Analysing IRI Marketing Dataset

Product Focus: Beer

By: Abhijeet Gulati, Conway Wong, John Gill, Jordan Levin, Kevin Dyer

Summary

The IRI data set was analyzed with the goal to forecast future sales of beer. Data from years 2008-2010 were utilized in creating daily and weekly models. During the exploration a number of techniques were used to derive the top features for making these predictions.

Methodology

The CRISP-DM methodology was applied with some modifications:

- Deployment stage was not performed.
- Evaluation was not a separate step but part of each modelling iteration.

Business Understanding

Goals that were chosen to focus analysis on:

1. Discover the important features for predicting transaction price of beer
2. Assessing if and which promotions lead to difference in sales revenue
3. Creating a model that could predict future sales revenue

Data Understanding

After reading the IRI Marketing Dataset file description outline, our group decided on the following sets of data to use in our analysis.

Panelist Data Sets

The single most important set of data for our project was the panelist data files. The initial analysis of the data revealed that the panelist data files from 2008 to 2011 allowed us to associate transactions with down to the minute granularity.

Store Data Sets

The other identified data files deemed critical for analysis were the store data files. These files contained the data about what promotions were used in a sale. It was hypothesized that store promotions/displays would be an important feature of beer sales during the modeling phase.

Delivery Stores

Although these files contain information about the regions of the stores which may be important features, it was determined these files would not be important for predicting daily sales. Since the chosen data from the panelist files were from only two regions, linking the data from the delivery store files was not necessary.

Product Attributes

Initial thoughts during data exploration was to use features from the product attribute files to predict sales of a particular beer item. After further investigation, it was determined that the number of transactions present in the panelist data was centered around a small subset of items. Therefore, the product attribute data files were not used in modeling.

Data Preparation

Parse/Process Panel Files

The panelist data files for years 2008 to 2011 were read into a pandas DataFrame, and feature engineering was performed to convert data into a statistically relevant format.

- MINUTE and WEEK => MONTH, DAY, YEAR, HOUR, MINUTE, DAYOFWEEK, and WEEKNUM
- COLUPC was converted into SY, GE, VEND, and ITEM to match the store format for merging data
- OUTLET was hot 1 encoded to split enumerations into integer fields:
    - 0=GK
    - 1=DK
    - 2=MK
    - 3=KK

After parsing and feature engineering, the panel DataFrame consisted of the following features:

```
PANID, WEEK, UNITS, OUTLET, DOLLARS, IRI_KEY, COLUPC, START_TIME_EPOCH_S,
END_TIME_EPOCH_S, TRANSACTION_TIME_EPOCH_S, MONTH, DAY, YEAR, HOUR_OF_DAY,
MINUTE, DAYOFWEEK, WEEKNUM, SY, GE, VEND, ITEM
```

Parse/Process Store Files

Store data files for years 2008 to 2011 were read into a pandas DataFrame with the resulting features:

```
WEEK, IRI_KEY, SY, GE, VEND, ITEM, F, D, PR
```

Merge Panel and Store Data

The panel and store data frames were merged into a single dataframe using key features WEEK, IRI_KEY, SY, GE, VEND, and ITEM. After merging, features F and D were hot-one encoded to split out enumerations to new integer fields.

Abnormalities/Outliers

During data preparation, it was discovered that only 15034 of the 47711 unique panelist (weekly) transactions could be joined with promotional features from the store data files. Many of the stores associated with panelist transactions were from manually reported stores (i.e. stores starting with ‘99’ as described in Section 3.6) which didn’t report any promotional information. It was decided to replace the missing promotional data with “unknown” enumerations.

Additionally, by only using panelist data from years 2008-2011, there were only 29 unique stores since the panelist data was from only two regions.

Finally, the data grocery store transactions made up 67891 of the 67965 total transactions:

- GK: 67891
- DK: 26
- MK: 25
- KK: 23

As a result, DK, MK, and KK store transactions were omitted from modeling.

The final data frame used in modeling contained approx. 68k rows and 30 features containing the item information and store information that specifically outlined promotions, price reductions and display.

Aggregation into Weeks

After final merge of data it was theorized that the data was organized in a manner best suited for doing modeling of daily predictions and not estimating the effect of promotional offers.
Using Excel, the data was aggregated into weekly bins. Any data that was non-numerical was counted instead of summed. This resulted in a single row for each week of the years 2008 to 2011. This new data set was used for modelling and assessing promotions.

Modeling and Evaluation

Predicting Daily Sales of Beer

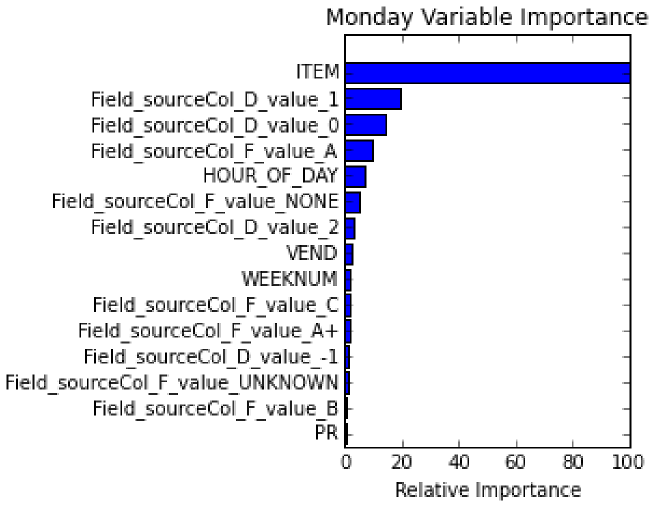
Feature Selection

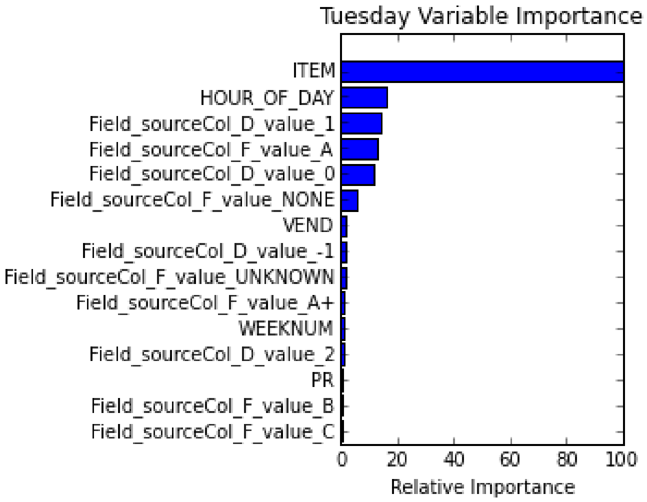
Prior to modeling, the following features were dropped from the data frame:

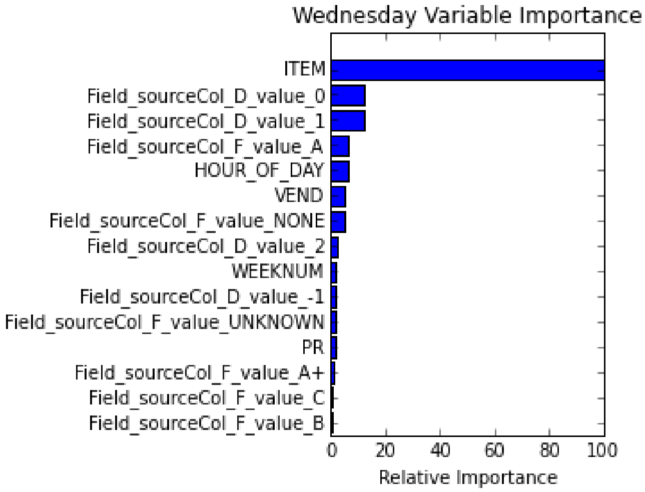
Reason	Feature
Unique Data Identifier	PANID WEEK IRI_KEY START_TIME_EPOCH_S END_TIME_EPOCH_S TRANSACTION_TIME_EPOCH_S COLUPC
Unnecessary (Using Day of Week)	DAY MINUTE MONTH YEAR
Negligible Variance	SY GE
Highly Correlated	UNITS

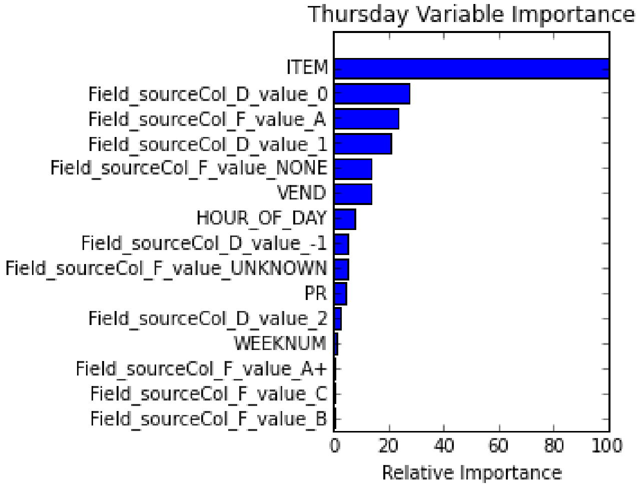
The resulting data frame was then separated by day of week and feature importance for each day were identified using the *sklearn.feature_selection.SelectKBest* and *sklearn.feature_selection.f_regression* class/functions. These functions were selected after comparing results from other regressions *GradientBoosting* and *RandomForest*.

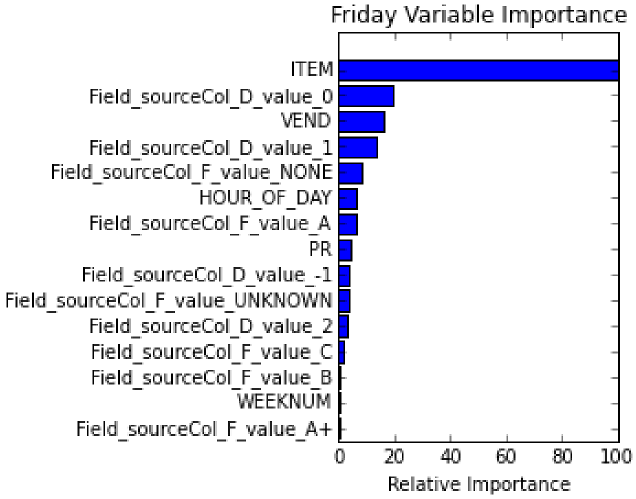
The following graphs show the relative importances of the various features by each day of the week:

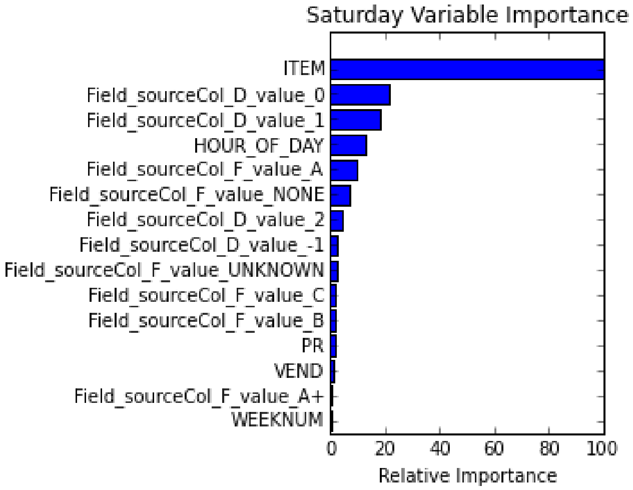

Monday Variable Importance
ITEM
Field_sourceCol_D_value_1
Field_sourceCol_D_value_0
Field_sourceCol_F_value_A
HOUR_OF_DAY
Field_sourceCol_F_value_NONE
Field_sourceCol_D_value_2
VEND
WEEKNUM
Field_sourceCol_F_value_C
Field_sourceCol_F_value_A+
Field_sourceCol_D_value_-1
Field_sourceCol_F_value_UNKNOWN
Field_sourceCol_F_value_B
PR
0
20
40
60
80
100
Relative Importance

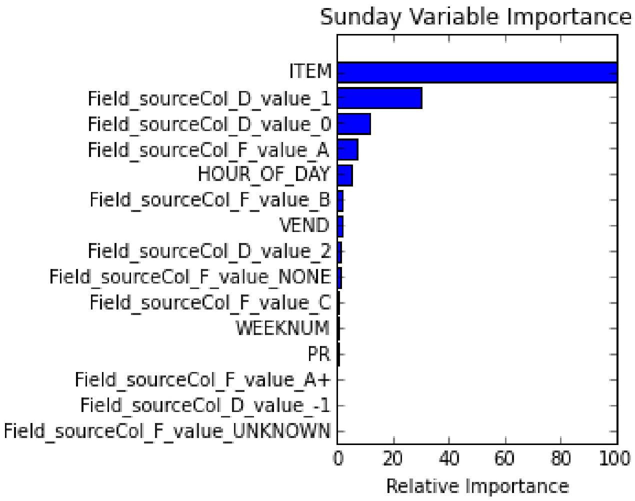

Tuesday Variable Importance
ITEM
HOUR_OF_DAY
Field_sourceCol_D_value_1
Field_sourceCol_F_value_A
Field_sourceCol_D_value_0
Field_sourceCol_F_value_NONE
VEND
Field_sourceCol_D_value_-1
Field_sourceCol_F_value_UNKNOWN
Field_sourceCol_F_value_A+
WEEKNUM
Field_sourceCol_D_value_2
PR
Field_sourceCol_F_value_B
Field_sourceCol_F_value_C
0
20
40
60
80
100
Relative Importance


Wednesday Variable Importance
ITEM
Field_sourceCol_D_value_0
Field_sourceCol_D_value_1
Field_sourceCol_F_value_A
HOUR_OF_DAY
VEND
Field_sourceCol_F_value_NONE
Field_sourceCol_D_value_2
WEEKNUM
Field_sourceCol_D_value_-1
Field_sourceCol_F_value_UNKNOWN
PR
Field_sourceCol_F_value_A+
Field_sourceCol_F_value_C
Field_sourceCol_F_value_B
0
20
40
60
80
100
Relative Importance


Thursday Variable Importance
ITEM
Field_sourceCol_D_value_0
Field_sourceCol_F_value_A
Field_sourceCol_D_value_1
Field_sourceCol_F_value_NONE
VEND
HOUR_OF_DAY
Field_sourceCol_D_value_-1
Field_sourceCol_F_value_UNKNOWN
PR
Field_sourceCol_D_value_2
WEEKNUM
Field_sourceCol_F_value_A+
Field_sourceCol_F_value_C
Field_sourceCol_F_value_B
0
20
40
60
80
100
Relative Importance


Friday Variable Importance
ITEM
Field_sourceCol_D_value_0
VEND
Field_sourceCol_D_value_1
Field_sourceCol_F_value_NONE
HOUR_OF_DAY
Field_sourceCol_F_value_A
PR
Field_sourceCol_D_value_-1
Field_sourceCol_F_value_UNKNOWN
Field_sourceCol_D_value_2
Field_sourceCol_F_value_C
Field_sourceCol_F_value_B
WEEKNUM
Field_sourceCol_F_value_A+
0
20
40
60
80
100
Relative Importance


Saturday Variable Importance
ITEM
Field_sourceCol_D_value_0
Field_sourceCol_D_value_1
HOUR_OF_DAY
Field_sourceCol_F_value_A
Field_sourceCol_F_value_NONE
Field_sourceCol_D_value_2
Field_sourceCol_D_value_-1
Field_sourceCol_F_value_UNKNOWN
Field_sourceCol_F_value_C
Field_sourceCol_F_value_B
PR
VEND
Field_sourceCol_F_value_A+
WEEKNUM
0
20
40
60
80
100
Relative Importance



Collecting the union of the top 5 features for each day, a new set of seven features were selected to be used for generating the final models:

```
Field_sourceCol_F_value_NONE
ITEM
VEND
Field_sourceCol_F_value_A
HOUR_OF_DAY
Field_sourceCol_D_value_1
Field_sourceCol_D_value_0
```

Using the seven features computed previously with the transactions from Saturdays only as a starting point, multiple regressor models were built to identify likely candidates to create a stacking ensemble. The following table outlines the R-squared score for the various regressors used:

Regressor	R^2
Linear	0.059
Ridge	0.0596
Decision Tree	0.448
Random Forest	0.424
Gradient Boosting	0.0283
Ada Boost	0.194
SVR (normalized data)	0.150
NuSVR (normalized data)	0.150

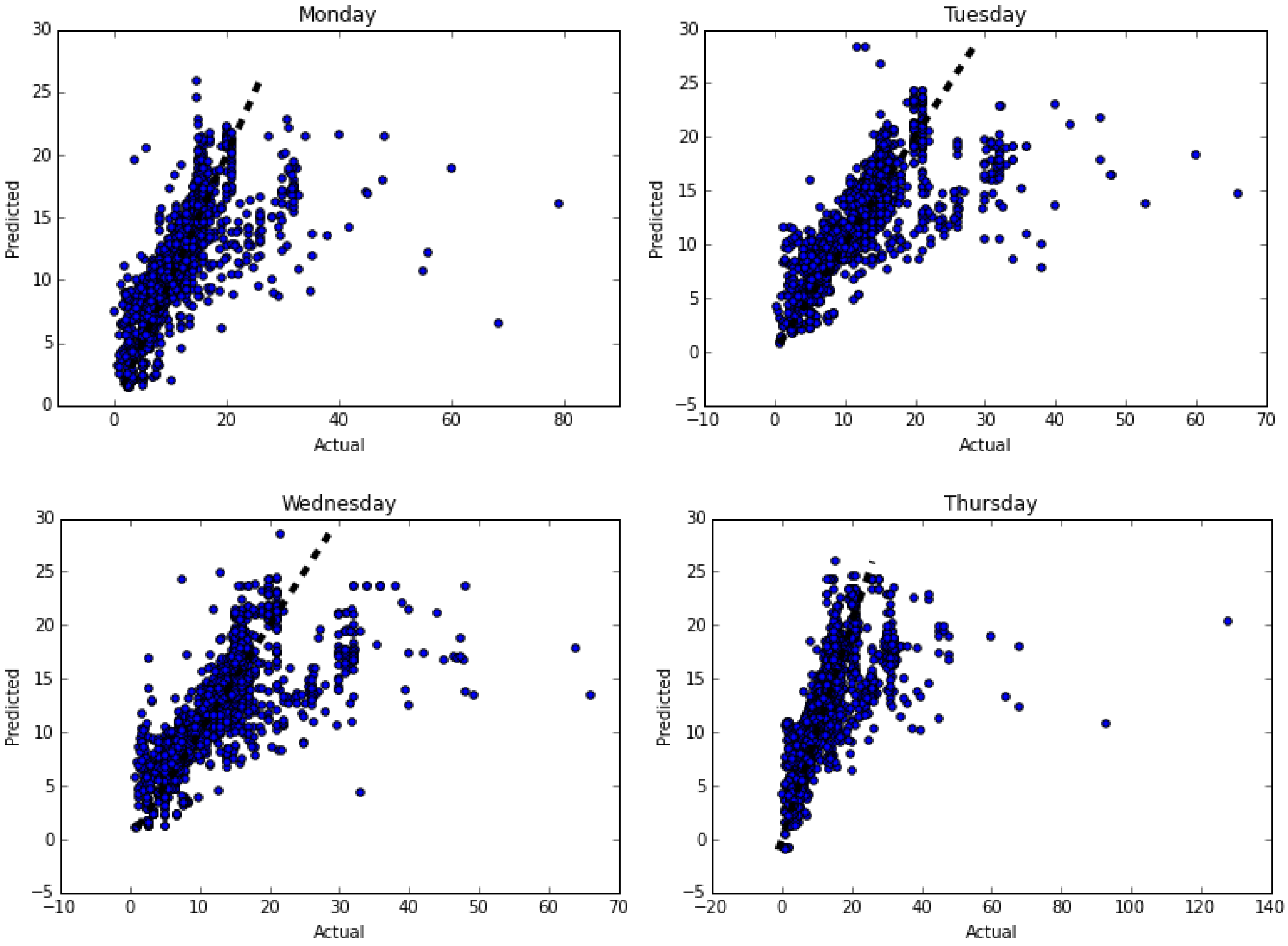
The following six were selected for ensembling because they are all weak learners:

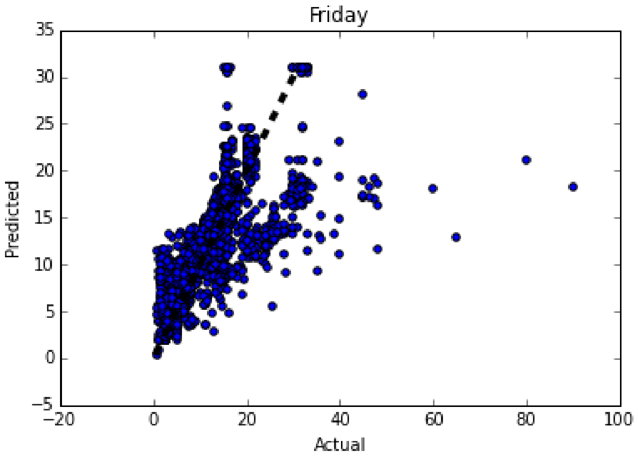
1. Random Forest
2. Decision Tree
3. Gradient Boosting
4. Ada Boost
5. NuSVR
6. LinearRegression used for the L1 combiner.

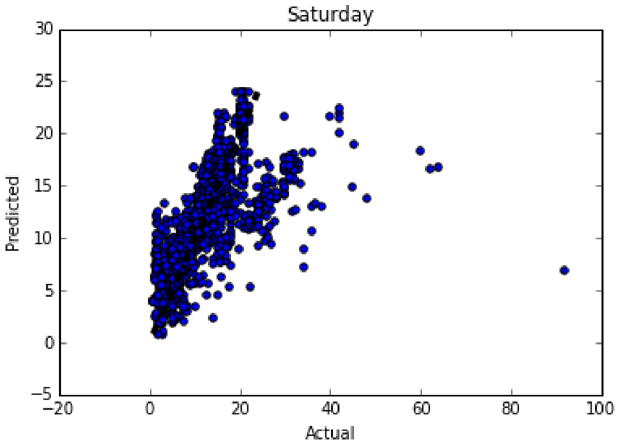
The following scores were obtained for each day when executing the Stacking Ensemble. Each of the stacking ensemble was tested using 10-fold validation.

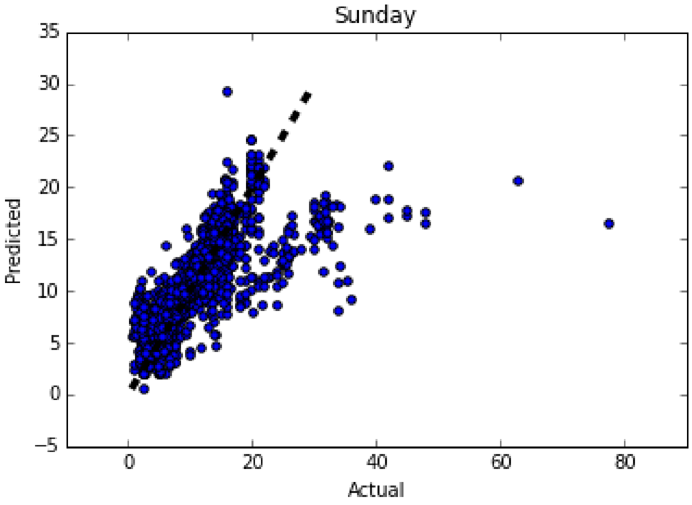
Day of Week	R^2
Monday	0.492
Tuesday	0.547
Wednesday	0.521
Thursday	0.480
Friday	0.552
Saturday	0.539
Sunday	0.578

The graphs depicting Actual vs Predicted for each day are displayed below:







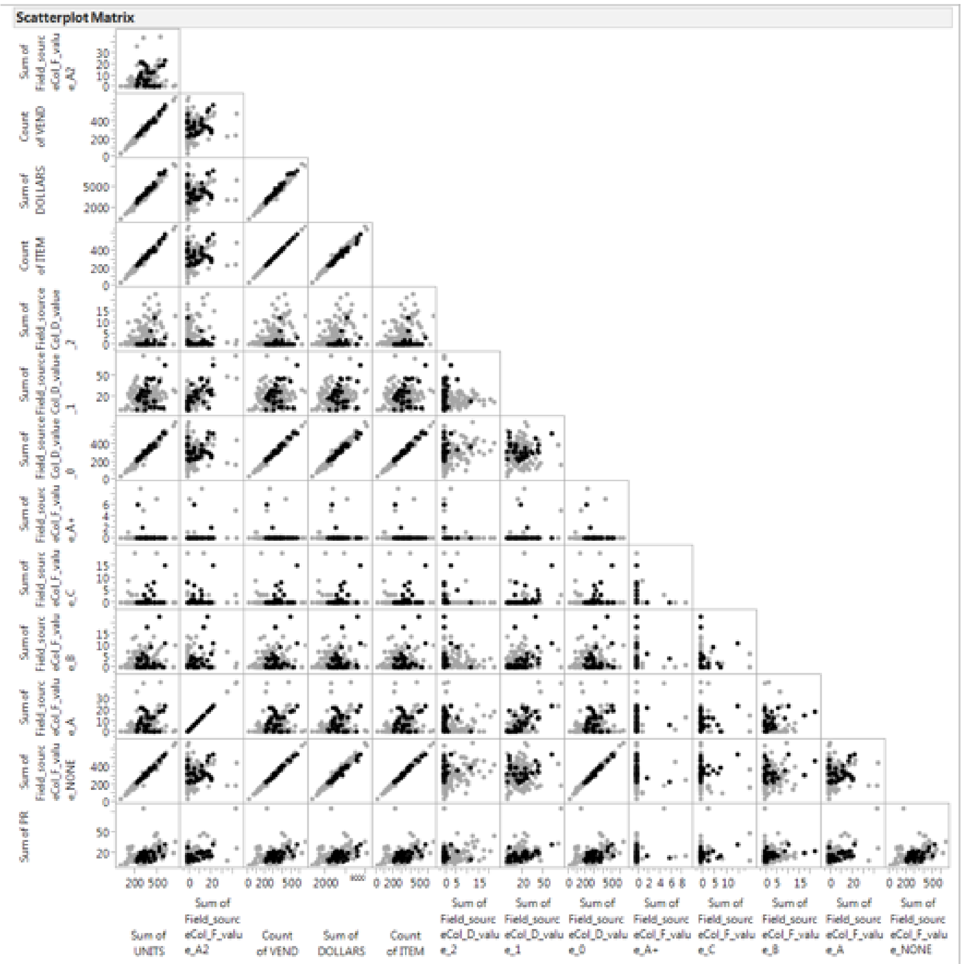


Determining Weekly Promotion Periods

In this portion of the analysis, it was assumed that “unknown” for promotions meant no promotion.

Feature relationships

JMP by SAS was used to identify features that are highly correlated. The following scatter matrix was generated:



Looking at the matrix it seemed that the following features had high correlations:

- Count of VEND with Sum of UNITS
- Sum of Dollars with Sum of UNITS
- Sum of Dollars with Count of VEND
- Count of ITEM with Sum of UNITS
- Count of ITEM with Count of SUM
- Count of ITEM with Sum of DOLLARS
- Sum of No Display with Sum of Units
- Sum of No Display with Count of VEND
- Sum of No Display with Sum of DOLLARS
- Sum of No Display with Count of ITEM

Feature Selection

Using Weka, two types of attribute selection routines were run. The purpose of this was not to identify and eliminate unimportant features, but to couple the information with the correlations as a guide to determine which features could be eliminated. For Best First + CFSsubset Eval and then WrapperSubsetEval using M5Rules and GreedyStepwise search, the following features tended to agree in importance reported by both algorithms:

- *Sum of UNITS* - important
- *Count of VEND* - important
- *Sum of Field_sourceCol_D_value_2* - not important

Predicting with all Features

As the values were all numerical, it was decided not to do any variable transformations. In Weka, certain models can handle all numeric values, others cannot (need nominal). Below are the models that were used, and their results--which were all quite good.

Model	R^2
Linear Regression	0.9887
M5P Rules	0.988
SMO Regression	0.9884
REP Tree	0.976
Rotation Forest with M5P	0.9882

Predicting with Subset of Features

After the previous predictions, knowledge of the correlations and the feature importances was used to drop certain features. The features that were dropped were Count of VENDORS and the promotions with did not have any displays (F=0 and D=0). The top four performing models from the previous phase were run again and generated similar results.

Model	R^2
Linear Regression	0.9882
M5P Rules	0.9882
SMO Regression	0.9885
Rotation Forest with M5P	0.9884

The linear regression and M5P rules produced almost the same model. The Linear Regression equation for feature coefficients is shown below:

```
Sum of dollars = 1.6956*Week number + 4.5507*Sum of UNITS +  4.051*Sum of
Field_sourceCol_F_value_A2 +  7.5224*Count of ITEM +  9.8466*Sum of
Field_sourceCol_F_value_B + 4.051*Sum of Field_sourceCol_F_value_A + -3.6757*Sum of PR
-252.876
```

Where Field_sourceCol_F_value_A means a *Promotion using a Large Size AD* and Field_sourceCol_F_value_A2 means *Promotion using a Retailer Coupon or Rebate* and PR means a *Price Reduction*.

Looking at the feature coefficients for the other models listed above there were variations in the coefficients of features but what was common across all models was that Promotions using ads, coupons and rebates always had positive coefficients indicating that they increased sales revenue. Compared to Price reduction which always had negative coefficients indicating that those promotions decreased sales revenue.

Predicting Future Sales per Week

The same preprocessed data from the Promotion Modeling was used to predict future sales. Data was divided as follows:

1. Data from years 2008 to 2010 was used a training data
2. Data from the year 2011 was used a testing data

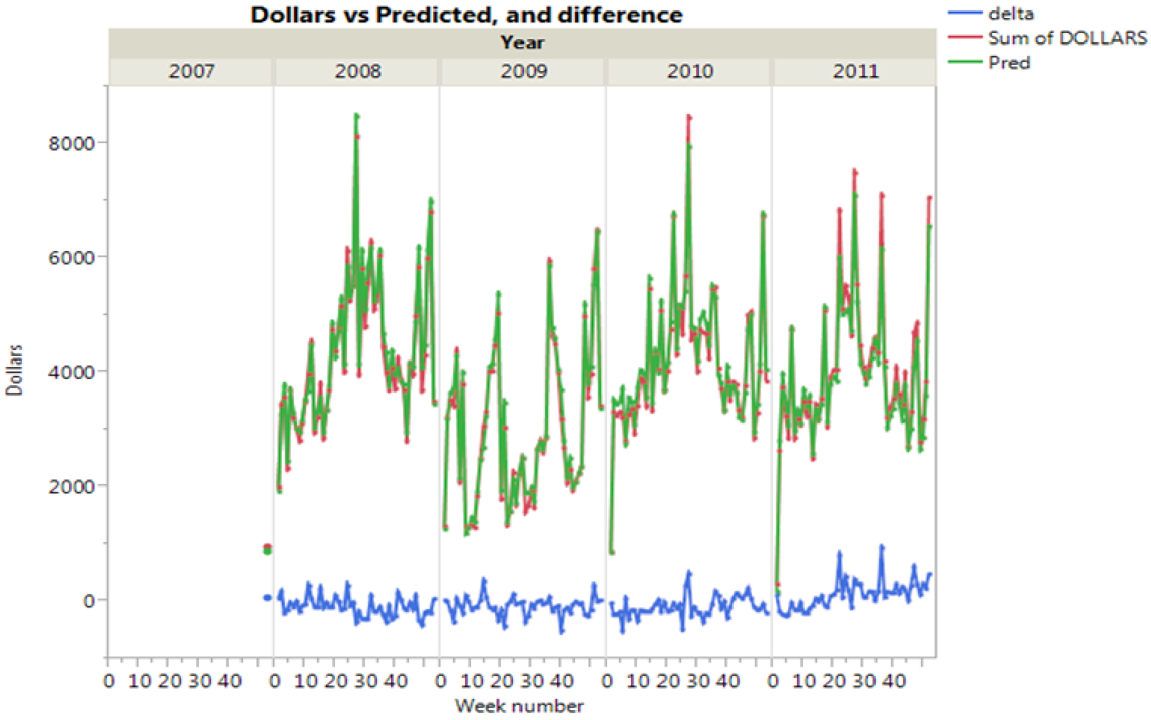
Model	R^2

Linear Regression	0.9924
M5P Rules	0.9924
REP Tree	0.976
SMO Regression	0.9921

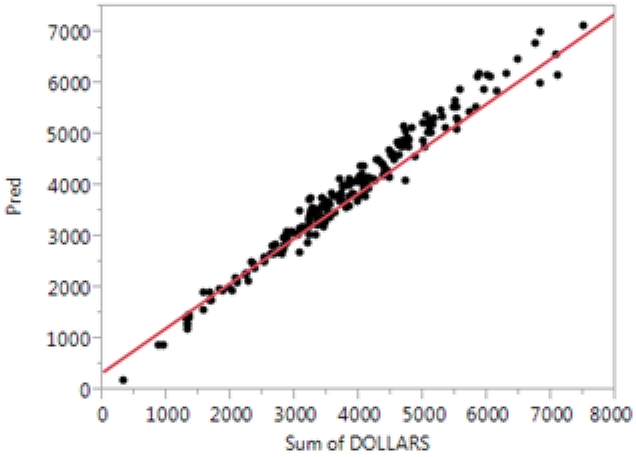
Comparing the results of modeling using three years versus four resulted in better results due to the fact that the three year dataset had less variation and therefore was easier to model.
Based on the results, the simplicity of the linear model was chosen to generate predictions for the fourth year. The resulting linear model used five features:

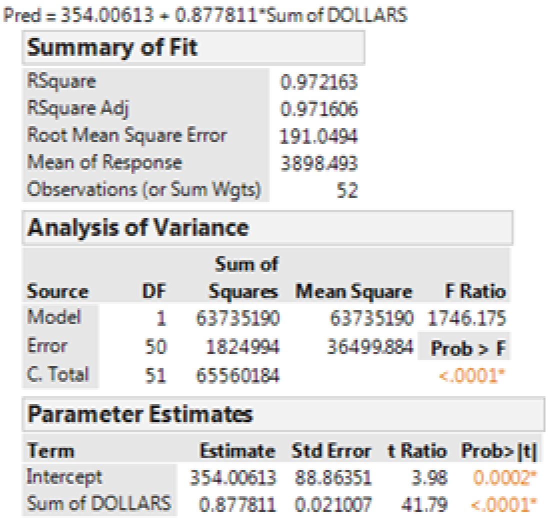
```
sum of dollars = 5.9039*Sum of UNITS + 5.8388*Count of ITEM + 5.8252*Sum of
Field_sourceCol_F_value_A + -3.0693*Sum of PR + -180.9078
```

The following plot shows the actual and predicted data using a model generated using years 2008-2010 data and predicting sales for years 2008-2011. The delta line at the bottom is the difference between the actual and predicted values.



By observation and intuition one can expect the accuracy of the model to be less in 2011 vs previous years. This is true, however, it is quite accurate even for 2011-- the R^2 is 0.97. The following figures show the detailed summary of the model accuracy:





Pred = 354.00613 + 0.877811*Sum of DOLLARS

Summary of Fit

RSquare	0.972163
RSquare Adj	0.971606
Root Mean Square Error	191.0494
Mean of Response	3898.493
Observations (or Sum Wgts)	52

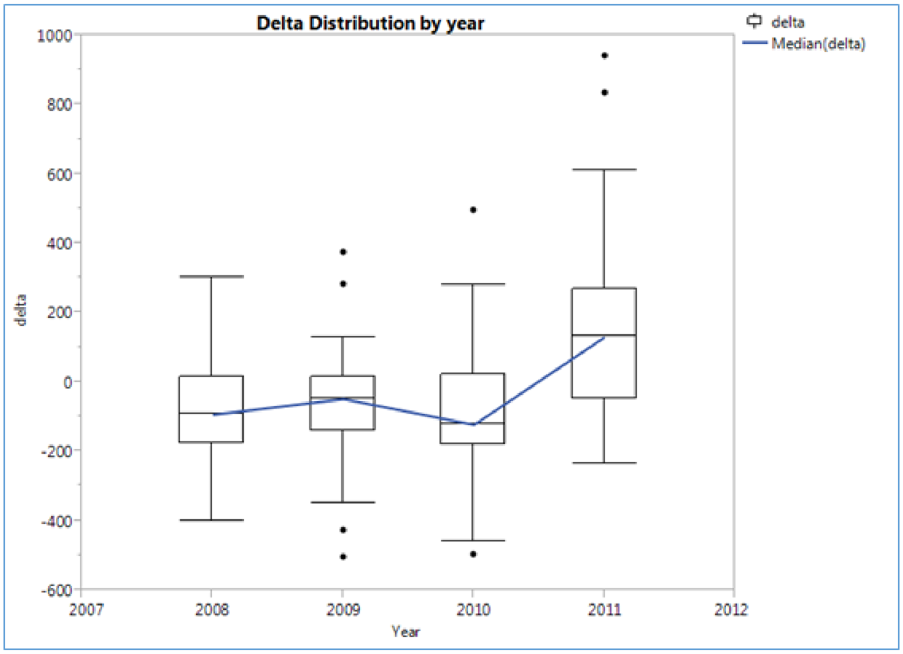
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	63735190	63735190	1746.175
Error	50	1824994	36499.884	Prob > F
C. Total	51	65560184		<.0001*

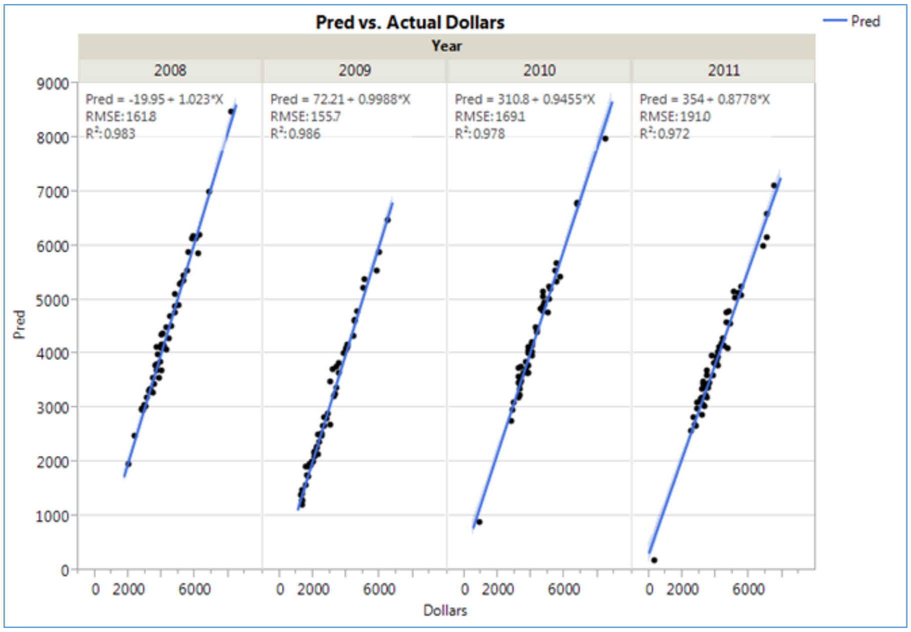
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	354.00613	88.86351	3.98	0.0002*
Sum of DOLLARS	0.877811	0.021007	41.79	<.0001*

The following chart shows the difference in performance for 2011. It shows a higher distribution of error compared to the other years.



The chart below shows the R^2 for each year of the model. One thing to notice is the RMSE is larger for 2011. It is not a majorly significant difference, but definitely noticeable.



Conclusion

Daily Predictions

Analysis of the data showed that the optimal indicators for predicting transaction price of beer are:

- Does the store have a promotion/display?
- Specific beer product offered at a store
- Vendors of beer a particular store offers
- Does the store have an advertisement out?
- What hour of the day does the store typically sell most beer

Weekly Sales

Analysis of the data showed that the optimal indicators for predicting weekly sales of beer are:

- Sum of units sold in a week
- Count of items sold in a week
- Use of Large Ad
- Avoid use of price rebates

Weekly Promotion Effects

Analysis of the data showed that the optimal indicators for predicting effect of weekly promotions of beer are:

- Advertisement/coupons/rebates increase sales
- Price reduction decreases sales

Note that the analysis of daily and weekly sales were performed using different feature sets. In particular, the daily predictions removed UNITS as a feature whereas weekly sale predictions did use UNITS as a feature.

Appendix

Weka Weekly model results

All features included

Linear Regression

Correlation coefficient	0.9887
Mean absolute error	147.949
Root mean squared error	203.0618
Relative absolute error	14.4225 %
Root relative squared error	14.9373 %
Total Number of Instances	212

Scheme:weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8

Relation: master_merged_preprocessed_bymonthnyear-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,6-weka.filters.unsupervised.attribute.Remove-R1

Sum of DOLLARS =

1.7721 * Week number +
4.4501 * Sum of UNITS +
7.5104 * Count of VEND +
7.5104 * Count of ITEM +
1.5229 * Sum of Field_sourceCol_D_value_0 +
-8.881 * Sum of Field_sourceCol_F_value_NONE +
-3.4921 * Sum of PR +
-240.8003

M5P rules

Just 1 Leaf !!, no branching. Vs Linear Regression, no item count,

Correlation coefficient	0.988
Mean absolute error	149.683
Root mean squared error	209.5124
Relative absolute error	14.5915 %
Root relative squared error	15.4118 %
Total Number of Instances	212

Scheme:weka.classifiers.trees.M5P -M 4.0

Relation: master_merged_preprocessed_bymonthnyear-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,6-weka.filters.unsupervised.attribute.Remove-R1

Sum of DOLLARS =
1.7043 * Week number
+ 4.4092 * Sum of UNITS
+ 8.1073 * Count of VEND
+ 6.1394 * Sum of Field_sourceCol_D_value_2
+ 6.9573 * Sum of Field_sourceCol_D_value_1
+ 8.4227 * Sum of Field_sourceCol_D_value_0
- 8.8155 * Sum of Field_sourceCol_F_value_NONE
- 3.3919 * Sum of PR
- 240.8668

SMOreg

Fast execution time

Correlation coefficient	0.9884
Mean absolute error	146.1806
Root mean squared error	209.3487
Relative absolute error	14.2501 %
Root relative squared error	15.3998 %
Total Number of Instances	212

Scheme:weka.classifiers.functions.SMOreg -C 1.0 -N 0 -I "weka.classifiers.functions.supportVector.RegSMOImproved -L 0.001 -W 1 -P 1.0E-12 -T 0.001 -V" -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"

Relation: master_merged_preprocessed_bymonthnyear-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,6-weka.filters.unsupervised.attribute.Remove-R1

weights (not support vectors):

```
+    0.0104 * (normalized) Week number
+    0.2036 * (normalized) Sum of UNITS
+    0.0348 * (normalized) Sum of Field_sourceCol_F_value_A2
+    0.1919 * (normalized) Count of VEND
+    0.1919 * (normalized) Count of ITEM
-    0.002  * (normalized) Sum of Field_sourceCol_D_value_2
+    0.018  * (normalized) Sum of Field_sourceCol_D_value_1
+    0.2014 * (normalized) Sum of Field_sourceCol_D_value_0
+    0.0099 * (normalized) Sum of Field_sourceCol_F_value_A+
+    0.0335 * (normalized) Sum of Field_sourceCol_F_value_C
+    0.0205 * (normalized) Sum of Field_sourceCol_F_value_B
+    0.0348 * (normalized) Sum of Field_sourceCol_F_value_A
+    0.1908 * (normalized) Sum of Field_sourceCol_F_value_NONE
-    0.037  * (normalized) Sum of PR
-    0.0125
```

Rotation Forest with M5P

Used PCA—again with M5p 1 model/leaf

See coefficients for variable importance.

Correlation coefficient	0.9882
Mean absolute error	150.2555
Root mean squared error	207.785
Relative absolute error	14.6473 %
Root relative squared error	15.2847 %
Total Number of Instances	212

Scheme:weka.classifiers.meta.RotationForest -G 3 -H 3 -P 50 -F "weka.filters.unsupervised.attribute.PrincipalComponents -R 1.0 -A 5 -M -1" -S 1 -I 10 -W weka.classifiers.trees.M5P -- -M 4.0

Relation: master_merged_preprocessed_bymonthnyear-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,6-weka.filters.unsupervised.attribute.Remove-R1

- Example LM

Sum of DOLLARS =

-2177.673 * 0.692Count of ITEM_0+0.692Count of VEND_1-0.204Sum of Field_sourceCol_F_value_C_2_0

+ 380.0482 * -0.979Sum of Field_sourceCol_F_value_C_2-0.144Count of ITEM_0-0.144Count of VEND_1_0

+ 2836.5433 * 0.68 Sum of Field_sourceCol_D_value_0_2+0.677Sum of Field_sourceCol_F_value_NONE_0+0.281Sum of Field_sourceCol_F_value_A2_1_1

- 861.4646 * 0.959Sum of Field_sourceCol_F_value_A2_1-0.211Sum of Field_sourceCol_F_value_NONE_0-0.187Sum of Field_sourceCol_D_value_0_2_1

- 1330.5247 * 0.709Sum of Field_sourceCol_D_value_0_2-0.705Sum of Field_sourceCol_F_value_NONE_0-0.017Sum of Field_sourceCol_F_value_A2_1_1

+ 474.8756 * 0.763Sum of Field_sourceCol_F_value_A_2+0.503Sum of UNITS_1+0.405Sum of Field_sourceCol_F_value_B_0_2

- 235.0855 * 0.747Sum of Field_sourceCol_F_value_B_0-0.663Sum of UNITS_1+0.04 Sum of Field_sourceCol_F_value_A_2_2

- 134.5265 * 0.645Sum of Field_sourceCol_F_value_A_2-0.554Sum of UNITS_1-0.527Sum of Field_sourceCol_F_value_B_0_2

+ 53.4333 * 0.643Sum of Field_sourceCol_D_value_2_1-0.624Sum of PR_2+0.444Week number_0_3

+ 88.4645 * 0.665Sum of Field_sourceCol_D_value_1_0+0.665Sum of Field_sourceCol_D_value_1_2+0.342Sum of Field_sourceCol_F_value_A+_1_4

+ 4114.794

Subset of Features

Linear Regression

Correlation coefficient	0.9882
Mean absolute error	150.416
Root mean squared error	207.3127
Relative absolute error	14.663 %
Root relative squared error	15.25 %
Total Number of Instances	212

Scheme:weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8

Relation: master_merged_preprocessed_bymonthnyear-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,6-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R4,8,13

Sum of DOLLARS=

1.6956 * Week number +

4.5507 * Sum of UNITS +

4.051 * Sum of Field_sourceCol_F_value_A2 +

7.5224 * Count of ITEM +

9.8466 * Sum of Field_sourceCol_F_value_B +

4.051 * Sum of Field_sourceCol_F_value_A +

-3.6757 * Sum of PR +

-252.876

M5P rules

Just 1 Leaf !!, no branching. Vs Linear Regression, no item count,

Correlation coefficient	0.9882
Mean absolute error	150.416
Root mean squared error	207.3127
Relative absolute error	14.663 %
Root relative squared error	15.25 %
Total Number of Instances	212

Scheme:weka.classifiers.rules.M5Rules -M 4.0

Relation: master_merged_preprocessed_bymonthnyear-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,6-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R4,8,13

Sum of DOLLARS =

1.6956 * Week number

+ 4.5507 * Sum of UNITS

+ 8.102 * Sum of Field_sourceCol_F_value_A2

+ 7.5224 * Count of ITEM

+ 9.8466 * Sum of Field_sourceCol_F_value_B

- 3.6757 * Sum of PR

- 252.876 [212/14.016%]

SMO Reg

Fast execution time

Correlation coefficient	0.9885
Mean absolute error	146.6427
Root mean squared error	208.6124
Relative absolute error	14.2952 %
Root relative squared error	15.3456 %
Total Number of Instances	212

Scheme:weka.classifiers.functions.SMOreg -C 1.0 -N 0 -I "weka.classifiers.functions.supportVector.RegSMOImproved -L 0.001 -W 1 -P 1.0E-12 -T 0.001 -V" -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"

Relation: master_merged_preprocessed_bymonthnyear-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,6-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R4,8,13

SMOreg

weights (not support vectors):

```
+    0.0111 * (normalized) Week number
+    0.325  * (normalized) Sum of UNITS
+    0.0297 * (normalized) Sum of Field_sourceCol_F_value_A2
+    0.6673 * (normalized) Count of ITEM
-    0.0154 * (normalized) Sum of Field_sourceCol_D_value_2
-    0.0178 * (normalized) Sum of Field_sourceCol_D_value_1
+    0.0091 * (normalized) Sum of Field_sourceCol_F_value_A+
+    0.0129 * (normalized) Sum of Field_sourceCol_F_value_C
+    0.0149 * (normalized) Sum of Field_sourceCol_F_value_B
+    0.0297 * (normalized) Sum of Field_sourceCol_F_value_A
-    0.0171 * (normalized) Sum of PR
-    0.0129
```

Rotation Forest with M5P

Used PCA—again with M5p 1 model/leaf

See coefficients for variable importance.

Correlation coefficient	0.9884
Mean absolute error	150.1257
Root mean squared error	206.1822
Relative absolute error	14.6347 %
Root relative squared error	15.1668 %
Total Number of Instances	212

Scheme:weka.classifiers.meta.RotationForest -G 3 -H 3 -P 50 -F "weka.filters.unsupervised.attribute.PrincipalComponents -R 1.0 -A 5 -M -1" -S 1 -I 10 -W weka.classifiers.trees.M5P -- -M 4.0

Relation: master_merged_preprocessed_bymonthnyear-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,6-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R4,8,13

- Example LM

```
Sum of DOLLARS =
        22.3335 * 0.687Sum of Field_sourceCol_F_value_C_0-0.682Sum of Field_sourceCol_D_value_1_2-0.252Sum of Field_sourceCol_D_value_2_1_0
        + 755.2766 * 0.721Sum of Field_sourceCol_F_value_A_2+0.681Sum of UNITS_0+0.131Sum of Field_sourceCol_F_value_A+_1_1
        + 93.2914 * -0.927Sum of Field_sourceCol_F_value_A+_1+0.343Sum of UNITS_0-0.155Sum of Field_sourceCol_F_value_A_2_1
        + 20.8848 * 0.684Sum of Field_sourceCol_F_value_B_0+0.572Week number_1+0.453Sum of PR_2_2
        - 46.5919 * 0.806Sum of PR_2-0.59Week number_1-0.041Sum of Field_sourceCol_F_value_B_0_2
        - 23.0274 * -0.729Sum of Field_sourceCol_F_value_B_0+0.569Week number_1+0.38 Sum of PR_2_2
        - 233.7609 * 0.69 Sum of Field_sourceCol_F_value_A_2+0.69 Sum of Field_sourceCol_F_value_A2_1+0.216Count of ITEM_0_3
        - 779.8045 * -0.976Count of ITEM_0+0.153Sum of Field_sourceCol_F_value_A_2+0.153Sum of Field_sourceCol_F_value_A2_1_3
        + 3748.694
```

Using 2008-2010 Data Only

Linear Regression

Correlation coefficient	0.9924
Mean absolute error	128.1557
Root mean squared error	169.7715
Relative absolute error	12.083 %
Root relative squared error	12.2304 %
Total Number of Instances	160

Scheme:weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8

Relation: master_merged_preprocessed_bymonthnyear3n2011removed-weka.filters.unsupervised.attribute.Remove-R1

Sum of DOLLARS =

5.9039 * Sum of UNITS +

5.8388 * Count of ITEM +

5.8252 * Sum of Field_sourceCol_F_value_A +

-3.0693 * Sum of PR +

-180.9078

M5P- Virtually same as Linear Regression (expected)

Correlation coefficient	0.9924
Mean absolute error	128.1557
Root mean squared error	169.7715
Relative absolute error	12.083 %
Root relative squared error	12.2304 %
Total Number of Instances	160

Scheme:weka.classifiers.trees.M5P -M 4.0

Relation: master_merged_preprocessed_bymonthnyear3n2011removed-weka.filters.unsupervised.attribute.Remove-R1

Sum of DOLLARS =

5.9039 * Sum of UNITS

+ 5.8252 * Sum of Field_sourceCol_F_value_A2

+ 5.8388 * Count of ITEM

- 3.0693 * Sum of PR

- 180.9078

REPTree (41 nodes)

Correlation coefficient	0.976
Mean absolute error	212.713
Root mean squared error	301.569
Relative absolute error	20.0553 %
Root relative squared error	21.7252 %
Total Number of Instances	160

Scheme:weka.classifiers.trees.M5P -M 4.0

Relation: master_merged_preprocessed_bymonthnyear3n2011removed-weka.filters.unsupervised.attribute.Remove-R1

Top 3 branching on:
Count of Item

Sum of Units

See splitting on FieldsourceFvalue_A2 and

Sum of PR in nodes.

SMO Reg

Correlation coefficient	0.9921
Mean absolute error	130.5956
Root mean squared error	176.0429
Relative absolute error	12.313 %
Root relative squared error	12.6822 %
Total Number of Instances	160

Scheme:weka.classifiers.functions.SMOreg -C 1.0 -N 0 -I "weka.classifiers.functions.supportVector.RegSMOImproved -L 0.001 -W 1 -P 1.0E-12 -T 0.001 -V" -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"

Relation: master_merged_preprocessed_bymonthnyear3n2011removed-weka.filters.unsupervised.attribute.Remove-R1

```
weights (not support vectors):
+      0.0011 * (normalized) Week number
+      0.45   * (normalized) Sum of UNITS
+      0.016  * (normalized) Sum of Field_sourceCol_F_value_A2
+      0.516  * (normalized) Count of ITEM
+      0.0133 * (normalized) Sum of Field_sourceCol_D_value_2
+      0.0067 * (normalized) Sum of Field_sourceCol_D_value_1
+      0.0022 * (normalized) Sum of Field_sourceCol_F_value_A+
+      0.0022 * (normalized) Sum of Field_sourceCol_F_value_C
-      0.0004 * (normalized) Sum of Field_sourceCol_F_value_B
+      0.016  * (normalized) Sum of Field_sourceCol_F_value_A
-      0.0276 * (normalized) Sum of PR
+      0.0007
```