AirBnB Dynamic Pricing and Wine Quality

Ryan Mokarian
Maziar Mohammad-Shahi
Maysam Mokarian

Abstract

As a host in airbnb it has been always a challenge to set the optimal price for the upcoming days. As a guest it is not always an easy decision to see if the property price is fair considering many factors such as property reviews, facilities, location and etc. Dynamic pricing can be leveraged to predict the optimal price in terms of profit for the hosts and value for the guests. Offering a high- quality wine at the right price is one of the most important goals of wineries. By analyzing different chemical factors, we are interested to know how chemical features are contributing to the wine quality to better predict and improve the quality.

1. Approach

We have chosen New York City Airbnb Open Data[1] and Wine Quality [2]. For both dataset our approach is to perform preprocessing and analysis on datasets such as outlier detection, handling missing values, feature selection and feature engineering. We will be experimenting three machine learning models namely Random Forest [3], Support Vector Machine [4] and Gradient Boosting [5].

For the airBnB data, our goal is to predict the price range of a property given the property attributes/features. Furthermore, for the Wine dataset the goal is to predict a rating of a wine given its features.

1.1. Outlier Detection

For the Airbnb data, before starting to work on preparing the features to use in the models, we looked at the prices to find possible inconsistencies. We noticed that there were a few entries that had prices of zero. For this reason, we removed those records from our data. Fortunately, there were only 11 such cases and therefore it did not hurt us in terms of the size of the training data. Additionally for the Wine dataset, we created histogram plots (Appendix, section 2) that compares red/white wine and we did not observe noticeable outlier in the frequency distribution of features by means of the data.

1.2. Missing Values

For The AirBNB data we noticed that "name", "host_name", "last_review" and "reviews_per_month" columns contains missing values. We have removed "name" and "host_name" columns as they are not informative fields in our dataset. The review_per_month represents number of reviews by guest per month. Hence we have replaced all missing values for this column with 0. Furthermore "last_review" is a field of type datetime and for the missing values, we replaced them with 1900/01/01.There is no missing value in the Wine dataset, as indicated in the "Red/White wine data description" tables in the Appendix.

1.3. Feature Selection

We performed an analysis of the Airbnb data to figure out which features are the best predictors of price range. For the categorical values, we plotted the average price against each to see which ones affect it the most. As we can see in the plots in appendix 10, the fields neighbourhood group, room

type and neighbourhood all affect price. Specially, we noticed that the top 10 most expensive neighbourhoods are at least $400/night more expensive than the least expensive 10 neighbourhoods. For the features with continuous values, we first performed a correlation analysis on the data and obtained the matrix as shown in appendix 11. We see here that they are not as good of predictors as the categorical variables but some still have a small correlation. After performing recursive feature elimination on airBnB data, we removed several fields such as minimum_nights, calculated_host_listing_counts, number_of_reviews and reviews_per_month.
Based on this analysis, we selected all the categorical variables alongside the numbered variables: latitude, longitude and the annual availability. For the wine data we utilizing all the features presented in the dataset.

Feature Engineering

In the AirBnB data, "last_review" is of type datetime. We split this column into three new columns namely: "last_review_year", "last_review_month" and "last_review_day".
Additionally we have performed one-hot encoding on all the categorical variables: "neighborhoot_group", "room_type" and "neighborhood" fields. We have also categorized the price field to 5 different categories: xx_150, 150_300, 300_450, 450_xx. Also for the Wine data , quality and feature values are normalized on a min-max scale and their mean value is shown in radar charts (Appendix , section 1). As radar chart indicates although the mean values of red and white wines are similar, there are noticeable differences in some features such as "total sulfur dioxide" and "density". Meanwhile, correlation between different features of Wine have been illustrated on the Heat maps (Appendix , section 3). As both maps show, consumers rating on the quality of wins has the most positive correlation with the alcohol percentage feature.

2. Experiments

2.1. Random Forest

We experimented the preprocessed data with Random Forest Model [3]. We ran exhaustive search to find the best hyperparameters. For both dataset we tuned the hyper parameters of n_estmiators and max_depth.

2.1.1. Wine Data

For the wine data we have set the number of estimators (trees) to 0, 100, 200, 300, 400 and 500 for both Red and white wines. We additionally ran the model with setting up the max depth to : 0, 20, 40, 80 and 100. (Appendix , section 5)

2.1.2. AirBnB Data

For the AirBnB data we have set the number of estimators (trees) to 0, 25, 50, 75, 100, 125, 175 and 200. We additionally ran the model with setting up the max depth to : 0, 20, 40, 80 and 100. (Appendix , section 6)

2.2. Gradient Boosting

Another algorithm we used was Gradient Boosting using Sklearn's GradientBoostingClassifier. With the initial runs we saw some promising results with the default hyper parameters and decided to optimize further. We optimized the fields n_estimators (number of rounds) and the learning rate and were able to improve accuracy by about 2%. The learning curves are presented in the results section.

2.2.1. Wine Data

For the wine data we have set the number of rounds to 0, 100, 200, 300, 400 and 500 for both Red and white wines. We additionally ran the model with setting up the learning rate to : 0, 1, 2, 3, 4 and 5. (Appendix , section 7)

2.2.2. AirBnB Data

For the AirBnB data we have set the number of rounds to 0, 25, 50, 75, 100, 125, 175 and 200. We additionally ran the model with setting up the learning rate to : 0, 1, 2, 3, 4 and 5. (Appendix , section 8)

2.3. Support Vector Machine

2.3.1. Wine Data

For both red and white wine, we experimented running the model with various kernels rbf, linear, poly and sigmoid. We also ran them using various values of the regularization parameter C. (Appendix, section 9)

2.3.2. AirBnB Data

We experimented running the SVM package in Sklearn called SVC and got very low accuracy rates and decided not to further optimize the hyperparameters. This is because the initial results were about 30% lower than the first two algorithms.

3. Results

The results below are obtained by running predictions on an unused test set using the optimal hyperparameters found in the previous section.

3.1. AirBnB Data

Random Forest (max_depth=20, n_estimators=25)	Gradient Boosting (learning_rate=0.15, n_estimators=25)	SVM (kernel='linear', C=1)
0.758	**0.759**	0.43

3.2. Red Wine Data

Random Forest (max_depth=40, n_estimators=100)	Gradient Boosting (learning_rate=0.4, n_estimators=300)	SVM (kernel='linear', C=100)
0.675	0.64	0.56

3.3. White Wine Data

Random Forest (max_depth=30, n_estimators=25)	Gradient Boosting (learning_rate=0.2, n_estimators=300)	SVM ((kernel='linear', C=1000)
0.66	0.64	0.50

4. Comparison

4.1. AirBnB

The results show that tree based methods are better predictors of price range. We believe that this is because tree based methods are good at using categorical features which we have a lot of in our dataset. Observing our data we see that a very important field that affects price and was not part of the dataset was the time period when the property was being offered. This is very relevant because the exact same apartment will be more expensive during long weekends and holidays.

4.2. Wine Quality

The results here also show that the tree based methods are better predictors of wine quality. This could be because the data is not linearly separable. Random Forest seems to perform better on this dataset as opposed to Airbnb data where Random Forest and Gradient Boosting have very similar accuracy. Perhaps weak learners are not as good of predictors as fully grown trees which shows that we most probably need to combine multiple feature combinations at a time to better predict.

5. Acknowledgements

All members worked on most parts of the project collaboratively.

Maysam: Preprocessing, hyper parameter tuning, writing the report and poster
Maziar: Visualization, hyper parameter tuning, testing the final model, report
Ryan: Preprocessing, visualization, hyperparameter tuning, testing the final model, report

6. References

[1] Dgomonov. “New York City Airbnb Open Data.” *Kaggle*, 12 Aug. 2019, www.kaggle.com/dgomonov/new-york-city-airbnb-open-data.

[2] Wine Quality Datasets, www3.dsi.uminho.pt/pcortez/wine.

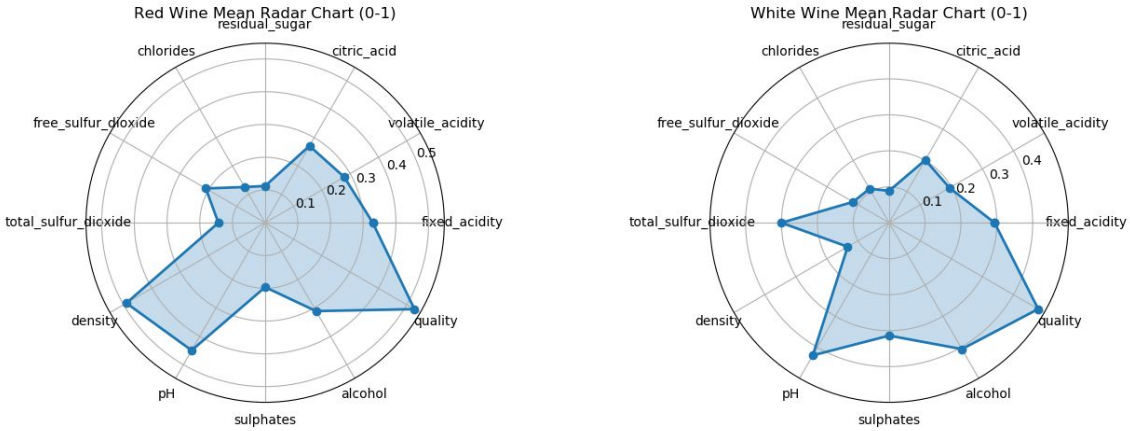
[3] Ho, Tin Kam. “Random Decision Forests.” *Proceedings of 3rd International Conference on Document Analysis and Recognition*, doi:10.1109/icdar.1995.598994.

[4] Cortes, Corinna, and Vladimir Vapnik. “Support-Vector Networks.” *Machine Learning*, vol. 20, no. 3, 1995, pp. 273–297., doi:10.1007/bf00994018.

[5] Friedman, Jerome H. “Stochastic Gradient Boosting.” *Computational Statistics & Data Analysis*, vol. 38, no. 4, 2002, pp. 367–378., doi:10.1016/s0167-9473(01)00065-2.

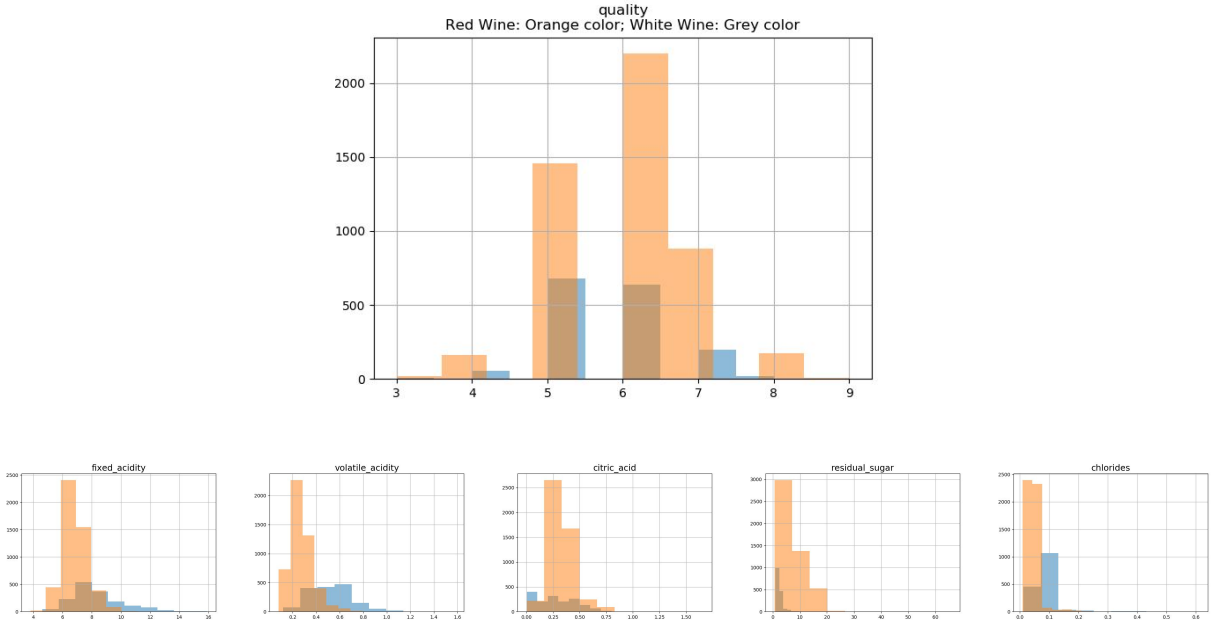
Appendix

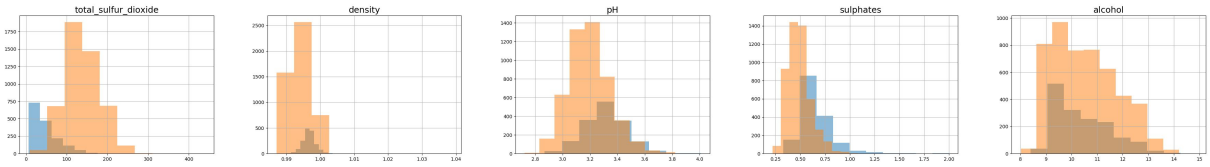
1. Radar chart mean for red/white wine



Radar chart Mean of White and Red Wins

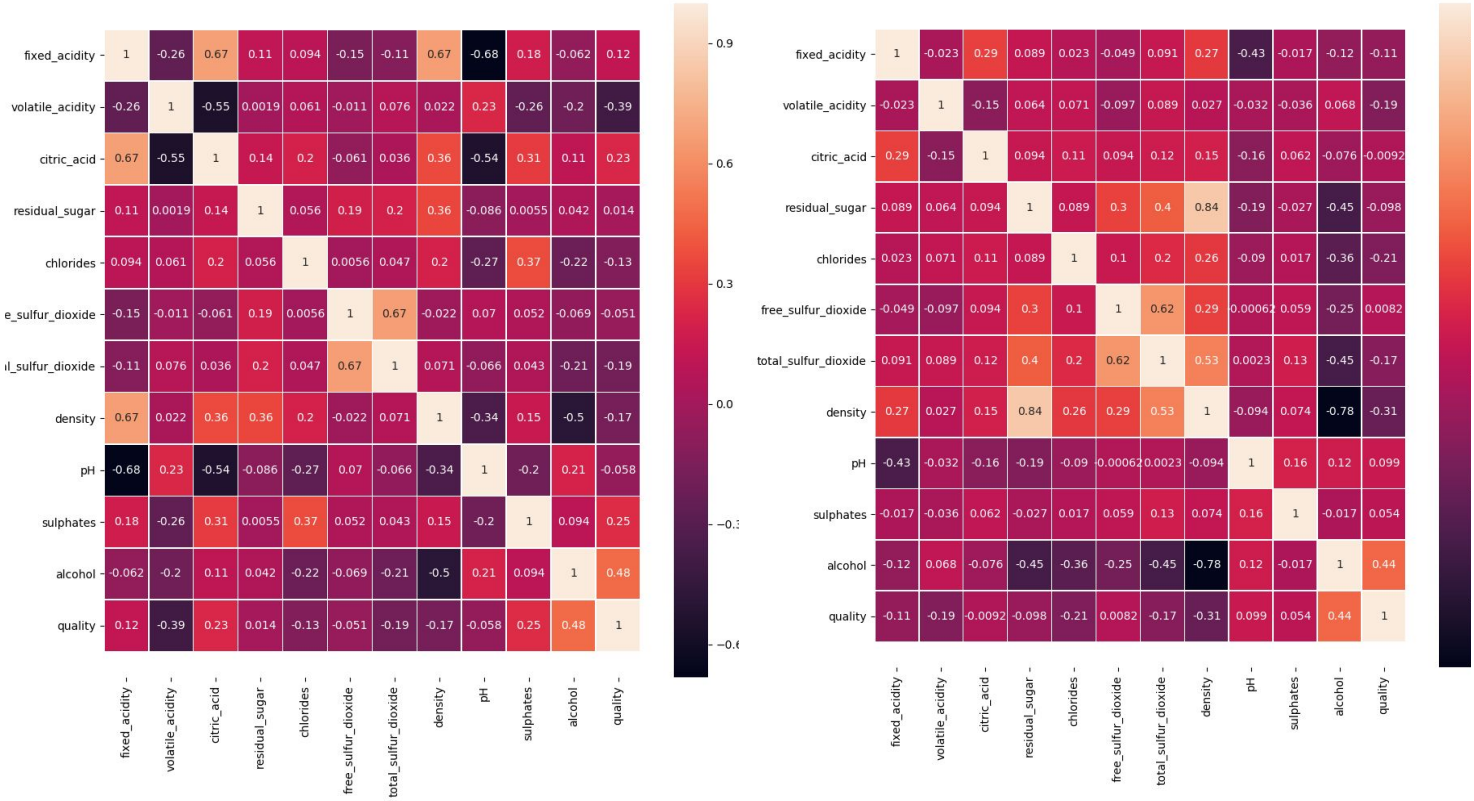
2. Histogram of red/white wine features





Histogram of red/white wine features

3. Red/White wine heatmap, correlation between features



Red (left)/White (right) wine heatmaps, correlation between features

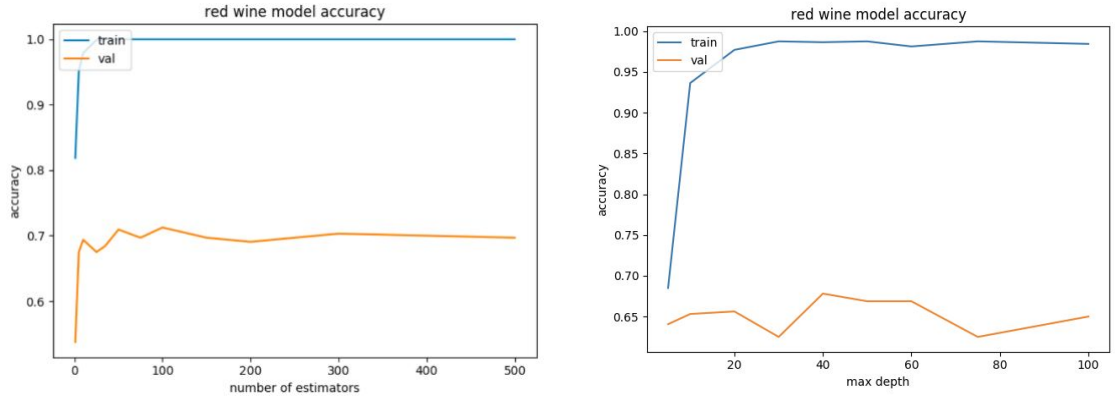
4. Red/White wine data description

Red Wine	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599
mean	8.32	0.53	0.27	2.54	0.09	15.87	46.47	1.00	3.31	0.66	10.42	5.64
std	1.74	0.18	0.19	1.41	0.05	10.46	32.90	0.00	0.15	0.17	1.07	0.81
min	4.6	0.12	0	0.9	0.012	1	6	0.9901	2.74	0.33	8.4	3
25%	7.1	0.39	0.09	1.9	0.07	7	22	0.9956	3.21	0.55	9.5	5
50%	7.9	0.52	0.26	2.2	0.079	14	38	0.9968	3.31	0.62	10.2	6
75%	9.2	0.64	0.42	2.6	0.09	21	62	0.9978	3.4	0.73	11.1	6
max	15.9	1.58	1	15.5	0.611	72	289	1.0037	4.01	2	14.9	8

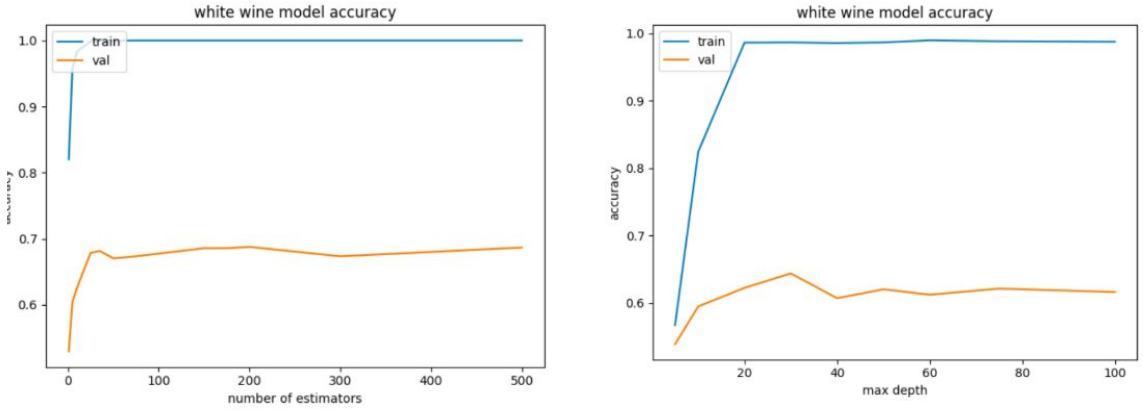
White Wine	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	4898	4898	4898	4898	4898	4898	4898	4898	4898	4898	4898	4898
mean	6.85	0.28	0.33	6.39	0.05	35.31	138.36	0.99	3.19	0.49	10.51	5.88
std	0.84	0.10	0.12	5.07	0.02	17.01	42.50	0.00	0.15	0.11	1.23	0.89
min	3.8	0.08	0	0.6	0.009	2	9	0.9871	2.72	0.22	8	3
25%	6.3	0.21	0.27	1.7	0.036	23	108	0.9917	3.09	0.41	9.5	5
50%	6.8	0.26	0.32	5.2	0.043	34	134	0.9937	3.18	0.47	10.4	6
75%	7.3	0.32	0.39	9.9	0.05	46	167	0.9961	3.28	0.55	11.4	6
max	14.2	1.1	1.66	65.8	0.346	289	440	1.039	3.82	1.08	14.2	9

Red/White wine data description

5. Random forest experiments with different hyperparameters for Wine dataset



Train/Test Accuracy for number of estimators and max depth (Random Forest, Red Wine)



Train/Test Accuracy for number of estimators and max depth (Random Forest, White Wine)

6. Random forest experiments with different hyperparameters for AirBnB dataset

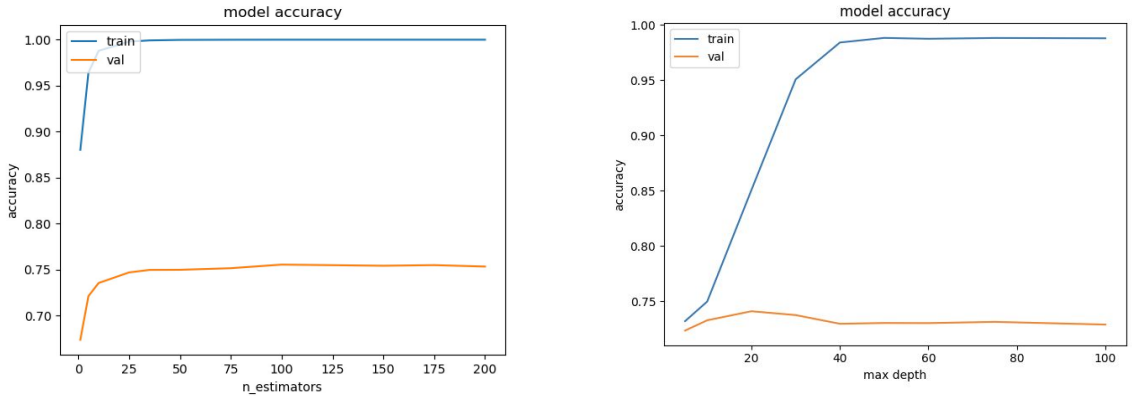
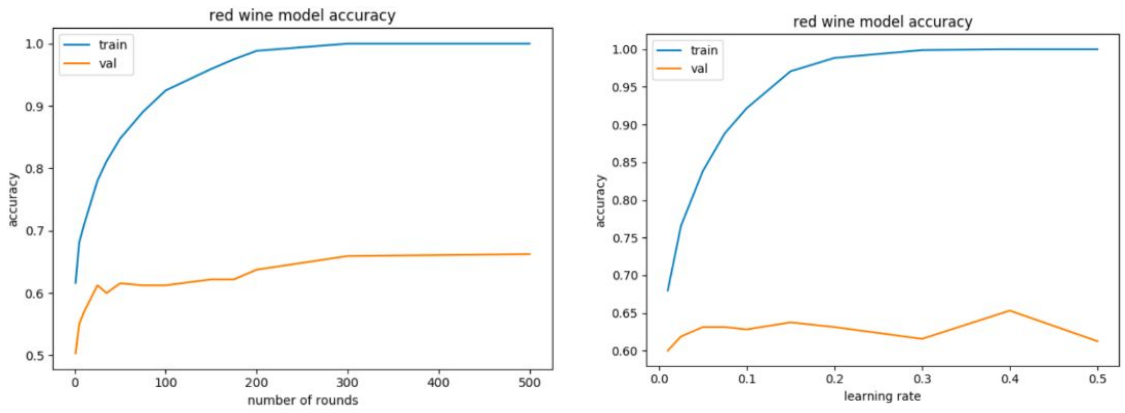
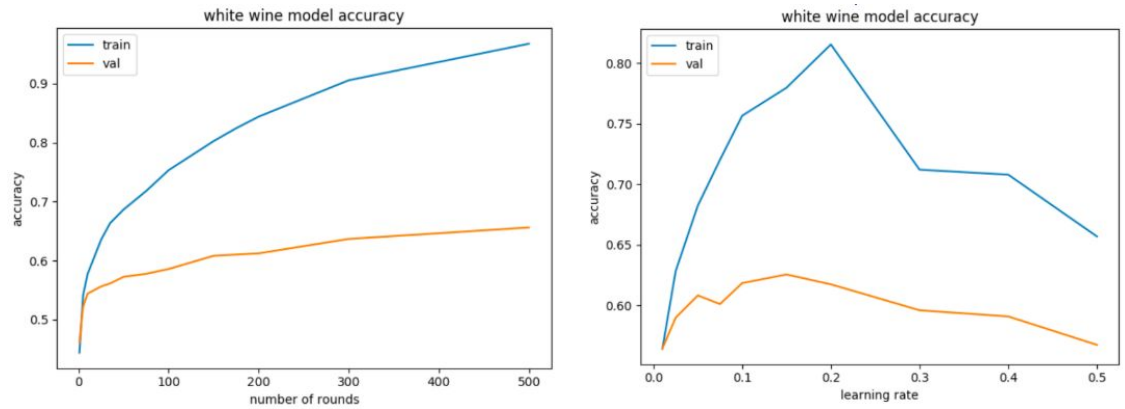


Figure 1 Train Test Accuracy for number of estimators and max depth

7. Gradient Boosting with different hyperparameters for Wine data

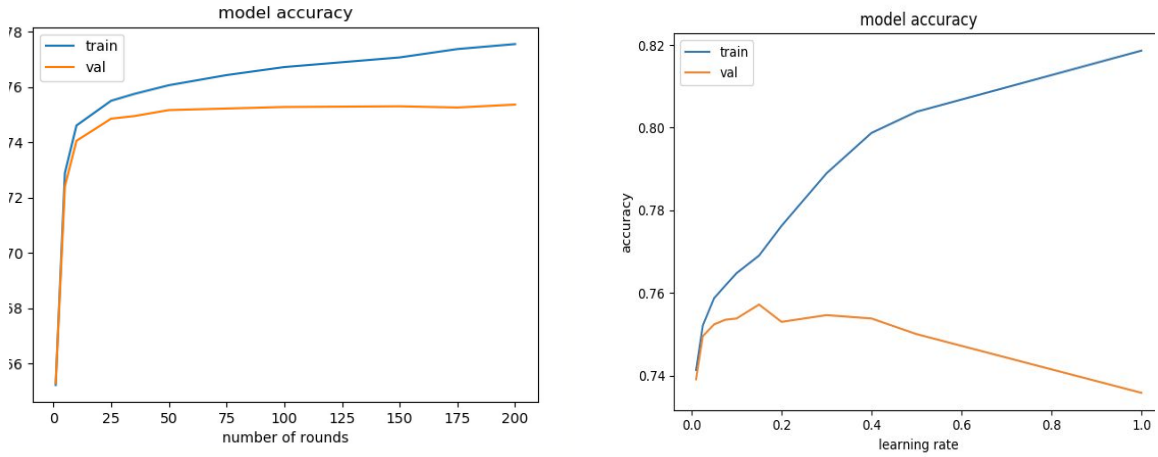


Train Test Accuracy for number of rounds and learning rate (Gradient Boosting, Red Wine)



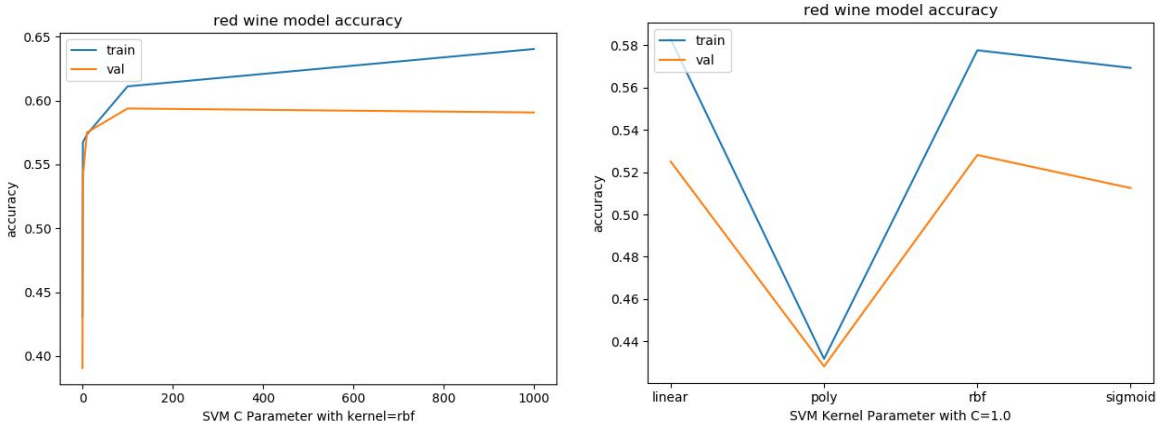
Train/Test Accuracy for number of rounds and learning rate (Gradient Boosting, White Wine)

8. Gradient Boosting with different hyperparameters for AirBnB data

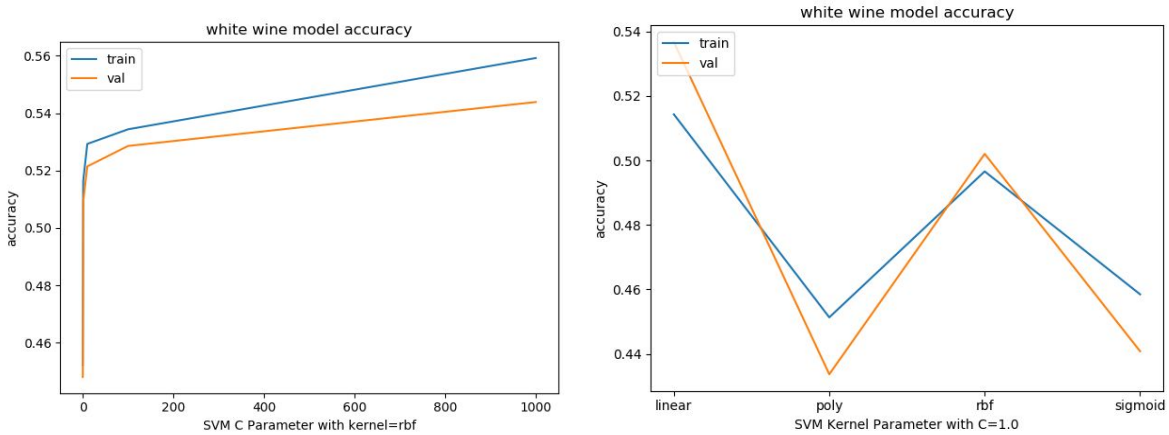


Train/Test Accuracy for number of rounds and learning rate

9. SVM with different hyperparameters for wine data

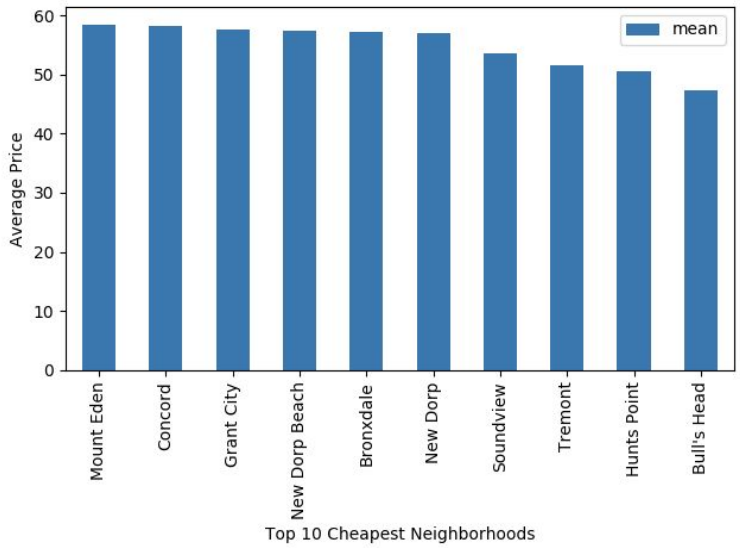


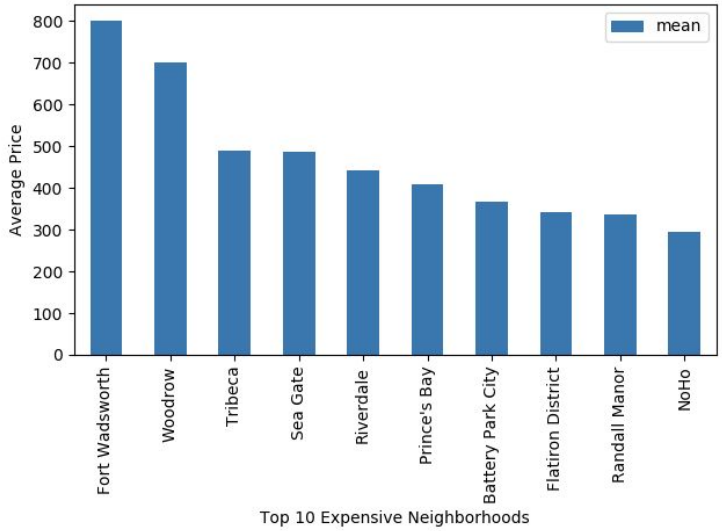
Train/Test Accuracy for C and Kernel parameters (SVM, Red Wine)

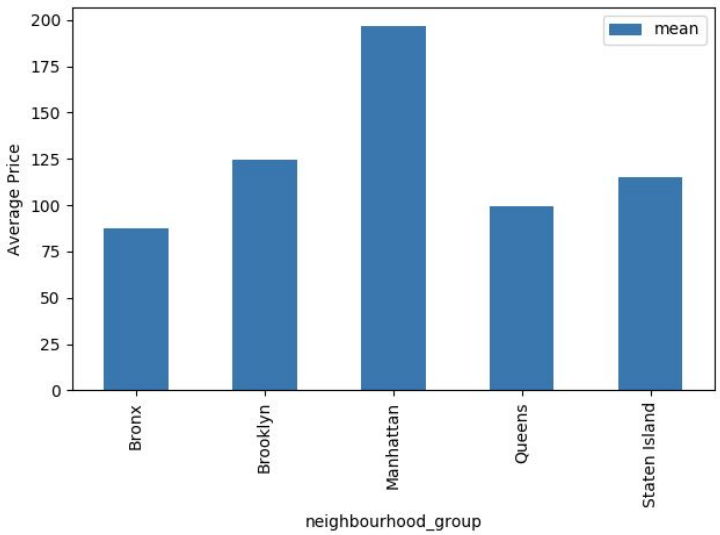


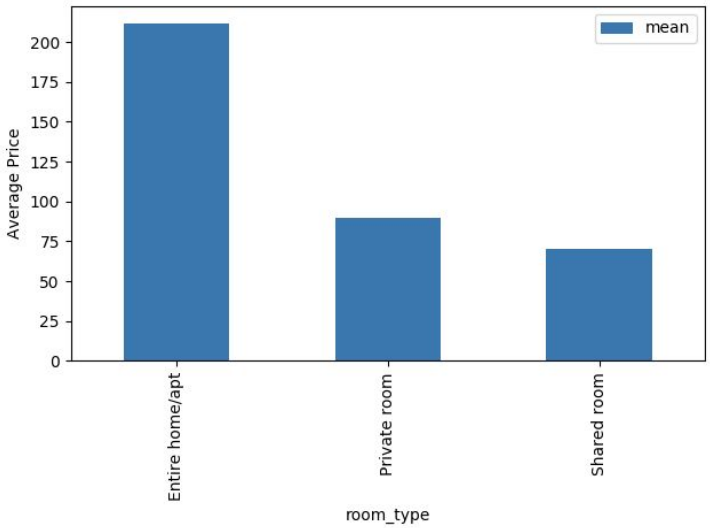
Train/Test Accuracy for C and Kernel parameters (SVM, White Wine)

10. AirBnB Categorical Data Analysis

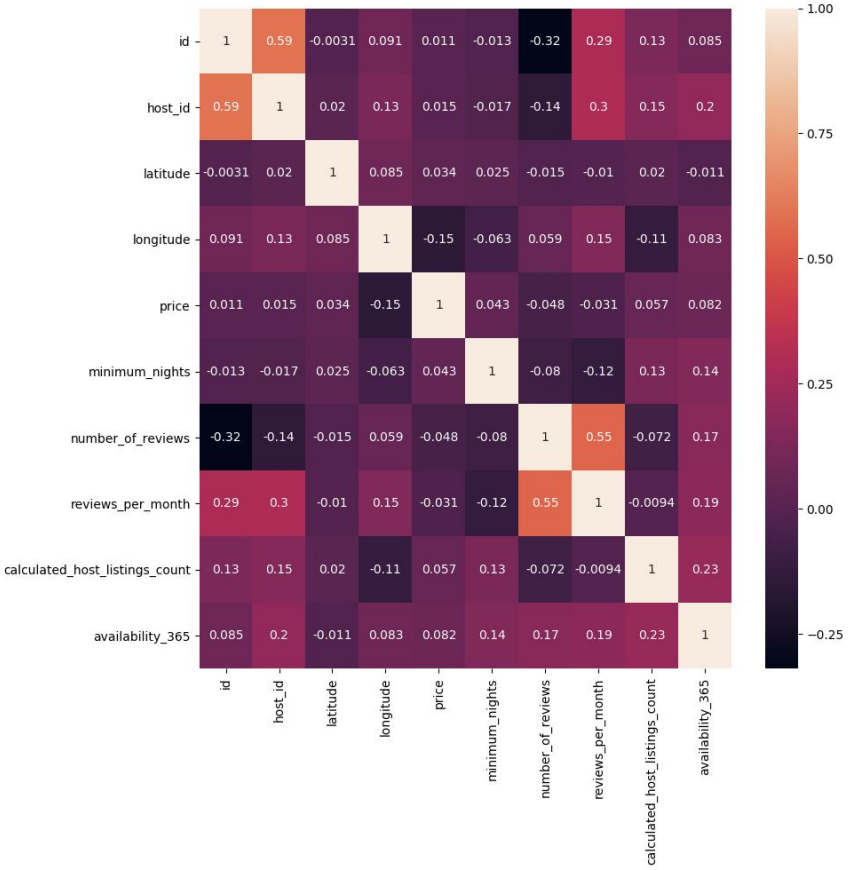





mean
Average Price
200
175
150
125
100
75
50
25
0
Bronx
Brooklyn
Manhattan
Queens
Staten Island
neighbourhood_group


mean
Average Price
200
175
150
125
100
75
50
25
0
Entire home/apt
Private room
Shared room
room_type

11. AirBnB real numbered features covariance heatmap



AirBnB real numbered features covariance heatmap