Homework 2

John Lukas Facile

3/2/2020

Question 1: Saratoga House Prices

Linear Model

In order to accurately predict housing prices given a house's unique features, It is important to examine trends in pricing. To this end, I will examine the existing(medium model) and attempt to improve its predictive power and accuracy. The original model is a linear model that takes into account only a few elements from the available dataset.

```
function (file, header = TRUE, sep = ",", quote = "\"", dec = ".",
    fill = TRUE, comment.char = "", ...)
read.table(file = file, header = header, sep = sep, quote = quote,
    dec = dec, fill = fill, comment.char = comment.char, ...)
<bytecode: 0x0000000012801758>
<environment: namespace:utils>
```

```
[1] 59924.62
```

This value indicates the RMSE of the initial linear model I will set out to improve on. I believe that I can lower this out of sample error by including more of the house elements included in the data set. Doing so yields an out of sample error of:

```
[1] 56638.5
```

This new linear model appears to improve on the previous model in terms of out sample error. This new model takes into account all values available except for whether or not there a sewer nearby, how many fireplaces there are

Important Factors

In order to determine what factors contribute more significantly to this model performing better in predicting house prices, I tried two methods. The first being examining the prediction error changes when I drop a certain variable. The variables that immediately stood out were land value anf lot size. Due to how slow this method was and my own personal propensity to overlook variables, I tested the variables using a lasso net in order to find out which variables appear to be more important.

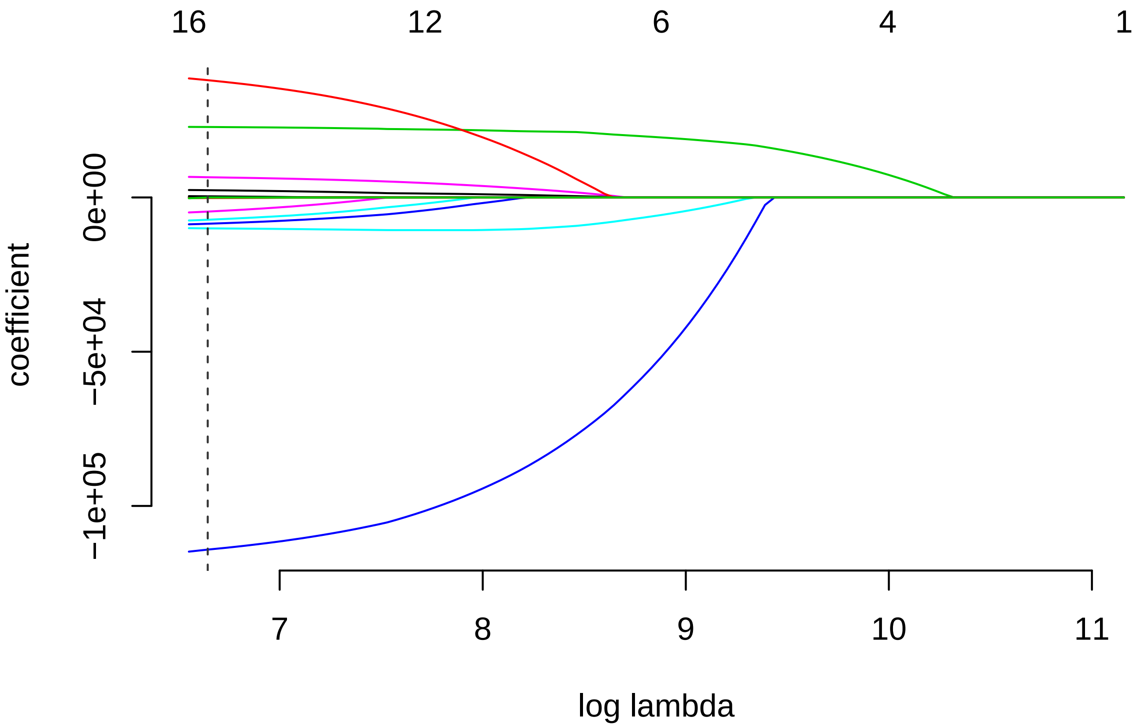
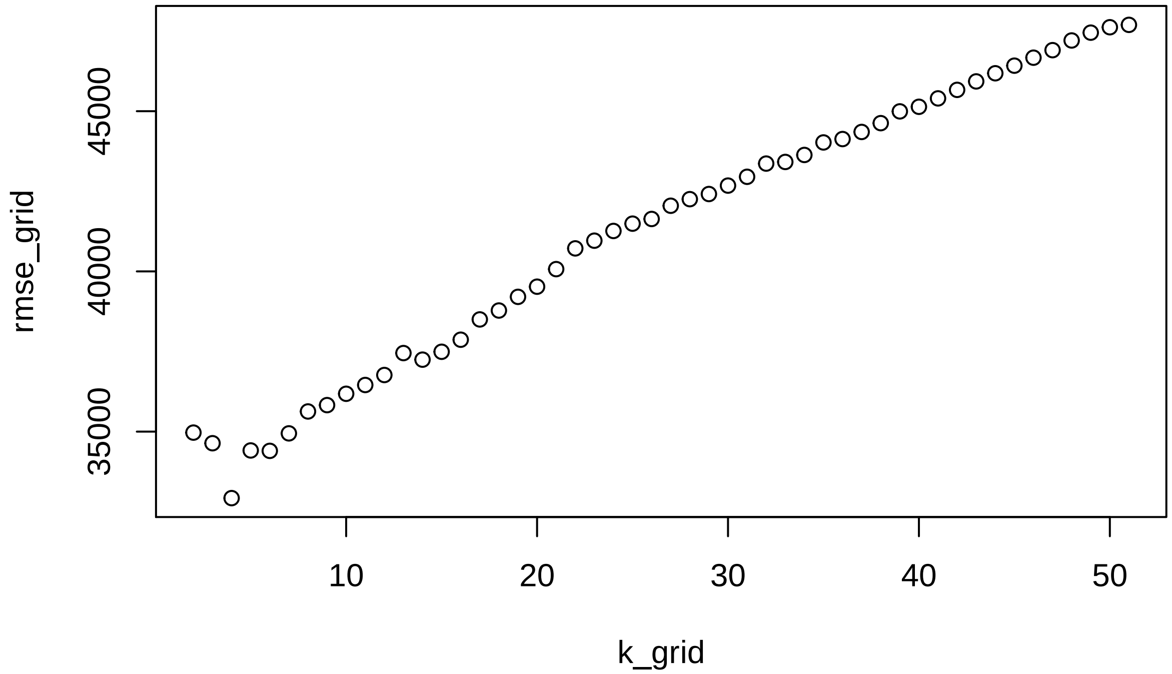


Table 1

intercept	
lotSize	
age	
landValue	
livingArea	
From this tab	le we can see what variables appear to “leap” first fron
## Nonparamet	
The best pred	ictive model for housing prices may not be a linear one, or a parametric one for that matter. In order to



For the following KNN model, I will use 3 as the value of K as it minimizes out of sample prediction error.

I will now test the new model across multiple train-test splits in order to get the average performance of this model.

```
     result
 Min.   :29475
 1st Qu.:33970
 Median :36763
 Mean   :36368
 3rd Qu.:38317
 Max.   :43096
```

This non parametric model appears to significantly outperform both previous models.
In summary, the final model is the most accurate among the three in predicting house prices in Saratoga.

Question 2: Hospital Audit

Conservative Radiologists

In this audit, I will be examining the performance of radiologists in deciding to recall and ultimately catch wether or not a patient has cancer. The first question I am going to answer is whether or not certain radiologists are more conservative, or more likely to recall a patient controlling for factors such as history, age, etc.

```
# weights:  17 (16 variable)
initial  value 684.136267
iter  10 value 411.871589
iter  20 value 400.100732
final  value 399.992800
converged
```

```
=============================================================
                                       price
-------------------------------------------------------------
lotSize                             8,449.341***
                                    (2,391.303)

landValue                             0.949***
                                      (0.051)

livingArea                           65.816***
                                      (5.293)

bedrooms                           -11,126.460***
                                    (3,026.706)

bathrooms                          25,314.600***
                                    (3,714.570)

rooms                               3,352.995***
                                    (1,127.066)

heatinghot water/steam             -11,427.580**
                                    (4,730.831)

heatingelectric                        74.192
                                    (13,781.720)

fuelelectric                         -8,777.481
                                    (13,581.960)

fueloil                              -3,177.794
                                    (5,467.326)

centralAirNo                        -8,966.554**
                                    (3,927.640)

Constant                           33,137.280***
                                    (8,371.595)

N                                      1,382
R2                                     0.624
Adjusted R2                            0.621
Residual Std. Error            60,811.270 (df = 1370)
F Statistic                  206.828*** (df = 11; 1370)
=============================================================
Notes:                 ***Significant at the 1 percent level.
```

```
                     **Significant at the 5 percent level.
                      *Significant at the 10 percent level.
```

```
      recall
cancer   0   1
     0 177  13
     1   3   4
```

```
      recall
cancer   0   1
     0 157  33
     1   2   5
```

I use a multinomial logit model with the specific intention to examine the changes in probability of recall for each radiologist. Based on the results, I construct confusion matrices to examine the relative differences between certain radiologists. Radiologist 89 recalled significantly more patients than radiologist 34. Thus, holding patient risk factors equal, some radiologists appear to recall patients at higher rate and could be considered more conservative.

Weighing Factors

In order to determine wether or not radiologists should be weighing certain risk factors differently, I am going to examine what happens to the impact on accuracy inclusion of different risk factors have on predicting cancer outcomes of patients.

```
========================================================
                                  cancer
--------------------------------------------------------
recall                           2.261***
                                 (0.348)

Constant                        -4.006***
                                 (0.261)

N                                  987
Log Likelihood                  -137.440
Akaike Inf. Crit.                278.881
========================================================
Notes:              ***Significant at the 1 percent level.
                     **Significant at the 5 percent level.
                      *Significant at the 10 percent level.
```

```
========================================================
                                  cancer
--------------------------------------------------------
recall                           2.257***
                                 (0.348)

history                           0.206
                                 (0.423)
```

```
Constant                        -4.045***
                                 (0.274)

N                                  987
Log Likelihood                   -137.326
Akaike Inf. Crit.                280.651
========================================================
Notes:            ***Significant at the 1 percent level.
                   **Significant at the 5 percent level.
                   *Significant at the 10 percent level.


===============================================================
                                         cancer
---------------------------------------------------------------
recall                                  2.293***
                                        (0.362)

ageage5059                               0.411
                                        (0.634)

ageage6069                               0.353
                                        (0.804)

ageage70plus                             1.369*
                                        (0.726)

history                                  0.210
                                        (0.433)

symptoms                                 0.036
                                        (0.702)

menopausepostmenoNoHT                    -0.175
                                        (0.454)

menopausepostmenounknown                 0.770
                                        (0.726)

menopausepremeno                         0.152
                                        (0.655)

densitydensity2                          0.723
                                        (1.075)

densitydensity3                          0.847
                                        (1.071)

densitydensity4                          1.952*
                                        (1.124)

Constant                               -5.576***
                                        (1.247)
```

```
N                                         987
Log Likelihood                          -130.588
Akaike Inf. Crit.                        287.177
=================================================================
Notes:                       ***Significant at the 1 percent level.
                              **Significant at the 5 percent level.
                              *Significant at the 10 percent level.
```

According to the data, taking family history into account actually yields worse results than a model where the decision to recall is the only factor. Going further, including all other risk factors also appears to reduce the predictive power of the model. If the radiologist were to appriately take into account these risk factors, the model including all of the risk factors should do improve on the model where the only consideration is patient recall, which is not what this data is showing.

Question 3: Viral Articles

Knowing the secret formula of why articles go viral and how one can consistently use that formula would be exceeding useful for marketing purpose. To this end I will be trying to estimate a model for predicting if an article would go viral. I will be comparing a model that estimates amount of shares and a probabilistic model that predicts wether or not a model exceeds a threshold that qualifies it as “going viral”.

Part 1: Modeling number of shares

For this first model, I will be estimating a model that predicts the amount of shares that an article would get given certain characteristics. Due to the distribution of number of shares of articles, I will be using the log of shares as my variable of interest.

```
[1] 0.9028189
```

```
[1] 0.4934416
```

```
[1] 0.7665532
```

The model predicted that 76% of the articles went viral while only 49% of the articles went viral.

Part 2: Predicting if an article is going to go viral

Now lets examine a probablistic model predicting whether or not an article exceeds a “viral” threshold.

```
n_TRUE n_TRUE n_TRUE n_TRUE
 16206  15848   3876   3714
```

The Logit model Appears to produce better results. The last Values printed is the confusion matrix.