisor for ASE	4886	4890
ERR	Error Function	1961.374	1879.165
FPE	Final Prediction Error	0.125347	.
MAX	Maximum Absolute Error	0.97575	0.969381
MSE	Mean Square Error	0.12433	0.117825
NOBS	Sum of Frequencies	2443	2445
NW	Number of Estimate Weights	20	.
RASE	Root Average Sum of Squares	0.351158	0.343257
RFPE	Root Final Prediction Error	0.354044	.
RMSE	Root Mean Squared Error	0.352604	0.343257
SBC	Schwarz's Bayesian Criterion	2117.393	.
SSE	Sum of Squared Errors	602.5008	576.1646
SUMW	Sum of Case Weights Times Freq	4886	4890
MISC	Misclassification Rate	0.167417	0.155828

With our base group (VP), our odds ratios show that:

- With every added customer who is an AVP, VPs are 35.2% less likely to buy a package.
- With every added customer who is an Executive, Execs are 208.4% more likely to buy a package.
- With every added customer who is a Manager, the chances of Managers buying a package is 21.2%.
- With every added customer who is a Senior Manager, Sr. Managers are 91.1% more likely to buy a package.
- With every additional company-invited type of outreach, company-invited customers are 39.9% more likely to buy.

Backward Regression

Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	2001.374	.
ASE	Average Squared Error	0.123312	0.117825
AVERR	Average Error Function	0.401427	0.384287
DFE	Degrees of Freedom for Error	2423	.
DFM	Model Degrees of Freedom	20	.
DFT	Total Degrees of Freedom	2443	.
DIV	Divisor for ASE	4886	4890
ERR	Error Function	1961.374	1879.165
FPE	Final Prediction Error	0.125347	.
MAX	Maximum Absolute Error	0.97575	0.969381
MSE	Mean Square Error	0.12433	0.117825
NOBS	Sum of Frequencies	2443	2445
NW	Number of Estimate Weights	20	.
RASE	Root Average Sum of Squares	0.351158	0.343257
RFPE	Root Final Prediction Error	0.354044	.
RMSE	Root Mean Squared Error	0.352604	0.343257
SBC	Schwarz's Bayesian Criterion	2117.393	.
SSE	Sum of Squared Errors	602.5008	576.1646
SUMW	Sum of Case Weights Times F...	4886	4890
MISC	Misclassification Rate	0.167417	0.155828

Odds Ratio Estimates

Effect		Point Estimate
Designation	AVP vs VP	0.648
Designation	Executive vs VP	3.084
Designation	Manager vs VP	1.212
Designation	Senior Manager vs VP	1.911
Gender	Fe Male vs Male	0.408
Gender	Female vs Male	0.733
IMP_TypeofContact	Company Invited vs Self Enquiry	1.399
LOG_REP_IMP_REP_NumberOfTrips		1.392
Occupation	Large Business vs Small Business	1.338
Occupation	Salaried vs Small Business	0.884
Passport	0 vs 1	0.228
REP_IMP_REP_Age		0.981
REP_IMP_REP_MonthlyIncome		1.000
REP_IMP_REP_NumberOfChildrenVisi		0.897
REP_IMP_REP_NumberOfFollowups		1.365
REP_IMP_REP_PreferredPropertySta		1.381
REP_MaritalStatus	Divorced vs Single	0.375
REP_MaritalStatus	Married vs Single	0.385
REP_REP_NumberOfPersonVisiting		0.932

The selected model, based on the error rate for the validation data, is the model trained in Step 0. It consists of the following effects:

Intercept Designation Gender IMP_TypeofContact LOG_REP_IMP_REP_NumberOfTrips Occupation Passport REP_IMP_REP_Age REP_IMP_REP_MonthlyIncome REP_IMP_REP_NumberOfChildrenVisi REP_IMP_REP_NumberOfFollowups REP_IMP_REP_PreferredPropertySta REP_MaritalStatus REP_REP_NumberOfPersonVisiting

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
2363.546	1961.374	402.1724	19	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Designation	4	44.4639	<.0001
Gender	2	11.6772	0.0029
IMP_TypeofContact	1	7.3023	0.0069
LOG_REP_IMP_REP_NumberOfTrips	1	4.8442	0.0277
Occupation	2	4.7016	0.0953
Passport	1	155.8074	<.0001
REP_IMP_REP_Age	1	6.4904	0.0108
REP_IMP_REP_MonthlyIncome	1	0.0000	0.9997
REP_IMP_REP_NumberOfChildrenVisi	1	1.5930	0.2069
REP_IMP_REP_NumberOfFollowups	1	23.3390	<.0001
REP_IMP_REP_PreferredPropertySta	1	21.0444	<.0001
REP_MaritalStatus	2	63.1689	<.0001
REP_REP_NumberOfPersonVisiting	1	0.4535	0.5007

- Valid ASE = 0.117825
- Valid misclassification rate = 0.155828
- The model is trained in Step 0.
- The imputed data shows the highest odds of buying our travel package is based on Type of Contact, then Number of Trips, then Preferred Property Star.

With a base group of VP, our odds ratios show that:

- With every added customer who is an AVP, VPs are 35.2% less likely to buy a package.
- With every added customer who is an Executive, Execs are 208.4% more likely to buy a package.
- With every added customer who is a Manager, the chances of Managers buying a package is 21.2%.
- With every added customer who is a Senior Manager, Sr. Managers are 91.1% more likely to buy a package.
- With every additional company-invited type of outreach, company-invited customers are 39.9% more likely to buy.

Forward Regression

- Valid ASE = 0.11765
- Valid misclassification rate = 0.156646
- The model is trained in Step 8.
- The imputed data shows the highest odds of buying our travel package is based on Preferred Property Star, Type of Contact and Number of Follow-ups.

With a base group of VP, our odds ratios show that:

- With every added customer who is an AVP, VPs are 32.1% less likely to buy a package.
- With every added customer who is an Executive, Execs are 226.3% more likely to buy a package.
- With every added customer who is a Manager, the chances of Managers buying a package is 28.5%.
- With every added customer who is a Senior Manager, Sr. Managers are 101.8% more likely to buy a package.
- With every additional company-invited type of outreach, company-invited customers are 37.7% more likely to buy.

Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	2002.119	
ASE	Average Squared Error	0.124151	0.11765
AVERR	Average Error Function	0.404036	0.384729
DFE	Degrees of Freedom for Error	2429	
DFM	Model Degrees of Freedom	14	
DFT	Total Degrees of Freedom	2443	
DIV	Divisor for ASE	4886	4890
ERR	Error Function	1974.119	1881.327
FPE	Final Prediction Error	0.125562	
MAX	Maximum Absolute Error	0.979086	0.979811
MSE	Mean Square Error	0.124867	0.11765
NOBS	Sum of Frequencies	2443	2445
NW	Number of Estimate Weights	14	
RASE	Root Average Sum of Squares	0.352351	0.343002
RFPE	Root Final Prediction Error	0.354376	
RMSE	Root Mean Squared Error	0.353365	0.343002
SBC	Schwarz's Bayesian Criterion	2083.333	
SSE	Sum of Squared Errors	606.6029	575.3089
SUMW	Sum of Case Weights Times Freq	4886	4890
MISC	Misclassification Rate	0.166598	0.156646

The selected model, based on the error rate for the validation data, is the model trained in Step 8. It consists of the following effects:

Intercept Designation Gender IMP_TypeofContact Passport REP_IMP_REP_Age REP_IMP_REP_NumberOfFollowups REP_IMP_REP_PreferredPropertySta REP_MaritalStatus

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
2363.546	1974.119	389.4271	13	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Designation	4	59.7338	<.0001
Gender	2	11.3188	0.0035
IMP_TypeofContact	1	6.7137	0.0096
Passport	1	156.0028	<.0001
REP_IMP_REP_Age	1	5.4476	0.0196
REP_IMP_REP_NumberOfFollowups	1	25.2930	<.0001
REP_IMP_REP_PreferredPropertySta	1	21.7566	<.0001
REP_MaritalStatus	2	64.8140	<.0001

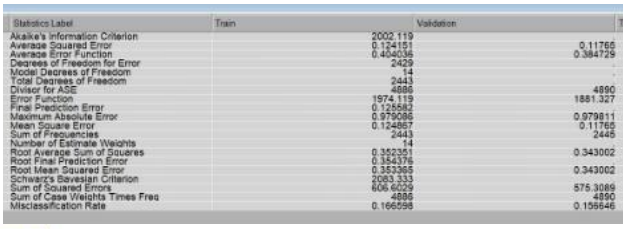
Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.5352	0.4795	54.35	<.0001
Designation	AVP	1	-0.7370	0.2637	7.81	0.0052
Designation	Executive	1	0.8329	0.1330	39.24	<.0001
Designation	Manager	1	-0.0990	0.1356	0.53	0.4654
Designation	Senior Manager	1	0.3527	0.1552	5.16	0.0230
Gender	Fe Male	1	-0.4611	0.2286	4.07	0.0437
Gender	Female	1	0.0737	0.1322	0.31	0.5772
IMP_TypeofContact	Company Invited	1	0.1599	0.0617	6.71	0.0096
Passport	0	1	-0.7353	0.0589	156.00	<.0001
REP_IMP_REP_Age		1	-0.0166	0.00711	5.45	0.0196
REP_IMP_REP_NumberOfFollowups		1	0.2908	0.0578	25.29	<.0001
REP_IMP_REP_PreferredPropertySta		1	0.3246	0.0696	21.76	<.0001
REP_MaritalStatus	Divorced	1	-0.3418	0.1037	10.86	0.0010
REP_MaritalStatus	Married	1	-0.3058	0.0826	13.70	0.0002

Odds Ratio Estimates

Effect		Point Estimate
Designation	AVP vs VP	0.679
Designation	Executive vs VP	3.263
Designation	Manager vs VP	1.285
Designation	Senior Manager vs VP	2.018
Gender	Fe Male vs Male	0.428
Gender	Female vs Male	0.731
IMP_TypeofContact	Company Invited vs Self Enquiry	1.377
Passport	0 vs 1	0.230
REP_IMP_REP_Age		0.984
REP_IMP_REP_NumberOfFollowups		1.337
REP_IMP_REP_PreferredPropertySta		1.383
REP_MaritalStatus	Divorced vs Single	0.372
REP_MaritalStatus	Married vs Single	0.385

Stepwise Regression



Statistics Label	Train	Validation
Akaike's Information Criterion	2002.119	
Average Squared Error	0.124151	0.11765
Average Error Function	0.404036	0.384729
Degrees of Freedom for Error	2429	
Model Degrees of Freedom	14	
Total Degrees of Freedom	2443	
Divisor for ASE	4886	4890
Error Function	1974.119	1881.327
Final Prediction Error	0.125582	
Maximum Absolute Error	0.979086	0.979811
Mean Square Error	0.124867	0.11765
Sum of Frequencies	2443	2445
Number of Estimate Weights	14	
Root Average Sum of Squares	0.352351	0.343002
Root Final Prediction Error	0.354376	
Root Mean Squared Error	0.353365	0.343002
Schwarz's Bayesian Criterion	2083.333	
Sum of Squared Errors	606.6029	575.3089
Sum of Case Weights Times Freq	4886	4890
Misclassification Rate	0.166598	0.156646

- Valid ASE = 0.11765
- Valid misclassification rate = 0.156646
- The model is trained in Step 8.
- The imputed data shows the highest odds of buying our travel package is based on Preferred Property Star, Type of Contact and Number of Follow-ups.

The selected model, based on the error rate for the validation data, is the model trained in Step 8. It consists of the following effects:

Intercept Designation Gender IMP_TypeofContact Passport REP_IMP_REP_Age REP_IMP_REP_NumberOfFollowups REP_IMP_REP_PreferredPropertySta REP_MaritalStatus

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
2363.546	1974.119	389.4271	13	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Designation	4	59.7338	<.0001
Gender	2	11.3188	0.0035
IMP_TypeofContact	1	6.7137	0.0096
Passport	1	156.0028	<.0001
REP_IMP_REP_Age	1	5.4476	0.0196
REP_IMP_REP_NumberOfFollowups	1	25.2930	<.0001
REP_IMP_REP_PreferredPropertySta	1	21.7566	<.0001
REP_MaritalStatus	2	64.8140	<.0001

Odds Ratio Estimates

Effect		Point Estimate
Designation	AVP vs VP	0.679
Designation	Executive vs VP	3.263
Designation	Manager vs VP	1.285
Designation	Senior Manager vs VP	2.018
Gender	Fe Male vs Male	0.428
Gender	Female vs Male	0.731
IMP_TypeofContact	Company Invited vs Self Enquiry	1.377
Passport	0 vs 1	0.230
REP_IMP_REP_Age		0.984
REP_IMP_REP_NumberOfFollowups		1.337
REP_IMP_REP_PreferredPropertySta		1.383
REP_MaritalStatus	Divorced vs Single	0.372
REP_MaritalStatus	Married vs Single	0.385

With a base group of VP, our odds ratios show that:

- With every added customer who is an AVP, VPs are 32.1% less likely to buy a package.
- With every added customer who is an Executive, Execs are 226.3% more likely to buy a package.
- With every added customer who is a Manager, the chances of Managers buying a package is 28.5%.
- With every added customer who is a Senior Manager, Sr. Managers are 101.8% more likely to buy a package.
- With every additional company-invited type of outreach, company-invited customers are 37.7% more likely to buy.

Based on valid ASEs and misclassification rates, we selected stepwise regression as our best model.

Neural Network

Data Massaging

Like our regression models, the neural network nodes are connected to the Impute and Replacement (Cap and Floor) nodes already discussed in the Logistic Regression section. Missing values and skewness have already been considered in that portion of our work.

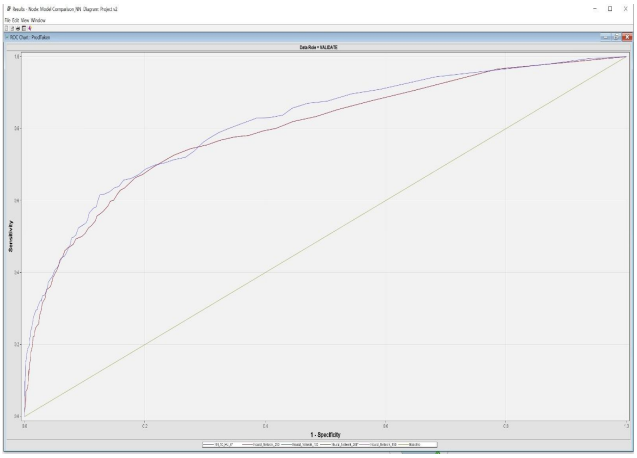
Neural network model selection criterion was set to Average Error so that the node selects the model with the least average error for validation of data.

Hidden Units

In order to discover the optimal number of hidden units, we then connected different neural networks to our prepared data. We ran models with:

- iterations = 50 and
- hidden units = 0, 2, 3, 4, 5, 6, 7, 8, 9 and 10

We found that the neural network with **6 hidden units** was the best model as it had the lowest valid average square error (0.113577), the highest ROC index (0.814, shared with the model with 8 hidden units) and the highest Gini coefficient (0.627). It also had one of the lower misclassification rates (0.15501, the fourth lowest out of 10) and the the fourth highest Kolmogorov-Smirnov statistic (0.492).



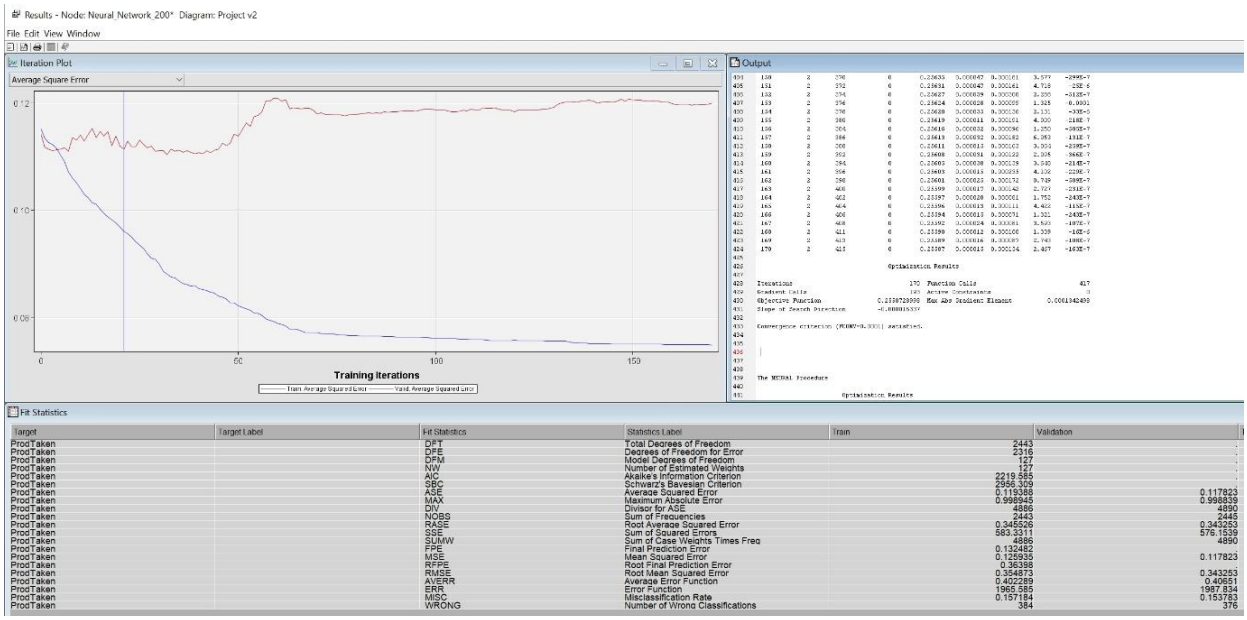
Model Description	Valid: Average Squared Error ▲	Valid: Misclassification Rate	Valid: Roc Index	Target Label	Valid: Gini Coefficient	Valid: Kolmogorov-Smirnov Statistic
NN 50 HU 6	0.113577	0.15501	0.814		0.627	0.492
NN 50 HU 7	0.114713	0.150511	0.812		0.624	0.468
NN 50 HU 8*	0.11554	0.155419	0.814		0.627	0.497
NN 50 HU No	0.116519	0.160736	0.8		0.6	0.493
NN 50 HU 3	0.116519	0.160736	0.8		0.6	0.493
NN 50 HU 9	0.117259	0.153374	0.803		0.606	0.46
NN 50 HU 10	0.118276	0.161963	0.815		0.63	0.492
NN 50 HU 5	0.120715	0.160327	0.791		0.581	0.444
NN 50 HU 2	0.126294	0.170961	0.774		0.549	0.397
NN 50 HU 4	0.147789	0.208589	0.762		0.524	0.38

Iterations

We then ran neural networks with 6 hidden units and various iterations to look at our iteration plots and check for convergence. We used:

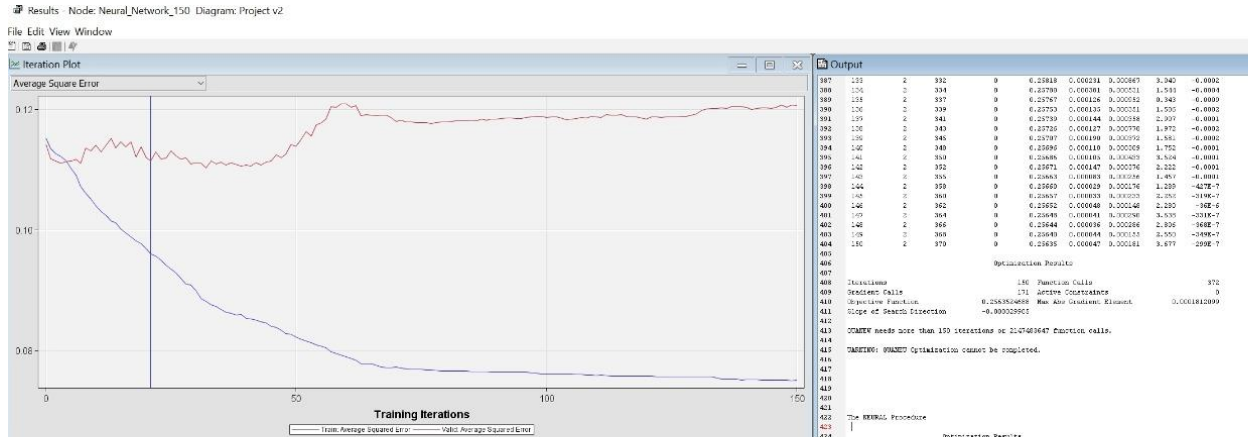
- hidden units = 6 and
- iterations = 50, 100, 150, 200 and 250

Our neural network converged at 200 iterations. However, the iteration plot was not ideal. (Please see the output for neural network node with 200 iterations on the next page.)



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
ProdTaken		DFT	Total Degrees of Freedom	2443	
ProdTaken		DFE	Degrees of Freedom for Error	2316	
ProdTaken		DFM	Model Degrees of Freedom	127	
ProdTaken		NW	Number of Estimated Weights	127	
ProdTaken		AIC	Akaike's Information Criterion	2219.585	
ProdTaken		SBC	Schwarz's Bayesian Criterion	2966.309	
ProdTaken		ASE	Average Squared Error	0.119388	0.117823
ProdTaken		MAX	Maximum Absolute Error	0.998945	0.998839
ProdTaken		DIV	Divisor for ASE	4886	4890
ProdTaken		NOBS	Sum of Frequencies	2443	2445
ProdTaken		RASE	Root Average Squared Error	0.345526	0.343253
ProdTaken		SSE	Sum of Squared Errors	583.3311	576.1539
ProdTaken		SUMW	Sum of Case Weights Times Freq	4885	4890
ProdTaken		FPE	Final Prediction Error	0.132482	
ProdTaken		MSE	Mean Squared Error	0.125935	0.117823
ProdTaken		RFPE	Root Final Prediction Error	0.36398	
ProdTaken		RMSE	Root Mean Squared Error	0.354873	0.343253
ProdTaken		AVERR	Average Error Function	0.402289	0.40651
ProdTaken		ERR	Error Function	1965.585	1987.834
ProdTaken		MISC	Misclassification Rate	0.157184	0.153783
ProdTaken		WRONG	Number of Wrong Classifications	384	376

The iteration plot of 150 iterations looked similar, though the model did not converge yet.

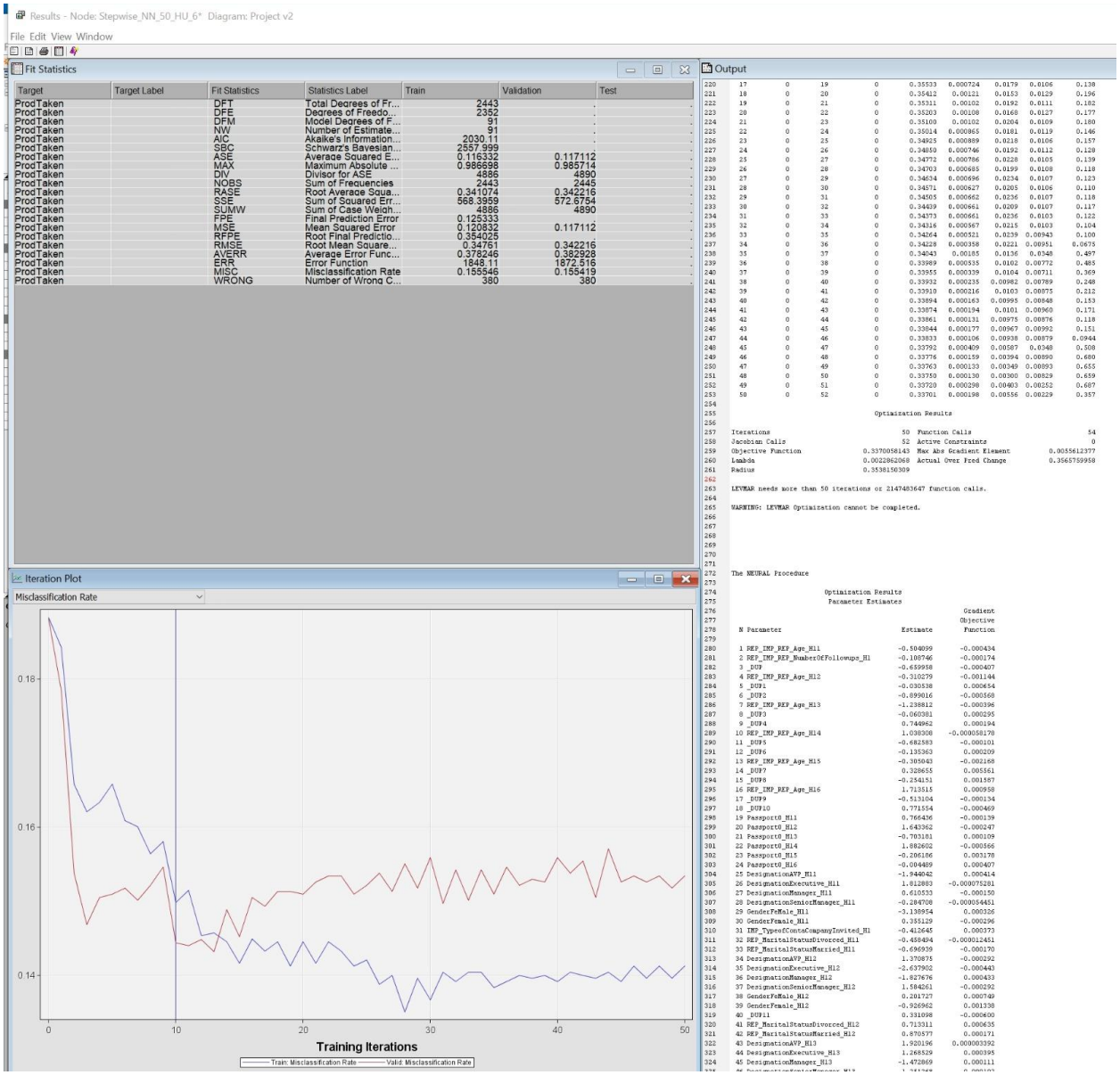


Our iteration plot for 50 iterations looked better and had better fit statistics (seen below), but again the model did not converge. We also noticed that the fit statistics did not change with the continuing iterations. This implied the need to reduce inputs to get a better model.

Model Description	Valid: Average Squared Error	Valid: Misclassification Rate	Valid: Roc Index	Valid: Gini Coefficient	Target Variable	Valid: Kolmogorov-Smirnov Statistic
Neural Network 150	0.117823	0.153783	0.797	0.595	ProdT...	0.479
Neural Network 200*	0.117823	0.153783	0.797	0.595	ProdT...	0.479
Neural Network 100	0.117823	0.153783	0.797	0.595	ProdT...	0.479
Neural Network 250	0.117823	0.153783	0.797	0.595	ProdT...	0.479
NN 50 HU 6*	0.113577	0.15501	0.814	0.627	ProdT...	0.492

Input Reduction Neural Network Node

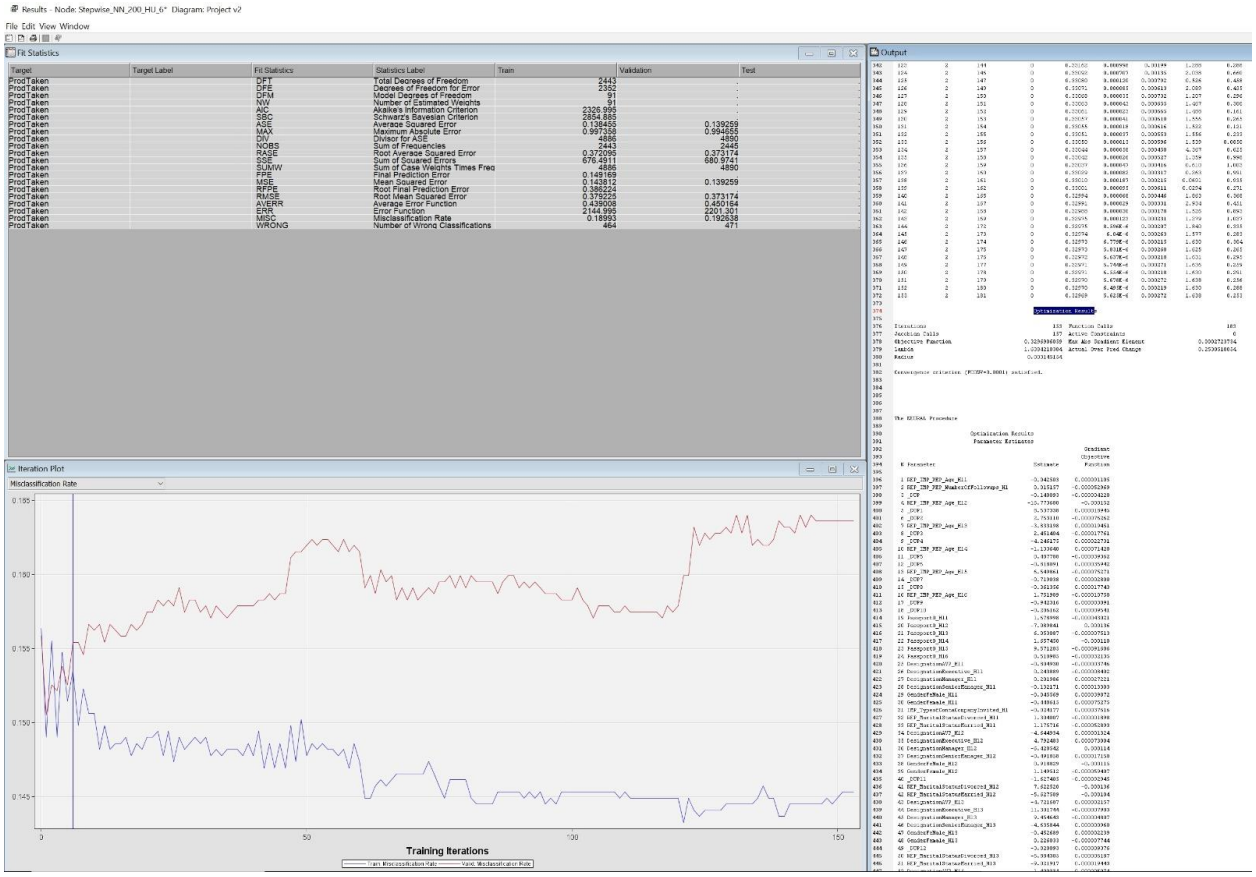
We then attached input reduction neural network nodes to our best regression model, the stepwise model. We ran our neural networks with 50 and 200 iterations.



The iteration plot had improved for the model with 50 iterations. However the valid ASE and valid misclassification rate had actually gotten higher (0.117112 and 0.155419 respectively as compared to the original neural network model with 50 iterations: 0.113577 and 0.15501 respectively).

A similar pattern was seen in the neural network model with 200 iterations with reduced inputs. The iteration plot improved (image on the next page), but the valid ASE and

misclassification rate got higher (0.139259 and 0.192638 respectively as compared to the original neural network model with 50 iterations: 0.117823 and 0.153783 respectively).



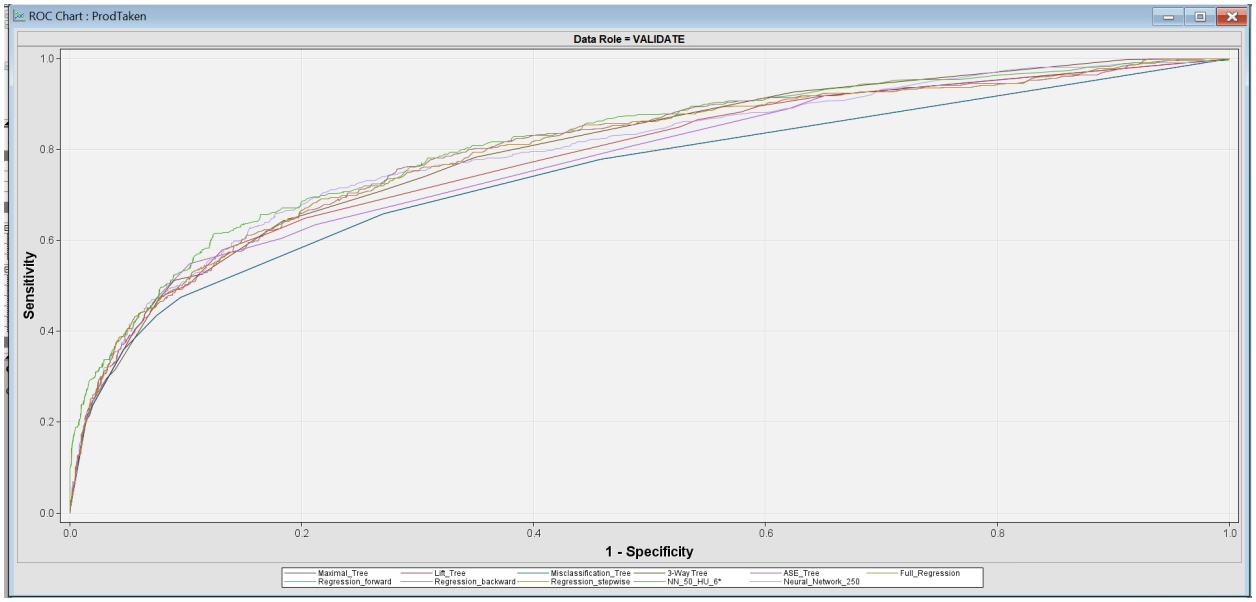
After comparing all the neural network models, we chose to discount input reduction altogether in the selection of the best model. With consideration for the valid ASEs, misclassification rates, Gini coefficients, ROC indexes and Kolmogorov-Smirnov statistics, **we chose the neural network with 50 iterations and 6 hidden units as the best model**.

Model Comparison

Model Node	Model Description	Valid: Misclassification Rate	Valid: Average Squared Error ▲	Valid: Gini Coefficient	Valid: Kolmogorov-Smirnov Statistic	Valid: Roc Index
Neural7	NN_50_HU_6*	0.15501	0.113577	0.627	0.492	0.814
Tree2	3-Way Tree	0.160736	0.116995	0.599	0.458	0.8
Neural16	Stepwise_NN_50_HU_6*	0.155419	0.117112	0.592	0.45	0.796
Reg	Regression_stepwise	0.156646	0.11765	0.597	0.475	0.799
Reg3	Regression_forward	0.156646	0.11765	0.597	0.475	0.799
Neural12	Neural_Network_150	0.153783	0.117823	0.595	0.487	0.797
Neural13	Neural_Network_200*	0.153783	0.117823	0.595	0.487	0.797
Neural14	Neural_Network_100	0.153783	0.117823	0.595	0.487	0.797
Neural2	Neural_Network_250	0.153783	0.117823	0.595	0.487	0.797
Reg4	Full_Regression	0.155828	0.117825	0.599	0.474	0.799
Reg2	Regression_backward	0.155828	0.117825	0.599	0.474	0.799
Tree	ASE_Tree	0.159918	0.118406	0.55	0.445	0.775
Tree4	Lift_Tree	0.159918	0.118446	0.567	0.447	0.783
Tree3	Misclassification_Tree	0.158282	0.123492	0.494	0.388	0.747
Tree5	Maximal_Tree	0.158282	0.123492	0.494	0.388	0.747
Neural15	Stepwise_NN_200_HU_6*	0.192638	0.139259	0.522	0.386	0.761

Ultimately, we chose to compare our models using:

- the lowest valid average squared error
- the highest valid Gini coefficient
- the lowest valid misclassification rate
- the highest valid Kolmogorov-Smirnov statistic, and
- the highest valid ROC index.



On the basis of these statistics, the neural network node with 50 iterations and 6 hidden units is our best model. It had the highest ROC index by far (0.814), the lowest valid average squared error (0.113577), and the highest Gini coefficient (0.627). While not

quite the lowest misclassification rate among all our models, it was not as high as some of our other models' rates.

The next best model based on these statistics was the 3-way decision tree. Its valid average squared error (0.116995), valid ROC index (0.8) and valid Gini coefficient (0.599) are on the higher end of the spectrum when comparing all the models' statistics.

Ultimately, **when considering the value of a model in terms of the business problem at hand, we chose the 3-way decision tree as the better model**. The neural network is only slightly better than this decision tree model purely in terms of the statistics, but is harder to interpret for business planning. On top of that, the 3-way decision tree has more actionable insights for the company's decision-making.

The company will be able to implement audience targeting in their marketing with the 3-way decision tree's results.

Conclusion

In conclusion, our neural network and 3-way decision tree models are the most reliable in predicting the customers who will buy the holiday travel package.

Purely from comparison of the statistics, the neural network is slightly better than the decision tree. However, when considering actionable insights for future marketing initiatives, the 3-way decision tree is more useful for the company.

Passport, job designation and age were critical variables affecting our decision tree models. While less important than these three, marital status and gender also contributed to further decision tree splits.

Based on these findings, these customers are most likely to buy the package:

- Single VPs, AVPs without a passport who are older than 20.5 years of age (or missing their age)
- Single executives without a passport who are older than 20.5 years of age (or missing their age) and have a monthly income equal to or more than $24,266.50 USD
- Divorced, married (or missing marital status) executives without a passport who have a monthly income equal to or more than $24,233 USD

As mentioned in the earlier sections, we recommend that it would be useful if the company could provide information on whether these trips are domestic or international. With Passport being such an important variable, we believe that would highly increase the accuracy of the model as it will identify whether the customer should have a passport in order to enjoy the trip package.

References

Susant_Achary. “Holiday_Package_Prediction.” *Kaggle*, August 2021. https://www.kaggle.com/susant4learning/holiday-package-purchase-prediction