CONTENTS

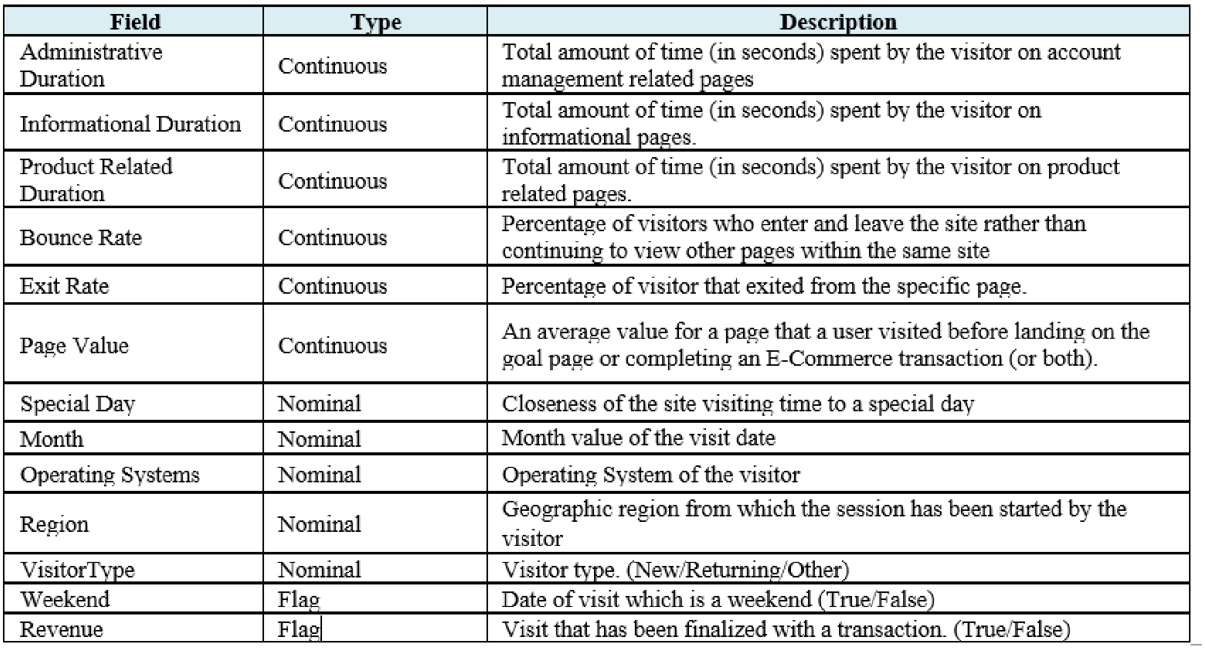
SECTION A: INTRODUCTION

In recent years, rapid growth of online purchase had created a new business channel for the brick-and-mortar retailers. As the internet is a cost effective medium to interact with and connect to the customers, many retailers are listing their products and services on their online store on the internet. This has resulted in cut-throat competition among the online retailers because shoppers are now able to compare the price easily on the internet without visiting the individual stores in town. Most importantly, many companies were spending significant budget digital marketing advertisement. To stay competitive on the internet, it is important for retailers to predict which prospective web visitors would tend to purchase. This allows retailers to optimise their marketing strategies for better positioning to their target audience. For instance, to launch specific month promotion campaign or targeted Electronic Direct Mail (EDM) marketing contents to specific prospective consumers to click, visit and purchase.

Through predictive modelling, netnography of online consumers were considered to better predict which prospective web consumers tend to purchase. Therefore, companies can better position their marketing efforts to increase revenue.

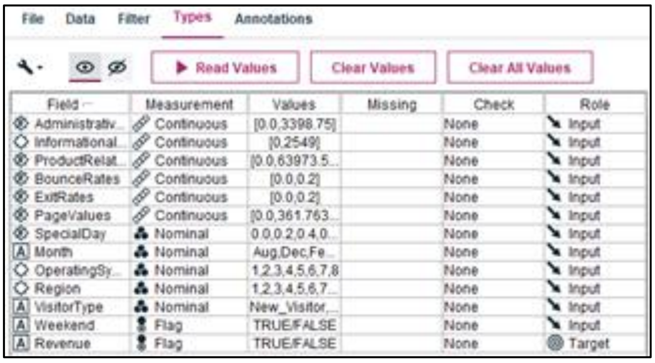
SECTION B: DATA UNDERSTANDING

There was a total of 12,330 records used in this model with 84.5% negative cases (10,422 records) and 15.5% positive cases (1,908 records). This result showed that online consumers were generally cautious of their purchase online. By considering the variables with more predictive impact on Revenue, we have selected the fields below. We have intentionally removed the other fields not listed due to logical reasoning.



Field	Type	Description
Administrative Duration	Continuous	Total amount of time (in seconds) spent by the visitor on account management related pages
Informational Duration	Continuous	Total amount of time (in seconds) spent by the visitor on informational pages.
Product Related Duration	Continuous	Total amount of time (in seconds) spent by the visitor on product related pages.
Bounce Rate	Continuous	Percentage of visitors who enter and leave the site rather than continuing to view other pages within the same site
Exit Rate	Continuous	Percentage of visitor that exited from the specific page.
Page Value	Continuous	An average value for a page that a user visited before landing on the goal page or completing an E-Commerce transaction (or both).
Special Day	Nominal	Closeness of the site visiting time to a special day
Month	Nominal	Month value of the visit date
Operating Systems	Nominal	Operating System of the visitor
Region	Nominal	Geographic region from which the session has been started by the visitor
VisitorType	Nominal	Visitor type. (New/Returning/Other)
Weekend	Flag	Date of visit which is a weekend (True/False)
Revenue	Flag	Visit that has been finalized with a transaction. (True/False)

Before modelling, we have prepared the following dataset “Online Shoppers Purchasing Intention Dataset.csv” by defining the roles and measurement for each field as showed in Figure 1. Revenue was set as “Target” so that all the predictive modelling in this study could be carried out.



Field	Measurement	Values	Missing	Check	Role
Administrativ...	Continuous	[0.0,3398.75]		None	Input
Informational...	Continuous	[0,2549]		None	Input
ProductRelat...	Continuous	[0.0,63973.5...		None	Input
BounceRates	Continuous	[0.0,0.2]		None	Input
ExitRates	Continuous	[0.0,0.2]		None	Input
PageValues	Continuous	[0.0,361.763...		None	Input
SpecialDay	Nominal	0.0,0.2,0.4,0...		None	Input
Month	Nominal	Aug,Dec,Fe...		None	Input
OperatingSy...	Nominal	1,2,3,4,5,6,7,8		None	Input
Region	Nominal	1,2,3,4,5,6,7...		None	Input
VisitorType	Nominal	New_Visitor,...		None	Input
Weekend	Flag	TRUE/FALSE		None	Input
Revenue	Flag	TRUE/FALSE		None	Target

Figure 1: Measurement and Role of Attributes

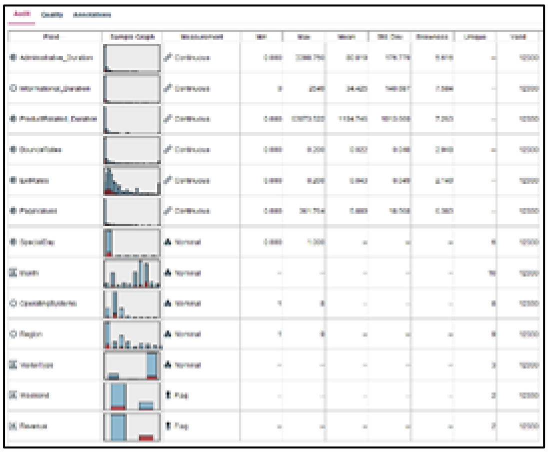
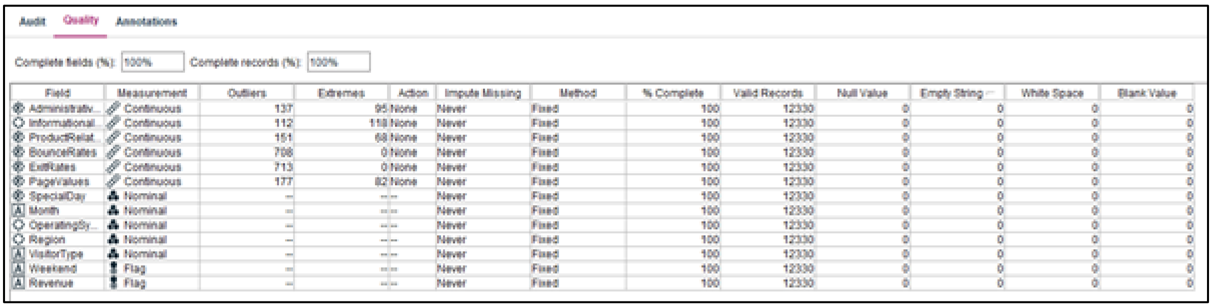


Figure 2: Data Audit – Audit

We also carried out Data Audit (Figure 2) to confirm the data quality for any missing values, outliers, and extremes. From Figure 3, it is evident that there are no missing/null values, but outliers and extremes were present on Administrative_Duration, Informational_Duration, ExitRates, PageValues ProductRelated_Duration, BounceRates. The number of outliers and extreme values are small and not significant relative to the data size of 12,330. Furthermore, Decision Trees such as Random Forest was robust to outliers, extreme outliers, and missing values. Hence, no data cleaning was performed to these values.



Complete fields (%): 100%
Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
Administrativ...	Continuous	137	95	None	Never	Fixed	100	12330	0	0	0	0
Informational...	Continuous	112	118	None	Never	Fixed	100	12330	0	0	0	0
ProductRelat...	Continuous	151	68	None	Never	Fixed	100	12330	0	0	0	0
BounceRates	Continuous	708	0	None	Never	Fixed	100	12330	0	0	0	0
ExitRates	Continuous	713	0	None	Never	Fixed	100	12330	0	0	0	0
PageValues	Continuous	177	82	None	Never	Fixed	100	12330	0	0	0	0
SpecialDay	Nominal	--	--	--	Never	Fixed	100	12330	0	0	0	0
Month	Nominal	--	--	--	Never	Fixed	100	12330	0	0	0	0
OperatingSy...	Nominal	--	--	--	Never	Fixed	100	12330	0	0	0	0
Region	Nominal	--	--	--	Never	Fixed	100	12330	0	0	0	0
VisitorType	Nominal	--	--	--	Never	Fixed	100	12330	0	0	0	0
Weekend	Flag	--	--	--	Never	Fixed	100	12330	0	0	0	0
Revenue	Flag	--	--	--	Never	Fixed	100	12330	0	0	0	0

Figure 3: Data Audit – Quality

SECTION C: BASELINE MODEL ILLUSTRATION

In this report, we have generated the CART model (using default settings) as the baseline model. The variables defined as the inputs and target were as shown in Figure 1. For the build of the baseline model, no pre-processing was done to remove outliers and extreme values due to CART’s robust nature. In addition, the dataset was not partitioned into training and testing sets for the model build. This means that the entire dataset was utilised when training the baseline model.

From the baseline model, the decision tree and decision rules were reflected in Figure 4 and Figure 6, respectively. With reference to Figure 5, the baseline model's accuracy of 90.06% and AUC of 0.842 (which is quite close to 1) implied a generally good accuracy of the model's performance. However, the Gini value of 0.685 suggested that splitting of the nodes were relatively impure and not homogeneous. With reference to Figure 7, the hit rates are 92.23% and 55.71% for "FALSE" and "TRUE" revenue cases, respectively. This suggests that the performance of the baseline was not consistent for predicting "TRUE" cases. With reference to Figure 8, this limitation could be a result of having a small proportion of "TRUE" cases (15%) in the original dataset relative to the number of "FALSE" cases being used to train the model.

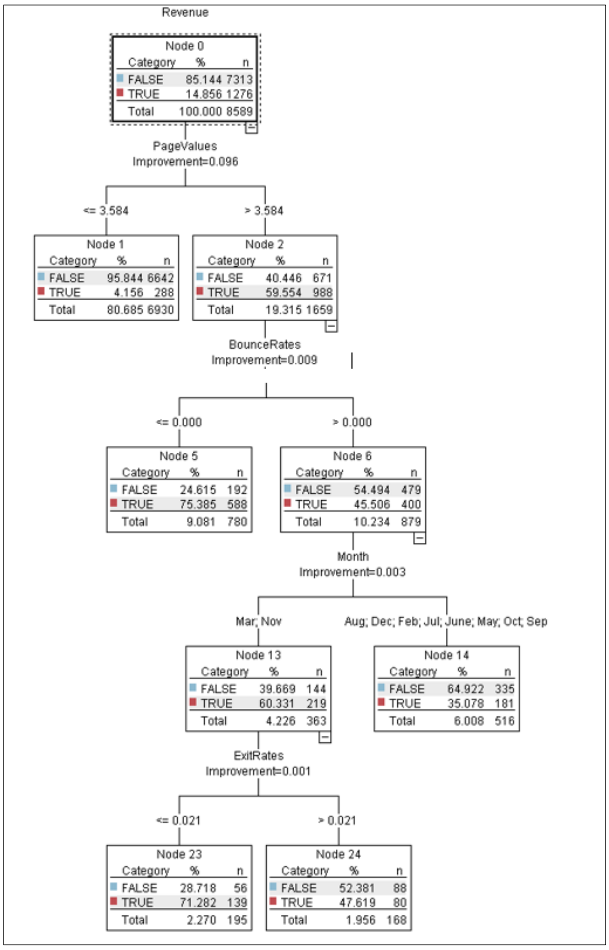
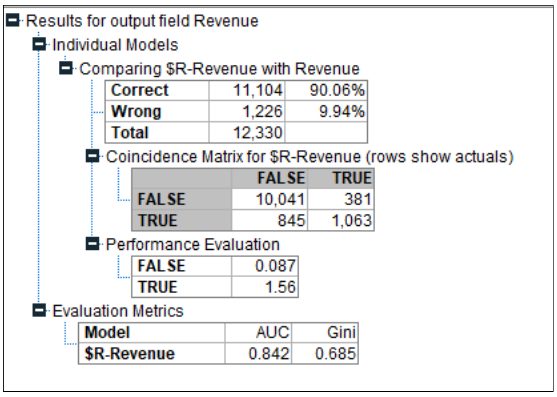


Figure 4: Baseline Model (CART – Default)



Results for output field Revenue

Individual Models

Comparing $R-Revenue with Revenue

Correct	11,104	90.06%
Wrong	1,226	9.94%
Total	12,330	

Coincidence Matrix for $R-Revenue (rows show actuals)

	FALSE	TRUE
FALSE	10,041	381
TRUE	845	1,063

Performance Evaluation

FALSE	0.087
TRUE	1.56

Evaluation Metrics

Model	AUC	Gini
$R-Revenue	0.842	0.685

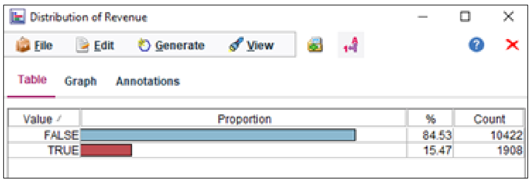
Figure 5: Baseline Model Performance

Node 1	If PageValuesImprovement <= 3.584, Revenue = FALSE
Node 5	If Page Values Improvement > 3.584 and BounceRates Improvement <= 0.000, Revenue = TRUE
Node 14	If Page Values Improvement > 3.584 and BounceRates Improvement > 0.000 and Month Improvement = Aug; Dec; Feb; Jul; June; May; Oct; Sep, Revenue = FALSE
Node 23	If Page Values Improvement > 3.584 and BounceRates Improvement > 0.000 and Month Improvement = Mar;Nov and ExitRates <= 0.021, Revenue = TRUE
Node 24	If Page Values Improvement > 3.584 and BounceRates Improvement > 0.000 and Month Improvement = Mar;Nov and ExitRates > 0.021, Revenue = FALSE

Figure 6: Decision Rules for Baseline

Hit Rate (FALSE)	10,041 / (10,041 + 845) = 92.23%
Hit Rate (TRUE)	1063 / (845 + 1063) = 55.71%

Figure 7: Hit Rate for Baseline Model



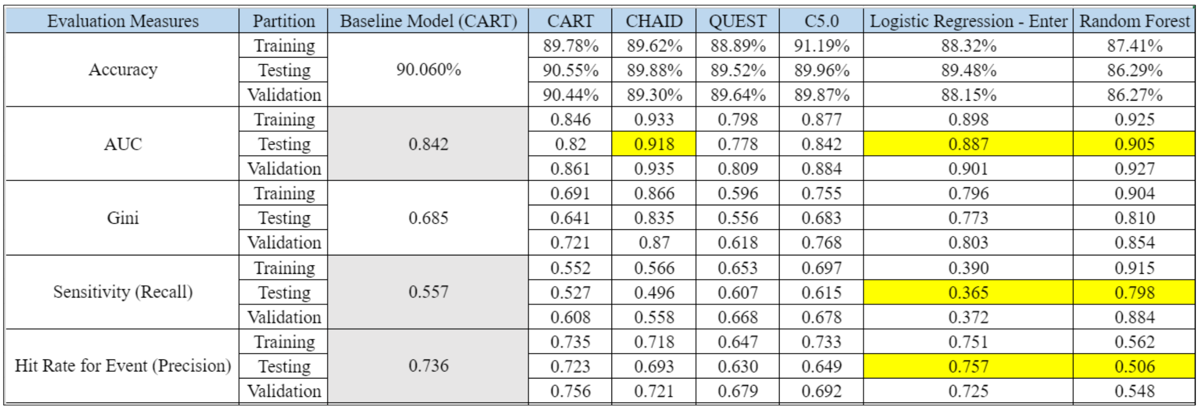
Value	Proportion	%	Count
FALSE		84.53	10422
TRUE		15.47	1908

Figure 8: Distribution of TRUE/FALSE in Dataset

SECTION D: NEW MODEL DESIGN & RESULT ILLUSTRATION

Data characteristics as well as the problem statement have driven our decision on the selection of the most relevant model from a multitude of models developed. With the objective of predicting online shopper's purchasing intention, the dependent variable of this analysis was binary - Yes / No. Multiple decision tree models applicable to binary dependent variable were explored; Baseline Model (CART without partition), CART, CHAID, QUEST, C5.0 & Random Forest. In addition, logistic regression has also been constructed. Modelling evaluations have been performed to identify the most appropriate model.

To ensure data consistency across all models, the same data preparation steps were performed for all 6 models (except for Baseline Model). Firstly, five inputs (Administrative, Informational, ProductRelated, Browser and TrafficType) were removed by logical reasoning that they were irrelevant. Secondly, both Weekend and Revenue variables were changed into 'Flag' measurement, while Revenue variable was indicated as the 'Target' role. The dataset was then partitioned into Training set (70%), Testing set (20%) and Validation set (10%).



Evaluation Measures	Partition	Baseline Model (CART)	CART	CHAID	QUEST	C5.0	Logistic Regression - Enter	Random Forest
Accuracy	Training	90.060%	89.78%	89.62%	88.89%	91.19%	88.32%	87.41%
	Testing		90.55%	89.88%	89.52%	89.96%	89.48%	86.29%
	Validation		90.44%	89.30%	89.64%	89.87%	88.15%	86.27%
AUC	Training	0.842	0.846	0.933	0.798	0.877	0.898	0.925
	Testing		0.82	0.918	0.778	0.842	0.887	0.905
	Validation		0.861	0.935	0.809	0.884	0.901	0.927
Gini	Training	0.685	0.691	0.866	0.596	0.755	0.796	0.904
	Testing		0.641	0.835	0.556	0.683	0.773	0.810
	Validation		0.721	0.87	0.618	0.768	0.803	0.854
Sensitivity (Recall)	Training	0.557	0.552	0.566	0.653	0.697	0.390	0.915
	Testing		0.527	0.496	0.607	0.615	0.365	0.798
	Validation		0.608	0.558	0.668	0.678	0.372	0.884
Hit Rate for Event (Precision)	Training	0.736	0.735	0.718	0.647	0.733	0.751	0.562
	Testing		0.723	0.693	0.630	0.649	0.757	0.506
	Validation		0.756	0.721	0.679	0.692	0.725	0.548

Figure 9: Model Performance Evaluation on Various Models

The five evaluation measures performed on the seven models were summarised in Figure 9 for easy performance comparison. The overall accuracy of all Testing set was in a narrow range (86.29% to 90.55%). Their corresponding high overall accuracy (about 90%) from Validating set validated the absence of overfitting and the use of a high 70% training set partition for all models (except no partitioning for Baseline Model).

CHAID, Logistic Regression and Random Forest outperformed Baseline Model with a higher AUC (>0.88 vs 0.842). These models also outperformed the Baseline Model in the Gini coefficient (0.835, 0.773 and 0.810 vs 0.685). All three models being closer to the perfect model (Gini coefficient of 1) than a random model (0) was good at distinguishing the true positive (those predicted to purchase who actually purchased) from the true negative (those predictive not to purchase who actually did not purchase) (Peñaloza, 2016). Recall and Precision were the most important measures for three reasons:

imbalanced dataset (Tao, 2020), binary classifier (Johnson, 2016) and the nature of the problem statement.

In terms of Precision, not only did Random Forest (0.506) being outperformed by Baseline Model (0.736) and Logistic Regression (0.757) but also by all other models. For Logistic Regression, it meant that for every 1000 predicted to purchase, 757 actually purchased in the future. This made Random Forest seemed to pale in comparison, in fact slightly better than tossing a coin. However, should one inspect on Recall, Random Forest (0.798) outperformed every model including the Baseline Model (0.557). Normally one could accept a trade-off between Recall and Precision but with respect to the business problem to improve revenue, Recall should be the primary measure. Comparison between two best models with greatest Recall-Precision Curve, Random Forest and Baseline Model could add clarity. For instance, for every 1000 actual purchasers, Random could correctly predict 798 vs Baseline Model (557). This meant 43.27% (798/557 -1) more revenue for Random Forest which could translate to a sizable absolute revenue. The inverse of Precision for both Random Forest and Baseline Model were 1.976 and 1.359. These were equivalent to marketing and sales effort spent on each True Positive event by acting upon the actionable insights of the models. Thus, Random Forest would expend (1.976 x 798) / (1.359 x 557) = 2.085 marketing and sales effort than Baseline Model. This was the trade-off for Random Forest to secure 43.27% more revenue than Baseline Model. The assumption that Random Forest was the best predictive model was that the trade-off would still result in higher profitability than Baseline Model even though it has the highest absolute Recall (highest revenue). A limitation of this study was that the monetary cost of marketing and sales effort was unknown.

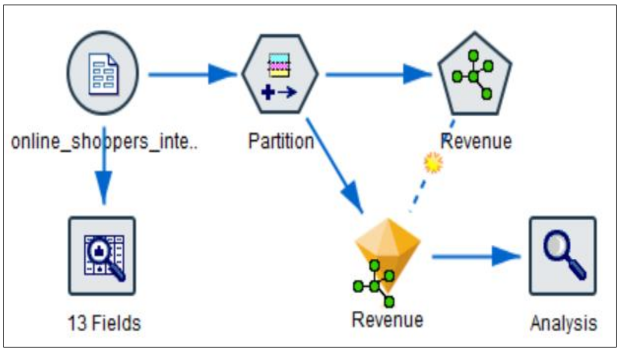


Figure 10: SPSS Modeler – Random Forest

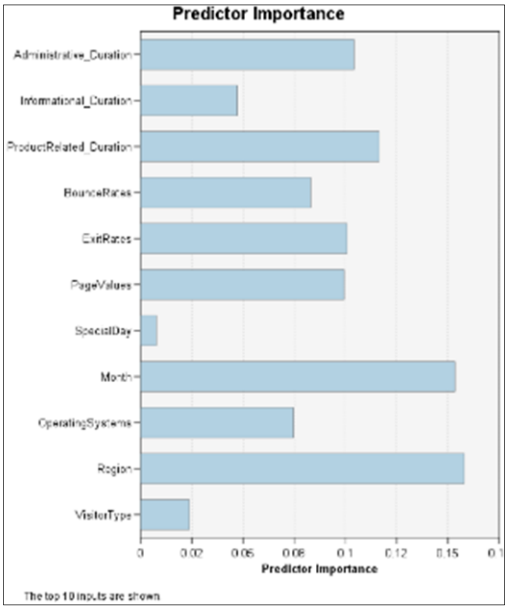
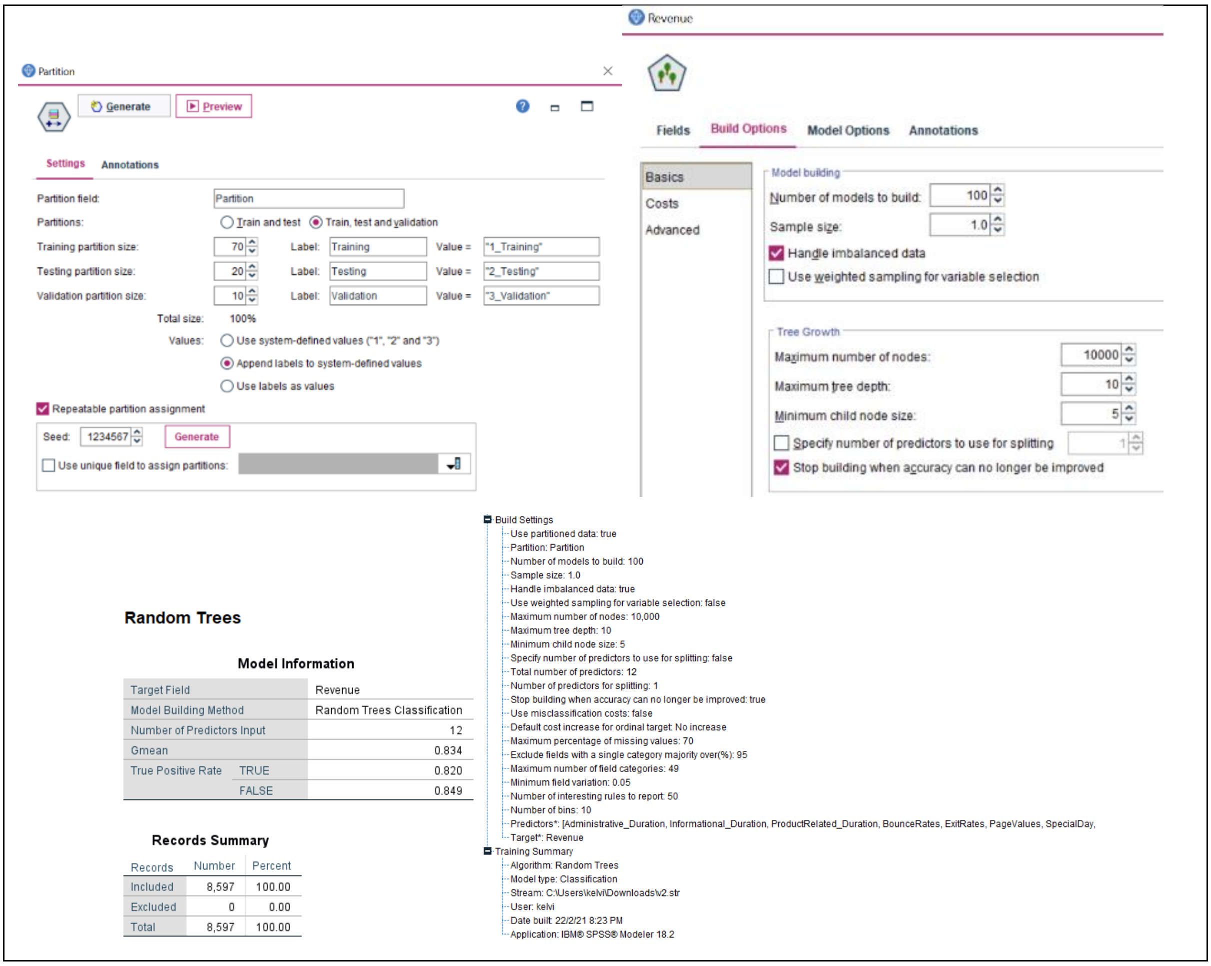


Figure 11: Predictor Importance – Random Forest

The final model proposed was the Random Forest which has been constructed using SPSS Modeler (illustrated in Figure 10 above). The parameter settings used to build the Random Forest model in the SPSS Modeler are shown in Figure 12 below.



Partition

Partition field: Partition

Partitions: ○ Train and test ● Train, test and validation

Training partition size:	70	Label:	Training	Value =	"1_Training"
Testing partition size:	20	Label:	Testing	Value =	"2_Testing"
Validation partition size:	10	Label:	Validation	Value =	"3_Validation"

Total size: 100%

Values: ○ Use system-defined values ("1", "2" and "3")
● Append labels to system-defined values
○ Use labels as values

☑ Repeatable partition assignment

Seed: 1234567 Generate

□ Use unique field to assign partitions:

Revenue

Model building

Number of models to build: 100

Sample size: 1.0

☑ Handle imbalanced data

□ Use weighted sampling for variable selection

Tree Growth

Maximum number of nodes: 10000

Maximum tree depth: 10

Minimum child node size: 5

□ Specify number of predictors to use for splitting 1

☑ Stop building when accuracy can no longer be improved

Random Trees

Model Information

Target Field		Revenue
Model Building Method		Random Trees Classification
Number of Predictors Input		12
Gmean		0.834
True Positive Rate	TRUE	0.820
	FALSE	0.849

Records Summary

Records	Number	Percent
Included	8,597	100.00
Excluded	0	0.00
Total	8,597	100.00

```
Build Settings
    Use partitioned data: true
    Partition: Partition
    Number of models to build: 100
    Sample size: 1.0
    Handle imbalanced data: true
    Use weighted sampling for variable selection: false
    Maximum number of nodes: 10,000
    Maximum tree depth: 10
    Minimum child node size: 5
    Specify number of predictors to use for splitting: false
    Total number of predictors: 12
    Number of predictors for splitting: 1
    Stop building when accuracy can no longer be improved: true
    Use misclassification costs: false
    Default cost increase for ordinal target: No increase
    Maximum percentage of missing values: 70
    Exclude fields with a single category majority over(%): 95
    Maximum number of field categories: 49
    Minimum field variation: 0.05
    Number of interesting rules to report: 50
    Number of bins: 10
    Predictors*: [Administrative_Duration, Informational_Duration, ProductRelated_Duration, BounceRates, ExitRates, PageValues, SpecialDay,
    Target*: Revenue
Training Summary
    Algorithm: Random Trees
    Model type: Classification
    Stream: C:\Users\kelvi\Downloads\v2.str
    User: kelvi
    Date built: 22/2/21 8:23 PM
    Application: IBM® SPSS® Modeler 18.2
```

Figure 12: Screenshots of Parameter Settings in SPSS Modeler (Random Forest)

Figure 11 (on page 7) illustrated the predictor importance of the inputs. The top five inputs with the greatest contribution to the accuracy of Random Forest were Region, followed by Month, ProductRelated_Duration, Administrative_Duration and ExitRates in decreasing importance. As such, the company needs to ensure these five input fields do not have missing values while SpecialDay and VisitorType may be ignored to further tune the model.

SECTION E: DISCUSSION & CONCLUDING REMARK

In conclusion, while the Baseline Model exhibited an overall accuracy of 90.06% and AUC of 0.842, its performance fell short in aspects such as Recall (0.557) and Gini (0.685). As such, 6 different models were generated and a comparison on their performance were conducted. For the data preparation, five irrelevant inputs were removed from the analysis. The dataset was then partitioned into Training (70%), Testing (20%) and Validation (10%) sets.

As the objective was to predict which prospective web visitors tend to purchase, we have proposed the use of the Random Forest and have accepted the model’s trade-off on the interpretability. Random Forest had outperformed the Baseline Model with a significantly higher AUC (0.905) and Recall (0.798), implying a better predictive accuracy and higher absolute revenue. This therefore empowered companies in predicting which prospective web visitors tend to purchase so that they could optimise their marketing strategies to improve revenue.

REFERENCES

American Cancer Society. (2019, February 17). *How Common Is Breast Cancer?* Retrieved from American Cancer Society: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

IBM. (2020). *IBM Knowledge Center: Verifying Data Quality* . Retrieved from: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_18.2.0/modeler_crispdm_ddita/clementine/crisp_help/crisp_verify_data_quality.html

IBM Cloud Education. (2020, December 7). *Random Forest*. Retrieved from IBM: https://www.ibm.com/cloud/learn/random-forest

Johnson, M. K. (2016). *Applied Predictive Modeling.* Michigan: Springer.

Peñaloza, R. (2016). *Gini Coefficient for Ordinal Categorical Data.* Brasília: University of Brasília.

Tao, C. (2020, February 26). *How to Evaluate a Classification Machine Learning Model*. Retrieved from Towards Data Science: https://towardsdatascience.com/how-to-evaluate-a-classification-machine-learning-model-d81901d491b1

Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, *31*(10), 6893-6908. Retrieved from https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset