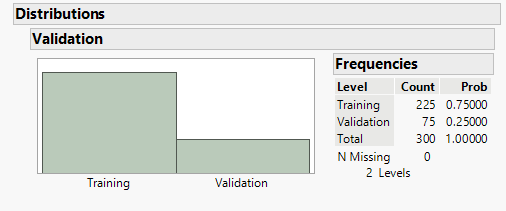
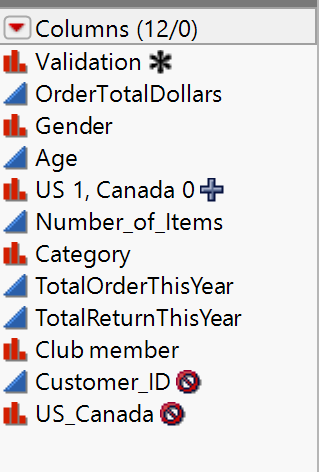
**Final Exam, INFO 3200, Spring 2022**

* In doing this exam, you are permitted to consult your text, notes, and any computer files you brought with you to the exam.
* During the exam, you are **not** permitted to discuss it, physically or electronically, with any other person.
* You have to take the exam alone and not have anybody in the room with you.
* Turning in your answers to the exam constitutes a representation that you neither gave nor received prohibited assistance during the exam.
* Turn in only one answer to each problem you do. If you hand in more than one answer to a problem, none of them will be graded.
* Name the files with your name (e.g. “YJLee.doc”).
* Submit both your word and JMP solution files on Canvas. The exam has 100 points.

The purpose of this task is to construct models **to predict the total dollar amount of order of a customer in a store (OrderTotalDollars)**:

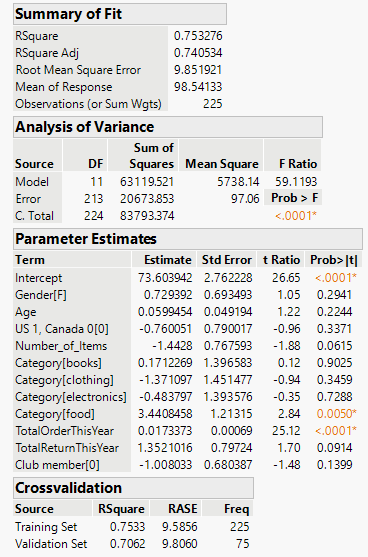
|  |  |  |
| --- | --- | --- |
| Variable | Description | Categorical? |
| Customer\_ID | Customer identification number in the store |  |
| Age | Age of customer |  |
| Gender | Gender of customer (M=male, F=female) | Yes |
| US\_Canada | 1 if the customer is from US, otherwise Canada | Yes |
| Number\_of\_Items | The number of items in the customer order |  |
| Category | Product category (books, clothing, electronics, food, or music) | Yes |
| TotalOrderThisYear | the total dollar amount of orders during this year from the customer |  |
| TotalReturnThisYear | the total number of returns during this year from the customer |  |
| Club member | 1 if the customer is club member, otherwise 0 | Yes |
| OrderTotalDollars (Response/Target) | the total dollar amount of the customer’s order in this transaction |  |

* Before performing any analysis make sure that the binary/categorical variables labeled in the table above are read as *Nominal* in JMP.
* Make sure you have *a validation column* before conducting any data-mining modeling.
* To support your answer, copy and paste a screenshot of each answer you create by JMP in each answer.

1. Show the distribution of your training and validation/testing datasets (5 pts)
2. You can remove any variables for particular models that are not useful. You can create a categorical variable from one of the continuous variables if you would like. Then use this as an alternative to the actual continuous variable. With your justifications, show what variables you have removed and what other data preparation you have conducted if you have done (5 pts)

The variables that I will not be including are Customer\_ID, as it is unique and will not help with prediction. I will also not be including US\_Canada, as I have created a new binary column for the variable named US 1, Canada 0. The rest of the variables I will use as they are as they all seem to be reasonable predictors.

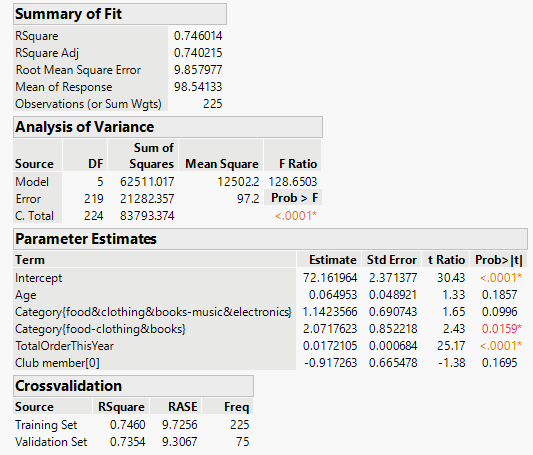
1. Create Two Linear Regression models (15 pts)
   1. The first model includes all variables you have chosen earlier



Model:

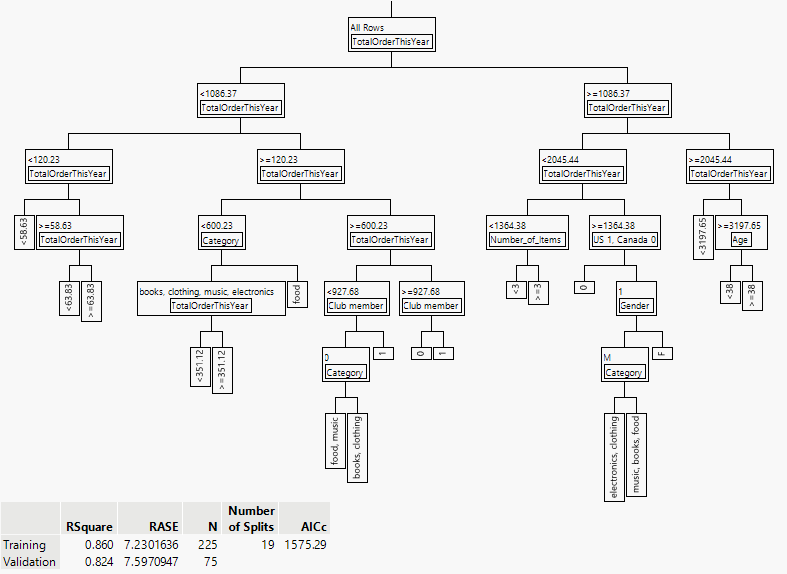
OrderTotalDollars = 73.603942 + 0.729392 \* Gender[F] + 0.0599454 \* Age -0.760051 \* US1, Canada0[0] – 1.4428 \* Number\_of\_Items + 0.1712269 \* Category[books] – 1.371097 \* Category[clothing] – 0.483797 \* Category[electronics] + 3.4408458 \* Category[food] + 0.0173373 \* TotalOrderThisYear + 1.3521016 \* TotalReturnThisYear – 1.008033 \* Club member[0]

* 1. The second one should be based on model selection (removing some variables based on p-value criteria)

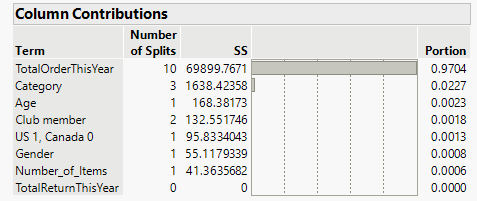
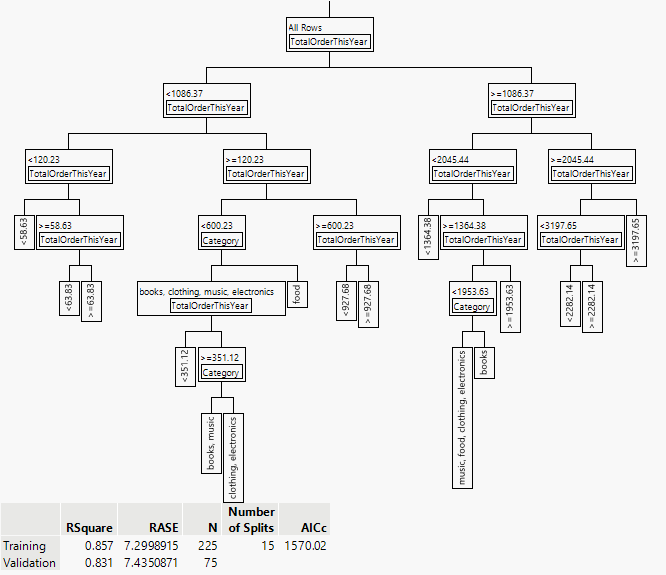


Model:

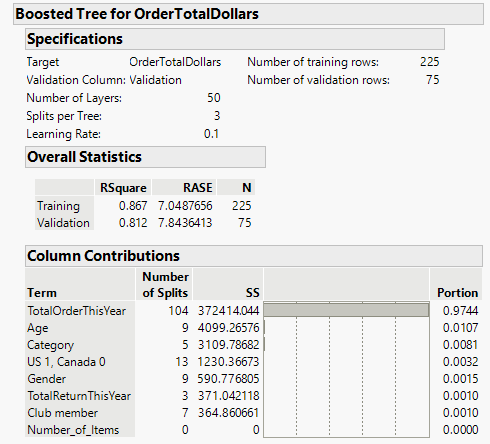
OrderTotalDollars = 72.161964 + 0.064953 \* Age + 1.1423566 \* Category{food&clothing&books-music&electronics} + 2.0717623 \* Category{food-clothing&books} + 0.0172105 \* TotalOrderThisYear -0.917263 \* Club member[0]

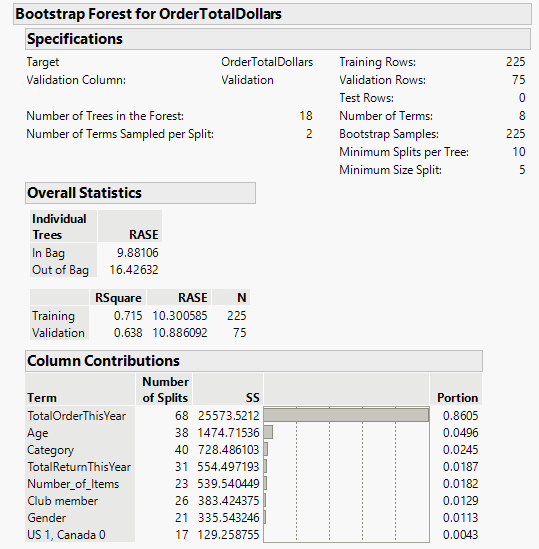
1. Create Two Regression Decision Trees (15 pts)
   1. The first Tree model includes all predictors you have chosen earlier
   2. The second Tree model includes most contributing predictors or pruning or splitting.

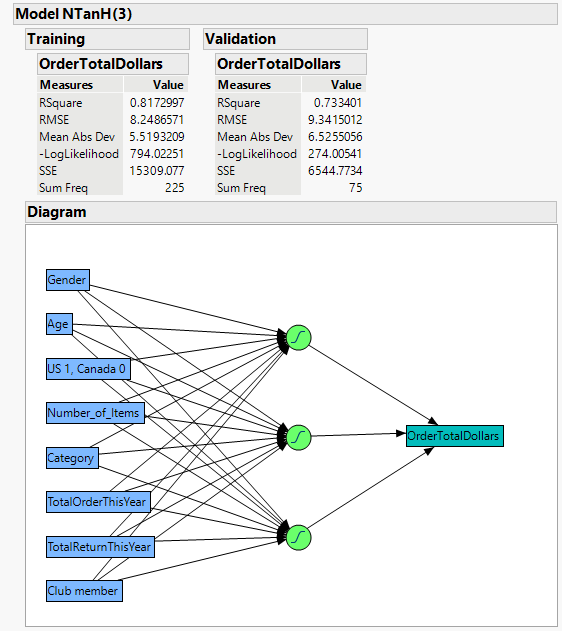
For second decision tree, I just used TotalOrderThisYear and Category, as those are the two variables that contributed the most to the model.



1. Create a Boosted Tree Model with all predictors you have chosen, and all default specification values (10 pts)



1. Create a Bootstrap Forest Model with all predictors you have chosen, and all default specification values (10 pts)
2. Create a Neural Network Model with all predictors you have chosen, and all default specification values (10 pts)

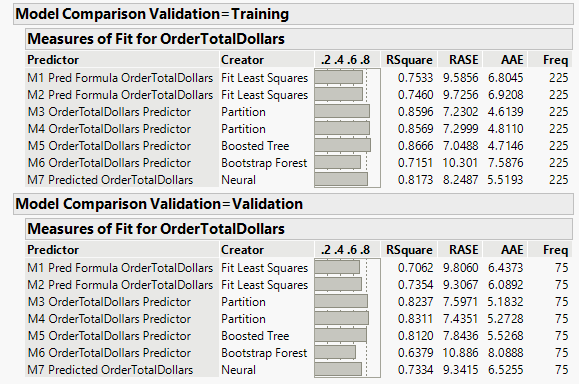


1. Please fill in the table below with predicted values of the first customer’s OrderTotalDollars (Customer\_ID =1) in the data from the seven models you have created (10 pts)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Customer\_ID=1 | M1 | M2 | M3 | M4 | M5 | M6 | M7 |
| Predicted  OrderTotalDollars | 109.0487 | 106.8690 | 106.4785 | 103.8714 | 105.6685 | 103.5958 | 108.8324 |

1. Based on a summary table of the different accuracy and error measures, use good reporting techniques to write up your results and compare and contrast how the models differed between your different techniques. Make sure to state which variables you used in your model and, where you can, state which variables were left in your model or what were the highest column contributions. (15 pts)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Predictors | R-Squared | RMSE(RASE) | MAE (Mean Abs Error = AAE) |
| Linear Regression #1 | Gender[F], Age, US 1, Canada 0[0], Number\_of\_Items, Category[books], Category[electronics], Category[food], TotalOrderThisYear, TotalReturnThisYear, Club member[0] | 0.7062 | 9.806 | 6.4373 |
| Linear Regression #2 | Age, Category{food&clothing&books-music&electronics}, Category{food-clothing&books}, TotalOrderThisYear, Club member[0] | 7.354 | 9.3067 | 6.0892 |
| Decision Tree #1 | Gender[F], Age, US 1, Canada 0[0], Number\_of\_Items, Category[books], Category[electronics], Category[food], TotalOrderThisYear, TotalReturnThisYear, Club member[0] | 0.8237 | 7.5971 | 5.1832 |
| Decision Tree #2 | TotalOrderThisYear, Category | **0.8311** | **7.4351** | **5.2728** |
| Boosted Tree | Gender[F], Age, US 1, Canada 0[0], Number\_of\_Items, Category[books], Category[electronics], Category[food], TotalOrderThisYear, TotalReturnThisYear, Club member[0] | 0.8120 | 7.8436 | 5.5268 |
| Bootstrap Forest | Gender[F], Age, US 1, Canada 0[0], Number\_of\_Items, Category[books], Category[electronics], Category[food], TotalOrderThisYear, TotalReturnThisYear, Club member[0] | 0.6379 | 10.886 | 8.0888 |
| Neural Network | Gender[F], Age, US 1, Canada 0[0], Number\_of\_Items, Category[books], Category[electronics], Category[food], TotalOrderThisYear, TotalReturnThisYear, Club member[0] | 0.7334 | 9.3415 | 6.5255 |



1. Conclude by stating which model you think is best and why (5 pts)

The best model to use to predict OrderTotalDollars is the Decision tree that only used the variables of TotalOrderThisYear and Category. This model resulted in the highest R-Squared value of 0.8311, the lowest root mean squared error of 7.4351, and the second smallest Mean Absolute error of 5.2728. This means that this decision tree model had the best fit for validation data, the lowest RMSE rate and second lowest MAE rate, which means it is one of the most accurate.