ILLINOIS
Gies College of Business

Model Development and Testing with Holdout Data

Sridhar Seshadri

Overview

Explanatory vs. Predictive Analytics

- Using a hold-out sample.

Regression for Predicting values

- Model Testing and Performance

Variable selection

Modeling Binary Response variables

- Logistic Regression

Explanatory Models

Goal: Explain relationship between predictors (explanatory variables) and target (response)

- Familiar use of regression in data analysis
- Model Goal: Fit the data well and understand the contribution of explanatory variables to the model
- "Goodness-of-fit": R2, residual analysis, p-values, visual tests, etc.

Predictive Analytics

Goal: predict target values in other data where we have values of predictors, but not target values

- Model Goal: Optimize prediction accuracy
- Develop model on training data
- Assess performance on validation (hold-out) data

Toyota Corolla Example

Predict prices of used Toyota Corollas based on their features

Prices of 1,436 used Toyota Corollas, with their information

Data are in Toyota.csv – available for download from website below.

Data Sources: https://www.biz.uiowa.edu/faculty/jledolter/datamining/datatext.html

Data and Variables (First 5 Rows)

Variable	Description
Price	Offer price in euros
Age	Age in months as of August 2004
Kilometers	Accumulated kilometers on odometer
Fuel type	Fuel type (*Petrol*, *Diesel*, *CNG*)
Horse Power	Horsepower
Metallic	Metallic color? (Yes = 1, No = 0)
Automatic	Automatic (Yes = 1, No = 0)
CC	Cylinder volume in cubic centimeters
Doors	Number of doors
Weight	Weight in kilograms

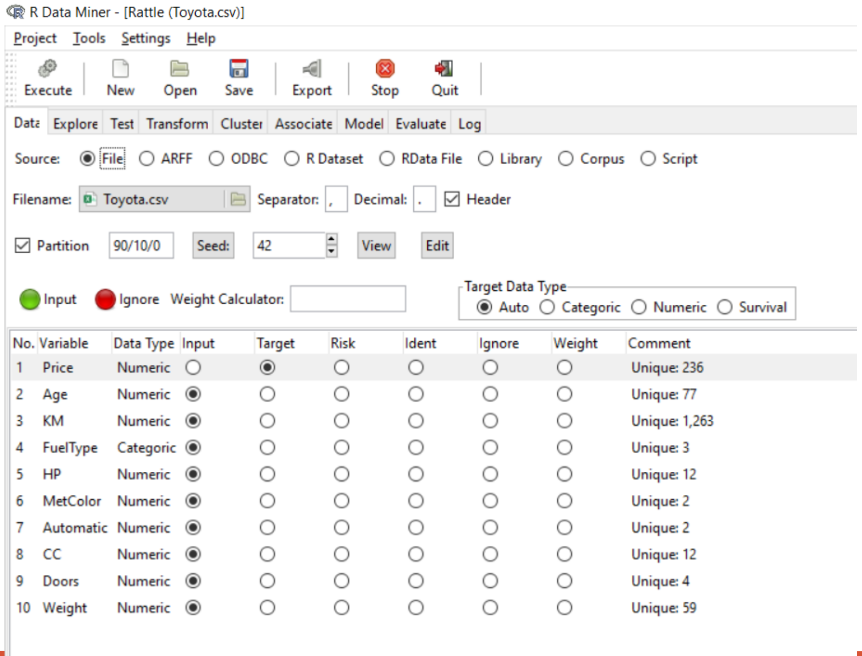
Price	Age	KM	FuelType	HP	MetColor	Automatic	CC	Doors	Weight
13500	23	46986	Diesel	90	1	0	2000	3	1165
13750	23	72937	Diesel	90	1	0	2000	3	1165
13950	24	41711	Diesel	90	1	0	2000	3	1165
14950	26	48000	Diesel	90	0	0	2000	3	1165
13750	30	38500	Diesel	90	0	0	2000	3	1170

Source: Rattle GUI / Togaware

Train, validate, test

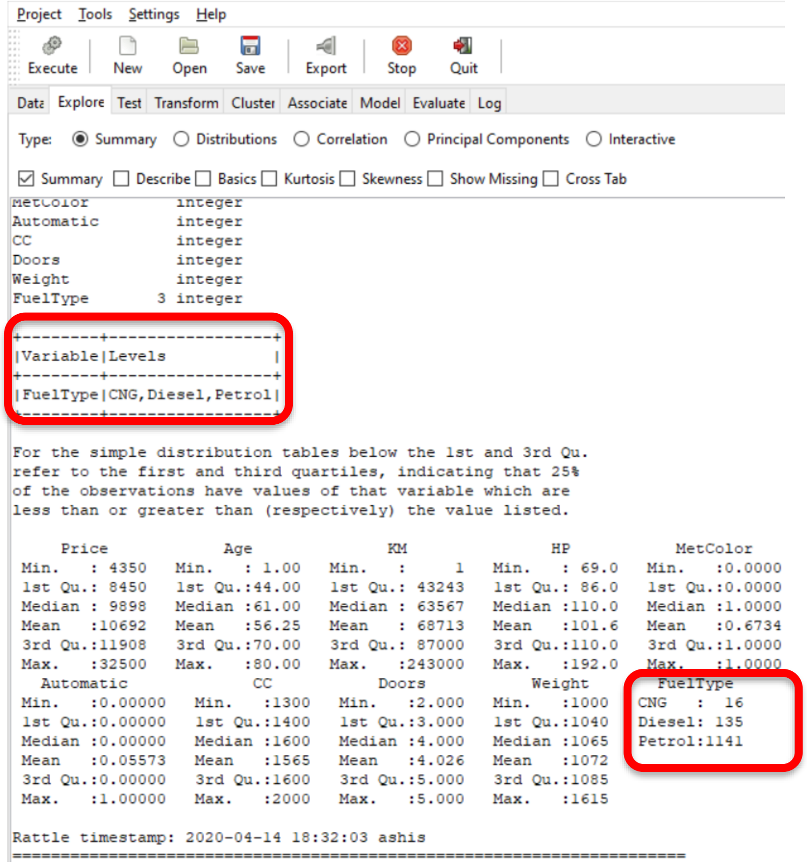
Out of 1,436 Observations, We randomly select 1,292 rows for Training and last 144 rows for Testing our predictions.

Please select 90/10/0 – this will create 90% train, 10 % validate and 0% on test



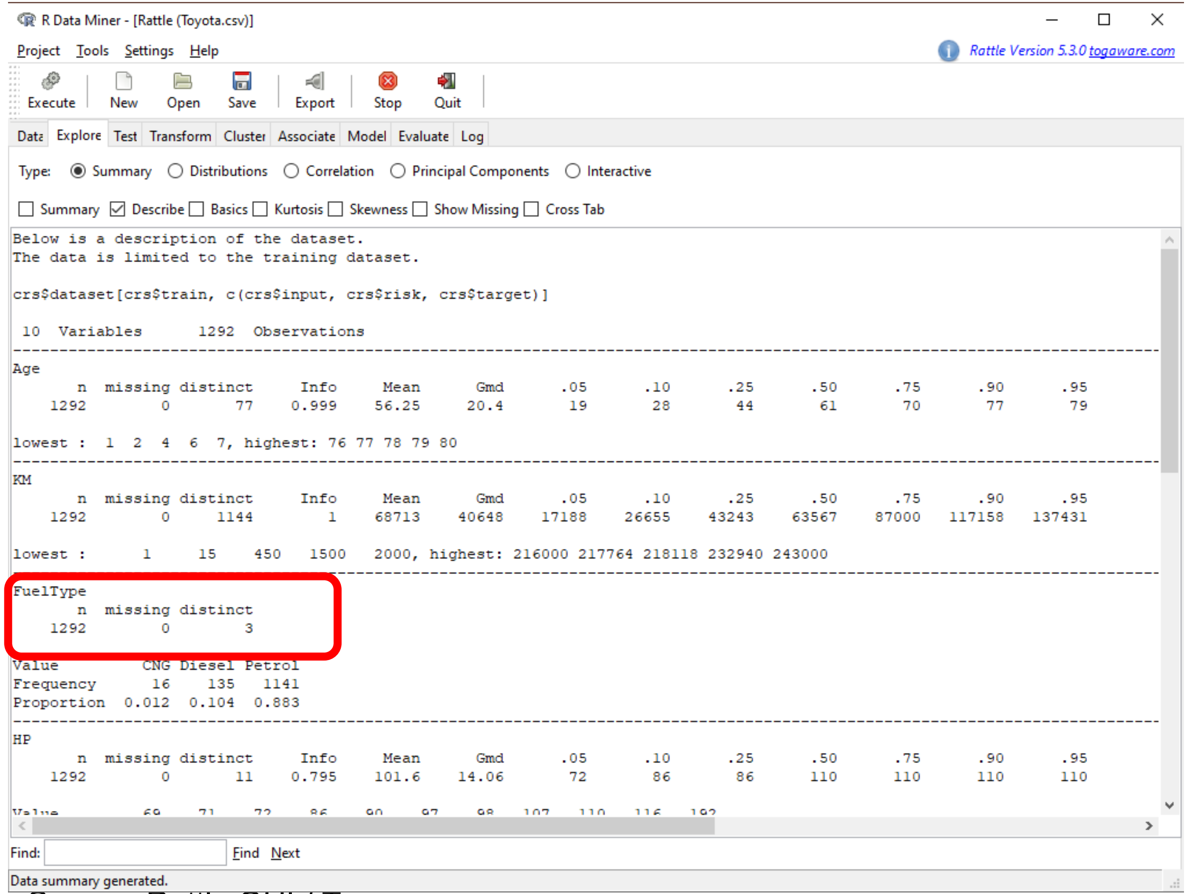
Source: Rattle GUI / Togaware

Categorical Predictors



Source: Rattle GUI / Togaware

Fuel Type has 3 categories



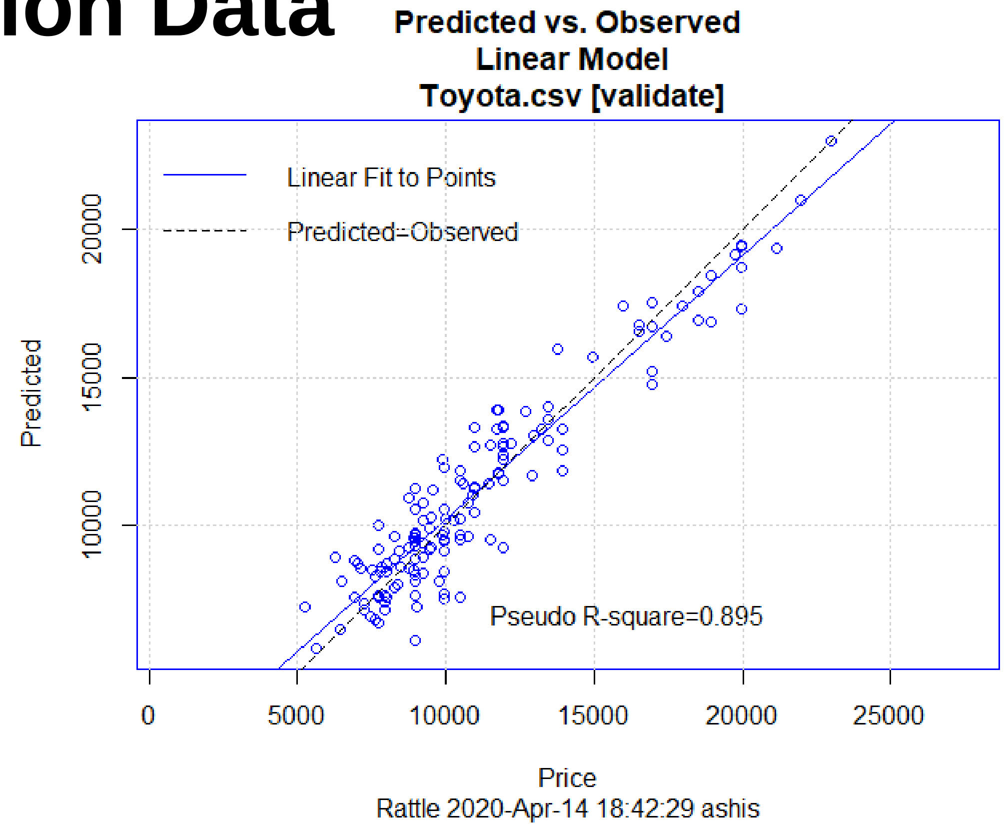
Source: Rattle GUI / Togaware

Estimate the regression model

```
Call:
lm(formula = Price ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)])

Residuals:
    Min      1Q  Median      3Q     Max
-9997.0  -741.4     3.5   750.5  6474.6

Coefficients:
                  Estimate   Std. Error t value Pr(>|t|)
(Intercept)    -2733.512690  1362.752613  -2.006  0.04508 *
Age             -123.176881     2.770001 -44.468  < 2e-16 ***
KM                -0.016251     0.001392 -11.677  < 2e-16 ***
FuelTypeDiesel  3539.895137   539.913528   6.556 7.97e-11 ***
FuelTypePetrol  1077.421607   346.283057   3.111  0.00190 **
HP                63.091432     5.915904  10.665  < 2e-16 ***
MetColor          34.338215    79.645115   0.431  0.66644
Automatic        434.004145   167.010650   2.599  0.00947 **
CC                -4.103270     0.571006  -7.186 1.13e-12 ***
Doors            -10.328114    42.474256  -0.243  0.80792
Weight            18.805412     1.254988  14.985  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1329 on 1281 degrees of freedom
Multiple R-squared:  0.8661,    Adjusted R-squared:  0.865
F-statistic: 828.2 on 10 and 1281 DF,  p-value: < 2.2e-16
```

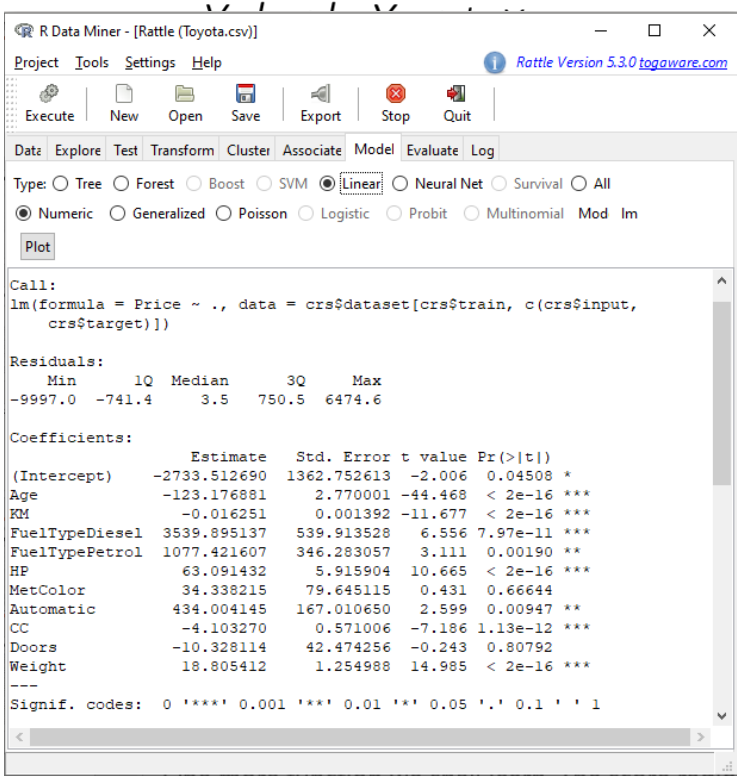
Source: Rattle GUI / Togaware

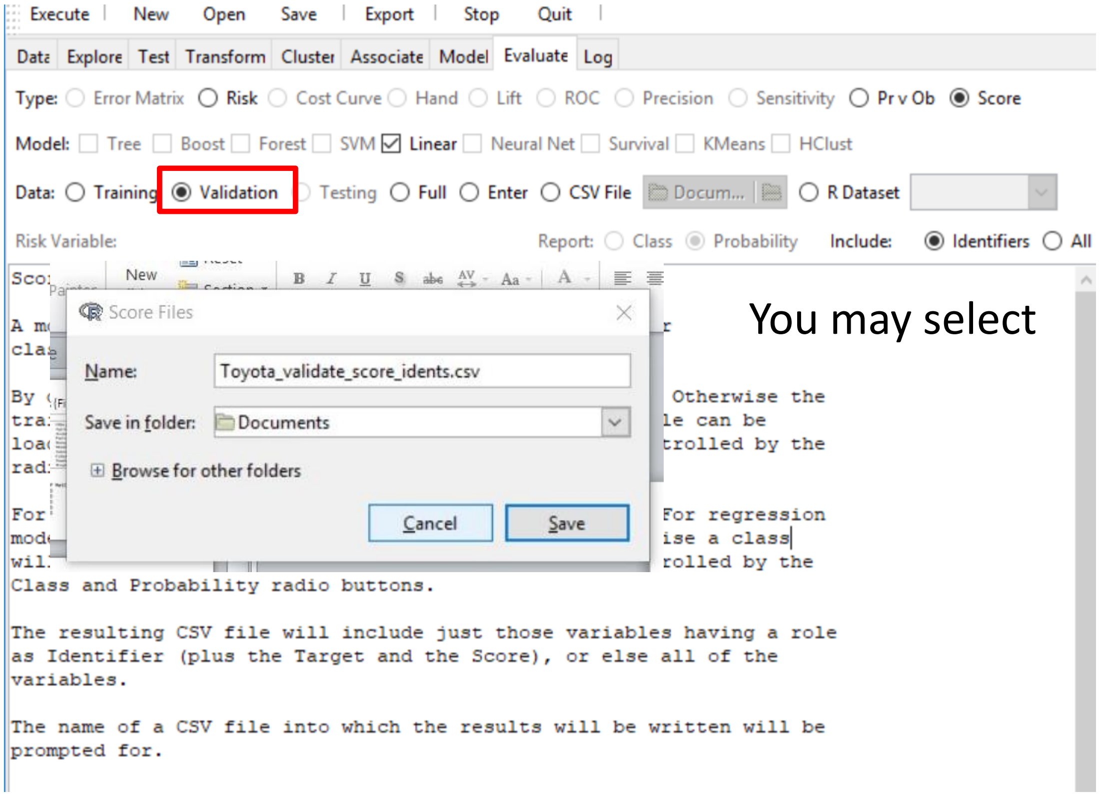
Predicted versus Observed on Validation Data



Source: Rattle GUI / Togaware

Predicted price and residuals: Use Score function under Evaluate





You may select

Source: Rattle GUI / Togaware

Predicted Prices on Test Set

J2 =SQRT(AVERAGE(H2:H145))

	A	B	E	F	G	H	J
1	Price	glm	Actual Price	Predicted	Residual	Residual Squared	
2	13750	15928.45	13750	15928.44526	-2178.445265	4745623.772	1204.27
3	16950	14760.86	16950	14760.8587	2189.141297	4792339.616	
4	14950	15674.63	14950	15674.6271	-724.6270991	525084.4327	
5	16950	16707.26	16950	16707.2631	242.7368982	58921.20175	
6	16950	17488.16	16950	17488.16082	-538.160817	289617.0649	
7	16950	15191.67	16950	15191.67441	1758.325587	3091708.87	
8	17450	16363.86	17450	16363.85846	1086.141541	1179703.447	
9	19950	19465.72	19950	19465.72249	484.2775114	234524.7081	
10	19950	19409.52	19950	19409.51625	540.4837546	292122.689	
11	18500	17885.11	18500	17885.1058	614.8942036	378094.8817	
12	21950	20982.05	21950	20982.04798	967.9520196	936931.1123	
13	19950	18690.79	19950	18690.78576	1259.214245	1585620.514	
14	18950	16852.35	18950	16852.34845	2097.651551	4400142.029	
15	16500	16543.58	16500	16543.57881	-43.57881369	1899.113002	
16	17950	17403.72	17950	17403.72109	546.2789142	298420.6521	
17	15950	17413.13	15950	17413.12982	-1463.129815	2140748.857	
18	16500	16749.08	16500	16749.08387	-249.0838735	62042.77606	
19	23000	22963.07	23000	22963.06631	36.93368726	1364.097255	
20	18500	16903.58	18500	16903.58481	1596.415192	2548541.466	
21	19950	17299.96	19950	17299.96187	2650.038129	7022702.086	
22	19750	19129.64	19750	19129.63786	620.3621398	384849.1844	
23	18950	18427.98	18950	18427.978	522.0219987	272506.9671	
24	21125	19325.89	21125	19325.89344	1799.106556	3236784.401	
25	11950	12257.39	11950	12257.38801	-307.3880078	94487.38731	
26	12950	13028.76	12950	13028.76448	-78.76447837	6203.843054	
27	11950	12413.43	11950	12413.43261	-463.4326104	214769.7844	

The saved csv has two columns: Price (the actual price), and glm(the predicted price.

Copied into Actual_Price and Predicted_Price

We create column (G) of residuals which is Predicted_price – Actual_Price

Source: Rattle GUI / Togaware

RMSE

Root mean
square error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}}$$

Number of observations

P is predicted
O is observed

Visual fit

Variable Selection

Variable selection pros and cons -- bias versus precision

Measures used commonly: Residual Mean Square error, Mallows Cp, Information Criteria (AIC and BIC), and adjusted R squared

- Parsimony preferred

Collinearity and variable selection -- caution

Why complete search an issue: 2^(number of variables) too much.

Subset Selection Techniques

Use Model_Selection.R

Forward selection
Backward elimination
Exhaustive

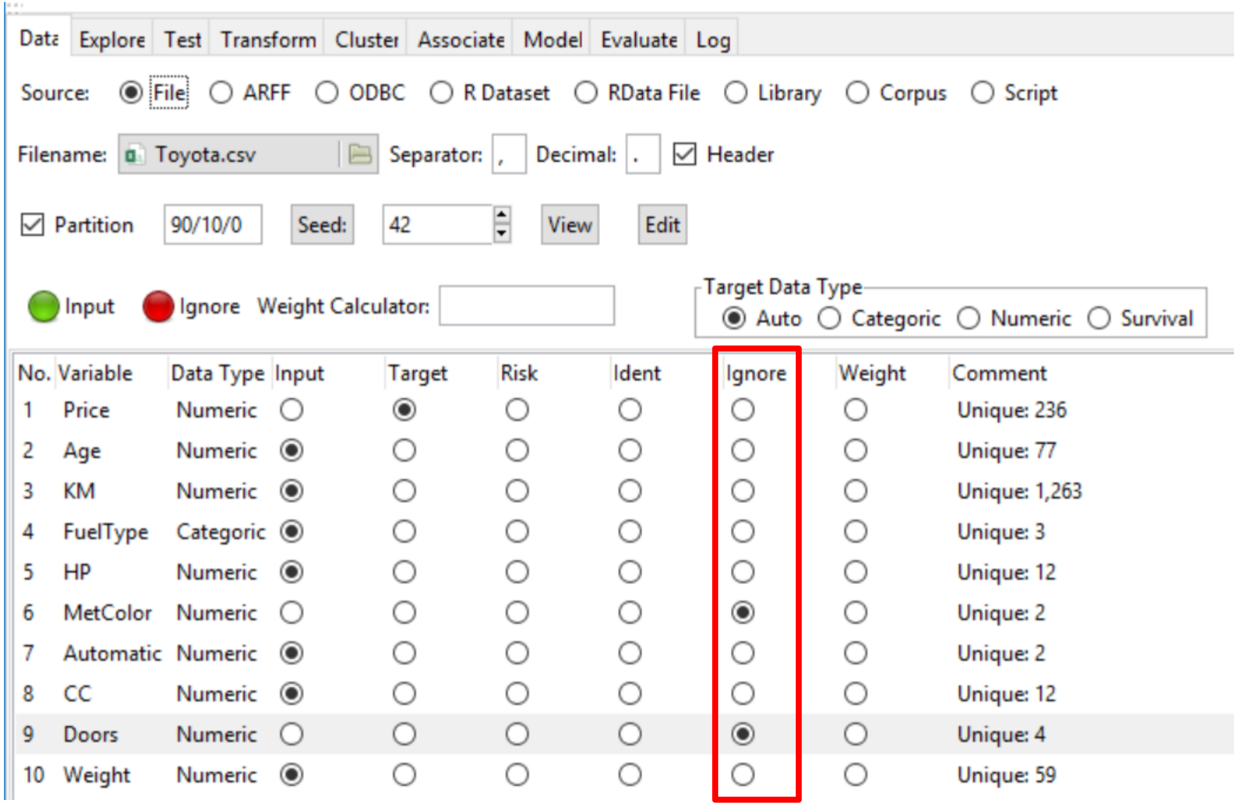
```
Toyota <- read.csv("Toyota.csv") # read data

install.packages("leaps")
library(leaps)

# nvmax represents the maximum number of
predictors to incorporate in the model
# method represents "exhaustive", "backward",
"forward"
models <- regsubsets(Price~., data = Toyota,
nvmax = 9, method="forward")
summary(models)

res.sum <- summary(models) # select the best
model based on the following criteria
data.frame(
  Adj.R2 = which.max(res.sum$adjr2))
```

The models output

```
Selection Algorithm: forward
         Age KM  FuelTypeDiesel FuelTypePetrol HP  Met     Color
Automatic
1  ( 1 ) "*" " " " "            " "            " " " "      " "
2  ( 1 ) "*" " " " "            " "            " " " "      " "
3  ( 1 ) "*" "*" " "            " "            " " " "      " "
4  ( 1 ) "*" "*" " "            " "            "*" " "      " "
5  ( 1 ) "*" "*" " "            " "            "*" " "      " "
6  ( 1 ) "*" "*" "*"            " "            "*" " "      " "
7  ( 1 ) "*" "*" "*"            "*"            "*" " "      " "
8  ( 1 ) "*" "*" "*"            "*"            "*" " "      "*"
9  ( 1 ) "*" "*" "*"            "*"            "*" "*"      "*"
         CC  Doors Weight
1  ( 1 ) " " " "   " "
2  ( 1 ) " " " "   "*"
3  ( 1 ) " " " "   "*"
4  ( 1 ) " " " "   "*"
5  ( 1 ) "*" " "   "*"
6  ( 1 ) "*" " "   "*"
7  ( 1 ) "*" " "   "*"
8  ( 1 ) "*" " "   "*"
9  ( 1 ) "*" " "   "*"
```

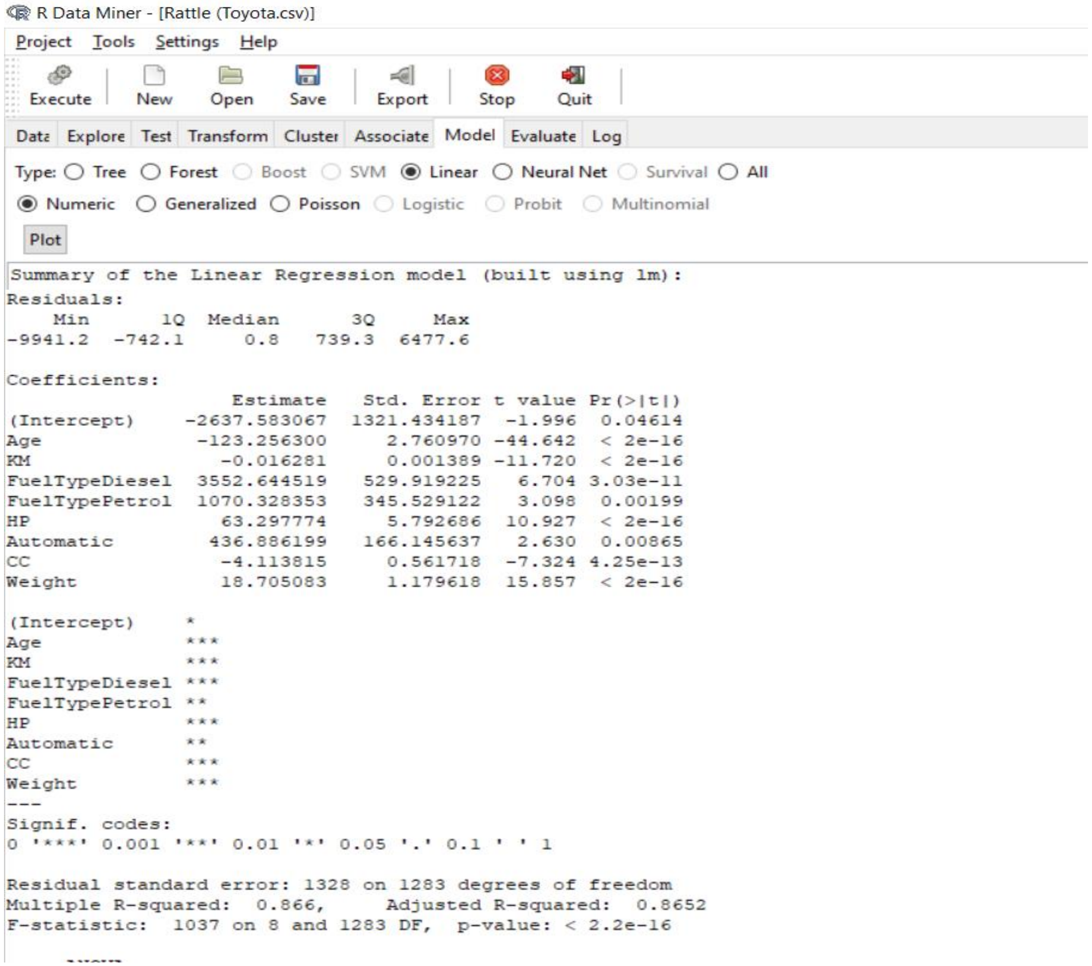
Source: Rattle GUI / Togaware

Fit a model with selected variables

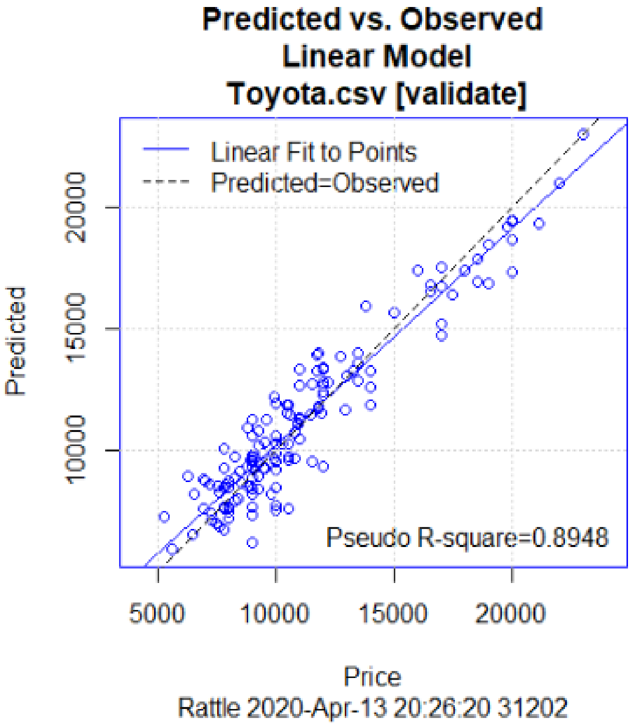


Source: Rattle GUI / Togaware

Adj R Square .8652 from .865, with simpler model.




```
Summary of the Linear Regression model (built using lm):
Residuals:
    Min      1Q  Median      3Q     Max
-9941.2  -742.1     0.8   739.3  6477.6

Coefficients:
                   Estimate   Std. Error t value Pr(>|t|)
(Intercept)    -2637.583067  1321.434187  -1.996  0.04614
Age             -123.256300     2.760970 -44.642  < 2e-16
KM                -0.016281     0.001389 -11.720  < 2e-16
FuelTypeDiesel  3552.644519   529.919225   6.704 3.03e-11
FuelTypePetrol  1070.328353   345.529122   3.098  0.00199
HP                63.297774     5.792686  10.927  < 2e-16
Automatic        436.886199   166.145637   2.630  0.00865
CC                -4.113815     0.561718  -7.324 4.25e-13
Weight            18.705083     1.179618  15.857  < 2e-16

(Intercept)    *
Age            ***
KM             ***
FuelTypeDiesel ***
FuelTypePetrol **
HP             ***
Automatic      **
CC             ***
Weight         ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1328 on 1283 degrees of freedom
Multiple R-squared:  0.866,    Adjusted R-squared:  0.8652
F-statistic:  1037 on 8 and 1283 DF,  p-value: < 2.2e-16
```



RMSE 1206 in Validate set

Source: Rattle GUI / Togaware

Further Improve the model

Higher predictive accuracy

Make it more robust

Other considerations

Dependence on sample

Inference

How about if the outcome variable is binary?

Win an auction or not?

Is it going to rain or not?

Will customer abandon contract?

Can we use regression?

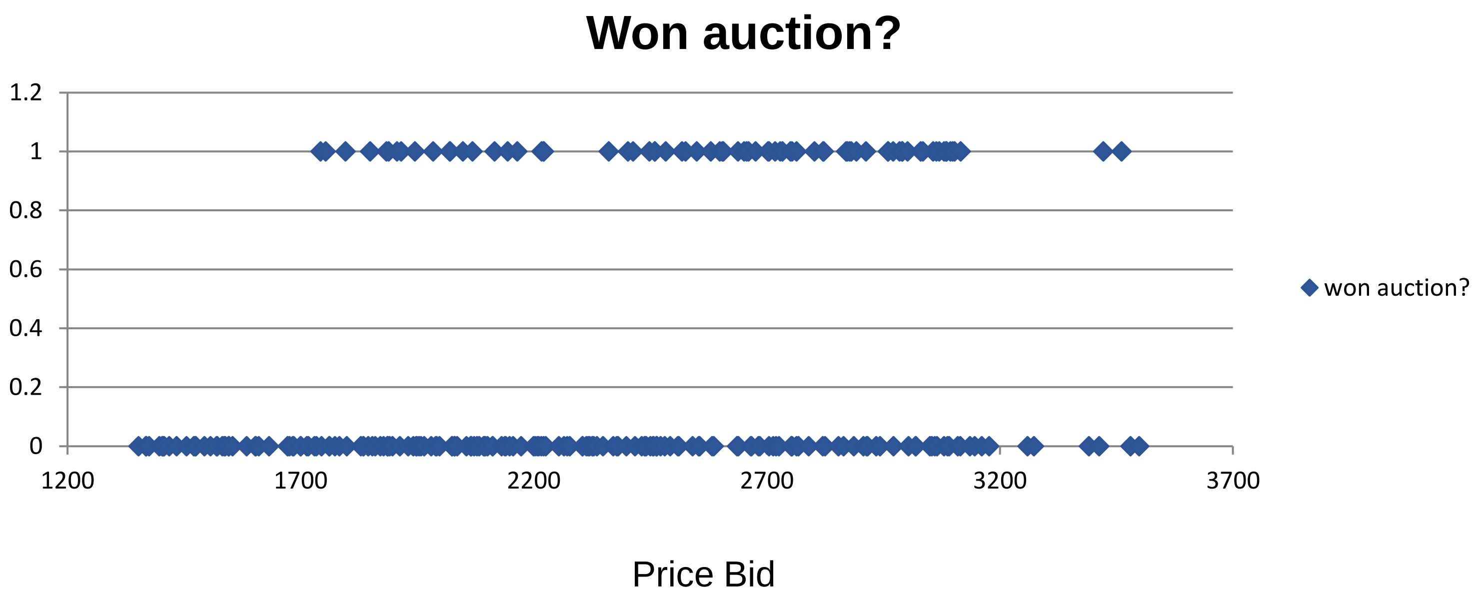
Ebay auctions for Atmos Clocks

Data Dictionary	
MSRP	Manufacturere suggested retail price
Price	Price Bid
Year	Year of manufacture
Model	Three models
Serviced	1 = Yes, 0 = No
Number of Bidders	Number of active bidders
Won auction	1 = Yes, 0 = No

Bid	MSRP	Price	MSRP-Price	Year	Model 528	Model 526	Model Baby	Serviced? (1/0)	Number of bidders	won auctic
1	4500	1604	2896	1986	1	0	0	0	3	0
2	4600	2140	2460	1976	0	1	0	0	1	0
3	4600	2116	2484	1977	0	1	0	0	18	1
4	4600	2483	2117	1980	0	1	0	0	8	1
5	4600	2726	1874	1982	0	1	0	0	16	0
6	4600	1984	2616	1987	0	1	0	0	14	1
7	4600	3030	1570	1985	0	1	0	1	3	1

Source: Rattle GUI / Togaware

Win Loss as a function of the bid



Source: Rattle GUI / Togaware

Introduce Logistic Regression

Extends idea of linear regression to situation where outcome variable is in binary category

Popular, particularly where a structured model is useful to explain or to predict

We focus on binary classification

i.e. $Y=0$ or $Y=1$

Steps: Logistic Function

f(x) = b0 + b1 x1 + b2 x2 + b3 x3 + … + bn xn

Features of the object.
One could be price.

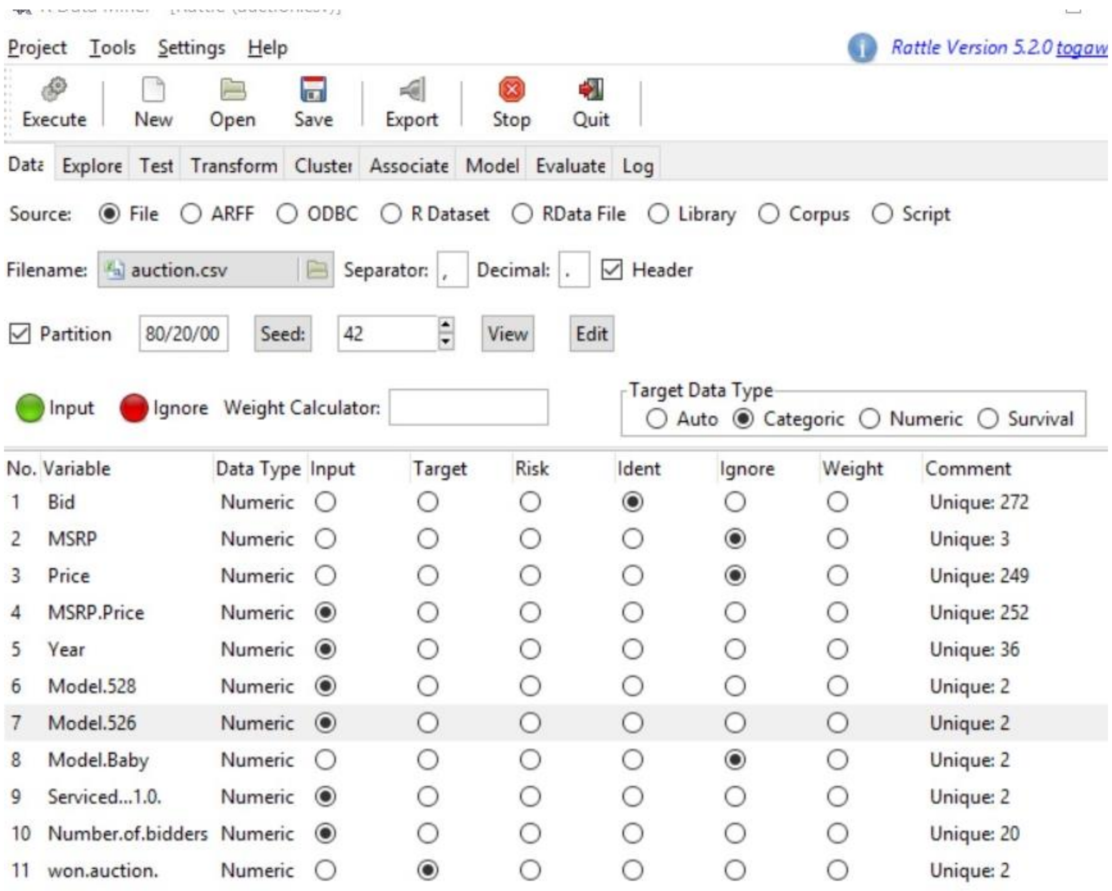
Probability(win auction), p

= exp(f(x))/[1 + exp(f(x))]

Or,

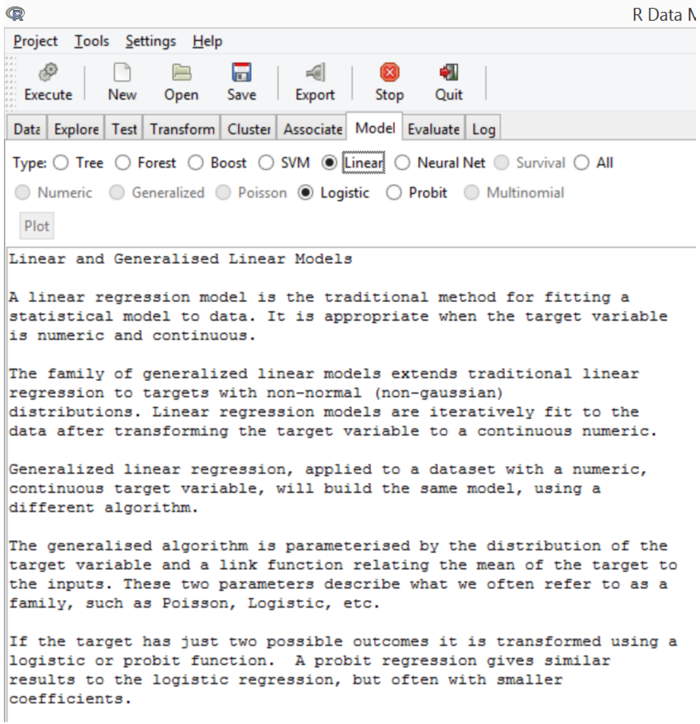
exp(f(x)) = p/(1-p) called the odds ratio

Rattle

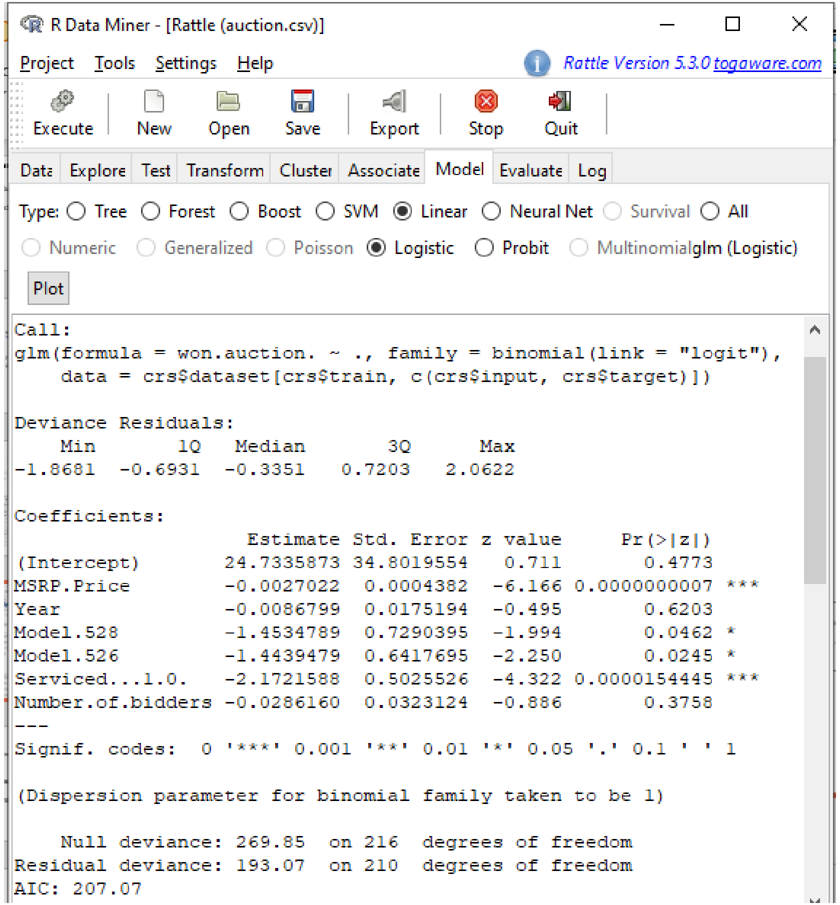


Source: Rattle GUI / Togaware

Example in Rattle:



Split used 80/20



Source: Rattle GUI / Togaware

Odds Ratio

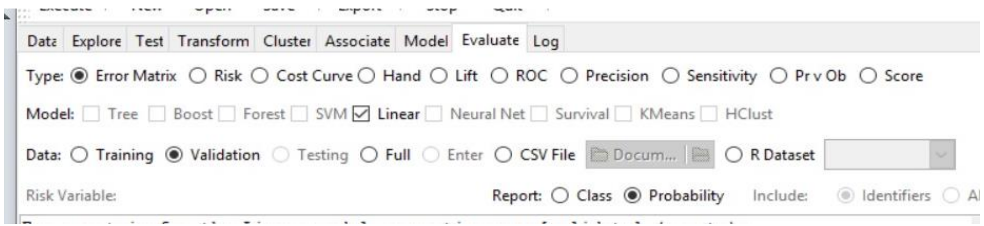
Ratio of prob win/(1-prob win)

If serviced, odds everything else same, drops to 11% of previous value. (exp(-2.172))

To recover the same log odds, price may have to increase approximately by $1000 (exp(1000*0.002))

To cancel exp(-2.172) drop price must contribute exp(+2) *approximately*

Model Performance



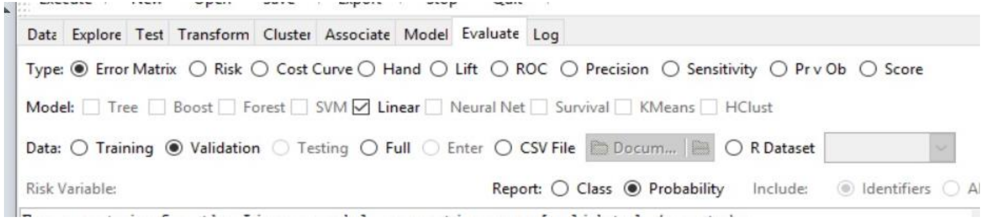
80% OF BIDS ARE LOST
BASE PREDICTION IS THE CLASS WITH HIGHEST PROBABILITY
(PREDICT ALWAYS LOSE AND WILL BE RIGHT 80%)

MODEL PREDICTS 74.5% RIGHT

BUT PREDICTS 7 OUT OF 11 WINS RIGHT WHEREAS BASE CASE IS ALWAYS INCORRECT FOR WINS

Source: Rattle GUI / Togaware

Model Performance



Error matrix for the Linear model on auction.csv [validate] (counts):

	Predicted		
Actual	0	1	Error
0	34	10	22.7
1	4	7	36.4

Error matrix for the Linear model on auction.csv [validate] (proportions):

	Predicted		
Actual	0	1	Error
0	61.8	18.2	22.7
1	7.3	12.7	36.4

Overall error: 25.5%,
Averaged class error: 29.55%

Source: Rattle GUI / Togaware

Logistic Regression Takeaways

Similar to linear regression, except that it is used with a categorical response

It can be used for both explanatory and predictive analytics

Logistic Regression Takeaways

Independent and response variables are related via a nonlinear function called the *logit*

As in linear regression, reduction of variables can be done via variable selection

Exercise for Logistic Regression

Use Boston Housing data from R library (don’t partition data). Use the mlbench data.

Run logistic regression to predict whether the tract bounds Charles river (CHAS variable) based on the following variables: medv and indus.

Exercise for Logistic Regression

Check your coefficients for indus and medv are 0.08177 and 0.07774.

Report Error matrix for the training data. It should be 6.9%.

References

Ledolter, J. (n.d.) Data Text. Retrieved from https://bit.ly/2vfZggf

Rattle
GUI / Togaware (https://rattle.togaware.com/)