**Data Preparation**

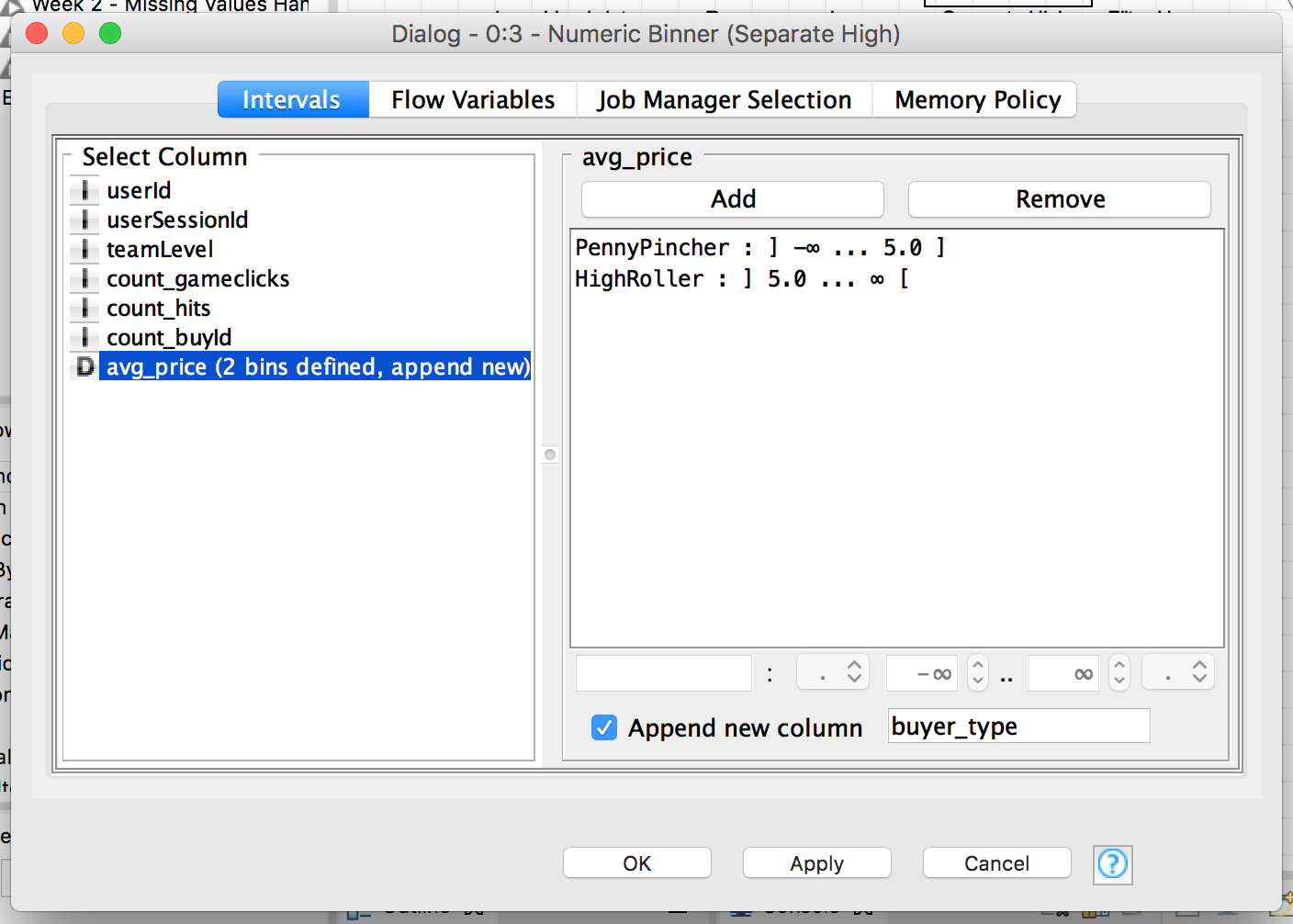
Analysis of combined\_data.csv

Sample Selection

|  |  |
| --- | --- |
| **Item** | **Amount** |
| # of Samples | 4619 |
| # of Samples with Purchases | 1411 |

Attribute Creation

A new categorical attribute was created to enable analysis of players as broken into 2 categories (HighRollers and PennyPinchers). A screenshot of the attribute follows:



High rollers are defined as those who purchase items over $5.00. Defining a new column based on the avg\_price allows us to classify users accordingly.

The creation of this new categorical attribute was necessary because:

Our goal is to understand the attributes of who makes large purchases. This categorical variable is what we are going to base our decision tree upon.

Attribute Selection

The following attributes were filtered from the dataset for the following reasons:

|  |  |
| --- | --- |
| **Attribute** | **Rationale for Filtering** |
| userId | Not relevant in determining the user’s behavior. |
| userSessionId | Not relevant in determining the user’s behavior. |
| avg\_price | This variable was used to create the categorical variable “buyer\_type”, the variable we are trying to predict based on other data elements. Therefore, we do not want to include this as a criteria in our decision tree model. |

**Data Partitioning and Modeling**

The data was partitioned into train and test datasets.

The **training** data set was used to create the decision tree model.

The trained model was then applied to the **test** dataset.

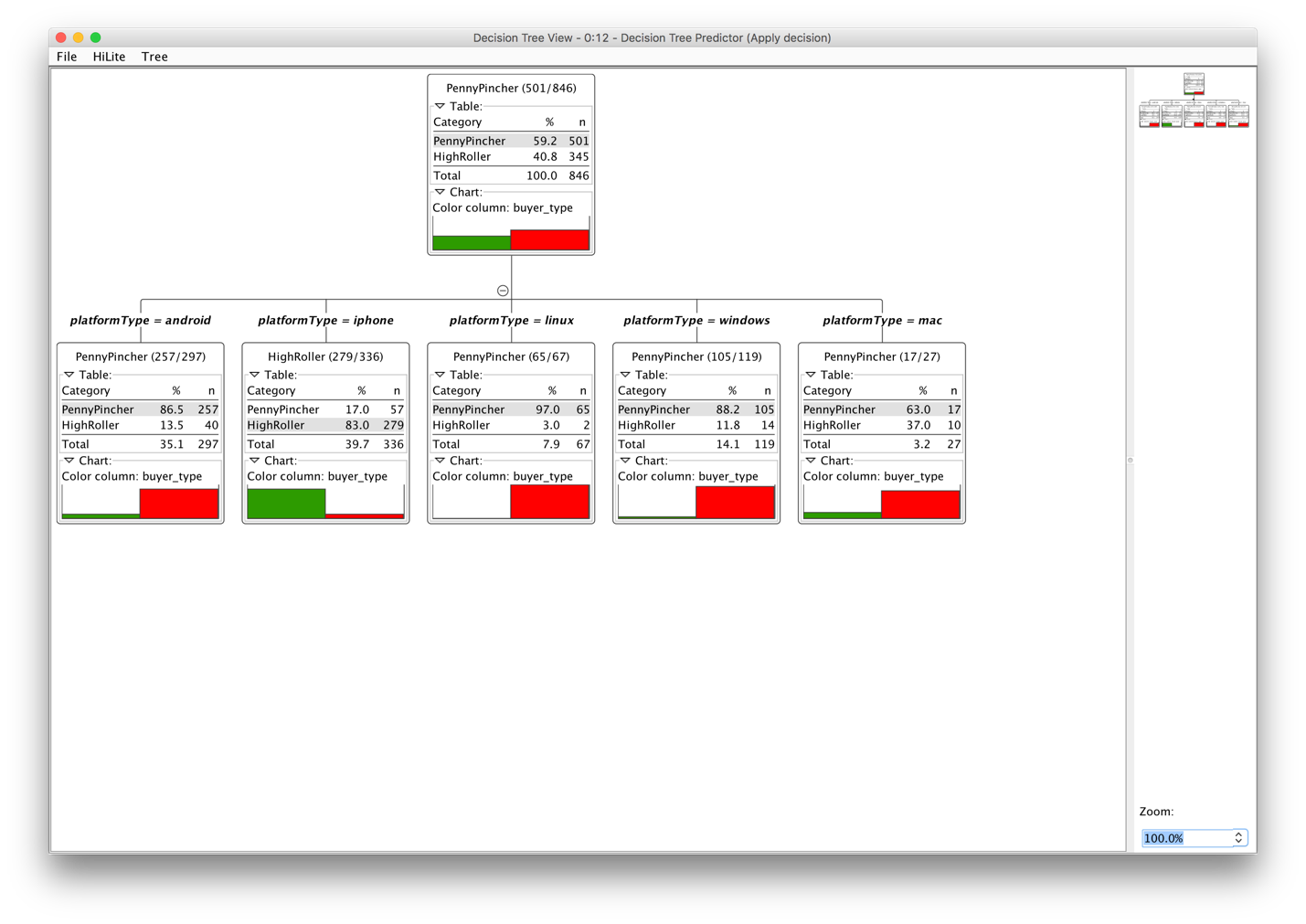
This is important because…

**Partitioning the data set into training and test data allows us to verify the accuracy of the trained model. If we used all the data for training the model, we would have no idea how the model performs.**

When partitioning the data using sampling, it is important to set the random seed because…

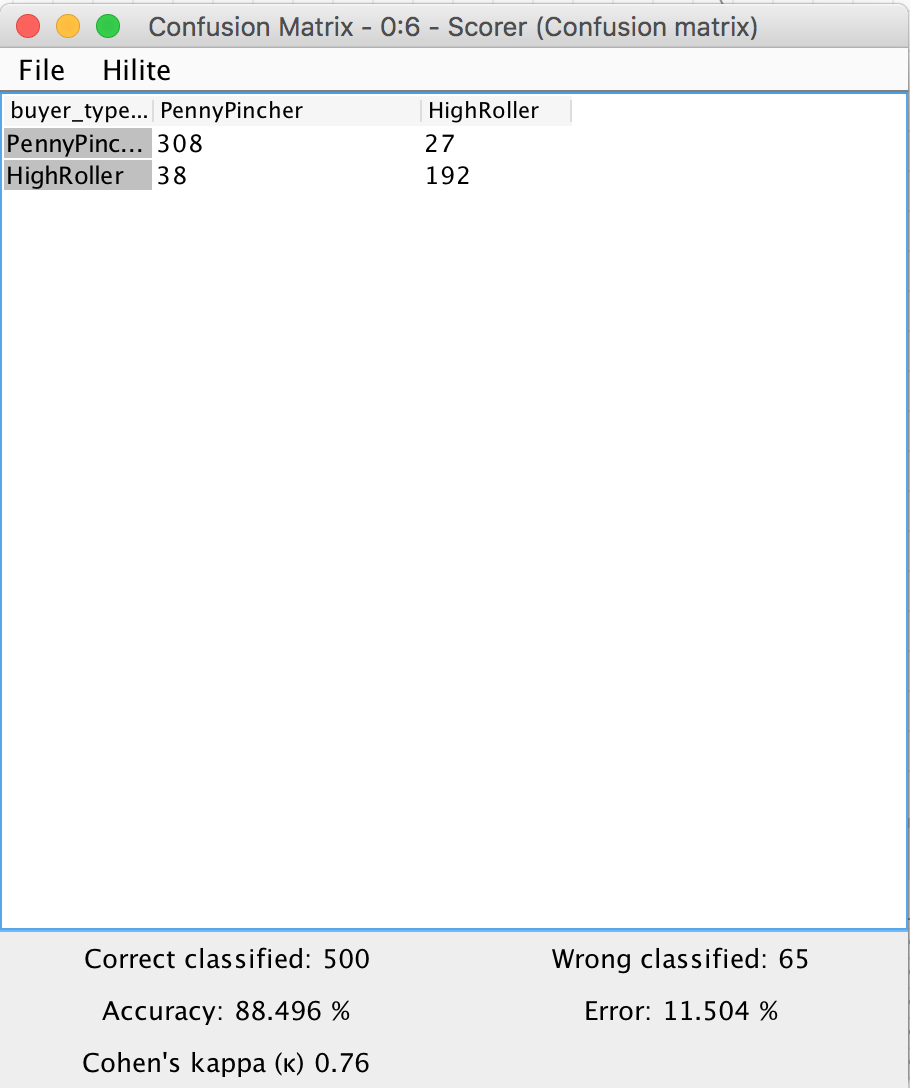
**A random seed will partition the data set consistently. This allows you to obtain reproducible results each time you run the partition. When validating the accuracy of different models, you need to keep the data used to test the models consistent. If the test data is inconsistent, you can’t compare the accuracy between models.**

A screenshot of the resulting decision tree can be seen below:



**Evaluation**

A screenshot of the confusion matrix can be seen below:

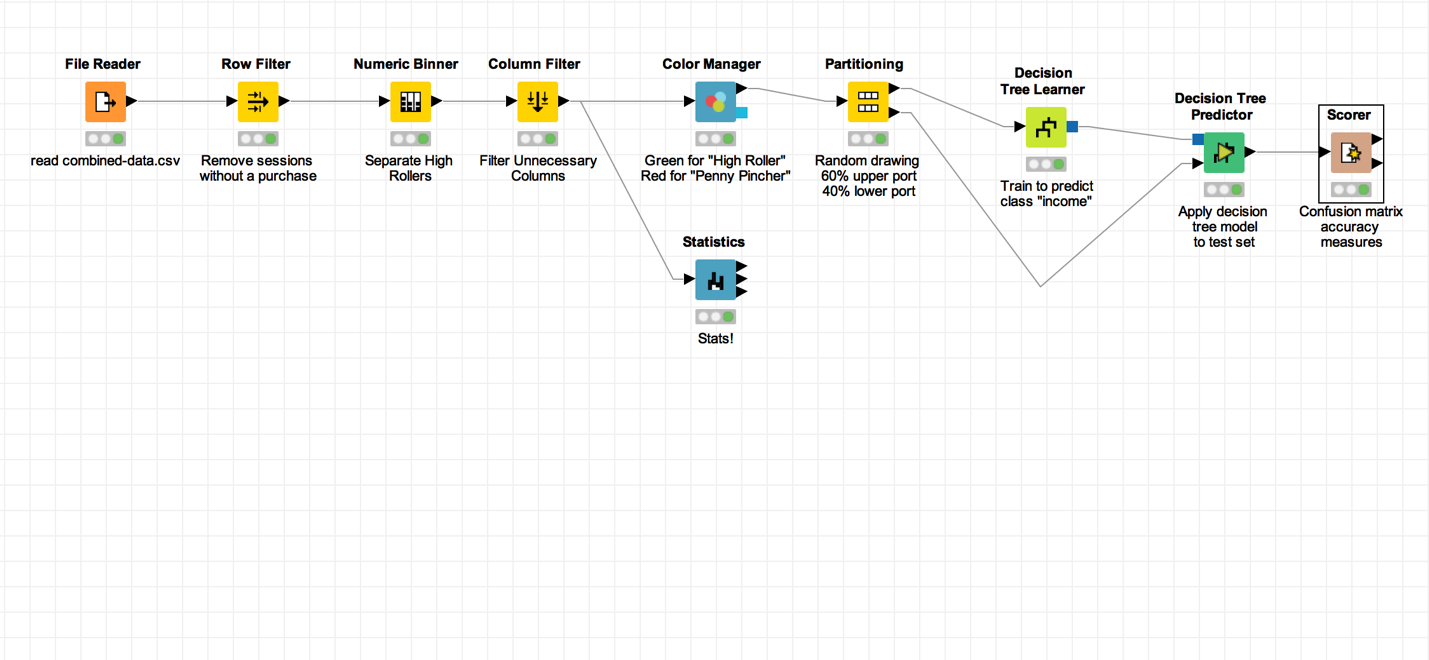


As seen in the screenshot above, the overall accuracy of the model is **88.496%**

**When the buyer type is PennyPincher, the model classified correctly 308 times and incorrectly 27 times. When the byer type is HighRoller, the model classified correctly 192 times and incorrectly 38 times.**

**Analysis Conclusions**

The final KNIME workflow is shown below:



What makes a HighRoller vs. a PennyPincher?

**The OS used. Users who are HighRoller use iOS. PennyPincher use Android, Mac, Win, Linux.**

|  |
| --- |
| **Specific Recommendations to Increase Revenue** |
| 1. Target promotions to iOS users. |
| 2. Target future product develop towards iOS. |