Data Mining Exercise 2

Kylie Taylor

3/15/2019

Saratoga House Prices

The output below reflects work we have done to determine which variables included in the Saratoga Houses data contained in the 'mosaic' package in R. The first model that we ran was the medium length model from class that is the model we are trying to preform better than. This model estimates house prices using the variables "lotSize", "age", "livingArea", "pctCollege", "bedrooms", "bathrooms", "fireplaces", "rooms", "heating", "fuel", and "centralAir". The factors found to be most significant in estimating price are "lotSize", "livingArea", "bedrooms", "bathrooms", "rooms", and "centralAir".

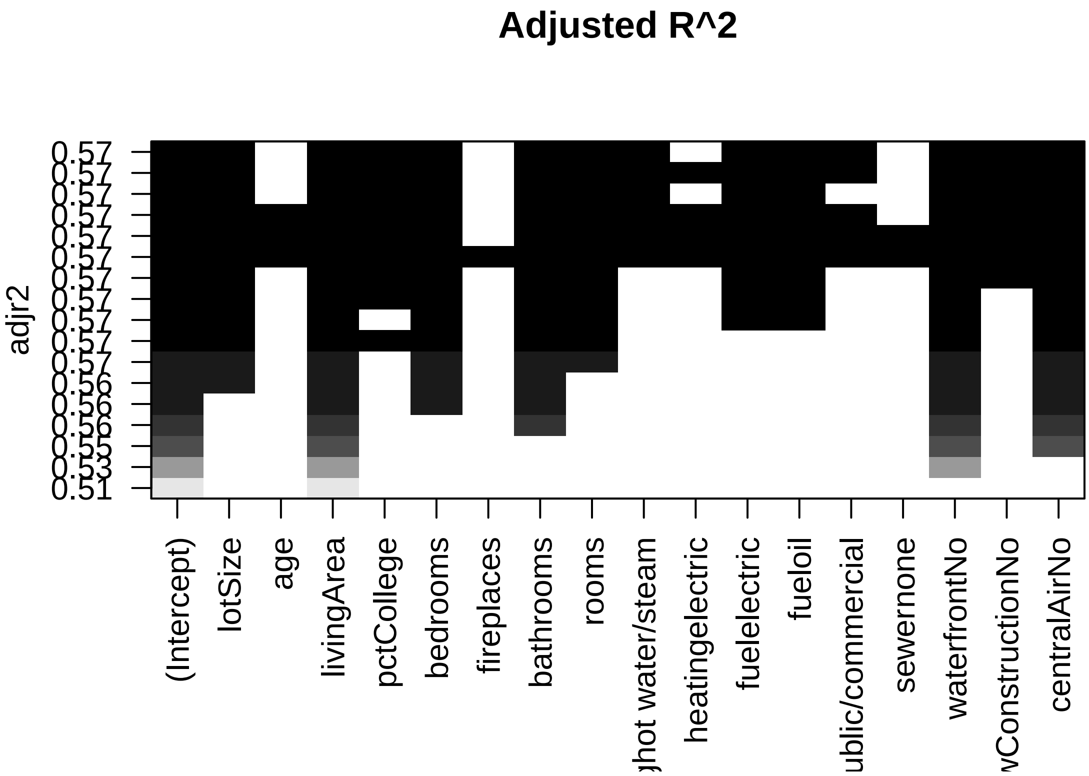
The output from the medium length model in class is below. We see that there is an $R^2 = 0.55$ and a RMSE of 66,767 (averaged over 1,000 sampled RMSE's).

```
##
## ==================================================
##                            Dependent variable:
##                        ---------------------------
##                                   price
## --------------------------------------------------
## lotSize                       9,554.533***
##                                (2,544.429)
##
## age                              21.648
##                                 (70.602)
##
## livingArea                      91.900***
##                                  (5.456)
##
## pctCollege                       142.357
##                                 (178.021)
##
## bedrooms                     -15,535.060***
##                                (3,104.743)
##
## fireplaces                      4,489.925
##                                (3,626.458)
##
## bathrooms                     24,143.050***
##                                (4,106.652)
##
## rooms                          2,821.262**
##                                (1,187.927)
##
## heatinghot water/steam         -3,847.374
##                                (5,164.614)
##
## heatingelectric                 5,358.532
##                               (16,050.490)
##
```

```
## fuelelectric                   -17,156.640
##                               (15,823.810)
##
## fueloil                        -6,126.069
##                                (5,830.788)
##
## centralAirNo                -19,116.380***
##                                (4,196.447)
##
## Constant                      34,027.180**
##                               (13,217.710)
##
## --------------------------------------------------
## Observations                       1,382
## R2                                 0.576
## Adjusted R2                        0.572
## Residual Std. Error      64,016.710 (df = 1368)
## F Statistic            143.162*** (df = 13; 1368)
## ==================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

```
##  result
## 66767.9
```

In an attempt to find a way to out-preform the model above, we used the package 'leaps' to run a best subsets on Saratoga houses. The best subsets algorithm cycles through all variables of interest in a data frame and combines them into separate regressions, then (in this case) compares the R^2 values of the many linear models that were generated by the algorithm. Below is a plot of the variables that best subsets suggests with the corresponding adjusted R^2. We see that if we include all the variables, the adjusted R^2 is about the same as if we were to leave a few superfluous variables, like "fireplace". A plot of adjusted R^2 on the y-axis and variables on the x-axis is below.



We created two models to compete with the medium model built in class, while also letting the best subsets inform us that there is no variable that should clearly be left out. Before running any models, we made the decision to remove the variables "landValue" from this analysis, as it is measuring almost the same thing as price. It would not be "fair" to include in any models, if our ultimate goal is to determine which factor explains house prices the best.

The first model we ran includes all the variables. The second model we ran includes all variables except "fireplace" and "sewer". The third and final model we tested was the same as the second plus interaction terms of "bedrooms*bathrooms", "age*heating" and "age*lotSize".

After averaging over 500 sampled RMSE's from the three models, we found that the second and the third model preform the best with equal RMSE's of 64,485.69.

```
##       V1       V2       V3       V4
## 66383.47 64669.74 64485.69 64485.69
```

We made the conclusion that the third model is the best, because the R^2 of the third model is the highest out of the three with an adjusted R^2 of 0.563. Therefore, this is the model we will be using for our analysis.

The linear model reveals that "lotSize", "livingArea", "bedrooms", "bathrooms", "rooms", "heating: water/steam", "waterFront", "newConstruction", "airCentral", "age*heating:water/steam", and "age*lotSize" are significant in estimating the value of a house in Saratoga.

```
lm.med <- lm(price ~ . - sewer - waterfront - landValue - newConstruction, data=saratoga_train)
lm <- lm(price ~. -landValue , data = saratoga_train)
lm1 <- lm(price ~  lotSize + age + pctCollege + livingArea + bedrooms + bathrooms + rooms + heating + fu
lm3 <- lm(price ~  lotSize + age + pctCollege + livingArea + bedrooms + bathrooms + rooms + heating + fu
stargazer::stargazer(lm, lm1, lm3, type = "text")
```

```
##
## ====================================================================================================
##                                                           Dependent variable:
##                            -------------------------------------------------------------------------
##                                                                  price
##                                       (1)                        (2)                        (3)
## ----------------------------------------------------------------------------------------------------
## lotSize                          11,096.190***               9,351.434***              19,153.640***
##                                   (2,664.271)                (2,547.569)                (3,422.286)
##
## age                                    21.152                     31.812                     1.205
##                                      (74.680)                   (74.366)                  (100.107)
##
## livingArea                            90.599***                  90.778***                  89.602***
##                                      (5.646)                    (5.497)                    (5.468)
##
## pctCollege                            359.585*                  428.263**                 487.775***
##                                     (188.346)                  (184.337)                  (182.844)
##
## bedrooms                         -14,339.490***             -14,899.030***             -11,629.180*
##                                   (3,202.119)                (3,192.762)                (6,600.502)
##
## fireplaces                            -322.839
##                                   (3,754.521)
##
## bathrooms                        21,101.200***              21,650.750***              27,587.590**
##                                   (4,279.775)                (4,245.668)               (11,369.360)
##
## rooms                             3,385.396***               3,354.934***              3,380.777***
##                                   (1,230.659)                (1,230.257)                (1,219.301)
##
## heatinghot water/steam           -12,108.100**              -11,811.890**              -36,207.400***
##                                   (5,341.608)                (5,327.365)                (7,939.717)
##
## heatingelectric                    -3,371.615                 -4,667.195                -12,079.240
##                                   (15,785.610)               (15,769.720)               (17,937.440)
##
## fuelelectric                      -18,843.770                -17,335.370                -19,410.730
##                                   (15,551.180)               (15,537.750)               (15,411.670)
##
## fueloil                           -11,593.690*              -16,015.130***             -11,502.340*
##                                   (6,224.651)                (5,902.767)                (5,896.753)
##
## sewerpublic/commercial            10,428.860**
##                                   (4,630.533)
##
## sewernone                          4,771.918
##                                   (22,451.840)
##
## waterfrontNo                    -151,261.400***            -152,258.300***            -154,706.400**
##                                   (20,330.630)               (20,334.280)               (20,388.270)
##
## newConstructionNo                 20,104.990**               18,938.040**              21,733.100**
##                                   (8,793.183)                (8,736.156)                (8,735.875)
##
## centralAirNo                     -16,060.220***             -16,750.350***             -14,959.090***
##                                   (4,408.626)                (4,380.722)                (4,368.289)
##
```

```
## bedrooms:bathrooms                                                                      -1,491.359
##                                                                                         (3,236.112)
##
## age:heatinghot water/steam                                                              591.569***
##                                                                                          (141.441)
##
## age:heatingelectric                                                                       433.324
##                                                                                          (371.211)
##
## lotSize:age                                                                             -358.340***
##                                                                                          (86.028)
##
## Constant                         150,664.200***             158,221.100***            142,979.700***
##                                   (25,233.750)               (24,989.290)               (31,221.130)
##
## ----------------------------------------------------------------------------------------------------
## Observations                         1,382                      1,382                     1,382
## R2                                   0.560                      0.558                     0.569
## Adjusted R2                          0.554                      0.553                     0.563
## Residual Std. Error         66,067.200 (df = 1364)     66,117.540 (df = 1367)   65,406.360 (df = 1
## F Statistic               101.967*** (df = 17; 1364) 123.266*** (df = 14; 1367) 99.853*** (df = 18;
## ====================================================================================================
## Note:                                                                           *p<0.1; **p<0.05; ***
```

After determining that the third model was the best, we ran a KNN regression on the same variables. In order to do this, we standardized the variables. This resulted in the RMSEs being on a much different scale (the z-scale to be exact) than the RMSEs from the linear models. We ran KNN's on 7 different K values; 5, 20, 50, 70, 100, 200, and 300.

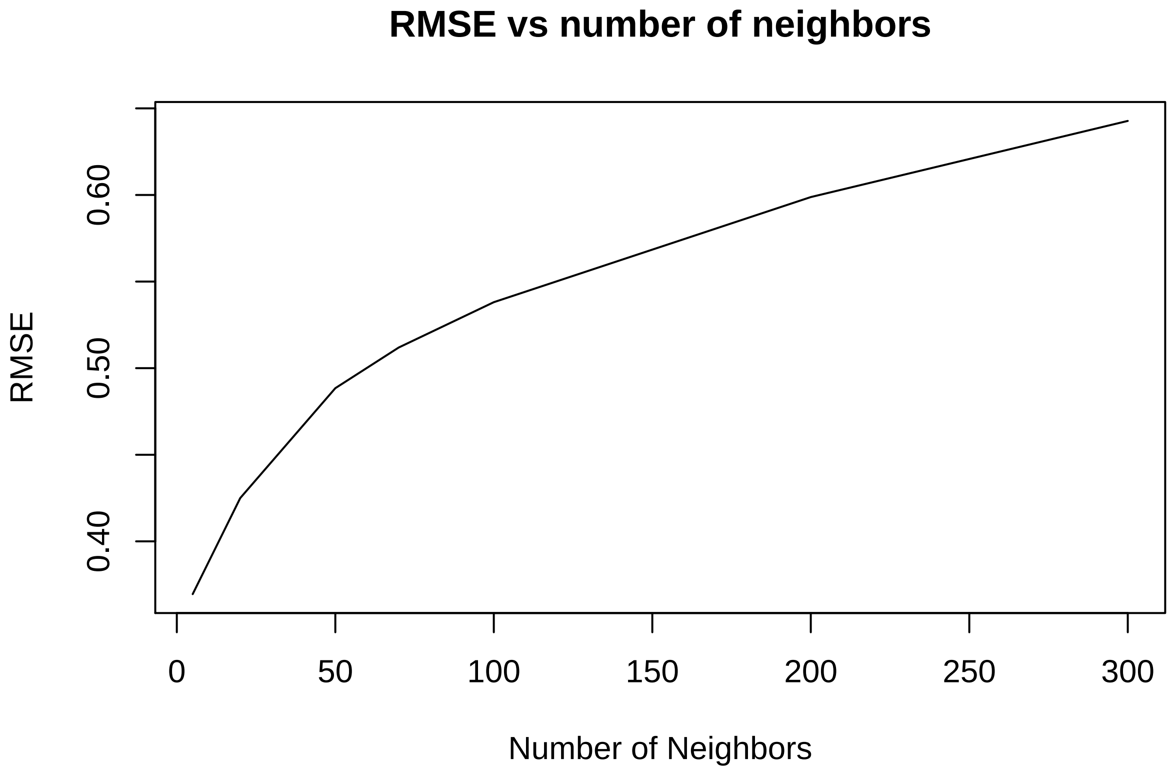
A KNN model with 5 nearest neighbors appears to have the smallest RMSE out of the 7 models tested. This is an interesting trade off, because a low K results in low variance, but high bias.

One way to compare if the KNN model does better than the linear model is to standardize the variables in the linear models and compare the RMSE's between the two methods. The standardized RMSE of the chosen linear model is 0.6536, this is significantly bigger than the RMSE for a KNN with 5 nearest neighbors of 0.3692. For this reason, we know that KNN is better preforming and therefore should be used.

```
## Warning: package 'psycho' was built under R version 3.5.2
```

```
## Warning: package 'FNN' was built under R version 3.5.2
```

```
##        V1        V2        V3        V4        V5        V6        V7
## 0.3695293 0.4249874 0.4884304 0.5119079 0.5380951 0.5987807 0.6427443
```



```
rmse_vals.std = do(500)*{
  n = nrow(SH)
  n_train = round(0.8*n)  # round to nearest integer
  n_test = n - n_train
  train_cases = sample.int(n, n_train, replace=FALSE)
  test_cases = setdiff(1:n, train_cases)
  SH_train = SH[train_cases,]
  SH_test = SH[test_cases,]
  lm3 <- lm(price ~  lotSize + age + pctCollege + livingArea + bedrooms + bathrooms + rooms + heating +
  yhat_test3 = predict(lm3, SH_test)
  rmse(SH_test$price, yhat_test3)
}

colMeans(rmse_vals.std)
```

```
##    result
## 0.6537636
```

To conclude what these models have told us, the price of a house is best determined by lot size, age, proximity to a college, size of living area, number of bedrooms, number of bathrooms, number of rooms, if there is water/steam heating, type of fuel to the house, if its waterfront, if its new construction, if there is central air, the number of bedrooms times the number of bathrooms, age times if there is water/steam heating, and age times the lot size. We suggest that the local taxing authority should use a KNN model in contrast to a linear model to predict the prices of homes in Saratoga, NY, as a KNN model has the lowest out of sample RMSEs. To be specific, our analysis found that a KNN with 5 nearest neighbors has the lowest RMSE out of all the KNN models we tested.

Hospital Audit

This problem asks us to examine the performance of radiologists at a hospital in Seattle, WA. The data contains 790 observations of patients with variables accounting for 5 different radiologists, whether the patient got cancer, and various other risk factors. The first step we took was to run some summary statistics on the data frame.

We found that approximately 3.75% of all patients were diagnosed with breast cancer, there was about a 15% recall rate overall, about 17.63% of all patients have family history of breast cancer, and about 4.8% of all patients have breast cancer symptoms. We also identified that the most common age of patients was between 40 and 49 years old, and the most common type of breast density is density 3, or heterogeneously dense, across all patients in this study.

Table 1: Table continues below

radiologist	cancer	recall	age
radiologist13:198	Min. :0.00000	Min. :0.0000	age4049 :287
radiologist34:197	1st Qu.:0.00000	1st Qu.:0.0000	age5059 :284
radiologist66:198	Median :0.00000	Median :0.0000	age6069 :199
radiologist89:197	Mean :0.03749	Mean :0.1499	age70plus:217
radiologist95:197	3rd Qu.:0.00000	3rd Qu.:0.0000	NA
NA	Max. :1.00000	Max. :1.0000	NA

history	symptoms	menopause	density
Min. :0.0000	Min. :0.00000	postmenoHT :321	density1: 89
1st Qu.:0.0000	1st Qu.:0.00000	postmenoNoHT :360	density2:332
Median :0.0000	Median :0.00000	postmenounknown: 35	density3:460
Mean :0.1763	Mean :0.04863	premeno :271	density4:106
3rd Qu.:0.0000	3rd Qu.:0.00000	NA	NA
Max. :1.0000	Max. :1.00000	NA	NA

We then created a confusion matrix displaying the cancer outcomes and recall decisions of all radiologists. We found that, as a group, the radiologists accurately make recall decisions about their patients in 85.7% of the instances recorded.

```
##       cancer
## recall   0   1
##      0 824  15
##      1 126  22
```

```
##
## Accuracy
## =====
## 0.857
## -----
```

Our ultimate goal is not to find the overall rates of these parameters, rather how the 5 radiologists compare across all the parameters listed above. In this problem, our variable of interest is “recall”, since the radiologist has little control over whether their patient is diagnosed with cancer, they do have control over if the patient is recalled (to hopefully catch cancer early on). Essentially, we are trying to determine which radiologist is best at their job.

The plot below reveals conditional probabilities of recall for each radiologist, given if the patient is diagnosed with cancer or not. We find that radiologists 89 and 95 have the highest rates of recall, given their patients got diagnosed with cancer, about 71.42%. The radiologist with the lowest recall given their patients' positive diagnosis were from radiologists 13 and 66, with a true positive rate of 50% (a coin flip).

```
##      Radiologist Cancer     Recall
## 1  radiologist13      0 0.13157895
## 2  radiologist34      0 0.06842105
## 3  radiologist66      0 0.17368421
## 4  radiologist89      0 0.17368421
## 5  radiologist95      0 0.11578947
## 6  radiologist13      1 0.50000000
## 7  radiologist34      1 0.57142857
## 8  radiologist66      1 0.50000000
## 9  radiologist89      1 0.71428571
## 10 radiologist95      1 0.71428571
```

Now to address the question at hand: are some radiologists more clinically conservative than others in recalling patients, holding patient risk factors equal?

Our tactic is to create a logistic model with recall as our outcome variable. We use a logistic model because our outcome variable, recall, is binary. To account for patient risk factors, we included all the factors in the data set as regressors in our model. We created a second model that included an interaction term for the estimated recall rate of each radiologist, given that a patient has breast cancer. This coefficient along with the coefficient on the radiologist variable should give us a good idea of which radiologist out of the 5 is the most conservative.

Output from the two models is below. We see that most significant determinant in recall is if the patient has cancer (intuitively). We also see that the model without the interaction term has a lower AIC than the one with the interaction term.

```
## 
## ============================================================
##                                      Dependent variable:    
##                                  ----------------------------
##                                             recall          
##                                       (1)            (2)    
## ------------------------------------------------------------
## radiologistradiologist34            -0.538        -0.634*   
##                                    (0.339)        (0.362)   
##                                                             
## radiologistradiologist66            0.401          0.442    
##                                    (0.287)        (0.297)   
##                                                             
## radiologistradiologist89            0.477*         0.438    
##                                    (0.289)        (0.298)   
##                                                             
## radiologistradiologist95            -0.032        -0.089    
##                                    (0.304)        (0.317)   
##                                                             
## cancer                             2.335***      2.043***   
##                                    (0.368)        (0.761)   
##                                                             
## ageage5059                          0.054          0.030    
##                                    (0.304)        (0.306)   
##                                                             
## ageage6069                          0.111          0.111    
##                                    (0.373)        (0.373)
##
## ageage70plus                        -0.084        -0.122
##                                    (0.382)        (0.388)
##
## history                             0.194          0.159
##                                    (0.241)        (0.245)
##
## symptoms                           0.728**        0.689*
##                                    (0.368)        (0.376)
##
## menopausepostmenoNoHT               -0.168        -0.192
##                                    (0.245)        (0.246)
##
## menopausepostmenounknown            0.264          0.181
##                                    (0.490)        (0.504)
##
## menopausepremeno                    0.333          0.302
##                                    (0.322)        (0.324)
##
## densitydensity2                    1.120**        1.143**
##                                    (0.542)        (0.545)
##
## densitydensity3                    1.301**        1.324**
##                                    (0.539)        (0.542)
##
## densitydensity4                     0.689          0.680
##                                    (0.612)        (0.616)
##
## radiologistradiologist34:cancer                     0.969
##                                                   (1.144)
##
## radiologistradiologist66:cancer                    -0.532
##                                                   (1.068)
##
## radiologistradiologist89:cancer                     0.507
##                                                   (1.165)
##
## radiologistradiologist95:cancer                     0.878
##                                                   (1.167)
##
## Constant                          -3.244***      -3.199***
##                                    (0.649)        (0.652)
##
## ------------------------------------------------------------
## Observations                         987            987
## Log Likelihood                    -379.882       -378.652
## Akaike Inf. Crit.                  793.765        799.305
## ============================================================
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```

To interpret the results from the glm, we take the exponent of all the coefficients of our model. The results reveal that radiologist 66 and radiologist 89 have 1.49 and 1.61 higher odds, respectively, of recalling patients than radiologist 13. Alternatively, we found that radiologist 34 and radiologist 95 have 0.58 and 0.96 higher odds, respectively, than radiologist 13 (low odds of recall! since odds < 1).

This means that radiologists 66 and 89 are more clinically conservative than radiologist 13, 34 and 95 when recalling patients, since they have a higher probability of recalling patients, thus recalling more patients.

Table 3: Table continues below

(Intercept)	radiologistradiologist34	radiologistradiologist66
0.03902	0.5837	1.494

Table 4: Table continues below

radiologistradiologist89	radiologistradiologist95	cancer	ageage5059
1.611	0.9683	10.33	1.056

Table 5: Table continues below

ageage6069	ageage70plus	history	symptoms	menopausepostmenoNoHT
1.118	0.9197	1.214	2.071	0.8451

Table 6: Table continues below

menopausepostmenounknown	menopausepremeno	densitydensity2
1.302	1.395	3.065

densitydensity3	densitydensity4
3.671	1.991

The next question of interest is when the radiologist is at the clinic, should they be weighting some clinical risk factors more heavily than they currently are?

For this question we are also dealing with a classification model, with a binary outcome variable, which means we will be using a logistic regression. Instead of wondering how individual radiologists performed, we are wondering if patient risk factors play a stronger role in determining cancer than radiologists thought. To do this, we compared a model with cancer as the outcome variable and recall as the explanatory variable to a model with cancer as the outcome variable and recall and all other risk factors as the explanatory variables. To determine if the radiologists are neglecting to account for patient's risk factors, we look at the fits of the models and which models estimate cancer the best.

The models reveal that the model that only regresses cancer on recall has a lower AIC than the model with all the risk factors included. This means that radiologists recalls are doing a relatively better job at estimating cancer diagnosis, therefore the data suggests that they do not need to be weighting some clinical factors more than they currently are.

```
##
## =======================================================
##                               Dependent variable:
##                          -----------------------------
##                            recall         cancer
##                              (1)       (2)       (3)
## ------------------------------------------------------
## radiologistradiologist34  -0.538               0.019
##                           (0.339)             (0.564)
##
## radiologistradiologist66   0.401              -0.370
##                           (0.287)             (0.541)
##
## radiologistradiologist89  0.477*              -0.233
##                           (0.289)             (0.570)
##
## radiologistradiologist95  -0.032              -0.385
##                           (0.304)             (0.578)
##
## cancer                   2.335***
##                           (0.368)
##
## ageage5059                 0.054               0.478
##                           (0.304)             (0.639)
##
## ageage6069                 0.111               0.398
##                           (0.373)             (0.813)
##
## ageage70plus              -0.084              1.436*
##                           (0.382)             (0.737)
##
## history                    0.194               0.247
##                           (0.241)             (0.439)
##
## symptoms                  0.728**             -0.008
##                           (0.368)             (0.716)
##
## menopausepostmenoNoHT     -0.168              -0.173
##                           (0.245)             (0.456)
##
## menopausepostmenounknown   0.264               0.820
##                           (0.490)             (0.728)
##
## menopausepremeno           0.333               0.230
##                           (0.322)             (0.662)
##
## densitydensity2           1.120**              0.718
##                           (0.542)             (1.080)
##
## densitydensity3           1.301**              0.835
##                           (0.539)             (1.081)
##
## densitydensity4            0.689              1.998*
##                           (0.612)             (1.134)
##
## recall                             2.261***  2.336***
##                                     (0.348)   (0.369)
##
```

```
## Constant                 -3.244*** -4.006*** -5.475***
##                           (0.649)   (0.261)   (1.309)
##
## ------------------------------------------------------
## Observations                987       987       987
## Log Likelihood           -379.882  -137.440  -130.133
## Akaike Inf. Crit.         793.765   278.881   294.266
## ======================================================
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

To summarize our findings, we identified that there are two radiologists that tend to be more conservative when recalling patients than other radiologists. To be exact, radiologists 66 and 89 have higher probabilities than radiologists 13, 34 and 95 of recalling patients. We also identified radiologist 89 to have on of the highest true positive rates of recall given cancer at a rate of approximately 71%. If the data were to recommend a radiologist based solely off recall rates, it would recommend radiologist 89 to do the patients' screenings. After determining how the radiologist preformed among themselves, we looked into if the radiologists might be ignoring some helpful information given by patients' risk factors. It was determined that the radiologists actually did a better job of diagnosing breast cancer when not accounting patients' risk factors. This conclusion was made because none of the variables in the model accounting for risk factors displayed statistical significance in determining cancer diagnosis, and the AIC of the cancer on recall model had a much lower AIC score. This information reveals that the radiologist's process for making a recall decision might actually be better if they do not look at the patient's risk factors. A possible reason for this is because a patient's risk factors may sway the radiologists decision to recall the patient, likely in the case that they will not recall them because their risks appear to be lower.

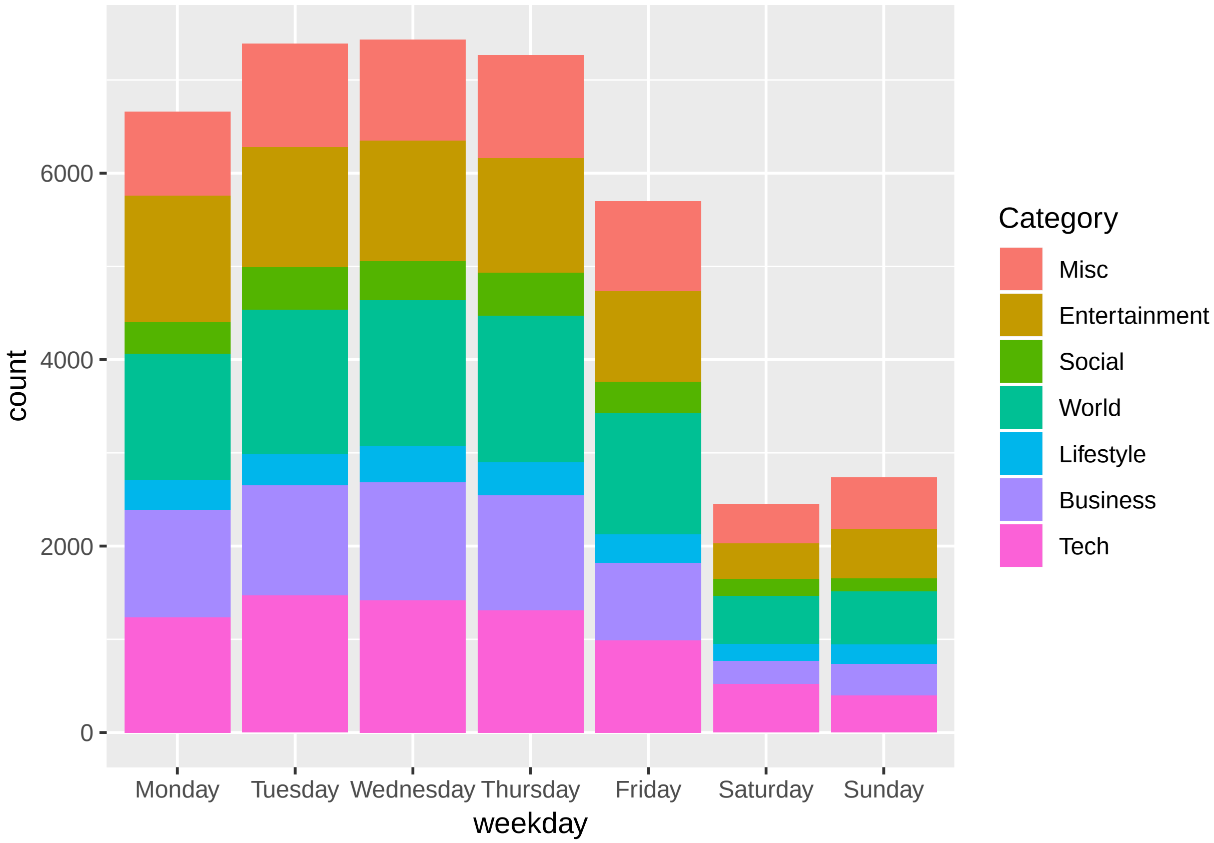
Viral Articles

The data set used in the analysis describes ~39,640 observations of articles published on Mashable.com's website. In the original data set there are 37 predictor variables that describe each article.

We added to the initial data set by generating dummy variables indicating the specific category of article into a single categorical variable column. We repeated this same transformation to indicate the day of the week each article was published.

To begin building our models, we decided to plot the data points in various visualizations to find any striking relationships across the data. Below are a few of the most interesting plots and our findings from them.

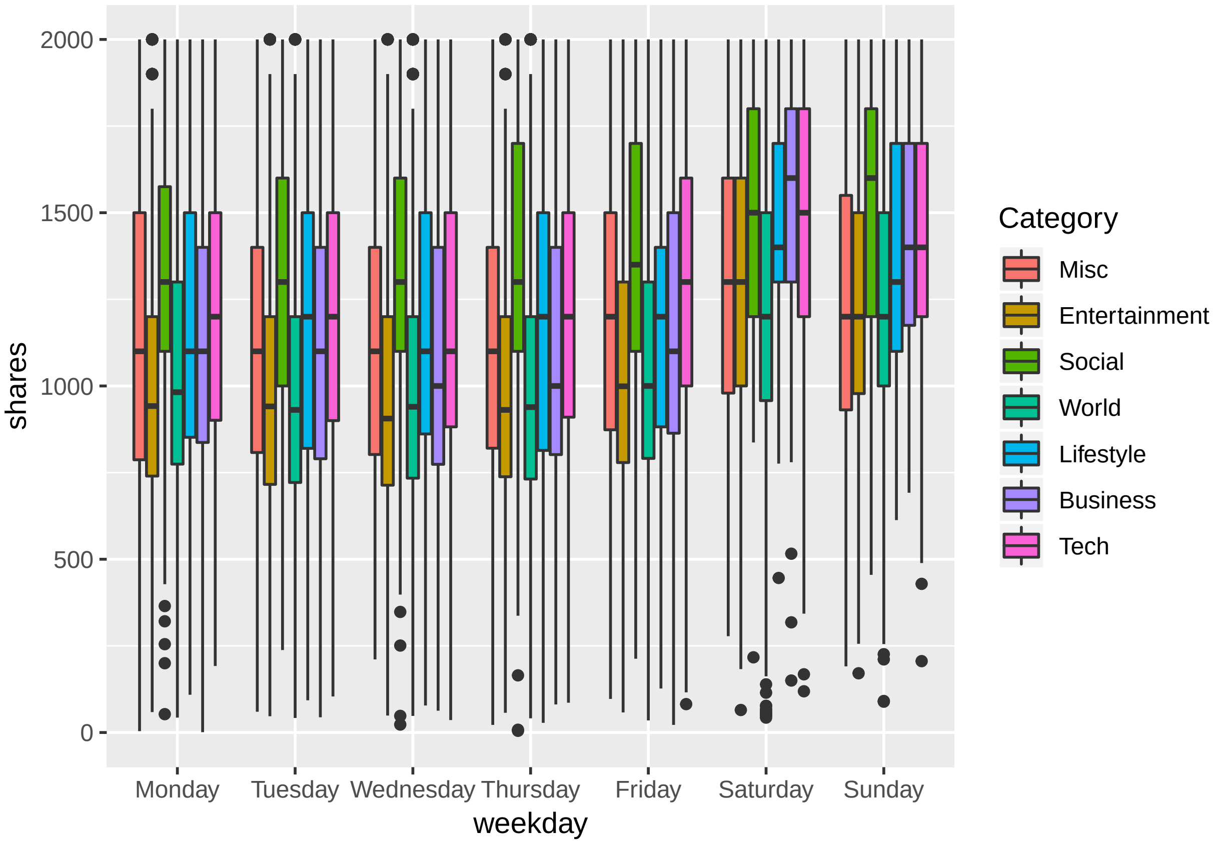
The first plot we created was a simple bar chart of number of shares of articles published on certain days of the week, with the color denoting the category of article.



Here we learned that our data set contained mostly articles that were published on Tuesday, Wednesday, or Thursday, and that were classified under the general category of World, followed by Technology.

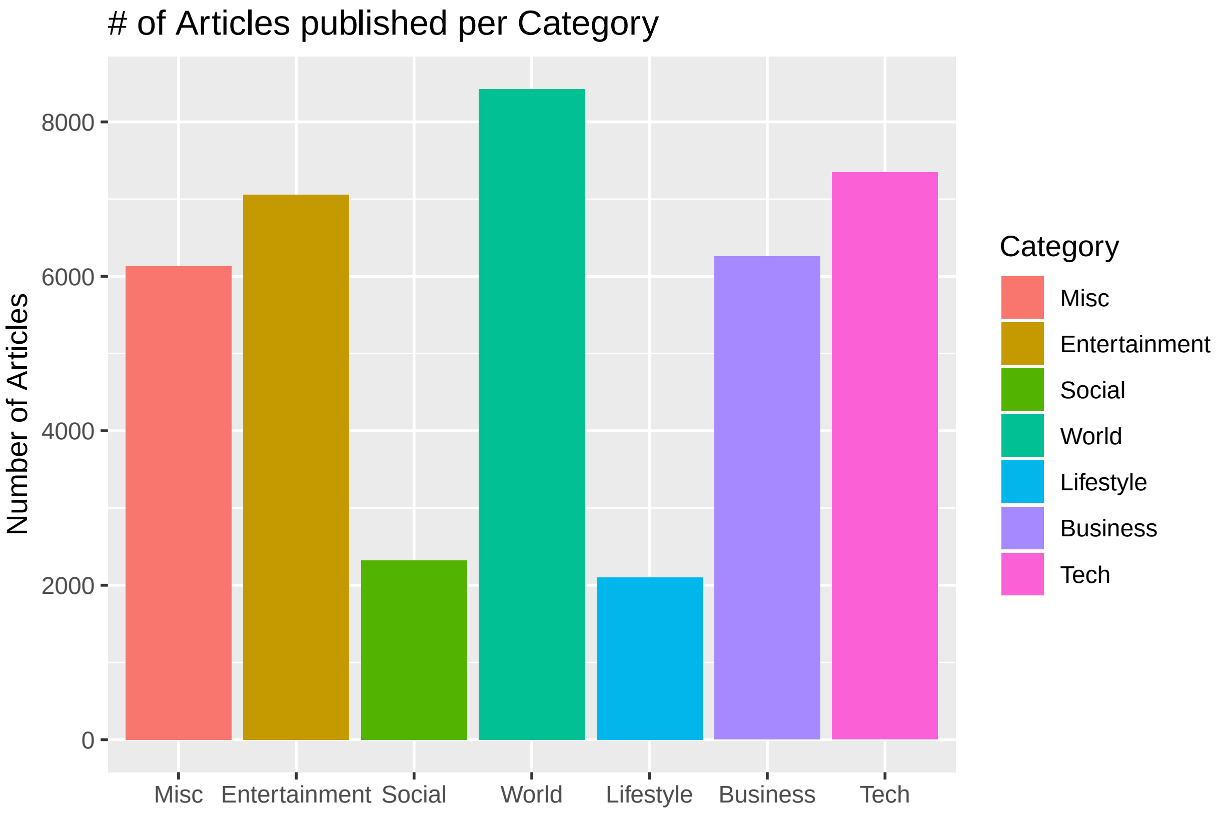
Given this information, we decided to create a box plot to explore the relationship between the number of shares across the days of the week for each particular type of category.

```
## Warning: Removed 13616 rows containing non-finite values (stat_boxplot).
```



Through this plot we found that across the week, shares of articles from all different types of sources is higher on the weekend. This led us to include "is_weekend" in our model.

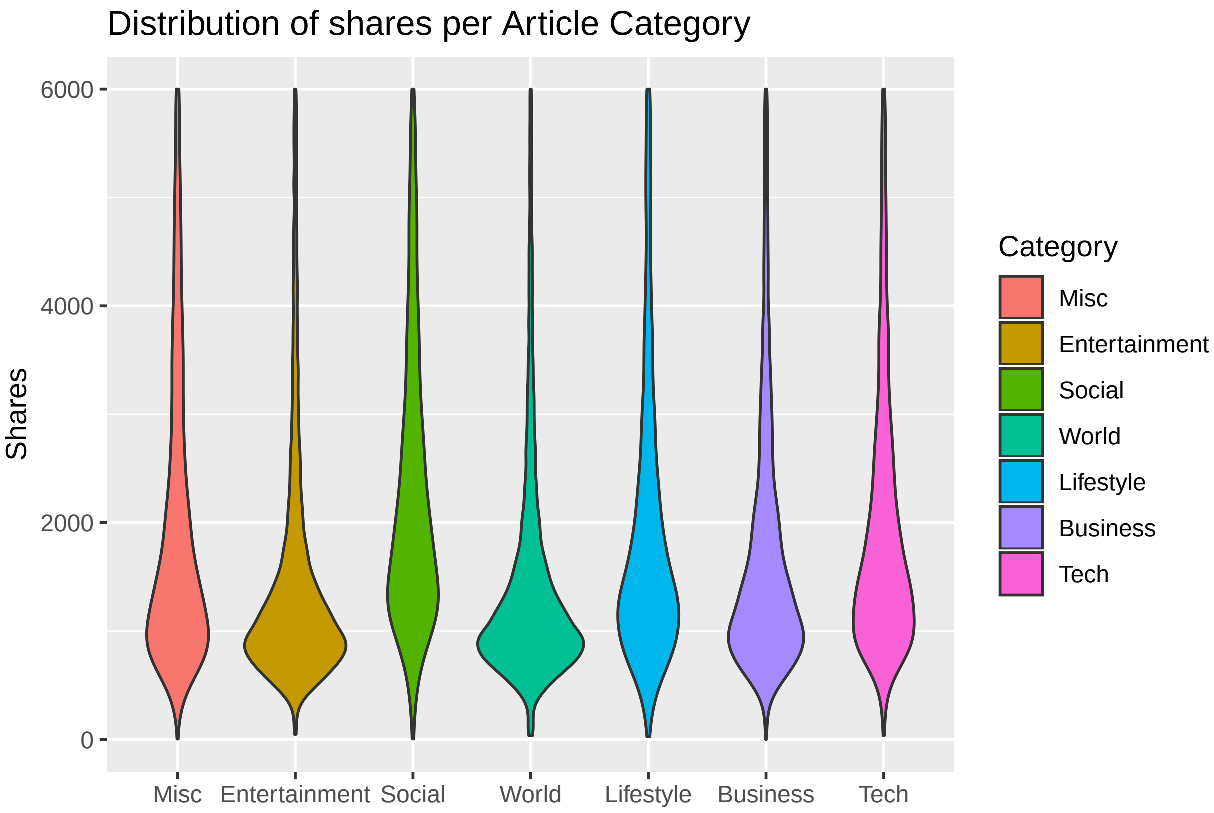
The next piece we were interested in was how many articles were actually published by each type of channel, or category of article. Below is a simple bar plot of number of articles by the different articles, where we find that "World" and "Tech" channels have published the most amount of articles.



Naturally following, we were interested in the distribution of those articles that were shared across all the channels.

We created the violin plots below and found that articles in the data set that were categorized as Social Media tended to have higher numbers of shares. The next categories that tended to demonstrate higher number of shares were Lifestyle and Tech articles.

```
## Warning: Removed 4103 rows containing non-finite values (stat_ydensity).
```



Given the information we gained visualizations, we hand-built a simple linear model including a mix of variables that were used to predict number of shares. The variables we chose to include were the following: number of hyperlinks, whether or not the article was published on the weekend, global rate of negative words, whether the article was Business, self reference average shares, whether the article was entertainment, number of keywords, and average negative polarity. Our visualizations showed a trend of higher shares on the weekend and seems to be very significant in our model. The particular type of article is a very deterministic part of predicting if an article will go viral as certain types of categories have higher shares. We tested this model against a few other possible specifications and found that the model described above is the strongest.

To begin, we start with a null model, which makes the prediction that every single article is viral. With this model, we will correctly predict that an article will go viral with a 49.34% success rate.

Our goal with this study is to preform better than just simply guessing that the article always goes viral. The first model we generated was a linear regression with shares as the outcome variable, and the factors listed above as the explanatory variables.

Testing Performance for best Linear Model (lm2)

Next, we averaged over 500 train-test splits and found that the in sample performance of the model has an RMSE of 11,685, a true positive rate of 0.994, a false positive rate of 0.991, a success rate of 49.43%.

```
##   result
## 11684.11
```

```
##    yhat
## y       0     1
##   0   133 15929
##   1   112 15541
##    result
## 0.4943072
```

The out of sample performance had a RMSE of 11,686, true positive rate of 0.993, a false positive rate of 0.994, and the same success rate of 49.43%. Both having a lift of 0.09% from the null.

```
##   result
## 11683.33
```

```
##    yhat
## y      0    1
##   0   22 4002
##   1   18 3887
```

```
##    result
## 0.4942671
```

We can conclude that the linear model for shares does not preform much better than the null.

Our next portion of the assignment was to determine if thresholding the shares so that any articles with shares over 1,400 would be considered viral and then running our analysis would render better results. In order to model for the binomial outcome variable "Viral", we need to use a logistic model. We stuck with the same variables in our model so we can compare to our linear model's Performance equally.

We proceeded to assess the performance of this model by averaging over 500 train/test splits.

Question 3 Part II: Classification Model using GLM Binomial

The logistic model has an in sample RMSE of 0.945, a true positive rate of 0.224, a false positive rate of 0.087, and a success rate of 0.5695, or 56.95%.

```
##    result
## 0.9445996
```

```
##    yhat
## y       0     1
##   0 14595  1385
##   1 12211  3524
```

```
##    result
## 0.5694197
```

The logistic model has an out of sample RMSE of 0.946, a true positive rate of 0.218, a false positive rate of 0.077, and a success rate of 0.5694, or 56.94%.

```
##    result
## 0.9459296
```

```
##    yhat
## y      0    1
##   0 3681  309
##   1 3082  857
```

```
##    result
## 0.5694297
```

This means that the logistic model has a lift of about 7.6% from the null. This is much better than the linear model. The logistic model appears to be more conservative with handing out predictions of viral than the linear model of shares, and the results appear to be slightly better because of it.

Based off of the data, the approach of thresholding first, then regress/classify second performs better than the alternative. This is likely because the linear model was likely being pulled and influenced by articles that had very extreme amounts of shares. Once we turned the variable shares into a binary variable, the magnitudes of those extreme observations of shares do not influence the model, since they now carry the same weight as an article that has barely over 1,400 shares.

Overall, our ability to predict if an article goes viral is very weak, barely better than simply guessing with a 50% chance at its best. We believe that this is due to the nature of why articles get shared and go viral. For an article to go viral is truly a random process, because at it's root, there is a person reading the article deciding if they like it enough to share it, which is inherently random.