Capstone Final Report

The problem that this capstone is to solve is one of an investing nature. A person or business that is looking to invest in Airbnb in both the US and Europe would want to know if one of the markets is more expensive than the other and if certain features are more important in one market than the other.

My approach to this problem was to evaluate the European and US regions separately, using the same features, and then compare the results. I gathered Airbnb data from nine cities in the US and 9 cities in Europe to ensure that the amount of data used was sufficient. Then I took steps to clean the data in preparation for Exploratory Data Analysis. After examining trends in the data, I then trained several models to determine which model would yield the best results. Finally, I used the chosen model to give predictions.

Findings

There were many different discoveries found in this project, some I would consider surprising. The first surprise finding was the range in the values of the features. The ‘price’ features for both regions varied greatly; the European Airbnb prices ranged from $0.00 to $150,368.00 while the US property prices ranged from $1 to $99999.00. The number of bathrooms had a large variance as well, with the US ranging between 0 and 36 bathrooms and in Europe the range was from 0 to 49 bathrooms. Even for a large residence it would be shocking to find forty-nine bathrooms. There were relationships that occurred that I expected to find, namely that price varied positively with the ‘accommodates’, ‘bathrooms’, ‘bedrooms’ and ‘beds’ features. On the other hand, I did not expect that the ‘Entire rental unit’ feature would have such a strong inverse correlation with all the features that start with ‘Private room’ (with the exception of ‘Private room in townhouse’).

The most important finding discovered from this capstone was the model performance. Since the data was continuous in nature, I chose the following models: Ordinary Linear Regression, Lasso, 2nd degree polynomial, 3rd degree polynomial, Gradient Boost and Random Forest Regressor. It is no surprise that that the numbers for a model in the European were similar to the numbers for that same model in the US region given that that data in the two regions turned out to be similar in structure. The criteria I used to evaluate the models was the (negative) RMSE, although I did use the coefficient of determination as well in some instances.

The first model, the Linear Regression model, performed the worst, but not by much. It gave a negative Mean Absolute Error of -41.65 and an r-squared score of .34 in the European region, and a nMAE of -41.565 and a .368 in the US region.

The next model I trained was the Lasso model. This model is similar to the Linear Regression but employs a penalty coefficient ‘alpha’. I used GridSearchCV in order to find the best alpha, which turned out to be alpha = .001 in the European region and .01 in the US region. For the European region, the best nMAE was -41.6507 and the r-squared score was .30. The metrics for the US region were an nMAE of -41.5615 and a coefficient of determination of .33. These results are roughly equivalent to those of the Linear Regression model.

The next model trained was the Polynomial Regression model. First, I ran a code that would determine the degrees I could use for my model and that resulted in 2nd and 3rd degree models. For the European region, the best MAE for the 2nd degree model was 37.416 and the best MAE for the 3rd degree model was 36.616. For the US region, the best MAE for the 2nd degree model was 38.04 and the best MAE for the 3rd degree model was 37.73. These results were better than those of the previous two models, but again not by much.

The next model I trained was the Gradient Boosting Regression model. Initially I trained Gradient Boosting Regression models using their default hyperparameter settings, and although the results were not production grade, they were much better than those of the previous 3 models. For the European region, the r-squared score was .464 and for the US region the r-squared score was .4804. In light of the fact that using GridSearchCV to find the optimal hyperparameters can take quite a long time, I trained 3 models where I randomly selected the values (just for the European data since the US data was similar in structure). The first model (using n\_estimators = 400, learning\_rate = .1 and max\_depth = 5) gave the following results: coefficient of determination = .5411 and negative mean squared error = -1834.645. The 2nd model (using n\_estimators = 600, learning\_rate = .1, max\_depth = 5) gave the following results: coefficient of determination = .5516, negative negative mean squared error = -1825.47. Finally, the third model (using n\_estimators = 600, learning-rate = .01, max\_depth = 5) gave the following results: coefficient of determination = .489, negative mean squared error = -1979.37. Since the results did not change much after changing the hyperparameters of the European models I only trained one model for the US data.

The final model that I trained was the Random Forest Regression model. For this model I trained models with different values of the ‘max\_depth’ hyperparameter. The results were interesting in that as the values of ‘max\_depth’ increased from 2 to 23, the coefficient of determination scores for the European region converged to roughly .8154, whereas the coefficient of determination for the US were still growing, although the rate of increase did slow down.

Given the correlation information revealed by the heatmap I also trained all the previously mentioned models on the data frame with the ‘private room’ feature removed. The results were equal to or worse than the results where all the available data features were used.

After looking at all the data, I chose the Random Forest Regression model. It is very plausible that the Gradient Boosting Regression model is just as good or even better than the Random Forest Regression, but it was faster to tune just one hyperparameter on the latter than to tune 3 hyperparameters on the former when good results require a great deal of computational time.

This project has two main uses: one is to determine which region is generating the largest profit in Airbnb properties, and the other is to determine which aspects of the properties contribute the most to a profit. Based on the findings, the European region has its prices varying widely between $0 and $70,000 whereas the US region has the bulk of its prices in the range of 0 to $32,000. Therefore, an investor in European Airbnb properties would look to make more revenue. As far as the areas that most related to the price, the investor would want to put most of the investment capital into how many people the property can accommodate, the number of bathrooms on the property, the number of bedrooms and beds, and allowing the renter to rent out the entire unit, instead of just one room, a couple of rooms or a level.