**Big Data Science**

**CSCI-GA 3033-001**

**Dr. Anasse Bari**

**Ryan Moriarty,** [**rpm295@nyu.edu**](mailto:rpm295@nyu.edu)

**Anthony Tse,** [**at4091@nyu.edu**](mailto:at4091@nyu.edu)

**William Jiang,** [**wj419@nyu.edu**](mailto:wj419@nyu.edu)

**Harnessing Big Data to Forecast NYC Rent Prices**

**Abstract**

In this project we aim to build a model that accurately predicts fluctuations in New York City rent prices within a given neighborhood based on a confluence of traditional and alternative data sources. We follow the cross-industry process for data mining (CRISP-DM) guidelines for conducting our research.

**Introduction**

There are an estimated 3.5 million rental units in New York City comprising 63% of available housing stock compared to the nationwide average of 37%. [1] NYC is the second most expensive rental market in the country with median rent prices of around $2,900. [2] This represents a large portion of a tenant’s disposable income as well as a large income source for real estate professionals. Accurately forecasting rent prices can inform renter’s decisions on where, when, and for how much to rent. They can help landlords and property firms maximize revenue by predicting market fluctuations and aid real estate agents in comparing the prospects of given properties.

We hypothesize that the following factors are leading indicators of rental prices within a given neighborhood: crime rates, number of taxi trips, and 311 complaints. By organizing these data sources into a common format and applying time series and machine learning models we will attempt to build a predictive model that’s useful to the NYC real estate community.

Our rent dataset will come from the website StreetEasy which has median monthly rent prices for 70 NYC neighborhoods dating back to 2010. As a test case, we will focus our analysis on the neighborhood of Williamsburg. For our exogenous variables, we will be using the 311 complaint, NYPD crime, and TLC taxi trip data sets made accessible through NYC’s open data website.

**Literature Review**

We read the paper *Twitter Mood Predicts the Stock Market* which informed us on how to apply alternative data to financial prediction. The authors analyzed the text content of daily Twitter feeds using mood tracking tools. They then ran Granger causality analysis to test whether public mood states are a leading indicator of DJIA prices. Our work aligns with theirs as as our goal is to discover whether quality of life factors within a neighborhood are predictive of rent prices. We perform a similar Granger causality analysis in testing the predictive effect of different time series.

To build our time series models we closely followed the textbook *Forecasting Principles and Practice* by Rob J. Hyndman and George Athanasopoulos. We adapted examples presented in their book to our own datasets. We also referenced the text *Practical Time Series Forecasting with R* by Galit Shmueli and Kenneth C. Lichtendahl.

To bolster our understanding of the application of big data to real estate, we read the paper *Predicting Housing Price Based on Ensemble Learning Algorithm* by Yajuan Tang et al. In it, we learned about the possibility of forecasting future prices using big data and which factors may correlate with housing prices. We also got some inspiration on how to preprocess our data.

**CRISP-DM Approach**

**Business Understanding**

The neighborhood of Williamsburg, Brooklyn, which is our use case for this study, has undergone significant gentrification since the late 1990s characterized by its art scene, nightlife, and hipster culture. Low rent prices were a main attraction for artists when they first began settling in the area, but rent prices have increased drastically as the neighborhood has become a popular place of residence. Today, average rents can range from $1,400 for a studio apartment to $4,000 for a two-bedroom apartment. [3] This is largely due to the area’s popular restaurants, convenient public transportation, and close proximity to Manhattan.

We believe that trends affecting rent prices in Williamsburg are representative of New York City as a whole. With a population of over 33,000, Williamsburg should give us enough of a sample to get meaningful results, which we can then extrapolate to the city at large. Depending on our results, we plan to iterate to other neighborhoods and boroughs, as a means of confirming our findings.

**Data Understanding**

All of the data that we sourced, the NYPD complaints, taxi data, and 311 service requests, are publicly available from NYC Open Data [8]. The data related to rental price is provided by StreetEasy [9]. The data from this source is structured and relatively clean. The data was mostly categorical, with some numerical data for longitude and latitude. The categorical data is most likely generated from a form and the coordinate data is most likely automatically generated. Each row of the data corresponds to a single incident, so it was necessary to transform the categorical data in a way in which it could be aggregated into monthly bins. We noticed that there was a sharp decline in taxi trips taken circa 2014. We attribute this to the rise in popularity of ride sharing applications such as Lyft and Uber. This decline is not visually correlated to any change in median rent price for Williamsburg.

**Data Preparation**

Our datasets were sourced from NYC’s open data website and StreetEasy’s NYC median rental price dataset. Each dataset was processed using a custom script written in R which, among other things, filtered for events based on latitude and longitude, dropped unnecessary columns, and aggregated data according to month. We’ve chosen to group the data by month because that is the format of the rental price dataset. The taxi datasets, which were up to 20 GB in size, had to be run on NYU’s high performance computing cluster.

The decision to initially limit all preprocessing, modelling, and evaluation steps to a single neighborhood was made because of the massive scale of the given datasets. Also, the NYPD complaint and the taxi trip datasets were not labeled by neighborhood, but by latitude and longitude. To avoid having the task of translating the coordinates to the neighborhoods for tens of millions of rows of data limit the progress of our project, we decided to initially narrow the scope of our project to the neighborhood of Williamsburg in Brooklyn. To do this, we used the filter function available on the Open Data website to restrict the data collected to Williamsburg’s geographic boundaries. These boundaries were estimated by setting limits on the longitude and latitude. Since Williamsburg is quite rectangular in shape we expect this will serve as a good approximation.

Since the columns of the 311 and NYPD complaints and Taxi data all contained categorical data and had multiple daily occurrences, they were converted into numeric data through expanding additional rows for each category encountered. For example, a record may contain “Noise Pollution” as its complaint type, and so can the next one. Another record may have “Rat Infestation” as another complaint type. The previous categorical complaint type column is now obviated by new columns created by each new complaint type, and then tallied across a given month. Doing so has allowed us to incorporate Linear Regression and Time Series analysis into our model. After aggregating the data, the resulting matrix had around 200 columns. Column regions where the data was most sparse were dropped, allowing us to selectively keep meaningful features for our model. This process was repeated across all 3 datasets, before being joined together into 1 final matrix with 58 features and 105 rows.

Before we settled on using Williamsburg as a test case, we wrote a K nearest neighbor implementation in Java as a means of mapping coordinates to neighborhoods. To do so, we obtained a list of every neighborhood in NYC, along with their respective coordinates. New data points were mapped to their single closest neighborhood using equirectangular approximation. In practice, it may be better use several representative points per neighborhood in order to capture nonlinear geographic boundaries.

**Modeling**

We took two different approaches to modeling our data, temporal models using a time-series methodologies and non-temporal models using machine learning algorithms.

1. Linear Regression

An initial approach that we used to forecast future median rent prices was forecasting with linear regression. We decided to first test the data using this more elementary method because it is simple to implement and the results are very interpretable. We were also influenced by our professor, as we were told that linear regression is still used for forecasting in Wall St. firms.

To maximize the interpretability of our data and to minimize the complexity of our linear regression model, we used a forward sequential feature selection algorithm to choose the most predictive features. This algorithm was implemented via RapidMiner, which sequentially chooses the features based on the degree in which it reduced training loss. To determine the best number of features to use, we performed tests within a range between 2 and 40 features to forecast the next month’s rent price. In these preliminary tests, we found that increasing the number of features reduces the observed training loss, but led to increases in the test loss. This likely occurred because having too many features in the linear regression model resulted in overfitting of the training data. Ultimately, we decided on using the best 5 features as this resulted in the lowest test error.

In addition to forecasting rent using linear regression with 5 features, we also modeled the data using single variable linear regression using rent from the recent month as the independent variable. We decided to do this because we found that this autoregressive model was already quite predictive of future rent prices, so we were hoping to investigate if the addition of data from alternative sources could improve our predictive accuracy.

With the approach described, we made forecasts of the rent price for 1, 2, and 3 months into the future. From our results, we found that the multiple linear regression approach did improve the predictive accuracy relative to the simple models (Table 1).The final features chosen for these models also shed some insights on data that could be used to predict the future rent prices. For example, we found that NYPD complaint related features were negatively correlated to the future rent price, which makes intuitive sense. On the other hand, some chosen features, such as the 311 feature indicating the number of damaged street sign complaints, are not easy to make sense of.

Table 1. Results for Linear Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Prediction Target | Most Recent Rent Only  Training Loss (RMSE) | Most Recent Rent Only  Test Loss (RMSE) | Top 5 Attributes  Training Loss (RMSE) | Top 5 Attributes  Test Loss (RMSE) |
| 1 Month in Future | 83.7 | 44.7 | 75.0 | 33.2 |
| 2nd Month in Future | 118.5 | 72.4 | 98.7 | 50.7 |
| 3nd Month in Future | 139.6 | 95.4 | 108.2 | 73.5 |

1. Time Series

The first step in any time series approach is to visualize the data. We first performed a classical additive decomposition of our predictive time series, which separates the dataset into its trend, seasonal, and remainder components. The trend reveals that rent prices increased until late 2013, then declined to a plateau before rising again in late 2015, after which it has been gradually declining. The series also has a bit of seasonality, hitting a trough early in the year, before rising roughly $100 during the late spring, then steadily declining.

We next calculated the autocorrelation and partial autocorrelation among the rent times series. There is a roughly 90% correlation between the current month’s rent price and the previous month’s. This implies we can achieve a fairly good accuracy level for short term prediction using historical rent prices alone.

To test different models on our data we broke the dataset into a training set from 2010-2016 and a test set from 2017-2018. Next we trained several models on our training data, including an autoregressive integrated moving average (ARIMA), a feedforward neural network (NNETAR), and an exponential smoothing model (ETS). We then forecasted each model 24 months into the future and computed the error of each compared to our test set. The ARIMA model performed best with an root mean squared error (RMSE) of 98.38, however all our models beat our baseline NAIVE prediction with drift, which had an RMSE of 182.12.

We then began incorporating exogenous time series to see if it improved our performance. The three time series we chose were the monthly data on the number of taxi trips taken, number of felonies committed, and number of 311 complaints. The number of taxi trips has been declining since late 2013, which we assume is due to the emergence of ride sharing apps like Uber. To model this growth, we included a piecewise linear trend as an intervention variable, which adds 2500 trips for each month after December 2013. Next, we generated a correlation matrix of the time series. Taxi trips had the highest correlation with rent prices at .4 followed by felonies, then 311 complaints.

Of course, correlation doesn’t imply causation. To account for this we made each time series stationary and then ran tests for Granger causality at lags from 1 to 24 months. The only statistically significant results we found were that felonies Granger cause rent prices at a lag of 2-4 months. Taking this into consideration we ran an ARIMAX model which used the 3-month lagged felony time series as an external regressor. The residuals from the model were roughly normally distributed and show no significant autocorrelation, which implies all trends are accounted for. Incorporating the felony data slightly improved our results over ARIMA alone, although the improvement is enough to be statistically significant.

3. Machine Learning

For prediction of a single month’s future rent price, we also modeled our data using a random forest model, a neural network, and deep learning. Of these models, only random forests models give interpretable results. Unfortunately, we were not able get better predictions using these models when compared to linear regression. The results are described as follows.

Random Forests: Random forests samples the dataset with replacement, then takes a subset of the features and branches based on their values. The depth of each tree can be specified, as can the number of trees in the forest. Random forests are an ensemble method, so the final rent prediction is determined as a mean of outputs from the trees in the forest. In testing with random forests, we also decided to use forward sequential feature selection for the best 5 features to split on. With model parameters of 100 trees with a maximum depth of 10 levels to predict the median rent price 1 month in the future, we obtained results of slightly better than simple linear regression. We were able to obtain a Training RMSE of 35.199 and a Test RMSE of 60.599. The chosen variables were mostly different from the ones chosen in linear regression. The following features were chosen: Rent Price (weight: 0.433), Petit Larceny (0.277), Grand Larceny of Motor Vehicle (0.131), Sanitation Condition (0.087), and Street Light Condition (0.072).

Additionally, we tried using Neural Networks in RapidMiner to forecast rent price. No feature selection was performed. Unsurprisingly, the training loss (RMSE: 52.5) was an order of magnitude less than the test loss (RMSE: 267.827). Reducing the number of features and adjusting the model parameters could improve this result.

Finally, we tried the Deep Learning Model in RapidMiner. Since we had the issue of overfitting for neural network model, we decided to use L1 regularization in an attempt to reduce overfitting. After testing several model parameters, we were not able to achieve a good performance. Ultimately we only achieved a training RMSE of 201.109 and a test RMSE of 95.384. It is possible that this could be improved by reducing dimensionality of the initial input. I would like to note that it is difficult to see if the loss is converging through training on the RapidMiner Platform.

**Evaluation**

Our main experimental results are that felonies Granger cause rent prices at a lag of 2 to 4 months. We also found that ARIMA models were able to forecast rent prices up to two years in advance with an RMSE of 98.38, and that an ARIMAX model using the lagged felony time series slightly improved upon that accuracy. In modeling our data using time series models, we learned that there is significant seasonality in rent price time series, and that prices tend to be lowest around February and highest around May. For shorter term forecasting with linear regression, models containing features with data from our alternative datasets gave slightly better forecasts than using past median rent data alone. Finally, we observed results from a random forest model that were slightly better than the results from linear regression in prediction of rent 1 month in the future.

**Deployment**

Our model is not robust enough to be deployed at the moment. The CRISP-DM is an iterative process, and our next steps would be to collect more data on more neighborhoods to refine and improve our model.

**Discussion and Conclusions**

Throughout the work that was performed in this project, we were able to implement several preprocessing methods and data forecasting methods, many of which were fairly predictive. We believe that an important next step to the project, if we were to continue, would be to expand our analysis to several other neighborhoods in New York. Since our data was trained and implemented only on data from Williamsburg, Brooklyn, we cannot assume that the models are flexible, and they would likely fail if we were to apply them with data from other New York neighborhoods. Additionally, we do realized that the movement of median rent price over time for Williamsburg is quite small (n = 109 months, mean = $3057, standard deviation = $183.7, coefficient of variance = %6). We find it questionable that the ability to predict the change that is expected within a 1 to 3 month time frame is too useful from an industry perspective, as real estate investments are typically longer term investments. However, our results could be used to inform a renter on the best time of year to sign a lease. A prospective renter who signs a lease during the winter is expected to save roughly $100 per month compared to one who signs during the Spring, which equates to savings of up to $1,200 per year.

**References**

[1] <https://www1.nyc.gov/assets/rentguidelinesboard/pdf/18HSR.pdf>

[2] Richter, Wolf. “Rental Prices Are Soaring around the US.” Business Insider, 2 July 2018, <https://www.businessinsider.com/rental-prices-are-soaring-around-the-us-2018-7>.

[3] [https://en.wikipedia.org/wiki/Williamsburg,\_Brooklyn](https://en.wikipedia.org/wiki/Williamsburg,_Brooklyn#Gentrification_and_2005_rezoning)

[4]Bollen, Johan, et al. *“Twitter Mood Predicts the Stock Market.”* Journal of Computational Science, vol. 2, no. 1, 14 Oct. 2010.

[5]Hyndman, Rob J., and George Athanasopoulos. *Forecasting: Principles and Practice.* OTexts, 2018.

[6] Shmueli, Galit, and Kenneth C. Lichtendahl. *Practical Time Series Forecasting with R*. Axelrod Schnall Publishers.

[7] Tang, Yajuan, et al. “Predicting Housing Price Based on Ensemble Learning Algorithm.” *IEEE Explore*, 28 Sept. 2018, doi:10.1109/IDAP.2018.8620781. <https://ieeexplore-ieee-org.proxy.library.nyu.edu/document/8620781>

[8] <https://opendata.cityofnewyork.us>

[9] [https://streeteasy.com/blog/data-dashboard](https://streeteasy.com/blog/data-dashboard/)