Predicting Real Estate Value in Manhattan
Midterm Report

Antong Su(as3657), Jiahui Yi(jy764), Xiaodan Fang(xf72)

October 28, 2017

1. **Project Goal & Applications**
   The goal of our project is to predict the full market value, including value of land, buildings and other resources of different tax lots in Manhattan, New York. The results of our project could contribute significantly to the following areas:

   - Predicting land prices during land valuations or auctions
   - Estimating values for companies' real estate assets
   - Providing references and suggestions for city planning

2. **Dataset Description**
   The dataset we use is part of the PLUTO dataset, provided by NYC city planning. It provides comprehensive geographic and land use information of Manhattan, with more than 42k data entries and 84 features. The 84 features include zipcode, number of buildings, year of construction, lot area and so on. Despite the considerable size of data entries, many data entries contain empty or incorrect fields that we need to remove. Some of the corrupted and missing data field examples are shown below:

   - Missing field address and unspecific address
   - Zipcodes pointing to areas outside Manhattan
   - Missing year of construction, or having year of alternation earlier than construction
   - Zero values, null values for some variables

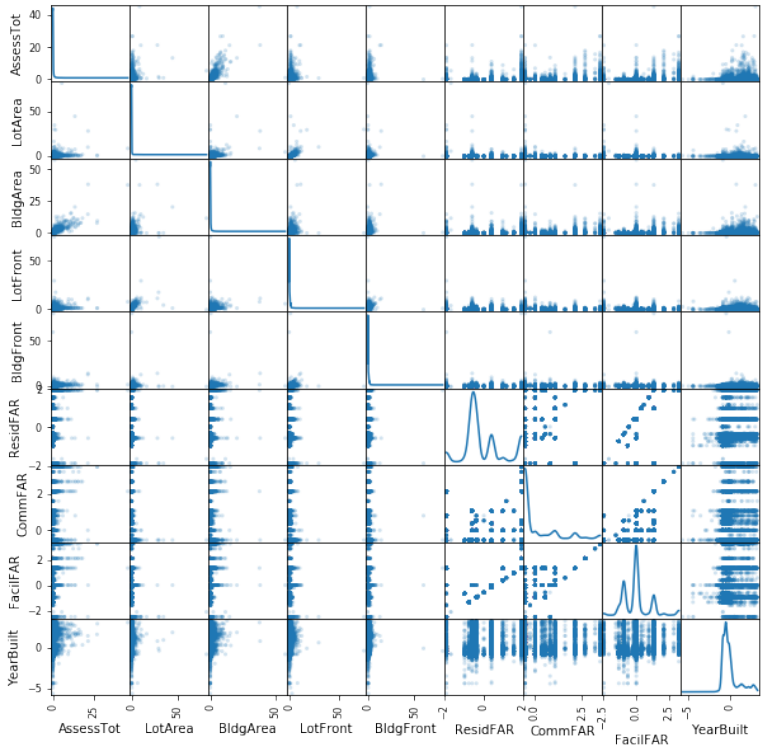
   After applying certain criteria to remove the entries with corrupted or missing data field, we get 16028 valid data entries.

3. **Exploratory Data Analysis & Feature Selection**
   Among 84 data features, some insignificant variables need to be eliminated. Thus, we screen all the variables and decide to first choose the following 13 variables in our analysis:

   - Nominal variables: Zipcode, LandUse, OwnerType, IfAlter
   - Continuous variables: LotArea, BldgArea, LotFront, BldgFront, ResidFAR, CommFAR, FacilFAR
   - Discrete variables: YearBuilt, YearAlter

   We plot the pairwise scatter plot of the 13 variables in Figure 1. We first eliminate variables that do not seem to have strong correlation with AssessTot, our dependent variable. Thus, we delete BuildingFront and LotFront. Secondly, we try to avoid over fitting by getting rid of variables that have strong correlations to each other. From Figure 1, it is obvious that FacilFAR is strongly correlated with CommFAR and ResiFAR. Thus we choose to delete it. Finally, we keep 10 varibles for further regression analysis.For nominal variables, we need to transform them to multiple dummy variables in order to construct the model. Here are the details of each variable we are going to use:

Figure 1: Pairwise scatter plot of potential variables



- Zipcode: The zipcode of the location. For the data transformation, we decide to split Manhattan into 10 zones by referring to NYC Neighborhood ZIP Code Definition. Therefore, the zipcode variable is mapped to 10 dummy variables Zones[1,...,10].
- LandUse: A code standing for the use of land. There are 9 types (e.g. residence, commercial, manufacturing) of land use. We map this variable to 9 dummy variables LandUse[1,...,9].
- OwnerType: Ownership of the land. There are 6 types (e.g. city ownership, private ownership, mixed ownership) of ownership. We map this variable to 6 dummy variables OwnType[1,...,6].
- LotArea: Total area of the land in sq feet.
- BldgArea: Total gross area in square feet.
- CommFAR: The Maximum Allowable Commercial Floor Area Ratio (FAR). Maximum 15%.
- ResiFAR: The Maximum Allowable Residential Floor Area Ratio (FAR). Maximum 10%.
- YearBuilt: The year construction of the building was completed.
- YearAlter: The year of the most recent alteration of the building
- IfAlter: If the building has been altered, the variable will be one; if not, 0.

4. **Preliminary Analysis**

For preliminary analysis, we first did a simple regression analysis on AssessTot against all selected continuous variables. In order to make our predicted w coefficients on the same scale, we scale all variables by substracting the mean and devided by the the standard deviation. Then we fit the model:

X = AssessTot, LotArea, BldgArea, ResidFAR, CommFAR, YearBuilt

Y = AssessTot

The result for regression coefficients is shown in Table 1:

Table 1: Coefficients For Regression Model 1

LotArea	BldgArea	ResidFAR	CommFAR	YearBuilt
-0.1436	1.0380	0.0031	0.0705	0.0260

The mean-squared error on training set is 0.36, on test set is 0.65. In comparison, the variance of our training set is 1.52, variance of test set is 0.08. The R2 score reported by sklearn is -6.42. This means our model is underfitting and the prediction error is relatively high. It is reasonable because we have not included any categorical variables yet. But we can still get some insights from the predicted coefficients.
We can see AssessTot decreases with LotArea, and increases greatly with BldgArea, which is reasonable because the total land price is the sum of all property values on this land. Lands with larger building areas like skyscrapers are usually located in city center and tend to have higher total prices than lands having larger area but no buildings on them, like parks, squares, etc. For ResidFAR and CommFAR, we can see AssessTot grows with both variables, but CommFAR has a stronger impact, which means the higher the maximum allowable commercial floor ratio, the higher the total land price. For YearBuilt, the positive coefficient means lands with newly built buildings tend to have higher total price.
In our second model, we add dummy variables IfAlter and Zones[1,...,10] transformed from ZipCode, and do the regression again. The coefficients are shown in Table 2 below:

Table 2: Coefficients For Regression Model 2

LotArea	BldgArea	ResidFAR	CommFAR	YearBuilt
-0.14420	1.0442	-0.0158	0.0929	0.0280
IfAlter	Central Harlem	Chelsea and Clinton	East Harlem	Gramercy Park and Murray Hill
0.0124	0.0510	0.0366	-0.0016	0.1112
Greenwich Village and Soho	Lower Manhattan	Lower East Side	Upper East Side	Upper West Side
-0.0324	-0.3352	0.0005	0.0868	0.0831
Inwood and Washington Heights				
0.0000				

The mean-squared error on training set is 0.35, on test set is 0.66, the R2 score reported by sklearn is -6.52. The positive coefficient for IfAlter suggests building alternations increase total real estate value. From the coefficients for zones, we can roughly observe a relationship between location and total tax plot market value. However, from the regression metrics, it seems our model still needs to be improved by adding more features and incorporating the categorical variables in a more intelligent way.

5. **Upcoming Plans**

   (a) Explore additional features from PLUTO dataset and other data sources, e.g. crime rate of an area specified by zipcode.

   (b) Try linear regression models with different regularization methods(ex. ridge, lasso), and use cross-validation to compare different models.

   (c) Try decision tree model to incorporate categorical variables in a smarter way.

   (d) Adapt our model to different boroughs in NYC and make comparisons about model effectiveness.