Introduction to Data Science
CSC59970 | CCNY
Professor Grant M. Long

Final Project
project_finding.md

Abdur Rafey
Dzhonibek Parmankulov
Hasibul Islam

Questions and Tasks:

(1) Data Usage.

(a) What outside data have you appended to the original data set? Why did you choose this data?

- We decided to append restaurant data. We used NYC Open data as a data source, we used restaurant zip code to match the zip code of rental apartment and accordingly attached the restaurant zip code, restaurant name and restaurant grade.

(b) Does the inclusion of this additional data raise any ethical considerations?

- One ethical consideration could be that the zip codes act like a sector divider. Thus, we get an idea of the common rent prices in different areas of New York City, which gives an indication of the social classes of tenants living in those areas.

(2) Data Exploration.

(a) What outliers present issues for your analysis? How have you chosen to handle them? Why?

- Outliers are luxury apartments with very high rent price and cheaply rented apartments.Since we are not allowed to delete the data we didn't remove the outliers.

(b) To what extent do missing values pose a challenge for your analysis? How have you chosen to handle them? Why?

- Missing values affected our prediction model in a negative way because it doesnt let us train with real values completely. Missing values do change the outlook of the conclusion of our analysis if not handled. We decided to use the median values of the other available data for the specific columns that had null values and filled them with those medians. We choose median over other statistics like mean because possible outliers don't have much effect at all on the median, whereas they can skew the mean quite a bit and possibly give a false analytical conclusion. We were planning to train model and predict missing values but didn't have enough time.

(c) Are there any other aspects of the data your exploration shows might be problematic?

- The problem with our model is that, due to the missing data in the training set, we were forced to use medians of the available data to fill in the null values so our final results don't completely reflect what a perfect model would be like, but it is still fairly close to the actual results.

(d) Create at least one visualization that demonstrates the predictive power of your data.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   rent   R-squared (uncentered):                   0.836
Model:                            OLS   Adj. R-squared (uncentered):              0.836
Method:                 Least Squares   F-statistic:                              6487.
Date:                Thu, 05 Dec 2019   Prob (F-statistic):                        0.00
Time:                        22:21:22   Log-Likelihood:                     -1.2528e+05
No. Observations:               14000   AIC:                                  2.506e+05
Df Residuals:                   13989   BIC:                                  2.507e+05
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
bedrooms       145.6283     19.185      7.591      0.000     108.024     183.233
year_built       1.4243      0.091     15.662      0.000       1.246       1.603
bathrooms     1622.6973     36.597     44.340      0.000    1550.963    1694.432
min_to_subway   -0.0023      0.006     -0.355      0.722      -0.015       0.010
size_sqft        2.4718      0.040     61.379      0.000       2.393       2.551
no_fee        -232.1212     34.133     -6.801      0.000    -299.026    -165.216
has_doorman    569.6259     53.495     10.648      0.000     464.768     674.483
addr_zip        -0.3815      0.016    -23.427      0.000      -0.413      -0.350
floor_count     18.6267      1.911      9.749      0.000      14.882      22.372
has_gym        498.2204     51.740      9.629      0.000     396.802     599.638
allows_pets    297.7480     34.079      8.737      0.000     230.950     364.547
==============================================================================
Omnibus:                    13693.453   Durbin-Watson:                   2.023
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          3379523.106
Skew:                           4.200   Prob(JB):                         0.00
Kurtosis:                      78.650   Cond. No.                     4.32e+04
==============================================================================
```

(3) Transformation and Modeling.

(a) Describe 5-10 features you think play the biggest role in your model.

These were the features that played the biggest role in our model

- bedrooms
- bathrooms
- size_sqft
- addr_zip
- floor_count

Number of bedrooms/bathrooms and floor_count had a positive correlation to the rent. Also, the addr_zip is a feature we added to divide our data into sectors so we could see what incomes are most prevalent in which areas.

• How did you create these features?

- We created these features by importing the training dataset.

• How do you know these features are playing key roles?

- These features are playing key roles because as bedrooms or bathrooms increase, the rent goes up, especially size_sqft. Also OLS table give us and idea which variable affects the outcome most.

(b) Describe how you are implementing your model. Why do you think this works well?

- WE are implementing it using features we chose and predicting the target which is rent. Initial training dataset with 12000 rental records and then we predicted 2000 records of test data to check for accuracy/mean squared error.

(c) Describe your methodology for selecting your model. Why do you think this type of model works well?

- Initially we started by using Linear Regression because we knew that we had continuous data to work with so that seemed to be the best option. Later, we tried Random Forest Tree and it gave a much smaller Mean Squared Error. We decided to try Gradient boosting Regressor and it worked in a very similar method and gave a slightly better result thus for submission of test3 rent we predicted using Ensemble Gradient Boosting REgressor. Reason it works well, due to its ability to navigate through more complex and larger datasets.

(4) Metrics, Validation, and Evaluation.
(a) How well do you think you model will perform on the hold out test set? How do you know? 3

- Linear Regression: Mean Squared Error using Linear Regression: 3313817.143868871

- Random Forest: Mean Squared Error using Random Forest Regressor: 1883630.165411068
- Gradient Boosting: Mean Squared Error using GradientBoostingRegressor: 1720228.7848562498

We think our model would perform fairly well on the hold out test set. Based on the results for each model we tried, we got closer to a perfect score of 1, so our model was pretty consistent and our mean squared error decreased with each trial and gradient boosting gave the best results.

(b) Is your model useful? Why or why not?

- Our model is useful as its variables correlate with each other at a fairly consistent rate based on the R-squared value.

(c) Are there any special cases in which your model works particularly well or particularly poorly?

- Our model works well on rentals with average rent price, but it is poorer for luxury rentals.

(d) Create at least one visualization that demonstrates the predictive power of your model.

- Linear Regression: Explained variance regression score function Best possible score is 1.0, lower values are worse. 0.5559738486087318
    - R^2 (coefficient of determination) regression score function: 0.5556196902015677
- Random Forest: Explained variance regression score function Best possible score is 1.0, lower values are worse. 0.7474068681056247
    - R^2 (coefficient of determination) regression score function: 0.7474066551922682
- Gradient Boosting: Explained variance regression score function Best possible score is 1.0, lower values are worse. 0.7693223390436726
    - R^2 (coefficient of determination) regression score function: 0.7693186536399759

(5) Conclusion

(a) How would you use this model?

- After analyzing our models accuracy, We can use this model to predict the rent for test3. We can use this model to determine if one sector of NYC would be best suited to your needs depending on the contributing environmental factors of that apartment location. For example, a higher rent based on better restaurants or close train stations can be seen differently from higher rent based on number of bedrooms/bathrooms.

(b) If you could have additional modeling features, what would they be?

- Since we are mostly taking into account environmental factors as the additional data for our experiment, we could also choose to use number of playgrounds near apartment or best rated schools near apartment if the tenants have children.

(c) Would you rather have more data, or more features?

- We would rather have more data because sometimes more data can give us an opportunity to test out more features and see the differences in the results if any. Also, having more data can also bring the accuracy closer to the expected value so using more features after adding extra data into a model could produce better results. This is

similar to the law of large numbers. As you perform more tests you get closer to an expected value.

Note: We have graphs and Restaurant finder feature based on external dataset, where you enter the zip code it tells you the good restaurants in your area.