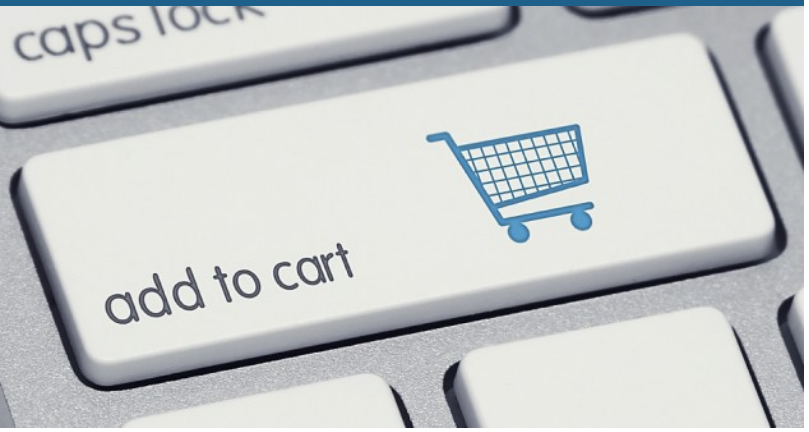
Predicting Online Shopping Purchasing Intention

Content



Part 1 Introduction

Part 2 Exploratory Data Analysis

Part 3 Tools, Techniques, Modeling

Part 4 Model Comparison

Part 5 Best Model

Part 6 Conclusion and Recommendations

PART 1

Introduction

- Business Problem
- Dataset Description
- Initial Exploration

ONLINE SHOPPING

11/11/2019

Online Shopping Festival

$ 38,340,000,000

Buyers ⇄ **Sellers**

Convenient

Employment
Profits

Business Problem

- Severe Competition
- Leave page without buying

Predictive Modeling

- Probability of Purchase
- Most Important Factors

Business Value

- Stand Out from Competition
- Profits, Brand, Reputation

2 Dataset Introduction and Variables

UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Online Shoppers Purchasing Intention Dataset

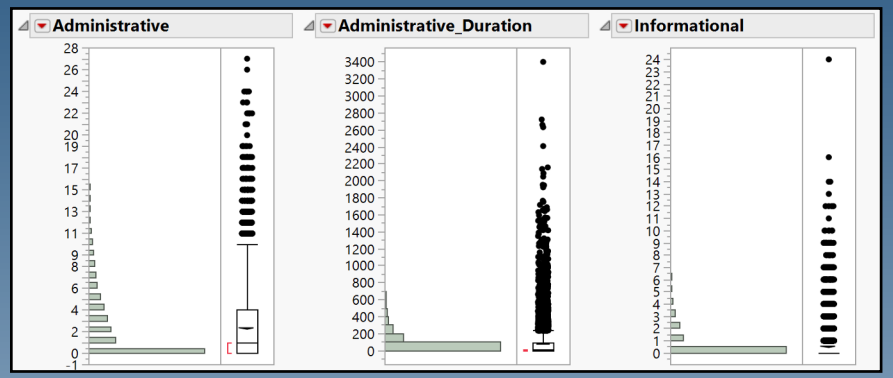
- 12,330 sessions
- 14 Numerical Attributes
- 4 Categorical Attributes
- Categorical Target: Revenue

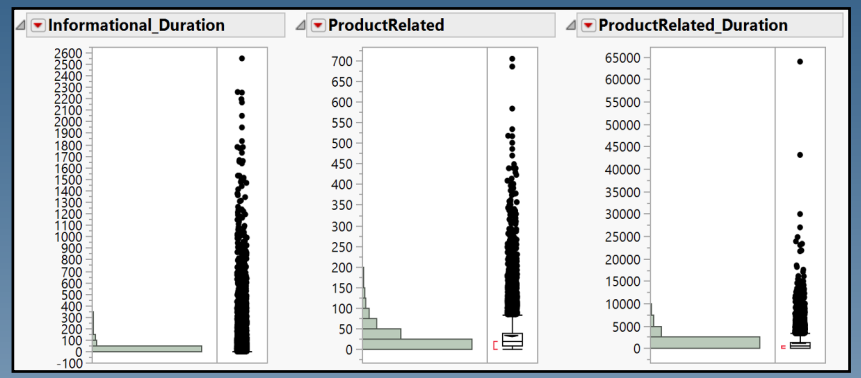
  TRUE: Purchase

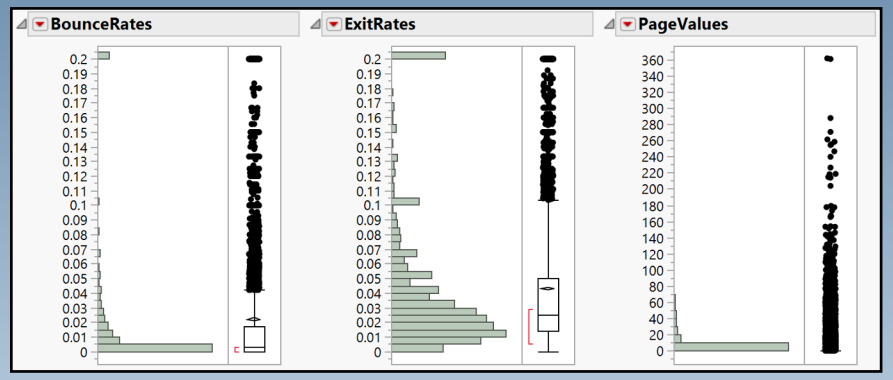
  FALSE: Not Purchase

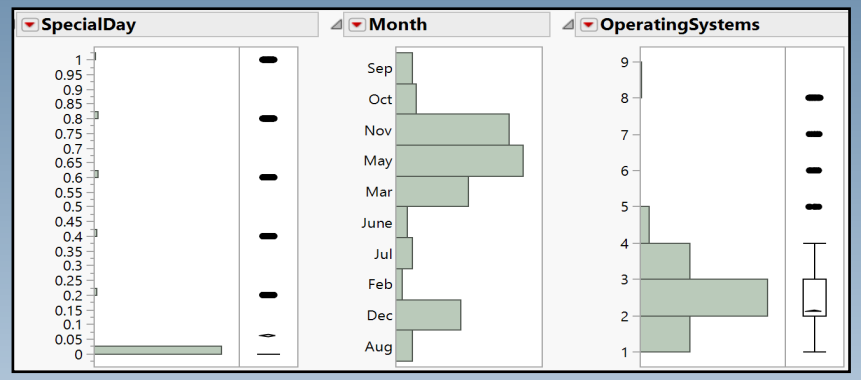
Variable Name	Variable Type	Description
Revenue	Categorical	Whether the session generated a revenue or not.
Administrative	Numerical	Number of Administrative pages visited by the visitor
Administative_Duration	Numerical	Total Time spent in Administrative pages by the visitor
Informational	Numerical	Number of Informational pages visited by the visitor
Informational_Duration	Numerical	Total Time spent in Informational pages by the visitor
ProductRelated	Numerical	Number of ProductRelated pages visited by the visitor
ProductRelated_Duration	Numerical	Total Time spent in ProductRelated pages by the visitor
BounceRates	Numerical	The percentage of visitors who enter the page then bounce without triggering other requests
ExitRates	Numerical	The percentage of the page is the last viewed in the session
PageValues	Numerical	The average value of a webpage that a customer visited before accomplishing the purchase.
SpecialDay	Numerical	The closeness of the visit time of a webpage to a specific special day or festival.
Month	Categorical	The month of year
OperationSystems	Numerical	The 8 different operation systems used by visitors
Browser	Numerical	The 13 different browsers used by visitors
Region	Numerical	The 9 different regions of visitors.
TrafficType	Numerical	The 20 different traffic types
VistorType	Categorical	Three different types of visitors.
Weekend	Categorical	The weekend or weekdays.

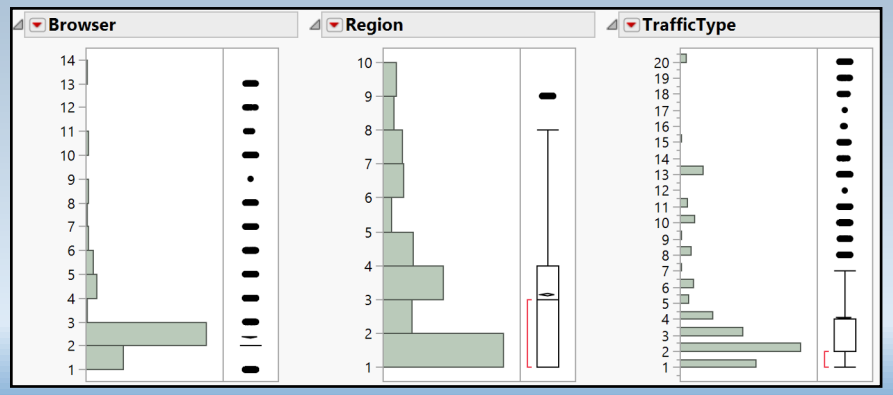
3 Original Variable Distributions











PART 2

Exploratory Data Analysis

- Correlation between variables
- Variable Transformation
- Outliers & Missing Value

1 Correlation among variables

Multivariate

Correlations

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	OperatingSystems	Browser	Region	TrafficType
Administrative	1.0000	0.6016	0.3769	0.2558	0.4311	0.3739	-0.2236	-0.3165	0.0990	-0.0948	-0.0063	-0.0250	-0.0055	-0.0336
Administrative_Duration	0.6016	1.0000	0.3027	0.2380	0.2891	0.3554	-0.1442	-0.2058	0.0676	-0.0733	-0.0073	-0.0154	-0.0056	-0.0144
Informational	0.3769	0.3027	1.0000	0.6190	0.3742	0.3875	-0.1161	-0.1637	0.0486	-0.0482	-0.0095	-0.0382	-0.0292	-0.0345
Informational_Duration	0.2558	0.2380	0.6190	1.0000	0.2800	0.3474	-0.0741	-0.1053	0.0309	-0.0306	-0.0096	-0.0193	-0.0271	-0.0247
ProductRelated	0.4311	0.2891	0.3742	0.2800	1.0000	0.8609	-0.2046	-0.2925	0.0563	-0.0240	0.0043	-0.0131	-0.0381	-0.0431
ProductRelated_Duration	0.3739	0.3554	0.3875	0.3474	0.8609	1.0000	-0.1845	-0.2520	0.0528	-0.0364	0.0030	-0.0074	-0.0331	-0.0364
BounceRates	-0.2236	-0.1442	-0.1161	-0.0741	-0.2046	-0.1845	1.0000	0.9130	-0.1194	0.0727	0.0238	-0.0158	-0.0065	0.0783
ExitRates	-0.3165	-0.2058	-0.1637	-0.1053	-0.2925	-0.2520	0.9130	1.0000	-0.1745	0.1022	0.0146	-0.0044	-0.0089	0.0786
PageValues	0.0990	0.0676	0.0486	0.0309	0.0563	0.0528	-0.1194	-0.1745	1.0000	-0.0635	0.0185	0.0456	0.0113	0.0125
SpecialDay	-0.0948	-0.0733	-0.0482	-0.0306	-0.0240	-0.0364	0.0727	0.1022	-0.0635	1.0000	0.0127	0.0035	-0.0161	0.0523
OperatingSystems	-0.0063	-0.0073	-0.0095	-0.0096	0.0043	0.0030	0.0238	0.0146	0.0185	0.0127	1.0000	0.2230	0.0768	0.1892
Browser	-0.0250	-0.0154	-0.0382	-0.0193	-0.0131	-0.0074	-0.0158	-0.0044	0.0456	0.0035	0.2230	1.0000	0.0974	0.1119
Region	-0.0055	-0.0056	-0.0292	-0.0271	-0.0381	-0.0331	-0.0065	-0.0089	0.0113	-0.0161	0.0768	0.0974	1.0000	0.0475
TrafficType	-0.0336	-0.0144	-0.0345	-0.0247	-0.0431	-0.0364	0.0783	0.0786	0.0125	0.0523	0.1892	0.1119	0.0475	1.0000

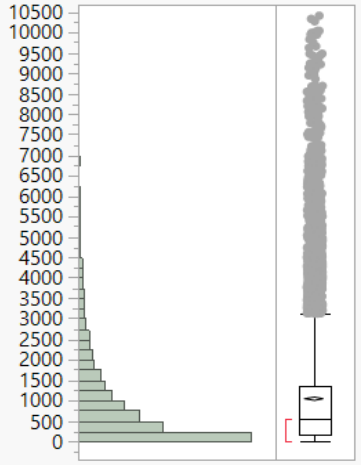
Product Relation variable
Product Related duration variable

Exit Rates variable
Bounce Rates variable

2 Variable Transformations

Distributions

ProductRelated_Duration



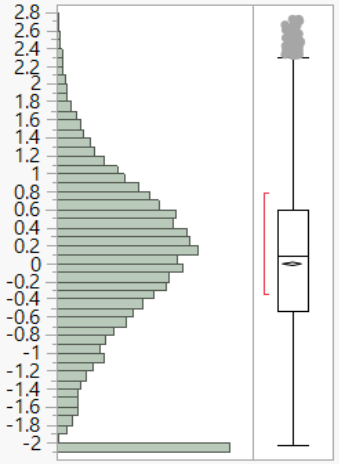
Quantiles

100.0%	maximum	10415.78364
99.5%		8140.4148927
97.5%		5064.1038755
90.0%		2618.9811146
75.0%	quartile	1350.4187503
50.0%	median	555.74583335
25.0%	quartile	168.69464285
10.0%		30.5
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	1038.9526
Std Dev	1371.8053
Std Err Mean	12.813338
Upper 95% Mean	1064.0689
Lower 95% Mean	1013.8362
N	11462

SHASH Transform ProductRelated_Duration



Quantiles

100.0%	maximum	2.7226971354
99.5%		2.3395972314
97.5%		1.7282430468
90.0%		1.0855660972
75.0%	quartile	0.5995500359
50.0%	median	0.0824074307
25.0%	quartile	-0.54130125
10.0%		-1.393075685
2.5%		-2.016337309
0.5%		-2.016337309
0.0%	minimum	-2.016337309

Summary Statistics

Mean	-0.013619
Std Dev	0.940488
Std Err Mean	0.0087846
Upper 95% Mean	0.0036004
Lower 95% Mean	-0.030838
N	11462

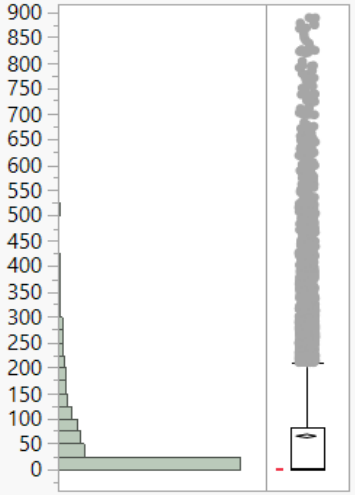
All rows with ProductRelated_Duration=0 is excluded

This effectively removes 756 rows of redundant data

2 Variable Transformations

Distributions

Administrative_Duration



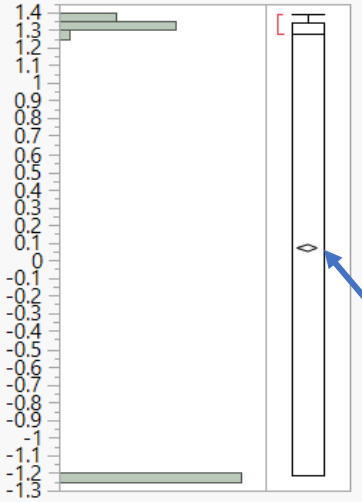
Quantiles

100.0%	maximum	889.1666667
99.5%		709.14368332
97.5%		423.0425
90.0%		200.39166668
75.0%	quartile	84
50.0%	median	3.25
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	65.434264
Std Dev	118.79356
Std Err Mean	1.1095904
Upper 95% Mean	67.609251
Lower 95% Mean	63.259277
N	11462

SHASH Transform Administrative_Duration



Quantiles

100.0%	maximum	1.3861795148
99.5%		1.3816797184
97.5%		1.3714406121
90.0%		1.3567182698
75.0%	quartile	1.3397163549
50.0%	median	1.2772705194
25.0%	quartile	-1.213365895
10.0%		-1.213365895
2.5%		-1.213365895
0.5%		-1.213365895
0.0%	minimum	-1.213365895

Summary Statistics

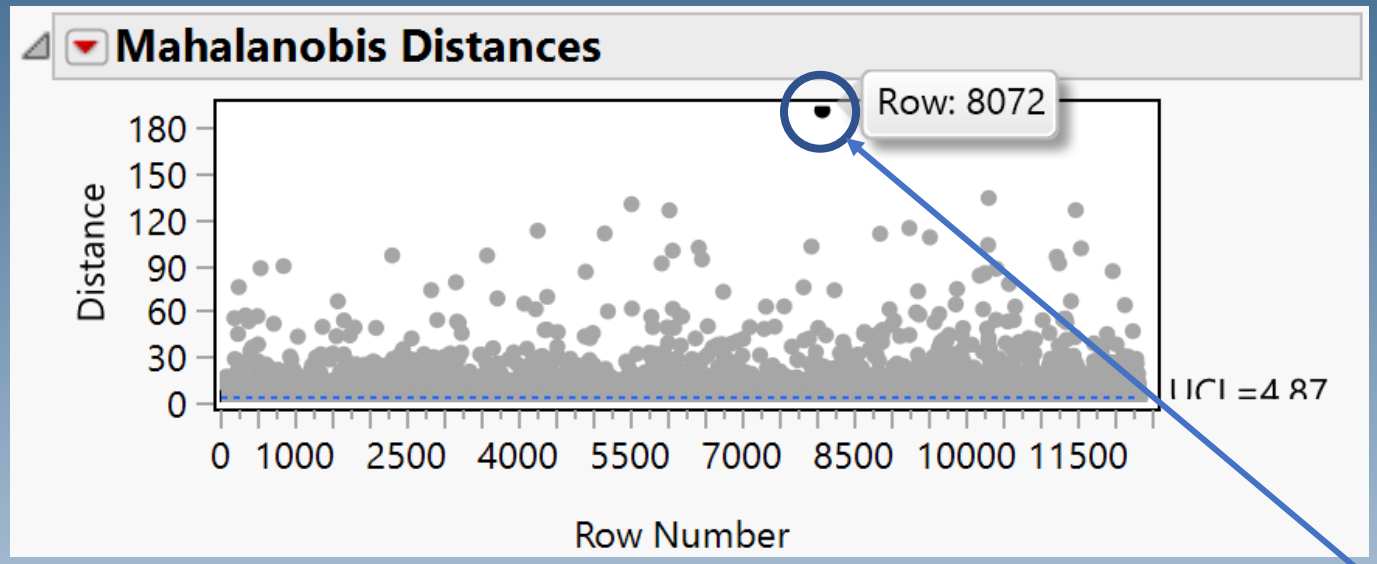
Mean	0.0706187
Std Dev	1.2756521
Std Err Mean	0.0119152
Upper 95% Mean	0.0939746
Lower 95% Mean	0.0472629
N	11462

Action:
Converted the variable into categorical one

Values > 0 are flagged as 1
Values of zeros are left unchanged

Distribution not even

3 Outliers & Missing Values



There is no Missing Value in this dataset.

This row has high duration values

PART 3

Tools, Techniques, Modeling

Tools & Techniques

- PCA
- Balancing Data Set
- Visualization

2 Principle Component Analysis

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13	Prin14
Administrative	0.38460	0.06417	0.03149	0.25426	0.33077	0.13762	0.24955	-0.00535	0.11901	0.01630	0.60277	0.44207	-0.13140	0.03484
Administrative_Duration	0.32128	0.10103	0.03823	0.34528	0.39681	0.19220	0.34674	-0.01461	0.17777	-0.00961	-0.56946	-0.28946	0.10473	0.01781
Informational	0.34675	0.29194	-0.02092	0.18123	-0.44736	-0.00828	0.02799	-0.01061	-0.01160	0.00681	0.38174	-0.64361	-0.00197	-0.00045
Informational_Duration	0.28717	0.30499	-0.01955	0.19262	-0.59435	-0.02353	0.00874	-0.01493	-0.01099	-0.02449	-0.35839	0.54921	0.04589	0.00484
ProductRelated	0.41188	0.15446	0.03572	-0.42593	0.20898	-0.14595	-0.23035	0.04639	-0.09045	-0.01123	0.04088	0.04027	0.68216	0.16331
ProductRelated_Duration	0.41354	0.18417	0.04025	-0.39668	0.18171	-0.15280	-0.23503	0.04188	-0.09253	-0.02509	-0.17677	-0.04319	-0.68288	-0.11868
BounceRates	-0.28224	0.60321	0.01131	0.09648	0.20121	-0.01382	-0.13454	0.09813	0.04376	0.01894	0.05805	0.03510	0.13840	-0.67276
ExitRates	-0.32884	0.57174	-0.00036	0.07129	0.14525	-0.02515	-0.09671	0.06532	0.05258	-0.00227	0.00554	-0.00143	-0.13647	0.70931
PageValues	0.09422	-0.18340	0.12524	0.28271	0.00178	0.32925	-0.53001	0.67969	0.01480	0.09185	-0.01999	-0.00778	-0.00467	0.03805
SpecialDay	-0.07217	0.10657	0.03923	-0.51931	-0.18101	0.27945	0.51667	0.50217	0.27790	0.06519	0.01695	0.00414	-0.01021	-0.01845
OperatingSystems	-0.01393	0.02168	0.59722	-0.05116	-0.05046	0.07813	-0.04859	-0.25986	0.08319	0.74450	-0.01491	0.00229	-0.00401	0.00747
Browser	-0.01769	-0.05227	0.55070	-0.01495	-0.06830	-0.11676	-0.15234	-0.10662	0.58455	-0.54571	0.02928	-0.01020	0.00576	-0.00961
Region	-0.01996	-0.06102	0.30196	0.19727	0.03860	-0.73352	0.29363	0.43097	-0.22799	0.04728	-0.00482	-0.00270	0.00421	0.00459
TrafficType	-0.05200	0.08361	0.47471	-0.04076	-0.00125	0.39209	0.14571	-0.05322	-0.67417	-0.36177	0.01022	0.00042	0.00302	0.00108

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	3.4053	24.324		24.324
2	1.6752	11.966		36.290
3	1.4003	10.002		46.292
4	1.0943	7.817		54.108
5	1.0272	7.337		61.446
6	0.9694	6.924		68.370
7	0.9595	6.854		75.224
8	0.9146	6.533		81.757
9	0.8604	6.146		87.902
10	0.7534	5.382		93.284
11	0.3930	2.807		96.091
12	0.3556	2.540		98.631
13	0.1143	0.816		99.447
14	0.0774	0.553		100.000

3 Balanced Dataset

- We balanced our data set as the negatives were much higher than the positives

KNN Balanced file
Source
Fit Nominal Logistic

Columns (30/1)
RowID
Administrative
Administrative_Duration
Administrat...uration_Cat
Informational
Informational_Duration
Information...uration_Cat
ProductRelated
ProductRelated_Duration
BounceRates
ExitRates
PageValues
SpecialDay
Month
Month_Asc
OperatingSystems
Browser

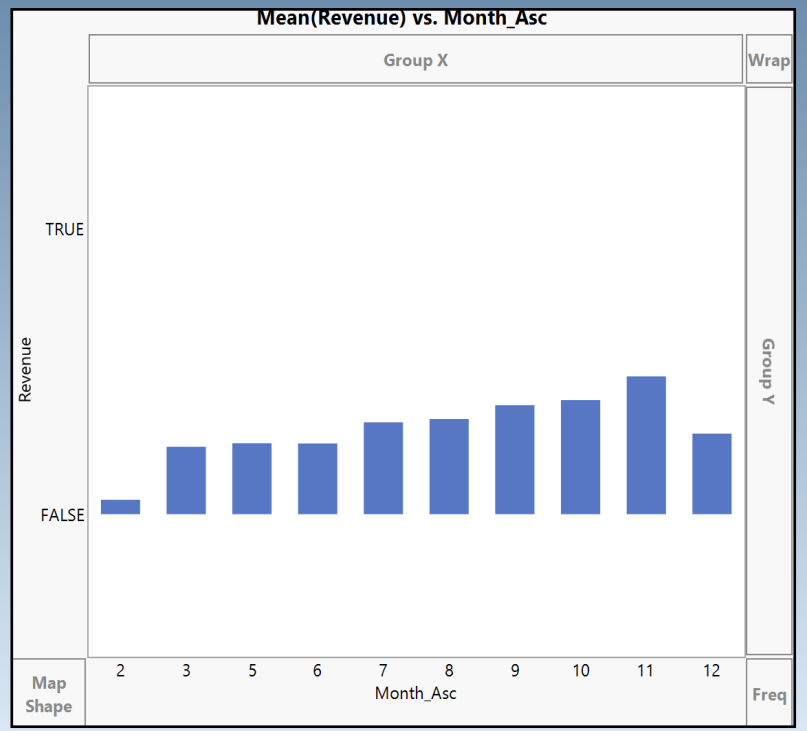
Rows

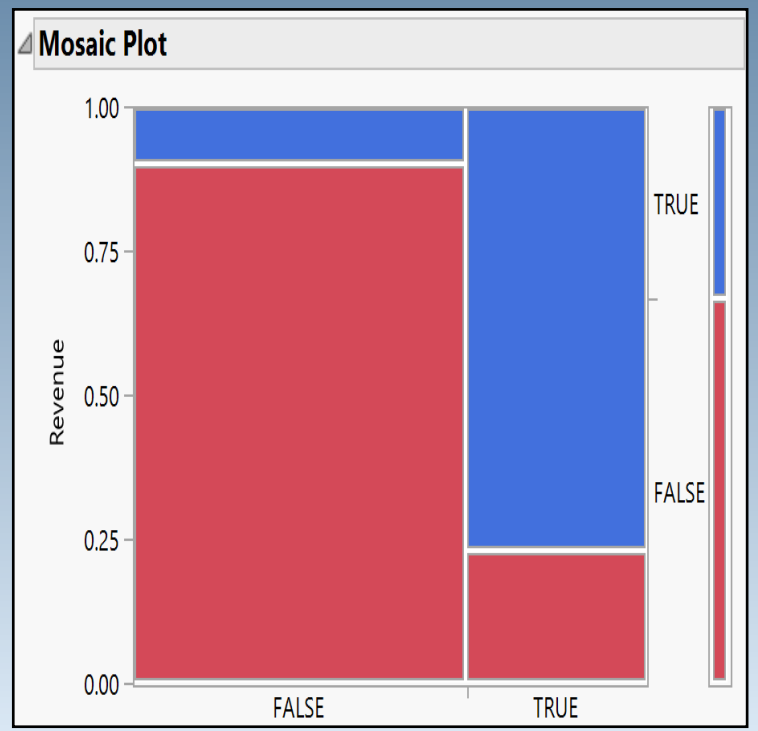
All rows	5,994
Selected	0
Excluded	310
Hidden	0
Labelled	0

	RowID	Administrative	Administrative_Duration	Administrative_Duration_Cat	Informational	Informational_Duration	Informational_Duration_Cat	ProductRelated	ProductRelated_Duration	E
1	3575	0	0	0	9	1779.166667	1	12	1886.5	
2	8662	0	0	0	0	0	0	2	0	
3	5462	7	56.3	1	0	0	0	15	315.4	
4	1152	1	19	1	0	0	0	13	748.6666667	
5	10574	0	0	0	0	0	0	7	242.75	
6	7129	7	234.4333333	1	0	0	0	85	6743.39876	
7	11855	8	447.125	1	3	168.375	1	142	4777.812554	
8	4454	3	105	1	0	0	0	35	999	
9	7029	1	5	1	0	0	0	5	95.1	
10	8892	0	0	0	0	0	0	30	946.5952381	
11	11485	12	244.25	1	0	0	0	110	4006.359314	
12	4616	2	29	1	1	14	1	106	3825.102381	
13	11460	1	11	1	1	10	1	74	3487.12471	
14	6799	2	128.8	1	2	49.3	1	13	700.55	
15	4363	4	45	1	0	0	0	21	819.8333333	
16	252	0	0	0	0	0	0	18	5188.5	
17	6075	8	131	1	0	0	0	32	5330.633333	
18	11231	0	0	0	0	0	0	1	0	
19	6401	1	8	1	0	0	0	47	1459.576667	
20	2472	3	49.33333333	1	2	346.1666667	1	9	194.3333333	
21	4497	2	820	1	0	0	0	28	2228.516667	
22	5135	0	0	0	0	0	0	1	0	
23	10724	2	48.5	1	0	0	0	5	396.1666667	
24	10689	0	0	0	0	0	0	9	470	

4 Visualization

- We utilized visualization techniques like boxplots, mosaic plots, bar graphs etc.





Modeling

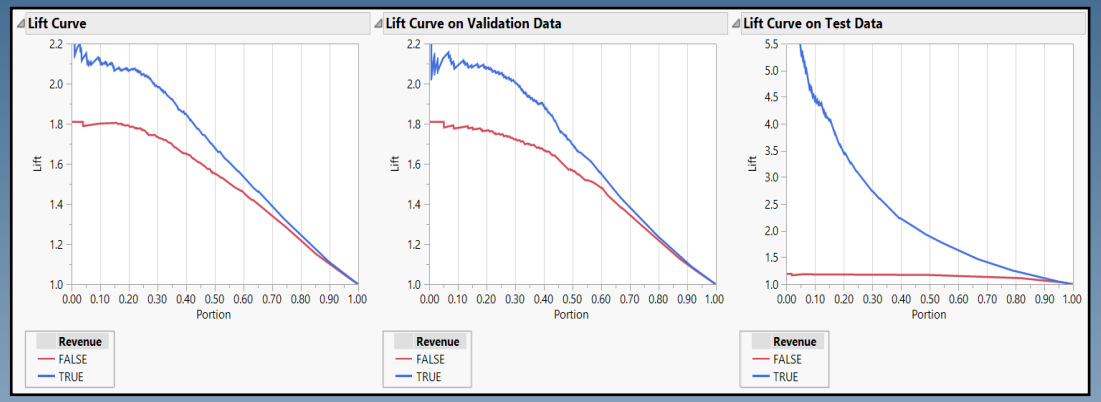
- Logistic Regression
- Decision Tree
- Boosted Tree
- KNN
- Bootstrap Forest
- Neural Network
- Naïve Bayes

Logistic Regression

Nominal Logistic Fit for Revenue

Effect Summary

Source	LogWorth	PValue
PageValues	155.705	0.00000
Month_Asc	18.488	0.00000
ProductRelated_Duration	8.901	0.00000
Johnson SI Transform ExitRates	4.617	0.00002
VisitorType	0.714	0.19327



Confusion Matrix

Training

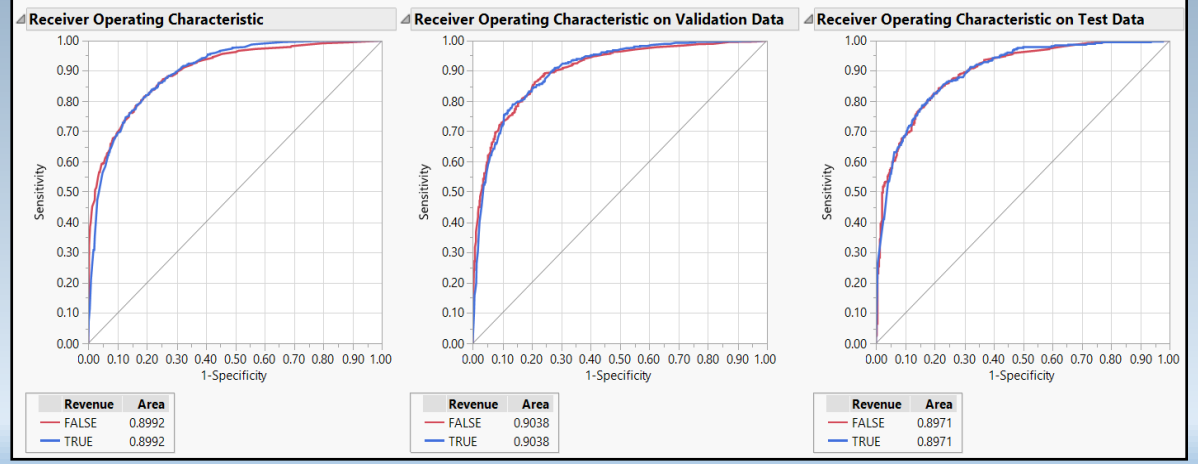
Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1053	118
TRUE	289	660

Validation

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	626	73
TRUE	149	417

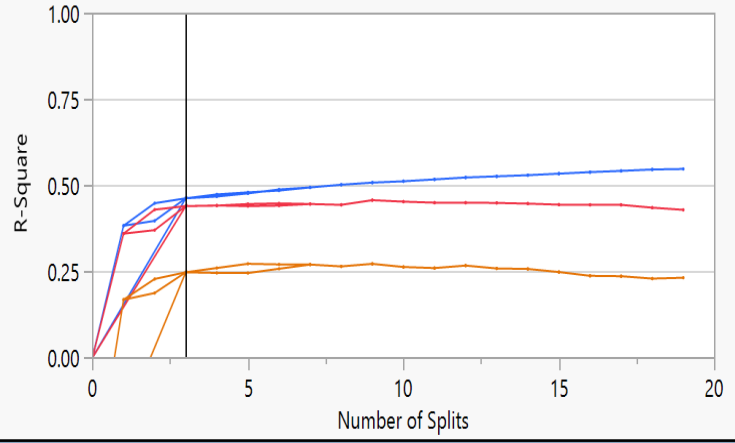
Test

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1754	177
TRUE	123	257



Decision Tree

Split History



Confusion Matrix

Training

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1017	154
TRUE	166	780

Validation

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	613	84
TRUE	118	450

Test

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1725	224
TRUE	73	308

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4627	0.4400	0.2479	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6300	0.6076	0.3361	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3694	0.3852	0.3350	∑ -Log(ρ[j])/n
RMSE	0.3439	0.3521	0.3256	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2370	0.2466	0.2110	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1512	0.1597	0.1275	∑ (ρ[j]≠ρMax)/n
N	2117	1265	2330	n

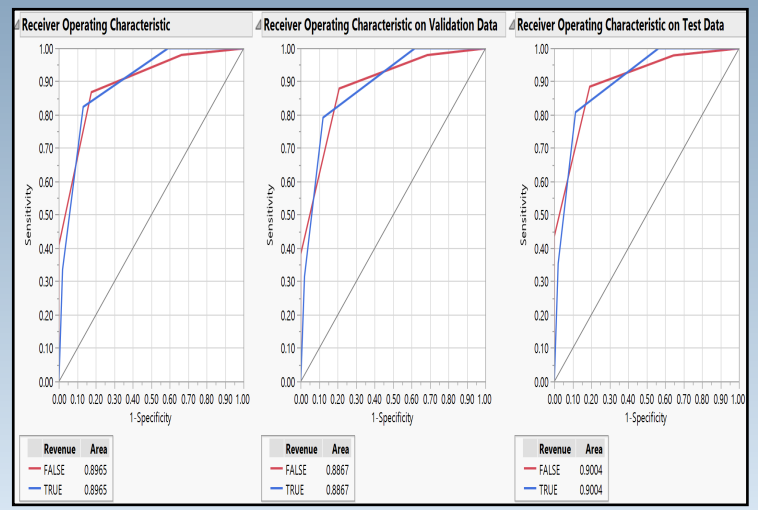
Leaf Report

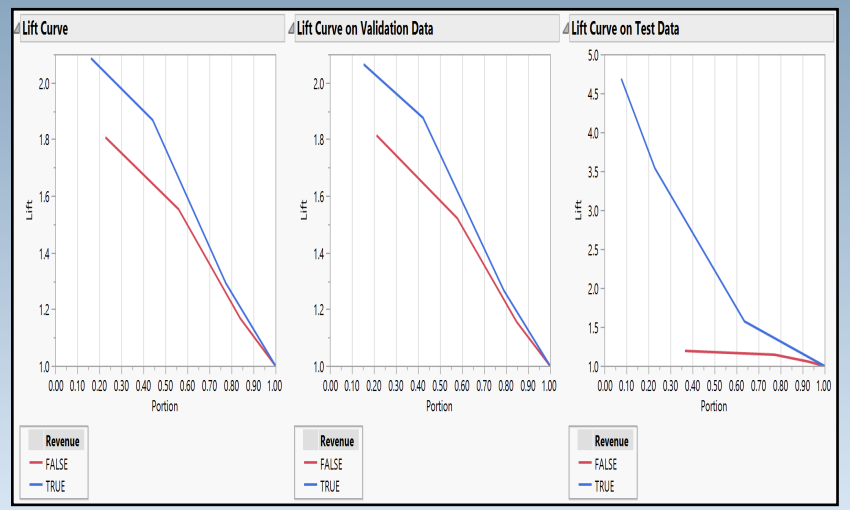
Response Prob

Leaf Label	FALSE	TRUE
PageValues>=0.067049546&Administrative<2	0.0688	0.9312
PageValues>=0.067049546&Administrative>=2	0.2214	0.7786
PageValues<0.067049546&Month_Asc(2, 3, 5)	0.9991	0.0009
PageValues<0.067049546&Month_Asc(6, 7, 8, 9, 10, 11, 12)	0.7636	0.2364

Response Counts

Leaf Label	FALSE	TRUE
PageValues>=0.067049546&Administrative<2	23	318
PageValues>=0.067049546&Administrative>=2	131	462
PageValues<0.067049546&Month_Asc(2, 3, 5)	480	0
PageValues<0.067049546&Month_Asc(6, 7, 8, 9, 10, 11, 12)	537	166





Boosted Tree

Column Contributions

Term	Number of Splits	G^2	Portion
Month_Asc	27	55330.8551	0.2555
PageValues	31	44139.4633	0.2038
Informational_Duration_Cat	14	19586.859	0.0905
VisitorType	8	17892.2808	0.0826
Browser	6	16849.5069	0.0778
OperatingSystems	6	15477.406	0.0715
Administrative_Duration_Cat	17	14059.6631	0.0649
Johnson SI Transform ExitRates	14	11934.5786	0.0551
TrafficType	4	11926.7574	0.0551
Weekend	17	3888.16624	0.0180
Administrative	4	3436.37582	0.0159
Region	2	2020.86031	0.0093

Confusion Matrix

Training

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1037	122
TRUE	162	788

Validation

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	593	88
TRUE	116	452

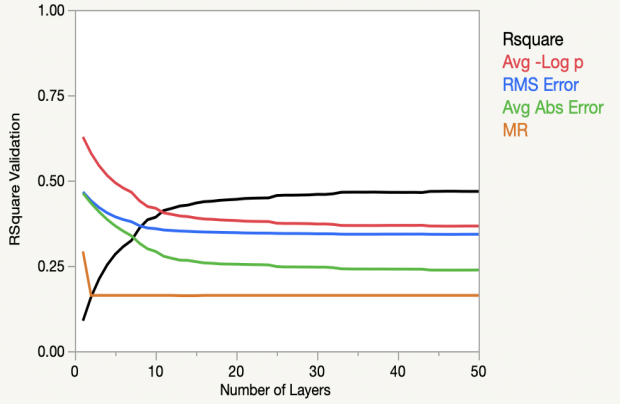
Test

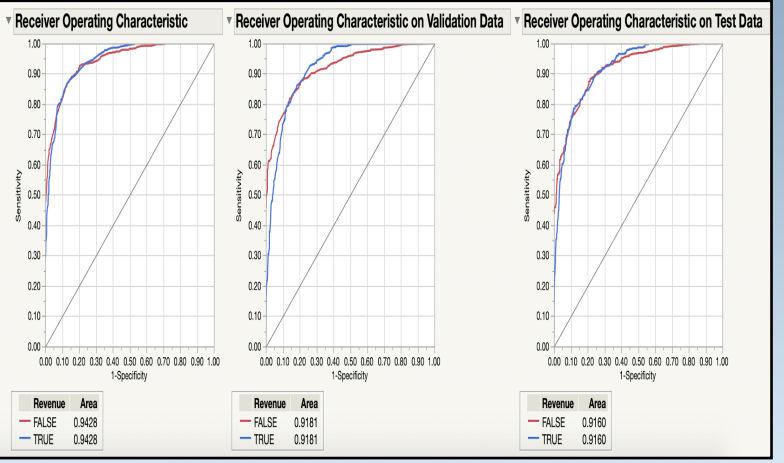
Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1676	242
TRUE	79	298

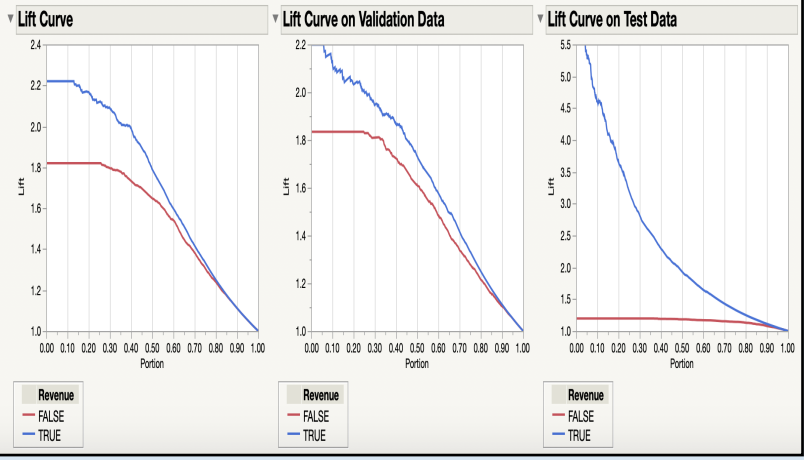
Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.5365	0.4681	0.2280	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6985	0.6356	0.3119	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3190	0.3665	0.3449	∑ -Log(ρ[j])/n
RMSE	0.3135	0.3425	0.3308	√∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2190	0.2378	0.2222	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1347	0.1633	0.1399	∑ (ρ[j]≠ρMax)/n
N	2109	1249	2295	n

Cumulative Validation





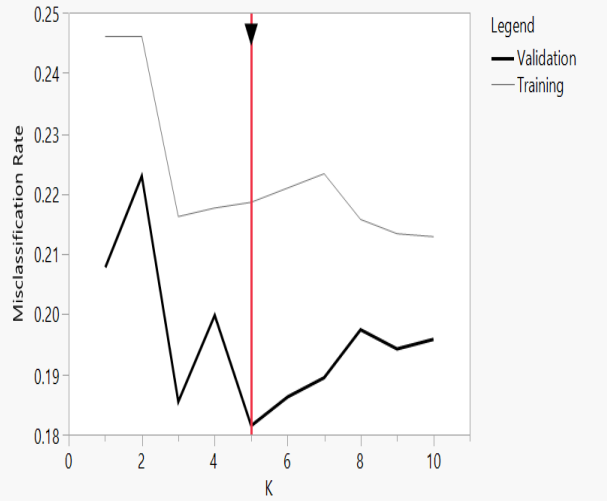


KNN

K Nearest Neighbors

Revenue

Model Selection



Training

K	Count	Misclassification Rate	Misclassifications
1	2109	0.24609	519
2	2109	0.24609	519
3	2109	0.21622	456
4	2109	0.21764	459
5	2109	0.21859	461
6	2109	0.22096	466
7	2109	0.22333	471
8	2109	0.21574	455
9	2109	0.21337	450
10	2109	0.21290	449 *

Validation

K	Count	Misclassification Rate	Misclassifications
1	1256	0.20780	261
2	1256	0.22293	280
3	1256	0.18551	233
4	1256	0.19984	251
5	1256	0.18153	228 *
6	1256	0.18631	234
7	1256	0.18949	238
8	1256	0.19745	248
9	1256	0.19427	244
10	1256	0.19586	246

Test

K	Count	Misclassification Rate	Misclassifications
1	2319	0.22682	526
2	2319	0.23372	542
3	2319	0.18758	435
4	2319	0.18844	437
5	2319	0.16947	393
6	2319	0.17335	402
7	2319	0.16214	376
8	2319	0.16645	386
9	2319	0.15179	352 *
10	2319	0.16171	375

Confusion Matrix for Best K=5

Training

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	962	199
TRUE	262	686

Validation

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	591	98
TRUE	130	437

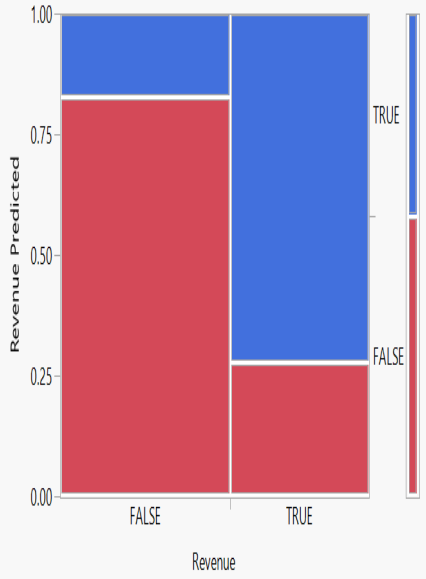
Test

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1643	296
TRUE	97	283

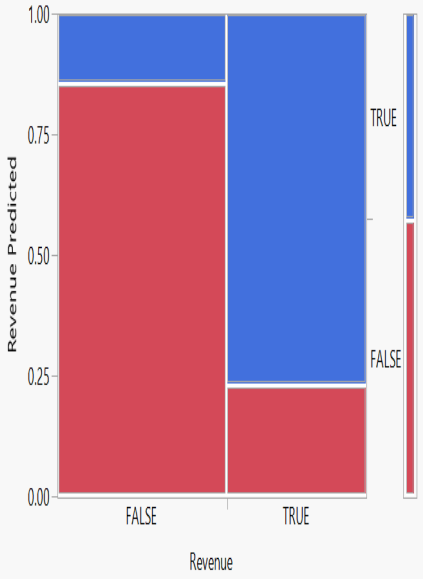
Mosaic Plot

Mosaic Plot for K=5

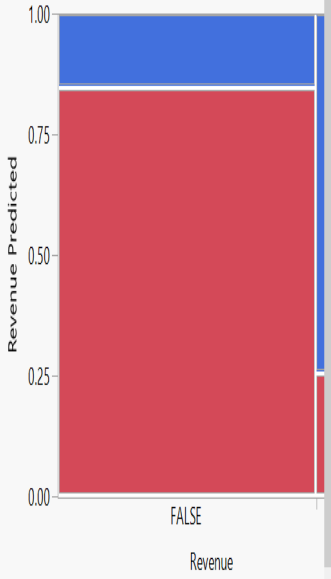
Training



Validation



Test



Bootstrap Forest

Bootstrap Forest for Revenue

Specifications

Target Column:	Revenue	Training Rows:	2120
Validation Column:	Validation Stratified by Revenue	Validation Rows:	1265
		Test Rows:	2311
Number of Trees in the Forest:	21	Number of Terms:	15
Number of Terms Sampled per Split:	3	Bootstrap Samples:	2120
		Minimum Splits per Tree:	3
		Minimum Size Split:	5

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.5137	0.4196	0.1576	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6780	0.5868	0.2223	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3344	0.3991	0.3765	∑ -Log(ρ[j])/n
RMSE	0.3075	0.3471	0.3367	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2614	0.2966	0.2779	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1208	0.1431	0.1285	∑ (ρ[j]≠ρMax)/n
N	2120	1265	2311	n

Confusion Matrix

Training

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1059	112
TRUE	144	805

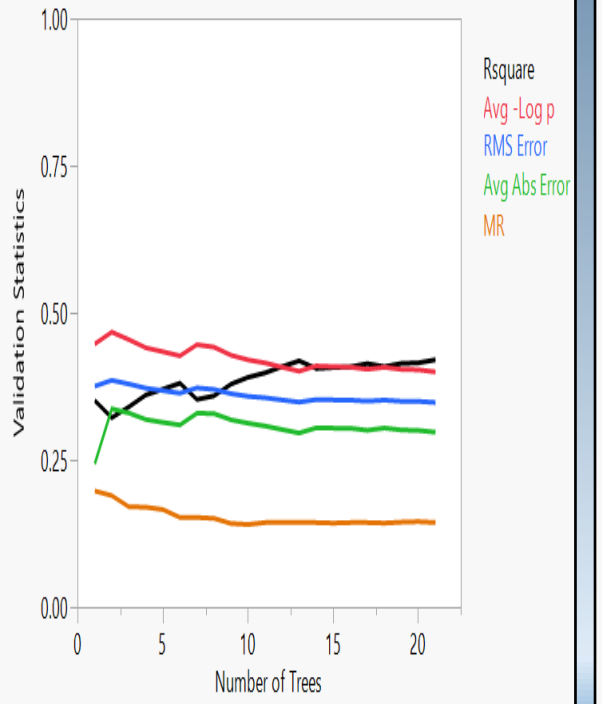
Validation

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	617	82
TRUE	99	467

Test

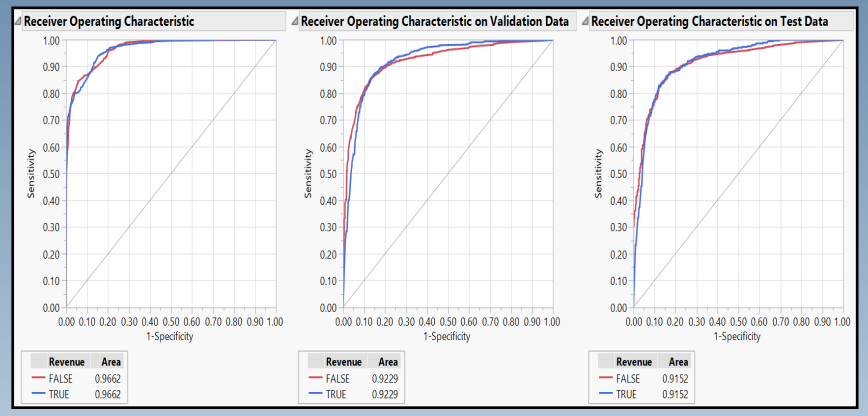
Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1706	225
TRUE	72	308

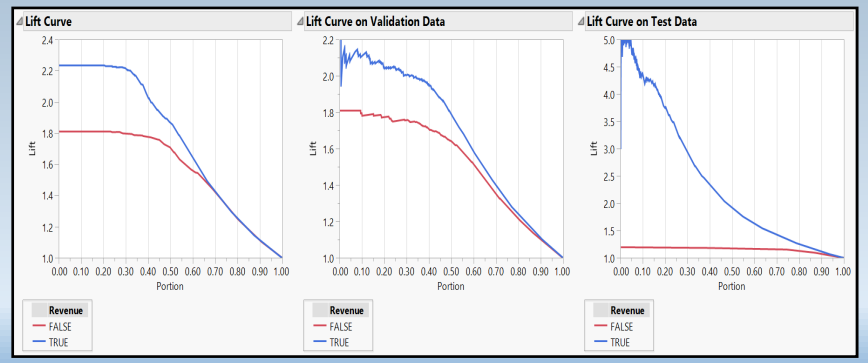
Cumulative Validation



Column Contributions

Term	Number of Splits	G^2	Portion
PageValues	210	458.154875	0.4802
Johnson Sl Transform ProductRelated_Duration	263	110.917306	0.1163
Johnson Sl Transform ExitRates	255	100.316693	0.1052
Month_Asc	161	77.3096502	0.0810
Administrative	218	52.8165003	0.0554
TrafficType	175	28.0287182	0.0294
Region	174	25.5572362	0.0268
VisitorType	90	20.340243	0.0213
Informational	114	17.141565	0.0180
Browser	129	13.2587902	0.0139
OperatingSystems	130	13.1717268	0.0138
SpecialDay	39	11.0087419	0.0115
Administrative_Duration_Cat	64	8.87650452	0.0093
Weekend	122	8.75507417	0.0092
Informational_Duration_Cat	75	8.35449596	0.0088





Neural Network

Cleaned-Validation_TRUE_PROP_0.5 - Neural of Revenue 5 - JMP Pro [2]

Neural

Validation Column: Validation Stratified by Revenue

Model Launch

Model NTanH(1)NLinear(1)NGaussian(1)

Training

Revenue

Measures	Value
Generalized RSquare	0.6817742
Entropy RSquare	0.517689
RMSE	0.3257098
Mean Abs Dev	0.212383
Misclassification Rate	0.1551642
-LogLikelihood	697.5592
Sum Freq	2101

Confusion Matrix

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1017	136
TRUE	190	758

Confusion Rates

Actual Revenue	Predicted Rate FALSE	Predicted Rate TRUE
FALSE	0.882	0.118
TRUE	0.200	0.800

Validation

Revenue

Measures	Value
Generalized RSquare	0.6498221
Entropy RSquare	0.4834701
RMSE	0.3368214
Mean Abs Dev	0.2173725
Misclassification Rate	0.1597168
-LogLikelihood	451.00962
Sum Freq	1271

Confusion Matrix

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	615	91
TRUE	112	453

Confusion Rates

Actual Revenue	Predicted Rate FALSE	Predicted Rate TRUE
FALSE	0.871	0.129
TRUE	0.198	0.802

Test

Revenue

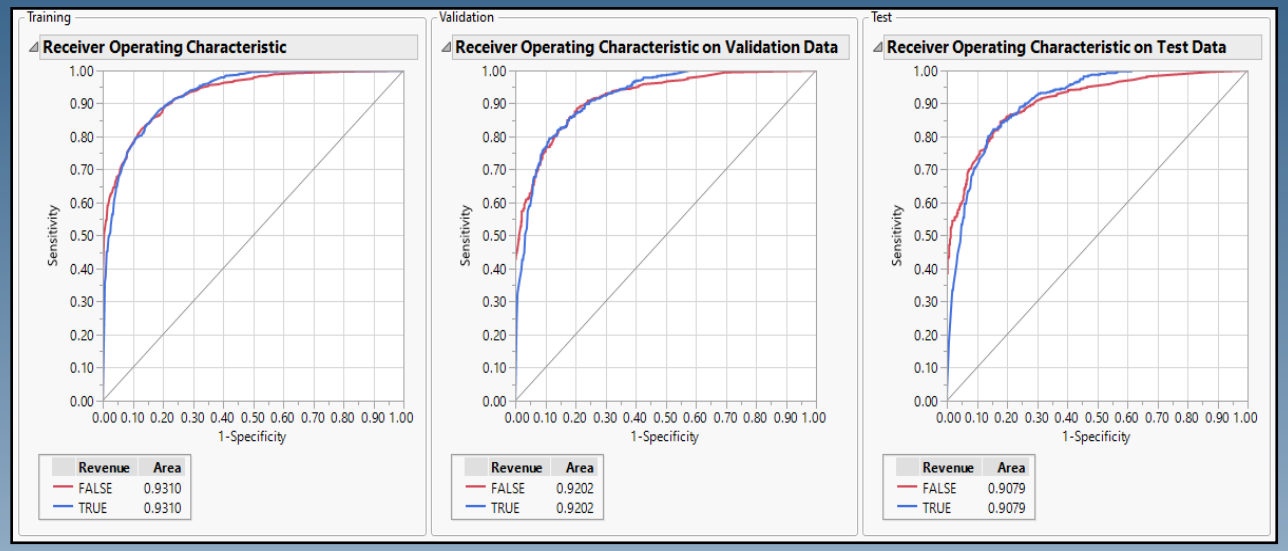
Measures	Value
Generalized RSquare	0.2158702
Entropy RSquare	0.152533
RMSE	0.344734
Mean Abs Dev	0.223424
Misclassification Rate	0.1547773
-LogLikelihood	878.3961
Sum Freq	2313

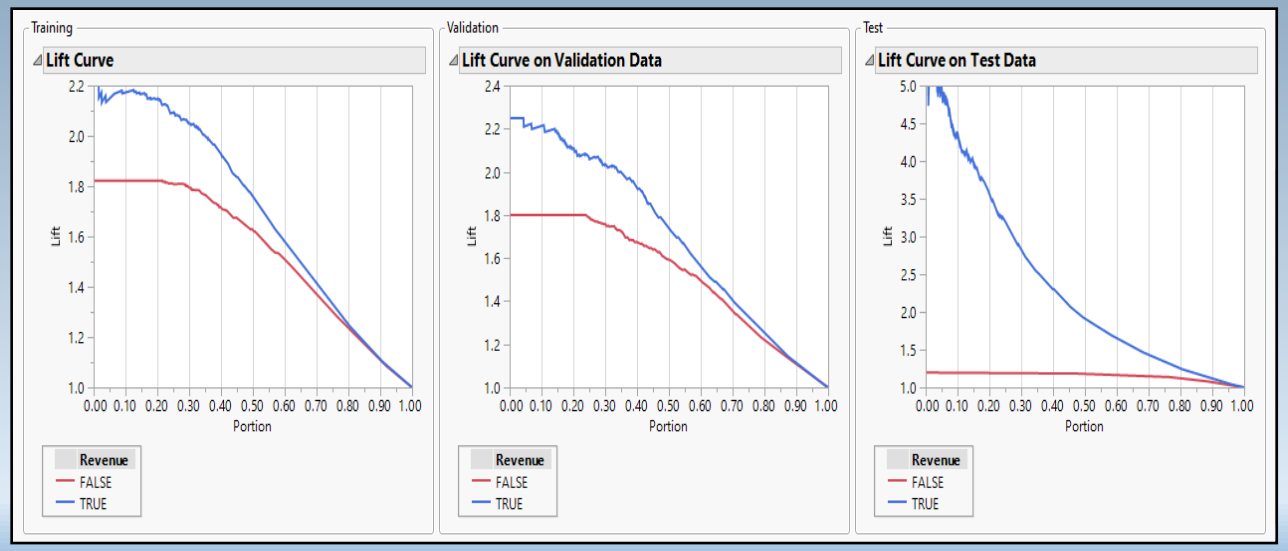
Confusion Matrix

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1647	284
TRUE	74	308

Confusion Rates

Actual Revenue	Predicted Rate FALSE	Predicted Rate TRUE
FALSE	0.853	0.147
TRUE	0.194	0.806





Naïve Bayes

Confusion Matrix

Training

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	948	223
TRUE	285	664

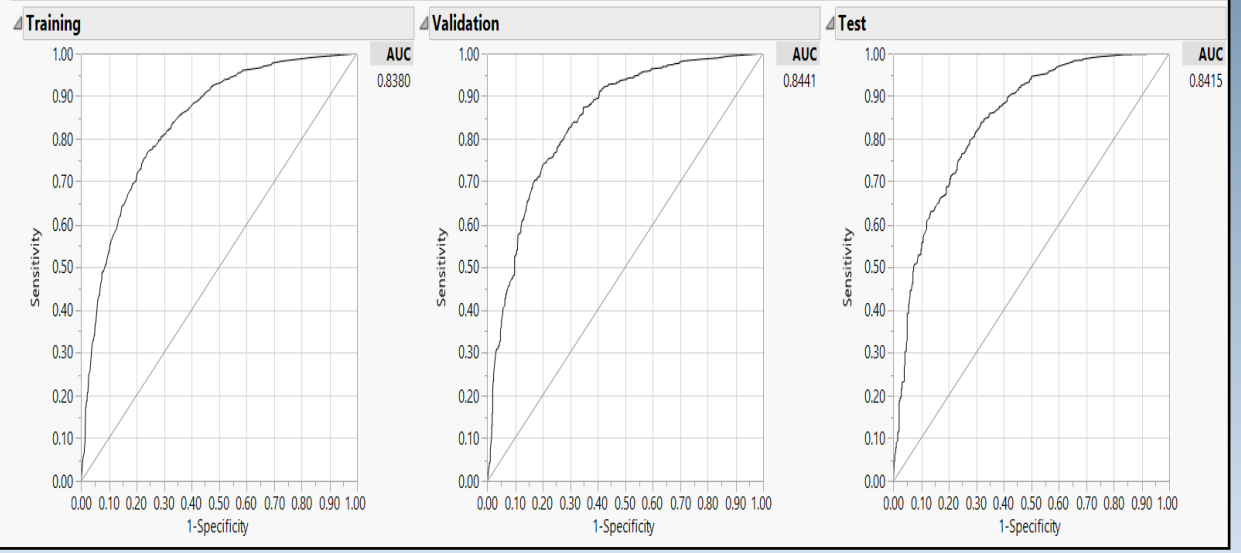
Validation

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	563	136
TRUE	158	408

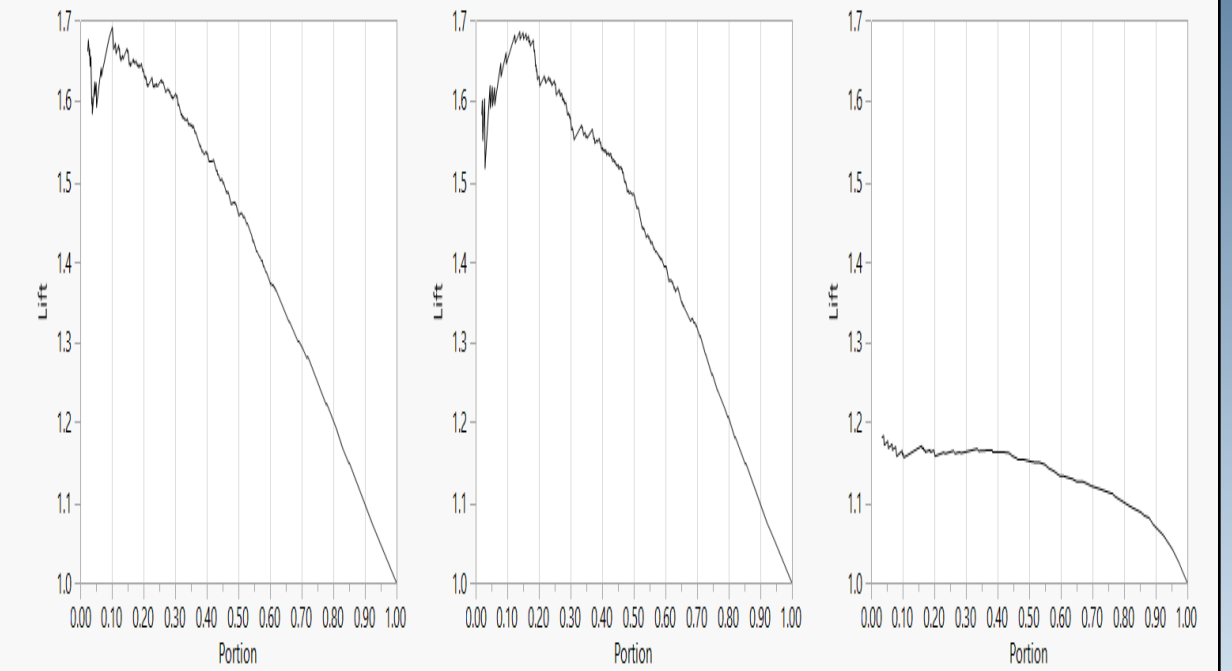
Test

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1561	370
TRUE	112	268

ROC Curve for Revenue = FALSE



Lift Curve for Revenue = FALSE

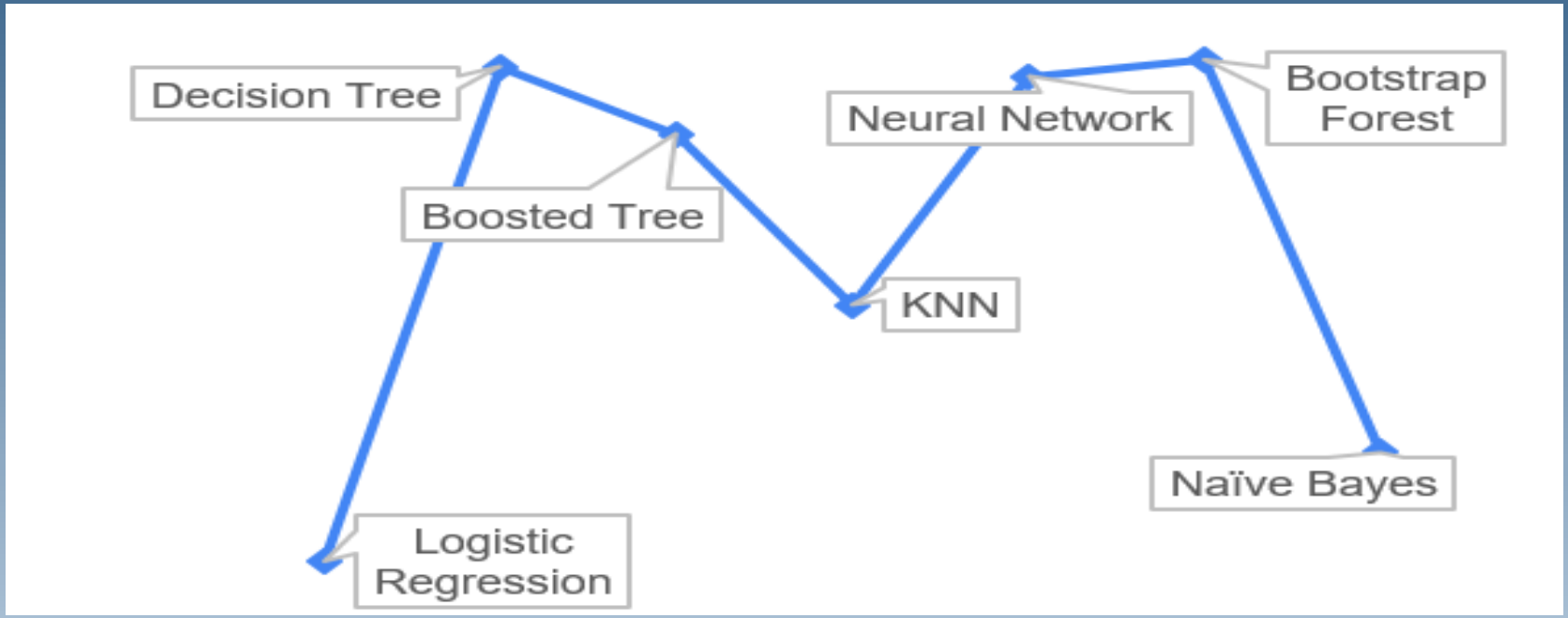


PART 4

Model Comparison

- Compare Accuracy
- Compare Complexity

Model Comparison



Potential Customers

We are focusing on identifying customers who are likely to purchase.

Model Comparison

Test data								
	Confusion Matrix				Accuracy	Accuracy of 1	Sensitivity	Misclassification
	TP	TF	FN	FP				
Logistic	257	1754	123	177	87.02%	59.22%	67.63%	12.98%
Decision Tree	308	1725	73	224	87.25%	57.89%	80.84%	12.75%
Boosted Tree	298	1676	79	242	86.01%	55.19%	79.04%	13.99%
KNN	283	1643	97	296	83.05%	48.88%	74.47%	16.95%
Neural Network	308	1647	74	284	84.52%	52.07%	80.60%	16.72%
Bootstrap Forest	308	1706	72	225	87.15%	57.79%	81.05%	12.85%
Naive Bayes	268	1561	112	370	79.14%	42.01%	70.53%	20.86%

Model Comparison

TP / (TP + FN)

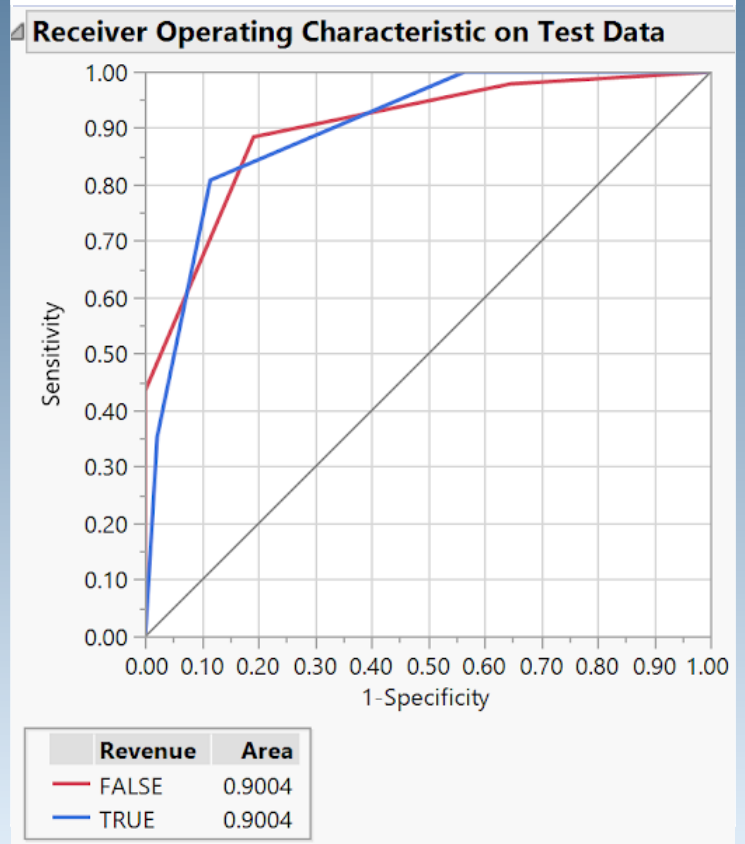
Test data								
	Confusion Matrix				Accuracy	Accuracy of 1	Sensitivity	Misclassification
	TP	TF	FN	FP				
Logistic	257	1754	123	177	87.02%	59.22%	67.63%	12.98%
Decision Tree	308	1725	73	224	87.25%	57.89%	80.84%	12.75%
Boosted Tree	298	1676	79	242	86.01%	55.19%	79.04%	13.99%
KNN	283	1643	97	296	83.05%	48.88%	74.47%	16.95%
Neural Network	308	1647	74	284	84.52%	52.07%	80.60%	16.72%
Bootstrap Forest	308	1706	72	225	87.15%	57.79%	81.05%	12.85%
Naive Bayes	268	1561	112	370	79.14%	42.01%	70.53%	20.86%

Model Comparison

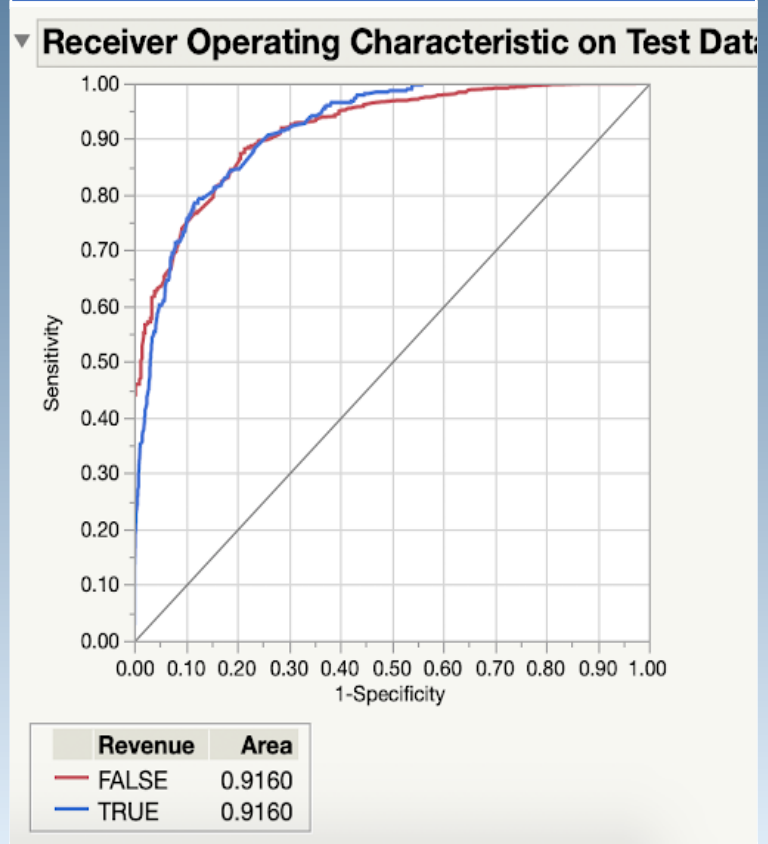
Test data								
	Confusion Matrix				Accuracy	Accuracy of 1	Sensitivity	Misclassification
	TP	TF	FN	FP				
Logistic	257	1754	123	177	87.02%	59.22%	67.63%	12.98%
Decision Tree	308	1725	73	224	87.25%	57.89%	80.84%	12.75%
Boosted Tree	298	1676	79	242	86.01%	55.19%	79.04%	13.99%
KNN	283	1643	97	296	83.05%	48.88%	74.47%	16.95%
Neural Network	308	1647	74	284	84.52%	52.07%	80.60%	16.72%
Bootstrap Forest	308	1706	72	225	87.15%	57.79%	81.05%	12.85%
Naive Bayes	268	1561	112	370	79.14%	42.01%	70.53%	20.86%

Area Under the Curve (AUC)

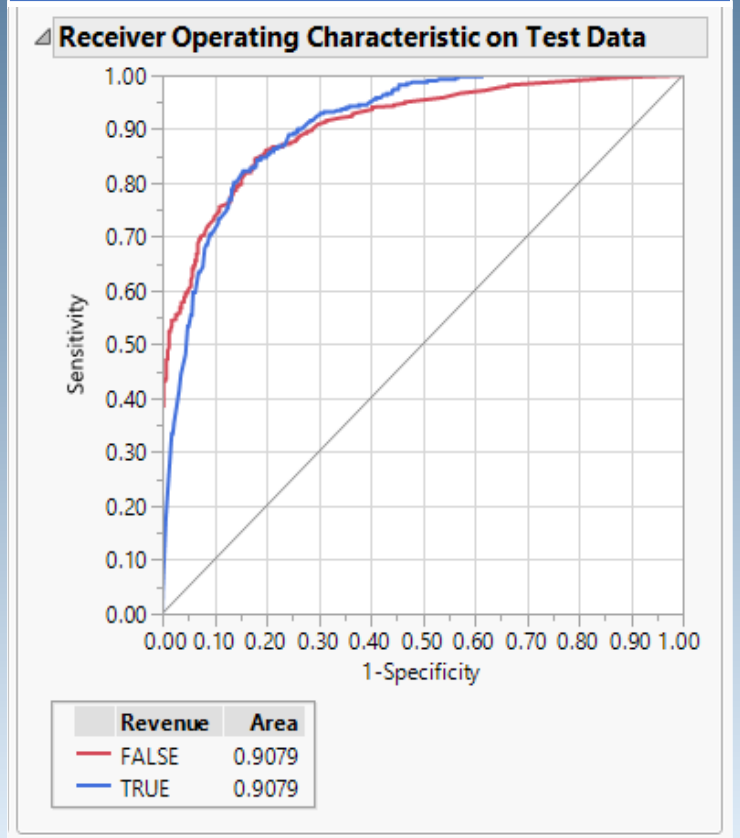
Decision Tree



Bootstrap Forest



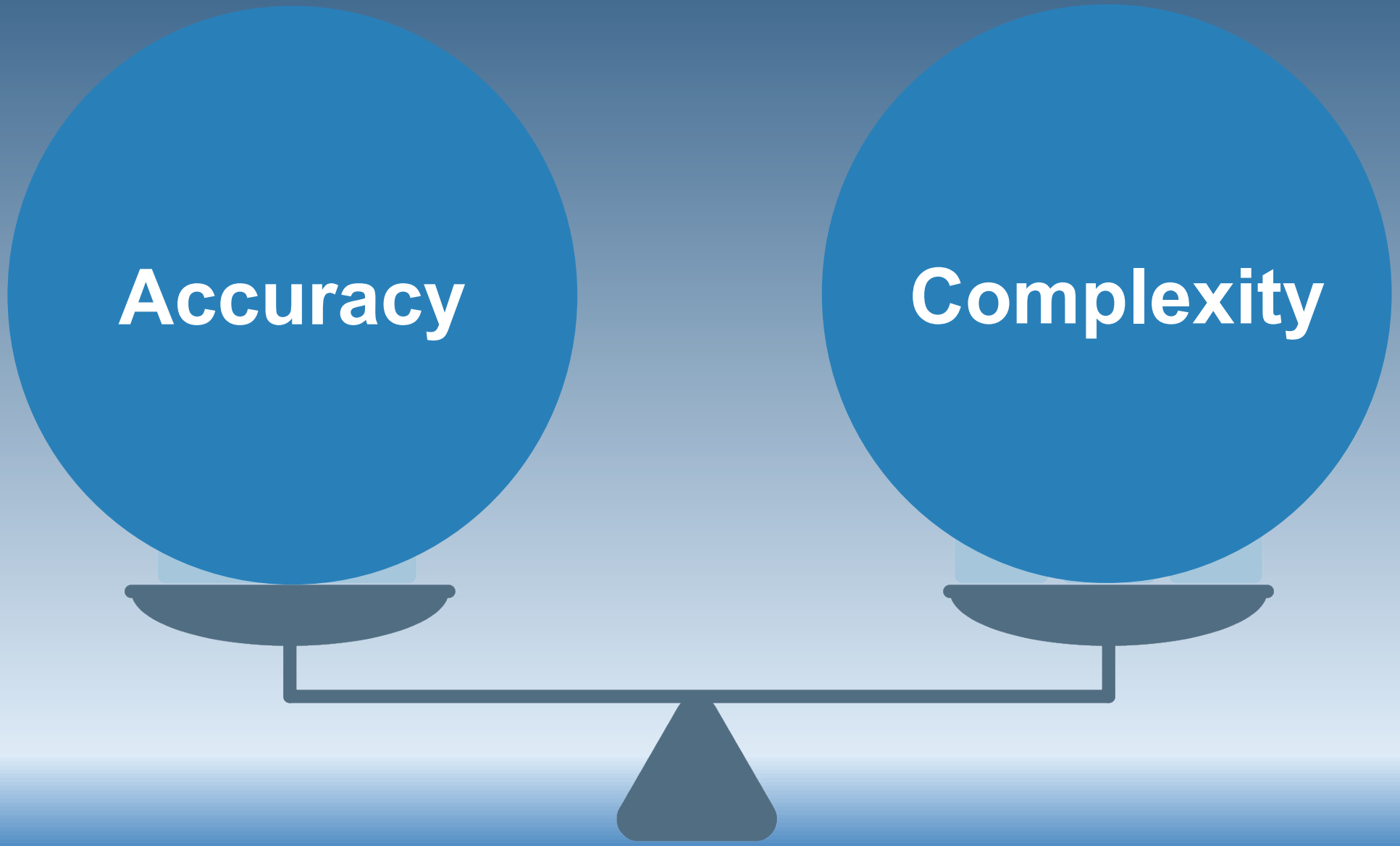
Neural Network



Model Comparison

Test data								
	Confusion Matrix				Accuracy	Accuracy of 1	Sensitivity	Misclassification
	TP	TF	FN	FP				
Logistic	257	1754	123	177	87.02%	59.22%	67.63%	12.98%
Decision Tree	308	1725	73	224	87.25%	57.89%	**80.84%**	12.75%
Boosted Tree	298	1676	79	242	86.01%	55.19%	79.04%	13.99%
KNN	283	1643	97	296	83.05%	48.88%	74.47%	16.95%
Neural Network	308	1647	74	284	84.52%	52.07%	**80.60%**	16.72%
Bootstrap Forest	308	1706	72	225	87.15%	57.79%	**81.05%**	12.85%
Naive Bayes	268	1561	112	370	79.14%	42.01%	70.53%	20.86%

Trade-off



Model Comparison

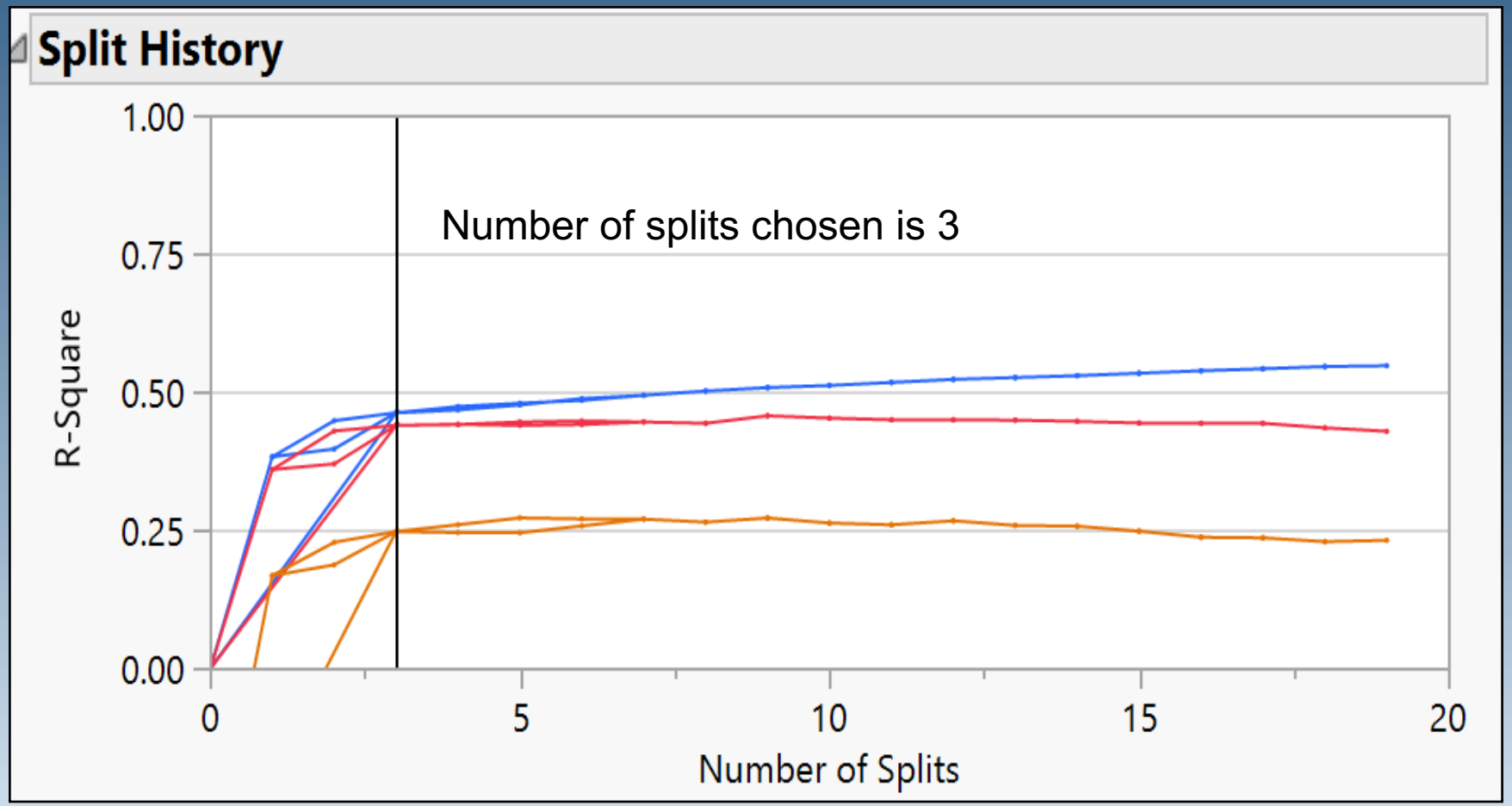
Test data								
	Confusion Matrix				Accuracy	Accuracy of 1	Sensitivity	Misclassification
	TP	TF	FN	FP				
Logistic	257	1754	123	177	87.02%	59.22%	67.63%	12.98%
Decision Tree	308	1725	73	224	87.25%	57.89%	**80.84%**	12.75%
Boosted Tree	298	1676	79	242	86.01%	55.19%	79.04%	13.99%
KNN	283	1643	97	296	83.05%	48.88%	74.47%	16.95%
Neural Network	308	1647	74	284	84.52%	52.07%	80.60%	16.72%
Bootstrap Forest	308	1706	72	225	87.15%	57.79%	**81.05%**	12.85%
Naive Bayes	268	1561	112	370	79.14%	42.01%	70.53%	20.86%

PART 5

Best Model

- Decision Tree
- Significant Variables

Decision Tree - Splits



Decision Tree – Performance

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4627	0.4400	0.2479	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6300	0.6076	0.3361	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3694	0.3852	0.3350	∑ -Log(ρ[j])/n
RMSE	0.3439	0.3521	0.3256	√ ∑(y[j]-ρ[j])^2/n
Mean Abs Dev	0.2370	0.2466	0.2110	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1512	0.1597	0.1275	∑ (ρ[j]≠ρMax)/n
N	2117	1265	2330	

Stable model

Drop due to change in event rate

Decision Tree – Confusion matrix

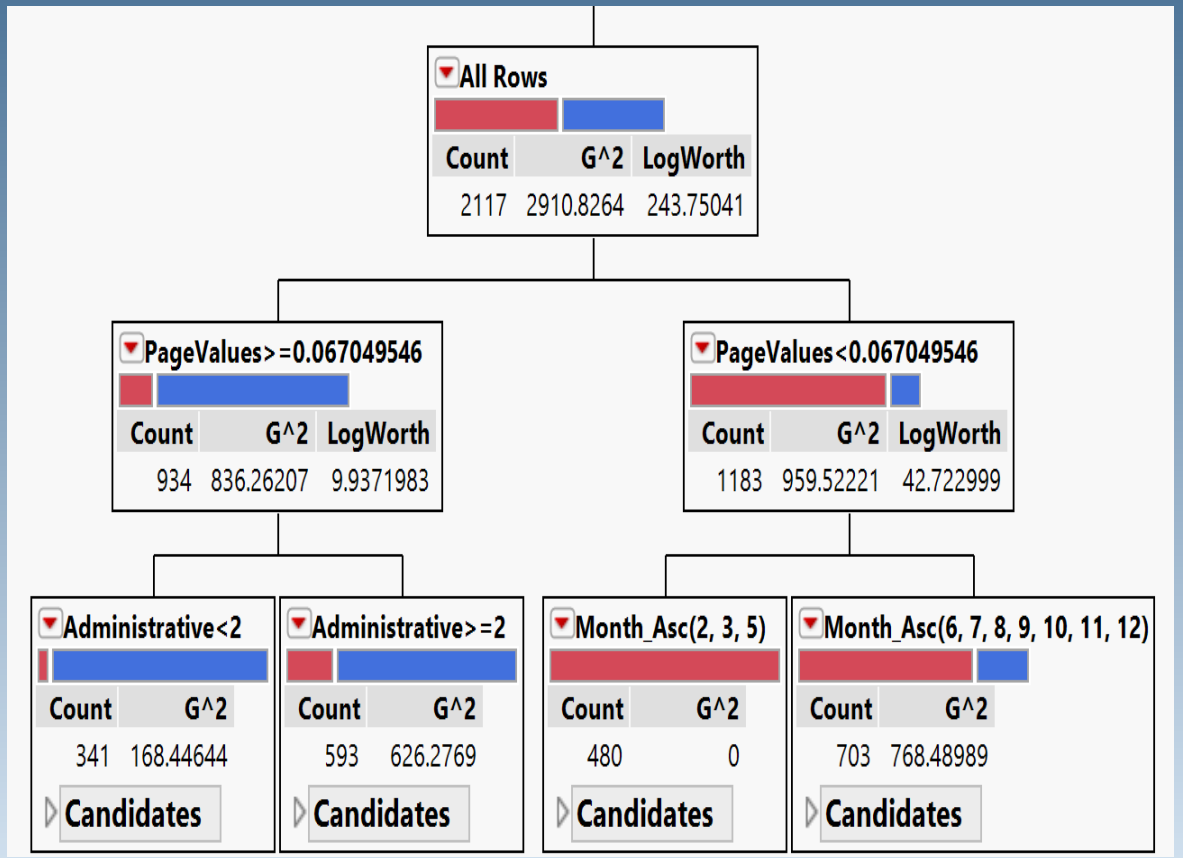
Test

Actual	Predicted Count	
Revenue	FALSE	TRUE
FALSE	1725	224
TRUE	73	308

Key metrics from confusion matrix:

- Percentage of 1 predictions = 22.8%
- Accuracy of 1 predictions = 57.8%
- Lift = 57.8/16.3 = 3.54
- Percentage of actual ones predicted right (sensitivity) = 80.84%

Decision Tree – splits & important variables



Column Contributions

Term	Number of Splits	G^2	Portion
PageValues	1	1115.0421	0.8274
Month_Asc	1	191.032322	0.1418
Administrative	1	41.5387351	0.0308
Administrative_Duration_Cat	0	0	0.0000
Informational	0	0	0.0000
Informational_Duration_Cat	0	0	0.0000
SpecialDay	0	0	0.0000
OperatingSystems	0	0	0.0000
Browser	0	0	0.0000
Region	0	0	0.0000
TrafficType	0	0	0.0000
VisitorType	0	0	0.0000
Weekend	0	0	0.0000
Johnson SI Transform ProductRelated_Duration	0	0	0.0000
Johnson SI Transform ExitRates	0	0	0.0000

PART 6

Business Finding & Recommendations

Business Finding & Recommendations

- Insights based on the below three impact variables:
  - Administrative Page visits – Reduce
  - Months – optimize marketing budgets
  - Page value – optimize flow

Q&A

Thank you for listening

References

- Annelieke's Analytics: What is Page Value in Google Analytics?
- https://yoast.com/what-is-page-value-in-google-analytics/

- Online Shoppers Purchasing Intention Dataset Data Set
  http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#

- Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018).Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks
- https://www.searchfit.com/2017/05/01/factors-influencing-online-shopping-behavior-consumers/

- Shmueli, G., Bruce, P.C., Stephens, M.L. et al. Online Shoppers Purchasing Intention Dataset Data Set. ISBN 978-1-118-87743-2

- Why it is important to work with a balanced classification dataset
- http://amsantac.co/blog/en/2016/09/20/balanced-image-classification-r.html

- 2019 T Mall "Double Eleven" turnover reached 268.4 billion yuan
- http:// xinhuanet.com/2019-11/12/c_1125219252.htm

Appendix

Model 1: Logistic regression.

Nominal Logistic Fit for Revenue

Effect Summary

Source	LogWorth	PValue
PageValues	216.701	0.00000
Johnson SI Transform ProductRelated_Duration	14.750	0.00000
VisitorType	4.834	0.00001
SpecialDay	3.332	0.00047
Johnson SI Transform ExitRates	2.663	0.00217
Browser	1.278	0.05272
Region	1.036	0.09201
Administrative_Duration_Cat	1.033	0.09275
Administrative	0.788	0.16279
Informational_Duration_Cat	0.489	0.32400
Weekend	0.414	0.38508
TrafficType	0.148	0.71099
OperatingSystems	0.119	0.76055
Informational	0.045	0.90223

Confusion Matrix

Training

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	4701	118
TRUE	617	364

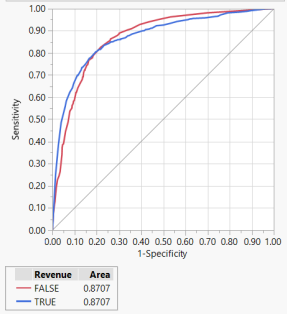
Validation

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	2827	56
TRUE	354	226

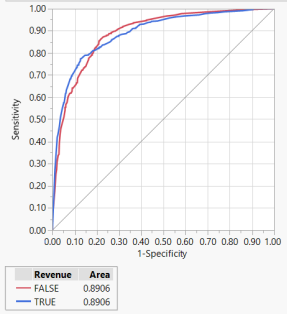
Test

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1928	49
TRUE	224	110

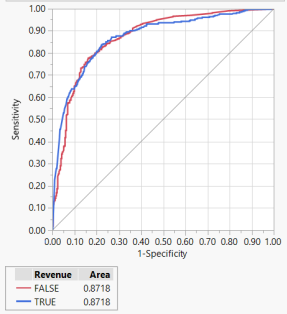
Receiver Operating Characteristic



Receiver Operating Characteristic on Validation Data



Receiver Operating Characteristic on Test Data



Because the Browser, Region, Administrative_Duration_Cat, Administrative, Informational_Duration_Cat, Weekend, Traffic Type, Operating Systems, and Informational have P Value larger than 0.05, they are not statistically significant, so we removed these variables and rerun the model.

Nominal Logistic Fit for Revenue

Effect Summary

Source	LogWorth	PValue
PageValues	218.206	0.00000
Johnson SI Transform ProductRelated_Duration	18.938	0.00000
VisitorType	4.792	0.00002
SpecialDay	3.351	0.00045
Johnson SI Transform ExitRates	3.131	0.00074

Confusion Matrix

Training

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	4703	116
TRUE	615	366

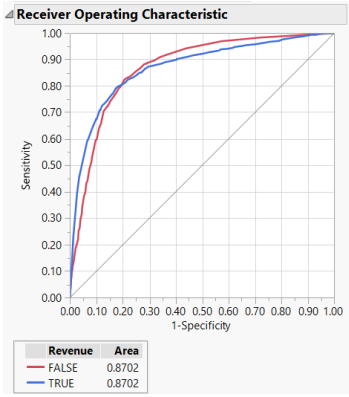
Validation

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	2826	57
TRUE	347	233

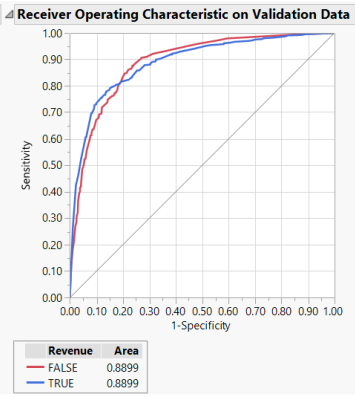
Test

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1927	50
TRUE	222	112

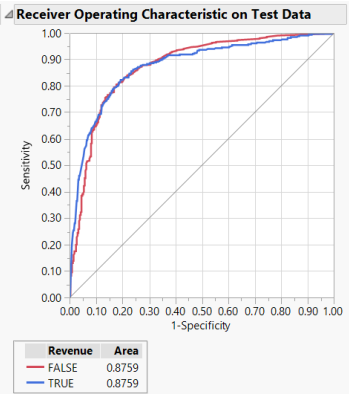
Receiver Operating Characteristic



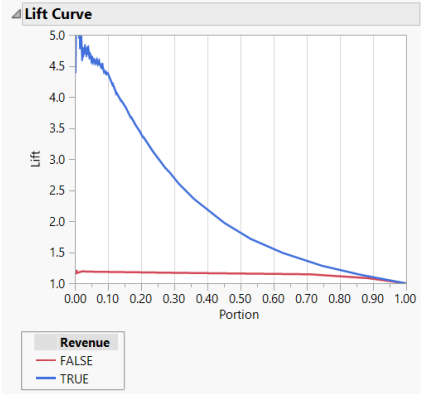
Receiver Operating Characteristic on Validation Data



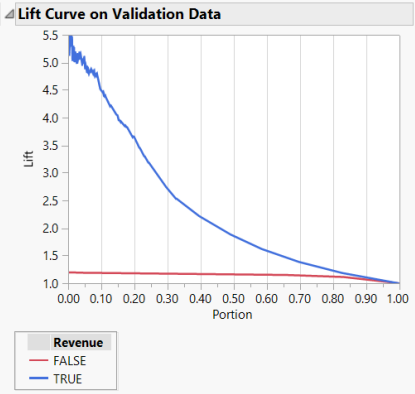
Receiver Operating Characteristic on Test Data



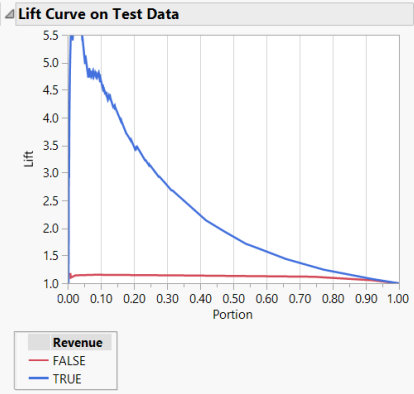
Lift Curve



Lift Curve on Validation Data

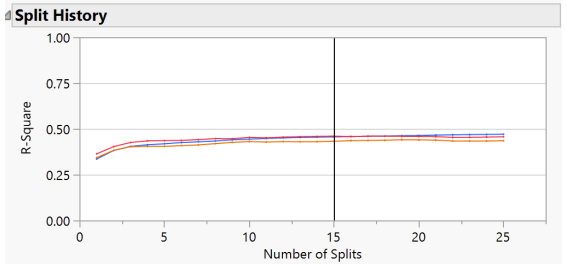


Lift Curve on Test Data



Appendix

Model 2: Decision Tree



By pruning the model, we can observe the performances of each choice of splits. In this case, FN tends to have a higher cost than FP, so the splits with lower FN should be chosen. Based on the table below, with the same level of errors, the one with the lowest complexity should be chosen, hence, the split 5 is finally chosen as the optimal.

Splits	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
FP	141	141	141	141	280	280	280	280	243	243	243	103	103	349	349
FN	250	250	250	250	128	128	128	128	162	162	162	298	298	101	101

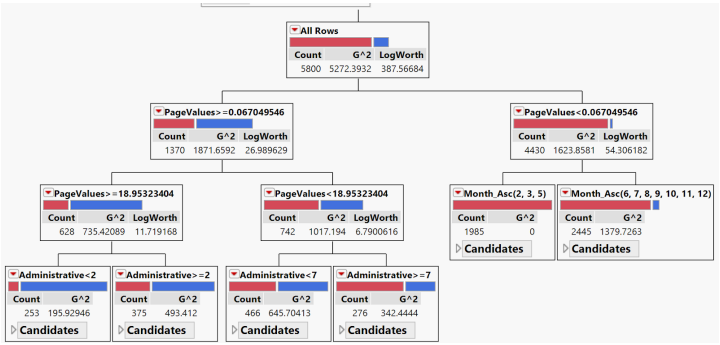
Leaf Report

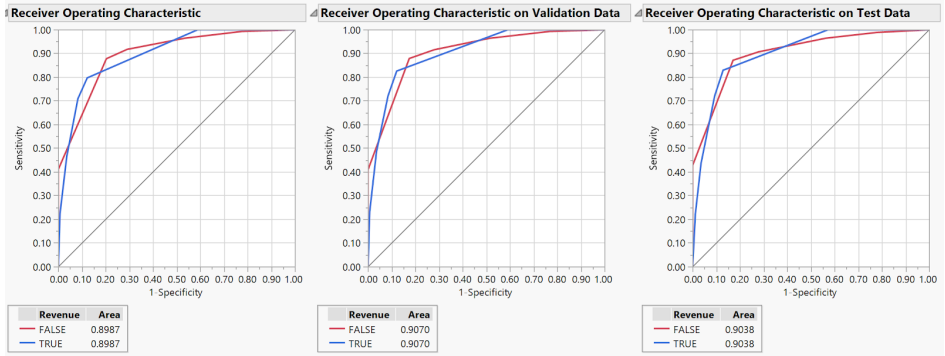
Response Prob

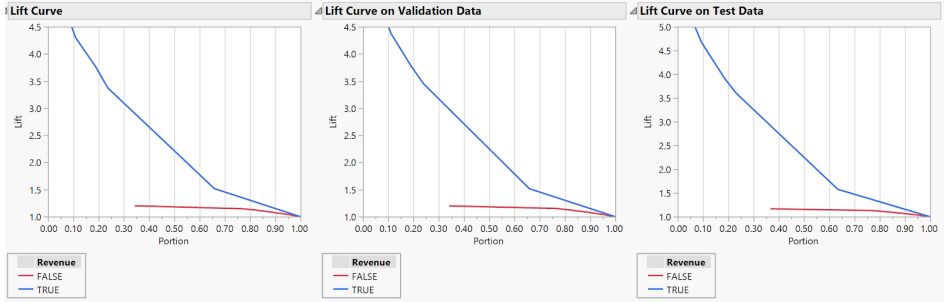
Leaf Label	FALSE	TRUE
PageValues>=0.067049546&PageValues>=18.95323404&Administrative<2	0.1328	0.8672
PageValues>=0.067049546&PageValues>=18.95323404&Administrative>=2	0.3690	0.6310
PageValues>=0.067049546&PageValues<18.95323404&Administrative<7	0.4877	0.5123
PageValues>=0.067049546&PageValues<18.95323404&Administrative>=7	0.6887	0.3113
PageValues<0.067049546&Month_Asc(2, 3, 5)	0.9999	0.0001
PageValues<0.067049546&Month_Asc(6, 7, 8, 9, 10, 11, 12)	0.9186	0.0814

Response Counts

Leaf Label	FALSE	TRUE
PageValues>=0.067049546&PageValues>=18.95323404&Administrative<2	33	220
PageValues>=0.067049546&PageValues>=18.95323404&Administrative>=2	138	237
PageValues>=0.067049546&PageValues<18.95323404&Administrative<7	227	239
PageValues>=0.067049546&PageValues<18.95323404&Administrative>=7	190	86
PageValues<0.067049546&Month_Asc(2, 3, 5)	1985	0
PageValues<0.067049546&Month_Asc(6, 7, 8, 9, 10, 11, 12)	2246	199







Confusion Matrix

Training

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	4421	398
TRUE	285	696

Validation

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	2640	243
TRUE	162	418

Test

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1793	184
TRUE	93	241

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4201	0.4373	0.4062	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5316	0.5487	0.5070	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2636	0.2543	0.2453	∑ -Log(ρ[j])/n
RMSE	0.2860	0.2808	0.2768	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1637	0.1611	0.1561	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1178	0.1170	0.1199	∑ (ρ[j]≠ρMax)/n
N	5800	3463	2311	n

Appendix

Model 3: Boosted Tree

Boosted Tree

Gradient-Boosted Trees Specification

Boosting

Number of Layers:	50
Splits per Tree:	3
Learning Rate:	0.1
Overfit Penalty:	0.0001
Minimum Size Split:	5

Multiple Fits

- [ ] Multiple Fits over Splits and Learning Rate
  - Max Splits Per Tree: 3
  - Max Learning Rate: 0.1
- [ ] Use Tuning Design Table

Stochastic Boosting

Row Sampling Rate	1.0000
Column Sampling Rate	1.0000

Reproducibility

- [ ] Suppress Multithreading
- Random Seed: 0

- [x] Early Stopping

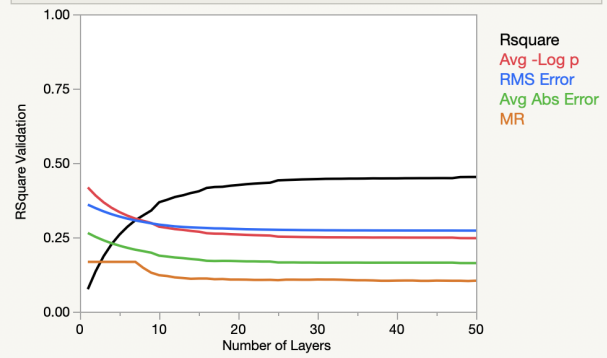
Cancel OK

Cumulative Validation

Column Contributions

Term	Number of Splits	G^2	Portion
Month_Asc	27	105932.544	0.2549
PageValues	36	100322.549	0.2414
VisitorType	15	64726.7031	0.1558
Administrative	10	27043.0297	0.0651
Browser	3	23677.7134	0.0570
Johnson SI Transform ProductRelated_Duration	11	20453.3452	0.0492
Region	7	19841.4044	0.0477
SpecialDay	3	15588.0301	0.0375
TrafficType	7	12919.8328	0.0311
Johnson SI Transform ExitRates	10	12182.0336	0.0293
Informational	3	6357.17507	0.0153
Weekend	8	2343.25222	0.0056
Administrative_Duration_Cat	5	2216.1513	0.0053
Informational_Duration_Cat	5	1951.52709	0.0047
OperatingSystems	0	0	0.0000

Cumulative Validation

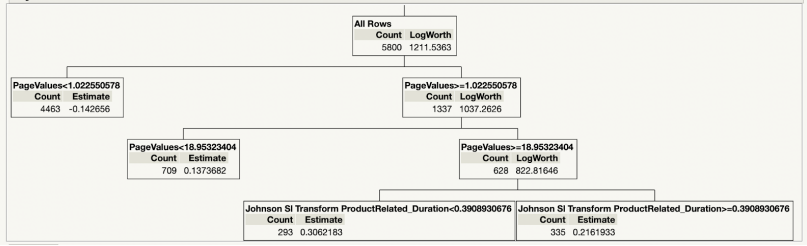


OperatingSystems has no contribution to the model. We consider remove it:

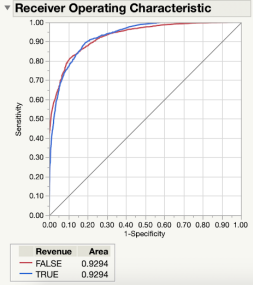
Column Contributions

Term	Number of Splits	G^2	Portion
Month_Asc	27	105932.544	0.2549
PageValues	36	100322.549	0.2414
VisitorType	15	64726.7031	0.1558
Administrative	10	27043.0297	0.0651
Browser	3	23677.7134	0.0570
Johnson SI Transform ProductRelated_Duration	11	20453.3452	0.0492
Region	7	19841.4044	0.0477
SpecialDay	3	15588.0301	0.0375
TrafficType	7	12919.8328	0.0311
Johnson SI Transform ExitRates	10	12182.0336	0.0293
Informational	3	6357.17507	0.0153
Weekend	8	2343.25222	0.0056
Administrative_Duration_Cat	5	2216.1513	0.0053
Informational_Duration_Cat	5	1951.52709	0.0047

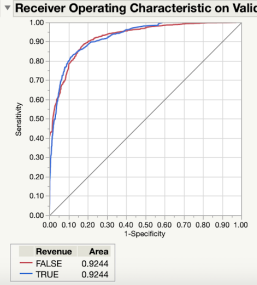
Layer1



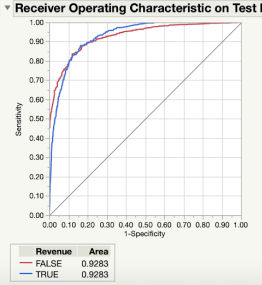
Receiver Operating Characteristic



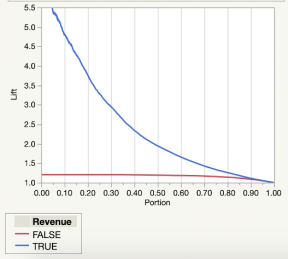
Receiver Operating Characteristic on Validation Data



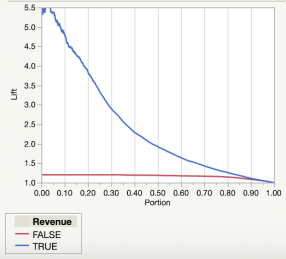
Receiver Operating Characteristic on Test Data



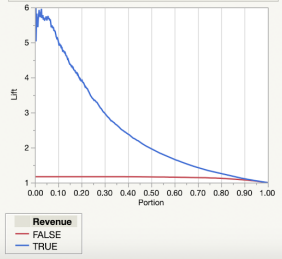
Lift Curve



Lift Curve on Validation Data



Lift Curve on Test Data



Confusion Matrix

Training

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	4653	166
TRUE	421	560

Validation

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	2768	115
TRUE	246	334

Test

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1903	74
TRUE	152	182

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4520	0.4530	0.4354	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5643	0.5647	0.5373	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2491	0.2472	0.2332	$\sum$ -Log(ρ[j])/n
RMSE	0.2746	0.2726	0.2659	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.1641	0.1634	0.1569	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.1012	0.1042	0.0978	$\sum$ (ρ[j]≠ρMax)/n
N	5800	3463	2311	n

Confusion Matrix

Appendix

Model 4: KNN

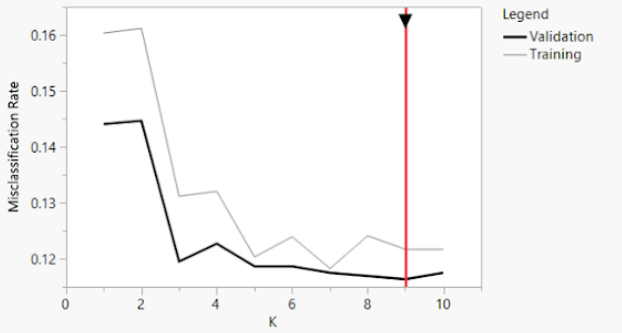
K=9 for validation with misclassification as **453** inclusive of False positives and false negatives.

From logistic regression, the optimal predictor variables are :
PageValues, ProductRelatedDuration, VisitorType, SpecialDay and ExitRates.

K Nearest Neighbors

Revenue

Model Selection



Training

K	Count	Misclassification Rate	Misclassifications
1	5800	0.16034	930
2	5800	0.16121	935
3	5800	0.13121	761
4	5800	0.13207	766
5	5800	0.12034	698
6	5800	0.12397	719
7	5800	0.11828	686 *
8	5800	0.12414	720
9	5800	0.12172	706
10	5800	0.12172	706

Validation

K	Count	Misclassification Rate	Misclassifications
1	3463	0.14409	499
2	3463	0.14467	501
3	3463	0.11955	414
4	3463	0.12273	425
5	3463	0.11868	411
6	3463	0.11868	411
7	3463	0.11753	407
8	3463	0.11695	405
9	3463	0.11637	403 *
10	3463	0.11753	407

Test

K	Count	Misclassification Rate	Misclassifications
1	2311	0.15578	360
2	2311	0.16443	380
3	2311	0.13371	309
4	2311	0.12722	294
5	2311	0.12073	279
6	2311	0.11640	269
7	2311	0.11986	277
8	2311	0.11467	265 *
9	2311	0.11856	274
10	2311	0.12203	282

Confusion Matrix for Best K=9

Training

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	4600	219
TRUE	487	494

Validation

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	2758	125
TRUE	278	302

Test

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1883	94
TRUE	180	154

Appendix

Model 5: Bootstrap forest

Bootstrap Forest for Revenue

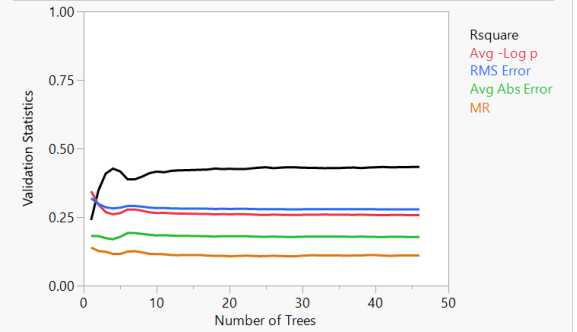
Specifications

Target Column:	Revenue	Training Rows:	5800
Validation Column:	Validation	Validation Rows:	3463
		Test Rows:	2311
Number of Trees in the Forest:	46	Number of Terms:	15
Number of Terms Sampled per Split:	3	Bootstrap Samples:	5800
		Minimum Splits per Tree:	10
		Minimum Size Split:	12

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4876	0.4323	0.4029	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5996	0.5436	0.5035	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2329	0.2565	0.2467	$\sum -\text{Log}(\rho[j])/n$
RMSE	0.2614	0.2774	0.2716	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.1672	0.1768	0.1713	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.0836	0.1097	0.1047	$\sum (\rho[j]\neq\rho Max)/n$
N	5800	3463	2311	n

Cumulative Validation



Column Contributions

Term	Number of Splits	G^2	Portion
PageValues	573	880.392394	0.5691
Johnson SI Transform ExitRates	641	133.719195	0.0864
Month_Asc	506	121.787657	0.0787
Johnson SI Transform ProductRelated_Duration	653	110.962114	0.0717
Administrative	522	64.6402315	0.0418
TrafficType	537	40.7119907	0.0263
Region	532	35.2682357	0.0228
VisitorType	228	32.20177	0.0208
Administrative_Duration_Cat	158	29.2004924	0.0189
OperatingSystems	418	21.1913271	0.0137
Informational	277	18.7926655	0.0121
SpecialDay	135	15.4463359	0.0100
Browser	323	15.2721673	0.0099
Informational_Duration_Cat	227	14.6496545	0.0095
Weekend	359	12.6847497	0.0082

Confusion Matrix

Training

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	4766	53
TRUE	432	549

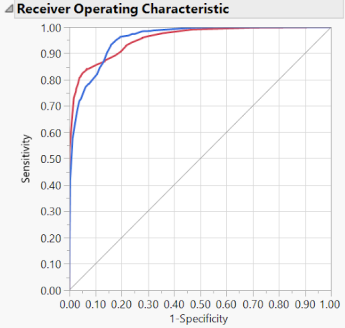
Validation

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	2812	71
TRUE	309	271

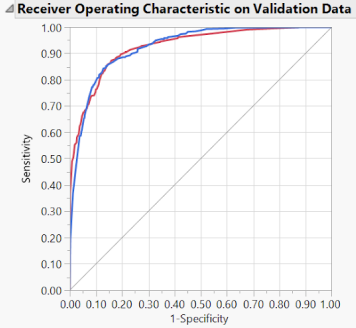
Test

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1933	44
TRUE	198	136

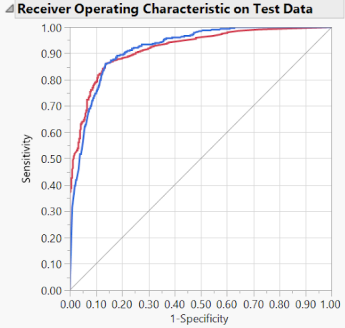
Receiver Operating Characteristic



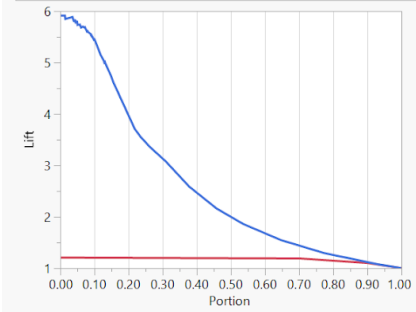
Receiver Operating Characteristic on Validation Data



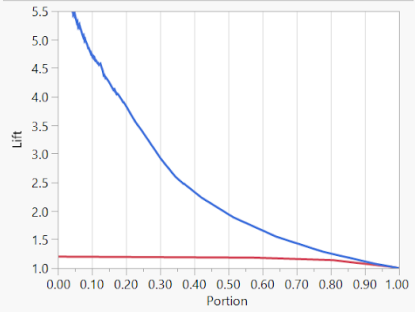
Receiver Operating Characteristic on Test Data



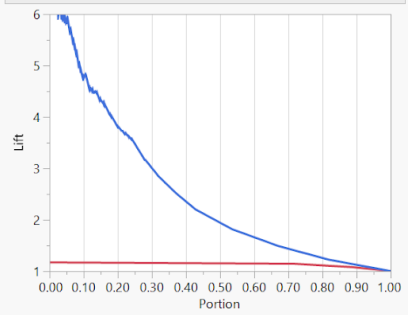
Lift Curve



Lift Curve on Validation Data



Lift Curve on Test Data



Appendix

Model 6: Neural Networks

(ii) 2 Layers with 3 Tan H each

Cleaned-Validation_TRUE_PROP_0.5 - Neural of Revenue 6 - JMP Pro [2]

Neural

Validation Column: Validation Stratified by Revenue

Model Launch

Model NTanH(3)NTanH2(3)

Training

Revenue

Measures	Value
Generalized RSquare	0.6603817
Entropy RSquare	0.4943747
RMSE	0.3312709
Mean Abs Dev	0.2208017
Misclassification Rate	0.1594479
-LogLikelihood	731.27828
Sum Freq	2101

Confusion Matrix

Actual Revenue	Predicted Count FALSE	TRUE
FALSE	1006	147
TRUE	188	760

Confusion Rates

Actual Revenue	Predicted Rate FALSE	TRUE
FALSE	0.873	0.127
TRUE	0.198	0.802

Validation

Revenue

Measures	Value
Generalized RSquare	0.6384425
Entropy RSquare	0.4715483
RMSE	0.3367237
Mean Abs Dev	0.2213203
Misclassification Rate	0.1455547
-LogLikelihood	461.41911
Sum Freq	1271

Confusion Matrix

Actual Revenue	Predicted Count FALSE	TRUE
FALSE	618	88
TRUE	97	468

Confusion Rates

Actual Revenue	Predicted Rate FALSE	TRUE
FALSE	0.875	0.125
TRUE	0.172	0.828

Test

Revenue

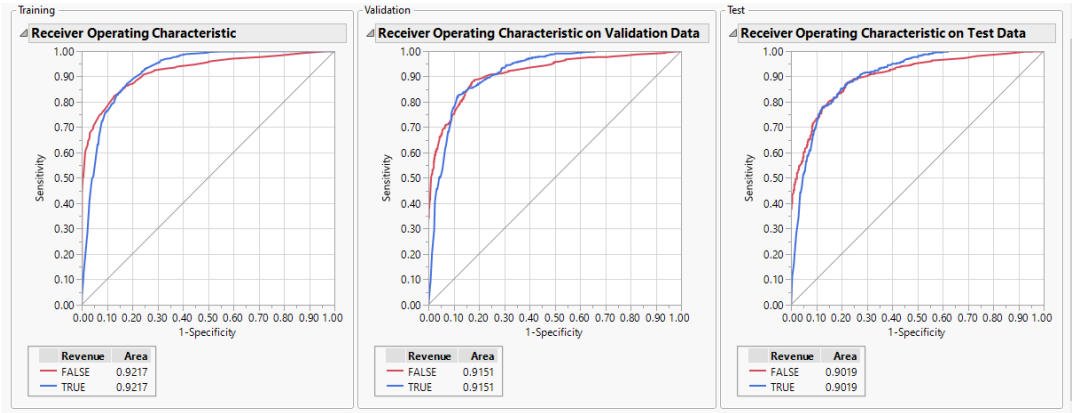
Measures	Value
Generalized RSquare	0.1884282
Entropy RSquare	0.1319458
RMSE	0.3537621
Mean Abs Dev	0.2274259
Misclassification Rate	0.1629918
-LogLikelihood	899.73467
Sum Freq	2313

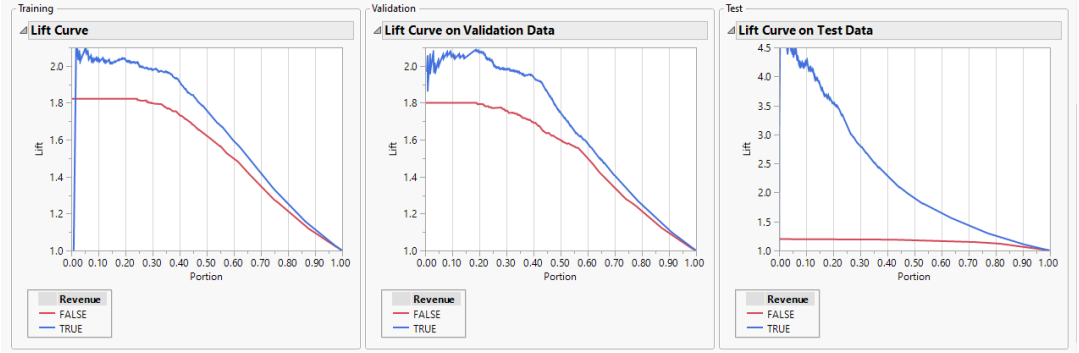
Confusion Matrix

Actual Revenue	Predicted Count FALSE	TRUE
FALSE	1633	298
TRUE	79	303

Confusion Rates

Actual Revenue	Predicted Rate FALSE	TRUE
FALSE	0.846	0.154
TRUE	0.207	0.793





Appendix

Model 7: Naïve Bayes

Naive Bayes

Fit Details

Revenue

Training

Count	Misclassification Rate	Misclassifications
5800	0.15793	916

Validation

Count	Misclassification Rate	Misclassifications
3463	0.13572	470

Test

Count	Misclassification Rate	Misclassifications
2311	0.14539	336

Confusion Matrix

Training

Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	4380	439
TRUE	477	504

Validation

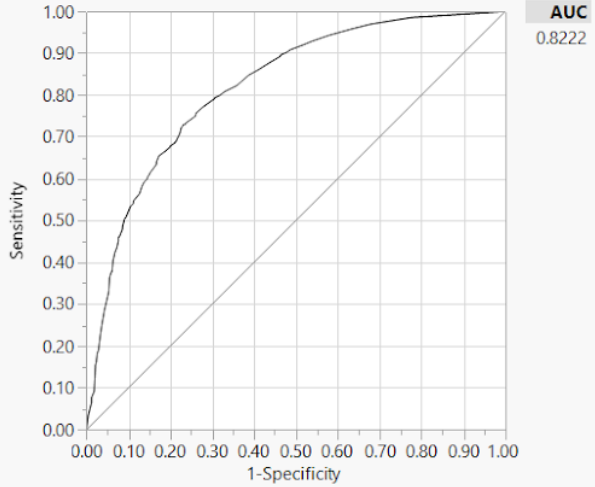
Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	2666	217
TRUE	253	327

Test

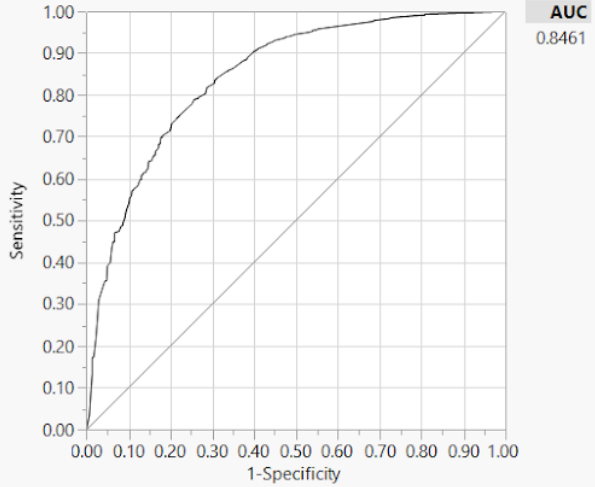
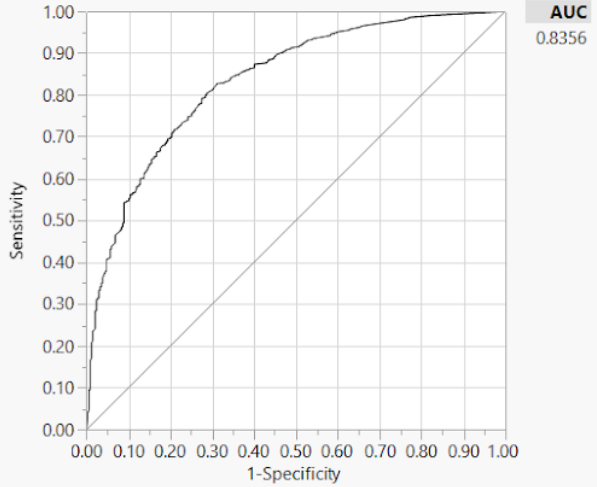
Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1804	173
TRUE	163	171

ROC Curve for Revenue = FALSE

Training



Validation



Test