Predicting Online Shopping Purchasing Intention

Table of Contents

Section 1 Executive Summary

The usage of e-commerce has increased significantly over the past few years, but the ratio of users visiting the website to view a product to the users that end up buying the product is always a concerning factor for the e-commerce companies. There are various factors such as ease of finding a product, post-purchase service, minimization of overall shopping effort, lower price and selection that impact the preference of online customers. Abandoning the shopping cart is also increasing due to the increase in the competition. Therefore, there is a need for companies to study the behavior of online shoppers and target them based on their buying patterns to avoid the threat of sales from their competitors. In this project, the likelihood of a customer making a purchase after visiting the website is predicted using the clickstream data and session data of the users visiting the website of one online retailer over a period of one year. Different predictive models were built to achieve this objective and a decision tree model was chosen for its performance, interpretability, and simplicity. Using the final decision tree model, 80.84% of the actual purchasing customers would be targeted. Based on the results, the company is recommended to utilize association rule to increase the purchasing of returning visitors who usually go through Administrative pages, identify and optimize page values to boost conversion rate, and adopt different marketing strategies in different months. By combining the model and recommended strategies, the company is expected to increase revenue and decrease customer churning, then increase customer satisfaction and company reputation.

Section 2 Problem Statement

A major proportion of users visiting an online shopping website browse through the catalog but do not happen to purchase the products. Some of these users come to buy and some of them are just purely browsing the catalog. It is very beneficial for the e-commerce company to identify the customers with purchase intent and nudge these customers towards completing the purchase. Predicting the likeness of the customers to purchase not only increases the revenue but also the brand value and reputation of the company as this would result in the effective use of time for the shoppers. Online shoppers' decisions are very hard to predict as they depend on the speculative thought process of individuals. Researchers have been focused on predicting the psychological state of consumers. The objective of this study is to predict the likeliness and classify the visitors into two groups – with purchase intent and no purchase intent. Prediction is made using the clickstream & pageview data tracked in the current and past sessions along with user information. Available data used consists of 18 variables which include the target variable of whether a visitor would make a purchase, number of administrative & informational pages visited, time spent on these pages, past bounce rates, exit rates, demographic region, visitor type, month.

Section 3 Methodology

Section 3.1 Sample

The dataset called 'Online Shoppers Purchasing Intention Dataset' is fetched from the UCI machine learning repository. The dataset describes the online shopping intention of 12,330 customers, among whom 84.5% (10,422) did not end with a purchase, and 15.5% (1908) ended with a purchase. This dataset is large enough to conduct the prediction with 14 numerical and 4 categorical attributes. The target Revenue is categorical, TRUE means the purchase occurred and FALSE is otherwise.

3.1.1 Data Dictionary

The initial data descriptions are as follows:

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 that 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.

Section 3.2 Explore

At the outset, while eyeballing the data in terms of variables, the problem statement can be categorized into a classification problem with **Revenue** being the target variable. Revenue here implies True or False, i.e., whether a session generated revenue from the purchasing of the customer or not. We find **no missing values** in the current data set.

Few of the interesting inferences we found in the dataset are as below:

- The dataset has 3 columns of duration in which a user is likely to spend his online screen time on the website, namely- Administrative, Informational and ProductRelated duration. The more the time spent by the user the more **likely** it is his intention to purchase a product is a basic rationale and this aligns with the current dataset context.

- There is a categorical predictor variable of ‘Month’, which signifies the purchases in any particular month. By exploring this variable, we notice an **increasing** trend in the purchasing intent of customers throughout the year.

- A few anomalies were detected in the duration variables, if we look up the distribution of these, we find many data points outside the box plots for Administrative, Informational and PageRelated duration variables.
    - This needs for transforming and/or standardizing variables as we cannot cater to removing many data points

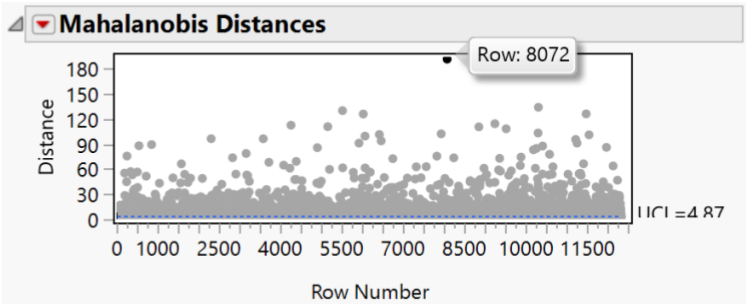
- Exploring the predictor variable 'VisitorType' gives more insights into the target revenues. On further examination of various visitor types - New visitor, Returning visitor and Others, the conversion rates of visitors are 25%, 14% and 19% respectively. One of the observations here is that returning users are least likely to purchase.

Section 3.3 Modify

We modified the data set to fit the needs of model building:

3.3.1. Detect and deal with outliers:

We use the Mahalanobis distance method to detect outliers:



We found one record that has extreme value. This user has cumulatively spent 70k seconds approximately (Administration duration, information duration, product duration) in one single session, which even if possible realistically, is very deviating from the normal trend so we excluded it.

After excluding that and rerun the Mahalanobis Distances, there are other possible outliers, but they are not as far as row 8072. They could be valuable for the predictive modeling, we decided to keep them.

3.3.2. Data reduction

We aim to distill complex data into simpler data.

3.3.2.1 Reduce variables

a. Principal Components Analysis

By conducting PCA, we aim to reduce the dimensionality of predictors. However, it increased the complexity of the data set but only reduced 2 variables. The complexity is outweighing the number of variables that are reduced, so we decided not to use PCA.

b. Multivariate Correlations

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

There are two pairs of variables with high correlations. Product Related and Product Related Duration; Exit rate and Bounce rate. We decided to keep one variable for each pair to avoid multicollinearity.

c. Logistic regression

To decide which variable to keep, we built a logistic regression and kept the variables with higher contribution in the model. We dropped Product Related and Bounce rate and retained Product Related Duration and Exit rate.

3.3.2.2 Reduce Rows

We excluded rows that ProductRelated_Duration=0. Because generally a customer will not purchase a good without clicking the product-related website, the Revenue in such sessions will be 0. Thirteen rows out of 756 rows with ProductRelated_Duration=0 are exceptions, but they account for only 1% and have little impact.

3.3.3 Data Transformation

3.3.3.1 Transform into normal distribution

Based on distributions, ProductRelated_Duration, ExitRates, Administrative_Duration and Informational_Duration are highly skewed, so we transformed these variables for better performance in predictions. But the transformation of Administrative_Duration and Informational_Duration is just creating a handful of distinct values. We decided to transform them into categorical variables.

3.3.3.2 Transform into categorical variable

We then converted Administrative_Duration and Informational_Duration into a 0/1 (binary valued) variables. All values greater than 1 are flagged as 1 and zeros are left unchanged.

Section 3.4 Modeling:

The number of rows with 'Revenue' as TRUE are very few compared to the number of with 'Revenue' as FALSE and our target prediction is TRUE Revenue. Besides modeling with original data, we created balanced datasets as well. In our model, the cost of False Negative is higher than that of False Positive, and this helps when tuning models. After finalizing the features required to create an optimal model which can predict the maximum correct results, we implemented following modeling techniques and show the best results of each after thorough validation comparisons:

3.4.1 Logistic Regression:

Logistic regression is performed with original data and balanced dataset, and on comparing various metrics like sensitivity, accuracy of models; balanced sampling model is better. We found Page Values, Month, ProductRelated_Duration, Visitor Type, and ExitRates as significant predictors.

Confusion Matrix:

Confusion Matrix

	Training		Validation		Test	
Actual Revenue	Predicted Count FALSE	TRUE	Predicted Count FALSE	TRUE	Predicted Count FALSE	TRUE
FALSE	1053	118	626	73	1754	177
TRUE	289	660	149	417	123	257

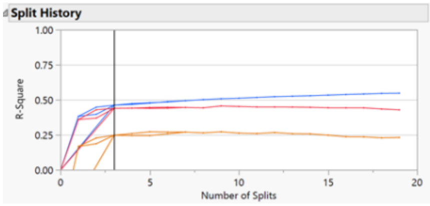
Area Under Curve (AUC): Training: 0.8992 Validation: 0.9038 Test: 0.8971

3.4.2 Decision Tree:

The tree with three splits is the one with the lowest complexity and high accuracy.
Confusion Matrix:

Confusion Matrix

	Training			Validation			Test	
Actual Revenue	Predicted Count FALSE	TRUE	Actual Revenue	Predicted Count FALSE	TRUE	Actual Revenue	Predicted Count FALSE	TRUE
FALSE	1017	154	FALSE	613	84	FALSE	1725	224
TRUE	166	780	TRUE	118	450	TRUE	73	308

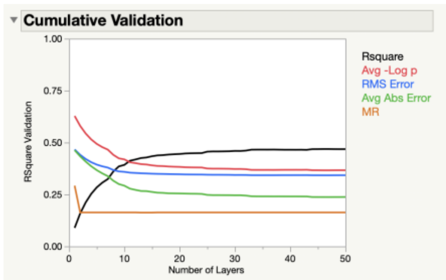


AUC: Training: 0.8965 Validation: 0.8867 Test: 0.9004

3.4.3 Boosted Tree:

The final boosted tree has 50 layers

Confusion Matrix:



Confusion Matrix

	Training			Validation			Test	
Actual Revenue	Predicted Count FALSE	TRUE	Actual Revenue	Predicted Count FALSE	TRUE	Actual Revenue	Predicted Count FALSE	TRUE
FALSE	1037	122	FALSE	593	88	FALSE	1676	242
TRUE	162	788	TRUE	116	452	TRUE	79	298

AUC: Training: 0.9428 Validation: 0.9181 Test: 0.9160

3.4.4 KNN:

The best number of K is 5 in terms of performance.

Confusion Matrix:

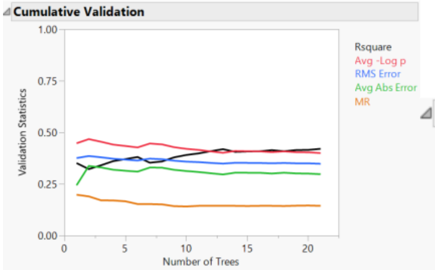
Confusion Matrix for Best K=5

	Training			Validation			Test	
Actual Revenue	Predicted Count FALSE	TRUE	Actual Revenue	Predicted Count FALSE	TRUE	Actual Revenue	Predicted Count FALSE	TRUE
FALSE	962	199	FALSE	591	98	FALSE	1643	296
TRUE	262	686	TRUE	130	437	TRUE	97	283

3.4.5 Bootstrap Forest:

The final number of trees in this forest 21, the number of terms sampled per split is 3.

Confusion Matrix:



Confusion Matrix

Training: Actual Revenue	Predicted Count FALSE	TRUE	Validation: Actual Revenue	Predicted Count FALSE	TRUE	Test: Actual Revenue	Predicted Count FALSE	TRUE
FALSE	1059	112	FALSE	617	82	FALSE	1706	225
TRUE	144	805	TRUE	99	467	TRUE	72	308

AUC: Training: 0.9662 Validation: 0.9229 Test: 0.9152

3.4.6 Neural Network:

The inputs are given by tweaking the hidden layer structure thereby changing the complexity of the model. It is observed that the most efficient model is the one with the complex hidden layer structure. ModelNTanH[1] NLinear[1] NGaussian[1]:

Confusion Matrix:

Confusion Matrix: Actual Revenue	Predicted Count FALSE	TRUE	Confusion Matrix: Actual Revenue	Predicted Count FALSE	TRUE	Confusion Matrix: Actual Revenue	Predicted Count FALSE	TRUE
FALSE	1017	136	FALSE	615	91	FALSE	1647	284
TRUE	190	758	TRUE	112	453	TRUE	74	308

AUC: Training:0.9310 Validation: 0.9202 Test: 0.9079

3.4.7 Naïve Bayes:

This method fits for datasets with more categorical variables.

Confusion Matrix:

Confusion Matrix

Training: Actual Revenue	Predicted Count FALSE	TRUE	Validation: Actual Revenue	Predicted Count FALSE	TRUE	Test: Actual Revenue	Predicted Count FALSE	TRUE
FALSE	948	223	FALSE	563	136	FALSE	1561	370
TRUE	285	664	TRUE	158	408	TRUE	112	268

AUC: Training: 0.8380 Validation: 0.8441 Test: 0.8415

Section 3.5 Assessment:

- Different models have been built based on original dataset, and equivalent balanced datasets are also created using Stratified Balanced Add-in.

- Among different assessment models, the criteria assumed to discern the models are Accuracy of model, Accuracy of 1s, Sensitivity, number of False Negatives, AUC, and misclassifications.

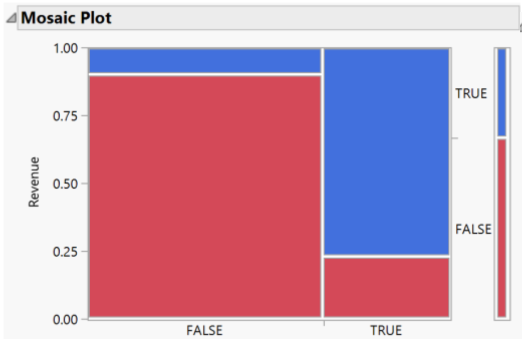
Balanced Dataset (Test Data)					
Model	**Confusion Matrix**		**Accuracy**	**Accuracy of 1**	**Sensitivity**
	FN	**FP**			
Logistic Regression	123	177	87.02%	59.22%	67.63%
Decision Tree	73	224	87.25%	57.89%	80.84%
Boosted Tree	79	242	86.01%	55.19%	79.04%
KNN	97	296	83.05%	48.88%	74.47%
Neural Network	74	284	84.52%	52.07%	80.60%
Bootstrap Forest	72	225	87.15%	57.79%	81.05%
Naive Bayes	112	370	79.14%	42.01%	70.53%

Balanced Data							
S. No.	**Model**	**AUC**			**Misclassification Rate**		
		Training	**Validation**	**Test**	**Training**	**Validation**	**Test**
1	Logistic Regression with predictor variables of high importance	89.92%	90.38%	89.71%	19.20%	17.55%	12.98%
2	Decision Tree	89.65%	88.67%	90.04%	15.12%	15.97%	12.75%
3	Boosted Tree	94.28%	91.81%	91.60%	13.47%	16.33%	13.99%
4	KNN with predictor variables of high importance	NA	NA	NA	21.86%	18.15%	16.95%
5	Bootstrap Forest with predictor variables of high importance	97.61%	93.21%	92.41%	10.19%	13.60%	12.59%
6	Neural Network with 1 TanH, 1 Linear and 1 Gaussian	93.10%	92.02%	90.79%	15.52%	15.97%	15.48%
7	Naïve Bayes	83.80%	84.41%	84.15%	23.96%	23.24%	20.86%

Section 4 Result:

Sensitivity is the most important metric for this business case as we are focused on identifying customers that are likely to purchase. On comparing sensitivity for all the 7 types of models built, we observe that bootstrap forest model has a better sensitivity. Decision tree and neural network also has very close sensitivity. Neural network is dropped from consideration due to its complexity and lack of interpretability which would make it difficult to draw insights on important variables and their relationship with target variable.

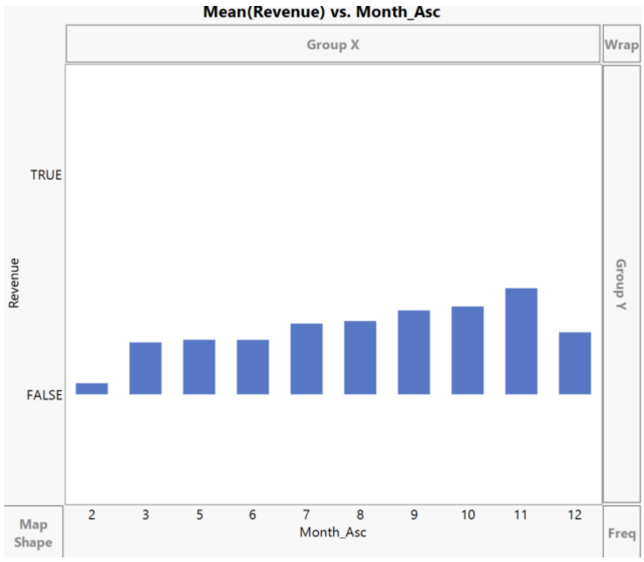
Decision tree and bootstrap forest models are compared on sensitivity, accuracy, AUC, RMSE, and **Decision Tree** is chosen as the final model because the improvement achieved in performance on all the metrics by bootstrap forest over decision tree model is very marginal at the loss of interpretability and increase in model complexity. The results of decision tree are visualized in the Mosaic Plot below, which shows a desirable accuracy of prediction.



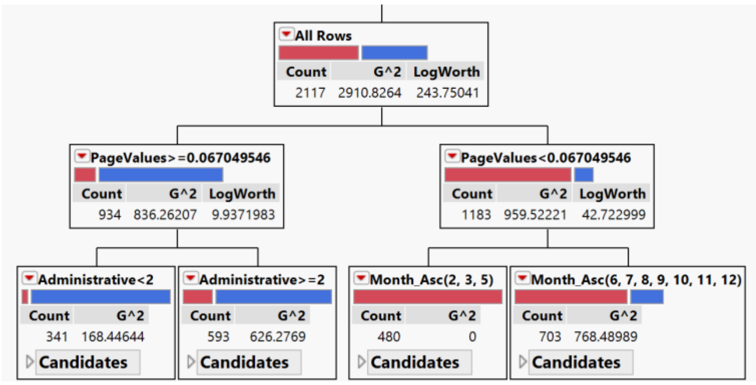
Column Contributions

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 Sl Transform ProductRelated_Duration	0	0	0.0000
Johnson Sl Transform ExitRates	0	0	0.0000

The most important variables are PageValues, Month, and Administrative. That is, if the average value for a web page that a user visited before completing an e-commerce transaction is higher, the probability of purchasing will be higher. Moreover, the probability of purchasing is also impacted by the month: there is an increasing trend of purchasing intention by month, except for the December. One possible reason is customers may concentrate their purchase during the Black Friday at the end of November.



Final model structure with interpretability:
This is the final tree model, which is simple and easy to explain. When facing the new data, it is easy to find the results.



Section 5 Conclusions and Recommendations

In the report, we conclude the Decision Model tree with 3 splits on the balanced dataset is the final choice. From the important variables, we understand that customers visiting administrative pages are less likely to purchase in that session. This could be because the customers have visited the website with the intention of tracking their account details (e.g.: tracking delivery of a previous purchase) and not to make a new purchase. Based on the Association Rule, the company can find inherent regularities in data by algorithms, for example, send customized messages and discount offers for the products that are most purchased or subsequently purchased to customers who have purchased on our website. For example, pop up advertisements and promotions of conditioners to those who have already purchased shampoos. This will help increase the purchasing probabilities of returning visitors.

Moreover, the company can optimize its strategies based on the goal value metric in the Google Analytics report. Given the knowledge that high page value is related to conversion, it could be a good idea to identify high-value pages with low traffic volume and driving more quality traffic to them to increase conversion rates and generate higher revenues. On the other hand, re-design the low-value pages with high traffic volume to increase qualities and then usability.

In addition, because purchase intentions of consumers are affected by months, in the lower sales months, the company can adopt more strategies, such as introducing new products, anti-season promotions, and price discrimination such as buy three get one free, or buy one and get one 50% off. In the months of higher sales, market competition may increase, and companies can adopt different strategies from other companies to increase competitiveness to attract and retain customers. For example, negotiate with the product suppliers to sell the product at the lowest price throughout the Internet to ensure the highest sales volume at low prices. If a company can monopolize the sale of a product on the web, it will get huge profits.

In total, the recommendation is to use the decision tree model to identify customers that are likely to purchase and target them with nudging techniques. We can conclude that using this model paired with the right strategy would result in higher revenue or profits through a better conversion rate. This will not only boost the profits of the company and prevent customer churn but also increase customer satisfaction and company reputation and brand value.

References

- 7 essential Google Analytics reports every marketer must know https://searchengineland.com/7-essential-google-analytics-reports-every-marketer-must-know-250412
- Analytics: What is Page Value in Google Analytics? https://yoast.com/what-is-page-value-in-google-analytics/
- Factors Influencing Online Shopping Behavior of Consumers https://www.searchfit.com/2017/05/01/factors-influencing-online-shopping-behavior-consumers/
- 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
- The Evolution and Impact Of Online Shopping. https://www.forbes.com/sites/mitsubishiheavyindustries/2017/02/02/the-evolution-and-impact-of-online-shopping/#58d874c834a3
- The Impact of Online Shopping On Society Information Technology Essay https://www.ukessays.com/essays/information-technology/the-impact-of-online-shopping-on-society-information-technology-essay.php

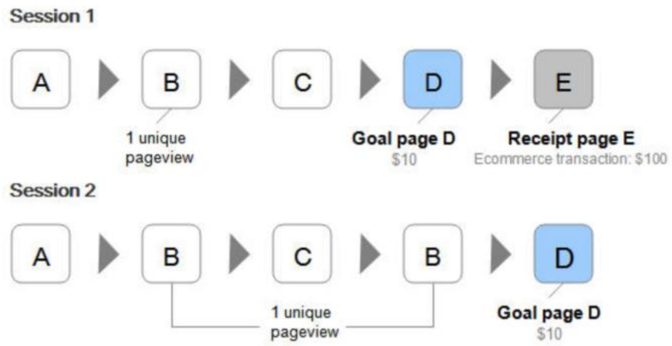
Appendix

The explanation of Page Value

Page Value is the average value for a page that a user visited before completing an Ecommerce transaction or landing on the goal page.

How Page Value is calculated

In each of two sessions below, the number of unique pageviews for Page B is 2. The Goal of Page D is also completed two times and Goal page D has a goal value of $10 with a sum of $20. In Session one, a transaction of $100 has taken place.



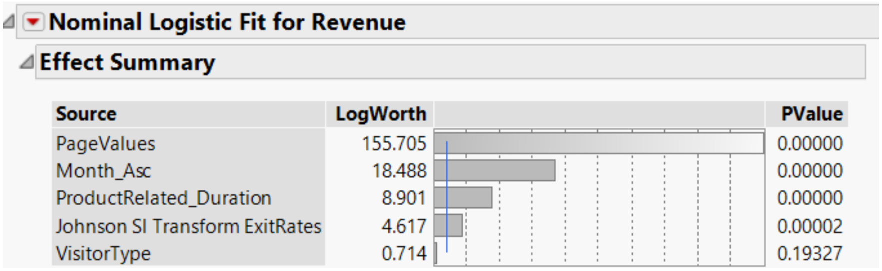
Calculation of Page Value of Page B:**(eCommerce Revenue + Total Goal Value) / Number of Unique Pageviews for Page B.** Therefore, ($100 + $20)/2, the page value of Page B is $60

Model Details

Partition and dataset: Training: 50%, Validation: 30%, and Test: 20%. The best model chosen for techniques is with the balanced data.

<u>Screenshots of the best models built using balanced data set:</u>

Model 1: Logistic regression.



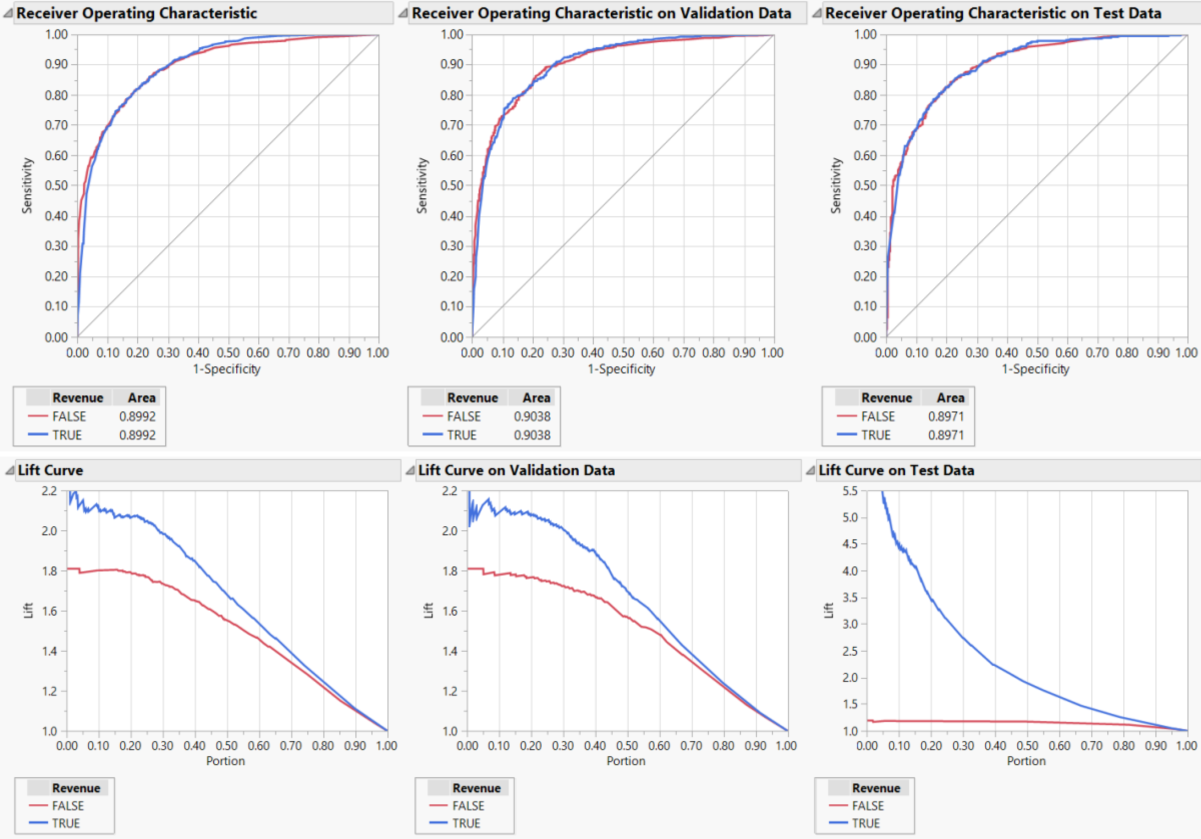
Nominal Logistic Fit for Revenue

Effect Summary

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

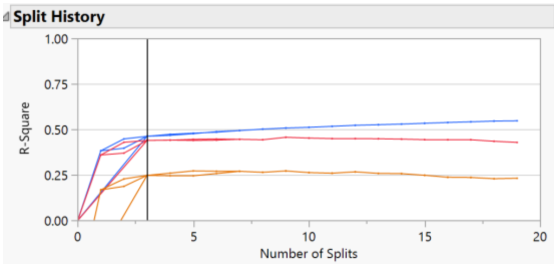
Confusion Matrix

Actual Revenue	Training Predicted Count FALSE	Training Predicted Count TRUE	Validation Predicted Count FALSE	Validation Predicted Count TRUE	Test Predicted Count FALSE	Test Predicted Count TRUE
FALSE	1053	118	626	73	1754	177
TRUE	289	660	149	417	123	257



Model 2: Decision Tree

In the new dataset we have 6001 rows. Use 3 splits because the performance did not increase much after 3 splits.



Confusion Matrix

Actual Revenue	Training Predicted Count FALSE	Training Predicted Count TRUE	Validation Predicted Count FALSE	Validation Predicted Count TRUE	Test Predicted Count FALSE	Test Predicted Count TRUE
FALSE	1017	154	613	84	1725	224
TRUE	166	780	118	450	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	$\sum$ -Log(ρ[j])/n
RMSE	0.3439	0.3521	0.3256	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2370	0.2466	0.2110	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.1512	0.1597	0.1275	$\sum$ (ρ[j]≠ρMax)/n
N	2117	1265	2330	n

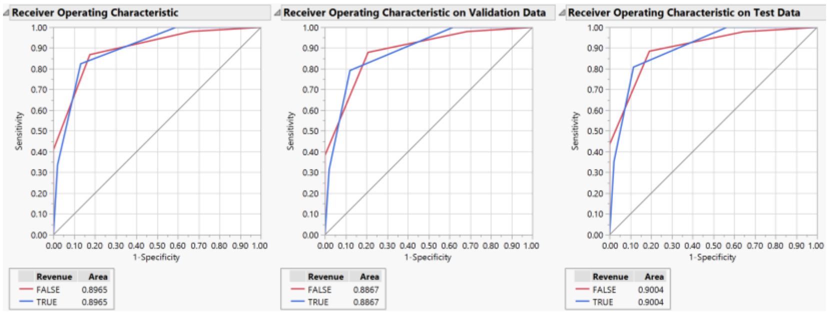
Leaf Report

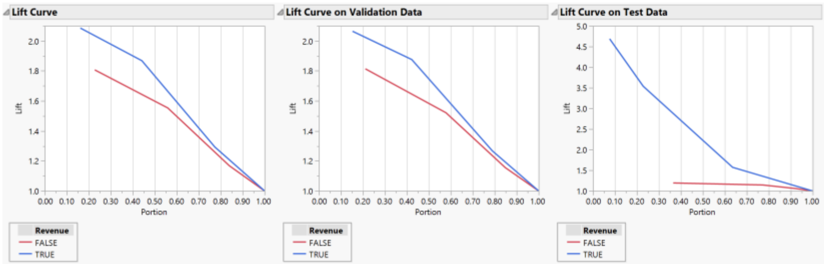
Response Prob

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





Model 3: Boosted Tree

Column Contributions

Term	Number of Splits	G^2	Portion
Month_Asc	27	53016.2137	0.2444
PageValues	30	44233.1877	0.2039
Informational_Duration_Cat	15	27238.8341	0.1256
VisitorType	9	18926.4151	0.0873
Browser	5	16878.4048	0.0778
OperatingSystems	6	14222.8606	0.0656
Johnson Sl Transform ExitRates	14	12235.4795	0.0564
Administrative_Duration_Cat	14	9926.28712	0.0458
TrafficType	4	8951.64499	0.0413
Weekend	18	6437.6542	0.0297
Region	3	3037.13345	0.0140
Administrative	3	1254.66163	0.0058
Johnson Sl Transform ProductRelated_Duration	2	551.331591	0.0025
Informational	0	0	0.0000
SpecialDay	0	0	0.0000

Remove informational and special day.

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 Sl 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

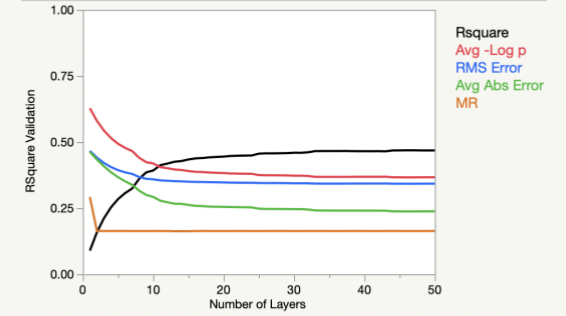
Confusion Matrix

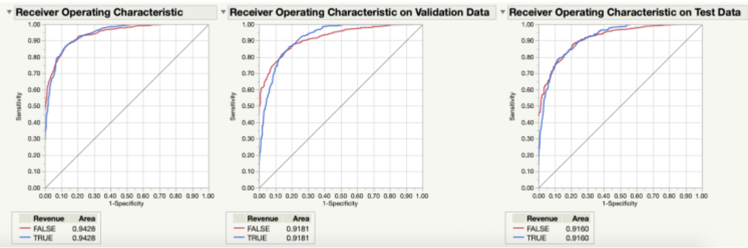
Training			Validation			Test		
Actual	Predicted Count		Actual	Predicted Count		Actual	Predicted Count	
Revenue	FALSE	TRUE	Revenue	FALSE	TRUE	Revenue	FALSE	TRUE
FALSE	1037	122	FALSE	593	88	FALSE	1676	242
TRUE	162	788	TRUE	116	452	TRUE	79	298

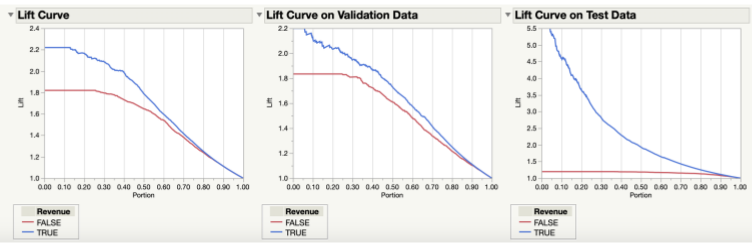
Overall Statistics

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







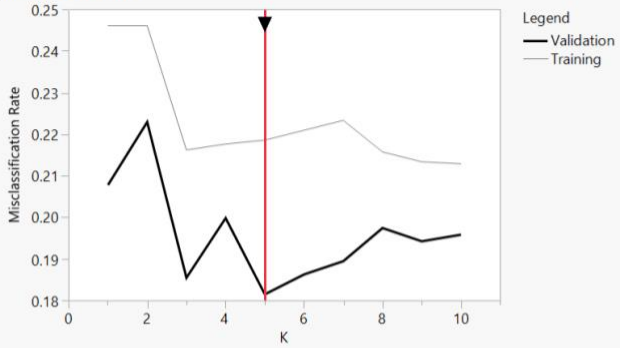
Model 4: KNN

Optimal predictors are page values, month_asc, Johnson Product Dur, Admin Dur cat, Info Dur cat. Using only these variables and generating the model below

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

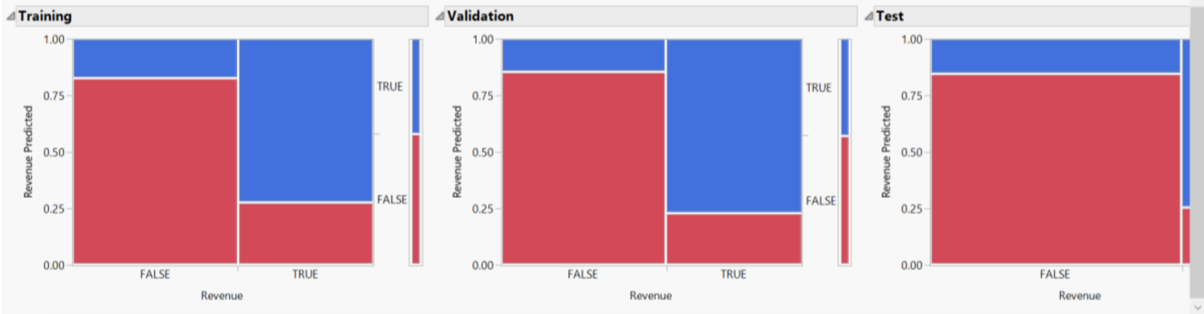
Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	591	98
TRUE	130	437

Test

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

Mosaic Plot

Mosaic Plot for K=5



Model 5: 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	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
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

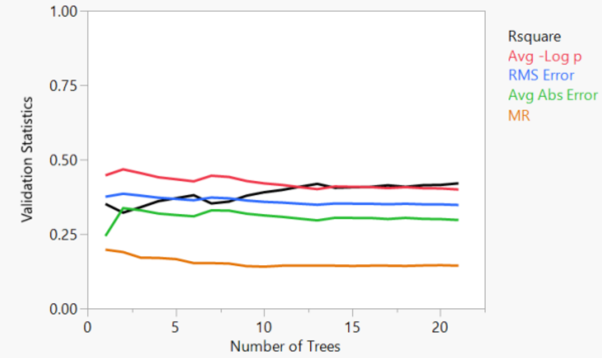
Validation

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

Test

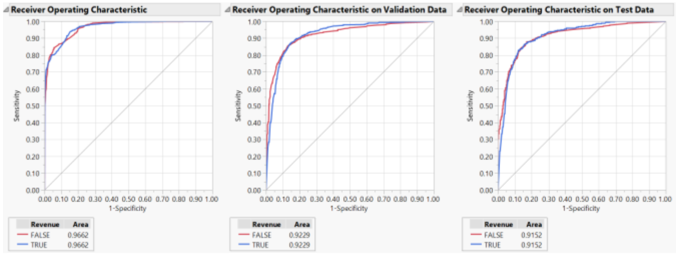
Actual Revenue	Predicted Count FALSE	Predicted Count TRUE
FALSE	1706	225
TRUE	72	308

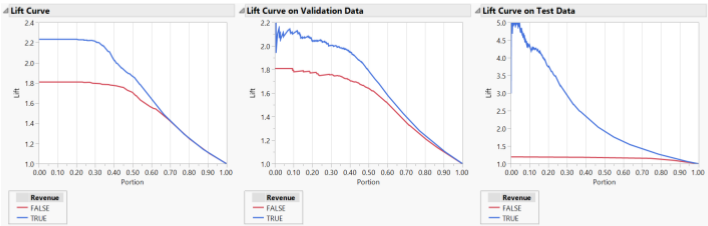
Cumulative Validation



Column Contributions

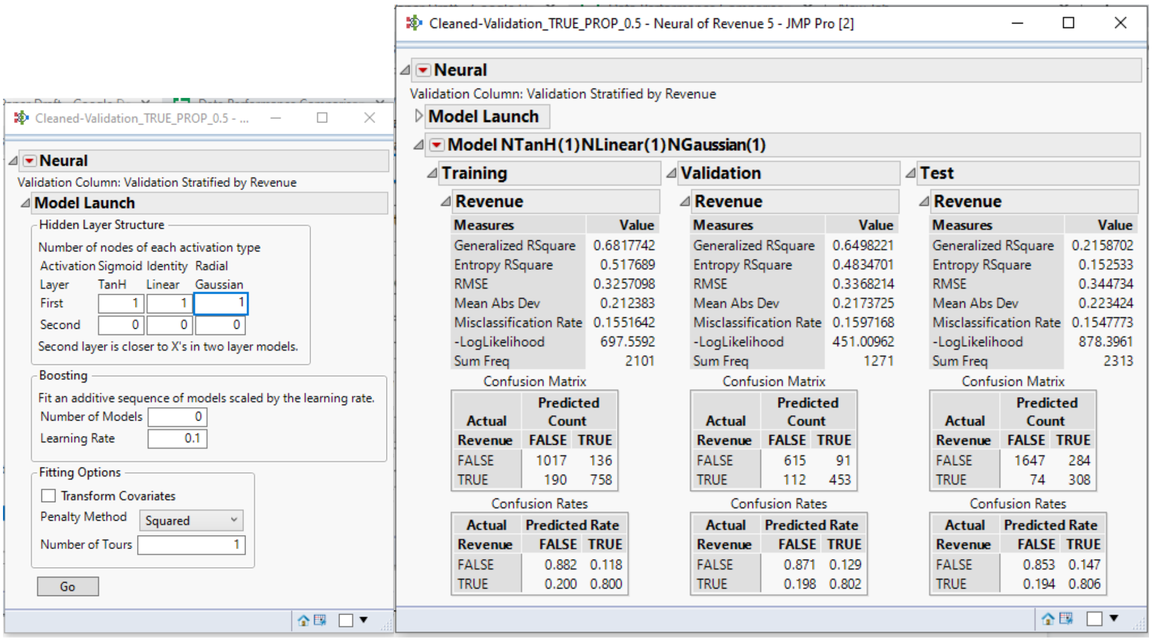
Term	Number of Splits	G^2	Portion
PageValues	210	458.154875	0.4802
Johnson Sl Transform ProductRelated_Duration	263	110.917306	0.1163
Johnson Sl Transform ExitRates	255	100.316693	0.1052
Month_Asc	161	77.3096502	0.0810
Administrative	218	52.8165003	0.0554
TrafficType	175	28.0287182	0.0294
Region	174	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





Model 6: Neural Networks

1 TanH, 1 Linear and 1 Gaussian



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	TRUE
FALSE	1017	136
TRUE	190	758

Confusion Rates

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

Confusion Rates

Actual Revenue	Predicted Rate FALSE	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

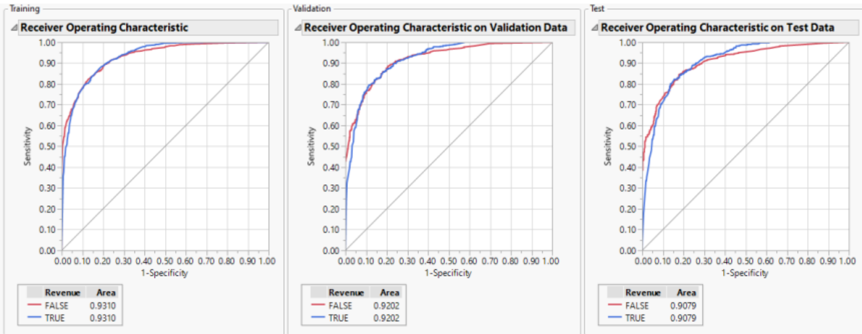
Confusion Matrix

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

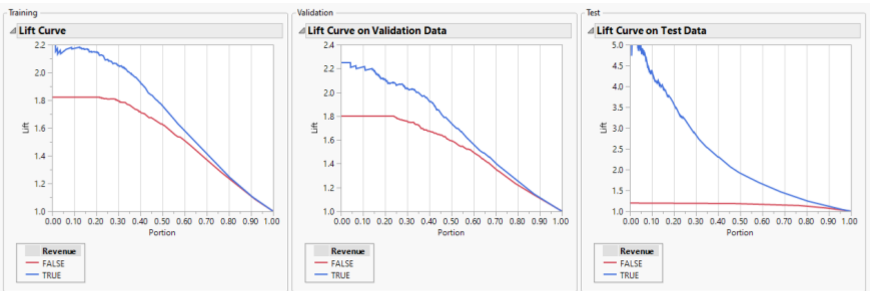
Confusion Rates

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

ROC Curve:



Lift Curve:



Model 7 :Naive Bayes

Confusion Matrix

Training

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

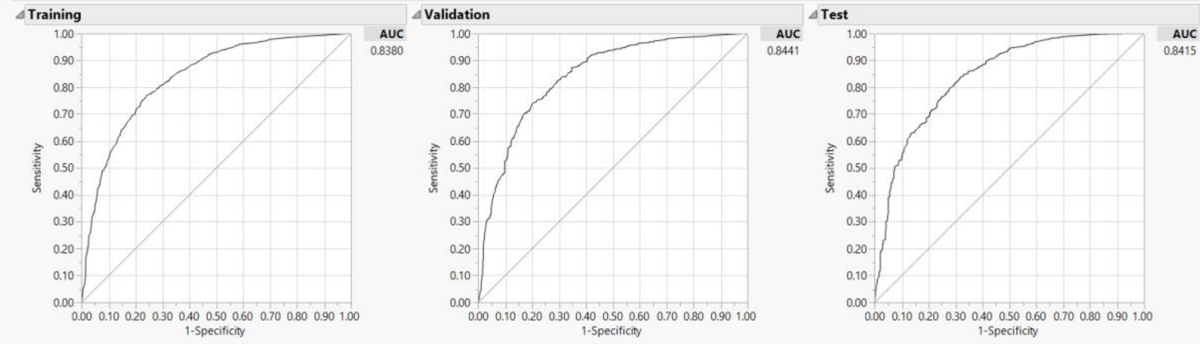
Validation

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