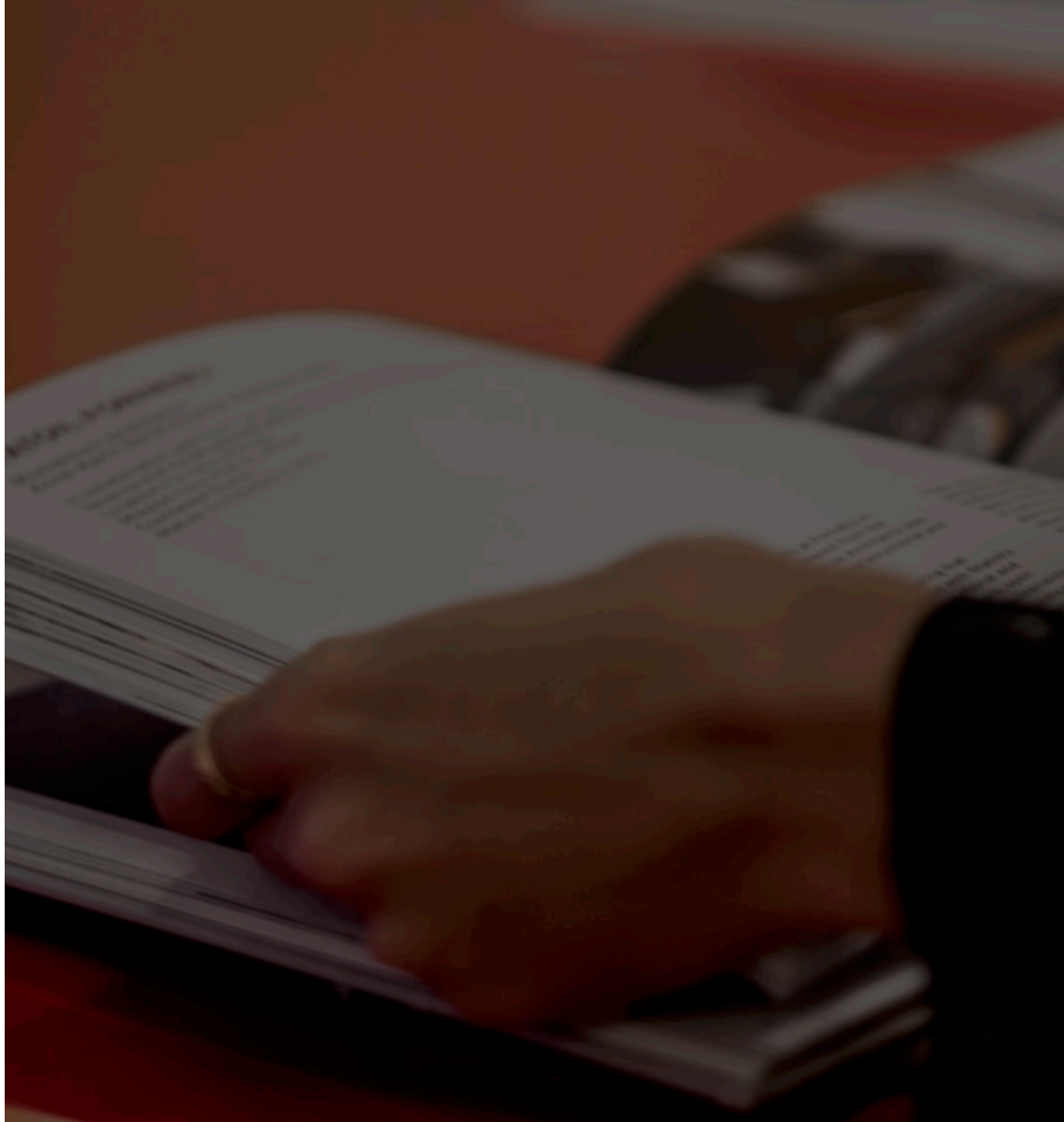
Predict Purchasing Intentions of Online Shoppers

Sean Robertson. Helen Phan

Goals

Predict the shopping intent of website visitors

- Purchase
- No purchase



Dataset

- "Online Shoppers Purchasing Intention" retrieved from UCI machine learning repository, data is from Columbia Sportswear Company
- No missing data
- 50% for training, 50% for validation

12,330

Online shopping sessions

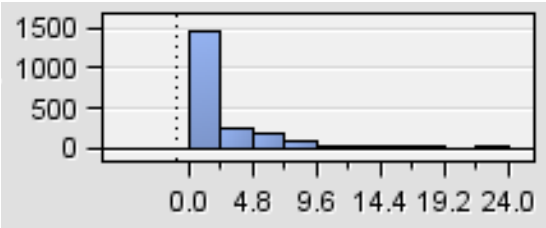
10

Numerical variables

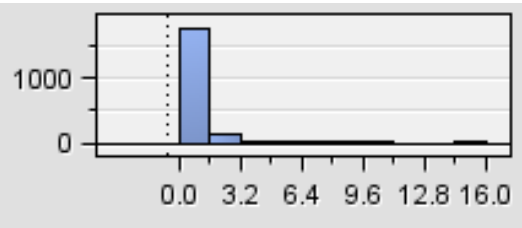
8

Categorical variables

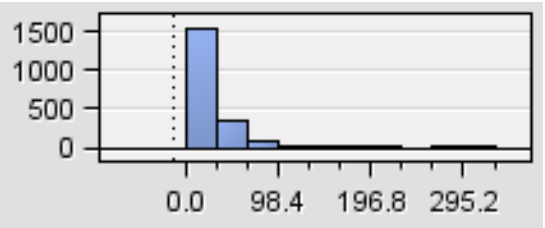
Dataset / 10 numerical variables



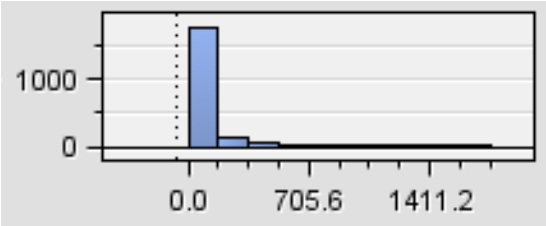
Administrative pages



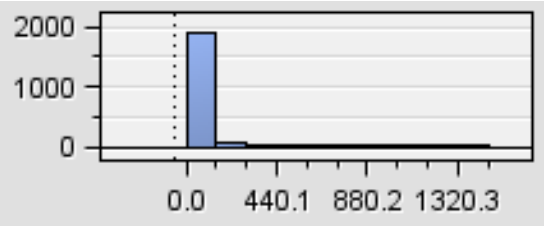
Informational pages



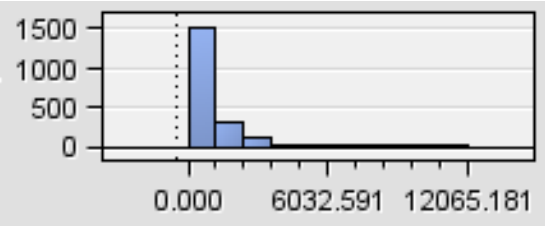
Product related pages



Time spent on administrative pages

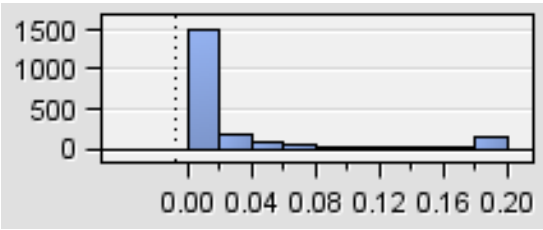


Time spent on informational pages

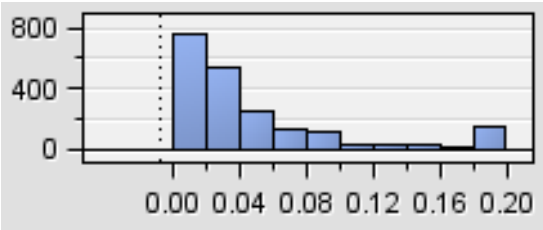


Time spent on product related pages

Dataset / 10 numerical variables

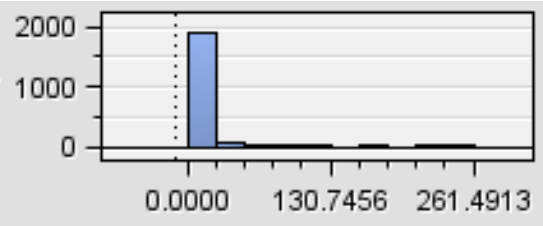


Bounce rate



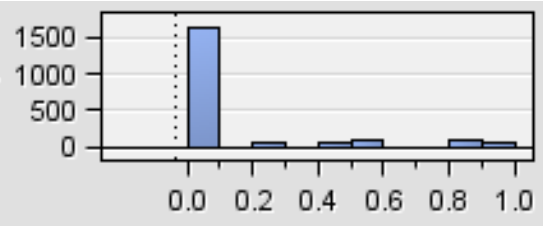
Exit rate

Dataset / 10 numerical variables



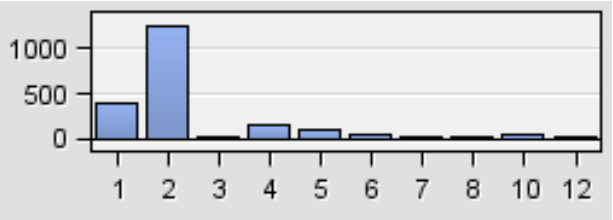
Page value

$$= \frac{e - commerce\ revenue + total\ goal\ value}{\#\ unique\ pageviews\ for\ given\ page}$$

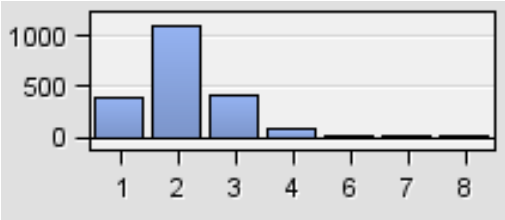


Special day

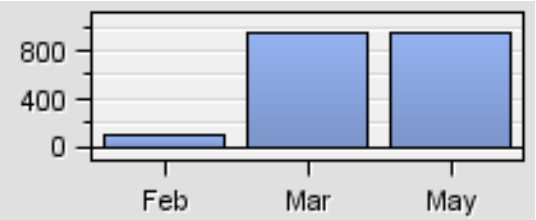
Dataset / 8 categorical variables



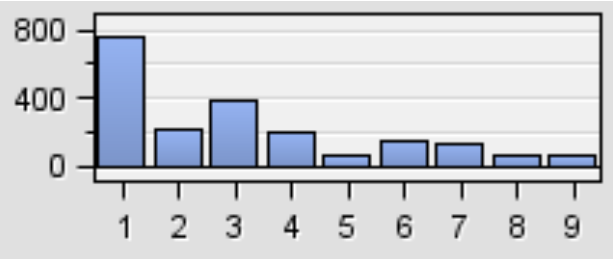
Browser



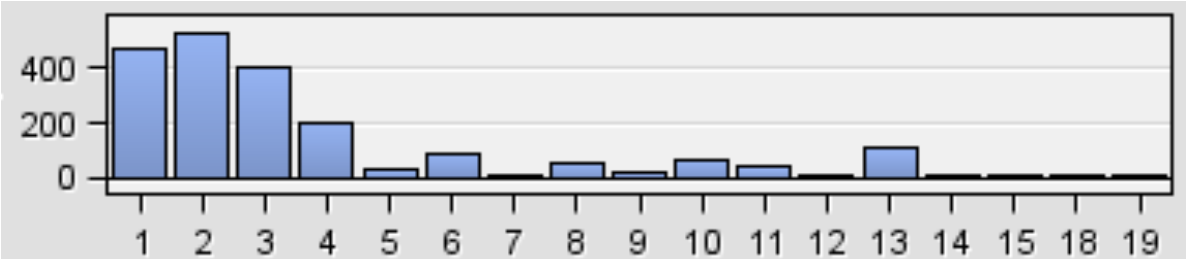
Operating system



Month



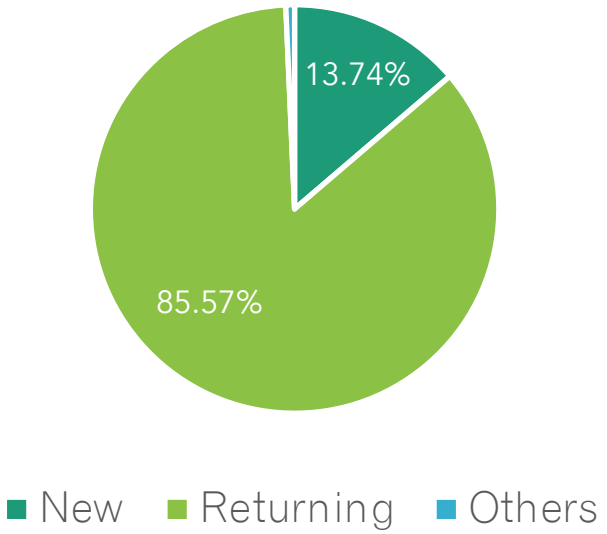
Region



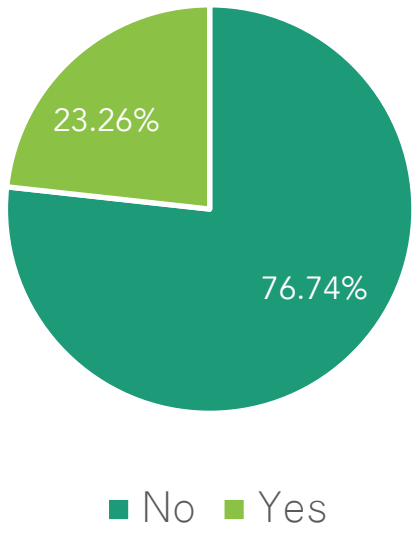
Traffic Source

Dataset / 8 categorical variables

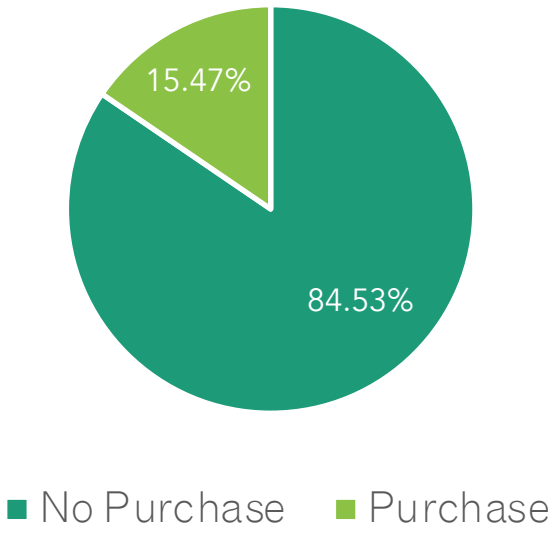
Visitor Type



Weekend



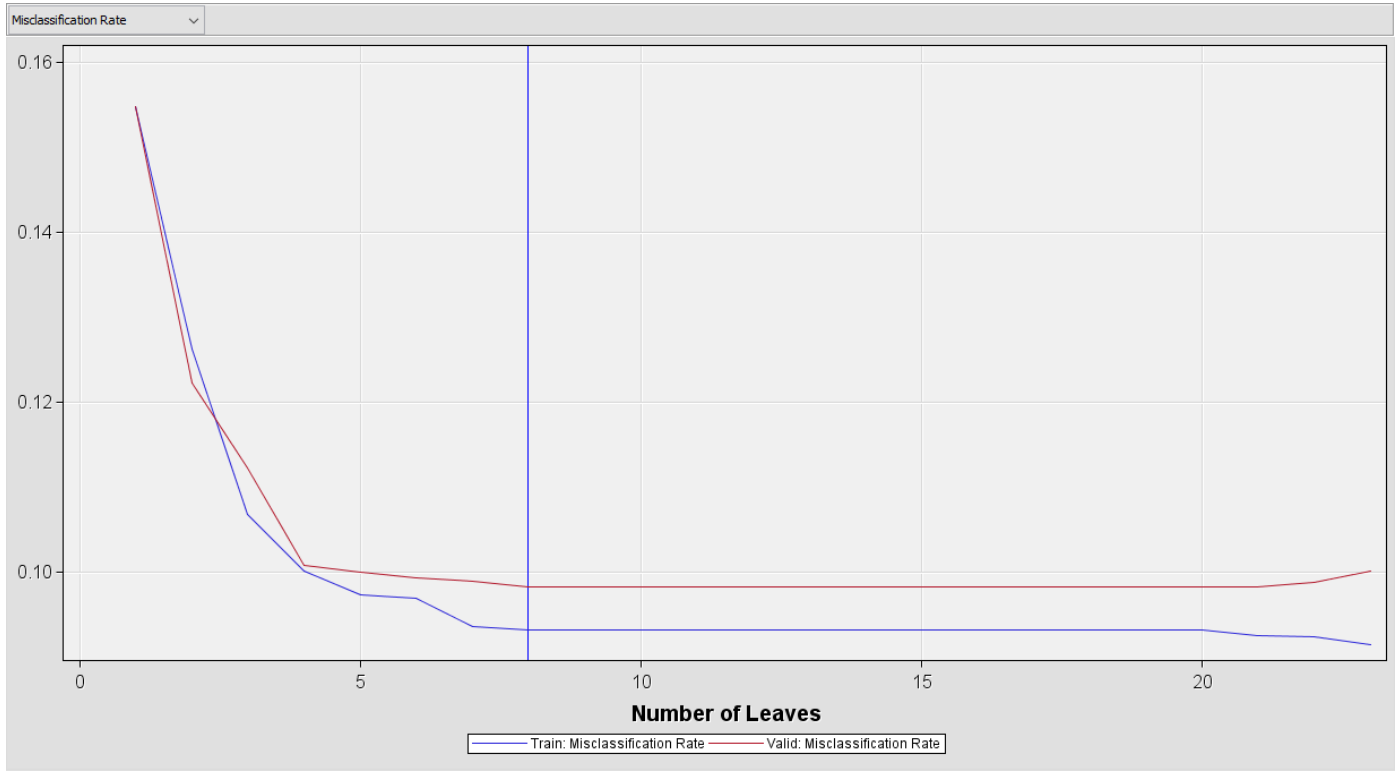
Revenue



Predict purchasing intention

1. Decision Tree
2. Logistic Regression
3. Neural Networks
4. Random Forest

Decision Tree

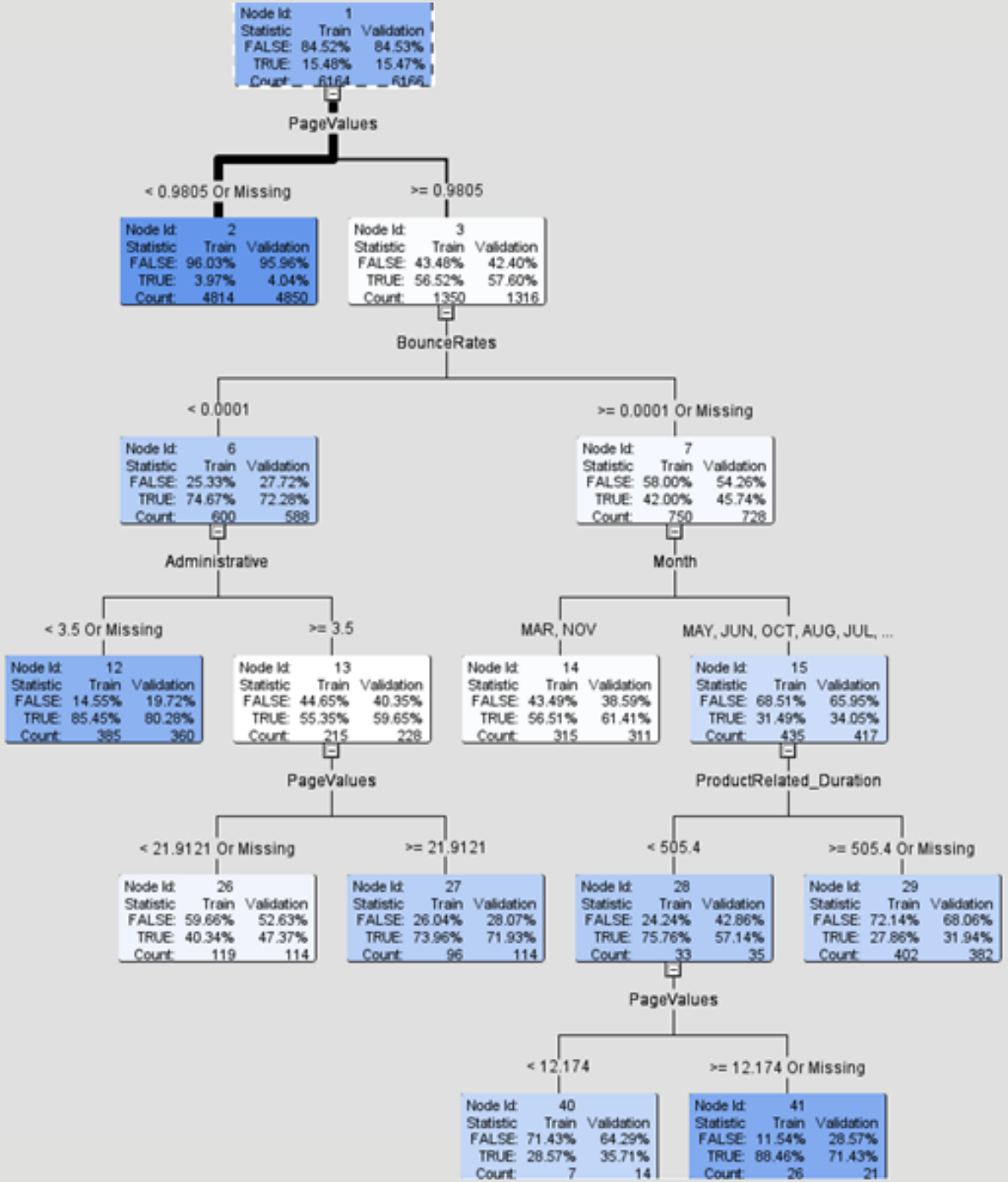


Subtree Assessment Plot

- Obtain the maximal decision tree
- Model overfitting

⬇

Optimal: 8-leaf tree



The optimal tree

- 41 nodes and 8 leaves
- Classification conditions
  - Page value
  - Bounce rate
  - Administrative page
  - Month
  - Time on product related pages
- Misclassification rate
  - Train data: 9.312%
  - Validation data: 9.828%

The optimal tree

Leaf	Classification Rules	% of train data		% of validation data		Purchasing Intention
		Y = False	Y = True	Y = False	Y = True	
1	Page value < 0.9805	96.03	3.97	95.96	4.04	No
2	Page value >= 0.9805 Bounce Rates < 0.0001 Administrative <3.5	14.55	85.45	19.72	80.28	Yes
3	Page value >= 0.9805 Bounce Rates < 0.0001 in Mar, Nov	43.49	56.51	38.59	61.41	Yes
4	Page value >= 0.9805 or <21.9121 Bounce Rates < 0.0001 Administrative >=3.5	59.66	40.34	52.63	47.37	No
5	Page value >= 21.9121 Bounce Rates < 0.0001 Administrative >=3.5	26.04	73.96	28.07	71.93	Yes

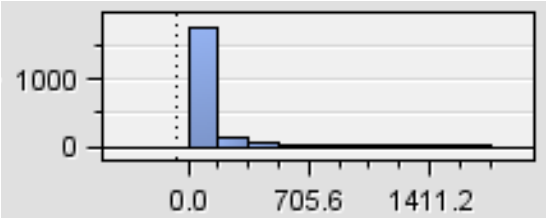
The optimal tree

Leaf	Classification Rules	% of train data		% of validation data		Purchasing Intention
		Y = False	Y = True	Y = False	Y = True	
6	Page value >= 0.9805 Bounce Rates < 0.0001 In between May and Oct Product Related Duration >= 505.4	72.14	27.86	68.06	31.94	No
7	Page value >= 0.9805 or <12.174 Bounce Rates < 0.0001 In between May and Oct Product Related Duration < 505.4	71.43	28.57	64.29	35.71	No
8	Page value >= 12.174 Bounce Rates < 0.0001 In between May and Oct Product Related Duration < 505.4	11.54	88.46	28.57	71.43	Yes

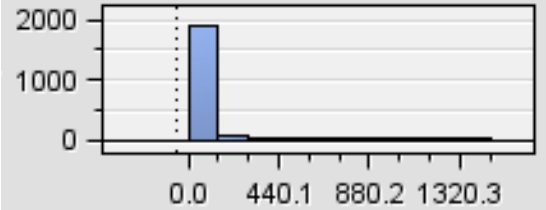
Logistic Regression

	Train	Validation
• Regression	11.162%	11.531%
• Regression after stepwise variable selection	11.291%	11.434%
• Regression after transformation	10.188%	10.282%
• Regression after transformation, categorial variables recoded and stepwise variable selection	10.399%	10.282%

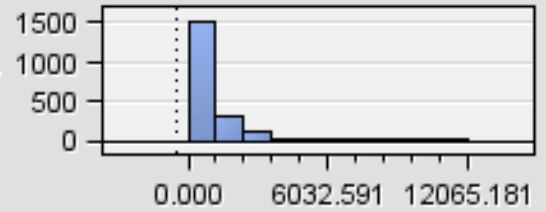
Time spent on administrative pages



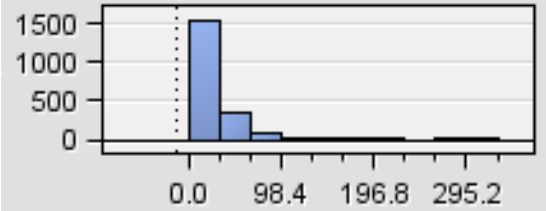
Time spent on informational pages



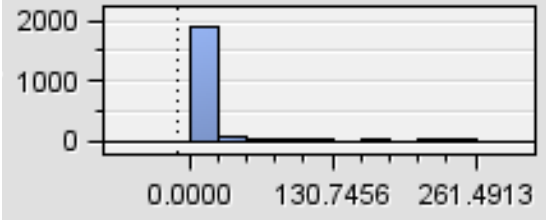
Time spent on product related pages



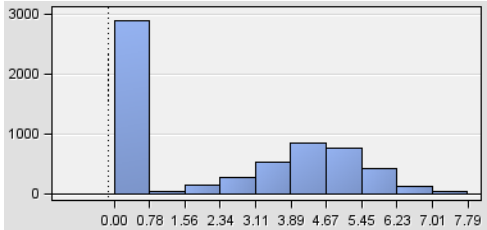
Product related pages

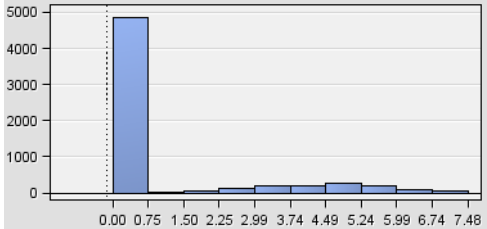


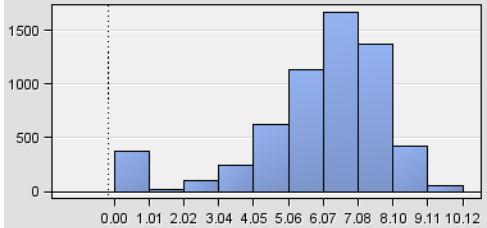
Page value

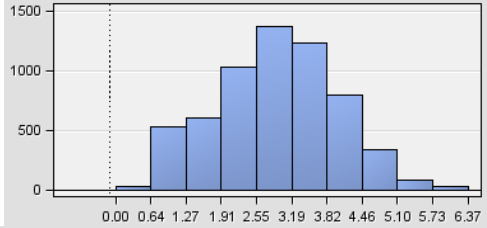


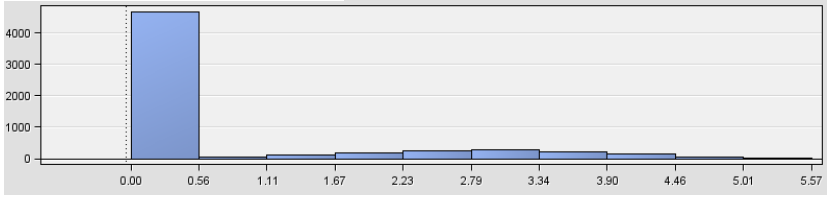
Log Transformation











Logistic Regression

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
5312.062	3063.412	2248.6495	11	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-0.3362	0.3674	0.84	0.3602		0.714
ExitRates		1	-16.8609	2.6374	40.87	<.0001	-0.4515	0.000
Indicator_Aug	0	1	-0.3385	0.1304	6.74	0.0094		0.713
Indicator_Nov	0	1	-0.6306	0.0540	136.28	<.0001		0.532
Indicator_Oct	0	1	-0.2350	0.1101	4.56	0.0328		0.791
Indicator_Sept	0	1	-0.4199	0.1154	13.23	0.0003		0.657
LOG_Administrative_Duration		1	-0.0688	0.0223	9.52	0.0020	-0.0896	0.934
LOG_PageValues		1	1.1359	0.0347	1070.42	<.0001	0.7949	3.114
New_Visitor	0	1	-0.2587	0.0664	15.16	<.0001		0.772
TrafficType_10	0	1	-0.3024	0.1116	7.35	0.0067		0.739
TrafficType_13	0	1	0.3589	0.1392	6.65	0.0099		1.432
TrafficType_20	0	1	-0.3888	0.1773	4.81	0.0283		0.678
TrafficType_8	0	1	-0.4175	0.1164	12.87	0.0003		0.659

Logistic Regression

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -.336 - 16.86\,(\text{ExitRates}) - .339\,(\text{Aug}) - .419(\text{Sept}) - .235(\text{Oct}) - .63\,(\text{Nov}) - .068(\text{log\_administrative duration}) + 1.136\,(\text{log\_pagevalues}) - .258\,(\text{New Visitor}) - .302(\text{traffic type\_10}) + .359\,(\text{traffic type\_13}) - .388(\text{traffic type\_20}) - .417(\text{traffic type\_8})$$

Variable	Odd ratio	% Change
ExitRates	0.001	-100%
Indicator_Aug	0.508	-49%
Indicator_Nov	0.283	-72%
Indicator_Oct	0.625	-38%
Indicator_Sept	0.432	-57%
LOG_Administrative _Duration	0.934	-7%

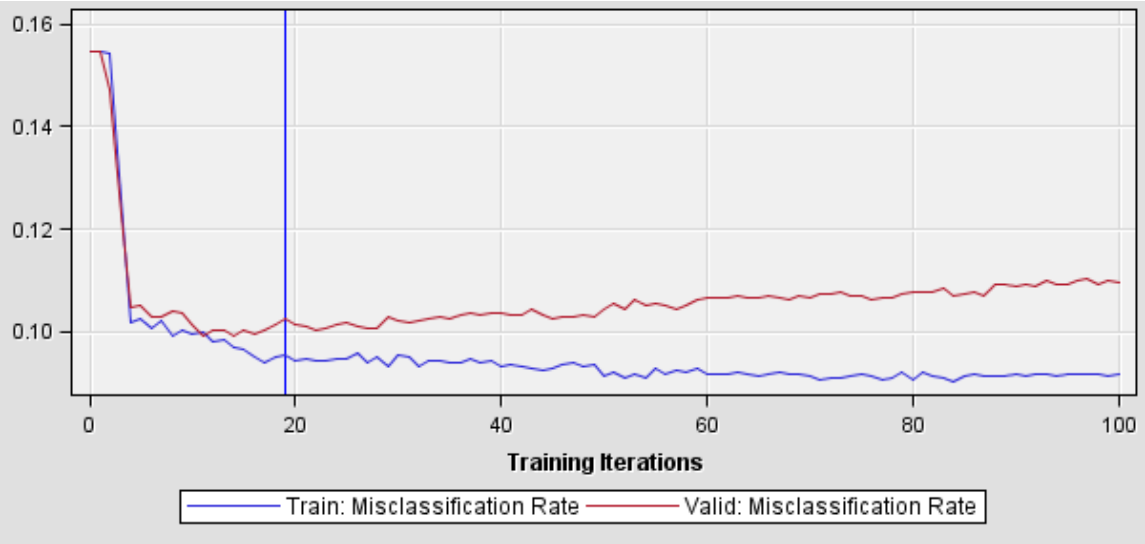
Variable	Odd ratio	% Change
LOG_PageValues	3.114	211%
New_Visitor	0.596	-40%
TrafficType_10	0.546	-45%
TrafficType_13	2.05	105%
TrafficType_20	0.46	-54%
TrafficType_8	0.434	-57%

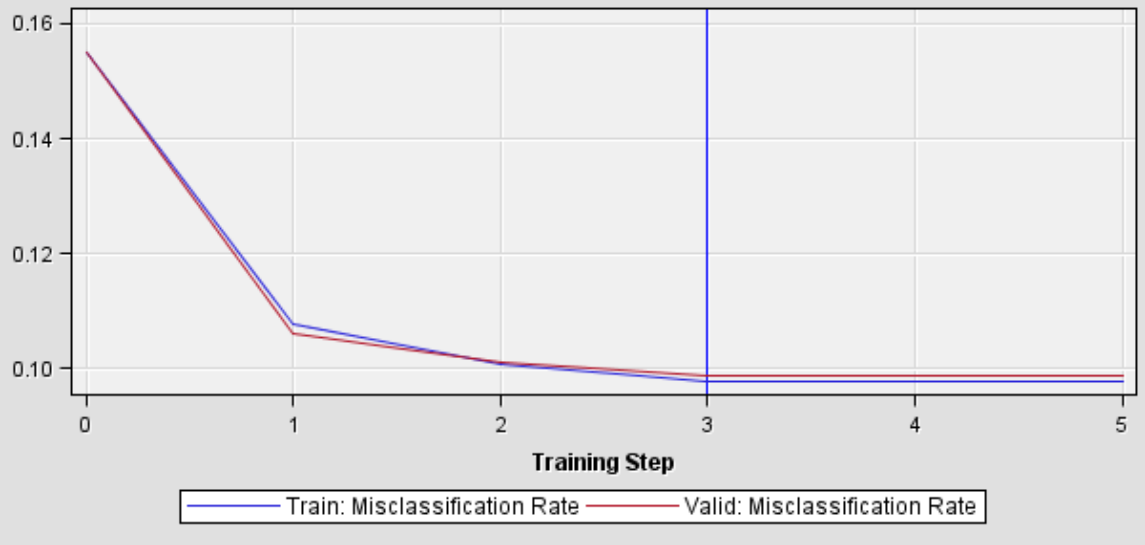
Neural Networks

	Train	Validation
• Auto Neural Network after stepwise	9.71%	10.08%
• Auto Neural Network after variable transformation and stepwise	9.79%	9.87%
• Auto Neural Network after variable transformation, recoding categorical variables and stepwise	10.61%	10.23%

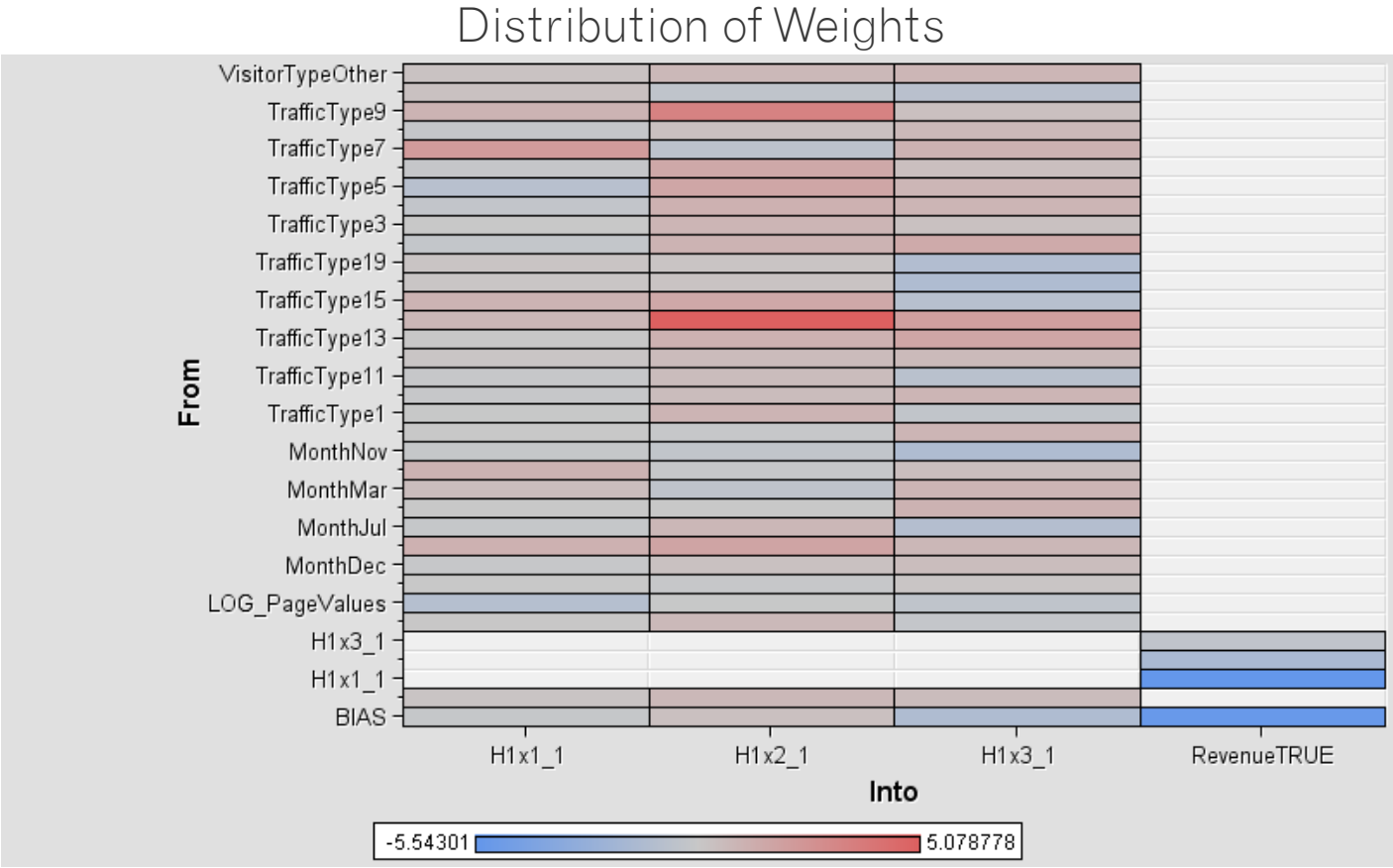
Neural Networks

- Neural Network (top) had 199 weights/parameters and 19 iterations to optimize Average Squared Error and Misclassification Rate
- Auto-Neural Network (bottom) had 100 weights and 3 hidden units
- Both models are large as a result of 18 variables in the model





Neural Networks

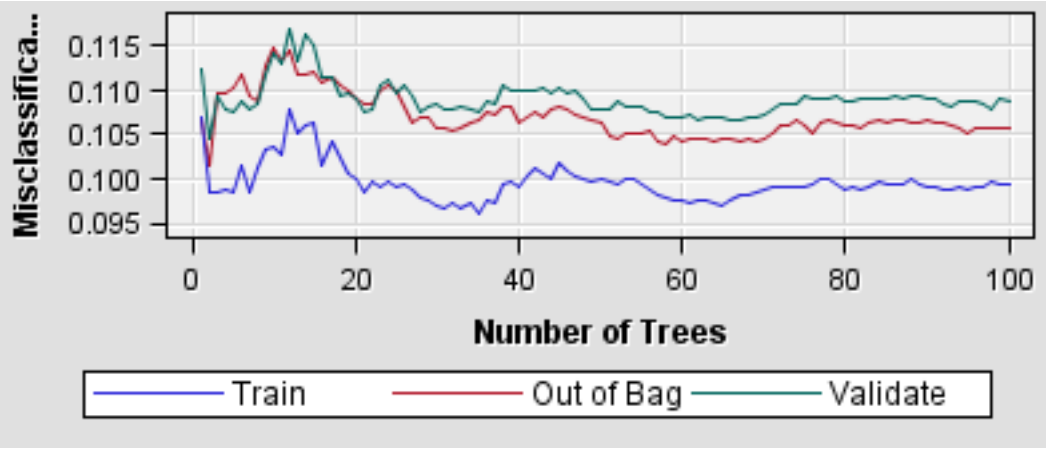


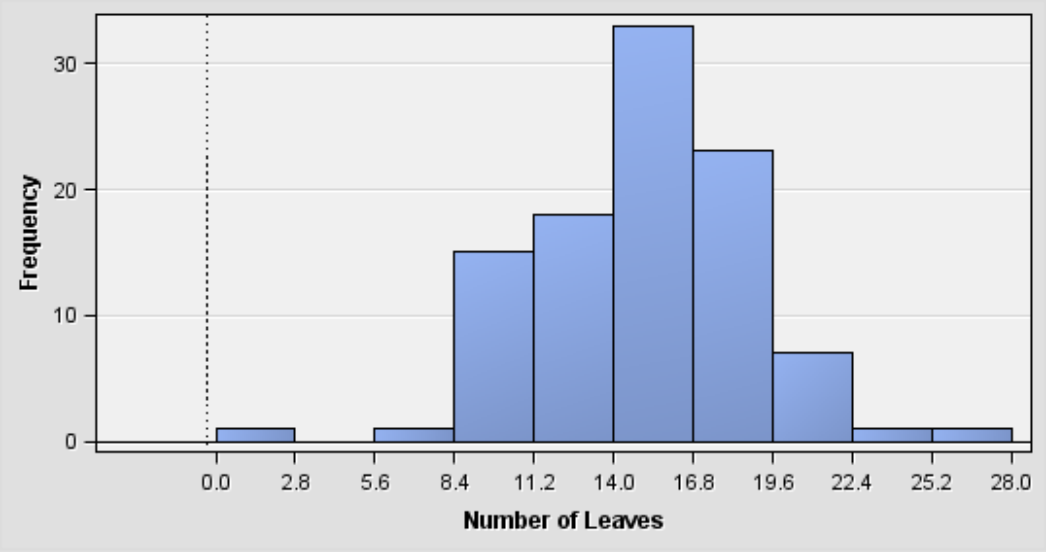
Random Forest

	Train	Validation
• Random Forest	9.61%	10.88%
• Random Forest after stepwise	9.08%	10.1%
• Random Forest after variable transformation	9.68%	10.78%
• Random Forest after variable transformation and stepwise	9.94%	10.88%
• Random Forest after variable transformation, recoding categorical variables and stepwise	8.19%	10.5%

Random Forest

- Misclassification Rate of training data converged to 9.94%
- Misclassification Rate of validation data converged to 10.9%
- No significant difference when using a forest larger than the default of 100 trees





Conclusion

#	Model	MISC Train	MISC Validation
1	Decision Tree	9.312%	9.828%
2	Regression after variable transformation, recoding categorical variables and stepwise	10.399%	10.282%
3	Auto Neural Network after variable transformation and stepwise	9.798%	9.876%
4	Random Forest after stepwise	9.085%	10.104%

Conclusion

When do people purchase?

- Visit the page that had high average value (Page value >= 0.9805)
- When they did not bounce from the site (Bounce Rates < 0.0001)
- Visit to their account management pages was low (<4), or if high was also paired with even higher average value (Page value >= 21.9805)

- Page value >= 0.9805 and Bounce Rates < 0.0001 and Administrative <3.5
- Page value >= 0.9805 and Bounce Rates < 0.0001 and in Mar, Nov
- Page value >= 21.9121 and Bounce Rates < 0.0001 and Administrative >=3.5
- Page value >= 12.174 and Bounce Rates < 0.0001 and In between May and Oct and Product Related Duration < 505.

Thank you!