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:

Binned Data - 0:13 - Numeric Binner

File Hilite Navigation View

Table "default" - Rows: 1411 | Spec - Columns: 9 | Properties | Flow Variables

Row ID	I userId	I userSe...	I teamLevel	S platfor...	I count_...	I count_...	I count_...	D avg_price	S avg_pri...
Row4	937	5652	1	android	39	0	1	1	PennyPinchers
Row11	1623	5659	1	iphone	129	9	1	10	HighRollers
Row13	83	5661	1	android	102	14	1	5	HighRollers
Row17	121	5665	1	android	39	4	1	3	PennyPinchers
Row18	462	5666	1	android	90	10	1	3	PennyPinchers
Row31	819	5679	1	iphone	51	8	1	20	HighRollers
Row49	2199	5697	1	android	51	6	2	2.5	PennyPinchers
Row50	1143	5698	1	android	47	5	2	2	PennyPinchers
Row58	1652	5706	1	android	46	7	1	1	PennyPinchers
Row61	2222	5709	1	iphone	41	6	1	20	HighRollers
Row68	374	5716	1	android	47	7	1	3	PennyPinchers
Row72	1535	5720	1	iphone	76	7	1	20	HighRollers
Row73	21	5721	1	android	52	2	1	3	PennyPinchers
Row101	2379	5749	1	android	62	9	1	3	PennyPinchers
Row122	1807	5770	1	iphone	177	25	2	7.5	HighRollers
Row127	868	5775	1	iphone	54	5	1	10	HighRollers
Row129	1567	5777	1	android	27	4	2	4	PennyPinchers
Row131	221	5779	1	iphone	37	2	1	20	HighRollers
Row135	2306	5783	1	android	67	5	1	1	PennyPinchers
Row137	1065	5785	1	iphone	37	5	2	11.5	HighRollers
Row140	827	5788	1	iphone	75	5	1	20	HighRollers
Row150	1304	5798	1	mac	71	9	2	11.5	HighRollers
Row158	1264	5806	1	linux	81	12	1	5	HighRollers
Row159	1026	5807	1	iphone	52	10	1	20	HighRollers
Row163	649	5811	1	linux	51	9	1	1	PennyPinchers
Row169	1958	5817	1	android	40	3	1	20	HighRollers
Row172	1300	5820	1	android	58	1	2	3	PennyPinchers
Row186	178	5834	1	iphone	54	6	1	20	HighRollers
Row196	670	5844	1	iphone	38	3	2	20	HighRollers
Row207	208	5855	1	iphone	32	3	1	20	HighRollers
Row210	157	5858	1	iphone	32	2	1	10	HighRollers
Row212	2221	5860	1	iphone	191	18	2	11.5	HighRollers
Row215	471	5863	1	iphone	45	6	2	15	HighRollers
Row218	1234	5866	1	android	46	3	1	10	HighRollers
Row222	371	5870	1	android	53	9	1	3	PennyPinchers
Row232	2146	5880	1	linux	46	7	1	2	PennyPinchers
Row239	935	5887	1	iphone	57	2	1	10	HighRollers
Row241	165	5889	1	iphone	49	3	1	5	HighRollers

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 avgprice.

Attribute Selection

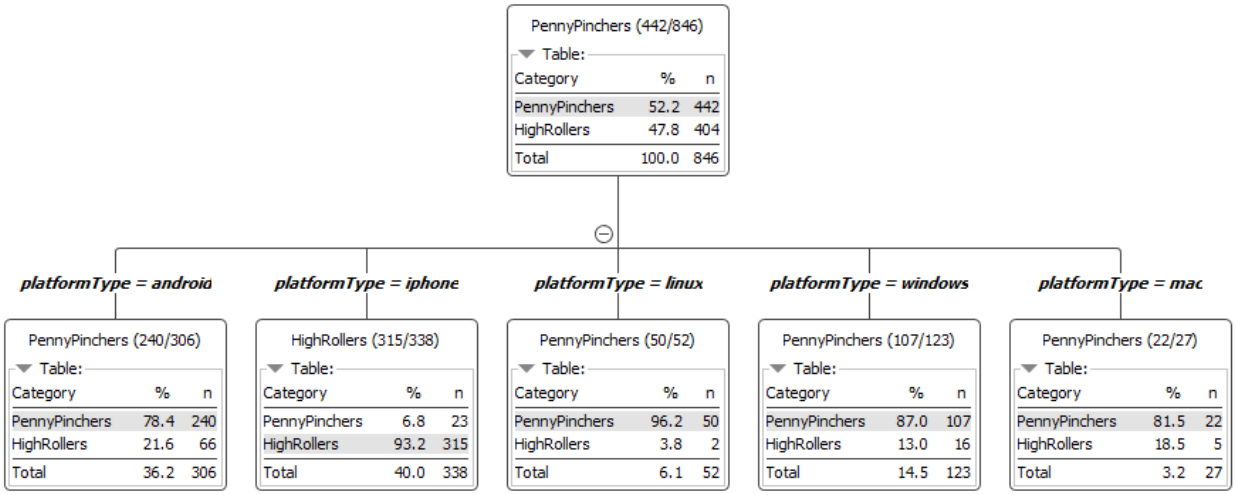
The following attributes were filtered from the dataset for the following reasons:

Attribute	**Rationale for Filtering**
avg_price	We don't need the average price anymore since we have a new
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

Data Partitioning and Modeling

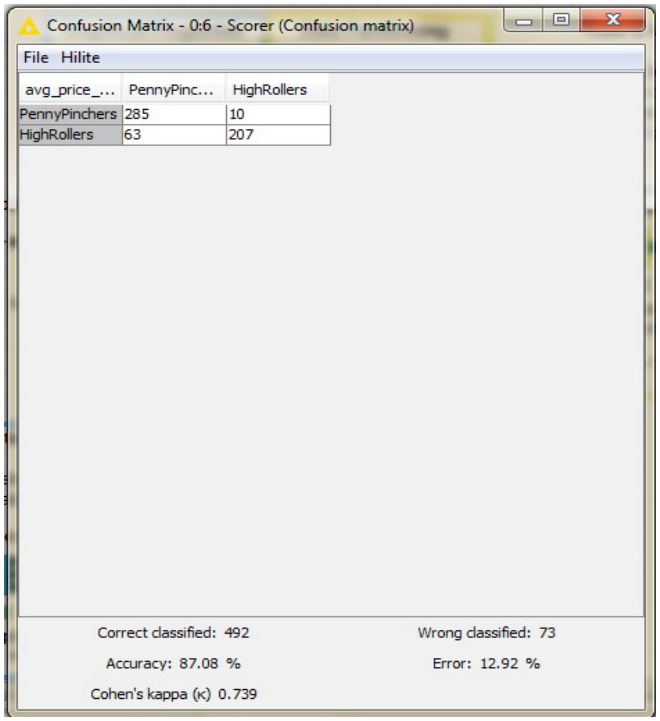
The data was partitioned into train and test datasets.
The <Fill In> data set was used to create the decision tree model.
The trained model was then applied to the test dataset.
This is important because when we do data analysis, we should test our model on a data set that was not used to train the model . After a model has been processed by using the training set, you test the model by making predictions against the test set.
When partitioning the data using sampling, it is important to set the random seed to make sure the partition is the same every time you run the program . That is needed when you need a reproducible result.

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



Evaluation

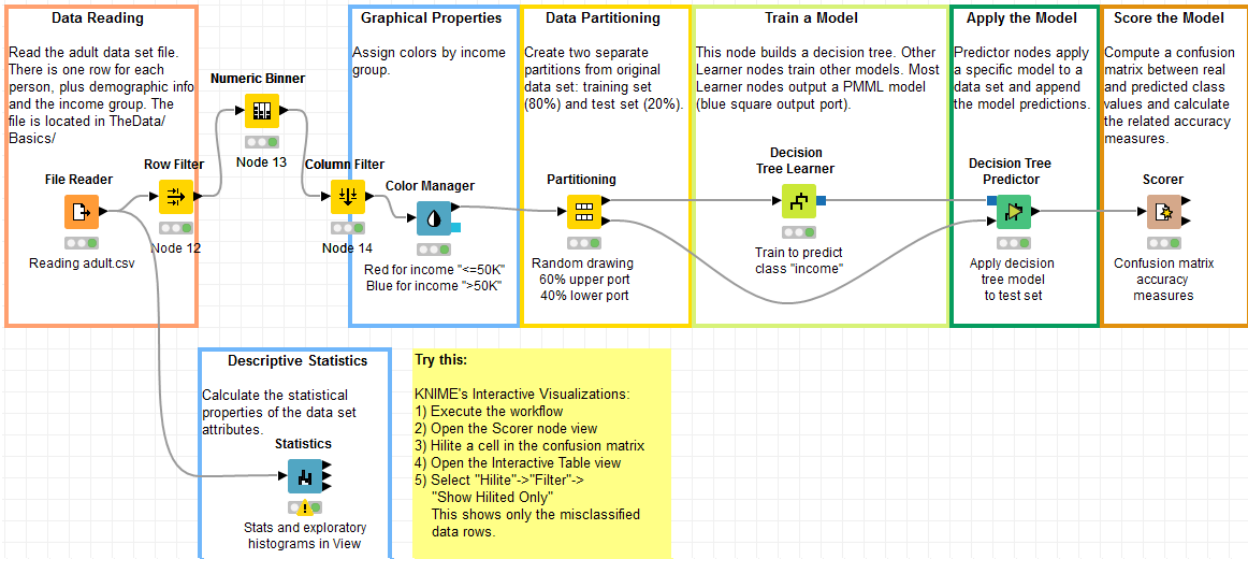
A screenshot of the confusion matrix can be seen below:



As seen in the screenshot above, the overall accuracy of the model is
207 HighRollers have been predicted correctly.
10 HighRollers have been predicted incorrectly.
285 PennyPinchers have been predicted correctly.
63 PennyPinchers have been predicted incorrectly.

Analysis Conclusions

The final KNIME workflow is shown below:



What makes a HighRoller vs. a PennyPincher?
iPhone users are HighRollers (93.2%) and other platformType users are PennyPinchers (6.8%).

Specific Recommendations to Increase Revenue
1. Show more ads to iPhone users.
2. Increase ads price for iPhone platform device.