Cornell University
School of Operations Research and Information Engineering

Predicting Real Estate Values in Manhattan

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

Abstract: The New York City Department of City Planning provides Primary Land Use Tax Lot Output data (PLUTO) of 84 fields about New York City extensive land use. This project analyzed meaningful data fields of real estates in Manhattan. Big messy data analyzing techniques were applied to provide insights on how these data fields impact the full market values of real estates. Models used in this project include quadratic and Huber loss models with l_1, l_2, smooth regularizers and without regularizers, and a two stage linear regression model. The result shows that buildings with higher lot frontage, number of floors, commercial area, residential area and retail area tend to have higher total market value. Besides, the location and type of land use also affect land value substantially. .

1. Introduction

1.1 Background. The aim of this project is to predict market value of real estates in Manhattan, New York. Understanding this dataset is meaningful, because we can use our analysis to predict building prices for auctions. Moreover, we can estimate asset values for real estate companies. Furthermore, this analysis can provide suggestions for city planning for New York state.

1.2 Data Description. The data we used for this project was found on NYC government open data website. The dataset includes more than 42,000 entries and 84 variables, including year built, Zip Code, address and more. Entries with empty or incorrect fields were removed, including but not limited to entries with unspecified address, null ZIP Code, and missing construction time.

1.3 Cleaning & Adding Features. After careful analysis of all variables, we chose 27 features as variables for preliminary analysis. All variables are from PLUTO dataset and some additional data sources. We combined features from other sources into our data set. Here is the summary of the variables we used for preliminary analysis and explanations of what each variable means:

Dependent Variables: AssessTot (Total Assessed Value)

Qualitative variables: Zipcode, LandUse (10 types of landuse), OwnerType (5 types of owner types) and IfAlter (1 or 0)

Numerical variables: LotArea (total area of the tax lot), LotFront (tax lot's frontage), BldgArea (building area of the tax lot), NumBldgs (the number of buildings on the tax lot), NumFloors (the number of floors of the tallest building on the tax lot), ComArea (floor area for commercial use), ResArea (floor area for residential use), OfficeArea (floor area for office use), RetailArea (floor area for retail use), GarageArea (floor area for garage use), StrgeArea (floor area for storage and loft use), FactryArea (floor area for factory use), OtherArea (floor area for other than Residential, Office, Retail, Garage, Storage, Loft or Factory use), ResidFAR (maximum allowable residential floor area ratio), CommFAR (maximum allowable commercial floor area ratio), FacilFAR (maximum allowable facility floor area ratio) and YearBuilt

Numerical variables from additional source: MedIncome (median household income), PopDensity (Population Density), NumofHouses (number of houses), EduRate (education rate), UnemployedRate

2 Visualization

2.1 Assessed Total Value Visualization. **Figure 2.1** shows the histogram of the dependent variable, Assessed Total Value without any transformation. The distribution is right skewed. Moreover, a lot of outliers are existing.

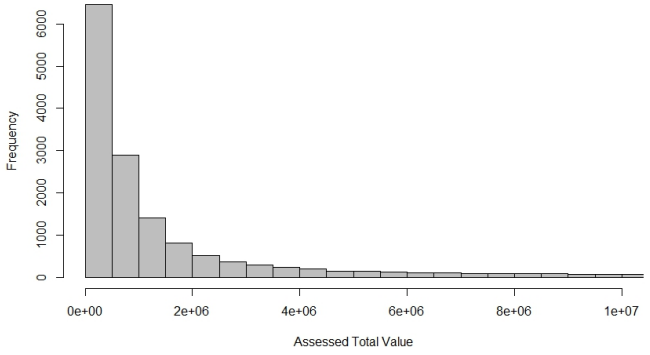

Figure 2.1 Histogram of Assessed Total Value

To make the dependent variable easier to work with, we perform natural log transformation. After the transformation, we could see in **Figure 2.2**, the data becomes more normally distributed. We could also see that outliers have much less effect now.

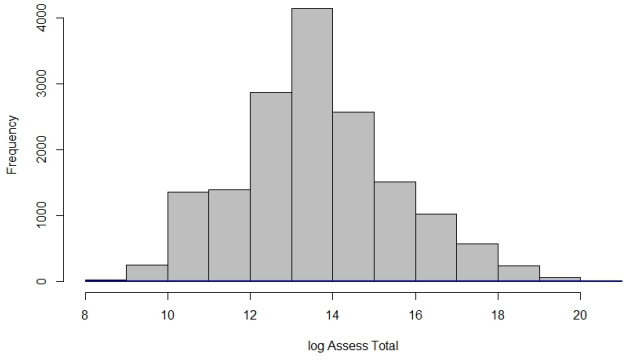

Figure 2.2 Histogram of Assessed Total Value after Log Transformation

2.2 Correlation Visualization. We first conducted correlation visualization to help us better understand the data and specifically, understand what features are significant, and which features have stronger correlation with Total Assessed Value. In order to do so, we split our variables into qualitative and numerical variables. As for numerical variables, we use a correlation hot plot and scatterplot to show what features have the greatest effect on Total Assessed Value.

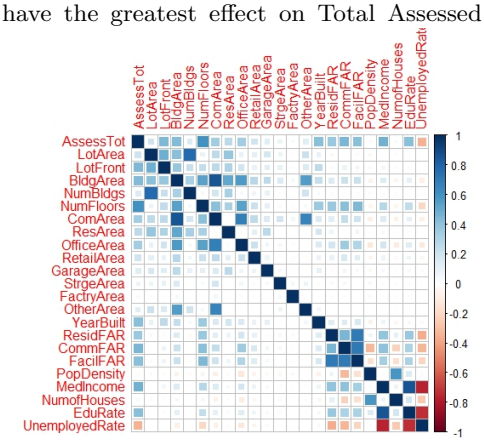

Figure 2.3 Variables Correlation Plot

As we can see in the plot, the dependent variable, AssessTot depends heavily on BldgArea, LotFront, NumFloors, YearBuilt, ResidFAR, CommFAR, FacilFAR, MedIncome, EduRate and UnemploymentRate. For the other variables (LotArea, PopDensity, NumofHouses and MarriedRate) that are not strongly correlated to AssessTot, we decided not to use them for modeling. Variable BldgArea is the sum of ComArea, ResArea, OfficeArea, RetailArea, GarageArea, StrgeArea, FactryArea and OtherArea, so we decided to use ComArea, ResArea, OfficeArea, RetailArea, GarageArea, StrgeArea, FactryArea and OtherArea, all of which can represent BldgArea. Besides that, there is strong correlation among ResidFAR, CommFAR and FacilFAR. Since AssessTot depends most strongly on CommFAR, we select it as one of the independent variables out of the three.

2.3 Numerical Variables Visualization. From **Section 2.2**, we already had some basic understanding of how our numerical variables relate to the dependent variable, AssessTot. Drawing scatter plot on one independent variable with dependent variable can help us further understand how strongly two variables are associated.

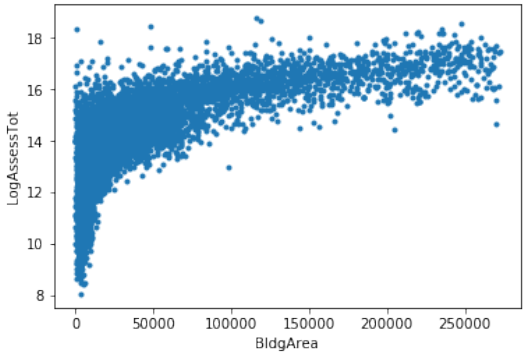

Figure 2.4 Building Area vs Total Assessed Value

Figure 2.4 shows strong correlation between independent variable, Building Area dependent variable, Total Assessed Value. The similar pattern can also be discovered from most of the variables we selected.

2.4 Qualitative Variables Visualization. For the qualitative variables, we use boxplots to view their influence on AssessTot. As Manhattan has around 50 different zipcodes, it is hard to view each zipcode as a specific area. Instead, we divide Manhattan into 10 zones and fit each zipcode into a zone by referring to NYC government website. As we can see in **Figure 2.5**, different zones have significant price differences. Lower Manhattan has the highest assessed value. While Central Harlem, East Harlem and WAHI (Washington Heights and Inwood) have the lowest assessed values.

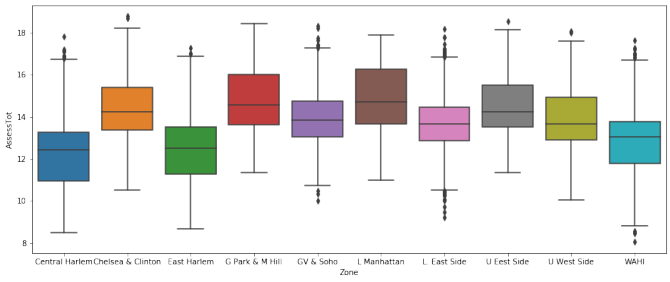

Figure 2.5 Zones vs Total Assessed Value

CODES	DECODES
Central Harlem	Central Harlem
Chelsea & Clinton	Chelsea and Hell's Kitchen
East Harlem	East Harlem
G Park & M Hill	Gramercy Park and Murray Hill
GV & Soho	Greenwich Village and Soho
L Manhattan	Lower Manhattan
L East Side	Lower East Side
U East Side	Upper East Side
U West Side	Upper West Side
WAHI	Inwood and Washington Heights

Table 1. Manhattan Zone Dictionary

As we look at the boxplot of Landuse VS Total Assessed Value in **Figure 2.6** and Owner Type vs Total Assessed Value in **Figure 2.8**, we can see the similar pattern as what **Figure 2.5** shows. Traditional residential properties (one & two family and multi family walk-up buildings) tend to have the lowest assessed value. Mixed residential and commercial buildings have plenty of outliers. When we look at **Figure 2.8**, we can see private ownership buildings also have plenty of outliers.

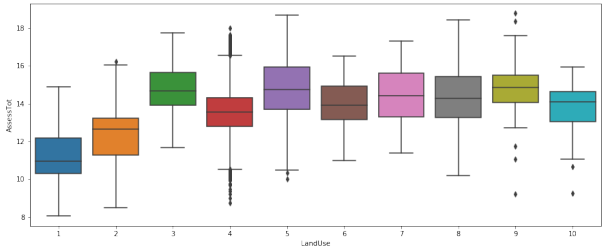

Figure 2.6 LandUse vs Total Assessed Value

CODES	DECODES
01	One &Two Family Buildings
02	Multi-Family Walk-Up Buildings
03	Multi-Family Elevator Buildings
04	Mixed Residential & Commercial Buildings
05	Commercial & Office Buildings
06	Industrial & Manufacturing
07	Transportation & Utility
08	Public Facilities & Institutions
09	Open Space & Outdoor Recreation
10	Parking Facilities

Figure 2.7 LandUse Dictionary

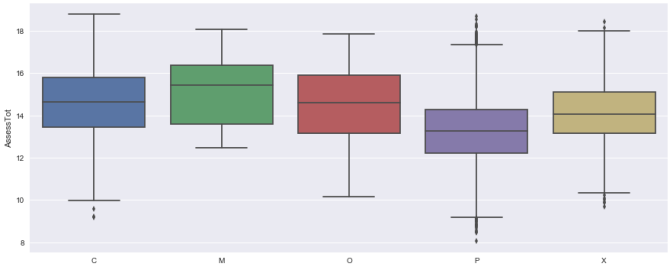

Figure 2.8 Owner Type vs Total Assessed Value

CODES	DECODES
C	City Ownership
M	Mixed City & Private Ownership
O	Other – Public Authority, State or Federal Ownership
P	Private Ownership – Either the tax lot has started an "in rem" action or it was once city owned.
X	Mixed (Excludes property with a C, M, O, or P ownership code). Fully tax exempt property that could be owned by the city, state, or federal government; a public authority; or a private institution

Figure 2.9 Owner Type Dictionary

In order to handle these qualitative variables, we transform them to dummy variables. For example, since we have 5 owner types, the variable "Owner Type" will be split into 5 variables, with only one variable being 1 instead of 0 to show what type of owner the tax lot has.

3. Model

3.1 l_2 Loss without Regularization. In the first step of fitting model, we created linear model by using quadratic loss without adding any regularizer. Later, we used this model as a basic model to compare with other models. The objective function is :

$$\text{minimize } \sum_{i=1}^{n}(y_i - w^T x_i)^2$$

3.2 l_2 Loss with l_1 Regularization. l_1 regularization gives a sparse model which can be used to quickly decide which factors influence the prediction of dependent variable most. Therefore, our first attempt with regularization was to use l_1 regularizer. The objective function is:

$$\text{minimize } \sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda \sum_{i=1}^{n}|w|$$

3.3 l_2 Loss with l_2 Regularization. Although l_1 regression gives a sparse model for quick prediction, it usually has higher test error because of higher bias. To reduce test error, we used l_2 regularization in this model. The objective function is:

$$\text{minimize } \sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda||w||_2^2$$

3.4 l_2 Loss with Smooth Regularization. In previous models, we observed that there are differences among the coefficients of qualitative variables. As a result, two real estates sharing very similar properties but one qualitative property (e.g. two similar building are only different in their zones) can have very huge predicting difference even though their actual assessed values are very close. In order to reduce such dramatic changes of coefficients of qualitative variables, we apply smooth regularizer to the original l_2 loss function. The objective function is shown below:

$$\text{minimize } \sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda||Dw||^2$$

where $D \in \mathbf{R}^{(d-1)*d}$ is the first order difference operator

$$D_{ij} = \begin{cases} 1, & \text{if } j = i \\ -1, & \text{if } j = i+1 \\ 0 & \text{else} \end{cases}$$

3.5 Huber Loss with l_2 Regularization. Another observation we had in the previous models is that there are some outliers that really skew the model considerably. In order to reduce the effect of outliers, we used a combination of l_2 loss and l_1 loss function, which is huber loss. Huber loss punishes small error by using l_2 loss and large error by using l_1 loss, therefore minimizing the effect of outliers with large error. The objective function is shown below:

$$\text{minimize } \sum_{i=1}^{n} \mathbf{huber}(y_i - w^T x_i) + \lambda||w||^2$$

where we define the huber function

$$\mathbf{huber}(z) = \begin{cases} \frac{1}{2}z^2, & \text{if } |z| \le 1 \\ |z| - \frac{1}{2}, & \text{if } |z| > 1 \end{cases}$$

3.6 Two Stage model with Huber Loss and l_2 Regularization. In this mode, we proposed a two stage method to fit dependent variable that cannot be fitted very well by a simple linear model. The method is described as follows. On the training set, We perform three linear regressions with Huber loss and l_2 regularization. The first one is fitted using all the training samples, which gives us the coefficient w. The second one is fitted using training samples with y values below a threshold y_t, which gives us w_1. And the third one is fitted using samples with y values above y_t, which gives us w_2.

$$w = argmin \sum_{i=1}^{n} \mathbf{huber}(y_i - w^T x_i) + \lambda||w||^2$$
$$w_1 = argmin \sum_{y_i \in \{y|y<=y_t\}} \mathbf{huber}(y_i - w^T x_i) + \lambda||w||^2$$
$$w_2 = argmin \sum_{y_i \in \{y|y>y_t\}} \mathbf{huber}(y_i - w^T x_i) + \lambda||w||^2$$

Then prediction is done in two stages. In the first stage, we predict the value using coefficient w. Then in the second stage, based on whether the prediction is below or above y_t, we use w_1 or w_2 to make the final prediction.

4. Numerical Experiment

4.1 Data Splitting and Preprocessing. We randomly split our data into 80% training set and 20% test set. In the following, cross validation is performed on our training set to select the best model parameters, and both training set error and test set error are reported.

For data preprocessing, we first removed the outliers in our dependent and numerical independent variables using the Interquartile Range(IQR) method. A value x is defined to be an outlier if $x < LB$ or $x > UB$ where

$$LB = Q1 - (Q3 - Q1) * \alpha$$
$$UB = Q3 + (Q3 - Q1) * \alpha$$

$Q1, Q3$ are the first and third quartiles and α is a paramter selected to be 10. Then we scaled our numerical independent variables to $[0, 1]$ using the min-max method

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

By doing so, the regression results of coefficients will be roughly on the same scale.

4.2 l_2 Loss without Regularization. We first performed l_2 loss linear regression without regularization on our data. The regression result of coefficients is shown in the graph below.

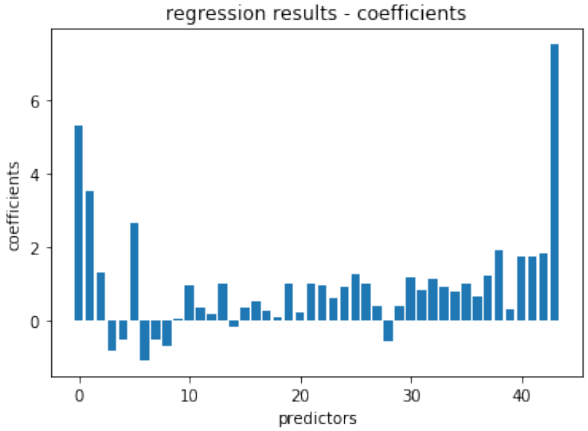

Figure 4.1 l_2 Loss without Regularization Coefficient Plot

Among all numerical independent variables, those having coefficients equal or larger than 1 are listed in the table below.

Variable	Coefficient
LotFront	5.32
NumFloors	3.51
ComArea	1.29
RetailArea	2.66
MedIncome	1.00
Constant	7.53

Table 2. Coefficients of some numerical variables

The result tells us that real estates with larger frontage area, commercial area, retail area, more floors and located in a higher income area tend to have higher total market value. It is not difficult to understand why higher buildings located in high income areas have larger market value. And also buildings with larger commercial area or retail area are usually shopping centers or hotels located in bustling regions. But it is really interesting that frontage area can have such big influence on total market value. This is partly because larger frontage area also suggests larger land area. Besides, buildings with larger frontage are more likely to be some commercial centers which also suggests higher market value.

The mean squared errors on training set and test set of this model are:

$$E_{train} = 0.6421$$
$$E_{test} = 0.6455$$

The residual plot is shown below.

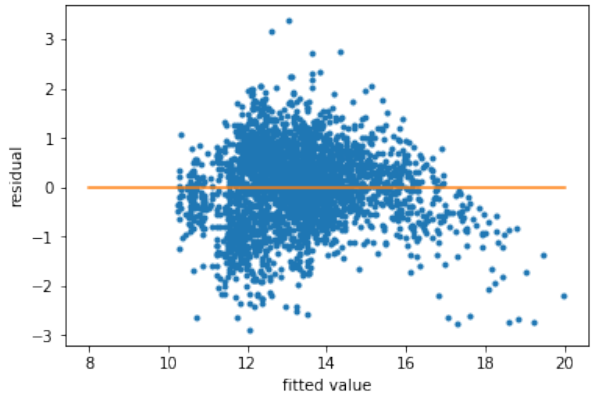

Figure 4.2 l_2 Loss without Regularization Residual Plot

From the residual plot, it can be seen that most of the residuals are nearly 0, but when the true value is high, our model tends to predict a higher value than true value, resulting to negative residuals when fitted value are approximately above 16 in the plot. This problem will be fixed later in **section 4.7** using the two parts model.

We also scaled back our prediction and plotted the predicted values against true values.

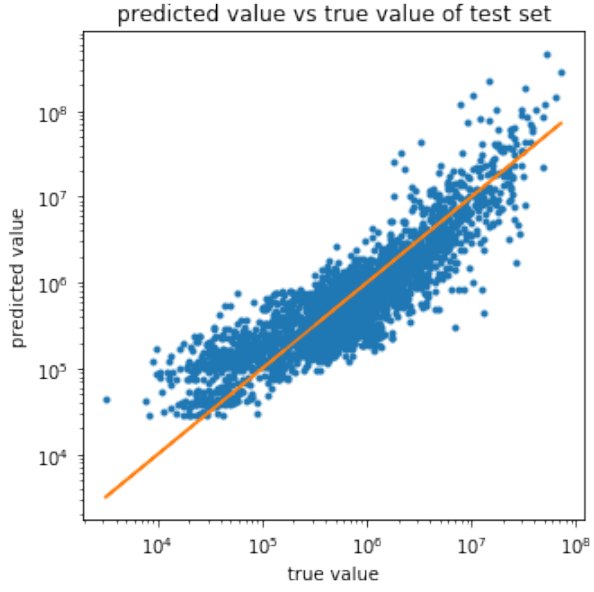

Figure 4.3 l_2 Loss without Regularization True Value vs Predicted Value

4.3 l2 Loss with l_1 Regularization. Now we are interested in what independent variables are more significant. So we performed l_1 regularization on our data and used 10-th fold cross-validation to select λ. The best model we get has $\lambda = 0.0005$ The regression result of coefficients is shown in the graph below.

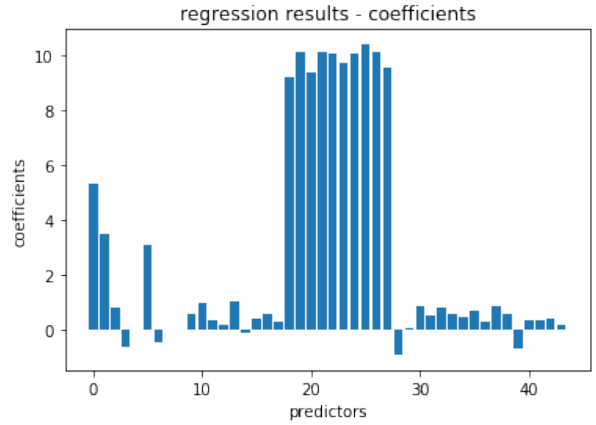

Figure 4.4 l_2 Loss with l_1 Regularization Coefficient Plot

The regression result shows that variables with 10 highest coefficient values are ten zone dummy variables. This means we could roughly predict total real estate value just by its location. The coefficients for ten zones are listed in the table below.

Variable	Coefficient
Central Harlem	5.8722
Chelsea & Clinton	6.7494
East Harlem	5.9853
Gramercy Park & Murray Hill	6.7577
Greenwich Village & Soho	6.7021
L Manhattan	6.3623
L East Side	6.6913
U Eest Side	7.0405
U West Side	6.7642
Inwood and Washington Heights	6.1785

Table 3. Coefficients of zone variables

Comparing the coefficients for zones with **Figure 2.5**, we can see these two results approximately align with each other, except the coefficient for L.Manhattan is low while in **Figure 2.5** the value for L.Manhattan is the highest. This is probably because buildings in L.Manhattan have higher values in some other independent variables which also have large coefficients.

The mean squared errors on training set and test set of this model are:

$$E_{train} = 0.6478$$
$$E_{test} = 0.6486$$

The residual plot and the scatter plot of predicted values against true values are similar with those in section 4.2, so we will not present them here due to constraint on page length.

4.4 l2 Loss with l2 Regularization. Although l_1 regularization can give us a sparse model which quickly decides what factors are most important, the prediction error is generally higher than using l_2 regularizor. So we performed l_2 regularization on our data and used 10-th fold cross-validation to select λ. The best model we get has $\lambda = 0.001$

The regression result of coefficients is shown in the graph below.

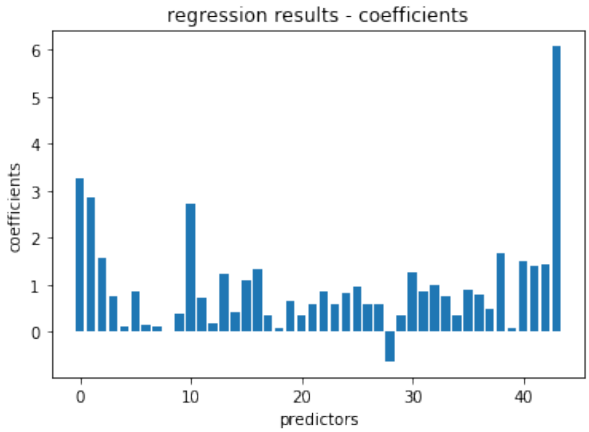

Figure 4.5 l_2 Loss with l_2 Regularization Coefficient Plot

The mean squared errors on training set and test set of this model are:

$$E_{train} = 0.6478$$
$$E_{test} = 0.6719$$

The residual plot is shown below.

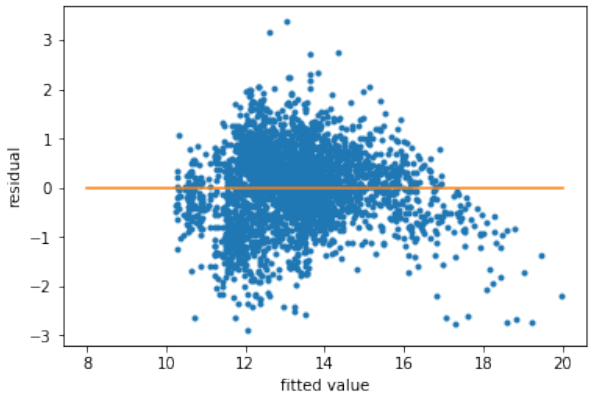

Figure 4.6 l_2 Loss with l_2 Regularization Residual Plot

The scatter plot of predicted values against true values is shown below.

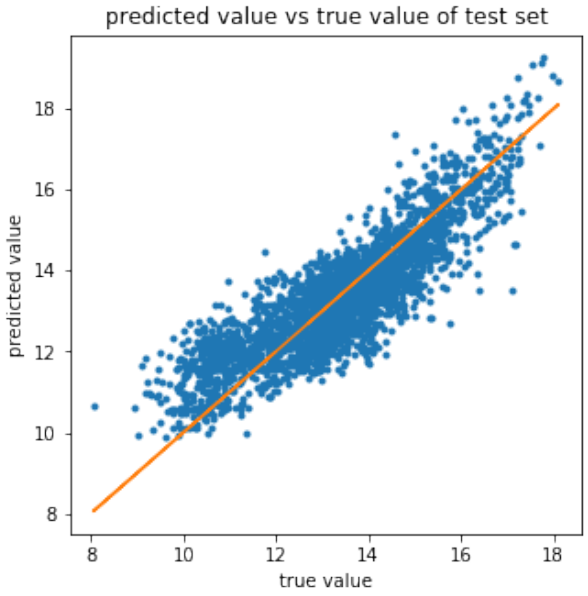

Figure 4.7 l_2 Loss with l_2 Regularization True Value vs Predicted Value

4.5 l_2 Loss with smooth Regularization. The model's independent variables contain three sets of categorical transferred dummy variables: zones, land use type, owner type. We are interested in whether these different categories really have different response on dependent variable. So in this section, we performed smooth regularization that penalizes pairwise difference in coefficients of these categories. We used 10-th fold cross-validation to select λ. The best model we get has $\lambda = 0.1$

We compare the regression coefficients of a model with no smooth regularization with the smoothed model in the following plot.

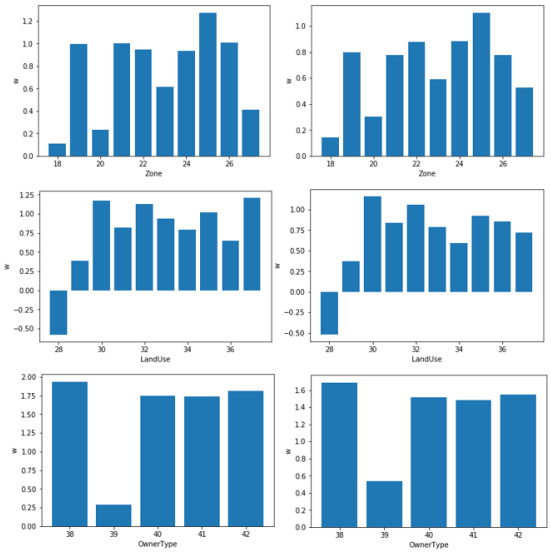

Figure 4.8 Comparison of Coefficients Between Regular (Left) and Smoothed Model (Right)

From this result we can see, for zones, buildings located in Central Harlem and East Harlem tend to have the lowest market value, while buildings in Upper East side have the highest market value. For land use type, we found One & Two Family Buildings and Multi-Family Walk-Up Buildings have the lowest value, while Commercial & Office Buildings have the highest value. For owner type, we found Mixed City & Private Ownership buildings have significantly lower value than other owner types. However, in Figure 2.8 you might find Mixed City & Private Ownership has a very high value. The inconsistency here is probably because our model still needs to be improved. As you can see in section 4.7, the coefficient value of Mixed City & Private Ownership becomes higher.

The mean squared errors on training set and test set of this model are:

$$E_{train} = 0.6748$$
$$E_{test} = 0.6731$$

The residual plot and scatter plot of predicted values against true values are similar to the models above. Due to constraint on page length, we do not show them here.

4.6 Huber Loss with l_2 Regularization. In the previous model, we are not fitting dependent variable with high values well. So we want to try Huber loss, which can reduce the influence of extreme values to regression results. The λ for the l_2 regularizor is 0.001.

The mean squared errors on training set and test set of this model are:

$$E_{train} = 0.6748$$
$$E_{test} = 0.6719$$

The residual plot is shown below. We can see the problem of fitting high values still exists. In the following section, we will propose a two stage model to solve this problem better.

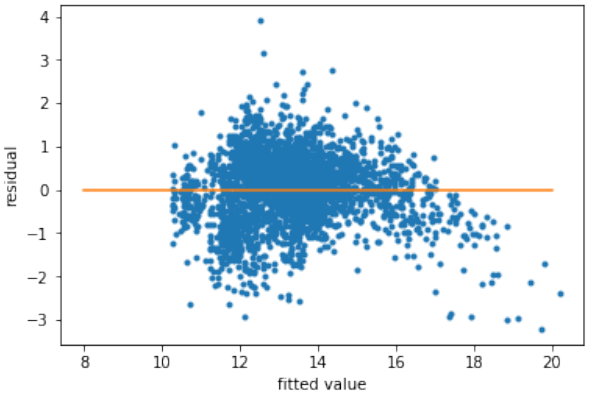

Figure 4.9 Huber Loss with l_2 Regularization Residual Plot

The scatter plot of predicted values against true values is shown below.

Figure 4.10 Huber Loss with l_2 Regularization True Value vs Predicted Value

4.7 Two stage model using Huber Loss with l_2 Regularization. In the following, we will show that the two stage model can significantly reduce fitting error, and improve prediction accuracy when the dependent variable has high value. We performed the two stage model using Huber Loss with l_2 Regularization on our training set. By doing 10th-fold cross validation, we select the threshold parameter y_t to be 14, and the λ for the l_2 regularizor to be 0.001.
The mean squared errors on training set and test set of this model are:

$$E_{train} = 0.519$$
$$E_{test} = 0.506$$

From this result, it is clear the prediction accuracy has been increased tremendously.

In the following, we briefly discuss the results of both w_1 and w_2 . The coefficient plots are shown below.

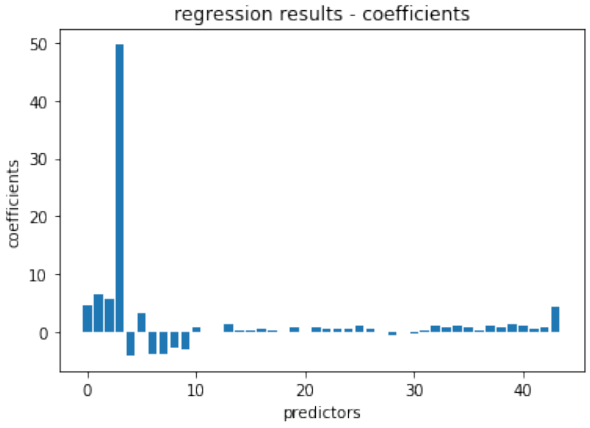

Figure 4.11 Two Stage Model with Huber Loss and l_2 Regularization Coefficient plot of w_1

According to the result of w_1 , the independent variables having relatively large coefficients are listed in the table below.

Variable	Coefficient
LotFront	4.61
NumFloors	6.61
ComArea	5.67
ResArea	49.68
RetailArea	3.12
Constant	4.17

Table 4. Coefficients of some numerical variables in w1

From **Table 4** we can see lot frontage area, number of floors, commercial area and retail area have high coefficients as expected. But it is interesting that Residential area has a coefficient as high as 49. This is probability because most of the buildings having low market value are family buildings. And among them, those having larger areas will have higher values.

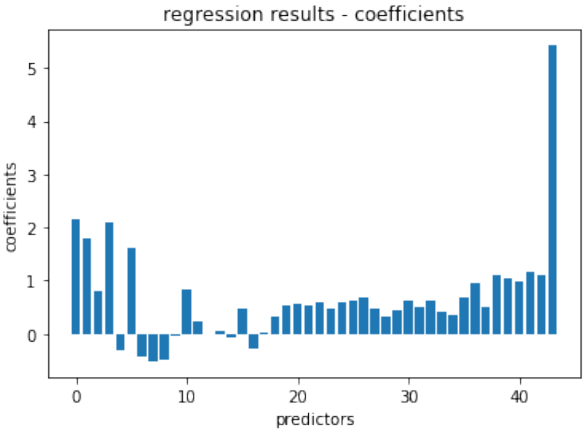

Figure 4.12 Two Stage Model with Huber Loss and l_2 Regularization Coefficient plot of w_2

For the prediction of real estates with high values, from **Figure 4.12**, we noticed that the highest coefficients are still LotFront, NumFloors, ResArea and RetailArea. The independent variables having relatively large coefficients are listed in the table below.

Variable	Coefficient
LotFront	2.14
NumFloors	1.81
ResArea	2.10
RetailArea	1.62
Constant	5.42

Table 5. Coefficients of some numerical variables in w2

The residual plot is shown below. We can see the 'tail' in our residual plot now disappears, meaning the problem of fitting high values has been solved.

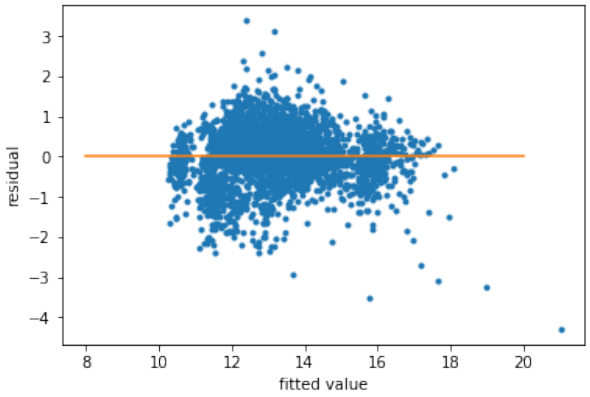

Figure 4.13 Two Stage Model with Huber Loss and l_2 Regularization Residual Plot

The scatter plot of predicted values against true values is shown below.

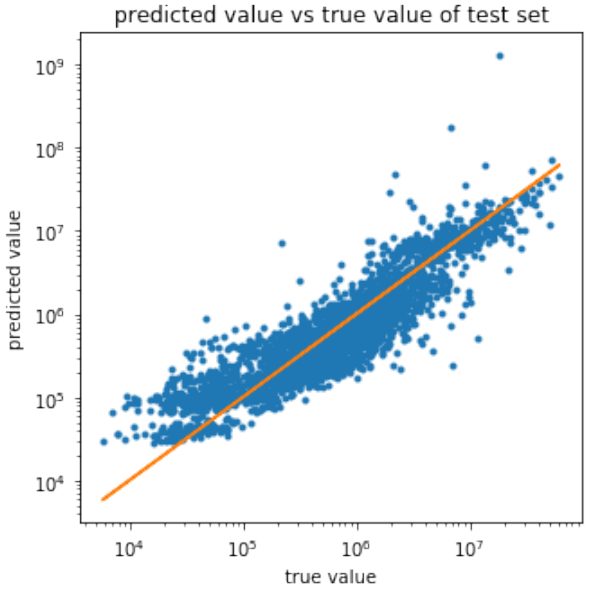

Figure 4.14 Two Stage Model with Huber Loss and l_2 Regularization True Value vs Predicted Value

5 Conclusions and Applications

In this study, we tried 6 different regression models to predict the total market value of real estates in Manhattan. Our best model is the two stage model with Huber loss and l_2 regularization, which yields a training error of 0.519, and a test error of 0.506.

We are willing to use our prediction model to provide references for real estates companies to evaluate land and building values. And some key findings from our model are listed below.

Our study shows the most important numerical variables influencing a real estate's total market values are: LotFront(lot frontage), NumFloors(number of floors), ComArea(commercial area), ResArea(residential area), RetailArea. Buildings with high values in these variables tend to have higher total market value. The result of lasso regression and smooth regression also tell us that the location of the real estate also has significant impact on its market value. Buildings located in Upper East area tends to have higher value, while those in Central Harlem or East Harlem have lower values. Smooth regression further reveals that Multi-Family Elevator Buildings and Commercial Buildings have higher prices. As for owner type, the regression result of the two stage model shows there is actually not much difference between these five owner types.

As for different regression methods, it can be seen that Huber loss can make the regression results more robust to outliers, and different regularizations can reduce the variance of prediction at the cost of increasing bias. Finally, a two stage model provides a simple way to refine the regression when the dependent variable cannot be described perfectly by a linear model.

6 Further Improvements

To further improve prediction accuracy, we could also try multivariate polynomial regression to fit the data better. Since some of our variables are correlated, principle component analysis can handle such case. But due to the time limit, we are not able to apply these techniques for the time being, and we plan to try it in the future.

ACKNOWLEDGMENTS.

1. "ZIP Code Definitions of New York City Neighborhoods." New York State Department of Health, Mar. 2006.
2. "Primary Land Use Tax Lot Output." NYC Open Data, 19 Jan. 2017.
3. Udell, Madeleine. ORIE 4741: Learning with Big Messy Data. Sept. 2017.