Predicting New York city residential real estate prices from local venues

Problem definition

Real estate risk management needs to be focused on a problem of predicting the market value of properties as they are unique and difficult to actively manage, i.e. buy and sell. The value of real estate assets should be an integral function of some internal and external features. In short, internal features are endogenous to the property, for example design of windows or a number of rooms. External features are exogenous, for example surrounding shops or transportation hubs.

The proposed study aims to model the market value of real estate assets using exogenous features while controlling for endogenous features. In other words, given a specific property type with standardized endogenous features, what nearby venues have the most effect on the value? The input to this model could be a relative frequency of certain types of venues mapped to a zip code. The output could be the zip code average price and monthly rent of a 2-bedroom residential unit.

Numerous stakeholders, including investors, developers, local authorities and everyday citizens, will benefit from an improved forecast that takes into account the effect of local venues. This is especially import at the moment because markets are just beginning to digest the economic consequences of massive shutdowns due to pandemic related social distancing. Ideally, the study will produce a list of local businesses that tend to drive the value of nearby residential properties. Stabilizing these crucial businesses should help improve home values, and thus support mortgage collaterals and local tax revenues.

Data sources

In order to effectively merge multiple data sets it is proposed to use zip code level granularity. This level allows for geospatial mapping and, if needed, aggregates up to neighborhoods and boroughs reasonably well. It is also important to focus on highly urbanized areas as such areas are expected to have more local businesses per zip code. Therefore, the ideal training data will have local businesses and average home values by zip

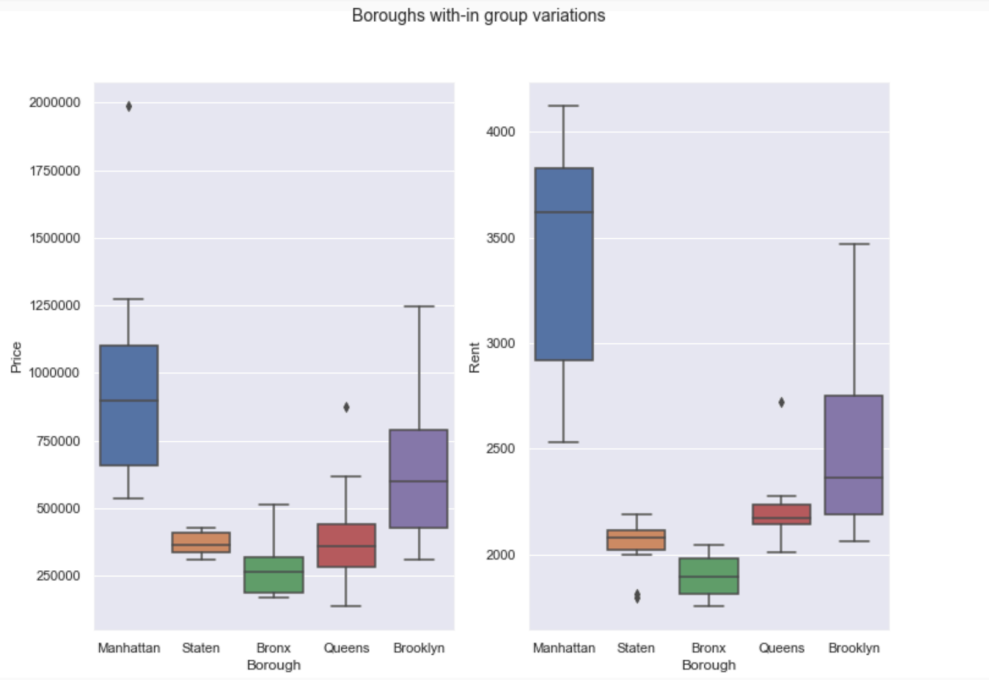
code for the following five New York boroughs: Manhattan, Bronx, Brooklyn, Queens, and Staten.

Most recent postal zip codes for New York city and geospatial mapping will be used as of end of May 2020, source: https://worldpostalcode.com/united-states/new-york/new-york-city. Zillow databases will be used to obtain monthly price and rent estimates for a 2-bedroom residential unit as of end of November 2019, source: https://www.zillow.com/research/data/. Foursquare APIs will be used to obtain information about local venues as of end of November 2019, source: https://api.foursquare.com/v2/venues/. The study can be reproduced for the following months to capture the effect of shutdowns.

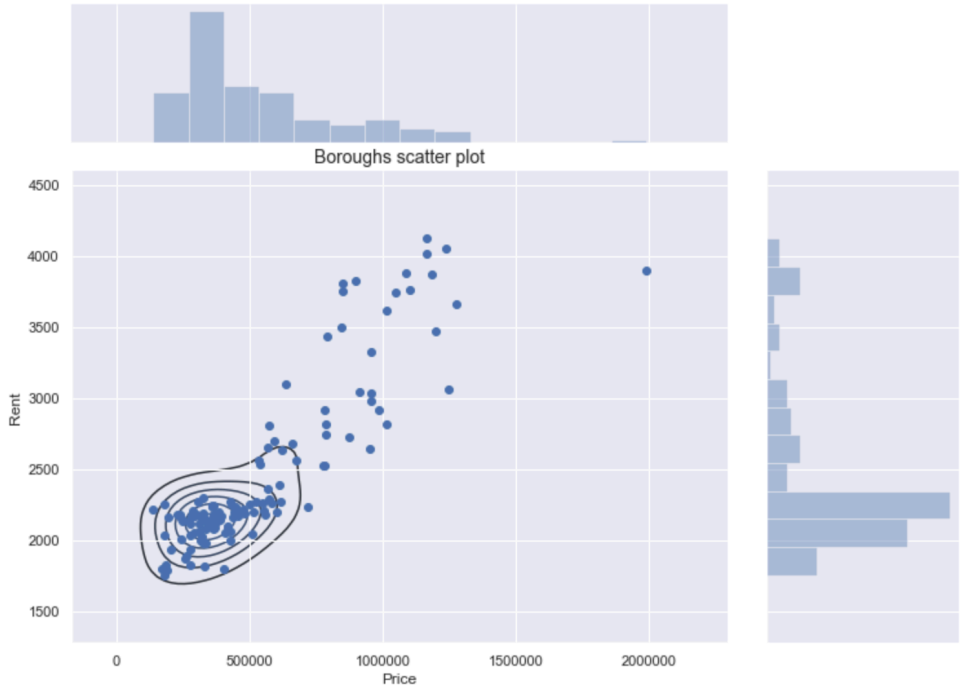
Benchmarking of the results can be performed against other alternative publicly available sources, such as https://www.realtor.com/estimates/ and https://www.neighborhoodscout.com.

Methodology

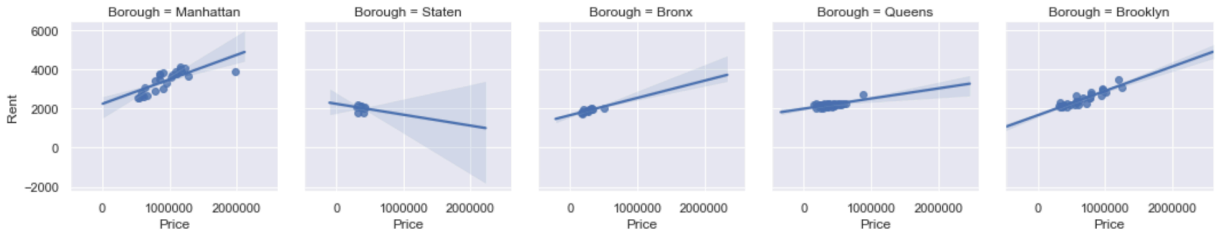
The original data set is not homogeneous at all. Box plots show that Manhattan and Brooklyn have the most variance what may adversely affect the model overall predictive power.



Histograms on a scatter plot of all 128 zip codes help identify the most reliable area that is inclosed in several contours. Corresponding statistics are supported by the most observations.



Regression plots also confirm the proposition that boroughs are likely different from each other in terms of what the drivers of home values are as slope coefficients are visibly different.



Taking into account all these visual results, one is proceeding with caution as it may be necessary to break down the observations into smaller clusters with less with-in group variations. Regrettably, this will reduce the number of observations. Some zip codes have more than 128 different types of venues in close proximity and thus any two clusters will have more features than observations.

Linear least squares regression model promises to produce the most meaningful results, but it does not work well with sparse data. Getting dummies from discrete features produces very sparse data and regularization may be needed. Another remedy, subset selection, may become extremely time consuming if the number of predictors is large. Also, the larger the search space, the greater is the chance of overfitting the train set and producing a really good model simply by random chance.

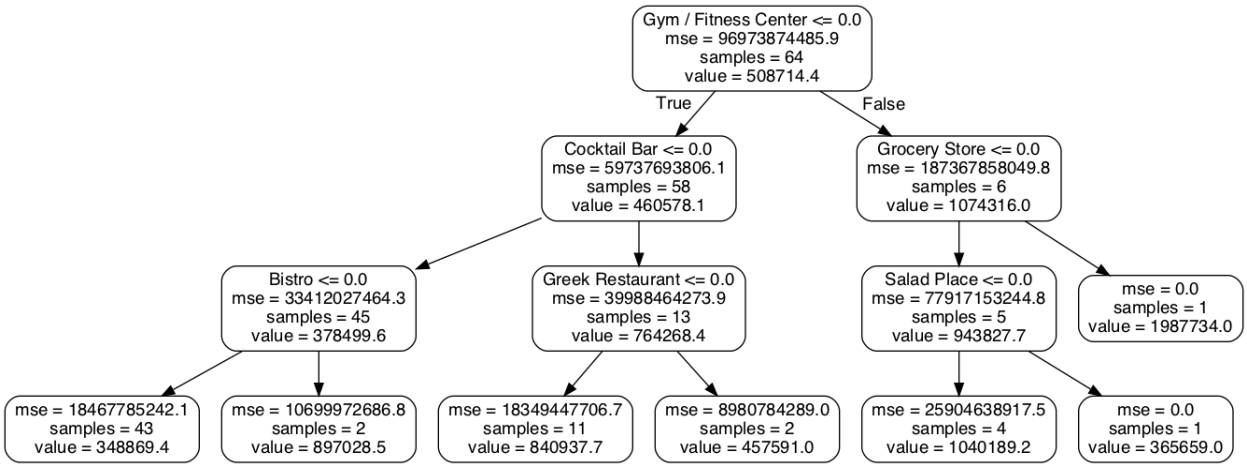
Therefore, looking at all the combinations of predictors is neither practical not beneficial. Instead one could focus on incremental improvement with every new predictor by using forward stepwise selection and adding one more predictor to the nested model and then use cross-validation to choose among the nested models. Regrettably, in absence of previous research, it is not clear what the good starting feature shall be and how it can be selected.

Another alternative, that does not require fitting coefficients with the least squares and making initial assumptions regarding feature importance, would be using shrinkage methods applied to scaled predictors. For example, the Lasso method produces sparse models as it effectively performs subset selection by shrinking predictors weights down to zeros. The downside is the loss of interpretability as the method does provide an explanation why certain features were dropped. Moreover, it is not clear how to interpret scaled predictors.

Given all these limitations, tree-based models seem to be the only good choice. Multiple trees can be combined in ensemble for better predicting power. For example, it is expected that the first split may be on a venue such as coffee shop, second split may be on grocery stores or restaurants, following splits may be on parks, gyms, doctor or business offices. Tree-based algorithms seem to be preferred choice for this task, as they should be able to provide feature importance measure based on amount of post-split regression RSS decrease. Important, that one does not have to make any initial assumptions regarding venues importance.

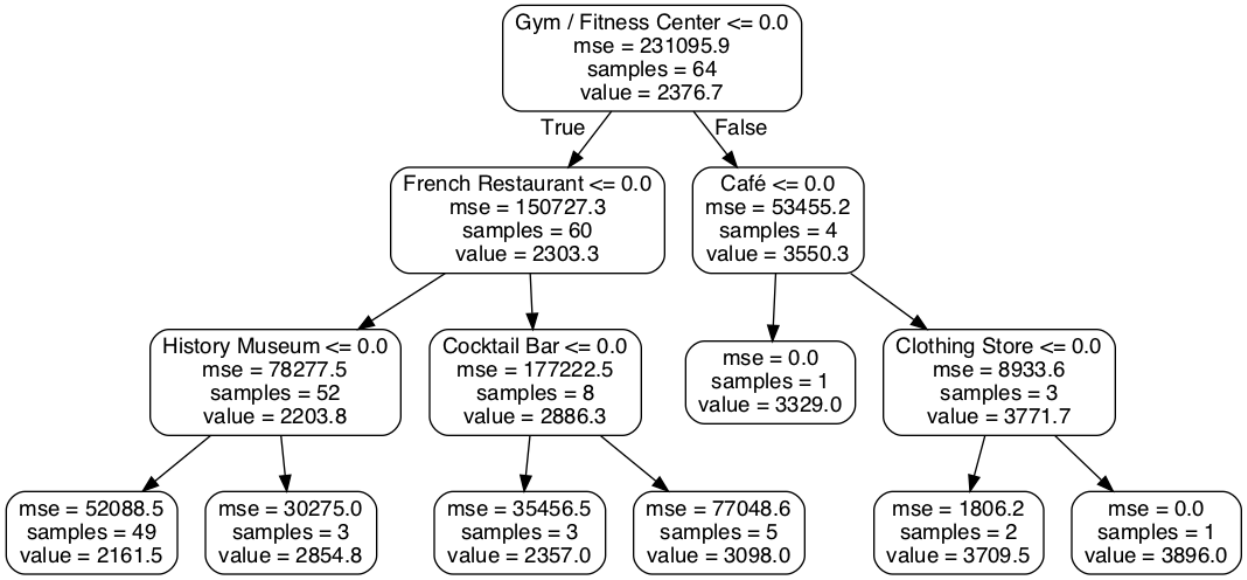
Moreover, although not always perfect, trees have good interpretability. They are easy to explain and mimic human decision-making. Finally, trees can be displayed graphically and can easily handle numerous features without scaling predictors. Below is an example of one of the trees that was

developed in random forest price regression, technique of averaging predictions of multiple trees grown by randomly dropping features.



Interpretation of this particular tree is very straightforward. The first split occurred for feature “Gym / Fitness Center” at the root of the tree for 64 samples with average price of $508,714. Properties that did not have gym are coded “False” and are further split on presence of “Grocery Store”. Similarly, those with “True” in the first node, are split on “Cocktail Bar”. Important, this is just on of possible trees in the random forest and overall feature importance is only assessed once all the trees are averaged.

Below is an example of tree from rent regression that follows exactly the same logic. In this case, “value” refers to the average monthly rent associated with a specific zip code.



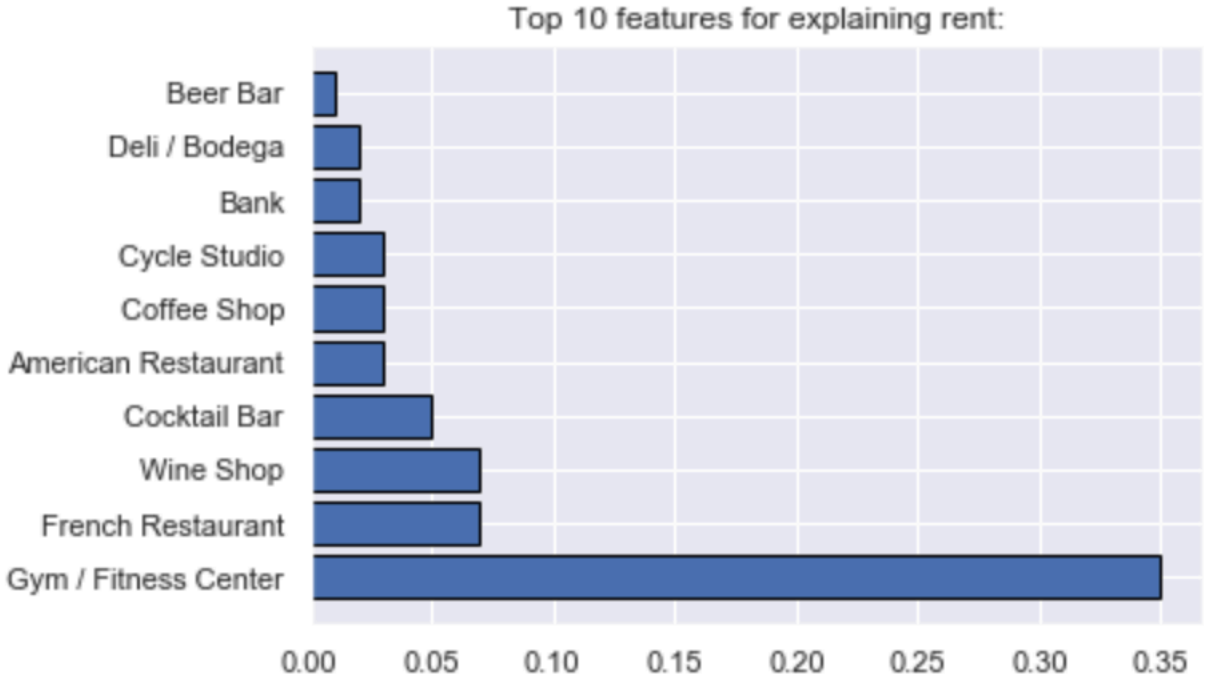
Results

Two random forest regression models were developed for targets “Price” and “Rent”. Models were trained and tested on 80/20 split of original 128 zip codes. Overall 10,000 random estimations were performed with a maximum depth of three for both regressions.

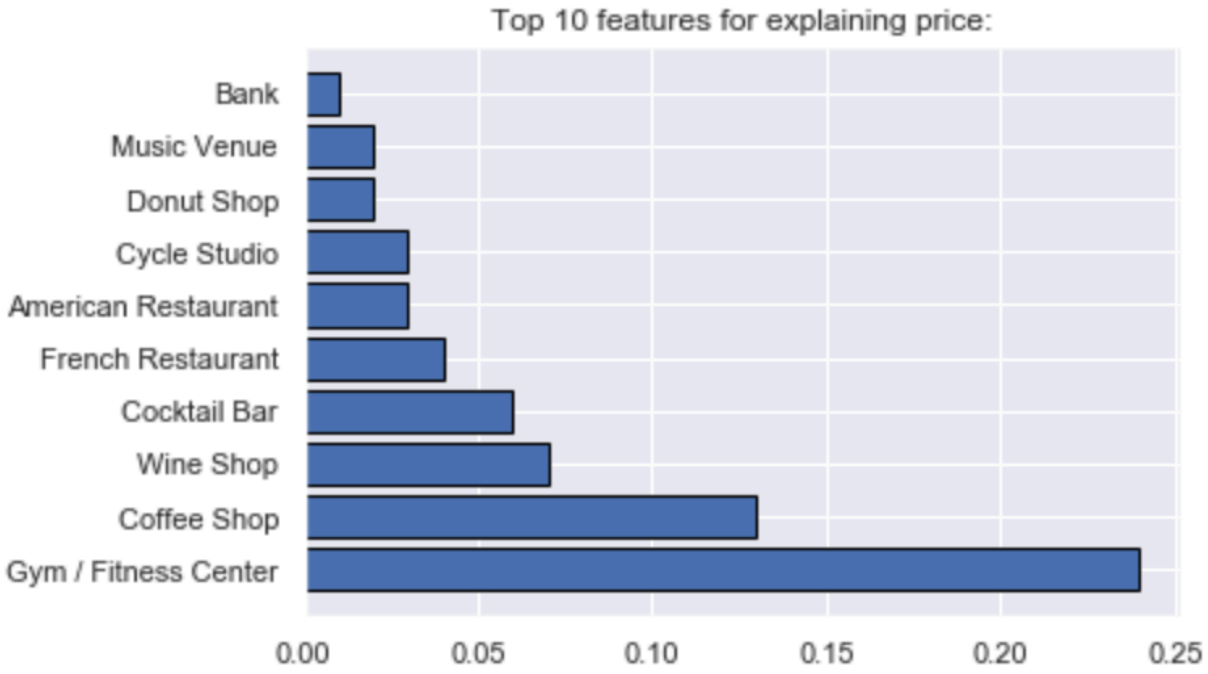
Performance was assessed based on R-squared and explained variance. Moreover, mean absolute error was used to calculate the dollar value of expected error.

Price predictions seem to be a little better, but noteworthy is that both models have very similar important features. As shown below, presence of nearby “Gym / Fitness Center” explains easily from 20% to 30% of home value. This dominant feature is followed by “Coffee Shop”, “Wine Shop”, “Cocktail Bar” , “French” or “American Restaurant”. Interestingly, that “Cycle Studio” and “Bank” also appears in both models, but “Music Venue” seems to be more important for predicting “Price” rather than “Rent”.

```
Performance measures for rent model
R2 score:  0.6
Explained variance:  0.61
Mean absolute error: 0.1
On average predictions are off by 260.0 $
```



```
Performance measures for price model
R2 score:  0.73
Explained variance:  0.73
Mean absolute error: 0.21
On average predictions are off by 106823.0 $
```



Discussion

From the findings of this study it appears that urban residents do put a lot of value on having fitness venues in close proximity. Moreover, coffee and dining options are important as well.

Generally speaking, buyers are willing to pay a premium for desirable exogenous features and thus home values increase as more or better features are added. Consequently, as more of such features are eliminated, value will go down. Prolonged shutdowns of local fitness or dining venues will likely negatively affect the value of residential properties.

Having identified promising features, allows utilizing least squares regression that can produce slope coefficients so that the exact dollar value of change in feature can be computed.

Conclusion and next steps

Clustering is very important as observations are not homogeneous. Much more observations (zip codes) are needed to accommodate sparse data. Possible solution would be to add urban zip codes from other big cities and repeat the analysis.