# SDS Exercise 2

## KNN Practice

To use K-nearest neighbors to build a predictive model for price, given mileage, separately for each of two trim levels: 350 and 65 AMG, we first create two datasets containing only the 350 model and the 65 AMG model. With train-test splits and applying different values of K, we report the root mean square error of each k and the relationship between RMSE and K as follows.

### 350 Model

#### RMSE for each value of K

$$k = 3 : RMSE_{out} = 1.1084 \times 10^4$$

$$k = 5 : RMSE_{out} = 9836$$

$$k = 10 : RMSE_{out} = 9422$$

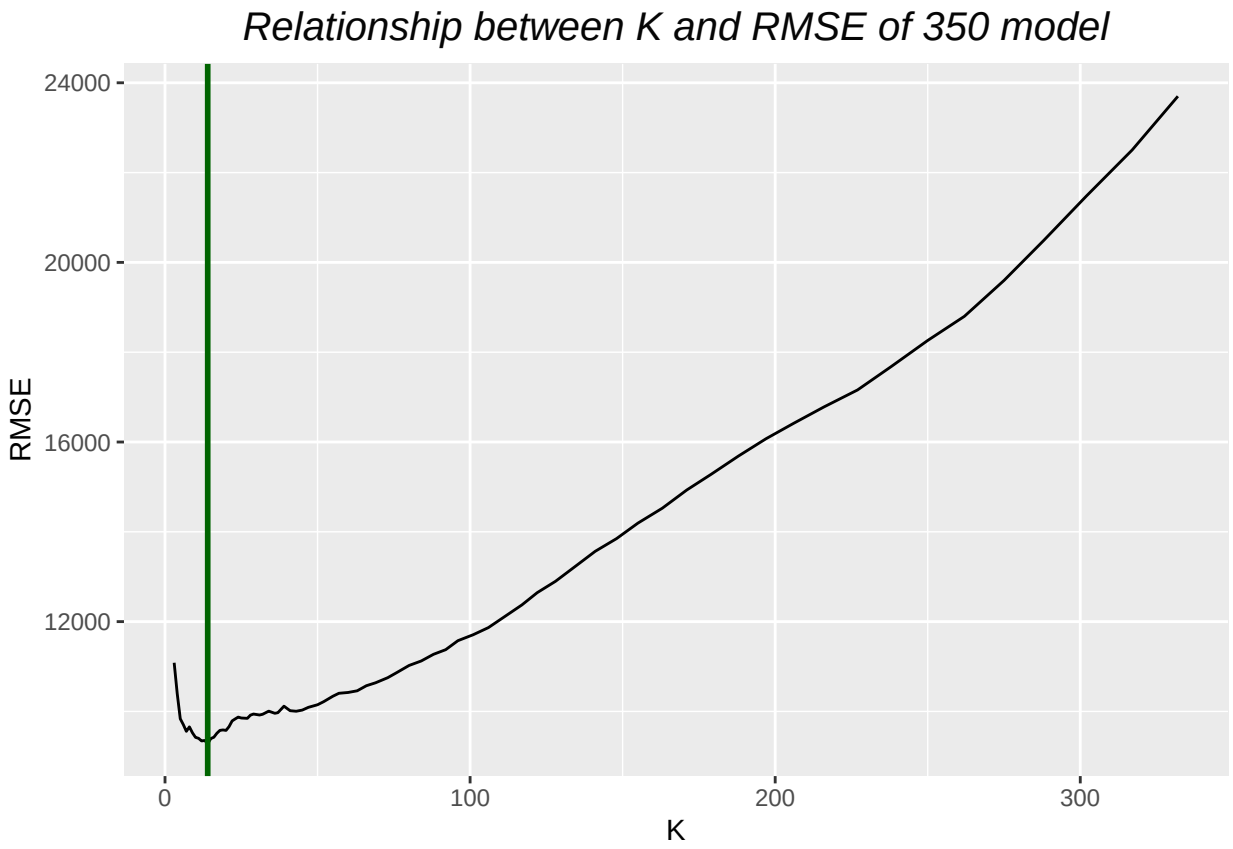
$$k = 20 : RMSE_{out} = 9578$$

$$k = 50 : RMSE_{out} = 1.0149 \times 10^4$$

$$k = 100 : RMSE_{out} = 1.1709 \times 10^4$$

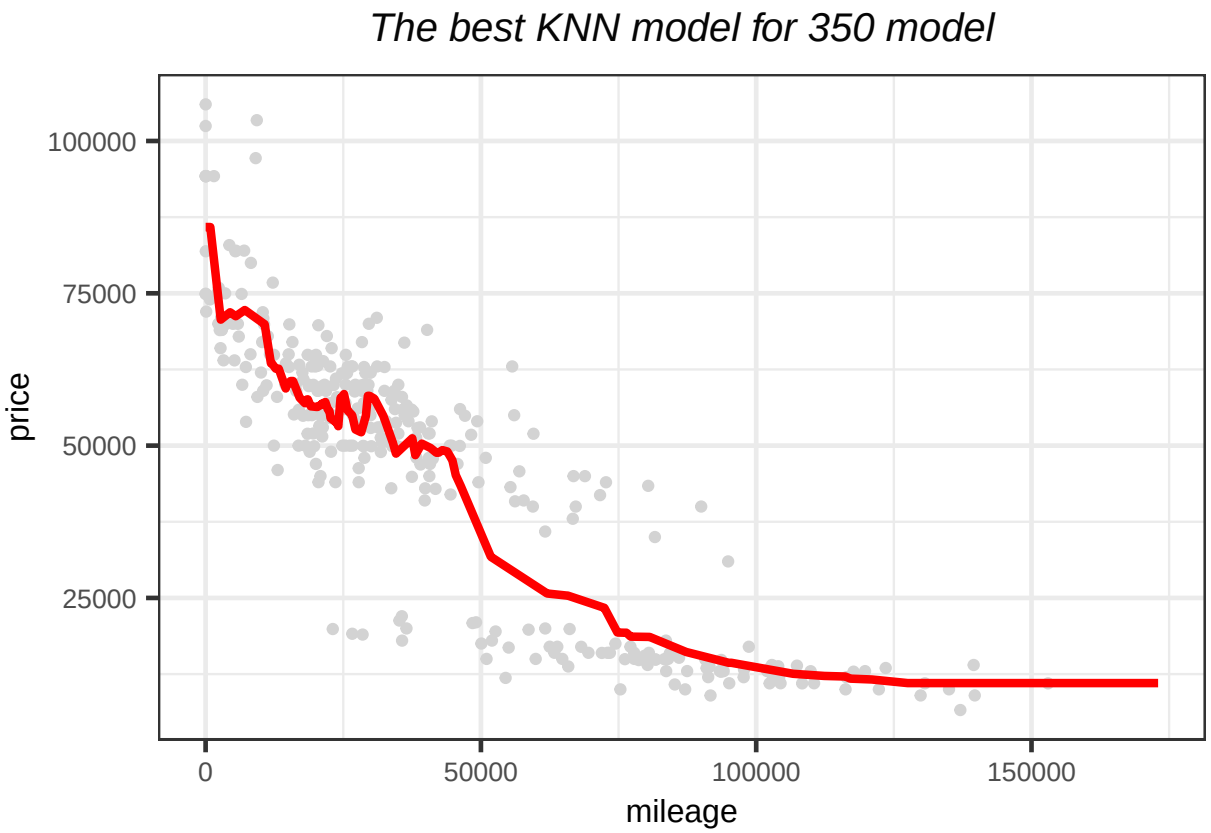
$$k = 332 : RMSE_{out} = 2.37 \times 10^4$$

**Plot the relationship between RMSE and K**




*The value of best k is* : 14

**Plot the best KNN model**




$$RMSE_{out} = 9281$$

## 65 AMG Model

**RMSE for each value of K**

$$k = 3 : RMSE_{out} = 1.9038 \times 10^4$$

$$k = 5 : RMSE_{out} = 1.5699 \times 10^4$$

$$k = 10 : RMSE_{out} = 1.3459 \times 10^4$$

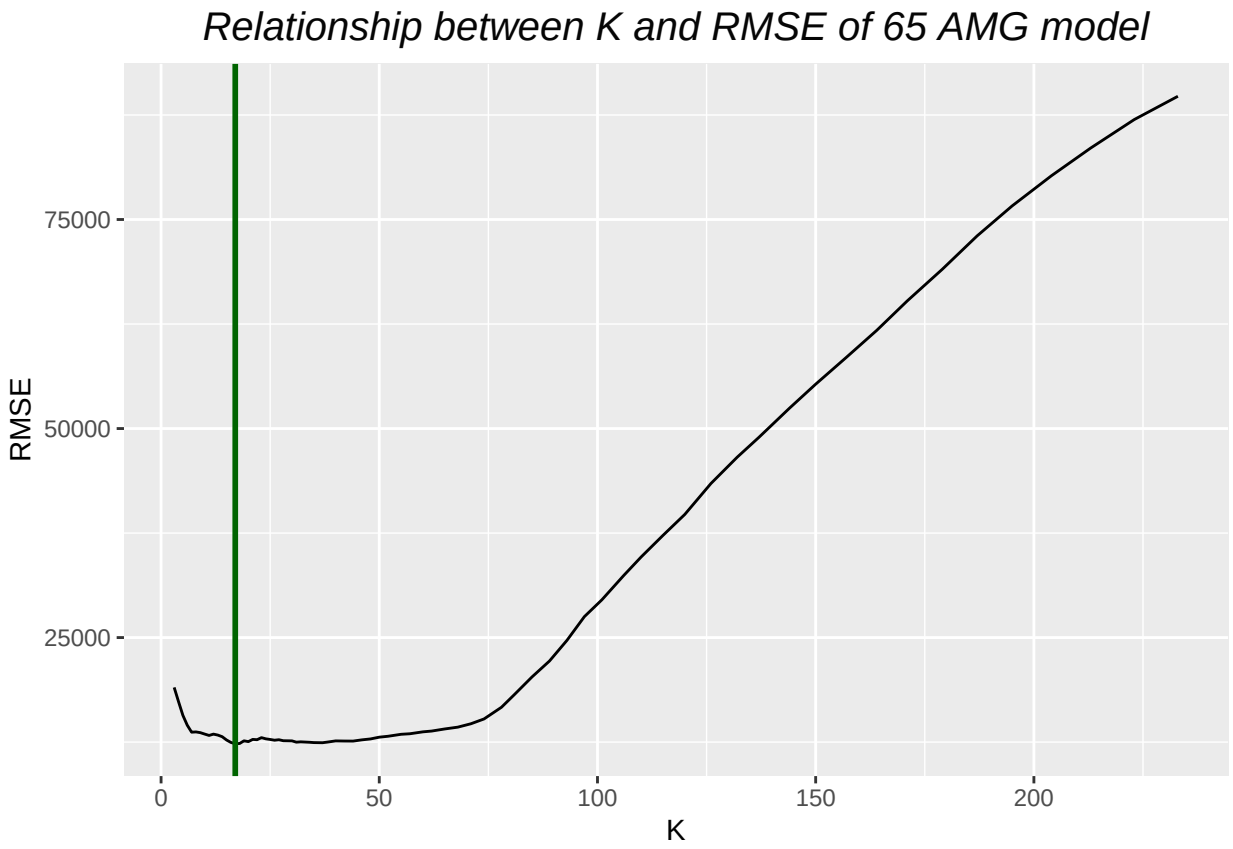
$$k = 20 : RMSE_{out} = 1.2556 \times 10^4$$

$$k = 50 : RMSE_{out} = 1.3086 \times 10^4$$

$$k = 100 : RMSE_{out} = 2.9015 \times 10^4$$

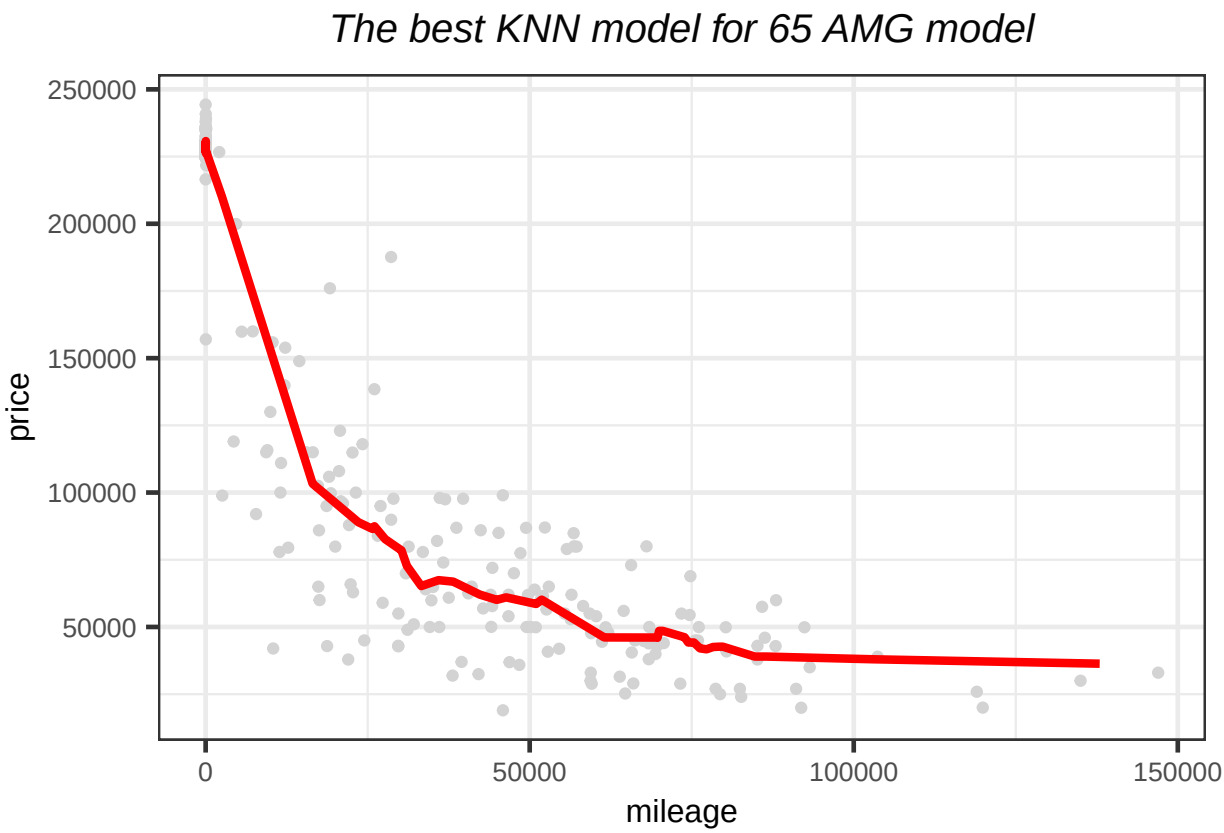
$$k = 233 : RMSE_{out} = 8.9739 \times 10^4$$

**Plot the relationship between RMSE and K**



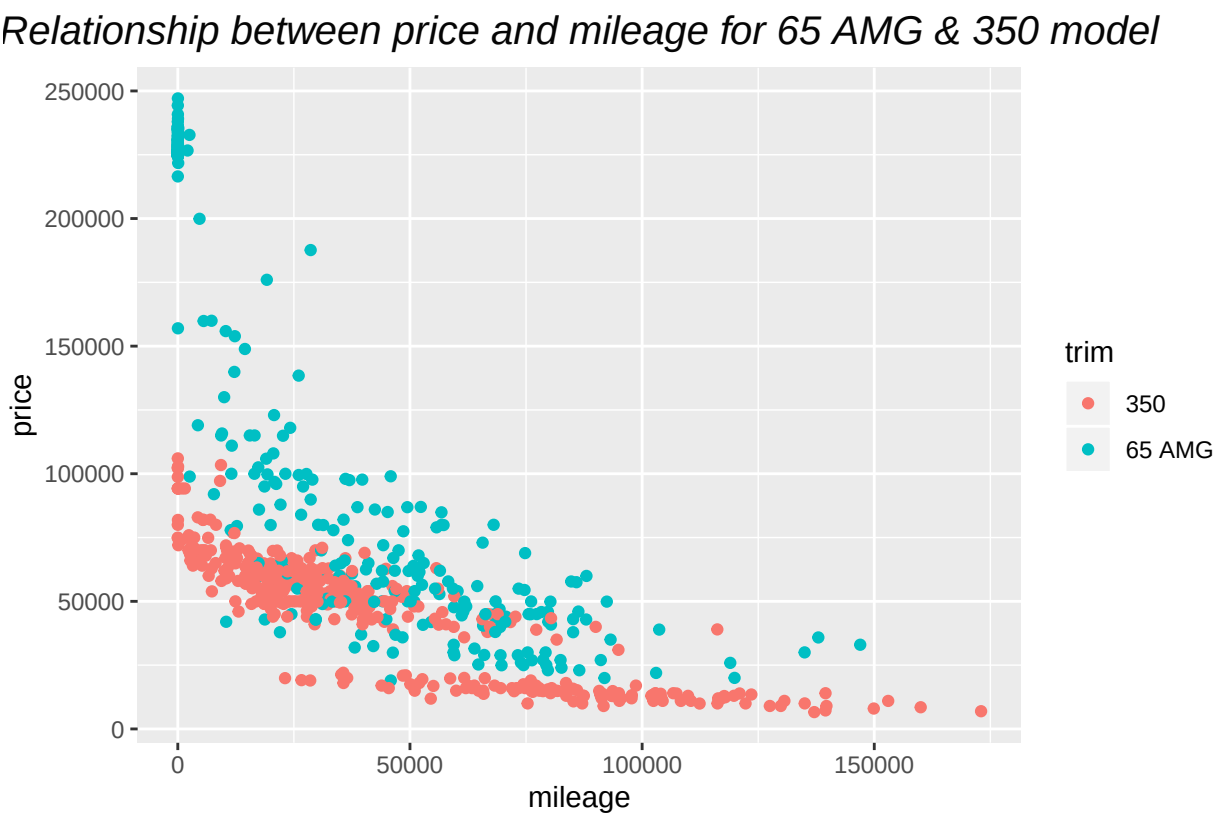

*The value of best k is* : 17

**Plot the best KNN model**




$$RMSE_{out} = 1.2306 \times 10^4$$

### Conclusion




*The price of 65 AMG models have greater variability than that of 350 model.*

With repeated trials, we investigate the relationship between RMSE and K, plot the best KNN model for each of the two trim levels. We also find that the optimal value of K is generally higher for 65 AMG model than it is for 350 model. This seems unusual as the conventional wisdom states that K value is generally correlated with the sample size, and 350 model has more observations than the 65 AMG model.

We try to explain this phenomenon with the scatter plot above, which shows that the price of 65 AMG model has greater variance than that of 350 model. This implies the data of 65 AMG model has more "noise". And when larger values of k are used, the KNN algorithm reduces the noise in the data, which results in better out-of-sample performance. On the contrary, 350 model does not have much "noise". Therefore, a smaller k value returns better out-of-sample predictions.

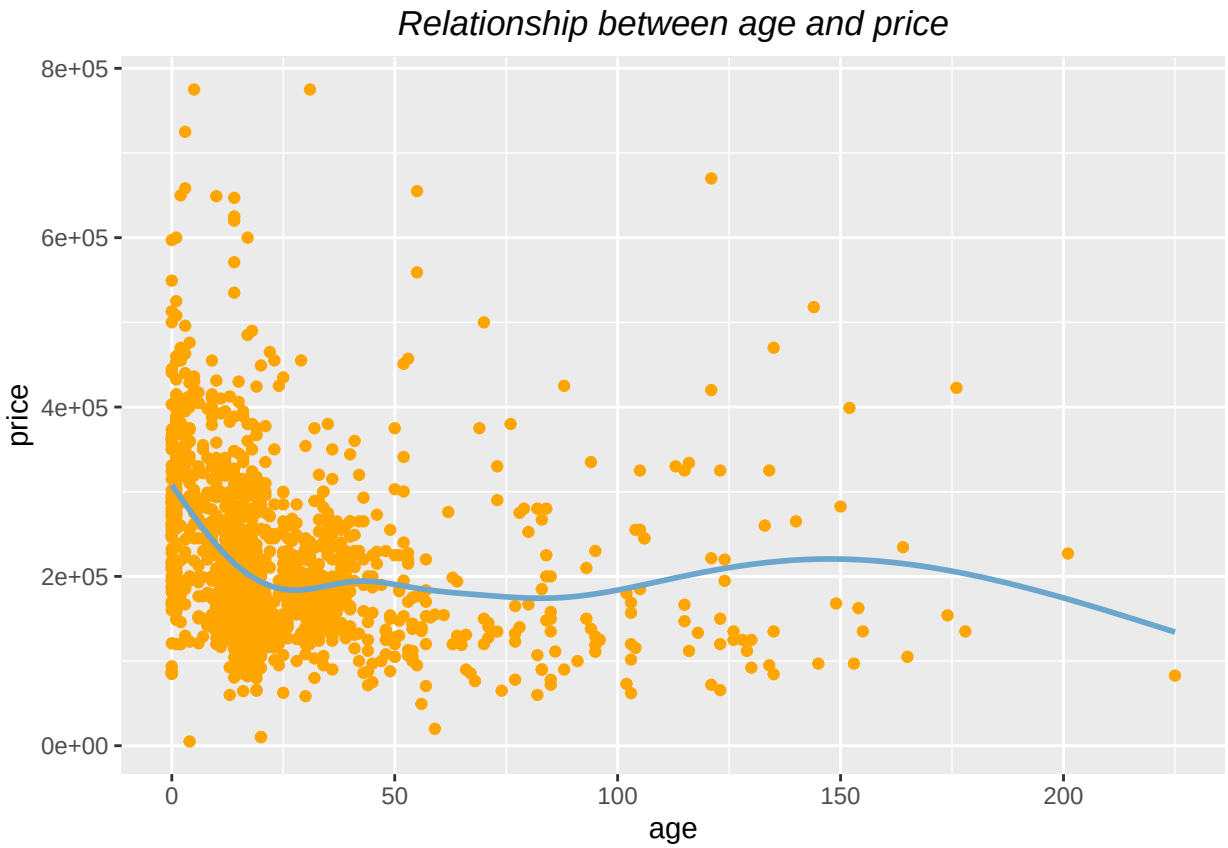
## Saratoga house prices

To investigate the relationship between house prices and key characteristics, we apply two methodologies: a hand-built linear model and a KNN model.

### Hand-built model

In an attempt to create a better performing model, we first look into the relationship between house prices and continuous variables and determine whether the relationship is linear or not. From real-life intuitions, we believe age to have a negative impact on the house price yet we expect a diminishing marginal impact of age on house prices as people are more sensitive to the ages of relatively new properties.

**Transformation: Age and Price**




The graph shows that age does seem to have a diminishing negative impact on price. Thus, to accommodate the diminishing marginal effect of age, we attempt two common types of transformation: log(age) and square root of age

**log(age) vs sqrt(age)**

```
## [1] 54337.92 54361.77 54648.79
```

The root mean square error of using log transformation is $5.4338 \times 10^4$, while for square root transformation is $5.4362 \times 10^4$ and for no transformation is $5.4649 \times 10^4$. With the log transformation possessing the smallest RMSE, we adopt log transformation of age.

**Finding variables & interactions with statistically significant impacts**

We further investigate the impact of other discrete/continuous variables on house prices and incorporate statistically significant variables into our hand-built model.

```
##
## Call:
## lm(formula = price ~ . - sqrtage - age, data = saratoga_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -236068  -35311   -5354   27813  459822
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.194e+05  2.170e+04   5.502 4.48e-08 ***
## lotSize                  8.216e+03  2.913e+03   2.820  0.00487 **
## landValue                9.293e-01  5.344e-02  17.390  < 2e-16 ***
## livingArea               6.851e+01  5.271e+00  12.999  < 2e-16 ***
## pctCollege              -1.082e+02  1.730e+02  -0.625  0.53196
## bedrooms                -6.577e+03  2.971e+03  -2.214  0.02702 *
## fireplaces               3.641e+03  3.385e+03   1.076  0.28220
## bathrooms                2.004e+04  3.801e+03   5.271 1.58e-07 ***
## rooms                    2.986e+03  1.096e+03   2.725  0.00651 **
## heatinghot water/steam -8.840e+03  4.750e+03  -1.861  0.06296 .
## heatingelectric          8.063e+03  1.343e+04   0.600  0.54839
## fuelelectric            -1.531e+04  1.316e+04  -1.164  0.24468
## fueloil                 -3.864e+03  5.740e+03  -0.673  0.50102
## sewerpublic/commercial   1.822e+03  4.168e+03   0.437  0.66199
## sewernone               -1.402e+04  1.924e+04  -0.728  0.46649
## waterfrontNo            -1.254e+05  1.638e+04  -7.654 3.66e-14 ***
## newConstructionNo        5.887e+04  9.272e+03   6.349 2.95e-10 ***
## centralAirNo            -1.121e+04  3.974e+03  -2.822  0.00484 **
## lgage                   -6.737e+03  2.059e+03  -3.271  0.00110 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59030 on 1363 degrees of freedom
## Multiple R-squared:  0.6576, Adjusted R-squared:  0.653
## F-statistic: 145.4 on 18 and 1363 DF,  p-value: < 2.2e-16
```

From the regression results above, we find lotSize, landValue, livingArea, bathrooms, rooms, heatinghot water/steam, waterfront, newConstruction, CentralAir and lgage to be statistically significant at 5% significance level. Next, we go on to add the effects of interactions into our model by incorporating preceding variables and their interactions:

```
##
## Call:
## lm(formula = price ~ (lotSize + landValue + livingArea + bathrooms +
##     rooms + waterfrontNo + heatingHotwater + newConstruction +
##     centralAir + lgage)^2, data = saratoga_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -236403  -32748   -3745   27838  419741
##
## Coefficients: (1 not defined because of singularities)
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       2.115e+04  1.228e+05   0.172  0.86325
## lotSize                           4.931e+04  7.312e+04   0.674  0.50018
## landValue                         2.702e-01  4.300e-01   0.628  0.52993
## livingArea                       -5.084e+01  7.999e+01  -0.636  0.52515
## bathrooms                         9.099e+04  6.115e+04   1.488  0.13702
## rooms                             1.619e+04  1.598e+04   1.013  0.31108
## waterfrontNo                     -3.387e+04  1.120e+05  -0.303  0.76231
```

```
## heatingHotwater                   -8.281e+03  9.546e+04  -0.087  0.93088
## newConstructionNo                  6.019e+04  4.718e+04   1.276  0.20228
## centralAirNo                       1.384e+05  6.373e+04   2.171  0.03009 *
## lgage                              7.126e+03  3.260e+04   0.219  0.82698
## lotSize:landValue                 -3.984e-01  1.020e-01  -3.904 9.92e-05 ***
## lotSize:livingArea                -1.789e+01  6.496e+00  -2.754  0.00597 **
## lotSize:bathrooms                 -2.258e+03  4.854e+03  -0.465  0.64187
## lotSize:rooms                      6.810e+02  1.665e+03   0.409  0.68271
## lotSize:waterfrontNo               8.489e+04  6.073e+04   1.398  0.16241
## lotSize:heatingHotwater            1.073e+04  6.350e+03   1.689  0.09142 .
## lotSize:newConstructionNo         -6.962e+04  3.755e+04  -1.854  0.06395 .
## lotSize:centralAirNo              -2.192e+04  8.047e+03  -2.724  0.00654 **
## lotSize:lgage                      7.425e+02  3.337e+03   0.223  0.82394
## landValue:livingArea              -7.417e-05  1.399e-04  -0.530  0.59600
## landValue:bathrooms                6.117e-02  1.259e-01   0.486  0.62706
## landValue:rooms                    9.488e-03  3.432e-02   0.276  0.78226
## landValue:waterfrontNo             1.277e-01  2.556e-01   0.499  0.61753
## landValue:heatingHotwater         -1.151e-01  1.618e-01  -0.711  0.47694
## landValue:newConstructionNo        2.689e-01  2.261e-01   1.189  0.23469
## landValue:centralAirNo            -3.441e-03  1.442e-01  -0.024  0.98096
## landValue:lgage                    1.672e-01  6.351e-02   2.632  0.00858 **
## livingArea:bathrooms               3.577e+00  6.134e+00   0.583  0.55988
## livingArea:rooms                   2.329e+00  1.714e+00   1.359  0.17436
## livingArea:waterfrontNo            1.191e+02  6.987e+01   1.705  0.08851 .
## livingArea:heatingHotwater         1.252e+01  1.353e+01   0.925  0.35490
## livingArea:newConstructionNo       2.874e+01  3.489e+01   0.824  0.41021
## livingArea:centralAirNo           -1.741e+01  1.180e+01  -1.475  0.14049
## livingArea:lgage                  -1.167e+01  5.579e+00  -2.091  0.03674 *
## bathrooms:rooms                   -3.118e+02  1.917e+03  -0.163  0.87083
## bathrooms:waterfrontNo            -1.015e+05  5.252e+04  -1.932  0.05356 .
## bathrooms:heatingHotwater         -1.932e+04  1.026e+04  -1.883  0.05987 .
## bathrooms:newConstructionNo        2.026e+04  2.767e+04   0.732  0.46423
## bathrooms:centralAirNo            -2.739e+03  8.416e+03  -0.325  0.74489
## bathrooms:lgage                    3.585e+03  4.177e+03   0.858  0.39091
## rooms:waterfrontNo                -1.376e+04  1.343e+04  -1.024  0.30593
## rooms:heatingHotwater             -5.077e+03  2.839e+03  -1.788  0.07394 .
## rooms:newConstructionNo           -9.875e+03  6.685e+03  -1.477  0.13984
## rooms:centralAirNo                 1.039e+02  2.348e+03   0.044  0.96470
## rooms:lgage                        1.946e+03  1.262e+03   1.542  0.12328
## waterfrontNo:heatingHotwater       1.315e+04  6.470e+04   0.203  0.83896
## waterfrontNo:newConstructionNo            NA         NA      NA       NA
## waterfrontNo:centralAirNo         -8.850e+04  5.695e+04  -1.554  0.12042
## waterfrontNo:lgage                -6.621e+03  1.994e+04  -0.332  0.73996
## heatingHotwater:newConstructionNo  3.375e+04  6.669e+04   0.506  0.61290
## heatingHotwater:centralAirNo       1.134e+03  1.397e+04   0.081  0.93531
## heatingHotwater:lgage              3.071e+02  6.318e+03   0.049  0.96124
## newConstructionNo:centralAirNo    -4.209e+04  1.999e+04  -2.105  0.03544 *
## newConstructionNo:lgage           -1.817e+04  2.385e+04  -0.762  0.44639
## centralAirNo:lgage                 8.176e+03  4.593e+03   1.780  0.07529 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56810 on 1327 degrees of freedom
## Multiple R-squared:  0.6912, Adjusted R-squared:  0.6787
```

```
## F-statistic: 55.01 on 54 and 1327 DF,  p-value: < 2.2e-16
```

From the regression results above, we find landValue, heatingHotwater lotSize:landValue, lotSize:livingArea, lotSize:rooms, lotSize:centralAirNo, landValue:waterfrontNo, landValue:heatingHotwater, landValue:lgage, livingArea:bathrooms, livingArea:waterfrontNo, bathrooms:newConstructionNo, centralAirNo:lgage to be statistically significant at 10% confidence level. Therefore, we build a new model using the all the independent variables & preceding interactions.

```
##
## Call:
## lm(formula = price ~ landValue + lotSize + landValue + livingArea +
##     rooms + centralAir + waterfrontNo + heatingHotwater + lgage +
##     bathrooms + newConstruction + lotSize:landValue + lotSize:livingArea +
##     lotSize:rooms + lotSize:centralAir + landValue:waterfrontNo +
##     landValue:heatingHotwater + landValue:lgage + livingArea:bathrooms +
##     livingArea:waterfrontNo + bathrooms:newConstruction + centralAir:lgage,
##     data = saratoga_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -237861  -33433   -3252   29177  431236
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.389e+05  6.031e+04   3.961 7.84e-05 ***
## landValue                    5.712e-01  1.891e-01   3.021 0.002566 **
## lotSize                      5.834e+04  1.125e+04   5.186 2.48e-07 ***
## livingArea                   1.860e+01  2.826e+01   0.658 0.510603
## rooms                        1.707e+03  1.272e+03   1.342 0.179879
## centralAirNo                -3.106e+04  1.000e+04  -3.106 0.001936 **
## waterfrontNo                -1.984e+05  4.439e+04  -4.468 8.54e-06 ***
## heatingHotwater             -4.158e+03  6.250e+03  -0.665 0.505971
## lgage                       -2.080e+04  3.389e+03  -6.138 1.09e-09 ***
## bathrooms                   -2.010e+03  1.651e+04  -0.122 0.903159
## newConstructionNo            5.196e+04  3.949e+04   1.316 0.188431
## landValue:lotSize           -3.702e-01  8.320e-02  -4.449 9.32e-06 ***
## lotSize:livingArea          -1.411e+01  5.510e+00  -2.562 0.010521 *
## lotSize:rooms                3.735e+02  1.619e+03   0.231 0.817534
## lotSize:centralAirNo        -1.879e+04  6.203e+03  -3.029 0.002503 **
## landValue:waterfrontNo       8.078e-02  1.896e-01   0.426 0.670084
## landValue:heatingHotwater   -1.990e-01  1.389e-01  -1.433 0.152088
## landValue:lgage              1.982e-01  4.011e-02   4.941 8.73e-07 ***
## livingArea:bathrooms         1.140e+01  3.127e+00   3.644 0.000278 ***
## livingArea:waterfrontNo      3.272e+01  2.715e+01   1.205 0.228332
## bathrooms:newConstructionNo  5.568e+02  1.484e+04   0.038 0.970072
## centralAirNo:lgage           9.112e+03  3.236e+03   2.816 0.004932 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57450 on 1360 degrees of freedom
## Multiple R-squared:  0.6764, Adjusted R-squared:  0.6714
## F-statistic: 135.4 on 21 and 1360 DF,  p-value: < 2.2e-16
```

From the regression results above, we further eliminate interactions that are not statistically significant,

namely, livingArea:waterfrontNo, bathrooms:newConstructionNo, landValue:waterfrontNo, lotSize:rooms.

```
##
## Call:
## lm(formula = price ~ lotSize + landValue + livingArea + rooms +
##     centralAir + waterfrontNo + heatingHotwater + lgage + bathrooms +
##     newConstruction + lotSize:landValue + lotSize:livingArea +
##     lotSize:centralAir + landValue:heatingHotwater + landValue:lgage +
##     livingArea:bathrooms + centralAir:lgage, data = saratoga_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -237785  -33508   -3615   29494  431297
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.785e+05  2.516e+04   7.094 2.09e-12 ***
## lotSize                     5.942e+04  1.047e+04   5.675 1.70e-08 ***
## landValue                   5.923e-01  1.089e-01   5.437 6.41e-08 ***
## livingArea                  5.096e+01  8.530e+00   5.974 2.95e-09 ***
## rooms                       1.826e+03  9.940e+02   1.837 0.066445 .
## centralAirNo               -3.166e+04  9.889e+03  -3.201 0.001401 **
## waterfrontNo               -1.351e+05  1.579e+04  -8.553  < 2e-16 ***
## heatingHotwater            -4.337e+03  6.215e+03  -0.698 0.485372
## lgage                      -2.122e+04  3.337e+03  -6.359 2.77e-10 ***
## bathrooms                  -1.788e+03  6.733e+03  -0.266 0.790594
## newConstructionNo           5.120e+04  9.035e+03   5.666 1.78e-08 ***
## lotSize:landValue          -3.624e-01  8.052e-02  -4.501 7.33e-06 ***
## lotSize:livingArea         -1.346e+01  4.073e+00  -3.305 0.000976 ***
## lotSize:centralAirNo       -1.871e+04  6.164e+03  -3.035 0.002447 **
## landValue:heatingHotwater -1.959e-01  1.379e-01  -1.420 0.155820
## landValue:lgage            2.119e-01  3.906e-02   5.427 6.78e-08 ***
## livingArea:bathrooms       1.152e+01  3.097e+00   3.719 0.000208 ***
## centralAirNo:lgage         9.223e+03  3.215e+03   2.869 0.004182 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57430 on 1364 degrees of freedom
## Multiple R-squared:  0.6757, Adjusted R-squared:  0.6716
## F-statistic: 167.2 on 17 and 1364 DF,  p-value: < 2.2e-16
```

Finally, from the regression results above, we eliminate variables heatingHotwater and bathrooms, and get our dominate model.

$$Price_{estimate} = \beta_0 \ + \ \beta_1 lotSize \ + \ \beta_2 landValue \ + \ \beta_3 livingArea \ + \beta_4 rooms \ + \beta_5 centralAir +$$

$$\beta_6 waterfrontNo \ + \ \beta_7 lgage \ + \ \beta_8 newConstuction \ + \ \beta_9 lotSize : landValue +$$

$$\beta_{10} lotSize : livingArea \ + \ \beta_{11} lotSize : centralAir \ + \beta_{12} landValue : heatingHotwater \ +$$

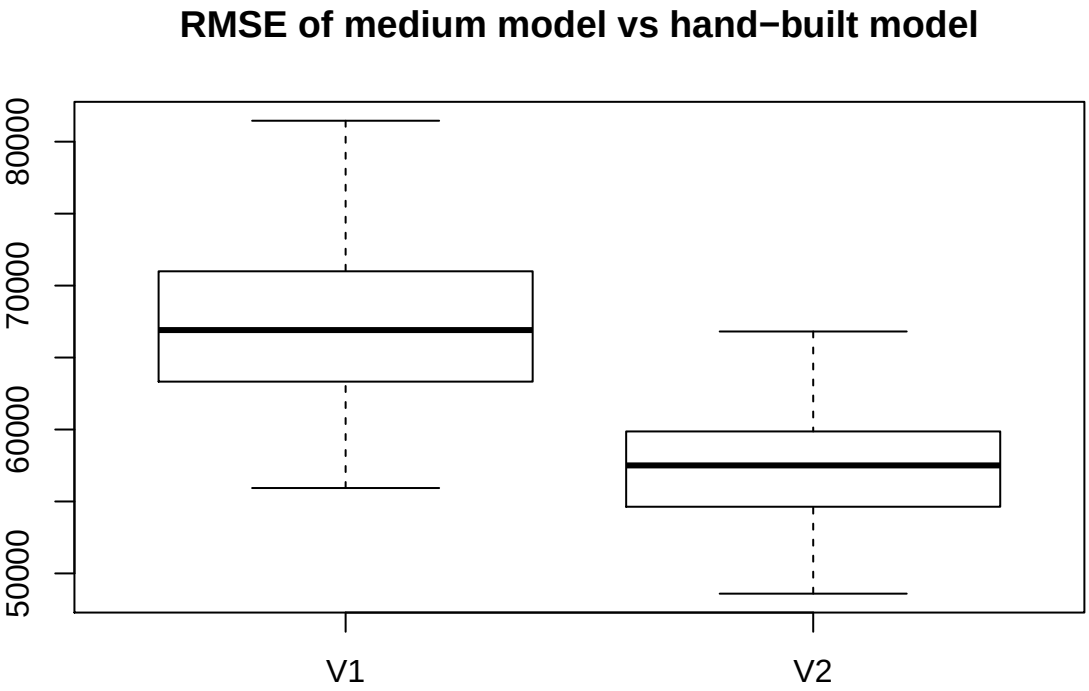
$$\beta_{13} landValue : lgage \ + \ \beta_{14} livingArea : bathrooms \ + \ \beta_{15} centralAir : lgage$$

```
##               (Intercept)                   lotSize                 landValue
```

```
##              1.766871e+05              5.912884e+04              5.890578e-01
##                livingArea                     rooms              centralAirNo
##              5.209899e+01              1.807386e+03             -3.193295e+04
##              waterfrontNo                     lgage         newConstructionNo
##             -1.350826e+05             -2.162476e+04              5.092245e+04
##         lotSize:landValue       lotSize:livingArea      lotSize:centralAirNo
##             -3.628189e-01             -1.337302e+01             -1.879277e+04
## landValue:heatingHotwater          landValue:lgage      livingArea:bathrooms
##             -2.640895e-01              2.197210e-01              1.073438e+01
##        centralAirNo:lgage
##              9.291778e+03
```

**Final model & Comparison**

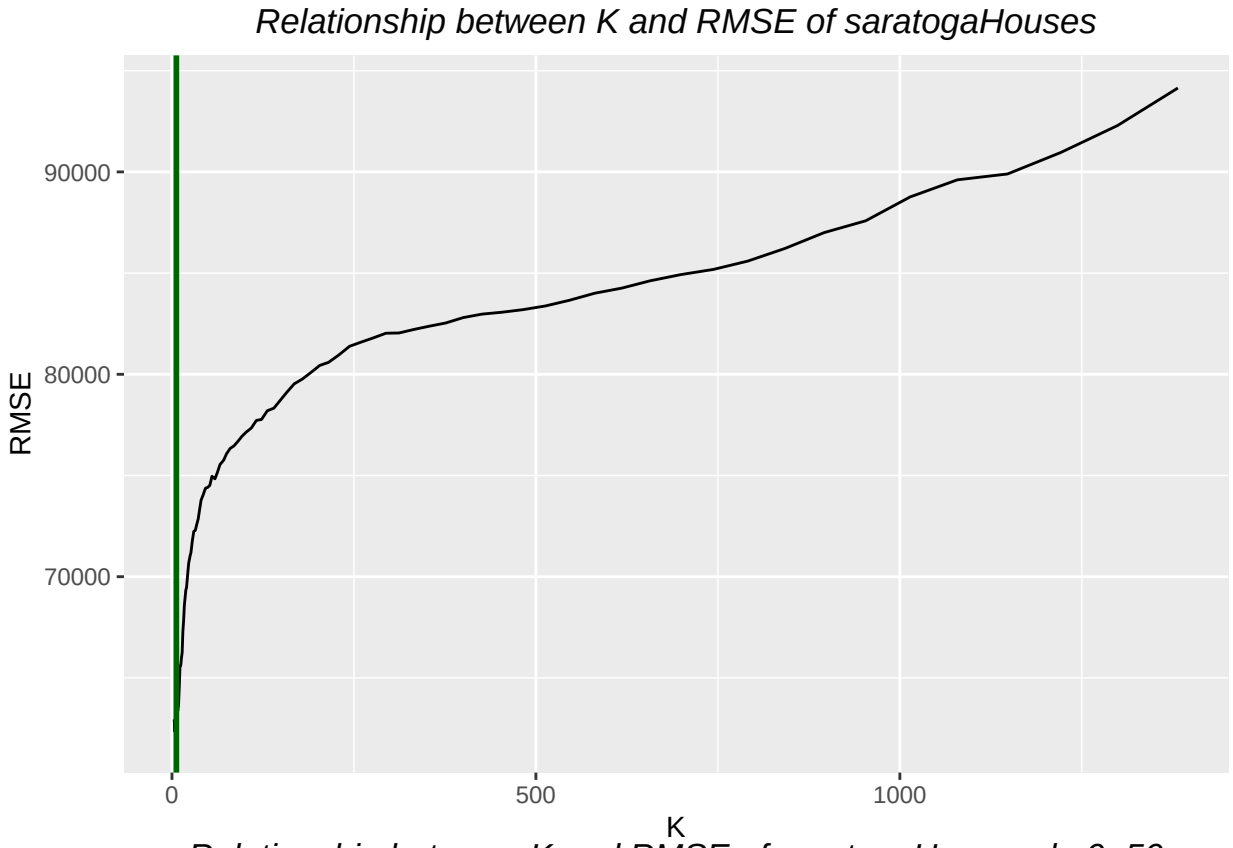
```
##       V1       V2
## 66981.90 57403.34
```




The boxplot above shows that our hand built model significantly outperforms the medium model.

## KNN model

In order to construct the best KNN model, we engage in multiple train-test splits and try to find the optimum k value.





From the graphs above, we find that when k = 6, the KNN model have the lowest root mean square error.

The calculation shows that on average, the lowest root mean square error for KNN model is $6.8557 \times 10^4$, which is higher than that for the hand-built model.

**Conclusions**

We have evaluated the available data regarding homes in the Saratoga area with a view to devising a model to value those homes for tax purposes. In doing so, we have focused explicitly on statistical evidence of relationships between home features and home price that appear in the available data set. Certain of those relationships are not intuitive, but if the available data is sufficiently robust, the model taken as a whole should provide a reasonable prediction of home value.

The data that we considered included 15 characteristics of homes, which we evaluated to assess predictive value with regard to the price of the home. Certain of those characteristics did not appear to be meaningfully predictive, and were therefore not included in the model. We also considered interactions between factors, and identified several that had predictive value, including an interaction involving one of the features that was not independently predictive. Predictive value was assessed based on the statistical significance of the predictive relationship at a 95% confidence interval. We identified eight characteristics that were individually predictive, and seven characteristic interactions that were predictive.

100 randomly sampled training/test splits were performed utilizing that model and compared to a similar number of training/test splits utilizing the medium model against which we bench-marked our analysis. We found that the mean squared error of the predictions utilizing the new model was materially lower than the benchmark error, as reflected in the boxplot. The root mean square error in our new model is $6.7133 \times 10^4$ while in the medium model, the root mean square error is $6.1212 \times 10^4$. That analysis suggests that the new model will be more effective in predicting the value of homes for tax purposes.

We also build KNN models that predicts house prices with documented features. By plotting the relationship between the out-of-sample root mean square errors and k values, we find that the best k value is 8 and the corresponding root mean square error is $6.8557 \times 10^4$, which is higher than the root mean square error of our hand-built linear model. The best-performing KNN model also, surprisingly, has a higher out-of-sample root mean square error than the medium model.

---

1.There were six such characteristics: the percentage of college graduates, the number of bedrooms, fireplaces and bathrooms, type of heating fuel and the presence of a sewer connection. This may appear surprising since he number of bedrooms and bathrooms are ordinarily thought of as value predictors, but two features that are included in the model, the living area and number of rooms, are surrogates for those qualities that, based on our analysis, appear to be a more reliable predictor of value.
2.The eight individual features are lot size, land value, living area, number of rooms, central air conditioning, whether the property is waterfront, the age of the building (considered on a logarithmic scale) and whether it is new construction. Seven interactions, five of which were interactions with either lot size or land value, were also included in the model on that basis. Another feature, the type of heating, was not predictive standing alone, but was predictive in combination with the land value of a property and included in the model on that basis.

# Predicting when articles go viral

In order to successfully predict whether an article would go viral, we attempt two approaches: the classification approach and the regression approach. In classification approach, we first classify the training sets into two categories with the 1400 threshold. Then, we establish a probability model for further predictions. In regression approach, we estimate the number of shares for each distinct article, and based on the predicted shares, classify the training sets. Both methodologies are expected to outperform the null model yet we

seek to find the best model for each approach and determine which approaches are more accurate in solving classification problems.

## Classification Approach

We first build a new dataset from online_news but taking away the "url" variables and adding a new variable, "viral", which defines whether the news is viral or not depending on the variable "shares" (1 is viral with shares over 1400, 0 is not viral with shares less than or equal to 1400). The new dataset is called "online_news1".

After performing train-test splits on the new dataset, we build baseline linear/logistic probability model with "viral" as dependent variable and all the remaining variables except "shares" as explanatory variables. Next, we apply the backward algorithm to select the variables that are better at predicting whether an article goes viral in both models and compare their in-sample performances.

### Logistic Probability Model

```
##    yhat
## y      0     1
##   0 14523  1648
##   1 11101  4443
```

```
## [1] 0.5980136
```

### Linear Probability Model

```
##    yhat
## y      0     1
##   0 10260  5911
##   1  5823  9721
```

```
## [1] 0.6300173
```

The in-sample performance of the best linear probability model is clearly better than the performance of the logistic probability model. The accuracy of the linear probability model is 0.63 while the accuracy of the logistic probability model is 0.598. Therefore, we choose the linear probability model for further calculations.

### Out-of-sample performance for linear probability model

```
##    yhat
## y      0    1
##   0 2533 1481
##   1 1472 1472
```

The above is the confusion matrix obtained from the average of 100 out-of-sample performances of linear probability model. From which we can report that: The true positive rate is 0.376. The false positive rate is 0.369. The false discovery rate is 0.377. The overall accuracy rate is 0.628. The overall error rate is 0.372.

**Comparison with the null model**

```
##
##     0     1
## 16074 15641
```

From the table, it's reasonable to assume that "not viral" is the more likely outcome. So a reasonable null model is the one that guesses "not viral" for every test-set instance.Then, we investigate the out-of-sample performance for the null model.

```
##
##    0    1
## 4008 3921
```

```
## [1] 0.5054862
```

In this particular train-test split. The accuracy of the model is 0.505. Our classification model returns an accuracy rate of 0.628. Its absolute improvement over the null model is approximately 12.212 percent. Its relative improvement, or lift over the null model is 1.242. Clearly, our classification model demonstrates significant improvements of accuracy compared to the baseline model.

## Regression Approach

**log(shares) vs shares**

We continue to use online_news1 as our dataset. However, instead of investigating variable "viral", when constructing baseline model, we use "shares"/"log(shares)" as our dependent variable and all the other variables except "viral" as explanatory variables. Next, we use backward algorithm to find the best linear model that predicts the shares/log transformation of shares with given features and compare the in-sample performance of the two models.

**shares model**

```
## [1] 0.4987861
```

$$\textit{The accuracy of the linear model for log(shares) is } 0.499$$

**log(shares) model**

```
## [1] 0.5864733
```

$$\textit{The accuracy of the linear model for log(shares) is } 0.586$$

Clearly, log(shares) model performs better at predicting whether an articles goes viral. Therefore, we employ log(shares) model in the following calculations and analysis.

**out-of-sample performance for the final regression model**

```
##    yhat
## y      0    1
##   0 2533 1481
##   1 1472 1472
```

The above is the confusion matrix obtained from the average of 100 out-of-sample performances of regression model. From which we can report that: The true positive rate is 0.143. The false positive rate is 0.675. The false discovery rate is 0.447. The overall accuracy rate is 0.587. The overall error rate is 0.413.

**Comparison with the null model**

```
##
##    0    1
## 4009 3920
```

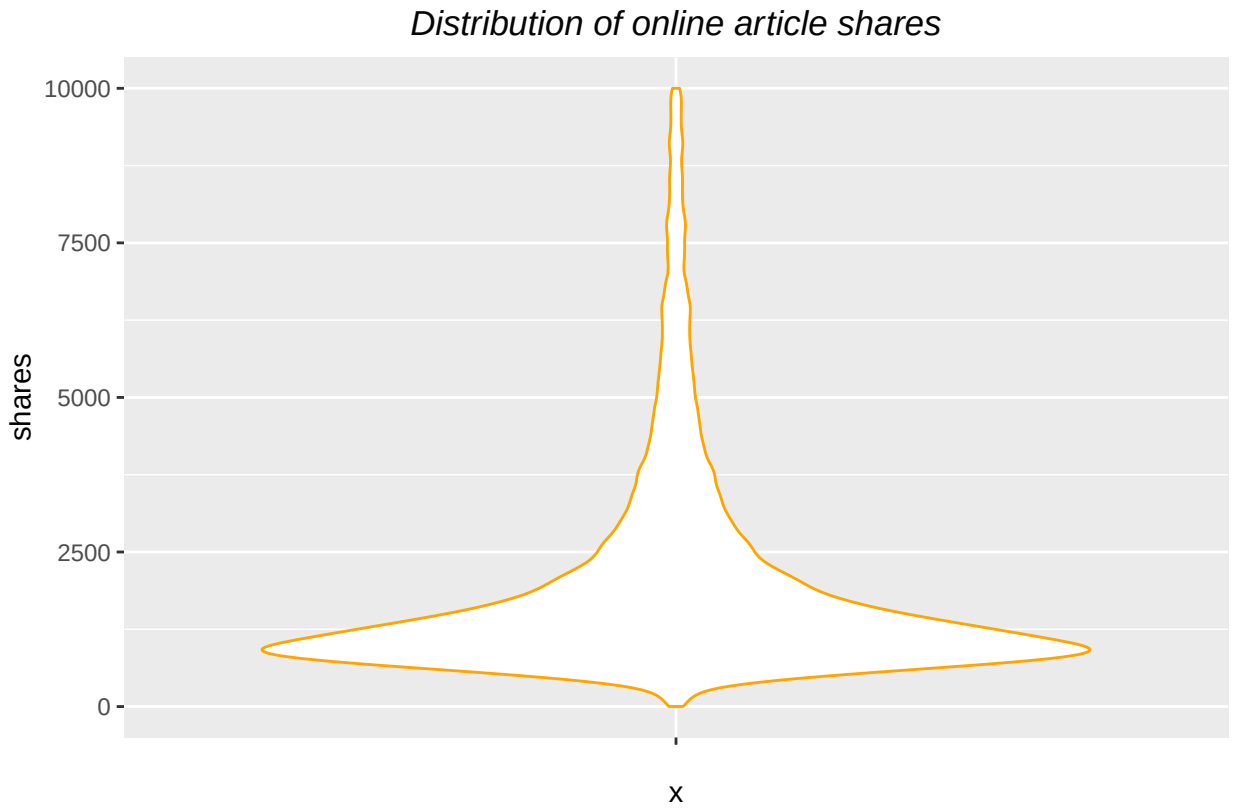
```
## [1] 0.5056123
```

We continue assume the null model classifies all articles as "not viral". Then, In this particular train-test split. The accuracy of the null model is 0.506. Our regression model returns an accuracy rate of 0.587. Its absolute improvement over the null model is approximately 8.164 percent. Its relative improvement, or lift over the null model is 1.161. The regression model shows a decent amount of improvements in accuracy over the null model.

**Conclusions**

In order to classify articles as "viral" and "not viral", we approach with two distinct methodologies - classification and regression. We determine that linear probability model performs better than the logistic probability model. We also discover that using log transformation of shares in our regression model improves its out-of-sample accuracy. Still, while both of our final regression and classification models demonstrate decent amount of improvements on accuracy over the null model, classification models do seem to outperform the regression model by around 5%.

The statistical intuition behind the disparity, as we suspect, is connected to the fundamental difference between the two models. Classification directly predicts a discrete class label whereas regression only predicts a continuous quantity. In classification, it doesn't matter if an article is shared over 1 million times or it's only shared 1401 times. They would all be classified as "viral". The classification method discards the informations that aren't useful in mere classification. But in regression models, the informations of how many shares each article has are retained, which disrupts the classification. With large outliers, the slope coefficients of explanatory variables are magnified. Consequently, the model tends to overestimate the exact number of shares. This leads to the high false positive rate as shown in previous sections and the reduced accuracy of the regression model predictions.




*Many articles have their online shares way above 1400, yet the exact numbers are not useful in classificatior*