Homework 2

Frank Chou, Milo Opdahl, Tejaswi Pukkalla

March 12, 2019

Question 1: Building a Better Model

Let's first look at the summary statistics of the data before going into developing regression models.

```
##      price           lotSize             age           landValue     
##  Min.   :  5000   Min.   : 0.0000   Min.   :  0.00   Min.   :   200  
##  1st Qu.:145000   1st Qu.: 0.1700   1st Qu.: 13.00   1st Qu.: 15100  
##  Median :189900   Median : 0.3700   Median : 19.00   Median : 25000  
##  Mean   :211967   Mean   : 0.5002   Mean   : 27.92   Mean   : 34557  
##  3rd Qu.:259000   3rd Qu.: 0.5400   3rd Qu.: 34.00   3rd Qu.: 40200  
##  Max.   :775000   Max.   :12.2000   Max.   :225.00   Max.   :412600  
##    livingArea     pctCollege       bedrooms       fireplaces    
##  Min.   : 616   Min.   :20.00   Min.   :1.000   Min.   :0.0000  
##  1st Qu.:1300   1st Qu.:52.00   1st Qu.:3.000   1st Qu.:0.0000  
##  Median :1634   Median :57.00   Median :3.000   Median :1.0000  
##  Mean   :1755   Mean   :55.57   Mean   :3.155   Mean   :0.6019  
##  3rd Qu.:2138   3rd Qu.:64.00   3rd Qu.:4.000   3rd Qu.:1.0000  
##  Max.   :5228   Max.   :82.00   Max.   :7.000   Max.   :4.0000  
##    bathrooms       rooms                    heating           fuel     
##  Min.   :0.0   Min.   : 2.000   hot air        :1121   gas     :1197  
##  1st Qu.:1.5   1st Qu.: 5.000   hot water/steam: 302   electric: 315  
##  Median :2.0   Median : 7.000   electric       : 305   oil     : 216  
##  Mean   :1.9   Mean   : 7.042                                         
##  3rd Qu.:2.5   3rd Qu.: 8.250                                         
##  Max.   :4.5   Max.   :12.000                                         
##                sewer      waterfront newConstruction centralAir
##  septic           : 503   Yes:  15   Yes:  81        Yes: 635  
##  public/commercial:1213   No :1713   No :1647        No :1093  
##  none             :  12                                        
##                                                                
##                                                                
## 
```

The average price of houses in Saratoga, NY is around $200,000. On average, houses are around 28 years old, with hardly 80 new constructions (out of more than 1500), with an estimated living area of 1755 sq feet, 3 bedrooms and 2 bathrooms.About 65% of these houses have hot air heating, 70% of them use gas fuel, about 63% of them do not use centrail air conditioning.Only 15 of these houses are waterfront properties.

We start by first building a baseline model using just bedrooms, bathrooms and Airconditioning and Waterfront property as variables and look at the regression coefficients.

```
##  (Intercept)     bedrooms centralAirNo    bathrooms waterfrontNo 
##    213283.89     23131.94    -33334.01     67795.61   -183620.03
```

We see that while increase in bedrooms and bathrooms increases the price, not having central air conditiong o=and not being waterfron tproperty drives down the price. This howeever, isn't very efficient. So then we add additonal effects such as land value, living area, age of the building, sewage type, fuel type, lot size and if the building is a new construction or not and look at the new regression coefficients.

```
##            (Intercept)                lotSize               bedrooms
```

```
##              217546.952             13850.891              1318.706
##              fireplaces             bathrooms                 rooms
##               14310.320             51277.008             11760.076
## heatinghot water/steam        heatingelectric           fuelelectric
##               -8486.597             -1423.725            -18190.593
##                 fueloil          waterfrontNo      newConstructionNo
##              -14174.595           -185623.747               751.993
##            centralAirNo
##              -24361.152
```

While this gives us a better picture, we still might be missing any interaction between these variables that might lead to conflation of estimator coefficients. Hence, we next run the regression taking into account the interactions possible. After that, we compare out of sample predictions to see how effective our regression model is and then calculate the average root mean square errors for these three regresssions.

```
## [1] 74156.68
```

```
## [1] 70366.34
```

```
## [1] 87183.3
```

Now let's improve the model by adding multiple interaction terms and repeating the regression a hundred times and taking an average of the root mean square errors.

```
##       V1       V2       V3
## 65917.05 65548.04 58019.35
```

Now, we see that our RMSE values are better than the ones we derived in class. However, to build a KNN model to better our outcomes, we first need to standardize our variables and rerun the linear regressions as well so that the RMSE values across different kinds of regressions are comparable. Let's look at the summary statistics of the standardized variables to ensure they have indeed all been standardized.

```
##       price.V1          lotSize.V1            age.V1
##  Min.   :-2.102436   Min.   :-0.715942   Min.   :-0.955703
##  1st Qu.:-0.680270   1st Qu.:-0.472626   1st Qu.:-0.510650
##  Median :-0.224161   Median :-0.186372   Median :-0.305241
##  Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.000000
##  3rd Qu.: 0.477780   3rd Qu.: 0.056944   3rd Qu.: 0.208282
##  Max.   : 5.719477   Max.   :16.745560   Max.   : 6.747141
##     landValue.V1       livingArea.V1       pctCollege.V1
##  Min.   :-0.981041   Min.   :-1.837249   Min.   :-3.441954
##  1st Qu.:-0.555584   1st Qu.:-0.733908   1st Qu.:-0.345254
##  Median :-0.272897   Median :-0.194336   Median : 0.138606
##  Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.000000
##  3rd Qu.: 0.161126   3rd Qu.: 0.617442   3rd Qu.: 0.816009
##  Max.   :10.794695   Max.   : 5.602234   Max.   : 2.557902
##      bedrooms.V1       fireplaces.V1       bathrooms.V1
##  Min.   :-2.635971   Min.   :-1.082269   Min.   :-2.886256
##  1st Qu.:-0.189042   1st Qu.:-1.082269   1st Qu.:-0.607841
##  Median :-0.189042   Median : 0.715962   Median : 0.151631
##  Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.000000
##  3rd Qu.: 1.034422   3rd Qu.: 0.715962   3rd Qu.: 0.911102
##  Max.   : 4.704815   Max.   : 6.110655   Max.   : 3.948989
##        rooms.V1                   heating           fuel
##  Min.   :-2.1764601   hot air        :1121   gas     :1197
##  1st Qu.:-0.8813764   hot water/steam: 302   electric: 315
##  Median :-0.0179873   electric       : 305   oil     : 216
##  Mean   : 0.0000000
```

```
##   3rd Qu.: 0.5216309
##   Max.   : 2.1404856
##                 sewer      waterfront newConstruction centralAir
##   septic           : 503   Yes:  15   Yes:  81        Yes: 635
##   public/commercial:1213   No :1713   No :1647        No :1093
##   none             :  12
##
##
##
```

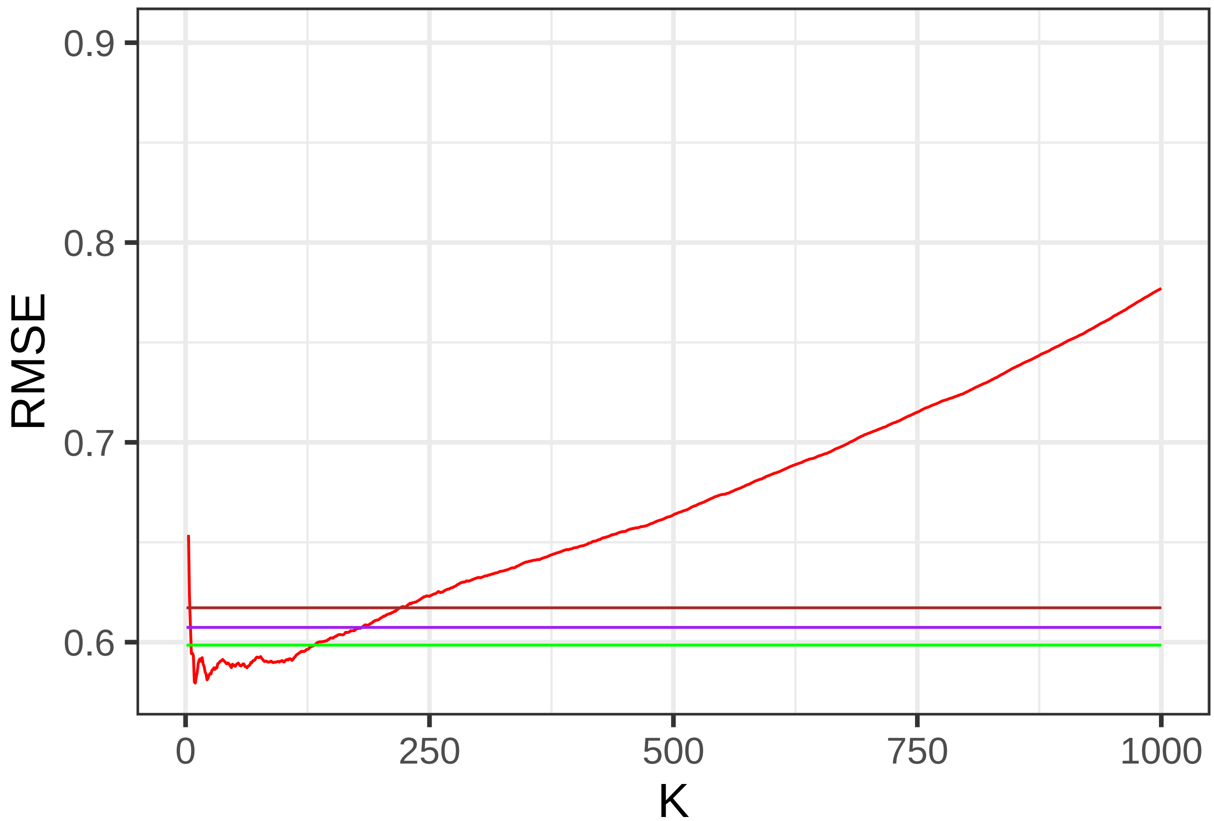
Let's now repeat the linear regressions as before and get the average RMSE values for the three regressions.

```
## [1] 0.598516
```

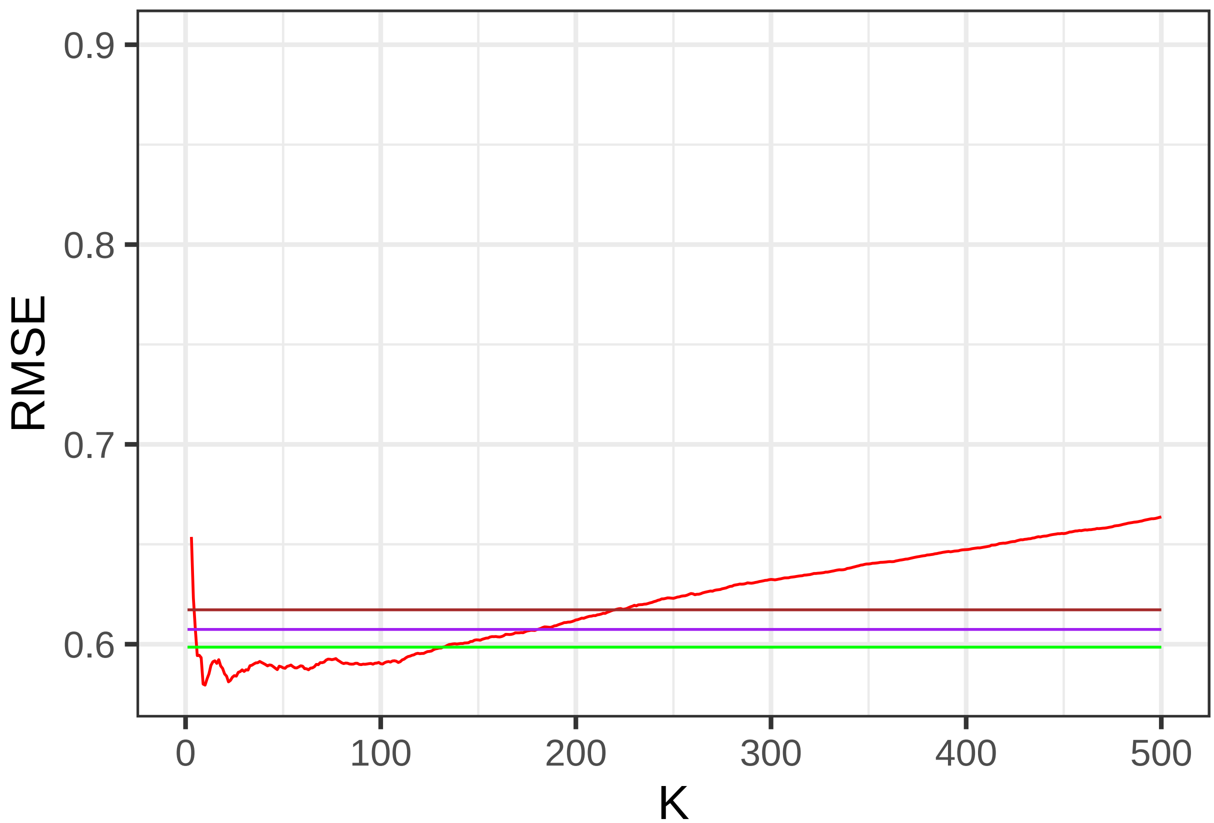
```
## [1] 0.6172167
```

```
## [1] 0.6073937
```

We are ready to run the KNN regression using the same variables.



Since we are only interested in looking at the data where RMSE does better than the linear models, let us narrow down our K values upto 500. We can see below that the optimal K value seems to be closer to 50.



Since, we have by now justified our usage of the variables in both the linear as well as the KNN regression by minimizing RMSE, let us have a look at the variables and their coefficients of regression once again.

```
##              (Intercept)                 bedrooms
##              -1.48485825              -0.07704319
##             centralAirNo                landValue
##              -0.16405867               0.29351677
##               livingArea                bathrooms
##               0.41482991               0.15827633
##             waterfrontNo                  lotSize
##              -1.50290716               0.08094998
##                      age              fuelelectric
##              -2.83957808              -0.26193175
##                  fueloil        newConstructionNo
##               0.08897765               3.11961799
##                    rooms         landValue:lotSize
##               0.08322276              -0.03948346
##          bathrooms:rooms     age:newConstructionNo
##               0.08134714               2.79133845
##       bedrooms:bathrooms centralAirNo:fuelelectric
##              -0.04669774               0.20154693
##     centralAirNo:fueloil    landValue:waterfrontNo
##              -0.18786621               0.02601707
```

Being a waterfront property increases the price steeply. Prices are also driven by the number of bathrooms and rooms and are negatively related with number of bedrooms although this would be because of the association between the bedrooms, bathrooms and rooms. Central Airconditioning is another driving factor in the determination of prices. Based on the type of fuel used, the price of house varies from high for gas to low for electricity. Hot air heating also drives the prices up as compared to electricity or water. Being a new construction however, starkly affects the price of the house.

Question 2: A Hospital Audit

Hospital Audits are important to determine the effectiveness of hospital operations from a objective standpoint. In this particular case, the goal is to determining the performance of radiologists using a statistical audit of their recent patient interactions - a crucial link between modern data-science and hospital operations. Two overall questions are posited:

1. First question: are some radiologists more clinically conservative than others in recalling patients, holding patient risk factors equal?
2. Second question: when the radiologists at this hospital interpret a mammogram to make a decision on whether to recall the patient, does the data suggest that they should be weighing some clinical risk factors more heavily than they currently are?

At the core of each question is reducing the number of false negatives - where a radiologist recommends a patient to conduct further tests and thereby allows a patient to begin immediately; and false positives - where a radiologist recommends further tests but ultimately turns out that there was no cancer. By introducing a statistical model, the goal is to augment the predictive capabilities of radiologist and offer a better standard of care for patients.

This audit is structured in four parts: first is a brief summary of the data and how it is structured, second is a demonstration and presentation of answering question one, third is a similar approach for question two, fourth is a conclusion of the audit's findings and recommendations for improvement of future radiologist performance or audit effectiveness.

Part One: Brief Summary of Data

```
##          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
##                       Max.   :1.00000   Max.   :1.0000
##      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
##   Max.   :1.0000   Max.   :1.00000
```

The data of mammograms used in this audit were selected from a Hospital in Seattle, Washington. At this hospital, five radiologists were selected at random for the audit - where about 200 mammograms were randomly selected from the hospital for each. For a total of 987 mammograms covering 7 parameters:

- age: 40-49*, 50-59, 60-69, 70 and older
- family history of breast cancer: 0=No*, 1=Yes
- history of breast biopsy/surgery: 0=No*, 1=Yes
- breast cancer symptoms: 0=No*, 1=Yes

- menopause/hormone-therapy status: Pre-menopausal, Post-menopausal & no hormone replacement therapy (HT), Post-menopausal & HT*, Post-menopausal & unknown HT
- previous mammogram: 0=No*, 1=Yes
- breast density classification: 1=Almost entirely fatty, 2=Scattered fibroglandular tissue*, 3=Heterogeneously dense, 4=Extremely dense

Of these factors, two are of special interest: [recall] and [cancer]. In the abstract [recall] can be explained as the following: upon seeing the medical history of a patient, they can either recommend either one of two actions: recall for further screening or not. It is presumed that radiologists utilize all of the information available before they make a decision. This implies that there is a inherent correlative factor between recall and patient history. On the other hand [cancer] is whether or not a patient, whether through the recall screening process, or through another pathway of discovery - develops cancer within a 12 month window after seeing the radiologist.

Part Two: Clinical Conservativism

Without knowing how patients are assigned to radiologists, it is presumed that the relationship is random at best, and preferential at worst. With a random assignment, we can presume that each radiologist chosen for the audit would have seen, on average, the same makeup of patients that would necessitate a mammogram. A random assignment would entail a random drawing of cancer patients from the overall total cancer patient pool from the population. If preferential - meaning that a patient approaches a radiologist and requests care and upon the approval of the radiologist, we see an issue of sampling error within the audit data; as there is a bias introduced between patient selection and radiologist. Radiologist may either self-select for more difficult cases or easier based on preference and patients self-select based on their estimate of the reputation of the radiologist within the medical community.

Regardless of assignment, the primary method of which we rank the clinical conservationism is to create a model that is trained on each of the radiologists' and then test the model on data from both the radiologist and other patients not seen by the radiologist in question. The goals behind this approach are twofold: one is to recreate a evaluation profile of the radiologist through a linear model of determining whether or not a patient should be recalled, two to determine whether or not a patient who is recalled or not develops cancer within a 12 month time frame.

The table below depicts the Root Mean Squared Error (RMSE) of each radiologist's model tested on a small sub-sample of the radiologist's test data and other radiologists' testing data.

```
##                         lm1        lm1.w         lm2      lm2.w
## radiologist13  0.362884528 0.363475760  0.43703167 0.41690651
## radiologist34  0.291038807 0.388776717  0.34675643 0.43351155
## radiologist66  0.387521391 0.369011936  0.50053818 0.43860951
## radiologist89  0.417197056 0.402194203  0.46984211 0.45024402
## radiologist95  0.348925933 0.382831323  0.42876230 0.51368693
## SuperRad       0.357458048 0.354011255  0.37116730 0.34973033
## Rad13.compare  0.005426480 0.009464505  0.06586437 0.06717617
## Rad34.compare -0.066419240 0.034765462 -0.02441087 0.08378122
## Rad66.compare  0.030063343 0.015000681  0.12937088 0.08887918
## Rad89.compare  0.059739008 0.048182949  0.09867481 0.10051369
## Rad95.compare -0.008532115 0.028820068  0.05759500 0.16395660
```

Example:

- **radiologist13**: we have a the same linear model, **lm1 = glm(recall ~ .-cancer, data=brca_train, maxit = maxit),** trained to 20% of radiologist13's sample data as well as the whole mammogram data - excluding radiologist13's.
- **SuperRad**: is a model trained on a 20% random sample of the whole data set and tested on the remainder of the whole data set. This pseudo-radiologist serves as the benchmark for comparing

radiologists to an artificial standard if one radiologist had access and saw all of the patients from the data set.

- **Rad13.compare**: is determined by subtracting the model RMSE result of ***radiologist13** by **SuperRad**. A positive value means that a model trained on **radiologist13's** training data did worse once it was tested on out of sample testing data and vice-versa.

```
##                        lm1       lm1.w         lm2      lm2.w
## radiologist13  0.362884528 0.363475760  0.43703167 0.41690651
## radiologist34  0.291038807 0.388776717  0.34675643 0.43351155
## radiologist66  0.387521391 0.369011936  0.50053818 0.43860951
## radiologist89  0.417197056 0.402194203  0.46984211 0.45024402
## radiologist95  0.348925933 0.382831323  0.42876230 0.51368693
## SuperRad       0.357458048 0.354011255  0.37116730 0.34973033
## Rad13.compare  0.005426480 0.009464505  0.06586437 0.06717617
## Rad34.compare -0.066419240 0.034765462 -0.02441087 0.08378122
## Rad66.compare  0.030063343 0.015000681  0.12937088 0.08887918
## Rad89.compare  0.059739008 0.048182949  0.09867481 0.10051369
## Rad95.compare -0.008532115 0.028820068  0.05759500 0.16395660
```

Just by viewing the table, it can be clearly discerned under **lm1** that on average, radiologists 13, 66, and 89 had worse performance than the benchmark **SuperRad** when looking at the RadXX.compare values for each radiologist; while 34 and 95 had better performance. But when we examine the results of each radiologists' model tested on the global data set, we find that on average, all radiologists were worse off. However **lm1** is a linear regression involving non-interacting variables from the data set. If we were to examine **lm2 <<- glm(recall ~ (.-cancer)^2,data=brca_train,maxit = maxit)** where we interact every variable with itself and another we find different results. Radiologist 95's model performance flips and becomes worse with 95's within-sample data. But once tested on the global data set, all radiologists' models performed worse than the benchmark. The takeaway from this analysis demonstrates that human radiologists, on average, are not as effective in determining whether or not a patient should be recalled than a statistical model. Although this might increase the number of false positives and false negatives, the overall increase in cancer detection would allow immediate treatment for true positives who otherwise would have gone undiagnosed. As for whether or not this behavior can be determined to be clinically conservative, meaning that radiologist will opt to recall a patient even if the clinical factors do not signal a need to recall, the distinction is minimal at best and hard to determine as all of the radiologists selected in the audit perform marginally better or worse than the benchmark.

Part Three: Weighing Different Clinical Risk Factors

We first approach this question by developing four linear models that attempts to predict cancer rates based on the parameters available in the data set.

- lm3 <<- glm(cancer ~ recall,data=brca_train,maxit = maxit)
- lm4 <<- glm(cancer ~ recall + history,data=brca_train,maxit = maxit)
- lm5 <<- glm(cancer ~ .,data=brca_train,maxit = maxit)
- lm6 <<- glm(cancer ~ (.)^2,data=brca_train,maxit = maxit)

Because the goal of this question is to determine whether or not radiologists are effectively utilizing all of a patient's clinical data to determine whether or not to recall a patient, we first examine **lm3** and **lm4**. Both are linear models designed to find the partial effect of whether or not a patient was recalled and if they developed cancer within the next 12 months. However the distinction is that **lm3** only has recall as its x variable while **lm4** has both recall and family history.

```r
summary(lm3)
```

```
##
## Call:
```

```
## glm(formula = cancer ~ recall, data = brca_train, maxit = maxit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.15789  -0.01923  -0.01923  -0.01923   0.98077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.019231   0.007238   2.657  0.00805 **
## recall      0.138664   0.019054   7.277 8.24e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.03541611)
##
##     Null deviance: 29.784  on 789  degrees of freedom
## Residual deviance: 27.908  on 788  degrees of freedom
## AIC: -393.14
##
## Number of Fisher Scoring iterations: 2
```

```r
summary(lm5)
```

```
##
## Call:
## glm(formula = cancer ~ . - recall, data = brca_train, maxit = maxit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.12691  -0.05138  -0.03509  -0.01804   0.99374
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.017696   0.036611  -0.483  0.62898
## radiologistradiologist34 -0.006550   0.021887  -0.299  0.76481
## radiologistradiologist66 -0.017135   0.021739  -0.788  0.43083
## radiologistradiologist89 -0.004308   0.022577  -0.191  0.84874
## radiologistradiologist95 -0.020006   0.021900  -0.913  0.36127
## ageage5059                0.018250   0.023882   0.764  0.44499
## ageage6069                0.014440   0.027999   0.516  0.60620
## ageage70plus              0.042110   0.027939   1.507  0.13217
## history                   0.011339   0.018208   0.623  0.53363
## symptoms                  0.009935   0.032965   0.301  0.76321
## menopausepostmenoNoHT    -0.001859   0.017293  -0.108  0.91441
## menopausepostmenounknown  0.062220   0.036526   1.703  0.08889 .
## menopausepremeno          0.017051   0.025245   0.675  0.49960
## densitydensity2           0.024038   0.025596   0.939  0.34796
## densitydensity3           0.046069   0.025642   1.797  0.07279 .
## densitydensity4           0.101038   0.032537   3.105  0.00197 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.03756466)
##
##     Null deviance: 29.784  on 789  degrees of freedom
```

```
## Residual deviance: 29.075  on 774  degrees of freedom
## AIC: -332.78
##
## Number of Fisher Scoring iterations: 2
```

By itself, we can see that the **recall** variable has a very significant (p-value close to 0) and large effect on whether or not a patient develops cancer. This makes sense because upon evaluating a patient, a radiologist will then determine whether or not the patient will be recalled and receive additional testing. Based on their experience and education, they will want to find the factors that most likely contributes to cancer. At the same time however, we also see significant (in terms of p-value and magnitude) effects from **age**, **menopause/hormone-therapy status**, and **breast density classification**. In light of these factors, a series of model efficacy tests were conducted to determine the effectiveness of different models.

```
##                        lm3        lm3.w           lm4        lm4.w
## radiologist13  0.324074711  0.336818945  0.3300584959  0.336718812
## radiologist34  0.216756564  0.309510048  0.2311163242  0.318019845
## radiologist66  0.375234897  0.334755153  0.3775737463  0.337464098
## radiologist89  0.387815334  0.324497040  0.4035942628  0.364026522
## radiologist95  0.294851580  0.320331592  0.3010532327  0.333142673
## SuperRad       0.329823184  0.330232865  0.3309978798  0.331266656
## Rad13.compare -0.005748473  0.006586081 -0.0009393839  0.005452156
## Rad34.compare -0.113066620 -0.020722816 -0.0998815556 -0.013246810
## Rad66.compare  0.045411713  0.004522288  0.0465758665  0.006197442
## Rad89.compare  0.057992150 -0.005735825  0.0725963830  0.032759867
## Rad95.compare -0.034971604 -0.009901273 -0.0299446471  0.001876017
##                        lm5        lm5.w         lm6      lm6.w
## radiologist13  0.370085434  0.378117422  0.40197358 0.40363778
## radiologist34  0.289220555  0.398030468  0.32324685 0.41812011
## radiologist66  0.404991514  0.362973180  0.41357450 0.38936605
## radiologist89  0.438920976  0.396836789  0.48686565 0.46158402
## radiologist95  0.348578189  0.387555312  0.39012341 0.43351195
## SuperRad       0.374639424  0.374439807  0.37833849 0.37508430
## Rad13.compare -0.004553991  0.003677615  0.02363509 0.02855348
## Rad34.compare -0.085418870  0.023590661 -0.05509164 0.04303581
## Rad66.compare  0.030352090 -0.011466627  0.03523601 0.01428175
## Rad89.compare  0.064281552  0.022396982  0.10852716 0.08649972
## Rad95.compare -0.026061235  0.013115505  0.01178492 0.05842765
```

Looking across **SuperRad** we see that the RMSE of each model remains fairly consistent throughout the different implementation and test of each model - except when we exclude **recall** in models **lm5** and **lm6** . The exclusion of **recall** has a meaningful impact models' ability to guess the cancer rate for each patient. Given this puzzling outcome, the next step would be to examine **lm5.w** and **lm6.w** where we take models that exclude **recall** - after all, as recall determinations occur after a radiologist sees a patient and not before, we cannot use it to predict cancer; and see which radiologist model performs the best. Iteration terms seems to be resulting in higher RMSE in the predictive model than by itself. Given the summary results from earlier regarding the significance of some variables over others, it can be concluded that radiologists weigh **age**, **menopause/hormone-therapy status**, and **breast density classification** as indicators of cancer than other factors excluding recall.

Part Four: Conclusion

Ultimately, it can be determined that human radiologists may appear to be more conservative than a statistical model, but the underlying analysis claims otherwise - the difference is small in nature and not of sufficient significance to sacrifice patient care for a more effective diagnosing mechanism. The number of false positives and false negatives remain small in comparison when the model changes from one to another.

```
## [1] "lm3 Confusion Table"

##    yhat
## y     0   1
##   0 161  30
##   1   2   4

## [1] "lm4 Confusion Table"

##    yhat
## y     0   1
##   0 160  31
##   1   2   4

## [1] "lm5 Confusion Table"

##    yhat
## y     0   1
##   0 187   4
##   1   6   0

## [1] "lm6 Confusion Table"

##    yhat
## y     0   1
##   0 165  26
##   1   5   1
```

- Pair-wise guesses and actual cancer results.
- (0,0) means that a patient did not have caner and was not recalled.
- (1,0) means that a patient had cancer but was not recalled.
- (0,1) means that a patient did not have cancer but was recalled.
- (1,1) means that a patient had cancer and was successfully recalled.

Question 3: Going Viral

In the digital age, where information is no longer a constraint but rather - a superfluousness asset, determining what will be popular is a contentious task in of itself. Factors observable and unobservable go into the underlying decision-making of drawing a user's attention towards the consumption of given content. At the core of this question is determining what factors will ultimately predict the 'virality' given a piece of content and its associating metadata. To better understand this phenomena, a data set of 39,797 articles were utilized to train and test models to this effect.

Methodology

Given the large data set, it was computationally impractical to run the models on the entirety of the data set. A compromise was reached where 1000 articles were randomly sampled per cycle of model testing. Thereby maintaining independent and identically distributed random variables among the samples. Six different linear models were trained on 80% of this sampled data and tested on the remaining 20%. As mentioned before, a Root Mean Squared Error value was established among the models and then they were tested for in-sample and out-of-sample accuracy. As for deciding which factors played a role in determining whether or not content went viral, linear regression models were created and promising variables selected for further testing.

The models used were the following

- lm1 <<- glm(shares ~ ., data=df_train, maxit = maxit)
- lm2 <<- glm(shares ~ weekday_is_friday + num_videos + data_channel_is_lifestyle + global_rate_negative_words, data=df_train, maxit = maxit)

- lm3 <<- glm(shares ~ . - weekday_is_friday - num_videos - data_channel_is_lifestyle - global_rate_negative_words, data=df_train, maxit = maxit)
- lm4 <<- glm(shares ~ (.)^2, data=df_train, maxit = maxit)
- lm5 <<- glm(shares ~ (weekday_is_friday + num_videos + data_channel_is_lifestyle + global_rate_negative_words)^2, data=df_train, maxit = maxit)
- lm6 <<- glm(shares ~ (. - weekday_is_friday - num_videos - data_channel_is_lifestyle - global_rate_negative_words)^2, data=df_train, maxit = maxit)

Results

Because of computational limitation of the underlying base model, only sampling 1000 from a population of about 40,000, different iterations of training/testing cycles yields different results. As a result, only a general sense of what factors makes content go viral can be obtained at this time.

```
##
## Call:
## glm(formula = shares ~ ., data = df_train, maxit = maxit)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -10252   -2468    -905     545  195037
##
## Coefficients: (2 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.376e+03  3.178e+03   0.748   0.4550
## n_tokens_title               -5.357e+01  1.683e+02  -0.318   0.7503
## n_tokens_content             -1.374e+00  1.166e+00  -1.178   0.2392
## num_hrefs                     2.286e+00  3.895e+01   0.059   0.9532
## num_self_hrefs               -4.415e+01  1.217e+02  -0.363   0.7168
## num_imgs                      1.102e+02  5.533e+01   1.992   0.0467 *
## num_videos                   -3.347e+01  1.139e+02  -0.294   0.7690
## average_token_length          1.959e+02  5.405e+02   0.363   0.7171
## num_keywords                  2.751e+02  1.864e+02   1.476   0.1404
## data_channel_is_lifestyle     3.286e+03  1.830e+03   1.795   0.0730 .
## data_channel_is_entertainment -2.757e+03  1.279e+03  -2.155   0.0315 *
## data_channel_is_bus          -1.393e+03  1.351e+03  -1.031   0.3027
## data_channel_is_socmed       -4.824e+02  1.797e+03  -0.268   0.7885
## data_channel_is_tech         -1.253e+03  1.351e+03  -0.928   0.3538
## data_channel_is_world        -2.812e+03  1.312e+03  -2.144   0.0323 *
## self_reference_min_shares    -2.786e-02  7.631e-02  -0.365   0.7152
## self_reference_max_shares    -2.521e-02  4.149e-02  -0.608   0.5437
## self_reference_avg_sharess    5.739e-02  1.151e-01   0.498   0.6183
## weekday_is_monday             1.123e+02  1.547e+03   0.073   0.9422
## weekday_is_tuesday           -2.177e+02  1.484e+03  -0.147   0.8834
## weekday_is_wednesday         -1.229e+03  1.492e+03  -0.824   0.4103
## weekday_is_thursday          -6.312e+02  1.511e+03  -0.418   0.6763
## weekday_is_friday            -1.750e+03  1.551e+03  -1.128   0.2596
## weekday_is_saturday          -9.790e+02  1.941e+03  -0.504   0.6142
## weekday_is_sunday                    NA         NA      NA       NA
## is_weekend                           NA         NA      NA       NA
## global_rate_positive_words   -1.061e+04  2.640e+04  -0.402   0.6878
## global_rate_negative_words   -6.778e+03  4.050e+04  -0.167   0.8671
## avg_positive_polarity         1.770e+03  6.086e+03   0.291   0.7713
## min_positive_polarity        -2.427e+03  6.526e+03  -0.372   0.7101
## max_positive_polarity        -2.206e+03  2.421e+03  -0.911   0.3626
```

```
## avg_negative_polarity          2.819e+03  6.842e+03   0.412   0.6804
## min_negative_polarity         -4.893e+03  2.707e+03  -1.808   0.0710 .
## max_negative_polarity         -7.156e+03  6.283e+03  -1.139   0.2551
## title_subjectivity             1.124e+03  1.342e+03   0.837   0.4026
## title_sentiment_polarity       2.958e+02  1.579e+03   0.187   0.8515
## abs_title_sentiment_polarity  -2.815e+02  2.208e+03  -0.127   0.8986
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 86845004)
##
##     Null deviance: 7.0913e+10  on 799  degrees of freedom
## Residual deviance: 6.6436e+10  on 765  degrees of freedom
## AIC: 16930
##
## Number of Fisher Scoring iterations: 2
```

The RMSE output for each model:

```
##           RMSE
## lm1   8004.990
## lm2   7669.190
## lm3   8044.068
## lm4 139862.939
## lm5   7784.611
## lm6 105677.383
```

Confusion Matrixes of each Model on the sample 1000 set.

```
## [1] "lm1 Confusion Matrix, training and testing"
```

```
##    yhat
## y     0   1
##   0  71 303
##   1  82 344
```

```
##    yhat
## y    0  1
##   0 25 74
##   1 10 91
```

```
## [1] "lm2 Confusion Matrix, training and testing"
```

```
##    yhat
## y     0   1
##   0  19 355
##   1  22 404
```

```
##    yhat
## y     0   1
##   0   6  93
##   1   1 100
```

```
## [1] "lm3 Confusion Matrix, training and testing"
```

```
##    yhat
## y     0   1
##   0  45 329
```

```
##   1  68 358
```

```
##    yhat
## y    0  1
##   0 16 83
##   1  7 94
```

```
## [1] "lm4 Confusion Matrix, training and testing"
```

```
##    yhat
## y     0   1
##   0 138 236
##   1 168 258
```

```
##    yhat
## y    0  1
##   0 47 52
##   1 28 73
```

```
## [1] "lm5 Confusion Matrix, training and testing"
```

```
##    yhat
## y     0   1
##   0  16 358
##   1  22 404
```

```
##    yhat
## y    0  1
##   0 11 88
##   1  2 99
```

```
## [1] "lm6 Confusion Matrix, training and testing"
```

```
##    yhat
## y     0   1
##   0 146 228
##   1 164 262
```

```
##    yhat
## y    0  1
##   0 46 53
##   1 38 63
```

Conclusion

In hindsight, it is believed that the method of regress first, then threshold second, would be the optimal mechanism of developing models that predict what articles will go viral. The reasoning behind this presumption is that just as a model needs what factors that will make an article succeed, it will also need to know what factors causes it to not succeed. There are useful information and metrics from failure data that needs to be considered when construing predictive models.