

Data Preparation

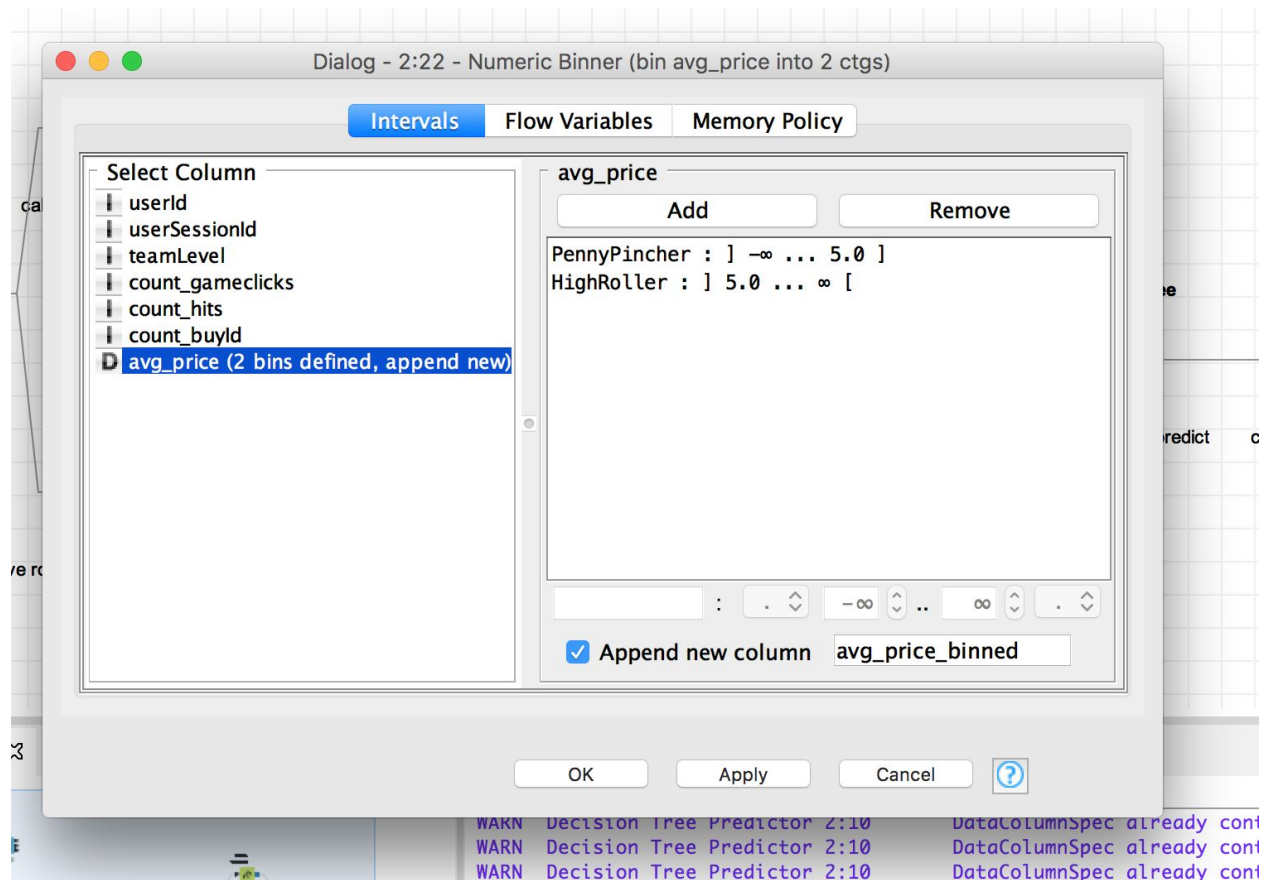
Analysis of combined_data.csv

Sample Selection

Item	Amount
# of Samples	4619
# of Samples with Purchases	1411

Attribute Creation

A new categorical attribute was created based on categorization of avg-price into 2 bins to enable analysis of players as broken HighRollers and PennyPinchers. A screenshot of the attribute follows:



New column named “avg_price_binned” is the new attribute where buyid > 5 belongs to “HighRollers” because the prices of them are over \$5, while buyid <=5 belongs to “PennyPinchers” because the prices of those are not over \$5.

The creation of this new categorical attribute was necessary because this is a classification problem, we should not use a continuous value field like avg-price.

Attribute Selection

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

Attribute	Rationale for Filtering
avg_price	We don't need the average price anymore since we have a new categorized attribute avg_price_binned based on this.
user_Id	Don't need this since it's just a computer generated number
user_Session_Id	Don't need this since it's just a computer generated number