Predicting and Quantify Development Influence

Jiali Zhou[1], Lizhen Tan[1] and Yuting Gui[1], Luc Wilson[2]

New York University[1], KPF Urban Interface[2]

Objectives

- Find underlying potential building attributes that impact a property's neighborhood price.
- Visualize the newly opened retailers' influence on its neighborhood price.

Introduction

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 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.

Dataset

Three datasets have been used in this project.

1. Primary Land Use Tax Lot Output (PLUTO)
2. Rolling Sales data
3. 311 NYC complaint dataset

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

Machine Learning Models
Feature engineering

Combine PLUTO, RollingSale, Complaint dataset, we applied a filtration to obtain only commercial properties, which gave us about 900 instances and 121 features for each instance.

1. Dropped duplicate features. (Dates)
2. Dropped features with lots of missing values.
3. Dropped meaningless features (Version number)

After dropped feature, we had 66 features in total.
Label: The binary labels indicate whether the newly opened properties affect their neighboring properties' average price or not. We used three different sets of neighbors to calculate labels, small (within 50m), medium (within 200m) and large(within 400m).

Machine Learning Models
Model selection

Machine learning models have been applied to the dataset. The model we used were LR, SVM, NB, RF, AdaBoost. The parameters were tuned in train set using cross-validation. The metrics of measuring all models was accuracy score,

Table 1: Test scores in different models

Dataset	Models	Val score	Test score
400_radius	LR	0.651	**0.591**
	NB	0.708	
	SVM	0.658	
	RF	0.753	
	Ada	0.678	

Important Result

Through our experiments in model training, the best model results we obtained was from **random forest(RF)**. Some important features are: coordinates, common area, year, building depth, retail area. However, with a really small dataset, any machine learning model is overfit prone.

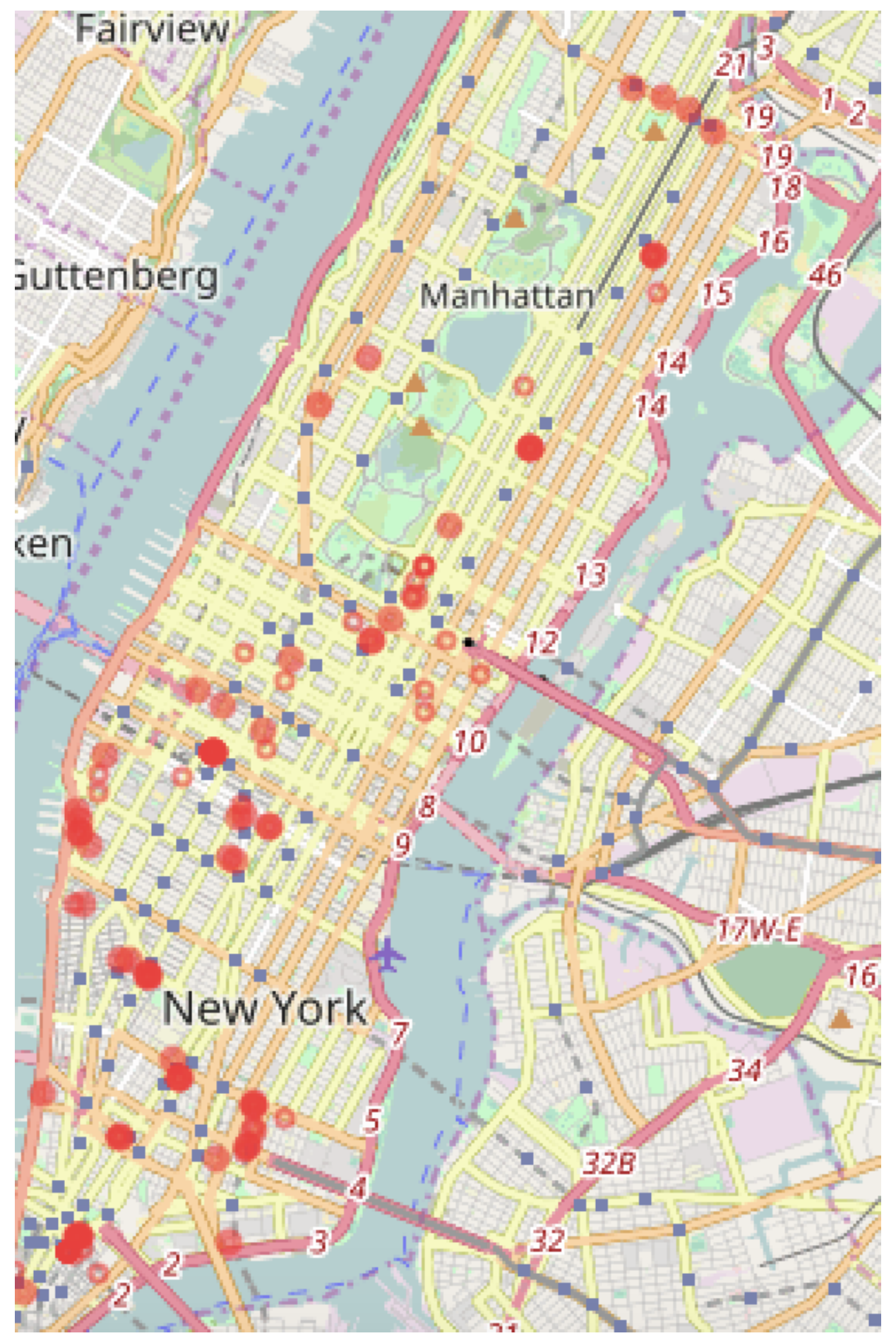

Figure 1: Properties have impact on Neighborhood with radius = 400 in 2014

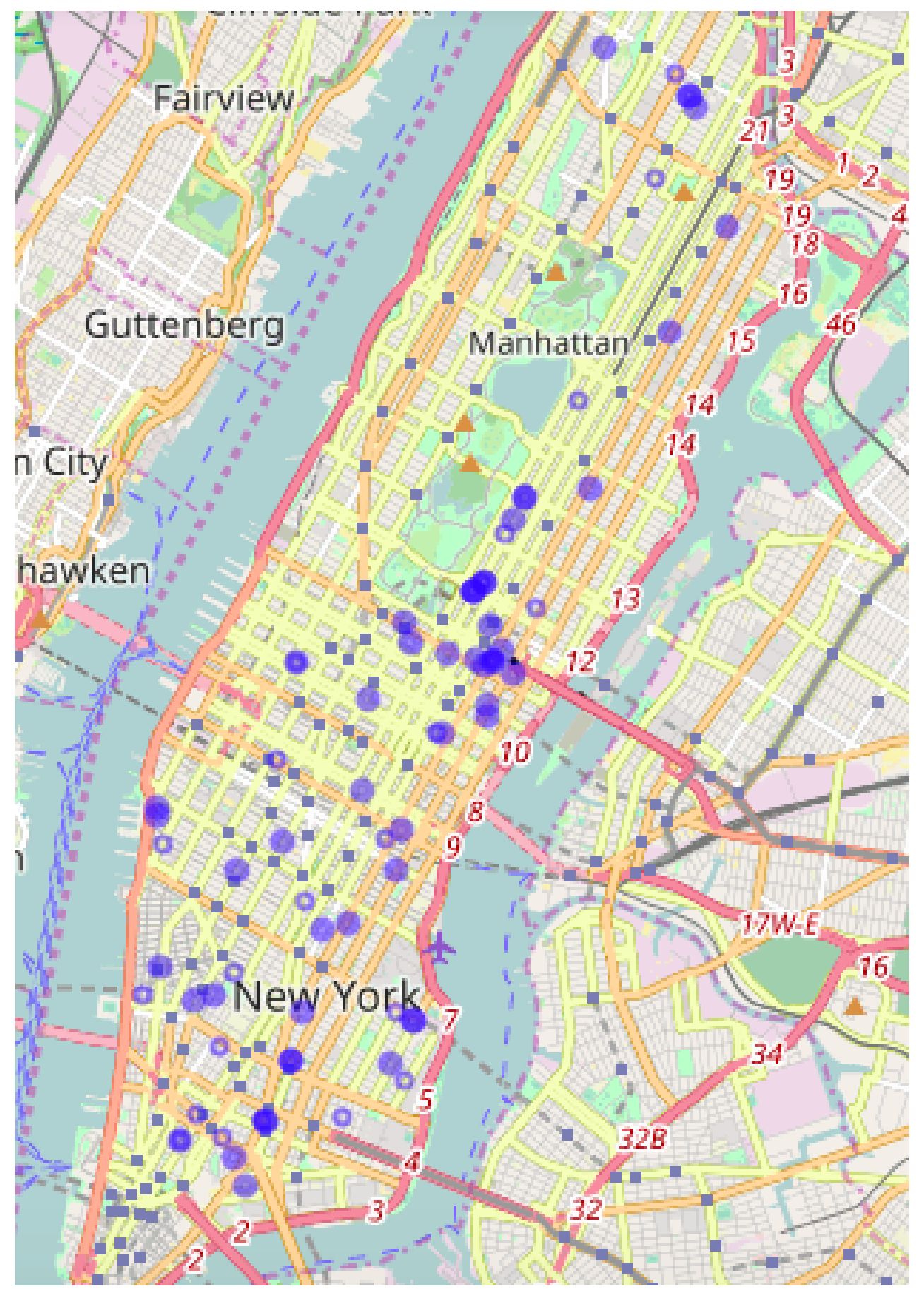

Figure 2: Properties do not have impact on Neighborhood with radius = 400 in 2014

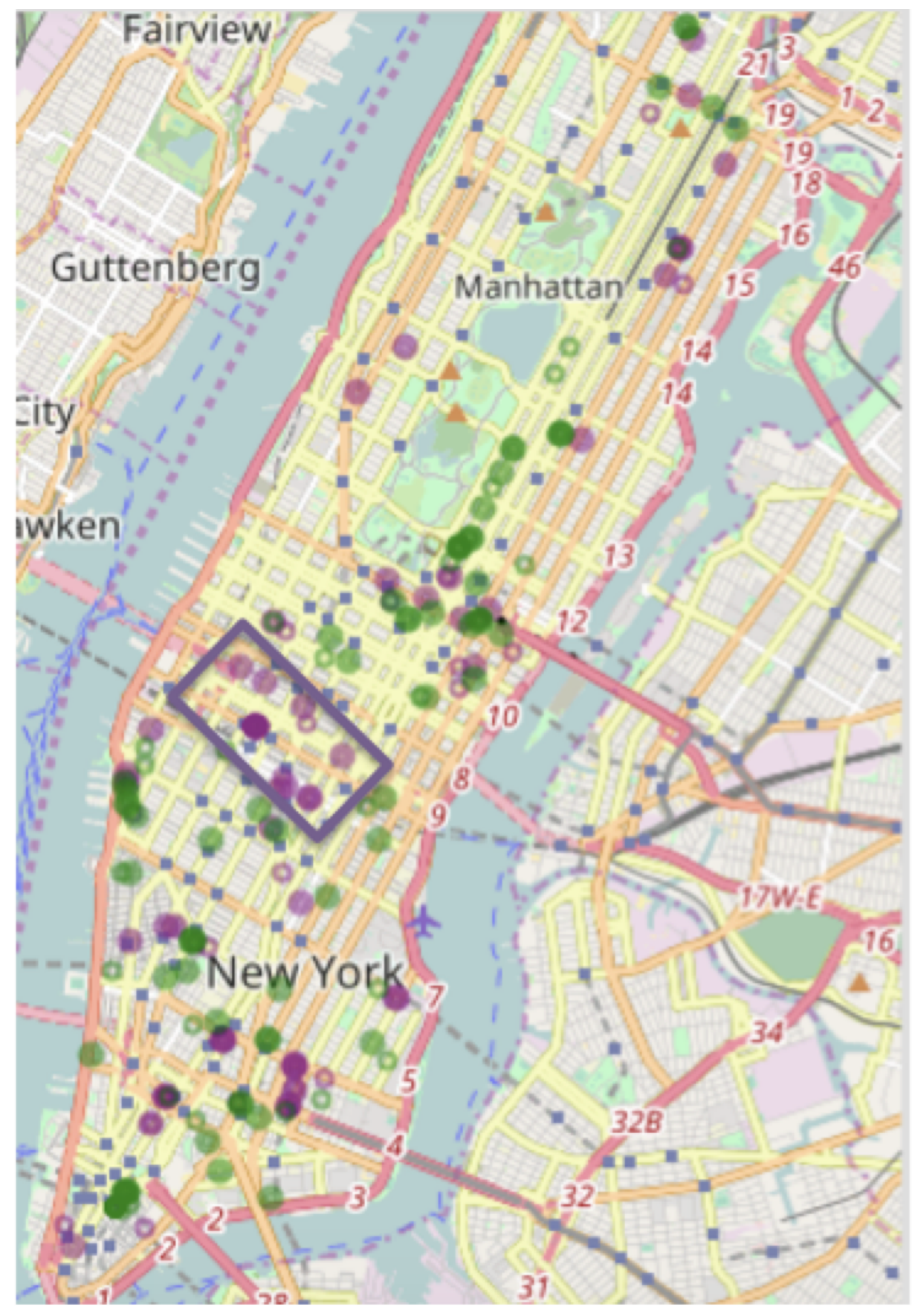

Figure 3: Predicted label for 2014 year's data with radius = 400, green-correctly classified, purple-misclassified

Picture Analysis

- Properties which have impacts on its neighborhood prices cluster in western Manhattan[Fig. 1], the area near the Hudson River. Properties have no impact cluster in eastern Manhattan, near the East River[Fig. 2]. (in 2014)
- There is no hard boundary between properties have impacts and those do not with respect to geolocation.
- Models tended to misclassify data points in midtown (shown in the purple square in Fig. 3). This may be due to the reason that more activities and changes in midtown every year. The frequent changes through years are hard to captured by the models.

Conclusion

Finding factors affecting real estate market is not an easy task. With an approach using only properties' attributes as features, the trained models did not give promising results. Our work showed that the property attributes can still be useful if we define a larger impact radius (>400 meters). Besides that, poor performance of models is also an alert for better features and larger data volume are needed.

Acknowledgements

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.

NYU Center for Data Science