Restaurant Health Inspections and Crime Statistics Predict the Real Estate Market in New York City

Rafael M. Moraes , Anasse Bari , Jiachen Zhu

Courant Institute of Mathematical Sciences
Computer Science Department
New York University
New York City
{rafael.moraes, abari, jiachen.zhu}@nyu.edu

Abstract. Predictions of apartments prices in New York City (NYC) have always been of interest to new homeowners, investors, Wall Street funds managers, and inhabitants of the city. In recent years, average prices have risen to the highest ever recorded rebounding after the 2008 economic recession. Although prices are trending up, not all apartments are. Different regions of the city have appreciated differently over time; knowing where to buy or sell is essential for all stakeholders. In this project, we propose a predictive analytics framework that analyzes new alternative data sources to extract predictive features of the NYC real estate market. Our experiments indicated that restaurant health inspection data and crime statistics can help predict apartments prices in NYC. The framework we introduce in this work uses an artificial recurrent neural network with Long Short-Term Memory (LSTM) units and incorporates the two latter predictive features to predict future prices of apartments. Empirical results show that feeding predictive features from (1) restaurant inspections data and (2) crime statistics to a neural network with LSTM units results in smaller errors than the traditional Autoregressive Integrated Moving Average (ARIMA) model, which is normally used for this type of regression. Predictive analytics based on non-linear models with features from alternative data sources can capture hidden relationships that linear models are not able not discover. The framework presented in this study has the potential to serve as a supplement to the traditional forecasting tools of real estate markets.

Keywords: Artificial Intelligence · Predictive Analytics · Supervised Learning · Recurrent Neural Networks · Open Data · Alternative Data · Wall Street · Real Estate Markets

1 Introduction

The price of apartments in New York City is a popular topic among homeowners, investors, real-estate agencies, city government and general inhabitants of the city. Average prices have risen in the last decades to the highest ever recorded in

recent years, despite the economic crisis in 2008-10 [19]. However, not all apartments are created equal: different regions of the city appreciated very differently over time, which compounded to very disparate prices of apartments that may have seemed to be similar in the past [12].

Apartments, as other real estate properties, are valued not based on some existing standard price or inflation; instead, market supply and demand determine how valuable each apartment in each region of the city is. Even in the same building, it is possible to have apartments whose prices varied differently when undergoing the same changes, for example: a new commercial building in front may boost the price of apartments in lower floors, but decrease the prices of those in higher floors (e.g. because they lost their nice view of the park). Therefore, predicting these prices is inherently hard, given the specificity of each apartment, building and location.

On the other hand, housing is a large part of a person's expenditures in life [21], so any help in forecasting can become a decisive factor for buyers and sellers, both to decide how valuable apartments are and the right time to act.

The contributions in this study can be summarized as: (i) a predictive analytics framework for valuation of apartments in New York City given historical buy and sell prices and two alternative data sources: restaurant health inspections data and crime statistics; (ii) an artificial neural network model for predicting future prices with new predictive features for real estate markets; (iii) we experimentally show that crime statistics and restaurant health inspections, when used as predictive features in an LSTM model, provide lower prediction error than a traditional forecasting model; and (iv) we show a use case where linear predictive algorithms can fail to model reality whereas non-linear models based on recurrent neural networks are able discover hidden complex relationships.

In the next section we introduce the preprocessing work we did on the data sets used in this study.

2 Datasets

2.1 Description

Since 2012, New York City has laws that require government agencies to make much of their data accessible to the public [10], improving transparency and enabling other to put these data to good use. With this legislation, the NYC Open Data was created, where thousands of different datasets related in some way to the city can be found and used by anyone. We describe three of those, which were used in this work:

Rolling Sales Data [20] Supported by the NYC's Dept. of Finance. Every time an apartment is sold in New York City, the operation needs to be registered with the citys Department of Finance (DoF), in order to generate the correct sales tax and proper documentation approval. Once the sale is registered, the DoF includes it as a record in its database and shares it with the public in the

Rolling Sales Data , which is updated frequently. This dataset contains tabular information about each apartment sold in the city since 2003, with fields such as: zip code, street address, apartment number, sale price, sale date, and tax class. We have chosen to use the data from 2008-2017, which contained around 197,000 sales records, of around 110,000 different apartments.

NYC Restaurant Inspection Results [17] Supported by the NYC's Dept. of Health and Mental Hygiene. Every year, each of NYC's 24,000 restaurants go through at least one unannounced public health inspection, which looks for hygiene violations, adherence to regulations and assigns a score based on its compliance to health standards; the commonly known restaurant grades A, B and C derive from the score obtained in this inspection. This dataset contains fields such as: restaurant ID (i.e. CAMIS), address, zip code, and each specific violation code and textual description by the inspector. We have chosen to use the data from 2014-2017, which contained around 380,000 records, where each is one violation registered in an inspection, amounting to multiple violations per inspection, on average.

Citywide Crime Statistics [18] Supported by NYC's Police Department (NYPD). They include all the arrests, shootings, complaints and crimes registered by the police around the city with details, such as: incident time and date, offense type (i.e. misdemeanor or felony), and location. We decided to use the Incident-level Complaint data, from 2006 to 2017, which contained more than 5 million records of misdemeanors and felonies committed across the city over these years. There are many details available for each record, such as a textual description and a categorization of the crime, details about the location and time it happened, and details about the suspect, among others.

2.2 Data Preprocessing

We performed several data preprocessing and data cleaning steps to the datasets used including feature extraction and feature selection.

In the Rolling Sales Data, in order to be able to match apartments that appear multiple times in the dataset, it was essential to standardize all the addresses. However, we have noticed that the addresses seemed to be typed manually at the source, leading to equal addresses being represented by different texts. For example: "10 West 15TH Styreet" and "10 W 15th ST". To homogenize them, we have used the Google Geocoding API [13], which receives an address and returns its standardized form (i.e. following Google's standardization).

A second problem faced was that many buy/sell operations had their prices listed too low when compared to other nearby apartments. For example, many operations had prices below $20,000. Therefore we discarded any record whose price was below $200,000. As a last step, zip codes of areas considered too small

were removed from the analysis - such as 10118, which points exclusively to the Empire State Building - resulting in a total of 36 zip codes analyzed.

Regarding the two other datasets, the score of inspections was averaged and the total number of crimes (i.e. summing misdemeanors and felonies) was consolidated: in both cases it was done per month and per zip code. Details about the crimes or suspects were not considered, only the total number of records per month and per zip code. A similar approach was used for the restaurant inspections: additional details of the records were not considered, only the final scores of the inspections, which were averaged per month and per zip code. No other preprocessing step was necessary in these datasets.

3 Methodology

3.1 Predicting Present Apartments' Prices

To create the final dataset we first filtered only apartments that appeared at least twice in the data and stored their prices and dates of sale. Then these were used to calculate the average price growth per month between dates that, when compounded, would yield the price difference observed. This was calculated for all matching apartments. We assumed that this price growth per month could be extended a little from before the first sale date to after the last one, totaling an additional 20% extension, 10% for each side. This seemed reasonable since the price of an apartment does not start or stop growing when it is sold, instead it can be seen as a smoother process over time. For example, if the difference between sales is 5 years, we assume the growth rate is defined and constant in the 6 months prior to the first sale and 6 months after the last sale. With these calculations done for each apartment, we clustered them by zip codes and calculated a corresponding average price growth of each zip code per month. We note that clustering based solely on zip code has a low granularity and may not achieve optimal results, but this was not necessary for the analysis shown here, so a more precise clustering is left as future work.

3.2 Predicting Future Apartments' Prices

We have reasoned that even the current price of an apartment is unknown and we have explained a simple method to obtain an estimate for it based on nearby apartments. However, stakeholders would benefit further if there were a reasonably reliable system that could predict where the prices are going. The first idea one can imagine is simply doing a time-series regression with the data we already have, but this, as we show, is not optimal. We propose using all three datasets (i.e. apartment prices, crime, restaurant violations) to predict future apartment prices. The idea is that the two other datasets are much noisier but they still contain a faint signal that can improve our predictions. This is arguably plausible given the complexity of big cities, where many measurable factors are interconnected and can reinforce each other with certain time delays.

As baseline, we use an ARIMA (Autoregressive Integrated Moving Average) model for comparison, which is normally used for time-series analysis [11]. Initially we propose a simple linear model that tries to combine the time-series but, as shown below, this does not yield good results. Our final proposed model consists of a Long Short-Term Memory (LSTM) [16] neural network model that is able to combine the three datasets in a non-linear fashion and reaches the best performance. To validate these results, we ran thousands of experiments by varying hyperparameters and comparing the loss metric.

The analyzed time series have data for a period of 48 months for each chosen zip code in Manhattan. We use the first 42 months as train data and the last 6 months as validation. The goal is to compare the predictions generated by both the ARIMA and LSTM (with and without additional data) models with the ground-truth values. We use the mean squared error (MSE) metric to make this comparison.

3.3 Data Linearity and Stationarity

When dealing with time series, it is a common assumption that the autoregressive nature of the stochastic process that generated it can be reasonably approximated by linear models. This is a mere simplification and certainly not the best approach, but it is a widely used procedure. Based on this assumption, the Wold Decomposition theorem [1] states that any time-series that is weakly (covariance) stationary [9] (i.e. mean and autocovariance do not change over time) can be approximated by a sum of a deterministic and a stochastic time series. We will first apply this idea to the datasets presented here and compare this linear approach with a more general non-linear model.

The first step is verifying if the data is stationary, which can be done by using the Augmented Dickey-Fuller test [15]. This will allow the use of models such as ARMA (AutoRegressive Moving Average), which assumes the points in the data are a weighted combination of the previous p points and some random noise. However, data that shows some signs of non-stationarity can be better approximated by the ARIMA (AutoRegressive Integrated Moving Average), which has an additional step to remove a possible non-stationarity.

3.4 Granger Causality and Cross-correlation

Given the three time series, the main question we are looking to answer is how much the two additional time series can help forecast future apartment prices. A structured way to verify this is by using Granger Causality [14], which is a statistical test that measures the predictive power of one time series into another and assumes they are linearly related. As the ARMA/ARIMA other approach mentioned above, this is the widely used by its simplicity, but it assumes the different datasets have a linear relationship among themselves.

4 Experiments

After generating the time series from the raw data, the ADF test is performed in R, where we see that the null hypothesis of data not being stationary cannot be rejected. To make all time series stationary, a first-order differentiation is used. After this, the ADF test is performed again, where we confirm that the data is now stationary, so the models can be applied.

4.1 Granger Causality

Once the data is stationary, we run the Granger-causality test by using the lmtest package in R. This was done with the "apartment prices"-"restaurant inspections" and "apartment prices"-"crime indices" pairs. For each pair, the Granger-causality test was performed twice: verifying the Granger-causality of the first time series into the second and vice versa. This null-hypothesis test is formulated such that a low p-value (i.e. assumed smaller than 0.05) would indicate the existence of Granger-causality (i.e. the inexistence could be rejected as null hypothesis). The results obtained are consistent with the opposite: with the exception of two zip codes (10006 and 10014) for the "apartment prices"-"crime indices" pair, all other tests returned a p-value greater than 0.05, indicating with high confidence that there is no Granger-causality.

4.2 ARIMA Model Applied to Base Time Series

The ARIMA model is commonly applied to regression problems involving time series because of its simplicity and clarity. It uses 3 parameters (P, D, Q) to determine the order of the regression: P is the number of parameters to use in the autoregressive part of the model; Q is the number of differences to use (to make sure the final time series is stationary); and Q is the number of parameters in the moving average window. Here we use a grid parameter search to show the best possible ARIMA model that could fit this problem: several values for p, d and q where tested and the MSE of the 6-month prediction was recorded. Here we present the histogram of the MSEs obtained for each of these configurations, but only showing those that obtained and MSE smaller than $2 \cdot 10^{-7}$. In this experiment, the MSE of the best ARIMA model was $1.1053 \cdot 10^{-7}$.

P	**D**	**Q**	**Mean Squared Error (1x10-7)**
2	0	4	**1.1053**
0	0	2	1.1066
2	0	0	1.1113

Table 1: Summary of the three best results obtained for ARIMA, with the corresponding model parameters.

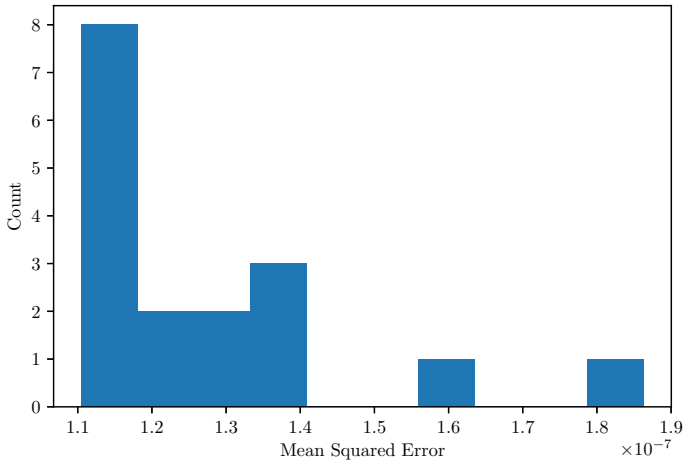

Fig. 1: Histogram showing the MSE performance of ARIMA models, which predict future apartment prices, with different hyperparameters.

4.3 LSTM Model Applied to Base Time Series

Here we use a Long Short-Term Memory neural network instead of the linear ARIMA model, following the same methodology: a large grid parameter search was used in order to find the best possible model to fit our data, based on the validation set. There were 48 different combinations of parameters and each of these was run 10 times with a different seed, totaling 480 runs. The best MSE obtained was $1.0943 \cdot 10^{-7}$. The histogram with the results is presented in the next item.

4.4 LSTM Model Applied to All Time Series

We now include the restaurant and crime time series into the LSTM model and perform the same parameter search. The best MSE obtained was $1.0906 \cdot 10^{-7}$. The histogram of the MSEs over all LSTM runs is shown in Fig. 2. Table 2 summarizes the findings.

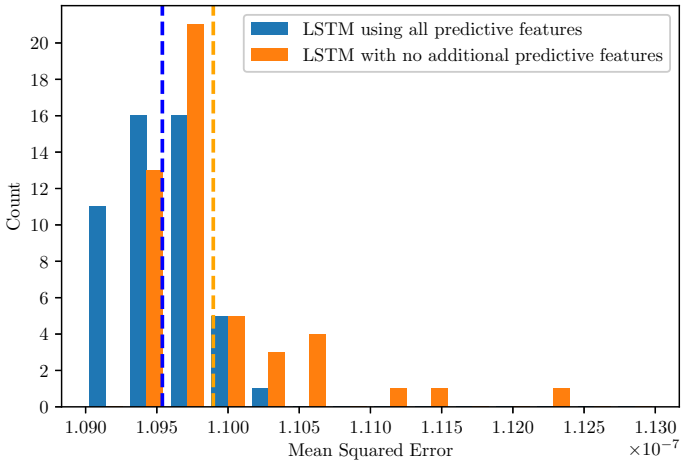

Fig. 2: Histogram showing the MSE performance of LSTM models, which predict future apartment prices, with different hyperparameters. Blue bars are the cases using all predictive features, extracted from the restaurant inspections, the crime statistics, and the apartment sales data. Bars in orange are cases when only using predictive features obtained from the apartment sales data. Dotted lines represent the average MSE of each set. Each model configuration was run 10 times and the average MSE of each is used here. A total of 98 configurations (980 runs) are shown.

Table 2: Summary of best results obtained for each model.

Model (best run)	**Mean Squared Error (1x10-7)**
ARIMA, only apt data	1.1053
LSTM, only apt data	1.0943
LSTM, all data	**1.0906**

5 Conclusion

In this study, we presented a predictive analytics framework that provides a methodology for analyzing alternative data sources and applying recurrent neural networks to help predict real estate markets. The experiments that we conducted indicate that engineering new predictive features such from new data

sources to real estate such as restaurant health inspection data and crime statistics has the potential to improve the accuracy of predictions of apartment prices using non-linear models such as recurrent neural networks. As future work, we plan to measure the predictive power of other alternative data sources related to real estate markets and to apply biologically inspired algorithms [2, 6–8] for better grouping apartments. We also plan to include new features by applying Emotion Artificial Intelligence [3–5] to big data sources such as Twitter and news articles in order to build a customer index for the real estate market of NYC.

Acknowledgments

We would like thank Jing Wang, who contributed with helpful discussions to the initial analysis of this work. Also, the NYU High Performance Computing team, especially Shenglong Wang, who was always available to help with technical issues in the computer cluster.

References

1. T. W Anderson. *The Statistical Analysis of Time Series.* John Wiley & Sons, Hoboken.
2. Anasse Bari, Mohamed Chaouchi, and Tommy Jung. *Predictive analytics for dummies.* John Wiley & Sons, 2016.
3. Anasse Bari and Lihao Liu. Probing the wisdom of apple, inc., crowds using alternative data sources. 2017.
4. Anasse Bari, Pantea Peidaee, Aniruddh Khera, Jianghao Zhu, and Hongting Chen. Predicting financial markets using the wisdom of crowds. In *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*, pages 334–340. IEEE, 2019.
5. Anasse Bari and Goktug Saatcioglu. Emotion artificial intelligence derived from ensemble learning. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 1763–1770. IEEE, 2018.
6. Abdelghani Bellaachia and Anasse Bari. Sfloscan: A biologically-inspired data mining framework for community identification in dynamic social networks. In *2011 IEEE Symposium on Swarm Intelligence*, pages 1–8. IEEE, 2011.
7. Abdelghani Bellaachia and Anasse Bari. Flock by leader: a novel machine learning biologically inspired clustering algorithm. In *International Conference in Swarm Intelligence*, pages 117–126. Springer, 2012.
8. Abdelghani Bellaachia and Anasse Bari. A flocking based data mining algorithm for detecting outliers in cancer gene expression microarray data. In *2012 International Conference on Information Retrieval & Knowledge Management*, pages 305–311. IEEE, 2012.
9. P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting.* Springer Texts in Statistics. Springer International Publishing, 2016.

10. Craig Campbell. New York City Open Data: A brief history. https://datasmart.ash.harvard.edu/news/article/new-york-city-open-data-a-brief-history-991, 2017. [Online; posted 08-March-2017].
11. Samarjit Das. *Time series analysis*. Princeton University Press, Princeton, NJ, 1994.
12. The Furman Center for Real Estate & Urban Policy. Trends in New York City housing price appreciation. pages 20–24, 2008.
13. Google. Google Geocoding API. https://developers.google.com/maps/documentation/geocoding/start.
14. Clive WJ Granger. Causality, cointegration, and control. *Journal of Economic Dynamics and Control*, 12(2-3):551–559, 1988.
15. William H. 1951 Greene. *Econometric analysis*. Pearson Prentice Hall, Upper Saddle River, NJ.
16. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
17. New York City Department of Mental Health and Mental Hygiene. NYC Restaurant Inspection Results. https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j.
18. New York Police Department. Citywide Crime Statistics. https://www1.nyc.gov/site/nypd/stats/crime-statistics/citywide-crime-stats.page, 2018.
19. Emily Nonko. Manhattan home prices have increased dramatically in a decade. 2017.
20. NYC Department of Finance. Rolling Sales Data. https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page, 2018.
21. Bureau of Labor Statistics. U.S. Department of Labor. Consumer Expenditures - 2017. https://www.bls.gov/news.release/cesan.nr0.htm, 2018.