Two Sigma Connect: Rental Listing Inquiries

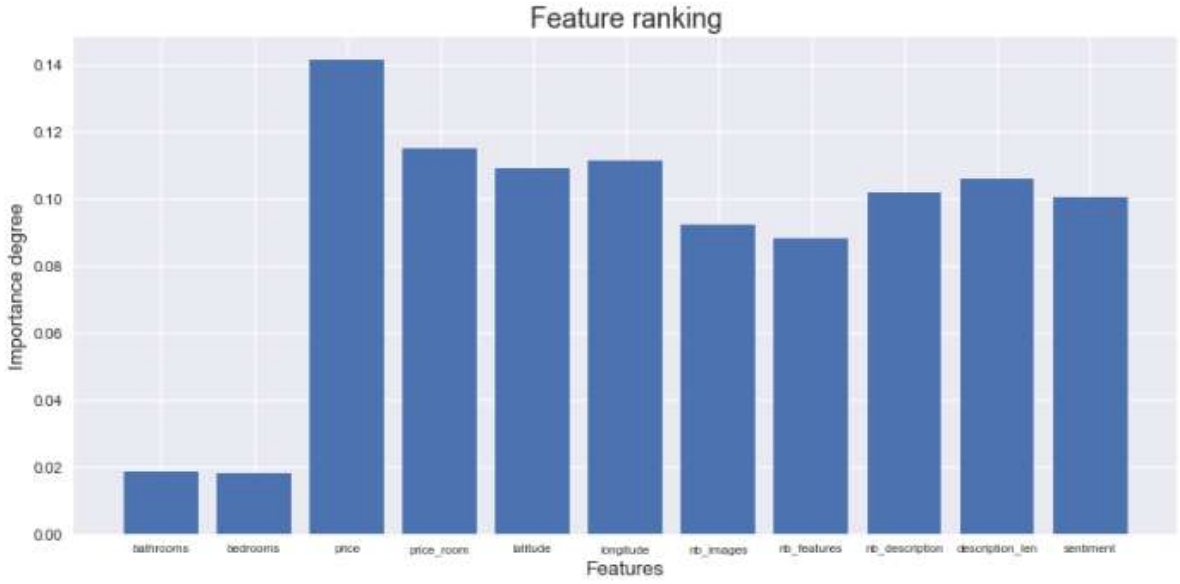
- This Machine learning model predicts the degree of popularity ("low," "medium," or "high") for a rental listing based on listings' attribute such as the number of rooms, location, price, etc.

Random forest classifier:

- Training data contains 44352 rows and 15 columns.
- Based on given data I want to predict the popularity of a rental listing.
- For this exercise I am dealing with numerical features only. For that reason, I converted photos to number of photos, description to length of description and features to number of features.
- I ignored building id, display address, manager id, listing id and street address for now.
- Training data was split into training set and test set into 80%-20% ratio.
- To avoid influence of any variable due big values eg. Price I normalized the data using standard scalar function from scikit learn preprocessing module.
- After all preprocessing and feature scaling was done I applied Radom Forest classifier with random parameters.
- I applied the generated model to test data. I got an accuracy score of 71%. In other words, based on the data we can predict the popularity of a rental listing with 71% confidence.
- With the use of Randomized Search CV algorithm, I tried to tune the hyperparameters and identify the best parameters for this model.
- I applied the best fit parameters back to the model. That increased the model accuracy by 1%.
- Then I applied the same model on Target data. I got an accuracy score of 69%. Which is very close to training accuracy score. That suggests our model is not overfitting the data and is predicting reasonably correct values.

Multiclass Classifier:

- Along with number of photos, number of features, number of words in the description and length of the description I also added price per room and sentiment columns.
- After preprocessing the data, I applied Extra Tree Classifier to identify which parameters influence target variable the most.



- Looks like price (price per room), location (longitude and latitude), description and photos influence the popularity of listing the most.
- I split the data into training and test set and applied multiple machine learning models with default parameters.
- Support vector machine and random forest overfit the data. I found huge difference between training accuracy score and testing accuracy score.
- Then I applied Gradient boost model. This model gave me approximately same accuracy (70%) as random forest model with hyperparameters tuning.

Conclusion:

Price and location influence the popularity of the rental listing the most. If you add long description and photos that helps.

Note:

With the help of powerful machine, I can tune the parameters a little more. With multiple grid search algorithm rounds we can find close to best fit parameters.

Also, I can make use of manager id, created date, address, image data to add more features to the model. From mangers id estimate manager's skill level. From create date find day, month and year. Check if some months has more influence on target variable. Eg. Check interest level in summer vs winter. From image data find out what kind of pictures are these. Eg. High resolution images, kitchen images, sunlight etc.

Github Links –

https://github.com/cricboy007/Springboard-Data-Science-Career-Track/blob/master/CapstoneProject1/Two%20Sigma/Multiclass%20Classifier.ipynb

https://github.com/cricboy007/Springboard-Data-Science-Career-Track/blob/master/CapstoneProject1/Two%20Sigma/Random%20forest%20classifier.ipynb