Week2_Assignment Report

The assignment for week 2 was to evaluate the responses of one thousand potential customers based on recent customer responses. The goal is to find the responses that will yield the highest profit and reject those that will not respond to the promotions. There was a total of twelve attributes minus the responses for the current customers. Data such as name, area, tool used to log-in, responses within four to six months, and sales were used to determine the customers that were likely to respond positively to a promotional campaign by the company.

Variables:

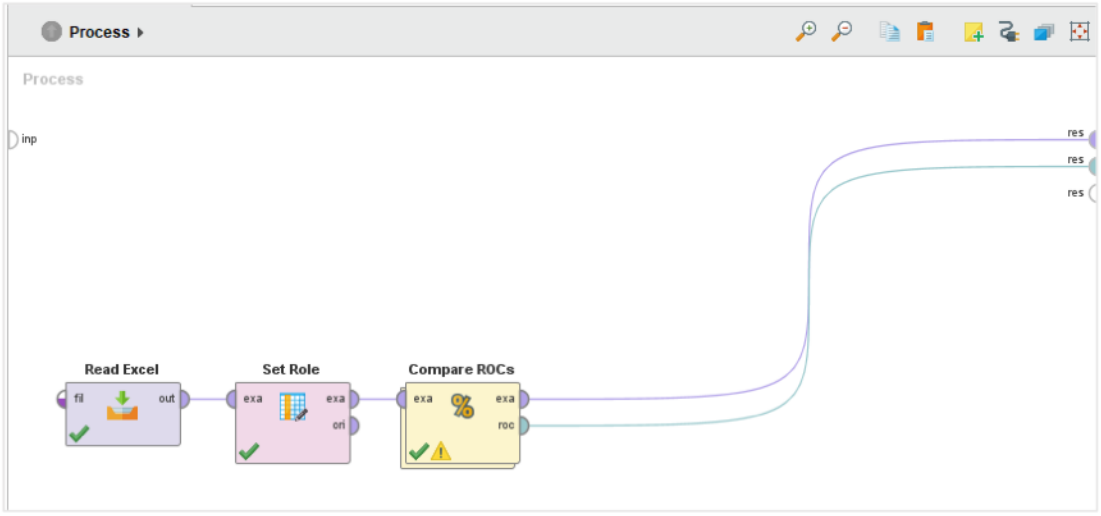
- Number of customers is 10,000.
- Average response rate is 10%.
- Cost per mail sent to each person is $1.
- Potential revenue per response is $20.

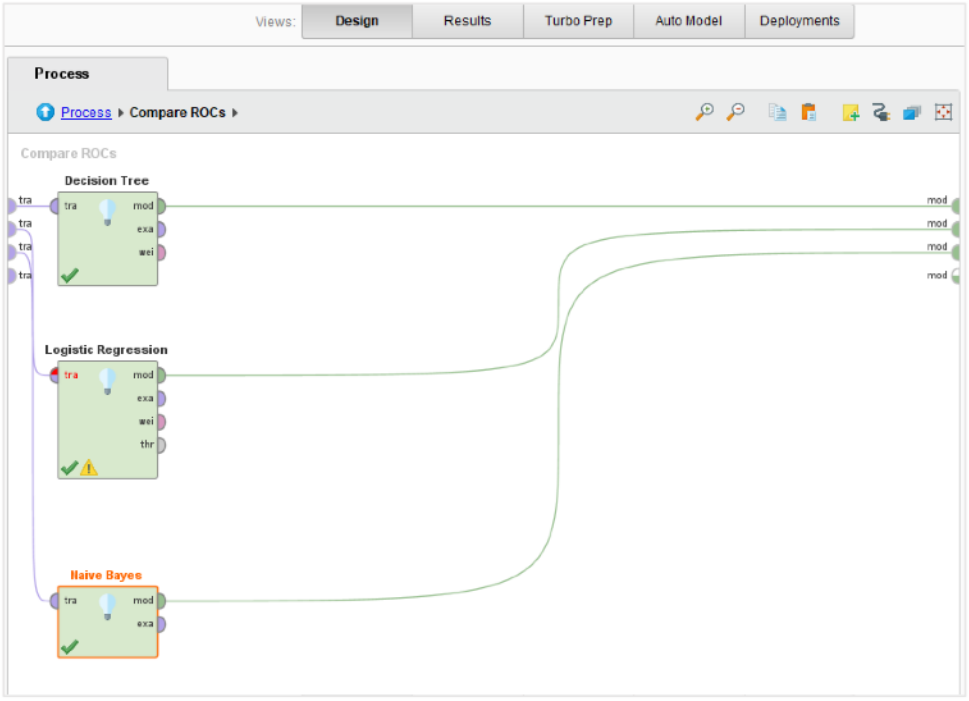
Profit:

- 10,000 x 10% = 1,000 responses
- 1,000 x $20 = $20,000
- Cost is $1 x 1,000 = $1,000
- $20,000 - $1,000 = $19,000 max profit

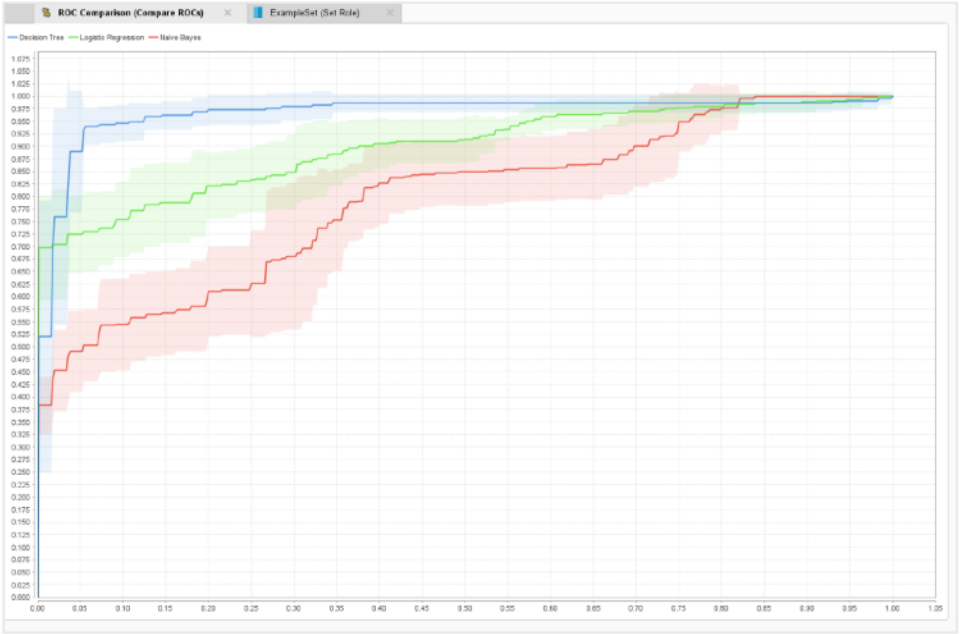
First I have listed screen shots and descriptions of what I followed in your instruction videos to produce the same results. Below those details, you will see my assessment, additional models and steps I implemented in order to try and achieve better results to benefit Best Buy.

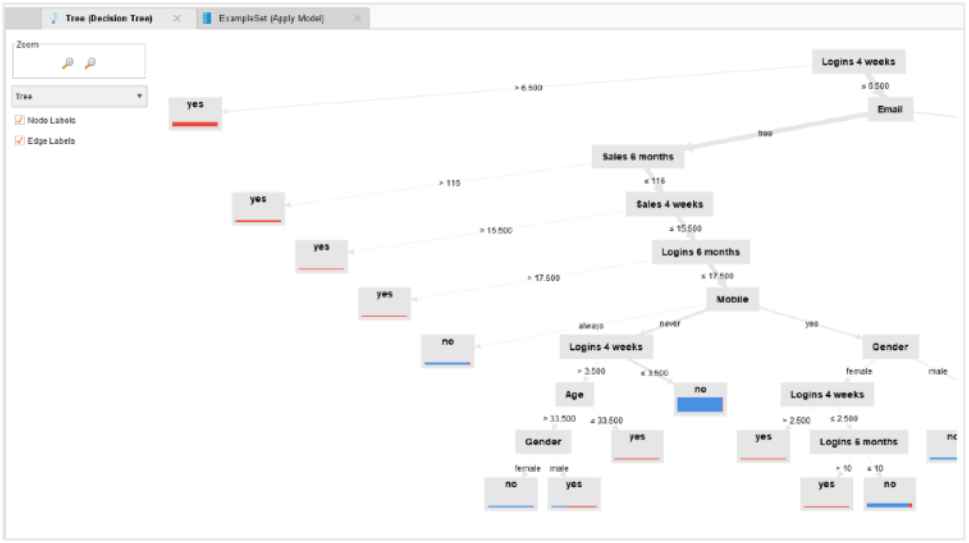
I imported the training data, set the role as "response" and compared ROC's Decision Tree, Logistic Regression, and Naïve Bayes:

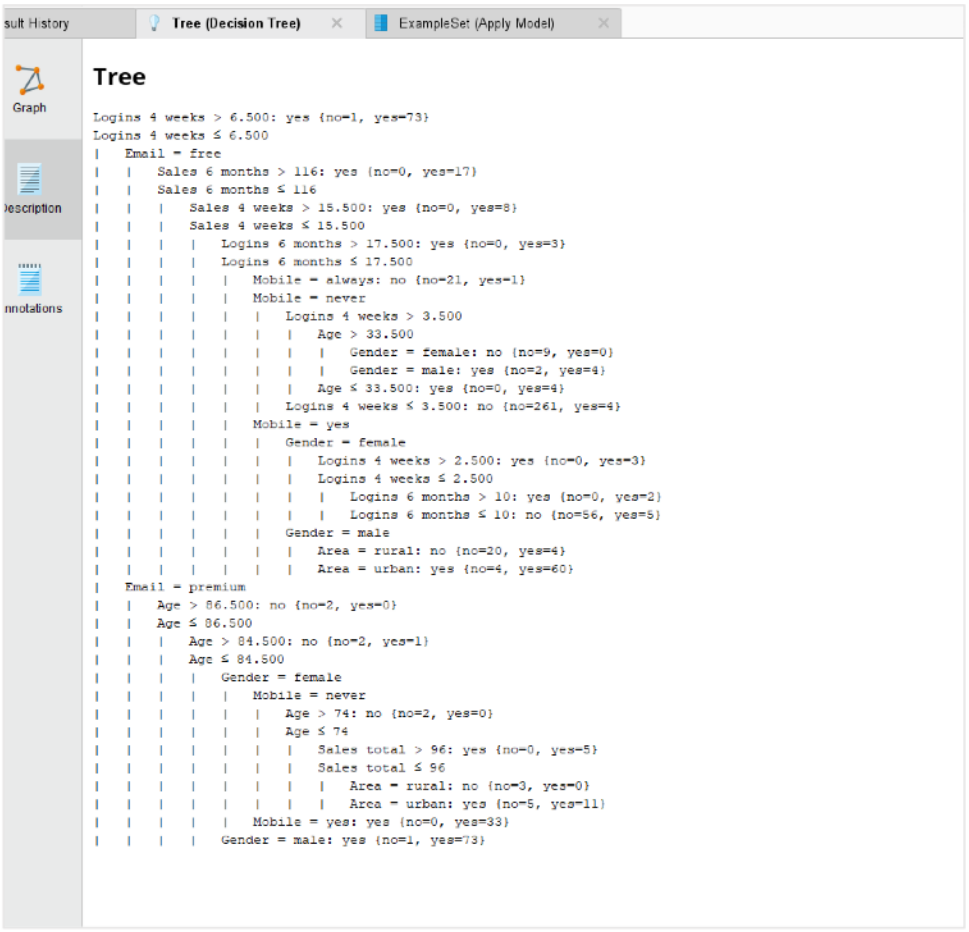




Next, I reviewed the ROC comparison, noting that the Decision Tree produced the most accurate results at the time.

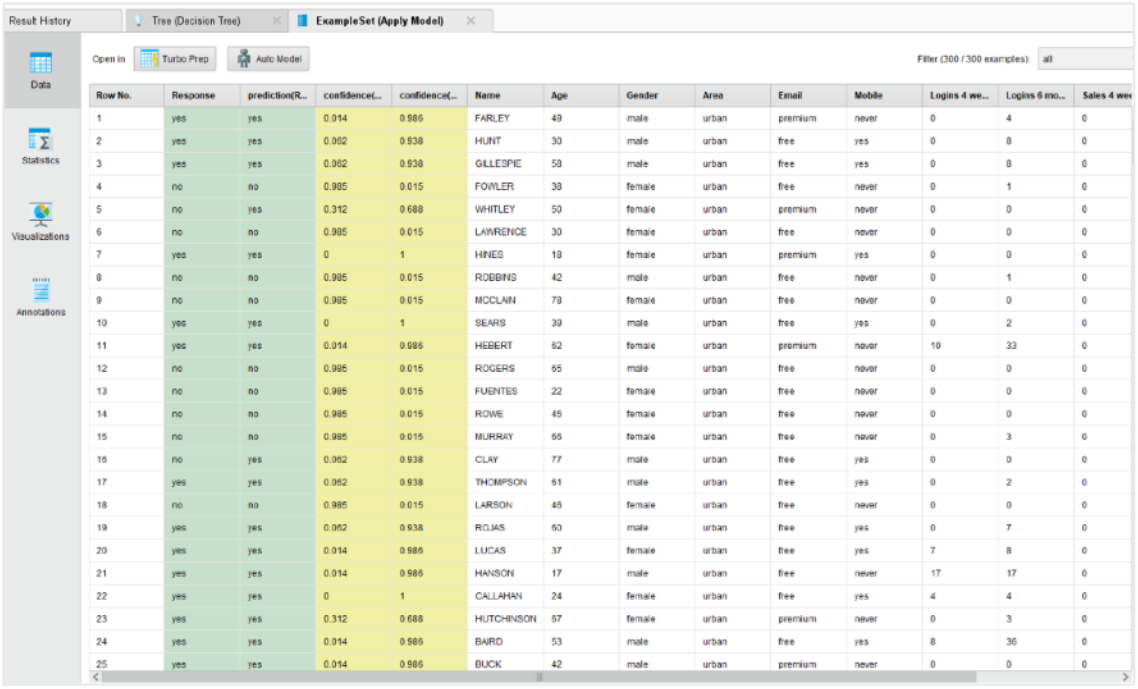






Tree

```
Logins 4 weeks > 6.500: yes {no=1, yes=73}
Logins 4 weeks ≤ 6.500
|   Email = free
|   |   Sales 6 months > 116: yes {no=0, yes=17}
|   |   Sales 6 months ≤ 116
|   |   |   Sales 4 weeks > 15.500: yes {no=0, yes=8}
|   |   |   Sales 4 weeks ≤ 15.500
|   |   |   |   Logins 6 months > 17.500: yes {no=0, yes=3}
|   |   |   |   Logins 6 months ≤ 17.500
|   |   |   |   |   Mobile = always: no {no=21, yes=1}
|   |   |   |   |   Mobile = never
|   |   |   |   |   |   Logins 4 weeks > 3.500
|   |   |   |   |   |   |   Age > 33.500
|   |   |   |   |   |   |   |   Gender = female: no {no=9, yes=0}
|   |   |   |   |   |   |   |   Gender = male: yes {no=2, yes=4}
|   |   |   |   |   |   |   Age ≤ 33.500: yes {no=0, yes=4}
|   |   |   |   |   |   Logins 4 weeks ≤ 3.500: no {no=261, yes=4}
|   |   |   |   |   Mobile = yes
|   |   |   |   |   |   Gender = female
|   |   |   |   |   |   |   Logins 4 weeks > 2.500: yes {no=0, yes=3}
|   |   |   |   |   |   |   Logins 4 weeks ≤ 2.500
|   |   |   |   |   |   |   |   Logins 6 months > 10: yes {no=0, yes=2}
|   |   |   |   |   |   |   |   Logins 6 months ≤ 10: no {no=56, yes=5}
|   |   |   |   |   |   Gender = male
|   |   |   |   |   |   |   Area = rural: no {no=20, yes=4}
|   |   |   |   |   |   |   Area = urban: yes {no=4, yes=60}
|   Email = premium
|   |   Age > 86.500: no {no=2, yes=0}
|   |   Age ≤ 86.500
|   |   |   Age > 84.500: no {no=2, yes=1}
|   |   |   Age ≤ 84.500
|   |   |   |   Gender = female
|   |   |   |   |   Mobile = never
|   |   |   |   |   |   Age > 74: no {no=2, yes=0}
|   |   |   |   |   |   Age ≤ 74
|   |   |   |   |   |   |   Sales total > 96: yes {no=0, yes=5}
|   |   |   |   |   |   |   Sales total ≤ 96
|   |   |   |   |   |   |   |   Area = rural: no {no=3, yes=0}
|   |   |   |   |   |   |   |   Area = urban: yes {no=5, yes=11}
|   |   |   |   |   Mobile = yes: yes {no=0, yes=33}
|   |   |   |   Gender = male: yes {no=1, yes=73}
```



Result History | Tree (Decision Tree) | ExampleSet (Apply Model)

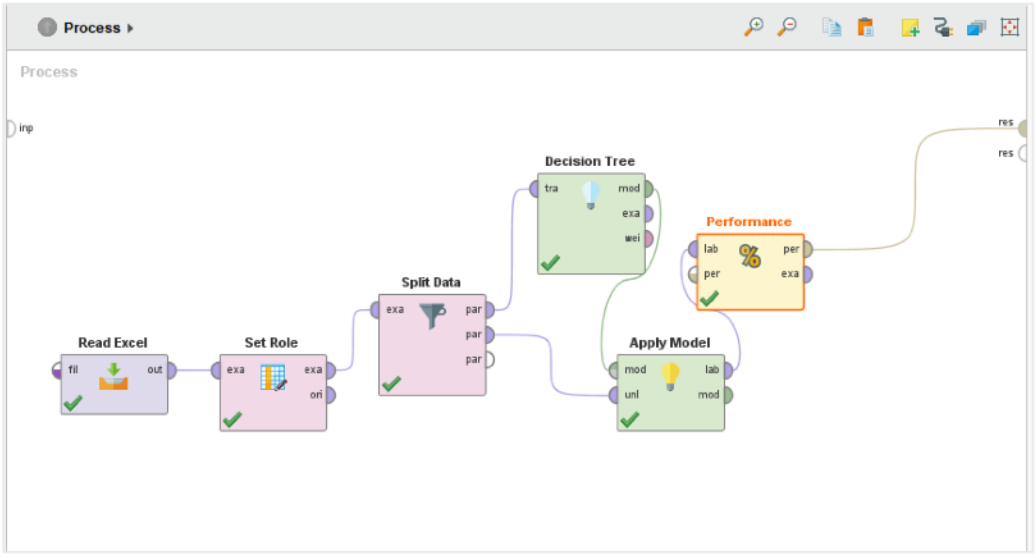
Data | Statistics | Visualizations | Annotations

Open in: Turbo Prep | Auto Model

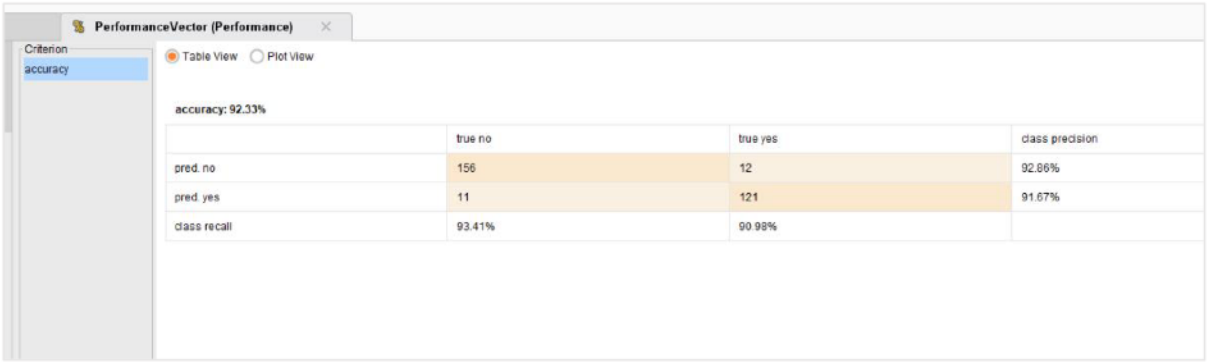
Filter (300 / 300 examples): all

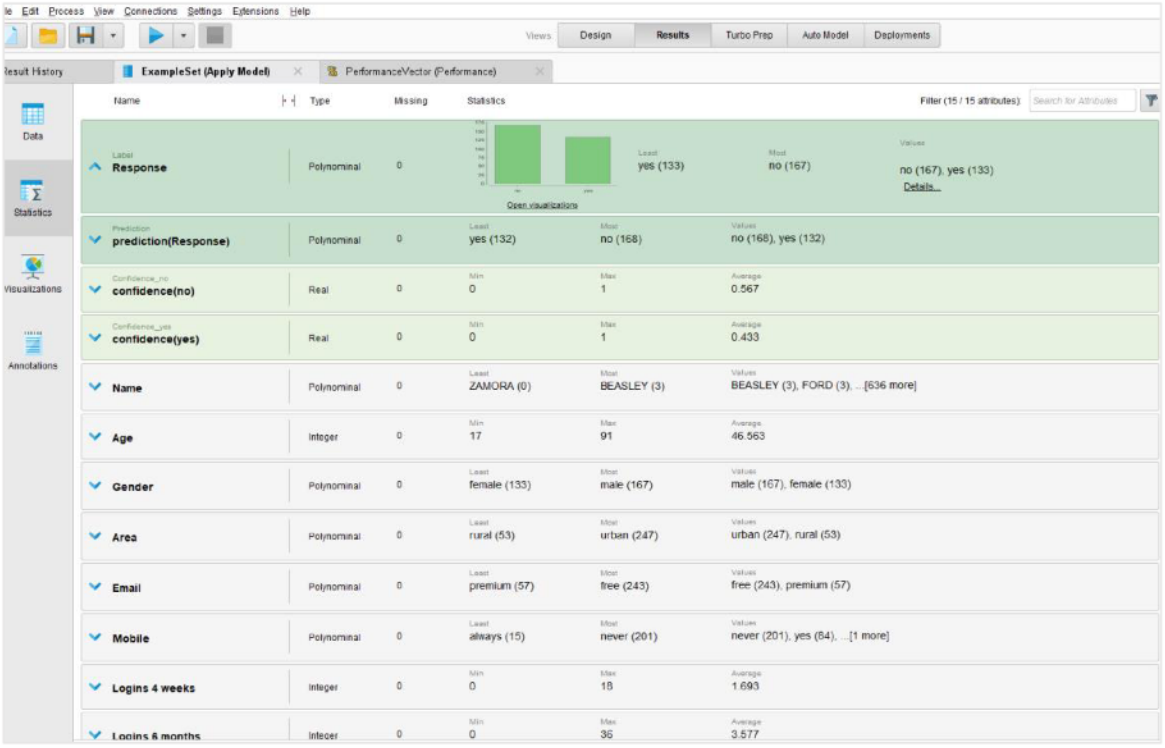
Row No.	Response	prediction(R...	confidence(...	confidence(...	Name	Age	Gender	Area	Email	Mobile	Logins 4 we...	Logins 6 mo...	Sales 4 we
1	yes	yes	0.014	0.986	FARLEY	49	male	urban	premium	never	0	4	0
2	yes	yes	0.062	0.938	HUNT	30	male	urban	free	yes	0	8	0
3	yes	yes	0.062	0.938	GILLESPIE	58	male	urban	free	yes	0	8	0
4	no	no	0.985	0.015	FOWLER	38	female	urban	free	never	0	1	0
5	no	yes	0.312	0.688	WHITLEY	50	female	urban	premium	never	0	0	0
6	no	no	0.985	0.015	LAWRENCE	30	female	urban	free	never	0	0	0
7	yes	yes	0	1	HINES	18	female	urban	premium	yes	0	0	0
8	no	no	0.985	0.015	ROBBINS	42	male	urban	free	never	0	1	0
9	no	no	0.985	0.015	MCCLAIN	78	female	urban	free	never	0	0	0
10	yes	yes	0	1	SEARS	39	male	urban	free	yes	0	2	0
11	yes	yes	0.014	0.986	HEBERT	62	female	urban	premium	never	10	33	0
12	no	no	0.985	0.015	ROGERS	65	male	urban	free	never	0	0	0
13	no	no	0.985	0.015	FUENTES	22	female	urban	free	never	0	0	0
14	no	no	0.985	0.015	ROWE	45	female	urban	free	never	0	0	0
15	no	no	0.985	0.015	MURRAY	66	female	urban	free	never	0	3	0
16	no	yes	0.062	0.938	CLAY	77	male	urban	free	yes	0	0	0
17	yes	yes	0.062	0.938	THOMPSON	61	male	urban	free	yes	0	2	0
18	no	no	0.985	0.015	LARSON	45	female	urban	free	never	0	0	0
19	yes	yes	0.062	0.938	ROJAS	60	male	urban	free	yes	0	7	0
20	yes	yes	0.014	0.986	LUCAS	37	female	urban	free	yes	7	8	0
21	yes	yes	0.014	0.986	HANSON	17	male	urban	free	never	17	17	0
22	yes	yes	0	1	CALLAHAN	24	female	urban	free	yes	4	4	0
23	yes	yes	0.312	0.688	HUTCHINSON	67	female	urban	premium	never	0	3	0
24	yes	yes	0.014	0.986	BAIRD	53	male	urban	free	yes	8	36	0
25	yes	yes	0.014	0.986	BUCK	42	male	urban	premium	never	0	0	0

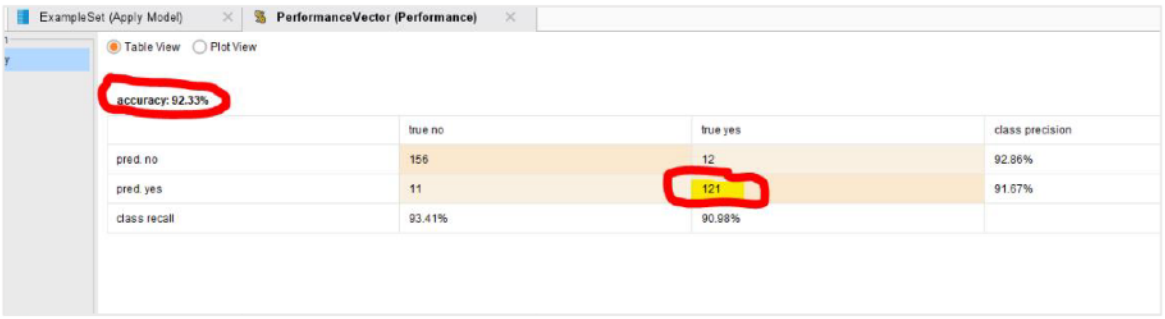
Next, I followed by deleting the Compare ROC operator and replaced Decision Tree, applied model and inserted Performance Classification operator:



I then viewed the confusion matrix, finding an overall accuracy of 92.33%, with a Class Recall of 90.98% for true yes.



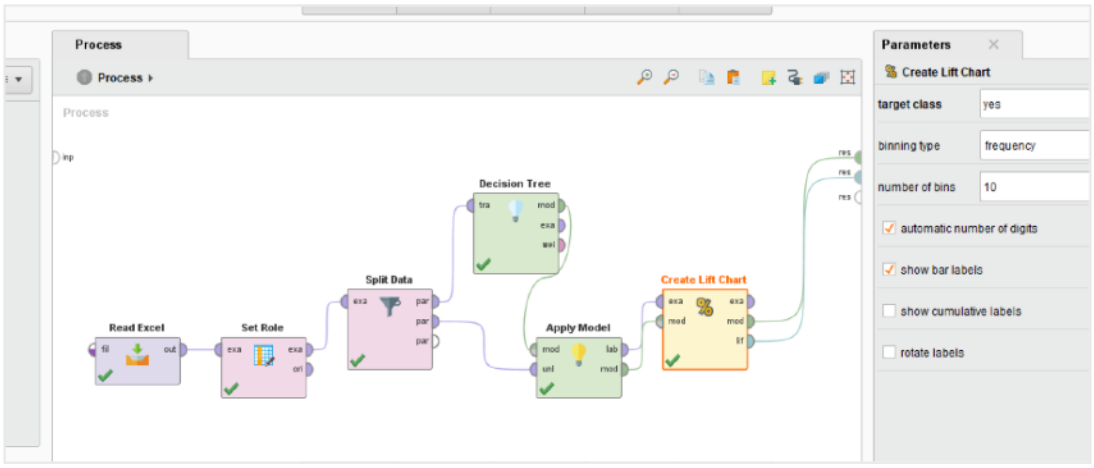




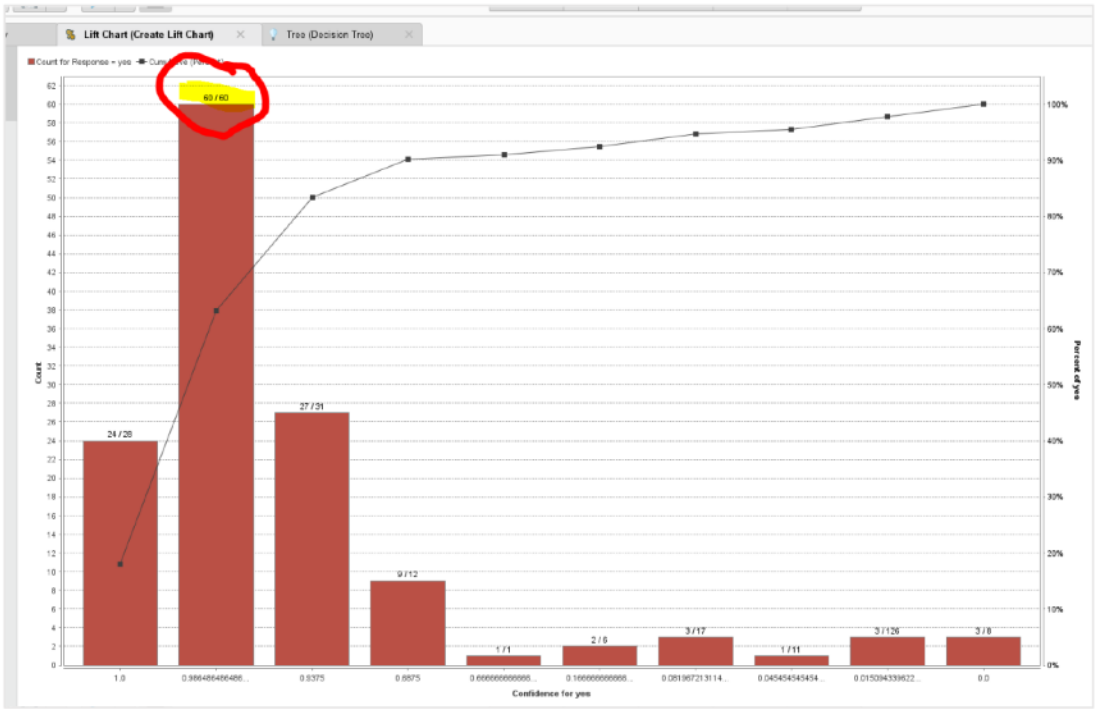
After reviewing the results above and more thoroughly examining the confusion matrix above, we realized the following:

- Although the model showed 132 responded "yes," we see in the Performance Vector that only 121 were true "yes" responses. However, the accuracy of 92.33% is acceptable, as Dr. Zhuhadar explained that these models will never produce 100%.

I next created a lift chart and used the target category as "yes."

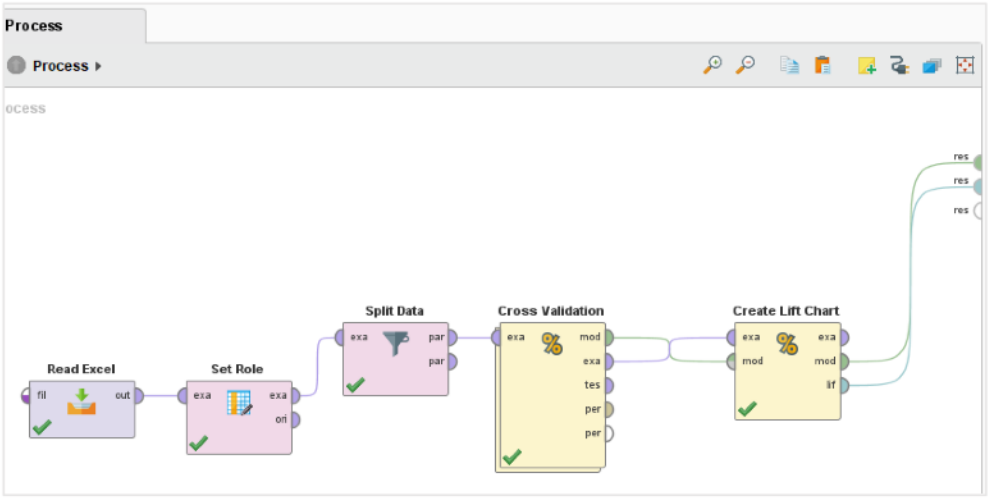


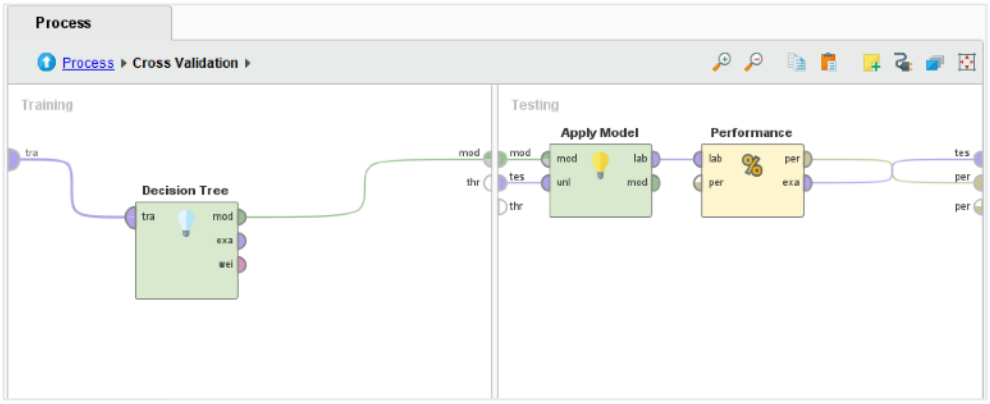
The results of the lift chart are below:

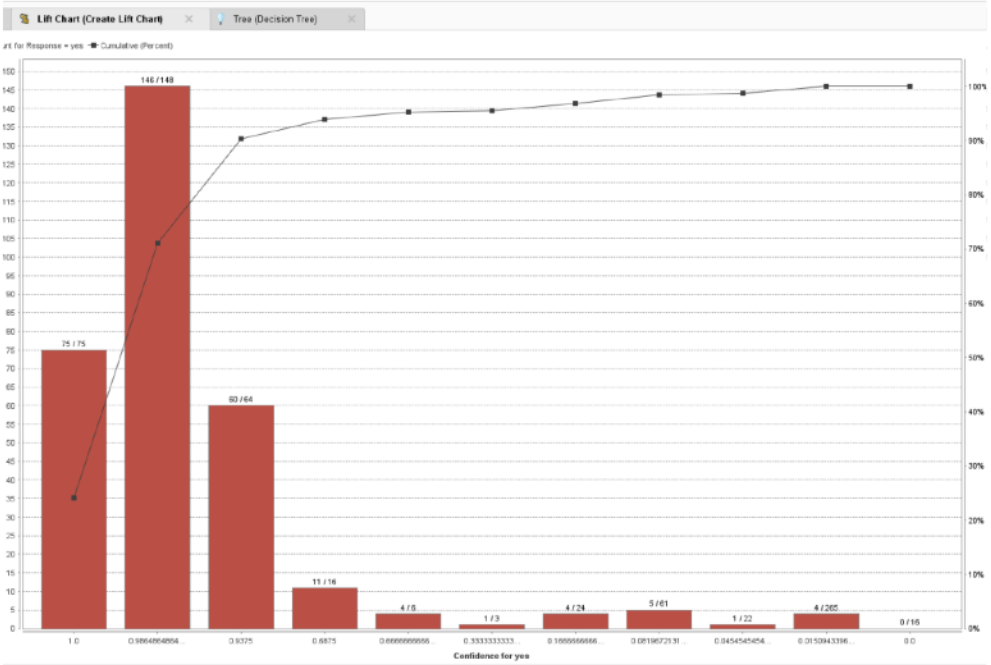


We see this is not a great model, as the highest represented result should be at the beginning of the chart – with the best model presenting a descending order bar graph.

After understanding this information, I inserted a Cross Validation operator and viewed a lift chart to determine better accuracy of the model, as displayed below:





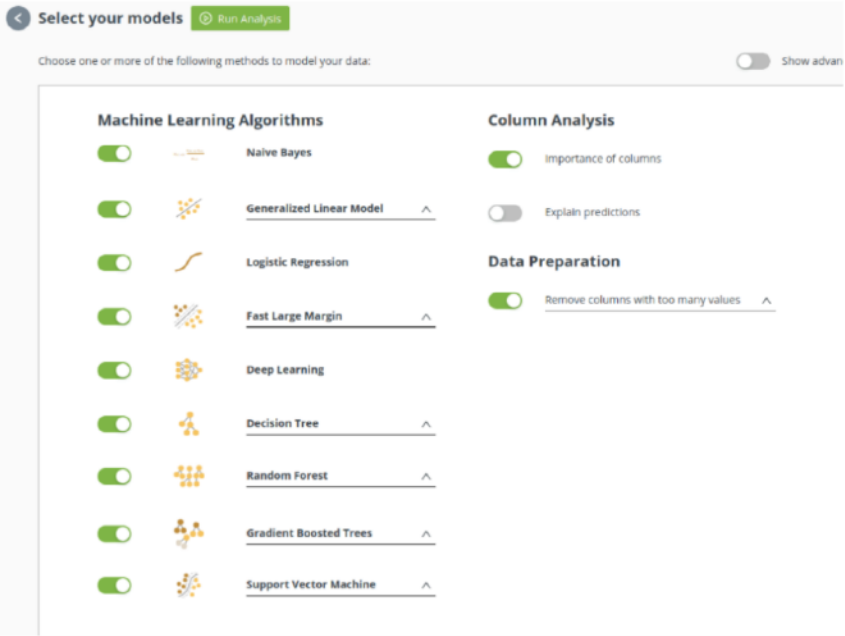


At this point of the assignment, I felt more confident to begin the journey to determine the most profitable target market audience for the direct marketing campaign. Below are my steps in the process:

Direct Marketing Campaign Recommendations

Best Buy – Bowling Green, KY

- What Business Problem are you trying to solve?
  - As the Manager of Best Buy Bowling Green, KY, I am trying to identify which customers would most likely utilize the upcoming promotional materials sent in the direct marketing campaign.
- Why solving this problem is important?
  - Generating additional revenue and from repeat customers is crucial and also creates brand loyalty, leading to more sustained profitability.
  - However, if we blindly mail thousands of promotional materials to customers, we will decrease our profit margins significantly, due to the sunk costs of postage/printing/labor by not making wise decisions based upon our target market.
- How did you solve this problem?
  - After watching Dr. Zhuhadar's lectures and completing the examples, I used RapidMiner to build multiple models in order to find the most accurate model to apply to the current customer database.
  - I additionally researched how to use predictive analytics for direct marketing at the following link : https://www.youtube.com/watch?v=dnQFmkXURgI.
  - Also, prior to beginning my models, I found the information that showed the best models to choose from below:



- Why would I consider your solution?
  - Based upon my research and comparison of predictive models, I can offer with a 97.67% accuracy that the customers selected will most likely utilize the direct marketing promotional materials to purchase items from Best Buy. Not choosing the solution could result in lower profit margins, and missed opportunities by mailing to those customers who have not shown a pattern of responding to direct mail marketing.

- Once the model was chosen and run on the current customer database, I further analyzed the data by running a pivot table and determining the following:
  - When filtering for the predicted value of "yes" and a confidence (yes) of 1,
    - Females in rural areas spent 27% more than females in urban areas, when comparing 4 weeks sales
    - Females in rural areas spent 64% higher than males in rural areas, when comparing 4 weeks sales

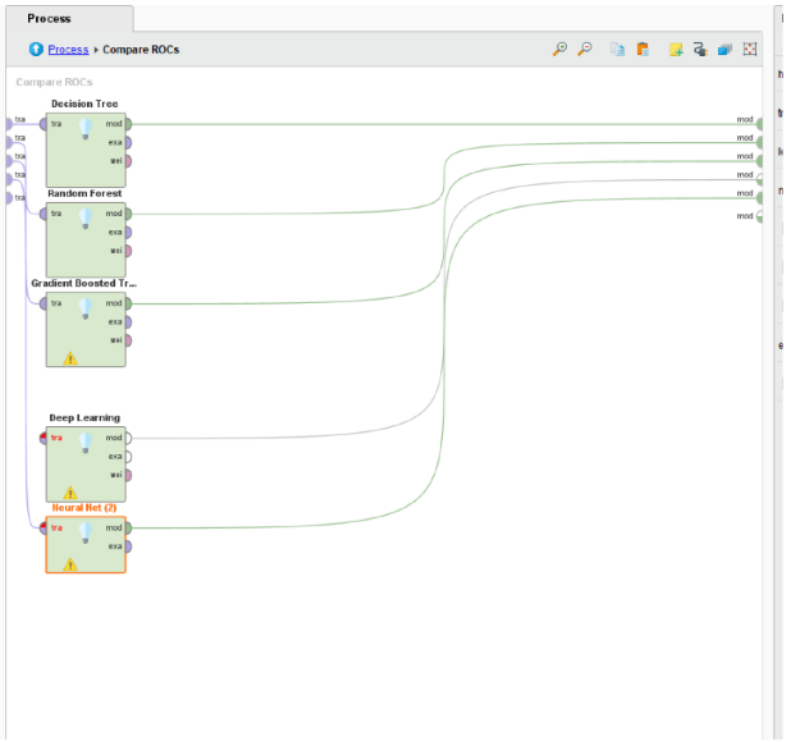
- Details of the pivot are both saved in the Excel file and are snipped below:

confidence(ye (Multiple Items

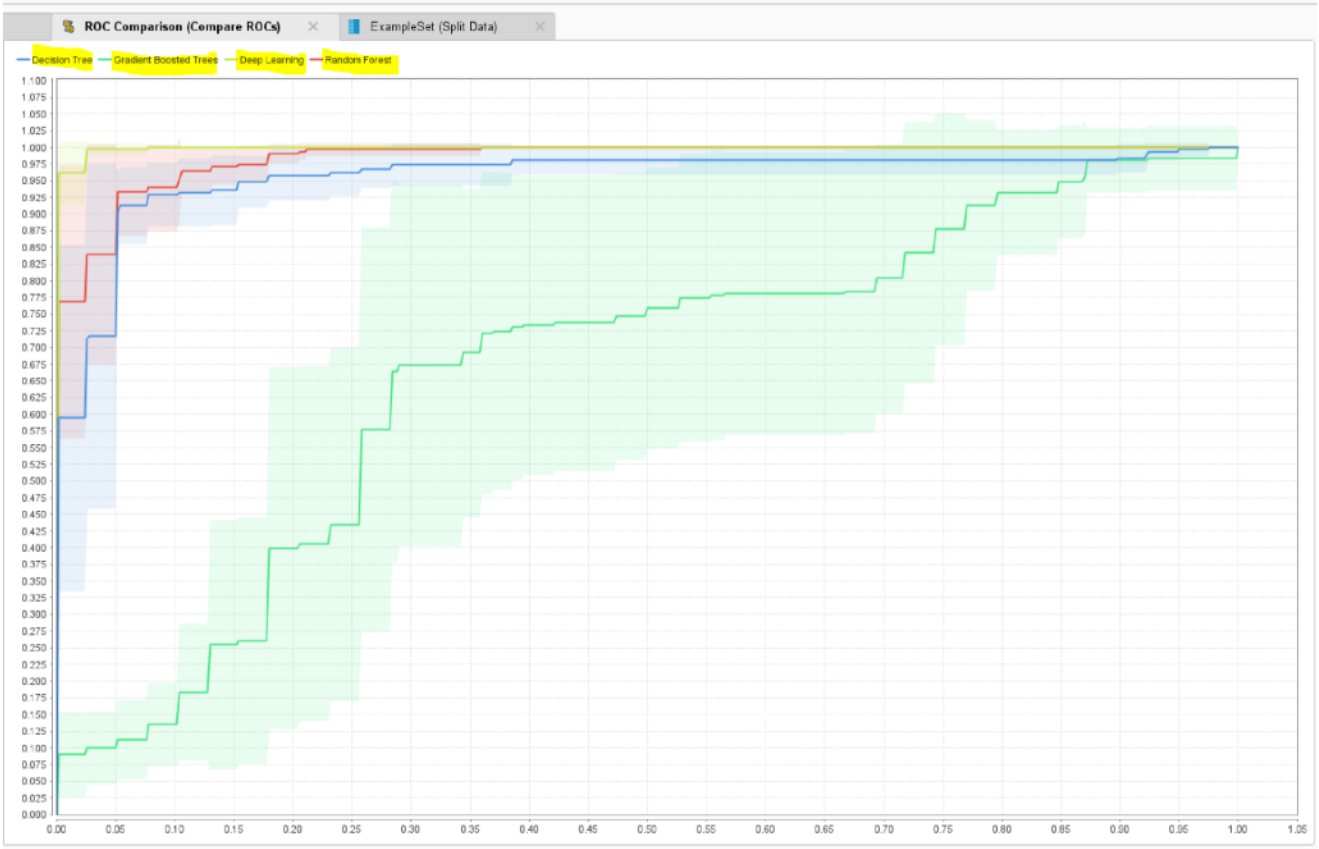
Gender	Area	Count of prediction(Response)	Sum of Sales 4 weeks	Sum of Sales 6 months	% spent per customer in 4 weeks	% spent per customer in 6 months			
female	rural	25	$ 546.00	$ 1,676.00	$ 21.84	$ 67.04	$ 5.81	27%	Female rural compared to female urban
	urban	99	$ 1,587.00	$ 5,345.00	$ 16.03	$ 53.99			
female Total		**124**	**$ 2,133.00**	**$ 7,021.00**					
male	rural	46	$ 358.00	$ 1,139.00	$ 7.78	$ 24.76	$ 14.06	64%	Female rural compared to male rural
	urban	238	$ 1,268.00	$ 5,265.00	$ 5.33	$ 22.12			
male Total		**284**	**$ 1,626.00**	**$ 6,404.00**					
Grand Total		**408**	**$ 3,759.00**	**$ 13,425.00**					

To begin the process – after importing the data and setting the role of "response," I then ran the Compare ROC operator on the following models for the Past Customers data:

- Decision Tree
- Random Forrest
- Gradient Boosted Trees
- Deep Learning
- Neural Net (did not work)

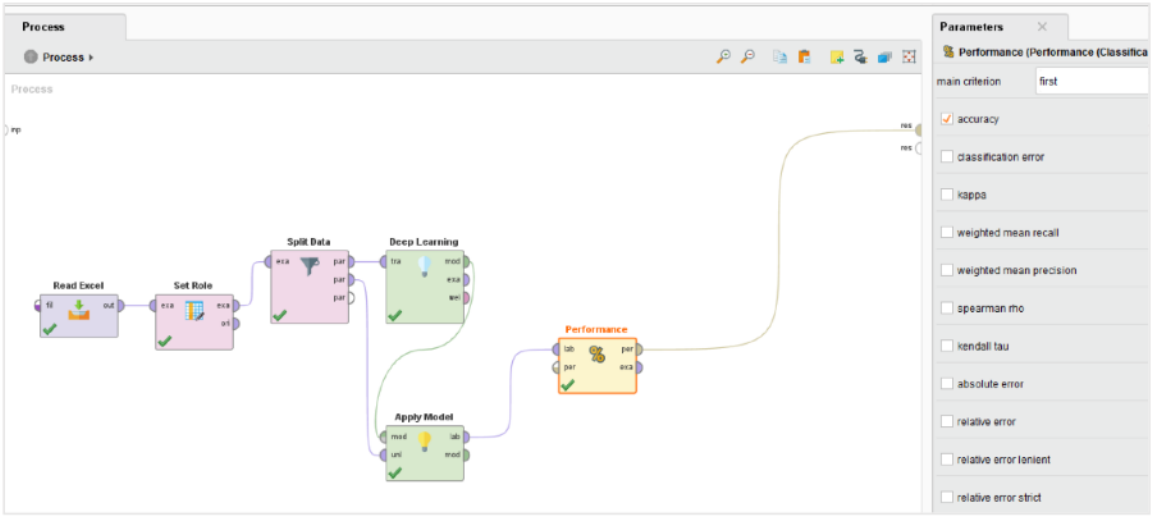


Based upon the data, the following ROC curve was displayed in the results, verifying that **Deep Learning** is the most accurate model to use in this case study:

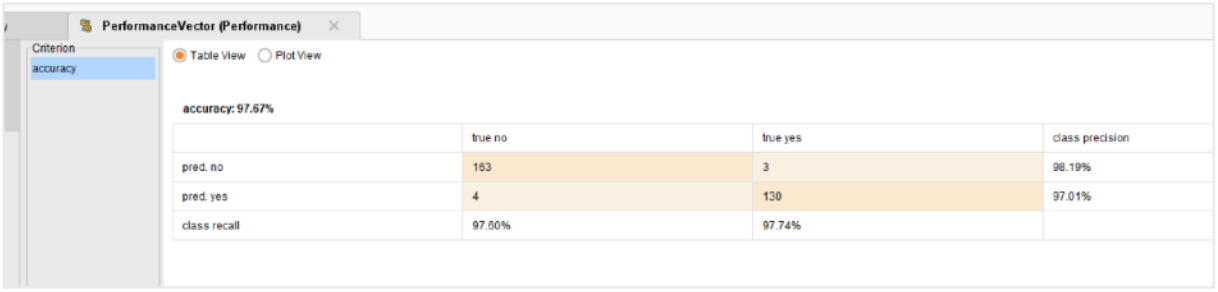


After determining Deep Learning was the model I wanted to use, I deleted the Compare ROC operator and inserted the Deep Learning operator and ran the Performance Classification operator.

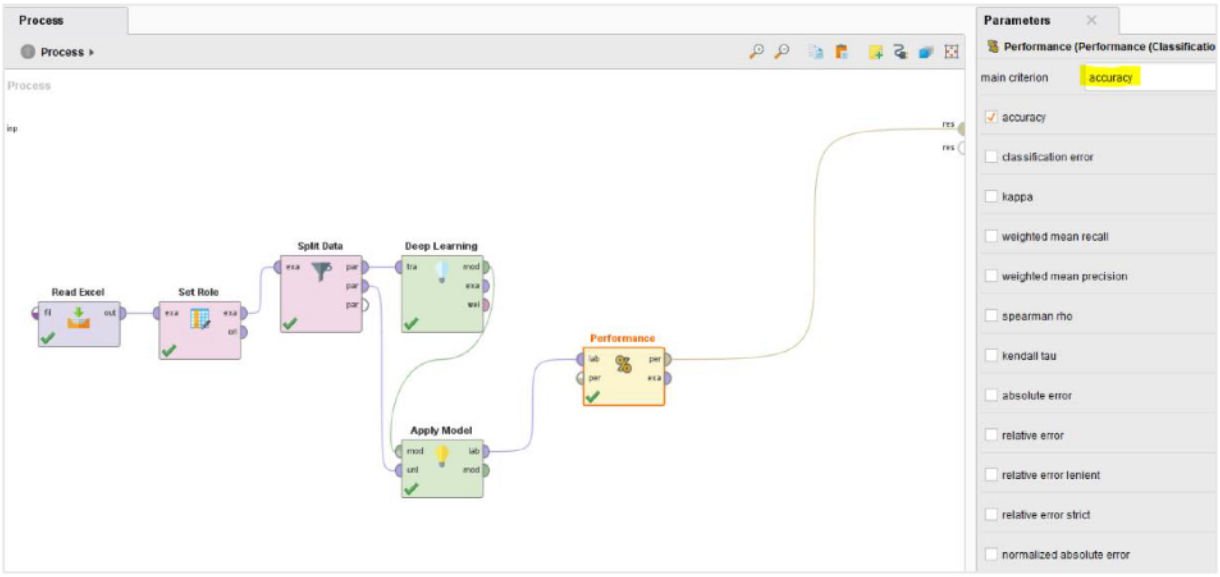
Below are a few different scenarios I ran for the Split Data ratios and corresponding Confusion Matrices:

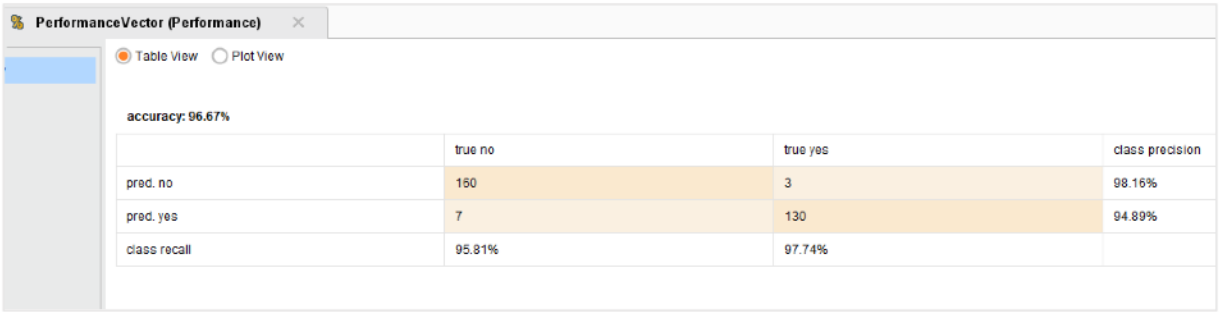


The Confusion Matrix



I also changed the parameters for the Performance Classification from Main Criterion "first" to "accuracy," and the results of the Confusion Matrix are below:





Therefore, I chose to place the parameter back to "first," as the class recall and precision were higher.

In addition, I changed the split data ratios to .80 & .20, but the accuracy dropped to 95.5%, as displayed below:

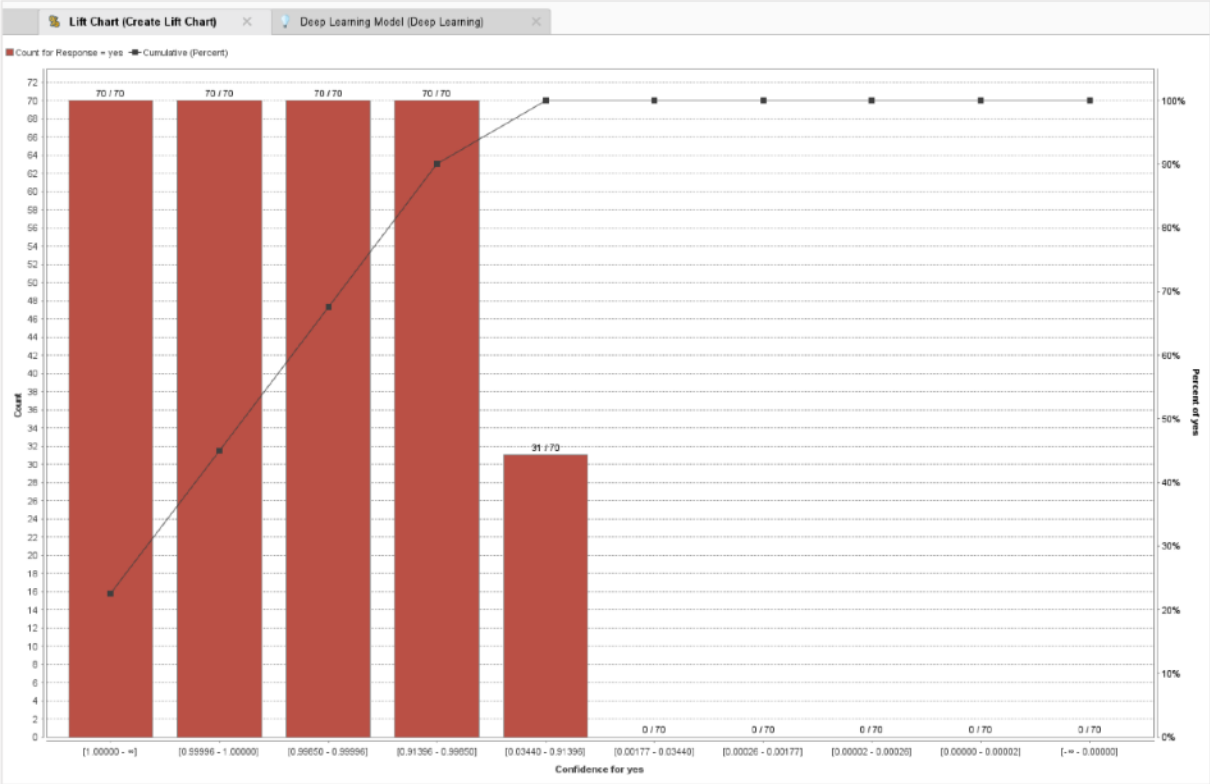
Table View Plot View

accuracy: 95.50%

	true no	true yes	class precision
pred. no	109	7	93.97%
pred. yes	2	82	97.62%
class recall	98.20%	92.13%	

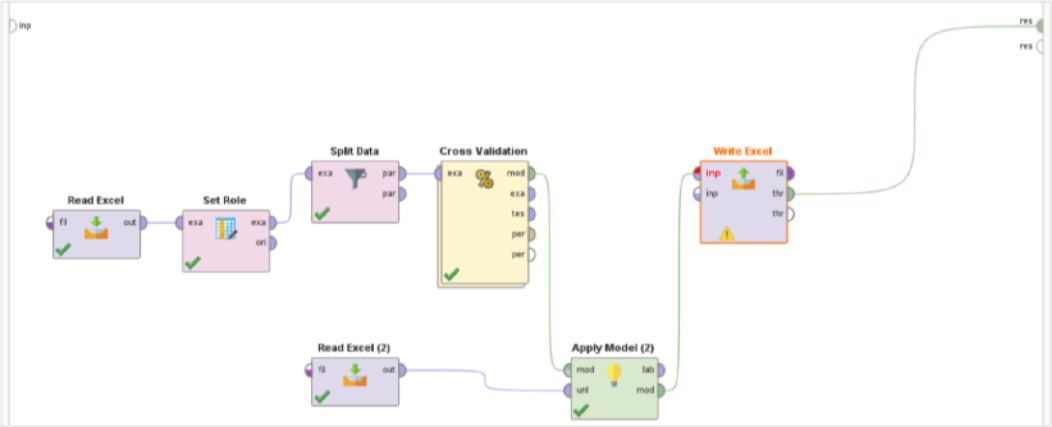
Therefore, I decided the best option was to use the ratio split of 70% / 30%.

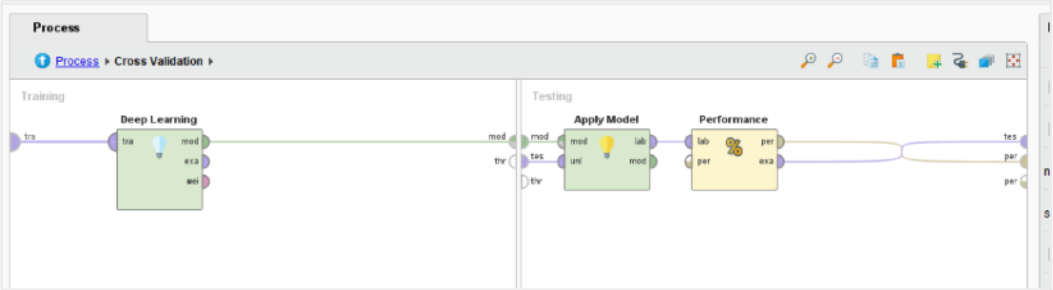
Next, I ran the Cross Validation operator to see if my results would be more accurate and ran a lift chart, as noted below:



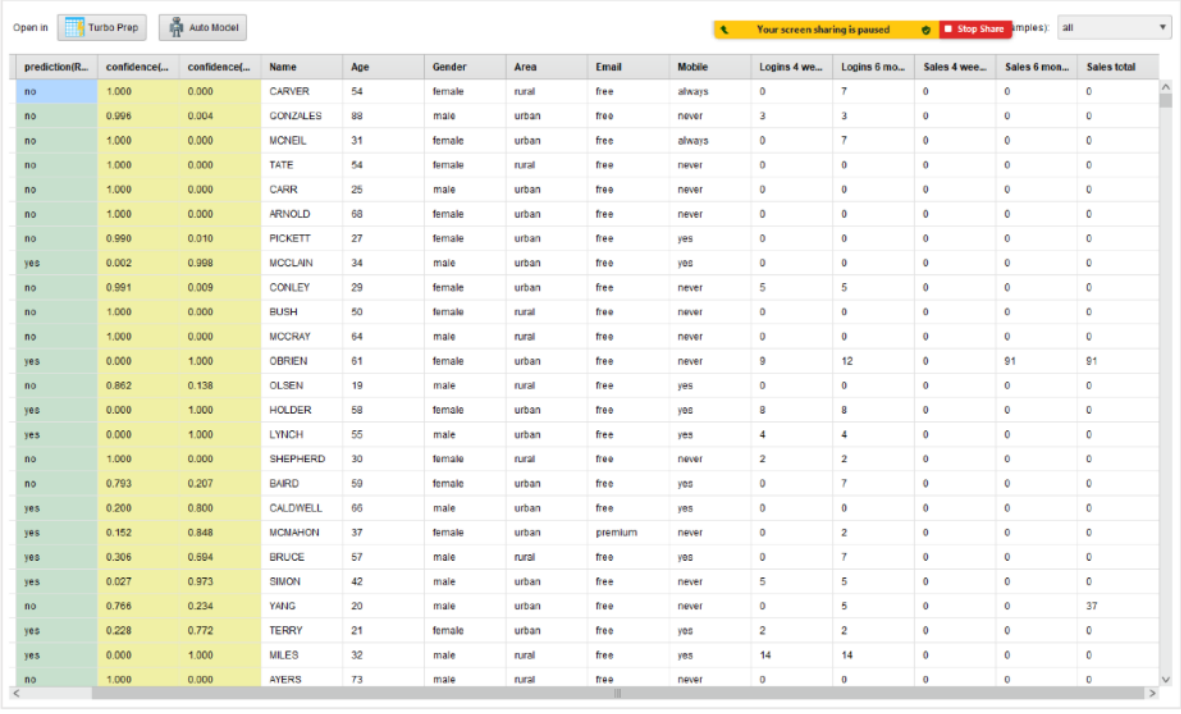
Each bar represents 10% of the customers we are planning to send promotional materials to. Based on our model, we want to look at the first three bars, as they are higher in % of confidence for "yes."

The final step in my process was to bring in the current customers' data, run the model and write the scored Excel file, as shown below:





Data results within RapidMiner prior to exporting to Excel:



prediction(R...	confidence(...	confidence(...	Name	Age	Gender	Area	Email	Mobile	Logins 4 we...	Logins 6 mo...	Sales 4 wee...	Sales 6 mon...	Sales total
no	1.000	0.000	CARVER	54	female	rural	free	always	0	7	0	0	0
no	0.996	0.004	GONZALES	88	male	urban	free	never	3	3	0	0	0
no	1.000	0.000	MCNEIL	31	female	urban	free	always	0	7	0	0	0
no	1.000	0.000	TATE	54	female	rural	free	never	0	0	0	0	0
no	1.000	0.000	CARR	25	male	urban	free	never	0	0	0	0	0
no	1.000	0.000	ARNOLD	68	female	urban	free	never	0	0	0	0	0
no	0.990	0.010	PICKETT	27	female	urban	free	yes	0	0	0	0	0
yes	0.002	0.998	MCCLAIN	34	male	urban	free	yes	0	0	0	0	0
no	0.991	0.009	CONLEY	29	female	urban	free	never	5	5	0	0	0
no	1.000	0.000	BUSH	50	female	rural	free	never	0	0	0	0	0
no	1.000	0.000	MCCRAY	64	male	rural	free	never	0	0	0	0	0
yes	0.000	1.000	OBRIEN	61	female	urban	free	never	9	12	0	91	91
no	0.862	0.138	OLSEN	19	male	rural	free	yes	0	0	0	0	0
yes	0.000	1.000	HOLDER	58	female	urban	free	yes	8	8	0	0	0
yes	0.000	1.000	LYNCH	55	male	urban	free	yes	4	4	0	0	0
no	1.000	0.000	SHEPHERD	30	female	rural	free	never	2	2	0	0	0
no	0.793	0.207	BAIRD	59	female	urban	free	yes	0	7	0	0	0
yes	0.200	0.800	CALDWELL	66	male	urban	free	yes	0	0	0	0	0
yes	0.152	0.848	MCMAHON	37	female	urban	premium	never	0	2	0	0	0
yes	0.306	0.694	BRUCE	57	male	rural	free	yes	0	7	0	0	0
yes	0.027	0.973	SIMON	42	male	urban	free	never	5	5	0	0	0
no	0.766	0.234	YANG	20	male	urban	free	never	0	5	0	0	37
yes	0.228	0.772	TERRY	21	female	urban	free	yes	2	2	0	0	0
yes	0.000	1.000	MILES	32	male	rural	free	yes	14	14	0	0	0
no	1.000	0.000	AYERS	73	male	rural	free	never	0	0	0	0	0

After writing the scored data to Excel, I then sorted and filtered the data for the predicted value of "yes", with a confidence (yes) of 1. I received 408 potential customers who would most likely utilize the promotional materials sent in the marketing campaign for Best Buy.

In summary, after further analyzing the predictive model tools, I would recommend the prioritization of the following for the direct marketing campaign:

- I would focus my efforts and corresponding direct marketing expense in descending order for the following:
    - Female in rural areas
    - Females in urban areas
    - Males in rural areas
    - Males in urban areas