Predict and Quantify Development Influence

Jiali Zhou[1], Lizhen Tan[1] and Yuting Gui[1]

***Abstract*— The real estate market is a key pillar industry in large cities. From everyday experience, high property values often indicate prosperity in the area. Many economists and researchers have worked on finding and analyzing factors which have impacts on the integrated view of the economic development [1],[2]. Yet many of the previous work focused on forecasting the housing price, while our work presented here is to make a detour to find underlying potential building attributes which give a newly introduced development power to exert an impact to its nearby real estate market in Manhattan, New York. We applied different basic machine learning models such as logistic regression(LR), support vector machine (SVM), random forest (RF), Naive Bayes (NB) and AdaBoost(Ada) to find if the approach using only real estate properties' attributes as factors would generate reasonable predictive models.**

I. INTRODUCTION

As one of the leading trade cities in the world, New York City has a significant impact on a lot of industries, such as finance, fashion, technology, and so on. Being such a prestigious city, its real estate market is an absorbing research field. In New York City, the value of the properties increases rapidly from year to year. However, this is not a uniform case from region to region within the city. In general, people think that the metropolitan area such as in midtown Manhattan have a higher housing value than other regions. Yet, our work has shown that, this may not be the case. In a certain area constraint, new introduced development may or may not bring any effect to the region housing values. In addition, even effect exists, it can be either positive or negative. Instead looking for the underlying economic factors, such as investment fund growth, cultural and political changes, etc. as described in [1], we proceeded our project under a machine learning scope to find how the property attributes as potential features contribute to the effect (effect is defined to be either positive or negative).

The Section II lists the major tasks we performed and tools we used for this project. Data collection, derivation and preprocessing are described in detail in Section III, this data preprocessing section mainly explained how we derived the wanted target variable (1= impact, both positive and negative; 0 = no impact) and data for visualization. Section IV describes a second-stage data preprocessing for modeling purpose, as well as models we used and the important features that we found from the well-tuned model. Those features are believed to work as important property factors affecting the housing value influence under the selected model setting. Section V analyzes and discusses the models, as well as the geolocation influence by visualizing data onto a map. We conclude our work in Section VII. Possible improvement and future work are described in Secion VII. Codes related to this project can be find in our GitHub page. [1]

II. TASKS

The two main tasks for this project are: visualization of the effect of newly opened properties on its neighbors price; identify the potential factors that affect the properties price.

III. DATA COLLECTION AND PREPROCESS

A. Data Sources

Three datasets have been used in this project. The first one is Primary Land Use Tax Lot Output (PLUTO) [3], it provides extensive land use and geographic data at the tax lot level. The second one is Rolling Sales data [4] from the NYC Department of Finance. These files lists properties that sold in the last twelve-month period in New York City for all tax classes. These files include the neighborhood, building type, square footage and other data. The third is the 311 NYC complaint dataset [5], it contains the count and type of complaints for each tax block.

B. Data format

Used attributes for Rollingsales data: borough, building class category, block, lot.

Used attributes for PLUTO data: We used nearly all of the features in the PLUTO dataset, but drop some which we think they are highly correlated to some other ones. A full set of used features can be found in Appendix I.

Used attributes for Complaint data: BIN number and complaint text.

C. Data Challenge

- Inconsistent data file format:
  PLUTO data file format differs from year to year. Attributes in PLUTO data also have variation from year to year. For example, year 2002 provides text file while year 2015 provides csv. Period 2002 provides 'Block', 'Lot', 'AssessTotal', while period 2015 provides Block','Lot','AssessTot' as attributes. Some files even doesnt include the name of columns. This inconsistency has increased the difficulty in getting required data through a large amount of data.
- Data missing:
  Entries in a file may not be complete, and some of them

[1] New York University

[1]https://github.com/NYU-CDS-Capstone-Project/Cider

do not provide meaningful information. For instance, our goal is to focus on data inside Manhattan, some key attributes for some entries are recorded as NA, which makes no sense to our agenda. We can drop these types of entries, however, if such data is of a large portion in a dataset, we may lose useful insight from the degenerate data. Data missing has a large effect in PLUTO data, since there is a gap in 2008. We have no way to get access to these data, so can only leave them out.

- Variation of coordinate representation:
  As the latitudes and longitudes in PLUTO data use standard coordinate translation between New York State Long Island (ESRI:102718) code, we make a transition between New York State Long Island (ESRI:102718) code and latitude/longitude (EPSG:4269) for PLUTO data in order to do data visualization in Jupyter notebook using Folium.
- Absence of archived data:
  In the project, we focused on using archived data which can map to each year and find how the data change from year to year raise difference in the target variable(whether a newly introduced development brings effect to nearby region). However, there is not a lot of open archived data. For example, we would like to find what types of buildings were around an interested location, using any api, such as Google Places api, only gives us data at present. Attempts to find such data by hand-Googling was not successful since it took a large amount of time and the web-searching did not give reasonable data. Since there is not many archived data, our data is very limited to the PLUTO data.

D. Data Preprocess and Data Derivation

This section describes the data preprocssing for derived-data. The derived-data is use for visualization and construction of target variable for machine learning models.

- Initial PLUTO data: P1. Initial rolling sales data: R1.
- Initial filtration of data: Select rolling sales data in certain category: here we select BUILDING CLASS AT PRESENT within category K but not equal to K9, reference to the glossary of the rolling dataset can be found at [6]. This subset of rolling sales data is called R2.
- Get key for merging: As we want to know which subset of PLUTO data has been sold out in specific year, we need a key for both dataset P1 and R2. Here we choose to concatenate borough, block, lot to form another column named bbl as a key for merging P1 and R2 from year to year. After merging, we get the sold properties PLUTO data for each year. We call this P2.
- Find K-category properties sold in each year:Get dictionary for all sold properties in Manhattan from 2002 to 2016: Initialize a dictionary bbl2price. For each row in P2, if row[bbl] in bbl2price, then append the (year, price) in the dictionarys item, else add a new (key, value) in the dictionary.
- Get average housing price within a distance: Based on dataset P2, pre-define a set of radii around the interested properties, then calculate the average housing price within the each of the radii to see how far a data point has affected in its nearby region. The same task is performed for all the years at hand, namely from 2002 to 2016(without year 2008). We call the dataset P3. Due to missing data issue, data instances without location, area or price value information were dropped.

IV. Model and Analysis

A. Data Description

We used three datasets to build models. The main difference between the three datasets is the label associated with each properties. The binary labels indicate whether the newly opened properties affect their neighboring properties' average price or not. To fairly measure this influence, we use three sets of neighbors, small, medium and large. The small neighborhood contains properties within a 50-meter distance from the center property, while the medium neighborhood and large neighborhood were defined to have distances of 200 meters and 400 meters respectively. We used 50_radius, 200_radius and 400_radius to represent the three datasets. The labels were assigned in the following procedure: if the ratio between **center property price increase rate** and **neighboring properties average price increase rate** was greater than 1.1 or less than 0.9, the center property would label to 1, otherwise 0. The features we used in this classification task were the properties PLUTO features, and a complaint count feature from 311 noise complaint dataset. See more details of the label assignment and features in Appendix II.

B. Model Data Preprocessing

The sizes of three datasets 50_radius, 200_radius and 400_radius are 891, 910 and 913 respectively (in the unit of number of instances). First, we split the whole dataset into train and test sets. The test set was data from year 2014, and the rest of the data were set to be historical training data. Since the label is defined using three consecutive years (namely, in order to find impact of a property which was introduced in 2014, we need data from 2013 to 2015 to see percentage change compared with the year before and the year after. As a result, since we have only data till 2015, the latest year we can construct the labels was only up to 2014). Since our dataset was so small, only around 900 instances, taking the 2014 data as testset has further restrain the training size, thus we use 5-fold cross validation for training. Both datasets had more than thirty percent of the instances with missing features, if simply drop all instances with missing features, a large number of potential useful information and features would also dropped. In order to maintain useful features as well as data completeness, we interpolated the missing features with reasonable value. For example, BsmtCode is a feature from PLUTO. It contains a numerical value which represents the basement type/grade of a property. 0 means no basement, 1 and 2 mean full

basement, 3 and 4 mean partial basement. For the missing values, we simply assigned a zero value to them, as we have no way to know what they are. A decimal interpolated value, such as an average of the feature was not used since such number has no meaning for the BsmtCode. After interpolate reasonable values of some features, the size of three datasets 50_radius, 200_radius and 400_radius are 890, 908, 911 respectively. The sizes were comparable to the original datasets.

C. Models

In this Models section, different feature engineering and machine learning algorithms have been applied into the datasets.

1) Feature engineering: The total feature size is 121 for all three datasets. First, we dropped duplicated features, for example multiple dates, and Floor Area Ratio. Second, we dropped features that contained more than ninety percent missing values. For example, Landmark Name, which recorded whether a property was a landmark and if it was, what its landmark name was. Third, we dropped meaningless features, such as Version of the PLUTO file. After dropping those features, we had 20 categorical features and 46 numerical features. The categorical features includes 'BldgClass','LandUse','OwnerType' and so on, and the numerical features includes complaint count, X_coord, Y_coord and so on. This whole set of features can be found in Appendix I. Next, we converted all categorical features to binary features, which gives us 1172, 1182, and 1185 features for the three dataset respectively. We also tried dropped all categorical features, just use the numerical features to fit the model. The results were shown in Table 1.

2) Model selection: Machine learning models have been applied to all datasets. To be more specific, the model we used were Logistic Regression(LR), Support Vector Machine(SVM), Naive Bayes(NB), Random Forest(RF), and AdaBoost(Ada). The parameters were tuned in train set using cross-validation. We then applied the tuned model with the highest average cross-validation accuracy score to the test set. The metrics of measuring all models was accuracy score, which is a ratio of correct predictions. It is chosen because all of the three dataset are roughly balanced with respect to the binary label. In Table 1, we reported the test score on different dataset with different feature engineering methods(either used all features or using only numerical features). The best model among all three datasets was achieved by random forest. Random forest is able to handle both numerical and categorical variables, but it is easy to overfit. More details of model analysis is in section V.

TABLE I
TEST SCORES IN DIFFERENT MODEL AND DATASETS

Dataset	Feature Eng	Models	Val score	Test score
50_radius	all feature	LR	0.606	
		NB	0.704	
		SVM	0.662	
		RF	**0.762**	**0.502**
		Ada	0.664	
	drop categorical	LR	0.613	
		NB	0.585	
		SVM	0.655	
		RF	0.741	
		Ada	0.662	
200_radius	all feature	LR	0.606	
		NB	0.652	
		SVM	0.653	
		RF	**0.717**	**0.510**
		Ada	0.675	
	drop categorical	LR	0.604	
		NB	0.579	
		SVM	0.652	
		RF	0.685	
		Ada	0.613	
400_radius	all feature	LR	0.651	
		NB	0.708	
		SVM	0.658	
		RF	0.753	
		Ada	0.678	
	drop categorical	LR	0.644	
		NB	0.658	
		SVM	0.658	
		RF	**0.788**	**0.591**
		Ada	0.644	

V. DISCUSSION AND ANALYSIS

As depicted from Table I, it was easy to find that, random forest gave the best validation score among all models through all datasets. However, even using such tuned models, the test score was quite low, only achieved accuracy slightly above 0.5, which was not much better than a random classifier. One issue was that our dataset was really small; and small datasets are overfit prone. In addition to this, smaller radii (50_radius and 200_radius) had better validation scores with all features, while larger distance (400_raidus) had a better validation score with only numerical features. This shows that the impact distance played an important role in our model setup. This may be due to the reason that, within a smaller radius,such as 50 meters, gave only information of the solely interested centered property. Then target label constructed by this small distance only shows the rate change within this interested property if it had a large land area. This can be inferred also from the test scores. As the distance increases, the test score deviates from a random guessing classifier.

Besides information obtained from the models, we can also get a clearer picture by using projection of data onto a map.

Fig. 1 shows instances with no impact on their neighboring housing prices while Fig. 2 shows instances which has impact (either positive or negative) on their neighboring housing prices. It is intriguing to find that, new commercial building transactions in the western part of Manhattan tend to have impact while the ones in eastern Manhattan do not. This might be caused by the newly development of the High Line

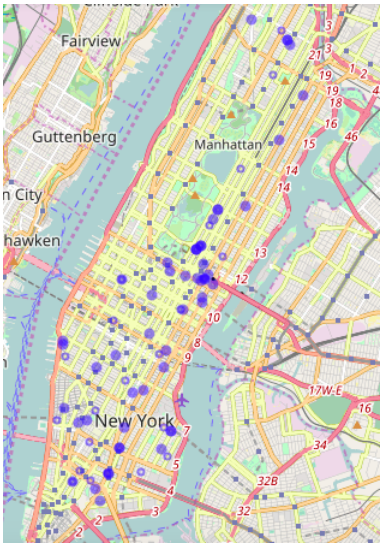

Fig. 1. Properties have no Impact on Neighborhood with radius = 400 in 2014

Fig. 2. Properties have impact on Neighborhood with radius = 400 in 2014

Park in the west which brings more activities to the region; such activities exert certain forces on the housing price in the area. Unlike the western Manhattan, the eastern Manhattan tends to be more stable as it is a residential area.

In order to see how our model does in the task, we also visualized 2014 year's data with radius = 400 on Manhattan to see the misclassified regions. For mislabeled instances, we plotted with purple circles, while corrected classified instances were plotted with green circles. Eighty four instances were mislabeled by the random forest model. The figure is shown in Fig. 3:

From the figure, we can find clusters for mislabeled instances in Times Square area. This is probably due to that this area's properties are sold frequently and since coordinates for x and y are shown as important features in training model, they are likely to be mislabeled due to that in year 2013

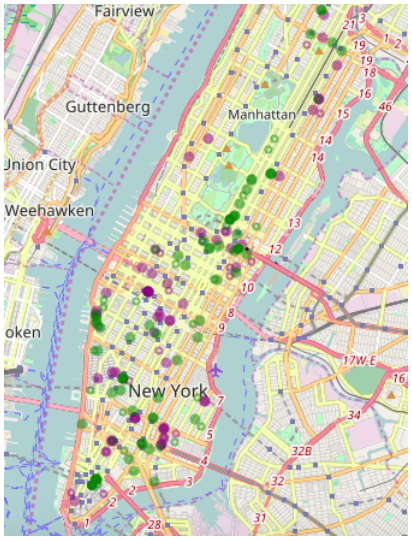

Fig. 3. Predicted label for 2014 year's data with radius = 400

the properties in this region are marked as no impact on neighbors. The figures are shown in Fig. 4 and Fig. 5. For Fig. 4, the blue circle means the sold property is labeled with no impact. For Fig. 5, we can clearly see that the properties in this region is mislabeled with 0 although it has influence on its neighborhood. This also tells us that active regions which have more potential driving forces are hard to capture by our basic models.

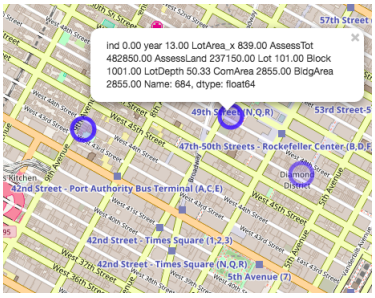

Fig. 4. Real label for properties in Times Square in 2013

VI. Acknowledgement

We would like to thank our advisor Mr. Luc Wilson for offering us valuable and inspiring suggestions and support along the way. And we would also like to thank Professor Claudio Silva and the teaching assistants in this class for preparing lectures.

VII. Conclusions and Future Work

New York City as a fast-paced city, its real estate market also runs in a fast-paced manner. Stable features such as most of the attributes gathered in the PLUTO data are not enough for analyzing the complicated market. The poor performance of our basic machine learning models also suggests that better features and bigger volume of dataset are

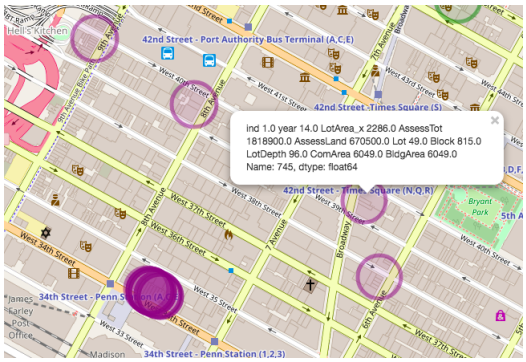

Fig. 5. Mislabeled properties in Times Square in 2014

needed for improvement. In order to improve the work, we should find a better way in interpolation for missing data. Techniques such as nearest neighbor interpolation may be a good candidate. Besides that, instead of finding impact using historical data(which may not have due to absence of archived data), one can focus on analyzing price difference between properties with similar attributes but in different regions in the same year, as recent data may be more easily found and obtained. Further on that, one can use refined features to construct causal inference to find possible potential factors which lead to the price difference.

APPENDIX I

Features used in PLUTO dataset

Numerical feature: 'XCoord_x', 'XCoord_y', 'FactryArea', 'BldgFront', 'Easements', 'LotArea', 'Lot', 'year', 'ind', 'ResArea', 'YearAlter1', 'StrgeArea', 'OfficeArea', 'NumFloors', 'NumBldgs', 'BldgDepth', 'AreaSource', 'LotFront', 'UnitsTotal', 'ExemptLand', 'BsmtCode', 'complaint', 'UnitsRes', 'YCoord_y', 'YCoord_x', 'CD', 'LotDepth', 'AssessLand', 'YearBuilt', 'BldgArea', 'Block', 'ComArea', 'BoroCode', 'LotType', 'YearAlter2', 'ZipCode', 'Council', 'assesstotal', 'APPBBL', 'ExemptTot', 'RetailArea', 'PolicePrct', 'GarageArea', 'SchoolDist', 'OtherArea', 'HealthCtr', 'Borough'

Categorical Features: 'AllZoning1', 'AllZoning2', 'BldgClass', 'BuiltCode', 'Ext', 'FireComp', 'IrrLotCode', 'LandUse', 'Overlay1', 'Overlay2', 'OwnerType', 'ProxCode', 'SPDist1', 'SPDist2', 'Sanborn', 'SplitZone', 'TaxMap', 'ZoneDist1', 'ZoneDist2', 'ZoneMap'

APPENDIX II

Details in getting target variable(indicator): For each sold property, define whether its neighborhood avg price increase ratio to Manhattan avg price increase ratio distinguishes in the whole distribution. For example, we want to analyze whether the 2003 sold properties has impact in radius 200 meters or not. We get the avg price of properties within 200 meter circle of sold properties in 2002, 2003 and 2004 separately, and get the ratio of the avg price from 2003 to 2002, 2004 to 2003, here we name it to be 03_to_02_ratio, 04_to_03_ratio. We also got the same ratio for the whole Manhattan. Then we set two thresholds: 0.9, 1.1. If 04_to_03_ratio_200_meter/03_to_02_ratio_200_meter to 04_to_03_ratio_whole_mahattan/03_to_02_ratio_whole_mahattan is between the two thresholds, then we say no impact exists (indicator=0), otherwise the impact exists (indicator=1).

ACKNOWLEDGMENT

We would like to thank our advisor Mr. Luc Wilson for offering us valuable and inspiring suggestions and support along the way. And we would also like to thank Professor Claudio Silva and the teaching assistants in this class for preparing lectures.

REFERENCES

[1] Burinskien, Marija, Vitalija Rudzkien, and Jrat Venckauskait. "Models of factors influencing the real estate price." (2011).
[2] Bork, Lasse, and Stig Vinther Mller. "Housing price forecastability: A factor analysis." EFA 2012 Copenhagen Meetings Paper. 2015.
[3] Neckerman, Kathryn M., et al. "Disparities in urban neighborhood conditions: evidence from GIS measures and field observation in New York City." Journal of Public Health Policy 30.1 (2009): S264-S285.
[4] http://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page
[5] https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9/data
[6] http://www1.nyc.gov/assets/finance/jump/hlpbldgcode.html