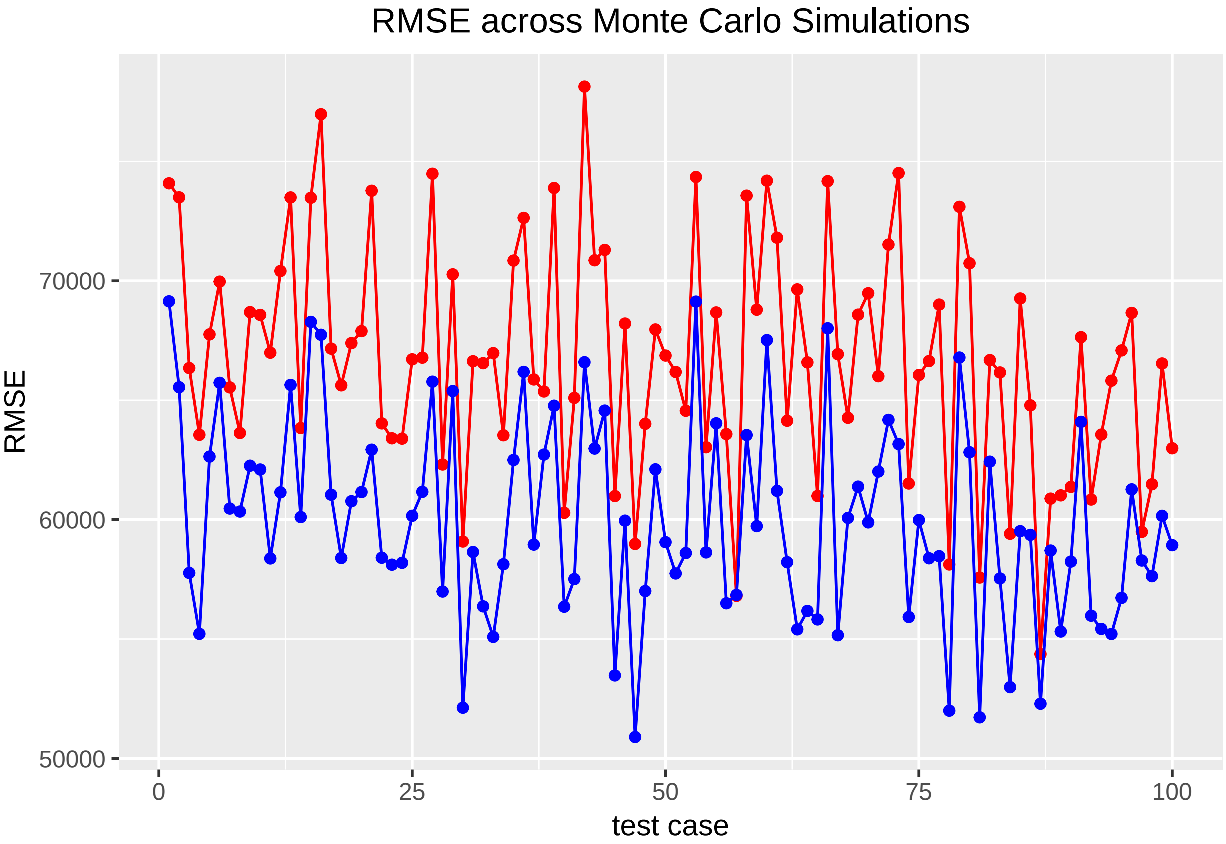
Homework 2

Linh Nguyen

3/15/2019

Question 1

The purpose of this report is to demonstrate the performance of modelling housing price in Saratoga. I compare baseline model, referring in class as "medium" model with my "hand-built" model. The criteria for evaluating performance is root mean squared error (RMSE).



As we observe, the RMSE for "hand-built" model has lower RMSE for most of Monte Carlo training-testing split compare to the baseline mode. The average RMSE is listed below:

Table 1: The Average RMSE

	AVG RMSE
baseline	66772.82
model	59960.84

By trials and errors, I test different interactions and combinations of factors such as land value, heating system, fuel/gas usage, number of rooms to modify the "hand-build" model to obtain the best model. In further details, my "hand-built" model has the following regression results:

Table 2: Regression Result for Hand-built Model

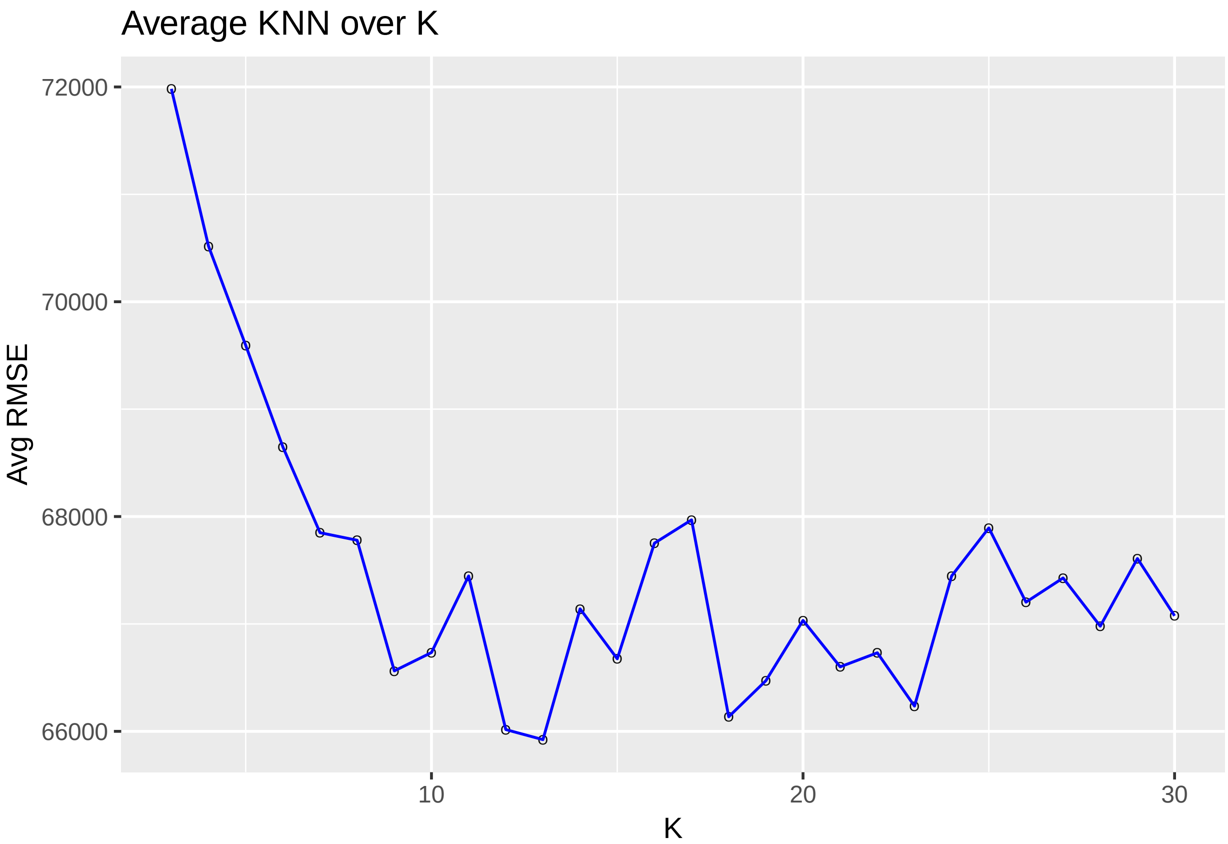
	coefficients.Estimate	coefficients.Std..Error	coefficients.t.value	coefficients.Pr...t..
(Intercept)	7.805011e+03	1.689418e+04	0.4619941	0.6441446
lotSize	1.175491e+04	3.113576e+03	3.7753739	0.0001653
age	-3.179614e+02	3.726324e+03	-0.0853284	0.9320103
landValue	8.441235e-01	6.091350e-02	13.8577403	0.0000000
livingArea	8.714490e+01	5.649386e+00	15.4255518	0.0000000
pctCollege	-5.611216e+02	2.135779e+02	-2.6272460	0.0086852
bedrooms	-7.843118e+03	3.485472e+03	-2.2502311	0.0245615
bathrooms	2.243555e+04	4.502193e+03	4.9832501	0.0000007
heatinghot water/steam	-1.755185e+04	6.456901e+03	-2.7183094	0.0066282
heatingelectric	-2.124778e+03	1.980800e+04	-0.1072687	0.9145885
fuelelectric	-2.006984e+04	1.989456e+04	-1.0088102	0.3132091
fueloil	4.667938e+03	7.482137e+03	0.6238777	0.5327915
newConstructionNo	5.464142e+04	8.091345e+03	6.7530701	0.0000000
centralAirNo	-2.155748e+04	5.285785e+03	-4.0783881	0.0000474
lotSize:age	-1.606309e+02	8.570529e+01	-1.8742233	0.0610708
age:landValue	3.022400e-03	1.332400e-03	2.2683458	0.0234329
age:livingArea	-3.710406e-01	1.253802e-01	-2.9593236	0.0031257
age:pctCollege	1.148853e+01	5.107529e+00	2.2493313	0.0246188
age:bedrooms	6.725127e+01	6.612520e+01	1.0170294	0.3092840
age:bathrooms	6.878274e+01	1.055346e+02	0.6517556	0.5146469
age:heatinghot water/steam	1.283908e+02	1.214300e+02	1.0573233	0.2905141
age:heatingelectric	3.379200e+02	8.164659e+02	0.4138813	0.6790132
age:fuelelectric	2.753738e+02	8.584404e+02	0.3207839	0.7484136
age:fueloil	-6.031577e+01	1.343781e+02	-0.4488513	0.6535961
age:newConstructionNo	-7.036706e+02	3.698974e+03	-0.1902340	0.8491485
age:centralAirNo	5.815197e+02	1.925704e+02	3.0197764	0.0025672

As we observe from the table, the most important factor is the land value. Looking at the very high t-value, we can conclude that this regressor is statistically significant in explaining the price of house. Similarly, the living area also plays an important role in explaining the price of house.

From the model, an interesting phenomenon is the age of the building interacts well with central heating system in influencing the price of the house as t-value for this coefficient suggests statistical significance. This is understandable that central heating system will be limited to outdated houses. The presence of central heating system indicates the relative comfort of having comfortable living environment, hence people will pay more for old buildings that include central heating system.

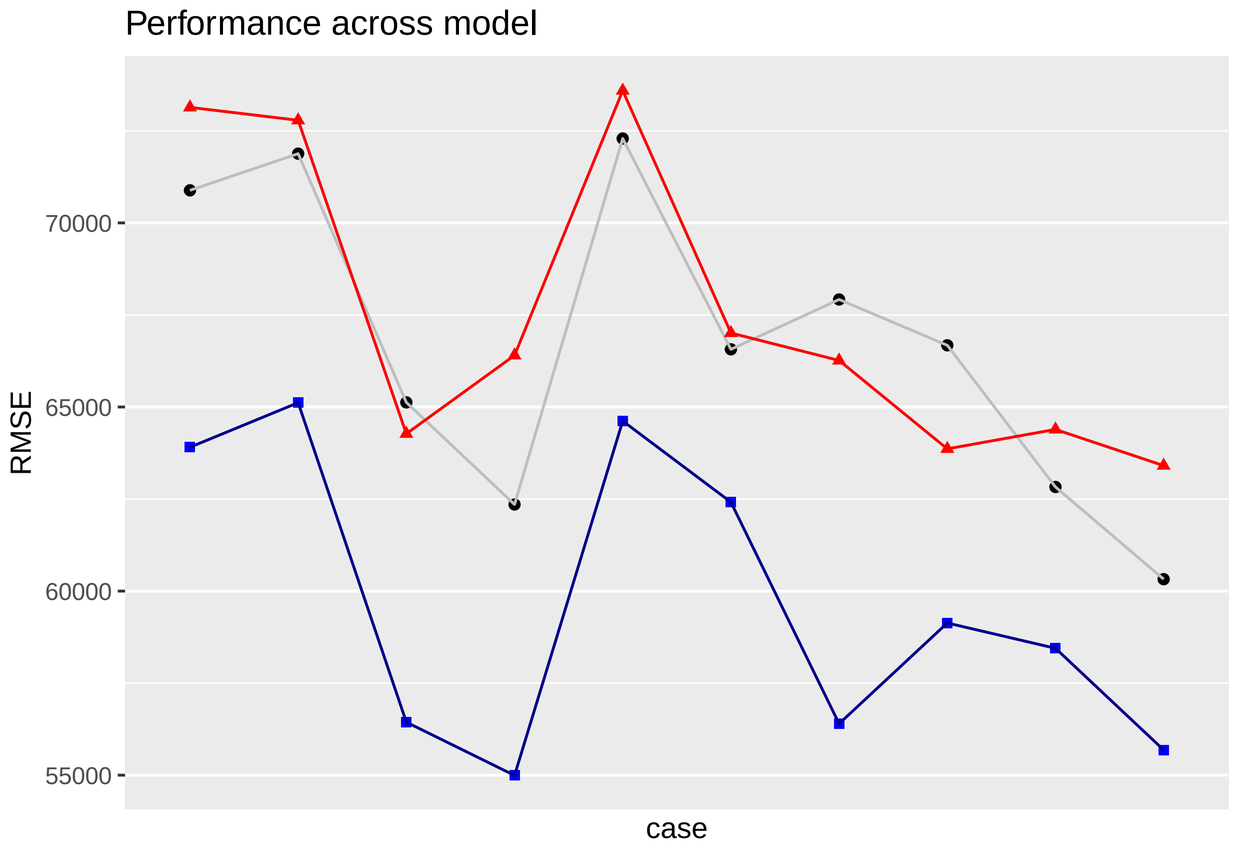
The indicator of new construction is also a significance price signal for housing according to my model.

Next, I compare my "hand-built" model with KNN model



The optimal K after repeating KNN regressions with different K on several train-test splits is 13. The lowest average RMSE obtained is 11

Final step to validate that my "hand-built" model outperform both KNN regression and baseline model is to test the three model on the same training-testing sample. I will repeat this process several times to eliminate the randomness occurs by random sampling.



As we observe, the RMSE for “hand-build” model (blue points) is lower than the KNN regression for most of 10 trials. The KNN regression (red points) and the baseline model (black points)’s performance are about the same. The average RMSE for these trials are:

Table 3: The Average RMSE

	AVG RMSE
baseline	66684.55
hand_built	59715.13
knn	67512.79

Question 2

This report address the two issues pose by the oncology unit. First, I will investigate if some radiologists are more clinically conservative than others.

I start by modelling recall decisions of each radiologists using 2 models:

- Model A: recalls based on patient’s history and radiologists
- Model B: recalls based on patient’s history and radiologists including their interactions

Table 4: The Average RMSE

	average RMSE
model A	0.4491371

	average RMSE
model B	0.4098985

Hence, model B performs better. I will apply model B to predict the decisions of radiologists on the same sets of patients

Table 5: Comparison between radiologists using model B

Radiologist	Recall probability
radiologist13	0.1257879
radiologist34	0.0852158
radiologist66	0.1702271
radiologist89	0.2076335
radiologist95	0.1323428

To cross validate that indeed Radiologist 66 and 89 are more conservative than others, I repeat the process with model A as well.

Table 6: Comparison between radiologists using model A

Radiologist	Recall probability
radiologist13	0.1390638
radiologist34	0.0887976
radiologist66	0.1860874
radiologist89	0.2006786
radiologist95	0.1327947

Even though the probability of recall is different, the ordering of conservatism remains the same. Hence, we can conclude that some radiologists are more conservative than others.

The second question I address is: weighing some clinical risk factors more heavily than they currently are is supported by the data

To validate this claim, I start with basic simple model, regressing patient's cancer outcome on the radiologist's recall decisions.

Then I add in clinical risk factors to form model A. Next, I add in risk factors and recall decisions interactions terms for model B. Last, I modify the weights of risk factors by dropping out less important factors such as density for model C. To summarize,

- Baseline Model: cancer ~ recall
- Model A: cancer ~ (.- radiologist)
- Model B: cancer ~ recall*(.- radiologist)
- Model C: cancer ~ recall + history + menopause

The table summarize the model performance in predicting cancer

Table 7: Average out-of-sample deviations

	Average Deviation for Different Models
Baseline	1.481726
Model A	1.494811

	Average Deviation for Different Models
Model B	1.427753
Model C	1.425003

Model C outperforms the baseline model (without any clinical risk factors) and model A(including all clinical risk factors) in terms of out of sample deviation, suggesting that some clinical risk factors should be put more weight than others. Specifically, history and menopause status are two important factors that should be considered to account for probability of having cancer

To further investigate the predictive power of model C, I report the accuracy, true positive rate, false positive rate and confusion matrix between model C and baseline model on the entire data set

Table 8: Confusion matrix of baseline model

	prediction = 0	prediction = 1
cancer = 0	915	35
cancer = 1	28	9

Table 9: Confusion matrix of model C

	prediction = 0	prediction = 1
cancer = 0	932	18
cancer = 1	32	5

Hence, the baseline model measure of performance are:

- Accuracy rate: 0.9361702
- False positive rate: 0.7954545
- True positive rate: 0.2432432

Compare to those of model C:

- Accuracy rate: 0.9493414
- False positive rate: 0.7826087
- True positive rate: 0.1351351

As false postive rate decreases and accuracy rate increases, suggesting model C with consideration of patient's history and menopause status will identify more precisely the cancer's outcome. We then can conclude that some clinical risk factors are more important than others suggesting by the data.

Question 3

2 models I consider for the first problem (regress first, threshold second):

- Model 1: without publication timings in the weekend, max/min polarity and global rate

  shares ~ (. - is_weekend - weekday_is_sunday - min_positive_polarity - max_positive_polarity - min_negative_polarity - max_negative_polarity - global_rate_positive_words - global_rate_negative_words - abs_title_sentiment_polarity)

- Model 2: emphasized on word counts and keywords $shares\ title + title^2 + content + content^2 + keyword + keywords^2 + content * keywords$

2 models I consider for the second problem (threshold first, classification second):

- Model 3: similar to model 1, but takes indicator viral as y-value

  viral ~ (. -share - is_weekend - weekday_is_sunday - min_positive_polarity - max_positive_polarity - min_negative_polarity - max_negative_polarity - global_rate_positive_words - global_rate_negative_words - abs_title_sentiment_polarity)

- Model 4: binomial logistic regression for classification of model 3 (including a logit link)

Following the suggestion of Dr. Scott in measuring the performance of models predicting virality of articles, the average confusion matrix, overall error rate, true positive rate and false positive rate for my best model are reported below.

Average confusion matrix for the 4 models:

Table 10: Confusion matrix for baseline model (model 1)

	prediction = 0	prediction = 1
viral = 0	87.96	3931.98
viral = 1	30.44	3878.62

Table 11: Confusion matrix for improved model (model 2)

	prediction = 0	prediction = 1
viral = 0	7.33	4012.61
viral = 1	4.35	3904.71

Table 12: Confusion matrix for classification model (model 3)

	prediction = 0	prediction = 1
viral = 0	2714.78	1305.16
viral = 1	1379.84	2529.22

Table 13: Confusion matrix for logit model (model 4)

	prediction = 0	prediction = 1
viral = 0	3595.42	424.52
viral = 1	2804.67	1104.39

Overall performance:

Table 14: Average performance of different models

	Model 1	Model 2	Model 3	Model 4
Error rate	0.4997377	0.5066162	0.3386303	0.4072632

	Model 1	Model 2	Model 3	Model 4
False positive rate	0.5034159	0.5068142	0.3403836	0.2776619
True positive rate	0.9922130	0.9988872	0.6470149	0.2825206

So my hand build model (model 2) perform better than the baseline model (model 1) to a small extent in the regress first, threshold second approach (using threshold: >1400 shares is "viral").

Model 3 and model 4 has a much better accuracy in prediction. Hence, threshold first, classification second approach performs better.

The reason behind this is because if we regress first, we are predicting an outcome with natural orderings. So a predicted value of 1399 is as close to 1400 as 1401 to 1400 in Euclidean normed space. However, when we threshold these predicted value, we impose a binary class (whether $\hat{y} \geqslant 1400$) that suggesting 1401 and 1399 means totally different things even though their outcomes in regression are close. So the decision whether an article is viral or not for predicted value in close proximity to threshold level 1400 shares is problematic.

When we threshold first, the outcome is categorized as one-hot vector. The ordering of outcome break downs as now values are classes and class does not exhibit relative proximity. It implies there is no natural orderings of the outcome. Hence threshold first reduces the problem of predicted values clustering around the threshold level causing prediction decisions to be wrong in some cases.