Prediction with Machine Learning: Assignment 2
Botagoz Amangeldiyeva

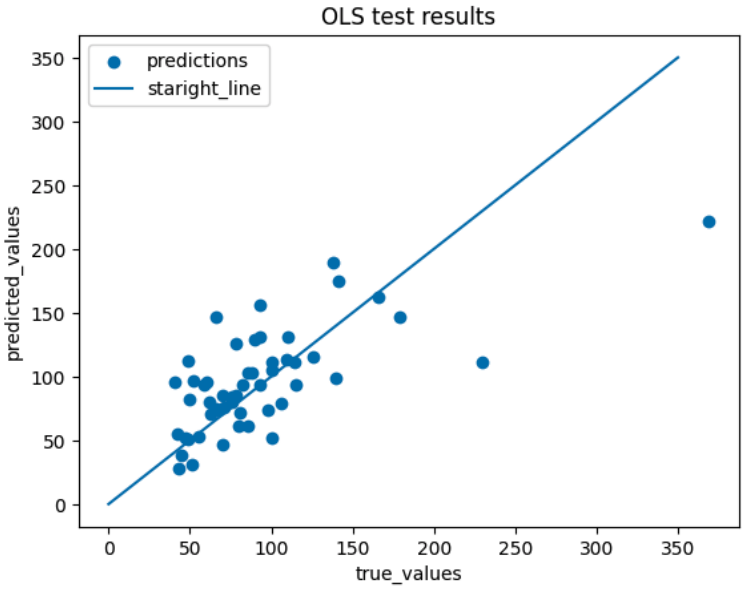
Summary Report

To build predictive models, we used 2 datasets: train data - dataset from Athens from 2018 till the end of 2022; and test data - January 2023. If one of our predictive models fits great, we can use it in this company to set prices for apartments hosting from 2 to 6 people in Athens.

As a first model, we chose OLS because first we make a **linearity assumption**. OLS is a classical linear regression model that assumes a linear relationship between the input features and the target variable. If the relationship between the features (e.g., amenities, location) and the apartment prices are approximately linear, OLS can capture this relationship well. Also, OLS provides coefficients for each feature, allowing for easy interpretation of the impact of each predictor on the target variable. This interpretability can be valuable for understanding the factors influencing Airbnb prices.

So, by running OLS regression, we got RMSE of 39. The RMSE of 39 indicates that, on average, the predicted apartment prices deviate from the actual prices by 39 units on the scale from 0 to 350. A lower RMSE is generally desirable, but in this case, a RMSE of 39 suggests that, on average, the model's predictions are within 39 units of the actual prices. If a deviation of 39 units is considered negligible in the context of apartment pricing, so the model should be acceptable.

If we plot the graph, we can see that it predicts pretty well.

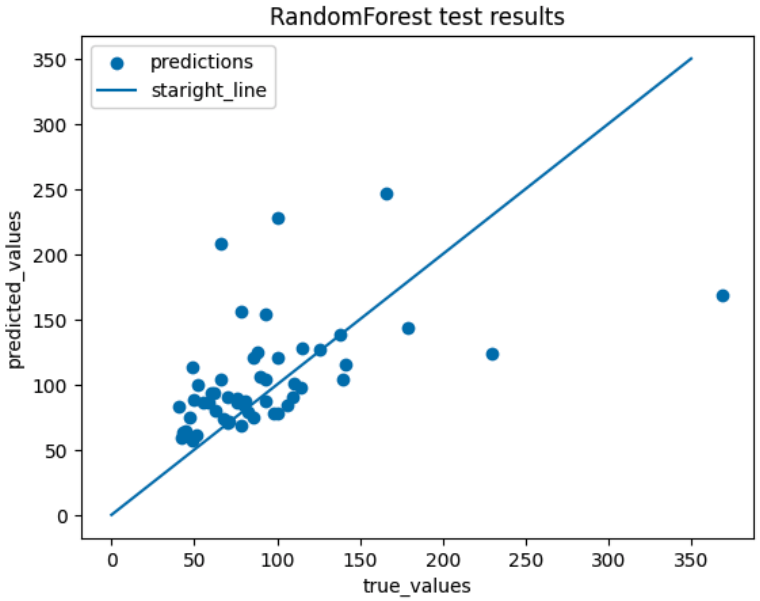


As a second model, we chose Random Forest. Here the assumption is non-linearity. Random Forest is an ensemble method that excels at capturing non-linear relationships and

interactions between features. In the context of Airbnb pricing, the relationship between various amenities, location, and price may not be strictly linear. Random Forest can handle complex relationships and interactions effectively. Also, Random Forest tends to be more robust to overfitting compared to a single decision tree. This is especially important when dealing with diverse datasets, where the risk of overfitting is higher.

RMSE is 50. The RMSE of 50 indicates that, on average, the predicted apartment prices deviate from the actual prices by 50 units on the scale from 0 to 350. That is higher than OLS, so OLS predicts with better precision.

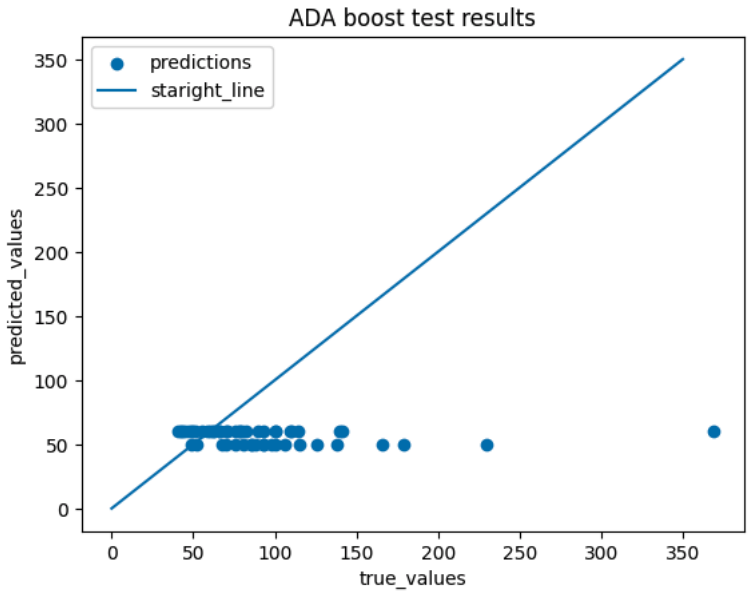
By looking at the graph, prediction is kinda similar to OLS.



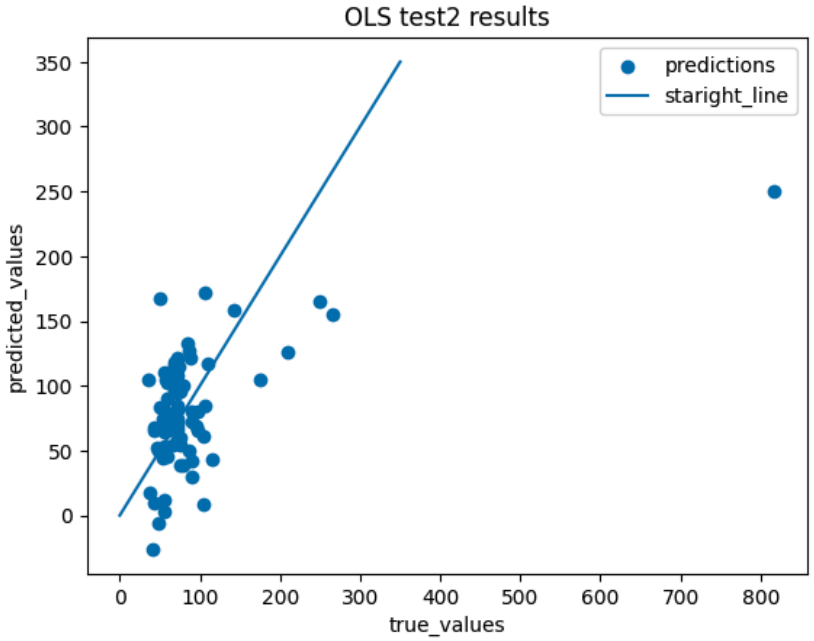
As the last model, we chose ADA boosting. AdaBoost is another ensemble method that focuses on combining weak learners (in this case, weak decision trees) to create a strong predictive model. If the relationship between features and prices is not entirely clear or if there are complex patterns, AdaBoost can adaptively improve predictive performance. And AdaBoost is known for handling class imbalance well, and it can be beneficial if your dataset has imbalanced distributions or if certain features are more critical in predicting prices than others.

RMSE is 65, which is almost double than OLS. The RMSE of 65 indicates that, on average, the predicted apartment prices deviate from the actual prices by 65 units on the scale from 0 to 350. That is way more higher than OLS, so OLS predicts with better precision.

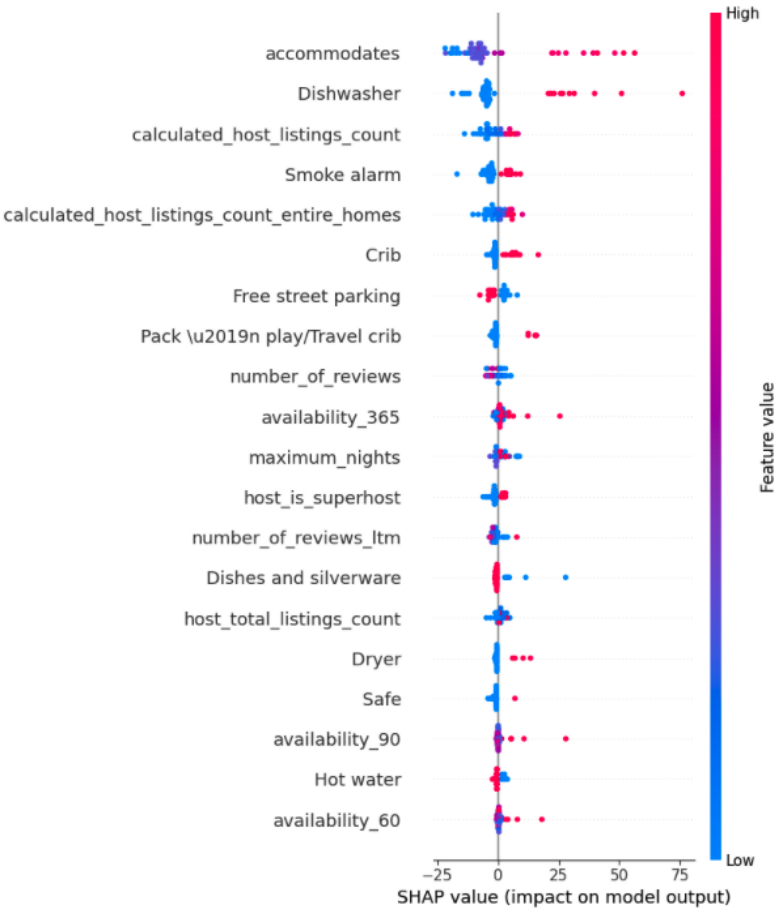
By looking at the graph, ADA boost predict very bad.



For the second task, we chose OLS, since it is the best model, and chose data for May 2023. Here RMSE is 75, which is very high. But if we look at the graph, we can see that here we have outlier. Either it is really outlier, or we don't have enough variables to explain this observations. So, we dropped it. And after that RMSE turned to 41, which is good. So, we can use OLS model to predict and set prices for your company.



As a last exercise, we took Shapley values to assess RF model. And we can see that "accommodates" and "dishwasher" variables explain the model the best. So they are very important in setting the prices.


accommodates
Dishwasher
calculated_host_listings_count
Smoke alarm
calculated_host_listings_count_entire_homes
Crib
Free street parking
Pack \u2019n play/Travel crib
number_of_reviews
availability_365
maximum_nights
host_is_superhost
number_of_reviews_ltm
Dishes and silverware
host_total_listings_count
Dryer
Safe
availability_90
Hot water
availability_60
−25
0
25
50
75
SHAP value (impact on model output)
High
Feature value
Low