Data: Holiday Package Prediction

BA 706 REVISION

BY

Risikat Hameed

Table of Contents

Table of Contents 1

Description of the Problem **2**

File Import **2**

Data Wrangling **2**

- Among 14 variables left 6 are Categorical (Nominal) variables: 3
- Data dictionary 4

Diagram **5**

Decision Tree **6**

- Comparison of Decision Tree Models 6
  - Probability Tree 6
  - 3-Way Tree 8
  - Lift Tree 10
  - Misclassification Tree 12
  - Maximal Tree 14

Logistic Regression **16**

- Data Massaging 16
- Comparison of Regressions 17
  - Backward Regression 17
  - Forward Regression 19
  - Stepwise Regression 20

Neural Network **21**

- Data Massaging 21
- Hidden Units 21
- Iterations 21
- Input Reduction Neural Network Node 23

Model Comparison **25**

Conclusion **26**

References

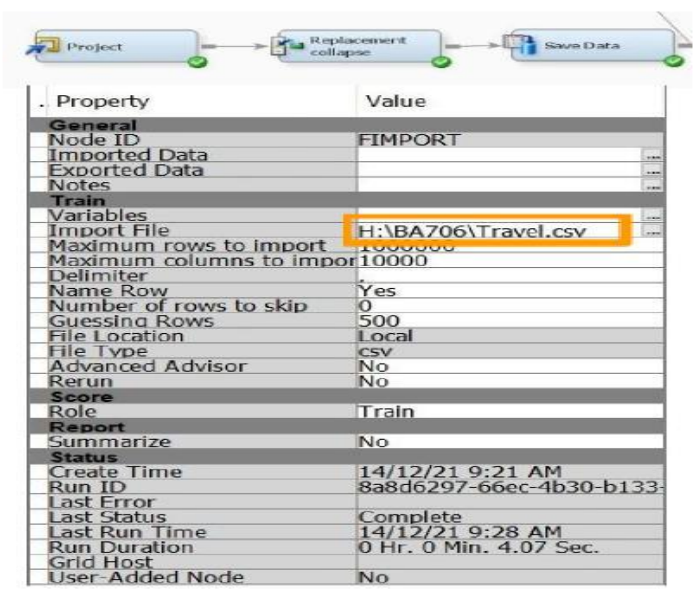
Description of the Problem

Introducing a new package offering is one way to broaden the customer base. The company currently offers five different types of packages: Basic, Standard, Deluxe, Super Deluxe, and King. Looking at data from the previous year, we discovered that 18% of customers purchased the packages. However, the marketing cost was quite high because customers were contacted at random without regard for the information available. The company is about to launch a new product called the Wellness Tourism Package. Wellness tourism is defined as travel that allows the traveler to maintain, improve, or begin a healthy lifestyle, as well as support or increase one's sense of well-being. However, this time the company wishes to leverage the available data of existing and potential customers in order to maximize marketing expenditure.

File Import

We imported the CSV file (under Train in the Property panel on the right) and saved it as an SAS dataset using the Export node ('Save Data'). This enables us to quickly connect the Export Node. The original (raw) dataset provided contains 19 input variables. We found no evidence of potential data leakage or data duplication.

The target variable for Holiday Package Prediction dataset is “ProdTaken”. It is a binary variable, which identifies whether the customer has purchased the product: 0 - No; 1 - Yes.



Property	Value
General	
Node ID	FIMPORT
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Import File	H:\BA706\Travel.csv
Maximum rows to import	1000000
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

Data Wrangling

After analyzing the data, I discarded 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.

4. PitchSatisfactionScore - Customer's satisfaction with marketing pitch. Ranges 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.

The other variables 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 Female)
3. Marital Status:

a. Divorced

b. Married

c. Single

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

Since the Unmarried and Single categories are the same, I used a spare node to reduce the Marital Status variable (see below for settings).

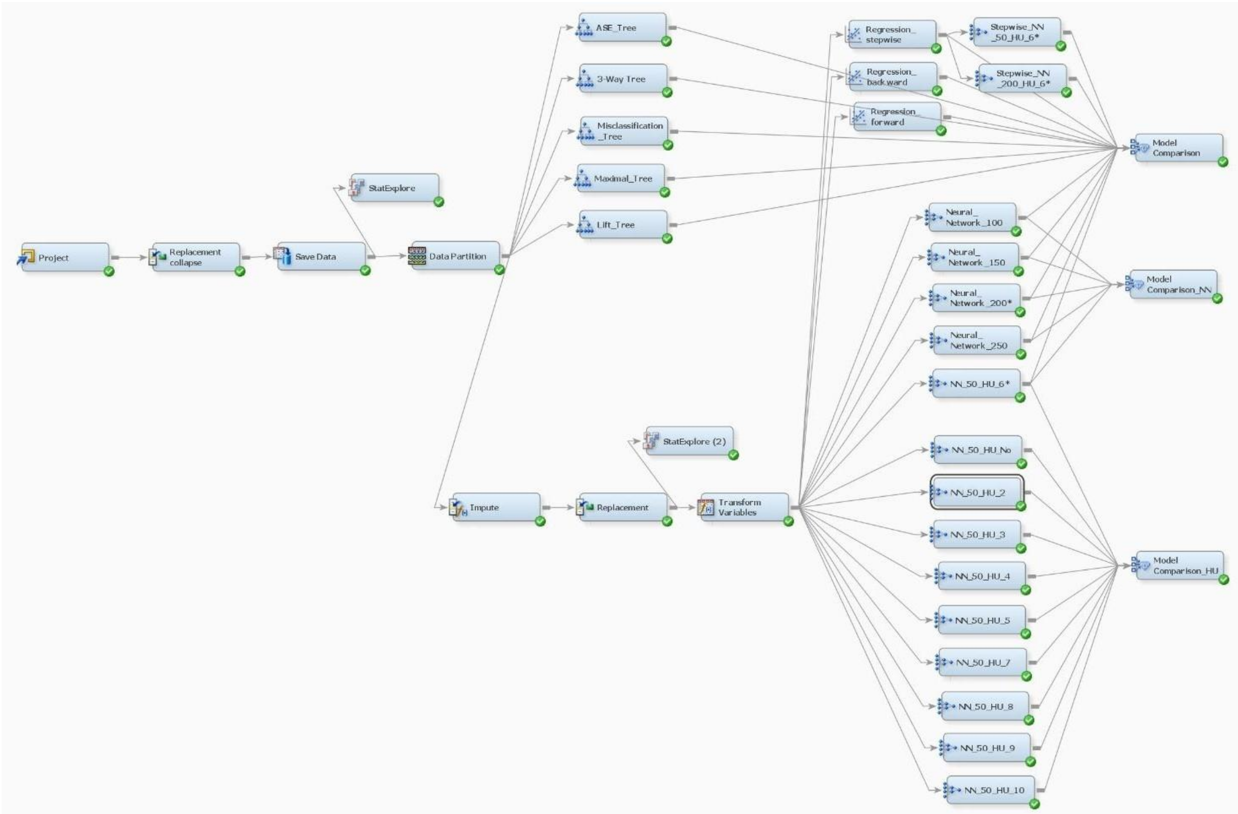
ProdTaken	_UNKNOWN_	_DEFAULT_	.		N		.
REP_MaritalStatus	Married			1148	C	Married	.
REP_MaritalStatus	Single			811	C	Single	.
REP_MaritalStatus	Divorced			484	C	Divorced	.
REP_MaritalStatus	_UNKNOWN_	_DEFAULT_	.		C		.

Next, I examined the source data. There are 4,888 observations in the dataset, along with missing values. The data is partitioned into training and validation data to optimize performance, the data is assigned 50% for training and 50% for validation.

General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0
Report	

	Name	Model Role	Measurement Level	Description
1	Age	Input	Interval	Age of the Customer
2	CityTier	Rejected	Interval	Level of destination city development.
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
19	ProductPitched	Rejected	Nominal	Advertised product during the marketing pitch
20	TypeofContact	Input	Nominal	How was customer interaction initiated?

SAS Diagram

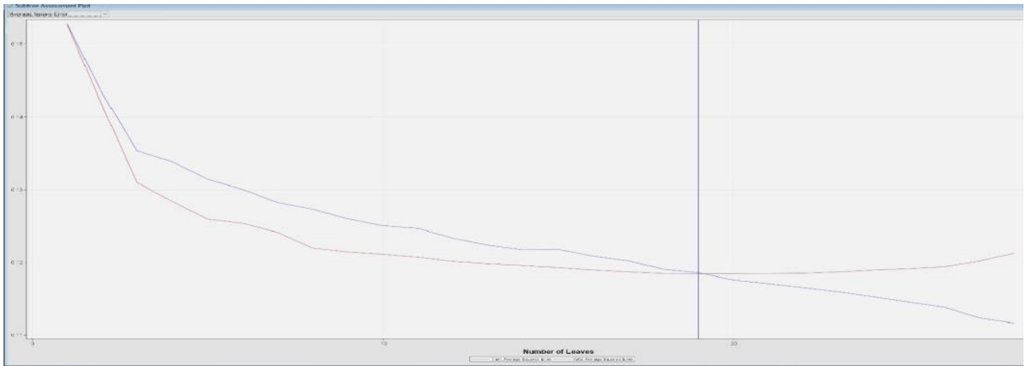


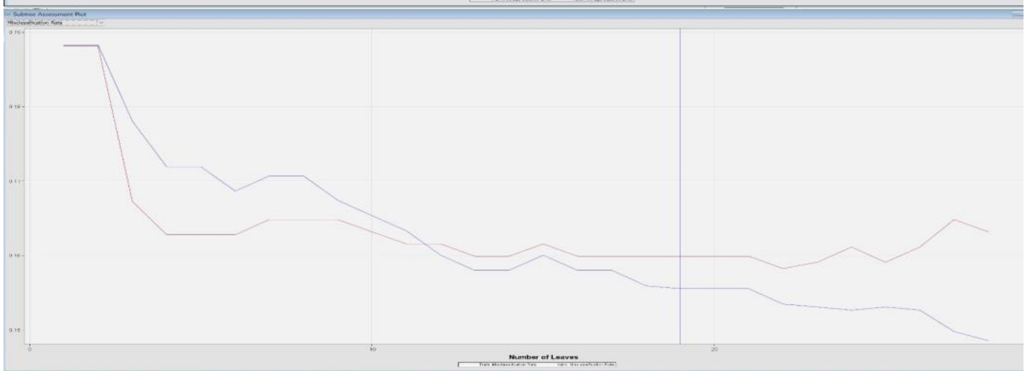
Decision Tree

After running the decision tree models (returning trees with the assessment measures below). I froze the model and disabled node training.

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

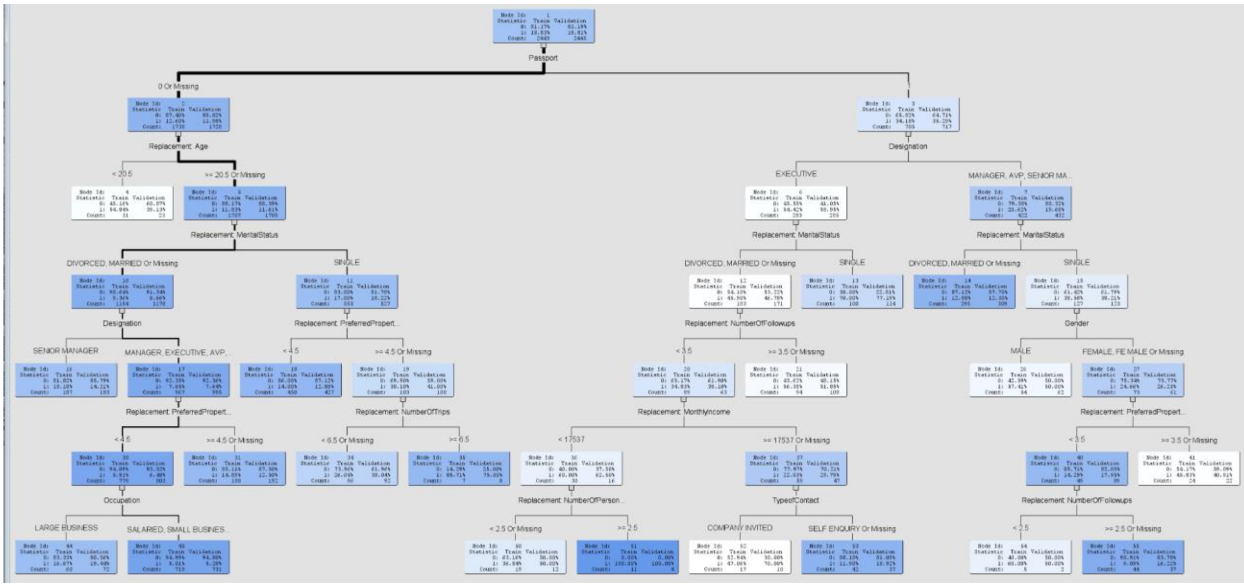
Probability Tree

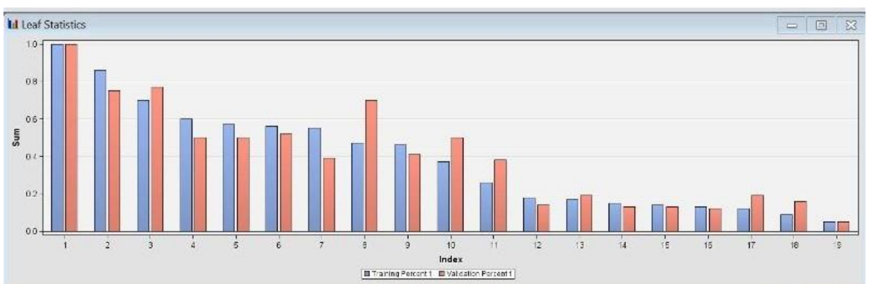




Use Frozen Tree	Yes
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf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

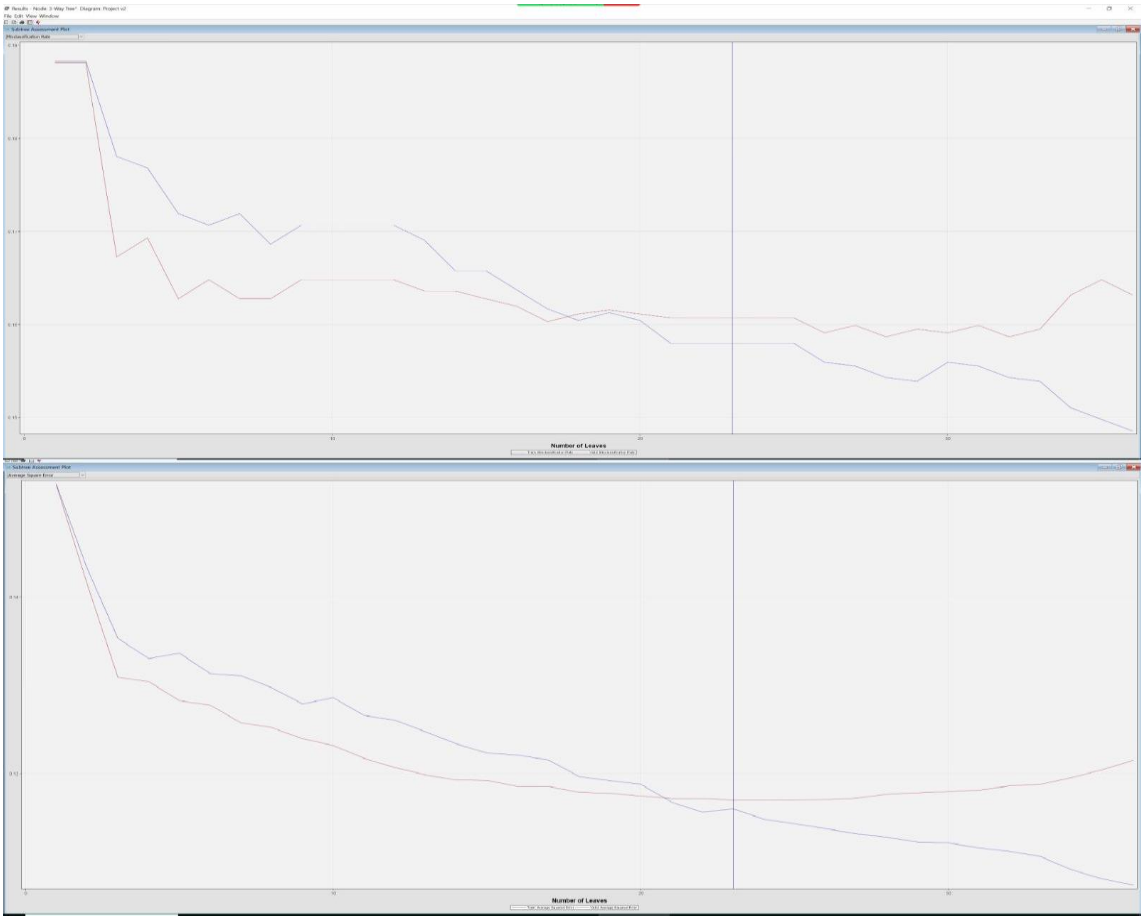
Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	2443	2445
MISC	Misclassification Rate	0.155546	0.159918
MAX	Maximum Absolute Error	0.94993	0.94993
SSE	Sum of Squared Errors	579.3291	579.0059
ASE	Average Squared Error	0.118569	0.118406
RASE	Root Average Squared...	0.344339	0.344102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



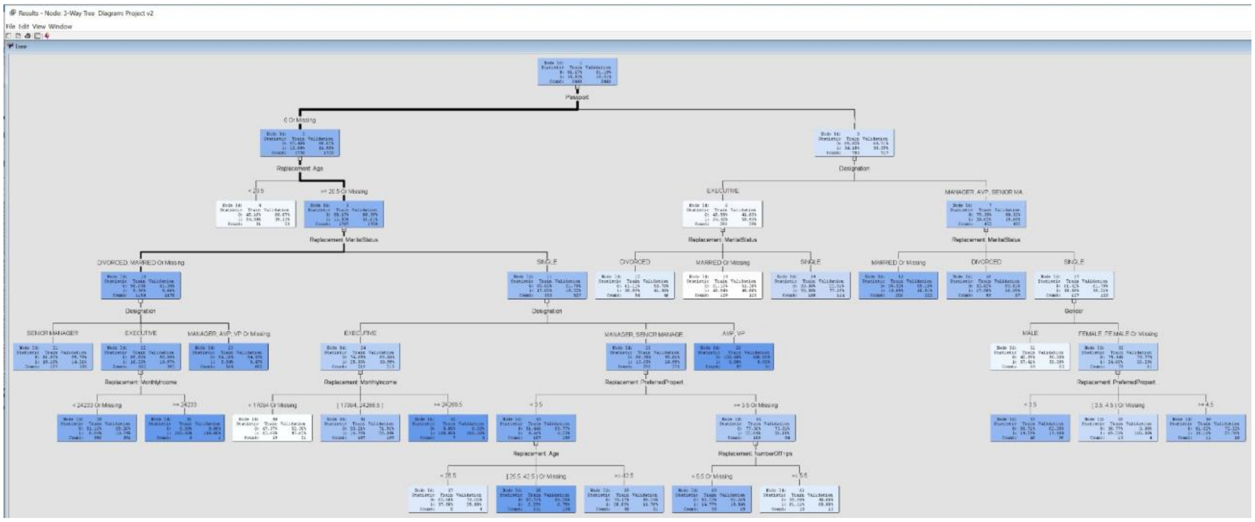


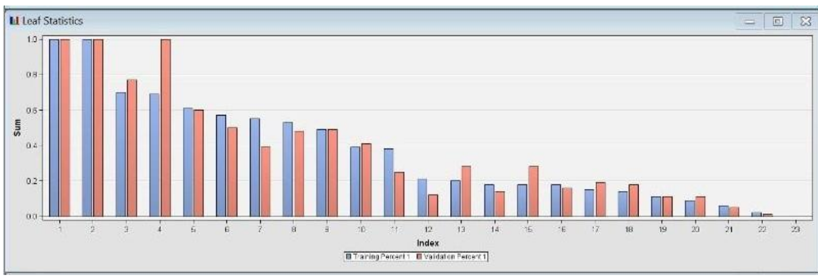
- The optimal number of leaves is 19.
- The variable used for the first split was Passport. The competing splits were Age and Designation. Other important variables are Marital Status, Gender, Preferred Property Score and Number of Follow-ups. As can be seen in the tree, variables like Monthly Income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

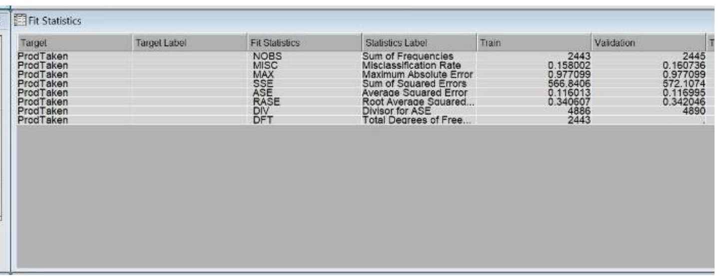
3-Way Tree



Tree Model Data Set	
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



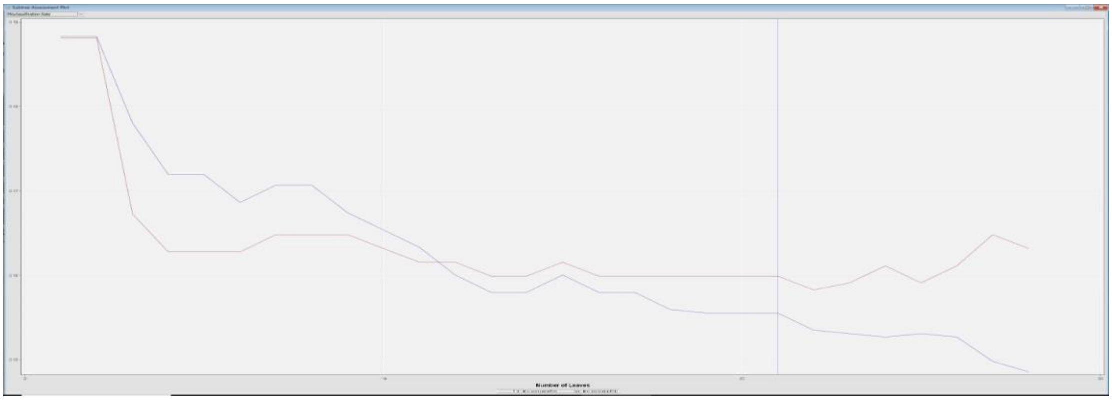


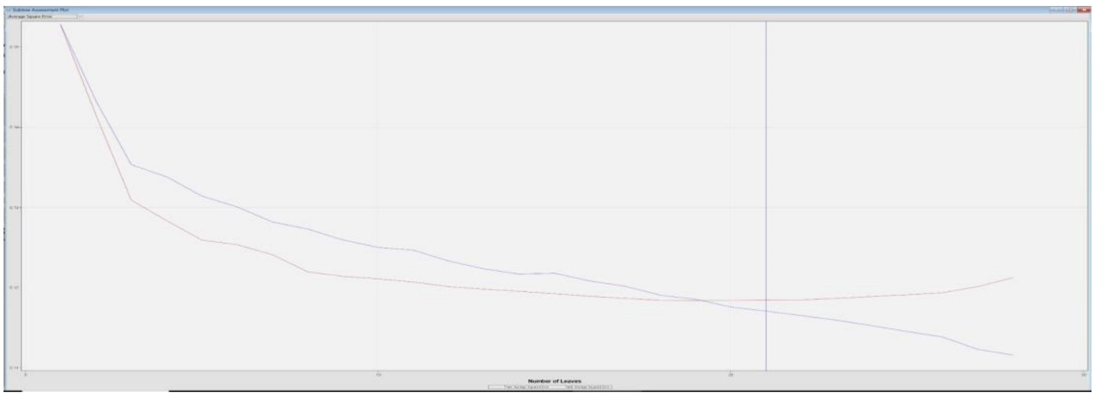

Fit Statistics

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	.

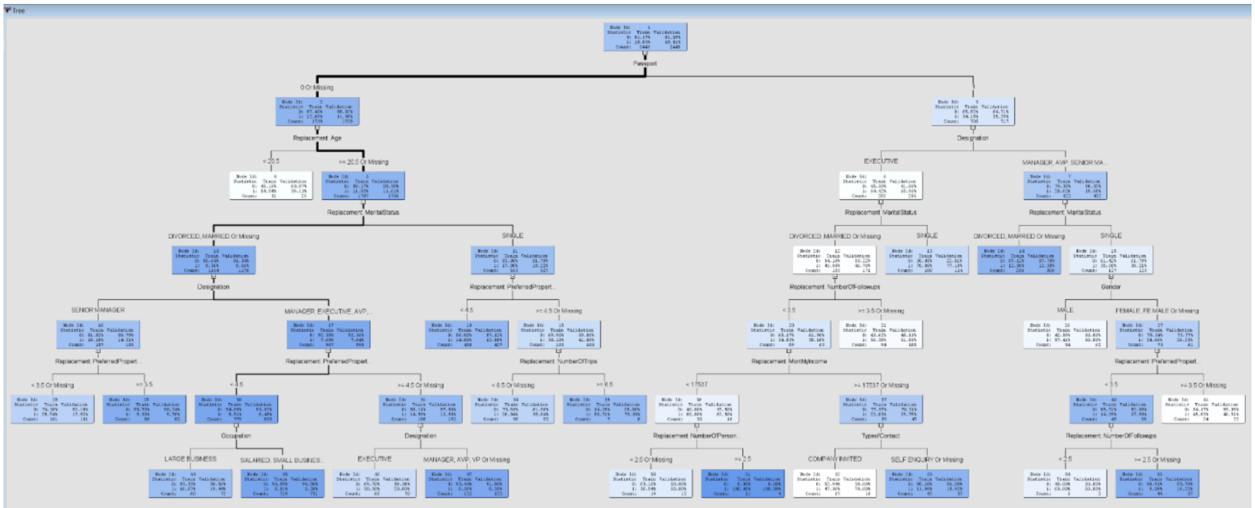
- The optimal number of leaves is 23.
- The variable used for the first split was Passport. The competing splits were Age and Designation. Another important variable is Marital Status. As can be seen in the tree, variables (Monthly Income, Preferred Property Score and Number of Trips) were included but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

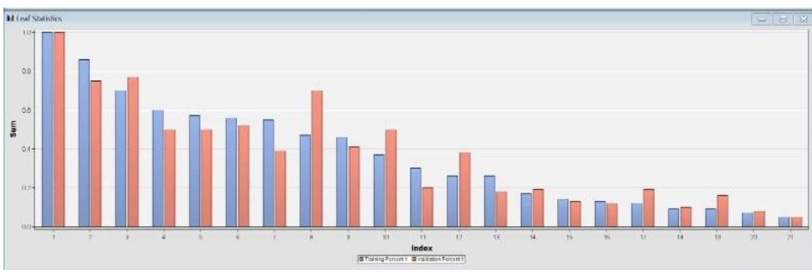
Lift Tree

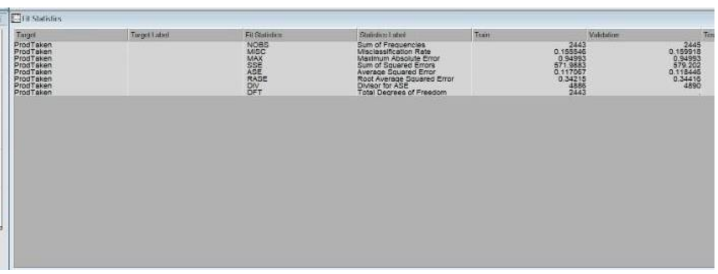




Use Frozen Tree	Yes
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf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



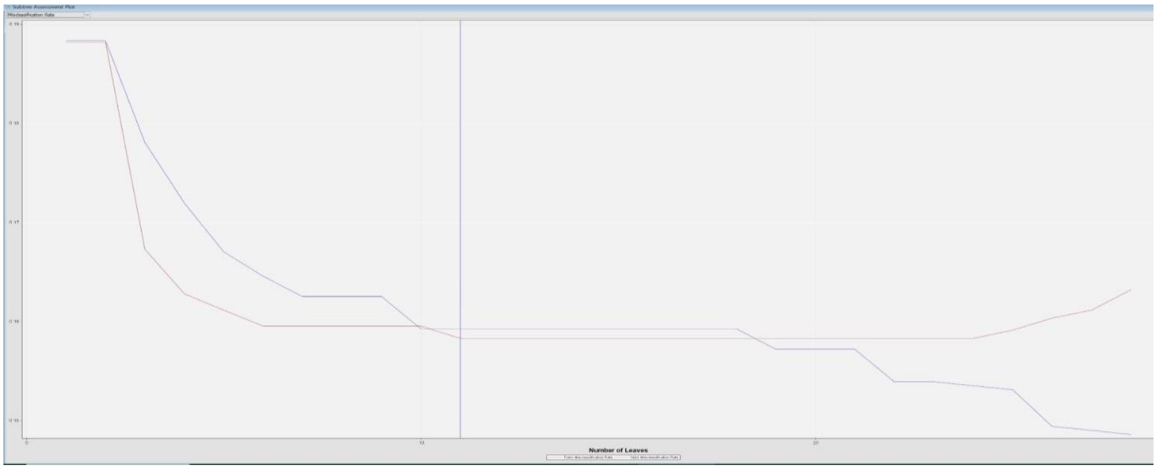


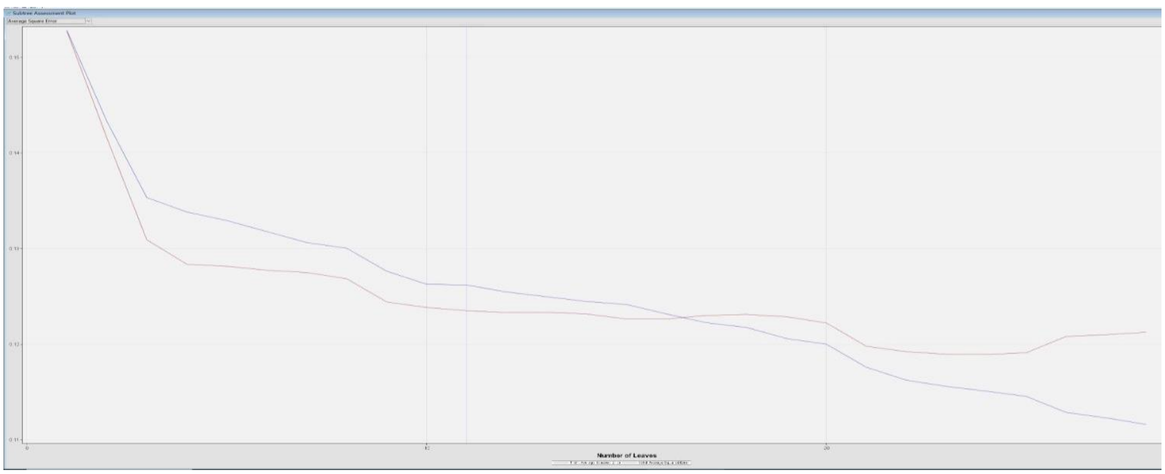


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
ProdTaken		NOBS	Sum of Frequencies	2443	2445
ProdTaken		MISC	Misclassification Rate	0.155546	0.159918
ProdTaken		MAX	Maximum Absolute Error	0.94993	0.94993
ProdTaken		SSE	Sum of Squared Errors	571.9883	579.202
ProdTaken		ASE	Average Squared Error	0.117067	0.118446
ProdTaken		RASE	Root Average Squared Error	0.34215	0.34416
ProdTaken		DIV	Divisor for ASE	4886	4890
ProdTaken		DFT	Total Degrees of Freedom	2443	

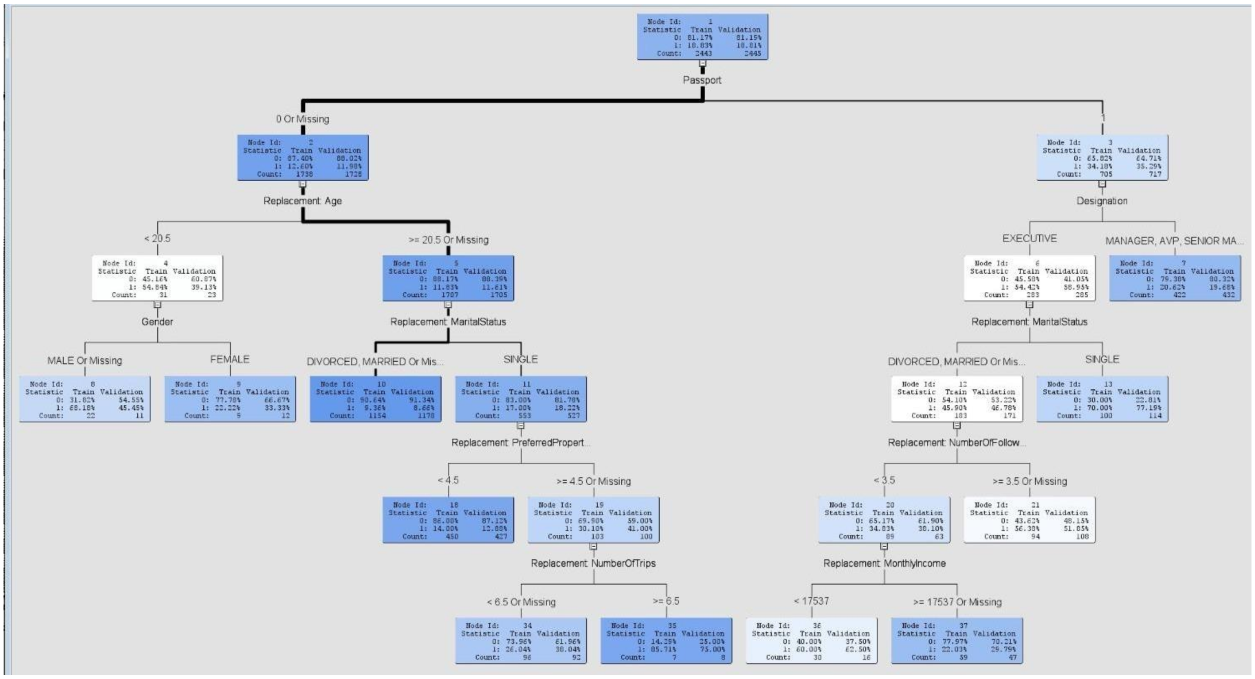
- The optimal number of leaves is 21.
- The variable used for the first split was Passport. The competing splits were Age and Designation. Another important variable is Marital Status. As can be seen in the tree, variables like Monthly Income, Gender, Preferred Property Score and Number of Follow-ups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

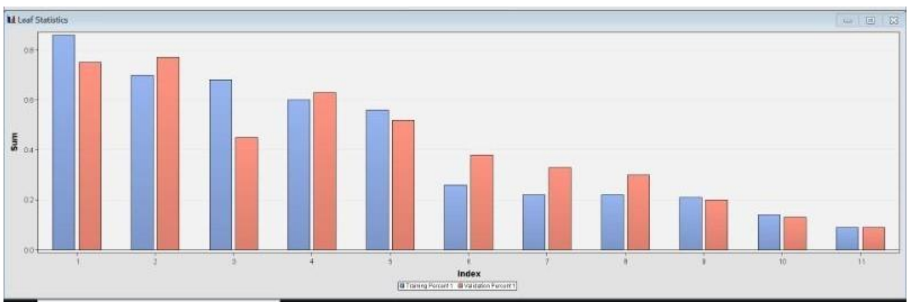
Misclassification Tree





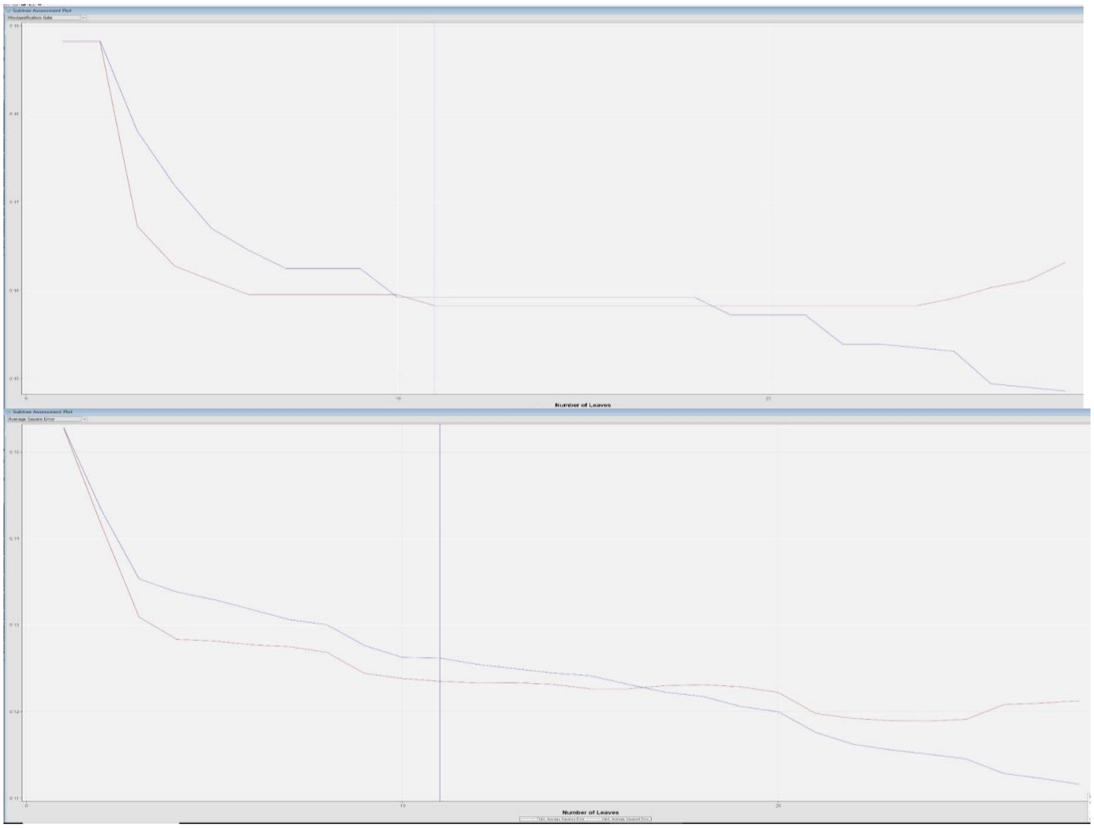
Use Frozen Tree	Yes
Use Multiple Targets	No
⊟Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
⊟Node	
Leaf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



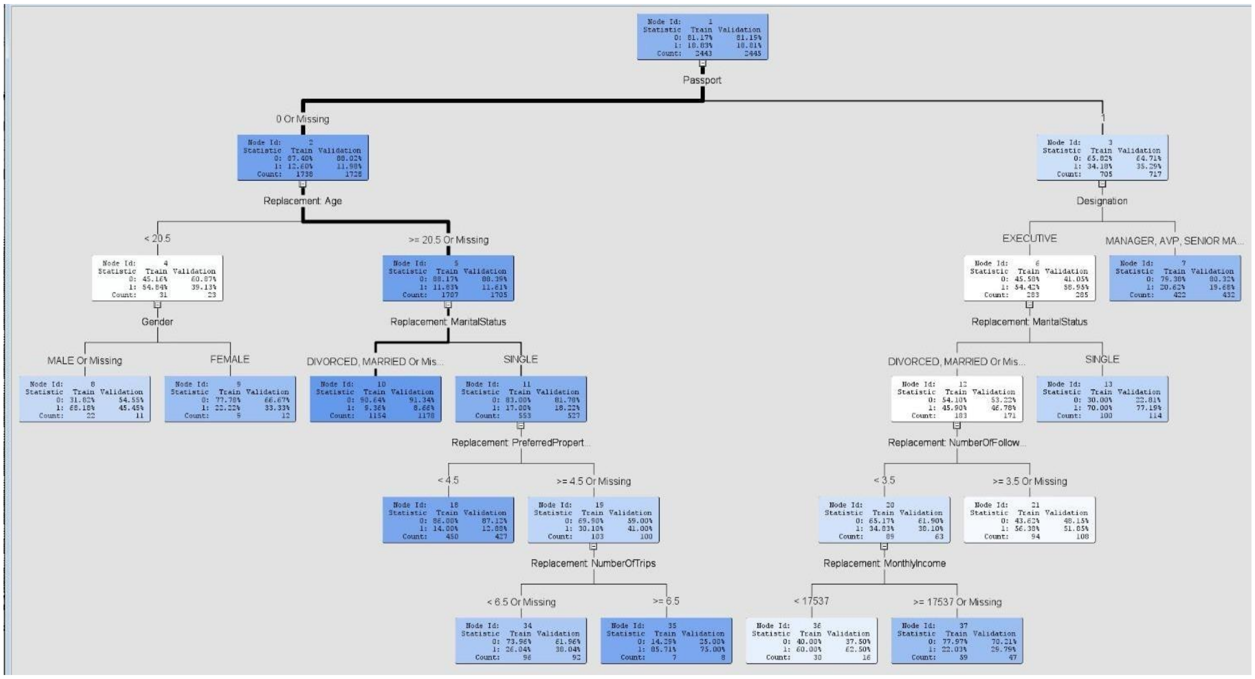


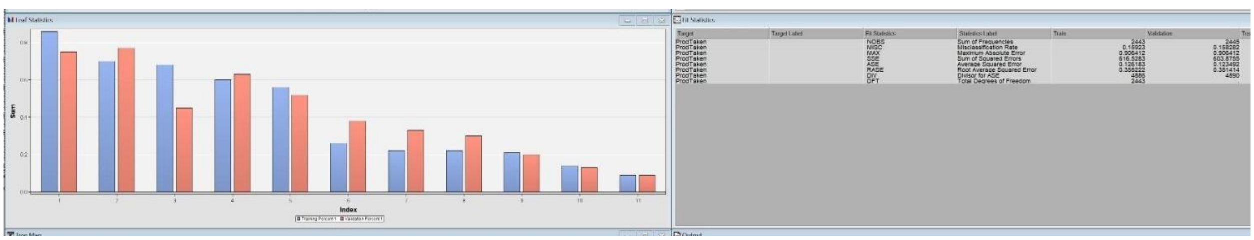
- The optimal number of leaves is 11.
- The variable used for the first split was Passport. The competing splits were Age and Designation. Other important variables are Marital Status and Gender. As can be seen in the tree, variables like Preferred Property Score and Number of Follow-Ups were included but less important.
- Valid average square error (ASE) = 0.123492
- Valid misclassification rate = 0.158282

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



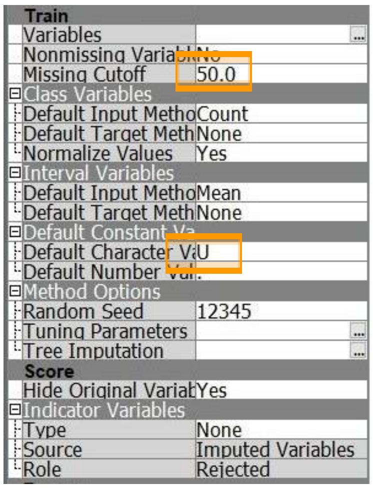


- The optimal number of leaves is 11.
- The variable used for the first split was Passport. The competing splits were Age and Designation. Other important variables are Marital Status and Gender. As can be seen in the tree, variables like Preferred Property Score and Number of Follow-Ups were included but less important.
- Valid average square error (ASE) = 0.123492
- Valid misclassification rate = 0.158282

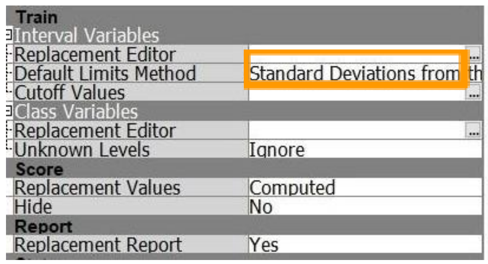
The 3-way Tree is the best decision tree, based on its lowest valid ASE and valid misclassification rate.

Logistic Regression

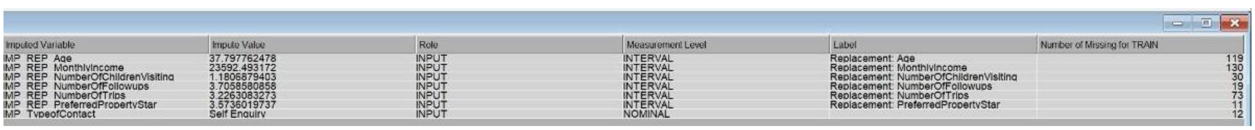
To prepare for linear regression, we used the Impute node (left) for the missing data. We then use an alternate node (on the right) with the default limiting method set to the value of the standard deviation from the mean. This is done to limit and rank outliers, reducing variables later we have to transform with logarithm. Using less logarithms will mean easier data communication for objects unfamiliar with data manipulation methods.



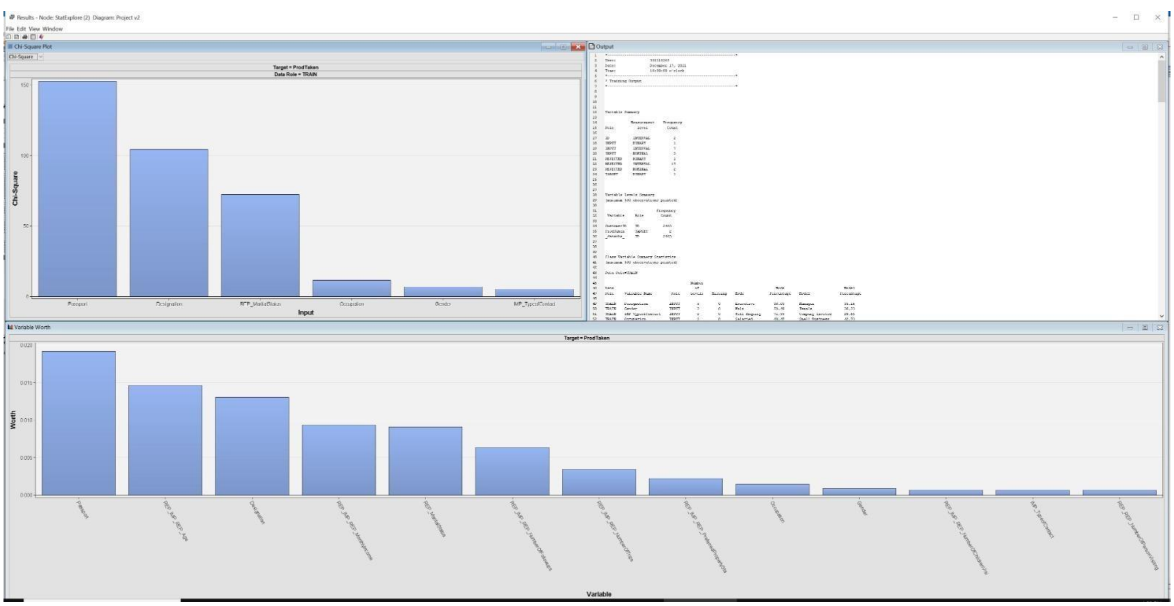
Train	
Variables	
Nonmissing Variab	No
Missing Cutoff	50.0
Class Variables	
Default Input Metho	Count
Default Target Meth	None
Normalize Values	Yes
Interval Variables	
Default Input Metho	Mean
Default Target Meth	None
Default Constant Va	
Default Character V	U
Default Number Val	
Method Options	
Random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

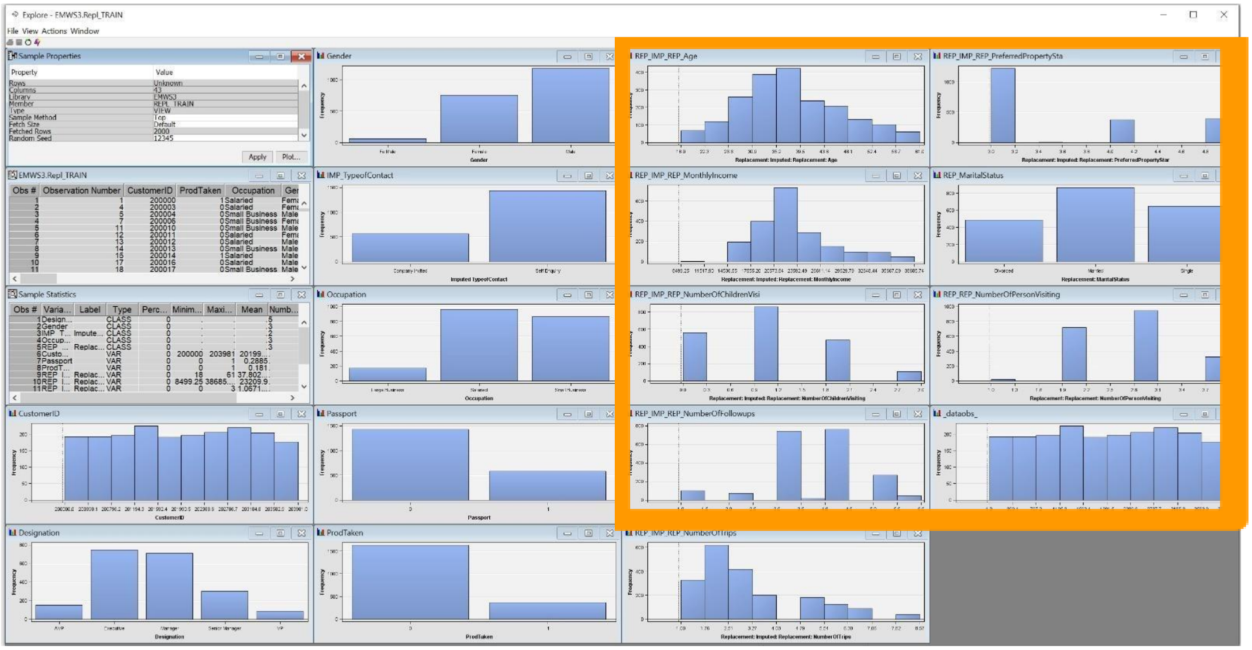


Train	
Interval Variables	
Replacement Editor	
Default Limits Method	Standard Deviations from th
Cutoff Values	
Class Variables	
Replacement Editor	
Unknown Levels	Ignore
Score	
Replacement Values	Computed
Hide	No
Report	
Replacement Report	Yes



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, the number of trips had skewed distribution.

REP_IMP_REP_NumberOfChildrenVisi	Default	4	Input	Interval
REP_IMP_REP_NumberOfFollowups	Default	4	Input	Interval
REP_IMP_REP_NumberOfTrips	Log	4	Input	Interval
REP_IMP_REP_PreferredPropertySta	Default	4	Input	Interval
REP_MaritalStatus	Default	4	Input	Nominal
REP_NumberOfPersonVisiting	Default	4	Rejected	Interval
REP_REP_NumberOfPersonVisiting	Default	4	Input	Interval

A Transform Node was applied to this variable.

Regression Results

For this dataset, backward, forward regression was used.

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

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

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
- 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.

Our odds ratios show that:

- With every added customer who is an AVP, VPs are 64.8% more 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

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	0.125582	
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

- 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.

Our odds ratios show that:

- With every added customer who is an AVP, VPs are 67.9% more 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.

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

```
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

- 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-

Our odds ratios show that:

- With every added customer who is an AVP, VPs are 67.9% more likely to buy a package.
- With every added customer who is an Executive, Executives 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, Stepwise regression is the best model.

Neural Network

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.

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

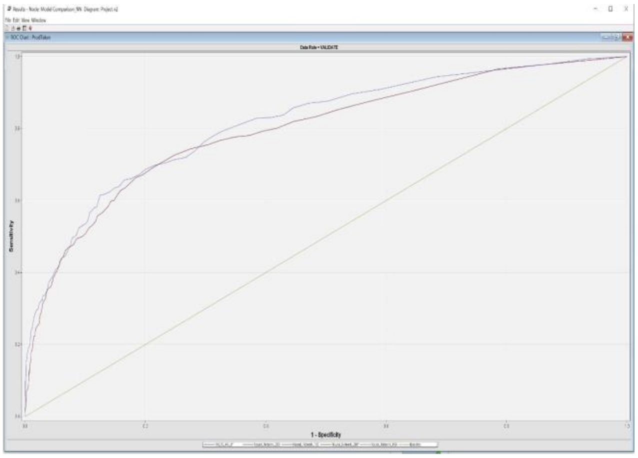
The result shows that the neural network with 6 hidden units was the best model as it had the lowest valid average square error (0.113577), the highest ROC index (0.814, shared with the model with 8 hidden units) and the highest Gini coefficient (0.627). It also had one of the lower misclassification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

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

The neural network converged at 200 iterations. However, the iteration plot was not ideal.



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 (

Input Reduction

Input reduction is attached to the best regression model (Stepwise) and ran with neural networks with 50 and 200 iterations.

Results - Node: Stepwise_NN_50_HU_6* Diagram: Project v2

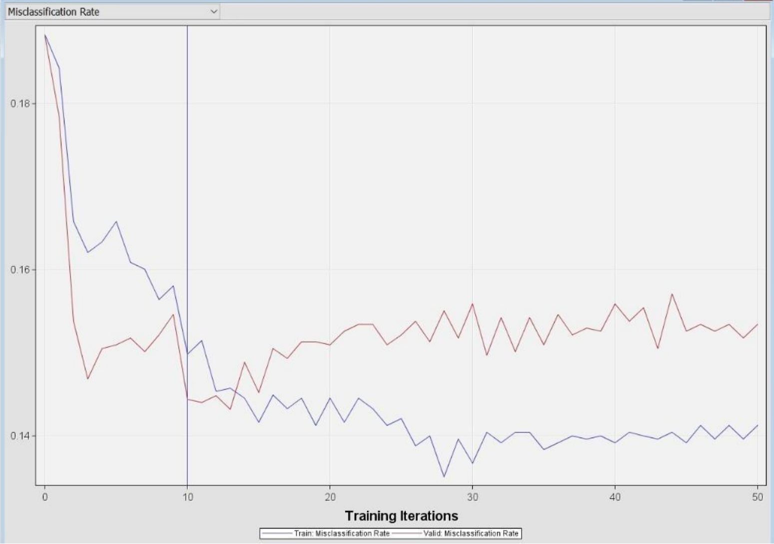
File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
ProdTaken		DFT	Total Degrees of Fr...	2443	.	.
ProdTaken		DFE	Degrees of Freedo...	2352	.	.
ProdTaken		DFM	Model Degrees of F...	91	.	.
ProdTaken		NW	Number of Estimate...	91	.	.
ProdTaken		AIC	Akaike's Information...	2030.11	.	.
ProdTaken		SBC	Schwarz's Bayesian...	2557.999	.	.
ProdTaken		ASE	Average Squared E...	0.116332	0.117112	.
ProdTaken		MAX	Maximum Absolute ...	0.986698	0.985714	.
ProdTaken		DIV	Divisor for ASE	4886	4890	.
ProdTaken		NOBS	Sum of Frequencies	2443	2445	.
ProdTaken		RASE	Root Average Squa...	0.341074	0.342216	.
ProdTaken		SSE	Sum of Squared Err...	568.3959	572.6754	.
ProdTaken		SUMW	Sum of Case Weigh...	4886	4890	.
ProdTaken		FPE	Final Prediction Error	0.125333	.	.
ProdTaken		MSE	Mean Squared Error	0.120832	0.117112	.
ProdTaken		RFPE	Root Final Predictio...	0.354025	.	.
ProdTaken		RMSE	Root Mean Square...	0.34761	0.342216	.
ProdTaken		AVERR	Average Error Func...	0.378246	0.382928	.
ProdTaken		ERR	Error Function	1848.11	1872.516	.
ProdTaken		MISC	Misclassification Rate	0.155546	0.155419	.
ProdTaken		WRONG	Number of Wrong C...	380	380	.

Iteration Plot

Misclassification Rate



Output

17	0	19	0	0.35533	0.000724	0.0179	0.0106	0.138
18	0	20	0	0.35412	0.00121	0.0153	0.0129	0.196
19	0	21	0	0.35311	0.00102	0.0192	0.0111	0.182
20	0	22	0	0.35203	0.00108	0.0168	0.0127	0.177
21	0	23	0	0.35100	0.00102	0.0204	0.0109	0.180
22	0	24	0	0.35014	0.000865	0.0181	0.0119	0.146
23	0	25	0	0.34925	0.000889	0.0218	0.0106	0.157
24	0	26	0	0.34850	0.000746	0.0192	0.0112	0.128
25	0	27	0	0.34772	0.000786	0.0228	0.0105	0.139
26	0	28	0	0.34703	0.000685	0.0199	0.0108	0.118
27	0	29	0	0.34634	0.000696	0.0234	0.0107	0.123
28	0	30	0	0.34571	0.000627	0.0205	0.0106	0.110
29	0	31	0	0.34505	0.000662	0.0236	0.0107	0.118
30	0	32	0	0.34439	0.000661	0.0209	0.0107	0.117
31	0	33	0	0.34373	0.000661	0.0236	0.0103	0.122
32	0	34	0	0.34316	0.000567	0.0215	0.0103	0.104
33	0	35	0	0.34264	0.000521	0.0239	0.00943	0.100
34	0	36	0	0.34228	0.000358	0.0221	0.00951	0.0675
35	0	37	0	0.34043	0.00185	0.0136	0.0348	0.497
36	0	38	0	0.33989	0.000535	0.0102	0.00772	0.485
37	0	39	0	0.33955	0.000339	0.0104	0.00711	0.369
38	0	40	0	0.33932	0.000235	0.00982	0.00789	0.248
39	0	41	0	0.33910	0.000216	0.0103	0.00875	0.212
40	0	42	0	0.33894	0.000163	0.00995	0.00848	0.153
41	0	43	0	0.33874	0.000194	0.0101	0.00960	0.171
42	0	44	0	0.33861	0.000131	0.00975	0.00876	0.118
43	0	45	0	0.33844	0.000177	0.00967	0.00992	0.151
44	0	46	0	0.33833	0.000106	0.00938	0.00879	0.0944
45	0	47	0	0.33792	0.000409	0.00587	0.0348	0.508
46	0	48	0	0.33776	0.000159	0.00394	0.00890	0.680
47	0	49	0	0.33763	0.000133	0.00349	0.00893	0.655
48	0	50	0	0.33750	0.000130	0.00300	0.00829	0.659
49	0	51	0	0.33720	0.000298	0.00403	0.00252	0.687
50	0	52	0	0.33701	0.000198	0.00556	0.00229	0.357

Optimization Results

Iterations	50	Function Calls	54
Jacobian Calls	52	Active Constraints	0
Objective Function	0.3370058143	Max Abs Gradient Element	0.0055612377
Lambda	0.0022862068	Actual Over Pred Change	0.3565759958
Radius	0.3538150309		

LEVMAR needs more than 50 iterations or 2147483647 function calls.

WARNING: LEVMAR Optimization cannot be completed.

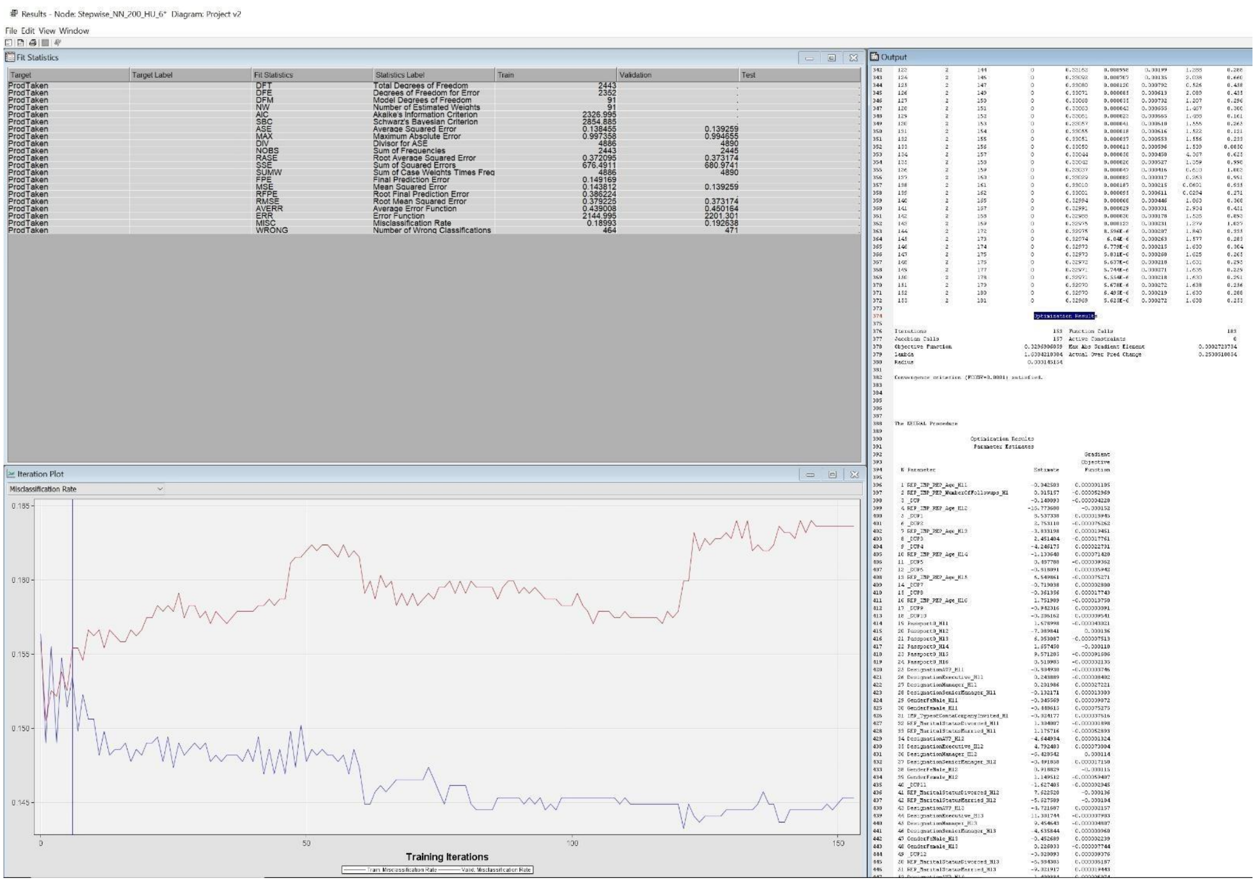
The NEURAL Procedure

Optimization Results
Parameter Estimates

N	Parameter	Estimate	Gradient Objective Function
1	REP_IMP_REP_Age_H11	-0.504099	-0.000434
2	REP_IMP_REP_NumberOfFollowups_H1	-0.108746	-0.000174
3	_DUP	-0.659958	-0.000407
4	REP_IMP_REP_Age_H12	-0.310279	-0.001144
5	_DUP1	-0.030538	0.000654
6	_DUP2	-0.899016	-0.000568
7	REP_IMP_REP_Age_H13	-1.238812	-0.000396
8	_DUP3	-0.060381	0.000295
9	_DUP4	0.744962	0.000194
10	REP_IMP_REP_Age_H14	1.038308	-0.000058178
11	_DUP5	-0.682583	-0.000101
12	_DUP6	-0.135363	0.000209
13	REP_IMP_REP_Age_H15	-0.305043	-0.002168
14	_DUP7	0.328655	0.005561
15	_DUP8	-0.254151	0.001587
16	REP_IMP_REP_Age_H16	1.713515	0.000958
17	_DUP9	-0.513104	-0.000134
18	_DUP10	0.771554	-0.000469
19	Passport0_H11	0.766436	-0.000139
20	Passport0_H12	1.643362	-0.000247
21	Passport0_H13	-0.703181	0.000109
22	Passport0_H14	1.882602	-0.000566
23	Passport0_H15	-0.206186	0.003178
24	Passport0_H16	-0.004489	0.000407
25	DesignationAVP_H11	-1.944042	0.000414
26	DesignationExecutive_H11	1.812883	-0.000075281
27	DesignationManager_H11	0.610533	-0.000150
28	DesignationSeniorManager_H11	-0.284708	-0.000054451
29	GenderFeMale_H11	-3.138954	0.000326
30	GenderFemale_H11	0.355129	-0.000296
31	IMP_TypeofContaCompanyInvited_H1	-0.412645	0.000373
32	REP_MaritalStatusDivorced_H11	-0.458494	-0.000012451
33	REP_MaritalStatusMarried_H11	-0.696939	-0.000170
34	DesignationAVP_H12	1.370875	-0.000292
35	DesignationExecutive_H12	-2.637902	-0.000443
36	DesignationManager_H12	-1.827676	0.000433
37	DesignationSeniorManager_H12	1.584261	-0.000292
38	GenderFeMale_H12	0.201727	0.000749
39	GenderFemale_H12	-0.926962	0.001338
40	_DUP11	0.331098	-0.000600
41	REP_MaritalStatusDivorced_H12	0.713311	0.000635
42	REP_MaritalStatusMarried_H12	0.870577	0.000171
43	DesignationAVP_H13	1.920196	0.000003392
44	DesignationExecutive_H13	1.268529	0.000395
45	DesignationManager_H13	-1.472869	0.000111

The iteration plot 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 observed in the neural network model with 200 iterations with reduced inputs. The iteration plot improved 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).

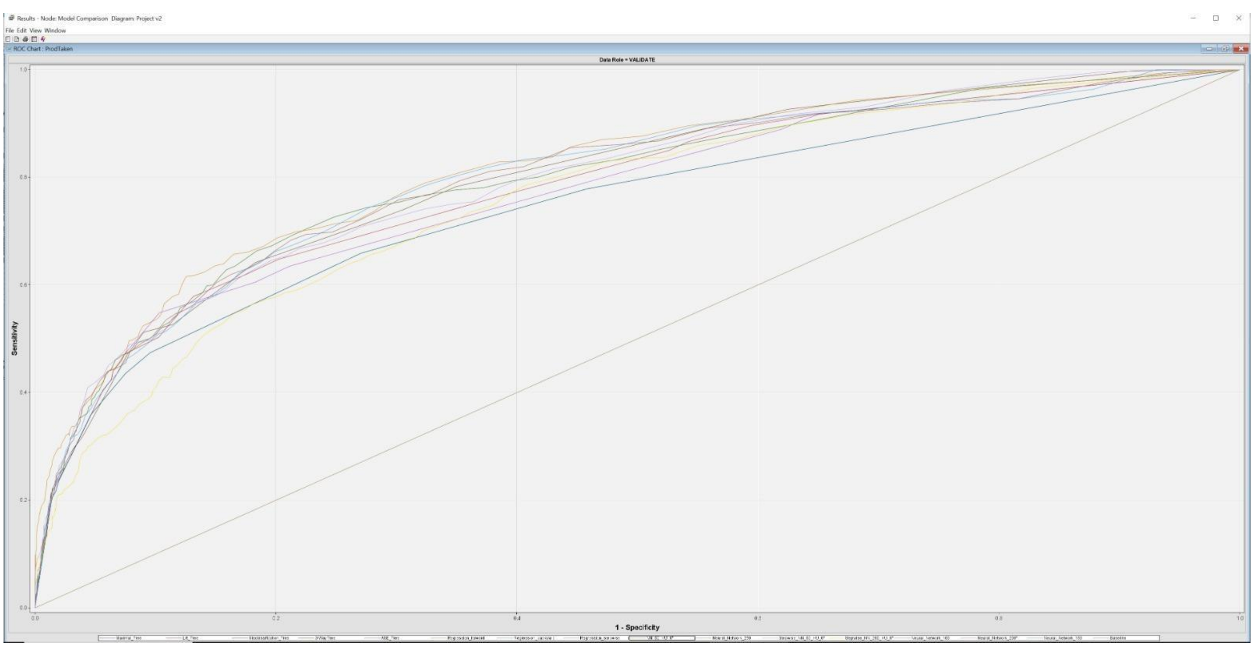


Model Comparison

The models are compared 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.

Model Node	Model Description	Valid: Average Squared Error ▲	Valid: Gini Coefficient	Valid: Misclassification Rate	Valid: Kolmogorov-Smirnov Statistic	Valid: Roc Index
Neural7	NN 50 HU 6*	0.113577	0.627	0.15501	0.492	0.814
Tree2	3-Way Tree	0.116995	0.599	0.160736	0.458	0.8
Neural16	Stepwise NN 50 HU 6*	0.117112	0.592	0.155419	0.449	0.796
Reg	Regression stepwise	0.11765	0.597	0.156646	0.47	0.799
Reg3	Regression forward	0.11765	0.597	0.156646	0.47	0.799
Neural12	Neural Network 150	0.117823	0.595	0.153783	0.479	0.797
Neural13	Neural Network 200*	0.117823	0.595	0.153783	0.479	0.797
Neural14	Neural Network 100	0.117823	0.595	0.153783	0.479	0.797
Neural2	Neural Network 250	0.117823	0.595	0.153783	0.479	0.797
Reg2	Regression backward	0.117825	0.599	0.155828	0.466	0.799
Tree	ASE Tree	0.118406	0.55	0.159918	0.445	0.775
Tree4	Lift Tree	0.118446	0.567	0.159918	0.447	0.783
Tree3	Misclassification Tree	0.123492	0.494	0.158282	0.388	0.747
Tree5	Maximal Tree	0.123492	0.494	0.158282	0.388	0.747
Neural15	Stepwise NN 200 HU 6*	0.139259	0.522	0.192638	0.383	0.761



On the basis of this comparison, the neural network node with 50 iterations and 6 hidden units is our best model. It has 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 others.
The next best model based on these statistics was the 3-way decision tree.

Conclusion

In conclusion, based on the analysis above; Passport, job designation and age are critical variables affecting this prediction model. 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 greater than or equal to $24,266.50 USD
- Divorced, married (or missing marital status) executives without a passport who have a monthly income more than or equal to $24,233 USD

As mentioned in the previous sections, it would be helpful if the company could provide information indicating whether these trips are domestic or international. Since passport is such an important variable, it will significantly increase the accuracy of the model, as it will determine whether the customer must have a passport to use the travel package

References

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