Kennesaw State University
IS 8935 BI: Traditional & Big Data Analytics
Dr. Reza Vaezi
Assignment 10
March 28, 2019

Cross Validation of Decision Tree Model for eReader Adoption

By Karis Kim

Executive Summary

The goal of this report is to evaluate the performance of the decision tree model in predicting the label eReader-Adoption category to find Innovator and Early Adopter using different parameters. A cross validation analysis revealed that the overall accuracy of the model was 58.85%, with a precision of 52.70% for Innovator prediction, 51.28% for Early Adopter prediction, and a recall of 39.80% for Innovator and 68.29% for Early adopter. The top/best predictor attributes are Website_Actvity and Age. The report recommends targeting marketing efforts for the next-gen eReaders to customers with more than regular website activity who are less than 60 years of age.

CONTENTS

BUSINESS UNDERSTANDING

The business is launching the next generation electronic book readers called the eReader. In order to target marketing to the people who are most likely to respond to ads and purchase the eReaders, the business is seeking to find which customers will buy the eReader early based on data from previous generation electronic book reader purchases. To identify early predictors of buying behavior and predict which customers will be early adopters of the next generation eReader, the business plans to employ a decision tree analysis. The goal of this report is to further empower the business by utilizing the cross-validation model to evaluate the accuracy of the decision tree model using different parameters.

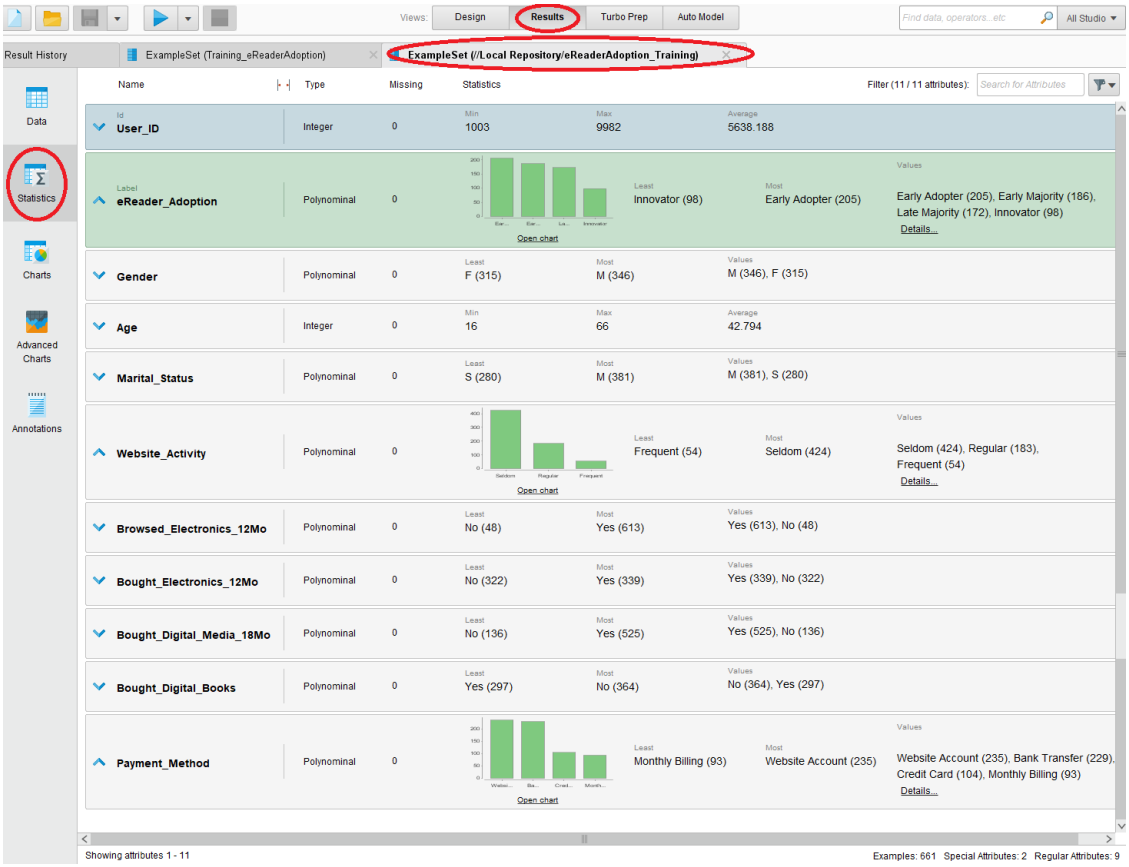
DATA UNDERSTANDING

The eReaderadoption_training data set contains 661 examples and 11 attributes. The *User_ID* attribute is an Id attribute that will not be considered in the decision tree analysis, and the *eReader_Adoption* attribute will be the label or predictor attribute.

ATTRIBUTE INFORMATION

	Attribute	Description
1	User_ID	Numeric unique ID for each customer with an account
2	eReader_Adoption	Label attribute. Based on customers who purchased previous-gen eReader. 1) **Innovator**: purchased within 1 wk of release 2) **Early Adopter**: purchased within 2-3 wk of release 3) **Early Majority**: purchased within 4-8 wk of release 4) **Late Majority**: purchased after 8 wks of release
3	Gender	M = male, F = female
4	Age	Age calculated by system date and recorded DOB
5	Marital_Status	M = married, S = single
6	Website_Activity	Categorized customer's website activity: Seldom, Regular, or Frequent
7	Browsed_Electronics_12Mo	Yes/No for whether customer browsed electronics in last year
8	Bought_Electronics_12Mo	Yes/No for whether customer bought electronics in last year
9	Bought_Digital_Media_18Mo	Yes/No for whether customer bought digital media (excluding eReaders) in last 18 mo
10	Bought_Digital_Books	Yes/No for whether customer ever bought a digital book
11	Payment_Method	Bank Transfer, Website Account, Credit Card, or Monthly Billing

Figure 1. Descriptive statistics of Training data set



DATA PREPARATION

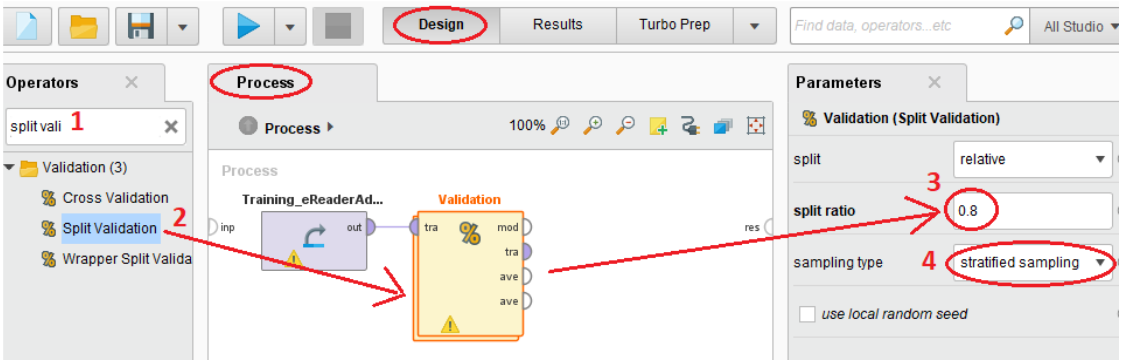
Data Type Transformation: At data import of the training set, the *User_ID* attribute was designated as the ID and the *eReader_adoption* attribute as the label. No transformation to data types was needed because decision tree model can accept all types of data.

Data Preparation of Missing Values: No missing values were found in the data set, and the decision tree model is not sensitive to missing or outlier values.

Split Data Set into Training and Testing using Split Validation: The eReaderadoption_training data set will be split into a training set (80%) and a testing set (20%), using the *Split Validation* operator in RapidMiner, to evaluate the accuracy of the decision tree model with cross-validation. The parameters in Split Validation operator allows user to custom set the split ratio and sampling type. If the Cross Validation operator were to be used, the default would be 70% training and 30% testing, and the default sampling type would be automatic, which typically uses stratified sampling. Stratified sampling ensures that samples in the training and testing will have equal distributions of class values.

Cross-validation takes the data set (in this case, the "eReaderadoption_training.csv" data set), divides it into the number of folds we designate for validation training and validation testing, and tests the model (in this case, the decision tree model) within the training data set, to evaluate the performance (or accuracy) of the model, before the model is deemed good enough to apply to an unlabeled, scoring data set.

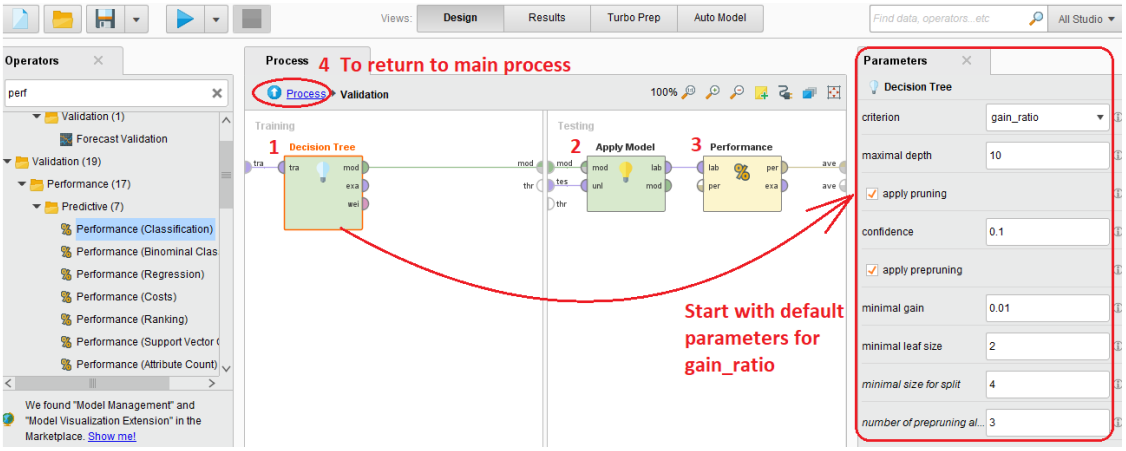
Figure 2. Split Validation operator dividing Training data set into 80/20 for training/testing



MODELING

Following data preparation, the cross-validation set up will continue by double clicking the Validation operator to enter the sub-process.

Figure 3. RapidMiner sub-process for cross-validation



DECISION TREE MODEL

Within the Validation sub-process, the Decision Tree operator is added to the Training sub-process window with parameters at default setting (see figure 3).

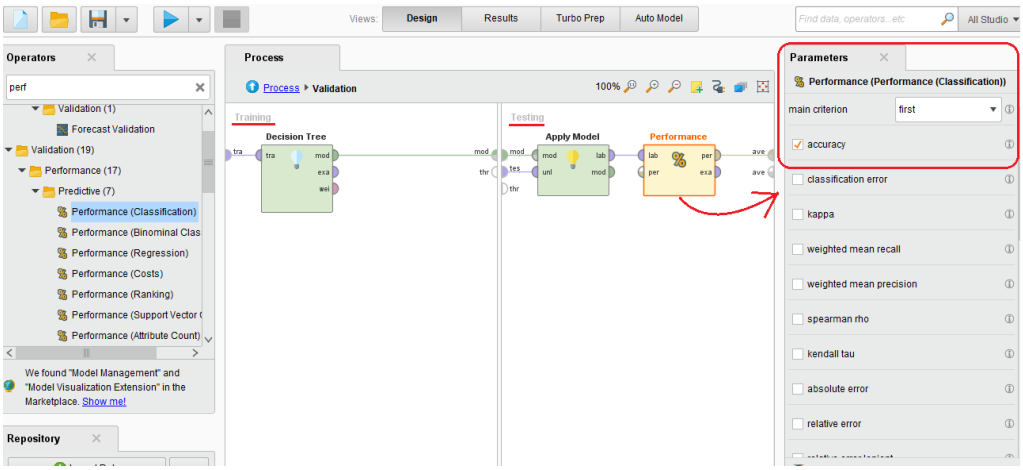
APPLY MODEL

Next, also within the Validation sub-process, the Apply Model operator is added to the Testing sub-process window to connect the training set (which is the 80% that was designated for Training in the Split Validator operator parameter in the Data Preparation stage; in figure 2) and the testing set (which is the 20% that was designated for Testing in the Split Validator operator parameter; in figure 2) to the unlabeled input port (see figure 3).

PERFORMANCE (CLASSIFICATION)

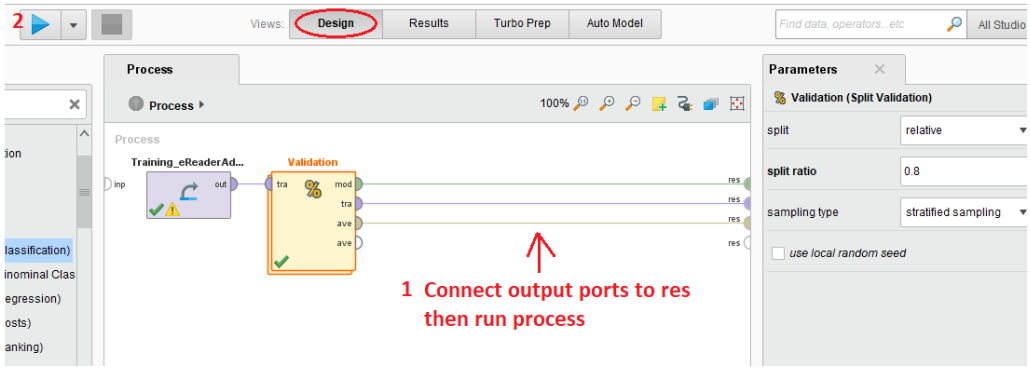
Then, in the Validation sub-process' Testing window, the Performance (Classification) operator is added. The parameters are set to default, which is for accuracy (see figure 3, 4). Of the various Performance operators, the one for Classification is chosen because the decision tree model predicts a classification in the target attribute.

Figure 4. Validation sub-process Testing window showing performance operator default parameters



To return to the main process after configuring the validation sub-process, click the Process as shown in Figure 3, step number4. Make the connections from the Validation operator's output ports to res ports, and click run.

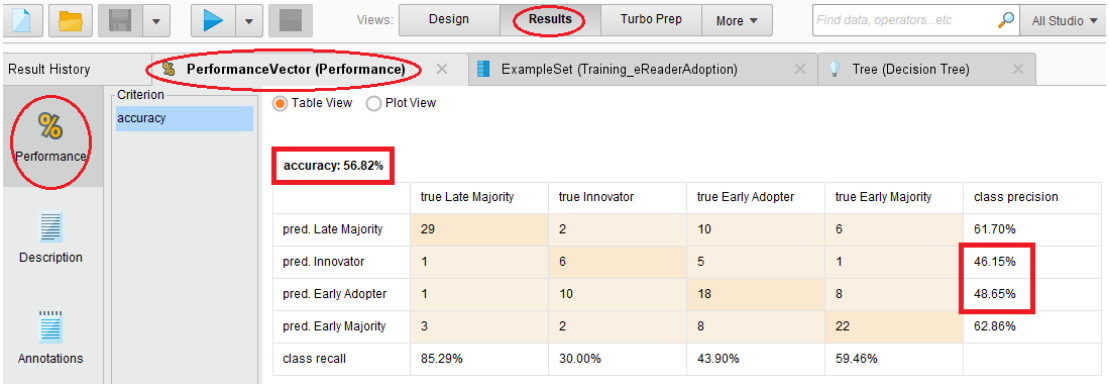
Figure 5. Main process view with all connections to run cross-validation on the nested decision tree model



RESULTS

The following Figure 6 shows the result of the decision tree evaluation, with overall accuracy of the model to be 56.82%. The precision for our desired prediction class, Innovator and Early Adopter, were below 50% while the highest precision was for Early Majority with 62.86%.

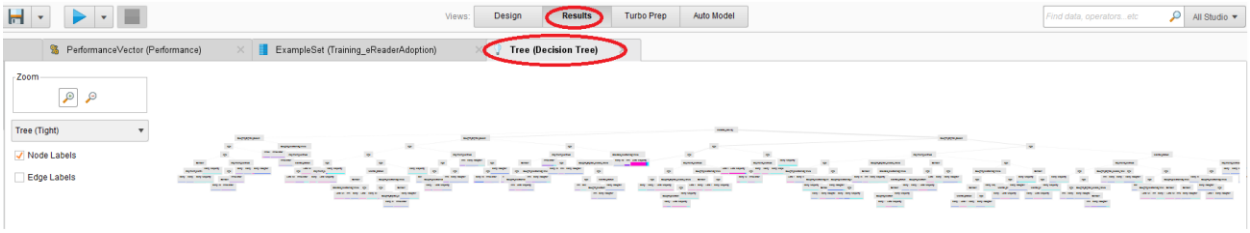
Figure 6. Results showing performance of the decision tree model in first iteration



	true Late Majority	true Innovator	true Early Adopter	true Early Majority	class precision
pred. Late Majority	29	2	10	6	61.70%
pred. Innovator	1	6	5	1	46.15%
pred. Early Adopter	1	10	18	8	48.65%
pred. Early Majority	3	2	8	22	62.86%
class recall	85.29%	30.00%	43.90%	59.46%	

The corresponding decision tree is shown in Figure 7.

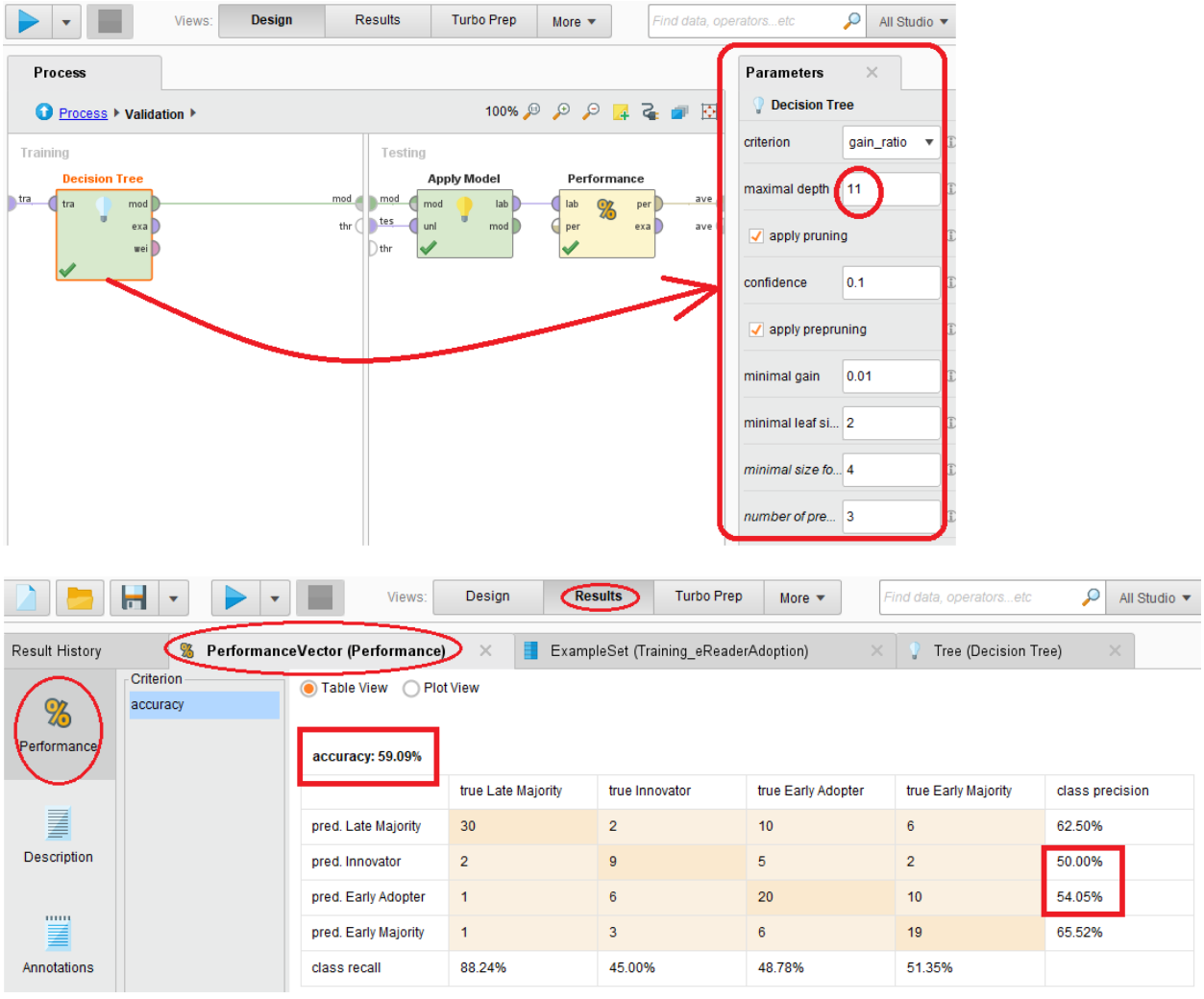
Figure 7. Results showing the corresponding decision tree for 56.82% overall accuracy in first iteration



SECOND ITERATION

So, another iteration will be performed with different decision tree parameters to evaluate the performance and find a better performing model (see figure 8).

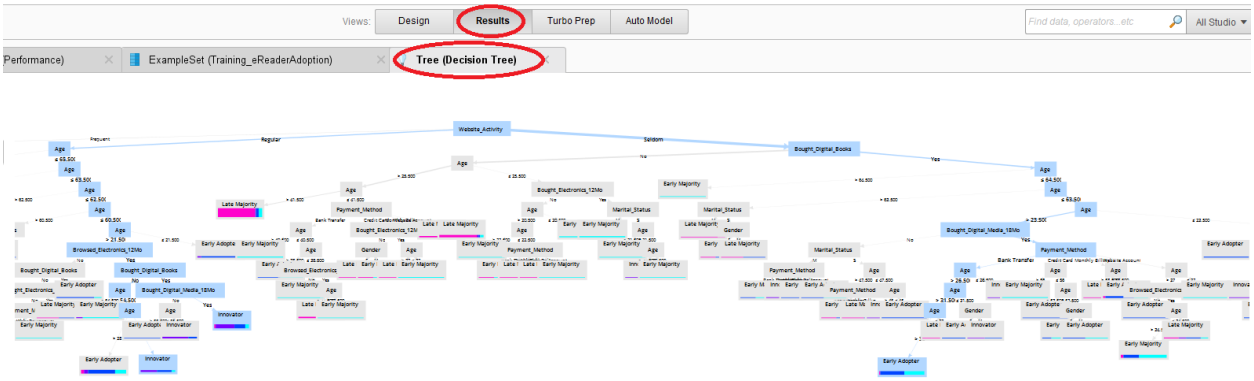
Figure 8. Second iteration showing performance of the decision tree model with different parameters and resulting accuracy



	true Late Majority	true Innovator	true Early Adopter	true Early Majority	class precision
pred. Late Majority	30	2	10	6	62.50%
pred. Innovator	2	9	5	2	50.00%
pred. Early Adopter	1	6	20	10	54.05%
pred. Early Majority	1	3	6	19	65.52%
class recall	88.24%	45.00%	48.78%	51.35%	

The result of changing the maximal depth from the default 10 to 11 yielded a higher overall accuracy from 56.82% to 59.09% and also higher precision for Innovator and Early Adopter (see figure 8). The corresponding decision tree is shown in Figure 9.

Figure 9. Results showing the corresponding decision tree for 59.09% overall accuracy in second iteration



THIRD ITERATION

An examination of the sample distribution in the confusion matrix of the second iteration showed an unbalance in that biased the percentages and prompted a third iteration with the Split Validation split ratio reset from 80/20 to 90/10 (see figure 10). The decision tree parameters remained the same as the second iteration. The third iteration resulted in an increased overall accuracy from 59.09% in the second iteration to 65.15% (see figure 11). This 90% for training also yielded the precision for Innovator to be raised from 50% to 75% and for Early Adopter to be raised from 54.05% to 61.11% (see figure 11).

Figure 10. Third iteration showing split validation parameter with split ration changed from 80% to 90%

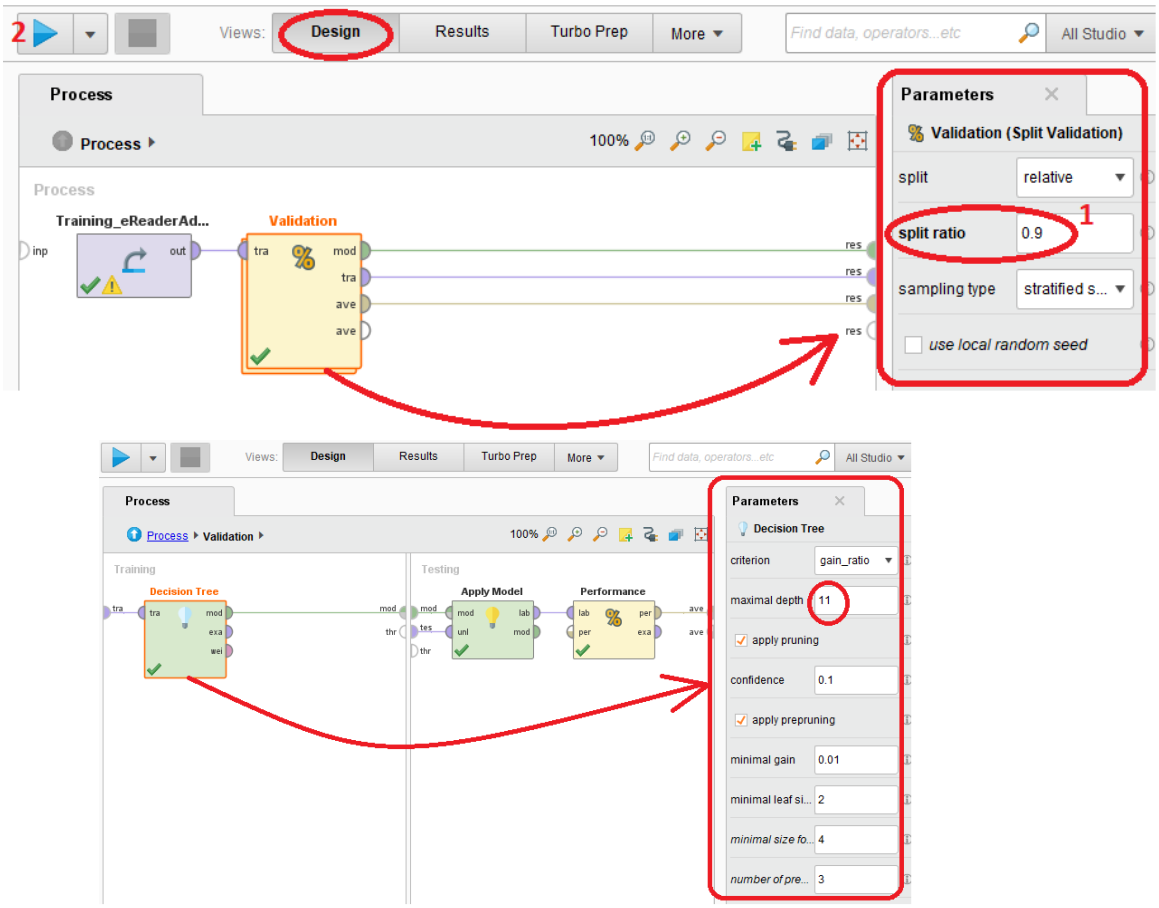
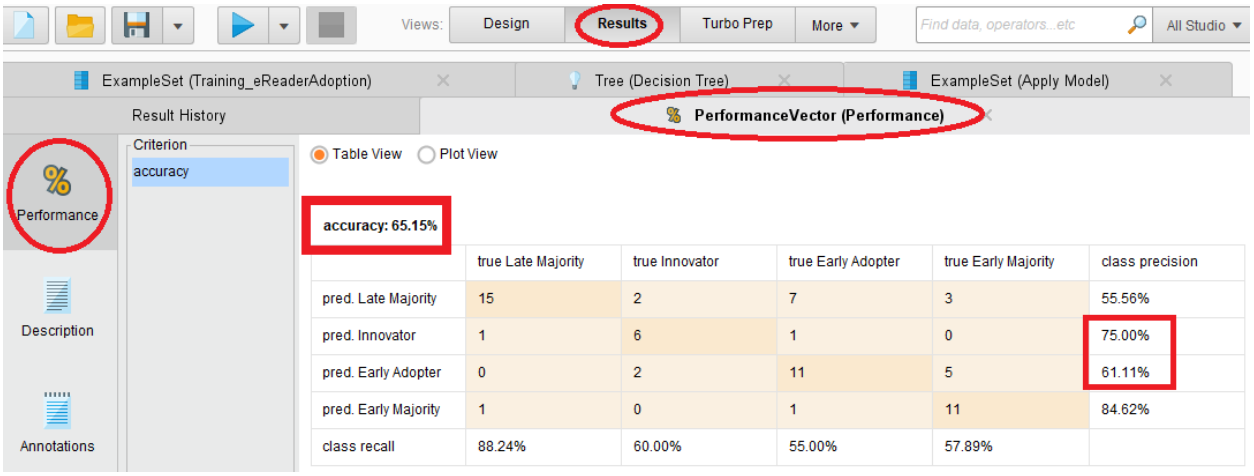
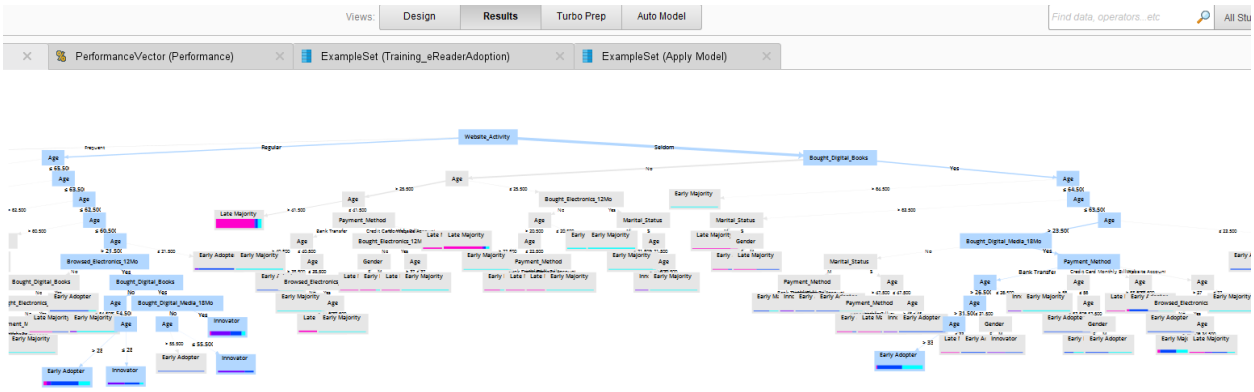


Figure 11. Results showing performance of the decision tree model in third iteration with 65.15% overall accuracy



	true Late Majority	true Innovator	true Early Adopter	true Early Majority	class precision
pred. Late Majority	15	2	7	3	55.56%
pred. Innovator	1	6	1	0	75.00%
pred. Early Adopter	0	2	11	5	61.11%
pred. Early Majority	1	0	1	11	84.62%
class recall	88.24%	60.00%	55.00%	57.89%	

Figure 12. Results showing the corresponding decision tree for 65.15% overall accuracy in third iteration



FOURTH ITERATION: CROSS-VALIDATION ON 10 FOLDS

The first and second iterations above were performed on a 80/20 split ratio, where the testing sample size was 132 (20% of the total 661 examples). The third iteration was performed on a 90/10 split ratio, where the testing sample size was 66 (10% of the total 661 examples). Smaller testing sample size yielded higher accuracy, which requires further evaluation. So, a Cross Validation operator was added to the main process to run the model and evaluation on 10 folds until each fold has been used as a testing set, and then the performance is derived from an aggregate of the 10 executions (see figure 13). This should provide a more reliable overall accuracy rating.

Figure 13. Process showing cross validation operator with parameter set to run 10 folds

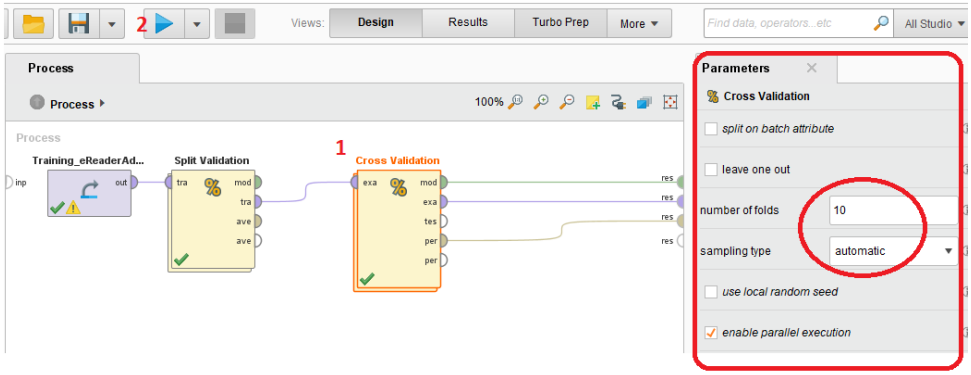


Figure 14. Sub-process showing decision tree model parameters for 10-fold cross validation

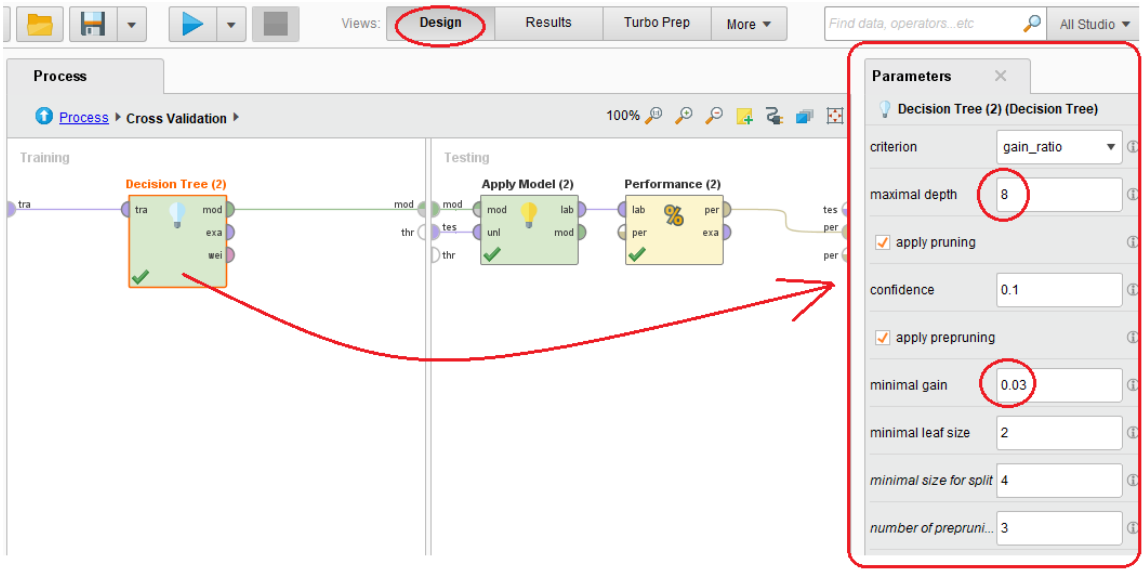
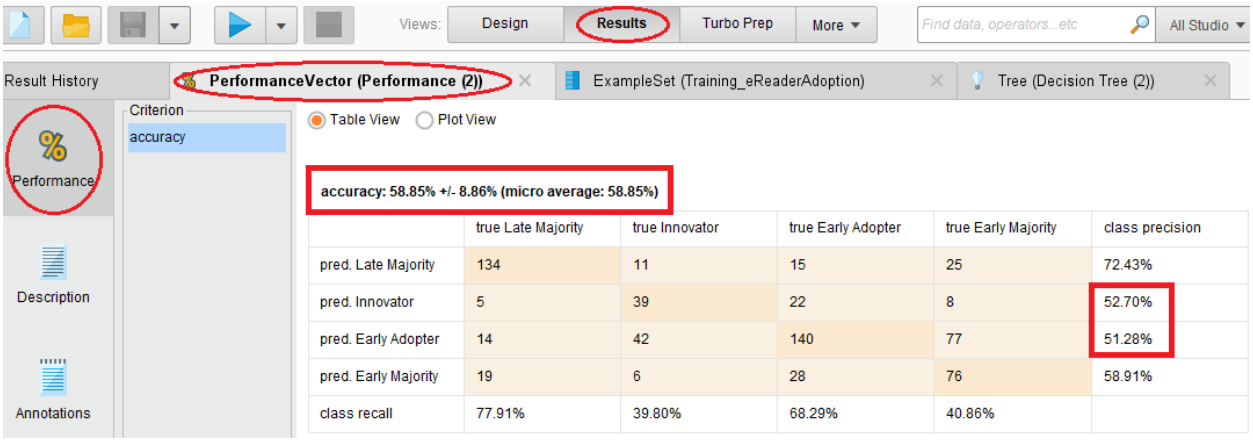
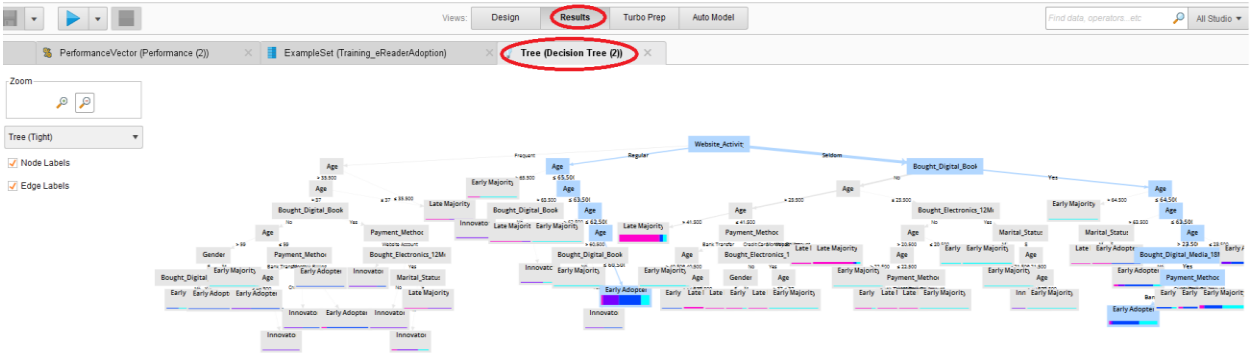


Figure 15. Results showing performance of the decision tree model with 10-fold cross validation



	true Late Majority	true Innovator	true Early Adopter	true Early Majority	class precision
pred. Late Majority	134	11	15	25	72.43%
pred. Innovator	5	39	22	8	52.70%
pred. Early Adopter	14	42	140	77	51.28%
pred. Early Majority	19	6	28	76	58.91%
class recall	77.91%	39.80%	68.29%	40.86%	

Figure 16. Results showing the corresponding decision tree from fourth iteration

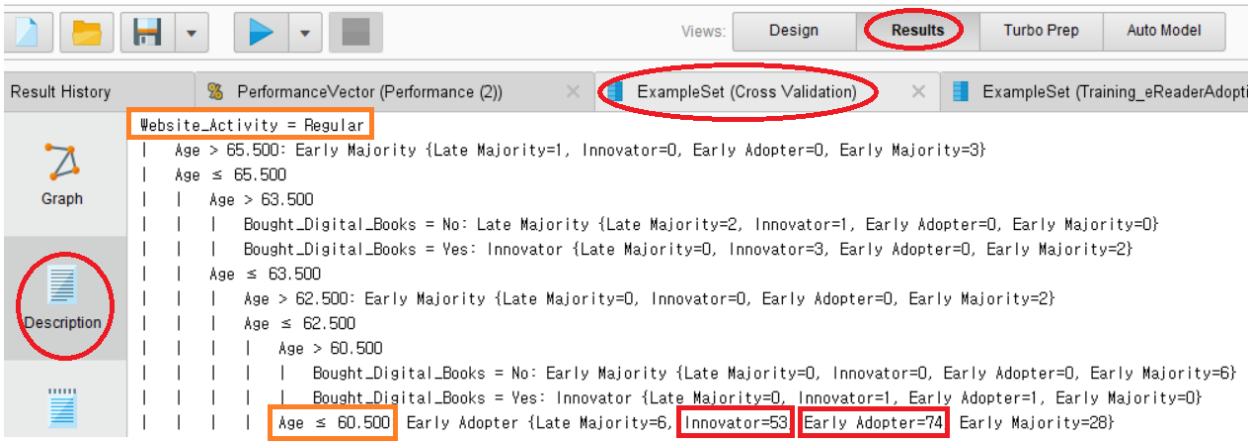


EVALUATION OF FINDINGS

A cross validation on 10 folds yielded an overall accuracy of 58.85%, a precision of 52.70% for Innovator and 51.28% for Early Adopter (see figure 15). Precision is the measure of how many predictions were correct (True Positive) out of all predictions (True Positive plus True Negative). Accuracy is an indicator of what percentage of predictions were correct (True Positive plus True Negative) out of all cases. Recall is the measure of how many were predicted correct out of total actual correct. In the 10 fold cross validation, the recall for Innovator was 39.80% and for Early adopter, 68.29% (see figure 15).

According to the 10-fold cross validation, the top/best predictor attributes are Website_Actvity (being a regular user) and Age (being less than or equal to 60.5 years old).

Figure 17. Cross Validation result showing top/best predictors



The decision tree model was most successful at predicting Late Majority customers with a precision of 72.43%; unfortunately, this business is not interested in predicting who will be Late Majority customers for the targeted marketing for early eReader adopters. However, since this prediction for Late Majority customers had a fairly high precision, it could at least help the business to eliminate true late majority customers from the scoring data set and narrow the focus on the remaining customer categories for the targeted early marketing.

Figure 18. Power BI list of predictions from decision tree analysis per User_ID with respective confidences

User_ID	prediction(eReader_Adoption)	confidence(Innovator)	confidence(Early Adopter)	confidence(Early Majority)	confidence(Late Majority)
1595	Innovator	1.00	0.00	0.00	0.00
1828	Innovator	1.00	0.00	0.00	0.00
3255	Innovator	1.00	0.00	0.00	0.00
3712	Innovator	1.00	0.00	0.00	0.00
4454	Innovator	1.00	0.00	0.00	0.00
4620	Innovator	1.00	0.00	0.00	0.00
5108	Innovator	1.00	0.00	0.00	0.00
7333	Innovator	1.00	0.00	0.00	0.00
8450	Innovator	0.89	0.00	0.11	0.00
1202	Innovator	0.88	0.00	0.00	0.13
8755	Innovator	0.88	0.00	0.00	0.13
6536	Innovator	0.86	0.00	0.00	0.14
7631	Innovator	0.86	0.00	0.00	0.14
6147	Innovator	0.83	0.00	0.00	0.17
7571	Innovator	0.83	0.00	0.00	0.17
3291	Innovator	0.83	0.10	0.07	0.00
3454	Innovator	0.83	0.10	0.07	0.00
8830	Innovator	0.83	0.10	0.07	0.00
2749	Innovator	0.80	0.20	0.00	0.00
9895	Innovator	0.80	0.20	0.00	0.00
1756	Innovator	0.75	0.00	0.25	0.00
5550	Innovator	0.75	0.25	0.00	0.00
7673	Innovator	0.75	0.25	0.00	0.00
2273	Innovator	0.71	0.17	0.07	0.05
6530	Innovator	0.71	0.17	0.07	0.05
2270	Innovator	0.67	0.33	0.00	0.00
2310	Innovator	0.67	0.11	0.22	0.00
4491	Innovator	0.67	0.33	0.00	0.00
6049	Innovator	0.67	0.11	0.22	0.00
7309	Innovator	0.67	0.33	0.00	0.00
9332	Innovator	0.67	0.33	0.00	0.00
1164	Innovator	0.61	0.31	0.07	0.02
2490	Innovator	0.61	0.31	0.07	0.02
3204	Innovator	0.61	0.31	0.07	0.02
5219	Innovator	0.61	0.31	0.07	0.02
5464	Innovator	0.61	0.31	0.07	0.02
6715	Innovator	0.61	0.31	0.07	0.02
7371	Innovator	0.61	0.31	0.07	0.02
7751	Innovator	0.61	0.31	0.07	0.02
2908	Innovator	0.60	0.40	0.00	0.00
4211	Innovator	0.60	0.00	0.20	0.20
4448	Innovator	0.60	0.00	0.20	0.20
5666	Innovator	0.60	0.40	0.00	0.00
2314	Innovator	0.56	0.31	0.10	0.03
2955	Innovator	0.56	0.31	0.10	0.03
6745	Innovator	0.56	0.31	0.10	0.03
9492	Innovator	0.56	0.31	0.10	0.03
9793	Innovator	0.56	0.31	0.10	0.03
9845	Innovator	0.56	0.31	0.10	0.03
9974	Innovator	0.56	0.31	0.10	0.03
9024	Early Adopter	0.55	0.90	0.48	0.07
2097	Innovator	0.55	0.34	0.08	0.03
4254	Innovator	0.55	0.34	0.08	0.03
4937	Innovator	0.55	0.34	0.08	0.03

BUSINES RECOMMENDATIONS

Evaluation of findings using cross validation suggest that the decision tree model finds customers with more than regular website activity who are less than 60 years of age will likely fall into the Innovator (who purchase within one week of release) or Early Adapter (who purchase within 2-3 weeks of release) category. So, the report recommends that the business target the marketing efforts for the next-gen eReader toward customers who fit that category, at least a couple of weeks prior to the release of the eReader.

Since the model was evaluated to have higher precision with predicting Late Majority (who purchase after 8 weeks of release), this report also recommends that the business could use these predictions of Late Majority to bypass those customers in targeting marketing efforts.