

IST 707 Final Project

Graham Latsa

Uday Datta

6/12/2019

Table of Contents

[Introduction and Objective 4](#_Toc11098008)

[Project Approach 4](#_Toc11098009)

[Dataset and Transformation 4](#_Toc11098010)

[Model Result Comparison 5](#_Toc11098011)

[Visualizations 6](#_Toc11098012)

[Conclusion 8](#_Toc11098013)

[Code 9](#_Toc11098014)

Table of Figures & Tables

[Table 1: Model Comparison Data 5](#_Toc11102368)

[Figure 1: Decision Tree Model Breakdown 6](#_Toc11102355)

[Figure 2: Density plot comparison of Model results 6](#_Toc11102356)

[Figure 3: Scatter plot comparison of Model results 7](#_Toc11102357)

[Figure 4: Parallel plot comparison of Model results 7](#_Toc11102358)

# Introduction and Objective

Real Estate is one of the most critical and fascinating concepts in society today. Broadly, it has a pervasive impact on all of our lives whether it’s economic, social, or political. Unfortunately, the world of real estate is very complex; there are vast data sources, indicators, and applications - therefore the goal of project is to cut through the complexity by examining how machine learning can be used to simplify data for the average person. Specifically, we would like to answer the following questions:

1. Can machine be learning used to classify real estate data based on features? (in this case the number of beds a dwelling has)
2. What algorithm most accurately classifies the data? Which least accurately classifies the data?
3. What are the tradeoffs with different models?

# Project Approach

We will leverage the decision trees, knn, svm, and random forest models. The models will be run at a cross validated with 10 folds and 3 repeats. Our measure of performance will be based on accuracy, kappa, and system time. For this analysis, R will be used, and all primarily default parameters will be leveraged, in order to provide a fair comparison vis a vis each model.

# Dataset and Transformation

The dataset we are leveraging is from the National Association of Realtors’ publicly available research dataset[[1]](#footnote-2). The dataset includes the following feature data:

PropertyType - Condominium or Single Family home type (nominal)

NoOfBdrs - The Number of Bedrooms (ordinal)

MedianListingPrice  - The Median listing price for each CBSA’s record (numeric)

MedianListingSqft - The Median Square footage for reach CBSA’s record (numeric)

ActiveListingCount – The number of listings per CBSA (numeric)

TotalListingCount – The total number of listings - active and cancelled (numeric)

MedianDaysonMarket – The median days on market for the listings (numeric)

ViewsMM – Month over month views – (numeric)

State – geographic location of the data – (nominal)

The original dataset was reduced to 6251[[2]](#footnote-3) records and was taken from April 2019. The training set consisted of 4167 records and the test set consisted of 2084 records

# Model Result Comparison

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Decision Trees | Naïve Bayes | KNN\* | SVM | RF | Decision Trees Tuned |
| Resampling method & other info | Repeated  Cross-Validation  With 10 fold and 3 repeats | Repeated  Cross-Validation  With 10 fold and 3 repeats | Repeated  Cross-Validation  With 10 fold and 3 repeats | Repeated  Cross-Validation  With 10 fold and 3 repeats  Kernel = SVM Linear | Repeated  Cross-Validation  With 10 fold and 3 repeats | Repeated  Cross-Validation  With 10 fold and 3 repeats with a tune length of 5 |
| Accuracy | 0.5581943 | 0.2438 | 0.5969 | 0.8316 | 0.8628 | 0.8373 |
| Kappa | 0.4280744 | 0.0135 | 0.4901 | 0.7871 | 0.8271 | 0.7943 |
| Run Time | 5.61 | 2217.61 | 65.20 | 546.64 | 1883.58 | 24.407 |
| Accuracy to Run time Ratio | 0.0994 | 0.000109 | 0.0091 | 0.002 | 0.0004 | 0.034 |
| Confusion  Matrix |  |  |  | A | A | cid:image001.png@01D522E3.13B30E80 |
| Top 3 features (variables) | MedianListingSqft  MedianListingPrice    TotalListingCount | MedianListingSqft  MedianListingPrice  TotalListingCount | MedianListingSqft  MedianListingPrice    TotalListingCount | MedianListingSqft  MedianListingPrice  TotalListingCount | MedianListingSqft  MedianListingPrice  TotalListingCount | MedianListingSqft  TotalListingCount  ActiveListingCount |

Table 1: Model Comparison Data

A

Near Identical prediction at the 2 bedroom range – but as the rooms increase, the random forest model is more accurate (as denoted by the blue arrow

\* Did not list final KNN tunned from code since it didn’t fit on page and wasn’t any more accurate than other models

# Visualizations

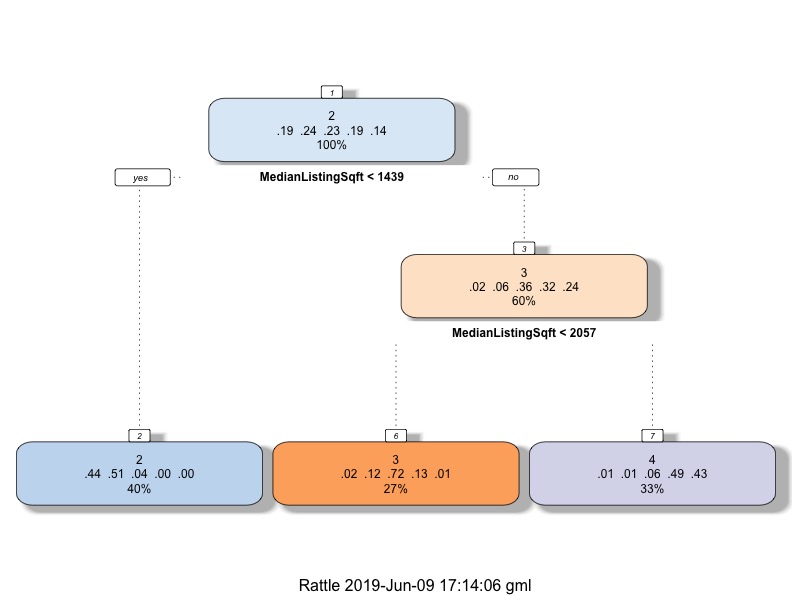


Figure 1: Decision Tree Model Breakdown

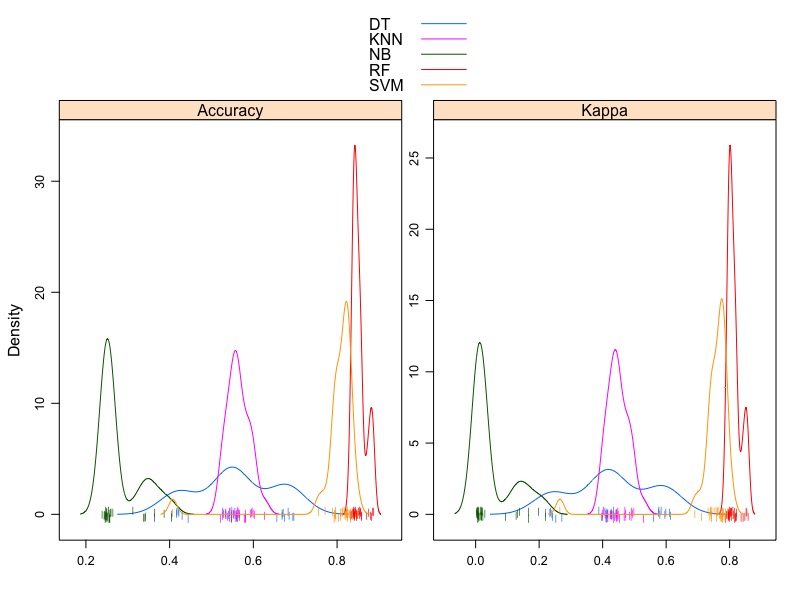


Figure 2: Density plot comparison of Model results

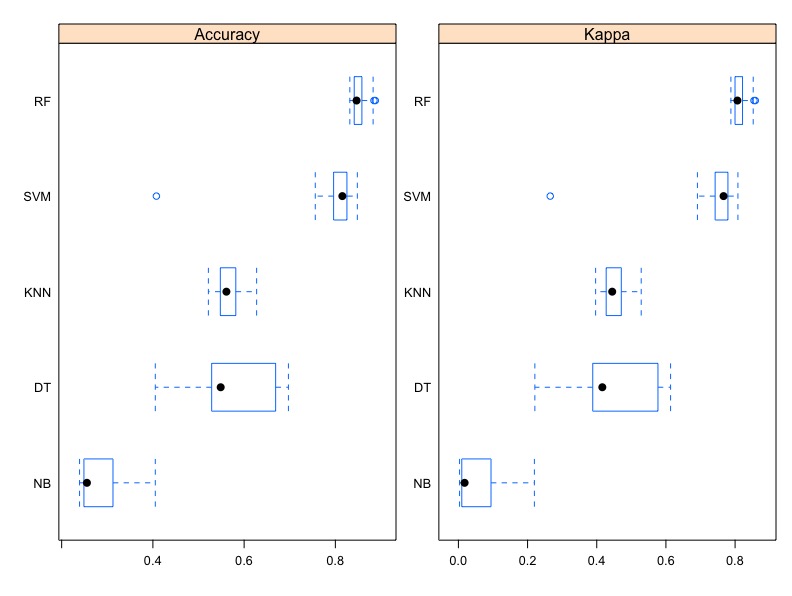


Figure 3: Scatter plot comparison of Model results

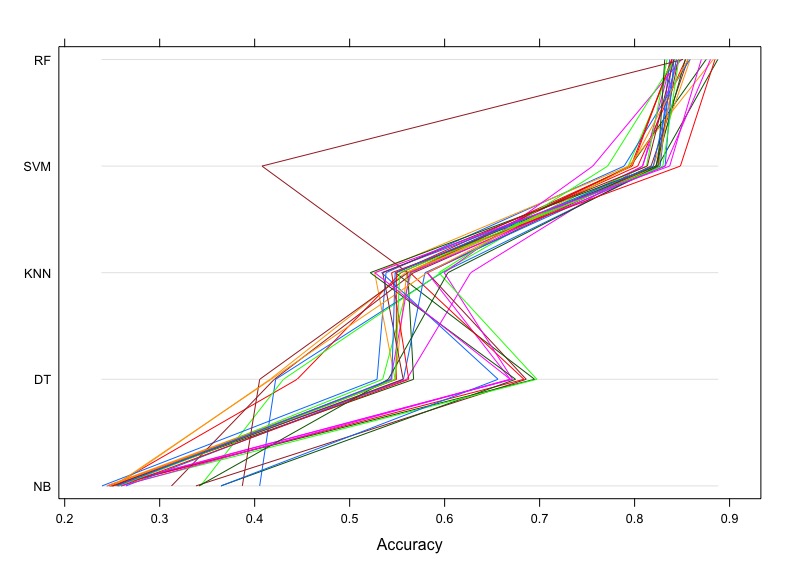


Figure 4: Parallel plot comparison of Model results

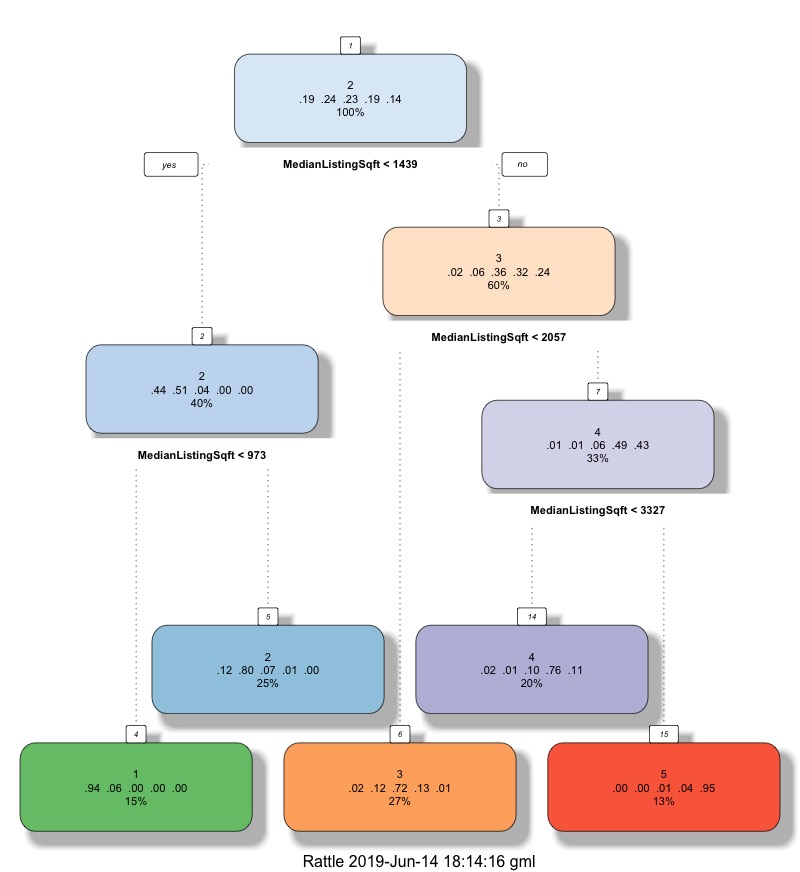


Figure Decision Tree Model Breakdown for tuned DT model

# Conclusion

The two top features (Area & Price) are positively correlated with number of rooms, which makes logical sense. The bigger the area, the higher the propensity for more rooms. If we examine price, the assumption is that more rooms equals a higher price, this is true in most scenarios, but this logic is flawed in some instances. For example, for areas like New York City (Manhattan) or San Francisco higher prices and number of rooms may not be directly related or are in some cases inversely related (Location is more of factor).

The decision tree provides quick results that are actually on par with one other algorithm (KNN) but only requires a fraction of the time. It is still weak in its overall accuracy, due to the fact that the algorithm only predicts rooms 2, 3, & 4, (see [results](#_Model_Result_Comparison) and the model above). The random forest model while the most accuracy consumed the 2nd highest amount of time. The SVM model provided a high degree of accuracy and was approximately 3.5 times faster without any hyper parameter tuning.

We believe that the SVM & Tuned Decision Tree Models are the most practical model to use. Both runs at least 1/3 of the time the Random Forest model does, with only a 3% sacrifice in accuracy. This means it’s more scalable and agile compared to the Random Forest Methods. We attempted to do some tuning of the algorithms, however with run times of close to 40 minutes it’s difficult to get quick results. Therefore, we have determined the SVM & Tuned Decision Tree Models are the best model for this scenario.

1. <https://www.realtor.com/research/data/> [↑](#footnote-ref-2)
2. Unfortunately, due to the volume of data and the memory intensive requirements we had to drop the CBSA code and Name – the extensive number of levels caused our run times to be extremely high (exceeding over 12 hours per model – over 800+ levels [↑](#footnote-ref-3)