

Determining Housing Rent

Jonathan Wong Hanzhang Bai Mas Zimmerman
Computer Science Department, University at Buffalo
White Rd, Amherst, NY 14260, United States
Jwong48@Buffalo.edu
Hbai@buffalo.edu
Massimoz@buffalo.edu

Abstract

Predicting the rent price for real estate can be difficult given the housing market is constantly changing. To help combat the difficult prediction, through linear regression, support vector regression, and SGD regressor, the price of a living space can be determined based on the other properties in the area. The lessor can utilize the average of the models' predictions to maximize profit margins instead of undervaluing or overvaluing their house. Along with helping a homeowner find the price of their estate, another application for these models is to help a tenant determine the market price of an estate given the type of housing and neighborhood. Of the three models tested, the most accurate model was the SGD regressor because of its ability to lower the mean squared error until convergence. A GUI was developed allowing users to input their type of estate and neighborhood. The results are dedicated to only NYC neighborhoods, hopefully expanding to other cities in the future.

Introduction

Today, the housing market across the United States is more volatile than ever before. Specifically in New York city, housing continues to exponentially grow in value, making it difficult to determine the rent price of a living space. To be successful in renting out real estate, the price of the house or apartment cannot be overpriced compared to other estates in the area. With the worries of overpricing, the lessor should also not underprice their house in order to maximize profit margins and invest into more real estate. The ideal price for an estate would be the average cost of all estates of the same type in a given neighborhood. Prices that stray away from the standard deviation produces fluctuations throughout the housing market, causing substantial losses every year. For this project, the property will be priced per night to match rates similar to those found on Airbnb. All predictions are calculated based on New York City listings.

Proposed Method

Using Scikit-learn, the three machine learning models that will be incorporated are linear regression, support vector regression, and Stochastic Gradient Descent Regression (SGDRegressor). Linear regression is an desirable model because of its simplicity and multifaceted regularization methods such as lasso regression and ridge regression. To improve upon the accuracy of linear regression, support vector regression can determine a better price by projecting the data in a higher dimension, iterating through lower dimensions, looking for the optimal price within a margin of tolerance. SGD regressor, further developing the predictions of the previous two models, uses the loss gradient of each sample and updates the model per iteration given a learning rate. With these three models together, the average determined price would be the most accurate, accounting fluctuations in the data. The initial data will be pre-processed, removing any impurities such as the name of airbnb, longitude, latitude, minimum nights to stay, number of reviews, and availability. The fields kept are the estate's room type, price, and neighborhood. After cleaning, the data is split in an 80/20 ratio where 80 percent of the data is used for training and

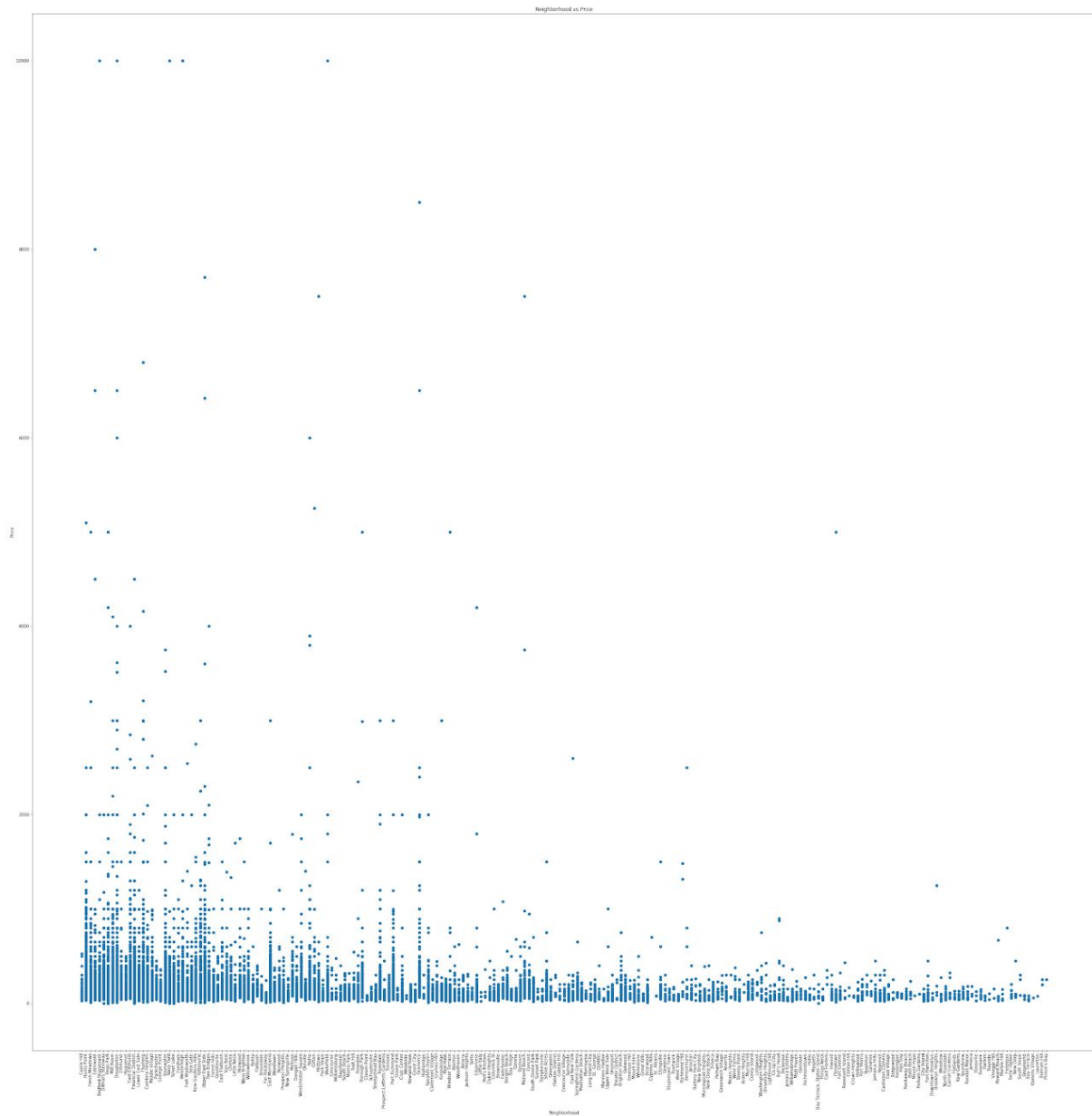
20 percent is used for testing the trained models. Each model is evaluated based on their mean squared errors and R2 scores after testing. The regularization parameters for each model are then tweaked, retrained and retested, lowering the mean squared error and raising the R2 score. This process is repeated until the models have achieved the least mean squared error capable and highest R2 score.

Experimental Data

All data gathered for this project comes from Kaggle, an online community of data scientists and machine learning practitioners. The dataset provides all New York City Airbnb listings in 2019. The original source can be found through Airbnb's public domain. Within the data includes the necessary metrics to form a conclusive price prediction such as neighborhood, room type and price. Each listing holds a listing id, host, host id, name of host, borough, neighborhood, latitude, longitude, room type, price, minimum nights, number of reviews, last review, reviews per month, amount of listings per host, and booking availability. There are 48,896 listings that span throughout the five boroughs of New York City.

Results

The scatter plot of all New York City neighborhoods' Airbnbs and their respective prices



- **Linear Regression (Ridge Regression)**

- Mean Squared Error : 70395.33277432474
- R2 Score : 0.04667786519054651
- Intercept: 218.88820330901277
- Coefficients: [-0.30626349 -60.97073996]

	Actual	Predicted
0	74	89.919315
1	55	90.759865
2	81	96.363530
3	86	200.066205
4	80	95.242797
...
9774	325	195.863457
9775	70	89.919315
9776	175	214.075367
9777	60	64.142458
9778	97	92.440964

- **Support Vector Regression**

- Mean Squared Error : 53945.801062577775
- R2 Score : 0.05547362439330006
- Intercept: [86.56500672]
- No Coefficients : Non-Linear Kernel

	Actual	Predicted
0	74	80.034320
1	55	81.551984
2	81	89.930676
3	86	118.388004
4	80	89.472183
...
9774	325	102.776839
9775	70	80.034320
9776	175	149.108415
9777	60	69.319971
9778	97	85.050578

- **Stochastic Gradient Descent Regression**

- Mean Squared Error :
51075.4791741854
- R2 Score :
0.0702526277346247
- Intercept : [150.71686433]
- Coefficients : [-12.1650281
-61.5590357]

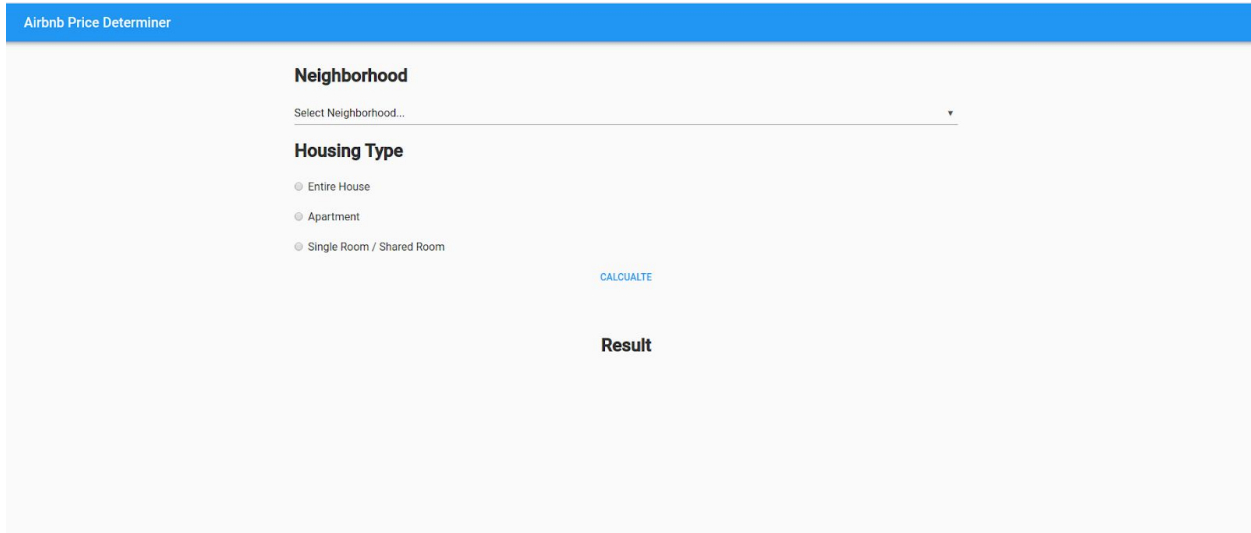
	Actual	Predicted
0	74	85.861793
1	55	86.915255
2	81	93.938334
3	86	197.745002
4	80	92.533718
...
9774	325	192.477693
9775	70	85.861793
9776	175	215.302698
9777	60	53.555631
9778	97	89.022179

Looking at the results, SGDRegressor performed the best out of the three models based on the mean squared errors and R2 scores. Linear regression displayed the worst performance even after using regularization methods such as ridge regression. Support vector regression gave similar results as SGDRegressor when increasing the regularization parameter. To predict an estate's price, the models are finding a correlation between each Airbnb's listing in different dimensions and boundaries. This correlation can be skewed by many outliers in the data within a given neighborhood. These outliers are caused by many owners not basing their rent price on the housing market but rather listing an arbitrary rent price through their emotions towards their estate. This random pricing results in large mean squared errors and low R2 scores.

The next step was to build a UI designed around the three models allowing users to input their housing data and retrieve an accurate rent price. The UI consists of a drop-down menu containing all 220 neighborhoods in New York City sorted in alphabetical order. Below the menu, there are three radio buttons detailing the different types of real estate : house, apartment, or shared room. Using these two features, the user can input their estate's qualifications and submit them to the three models. Each model will predict a price and the average will be printed to the user's screen. The UI is designed through Anvil and hosted on a web server allowing the front-end to communicate and retrieve the final price. In order for the UI to work, the jupyter notebook with a specific API key must be running prior.

Link to the UI

<https://APAOLi6YG344X4VA.anvil.app/BCVF2OHKAUZDZBNUQNGSYZQF>



The screenshot shows a web application titled "Airbnb Price Determiner" with a blue header. The main content area is light gray and contains two sections: "Neighborhood" and "Housing Type". The "Neighborhood" section has a text input field with the placeholder "Select Neighborhood..." and a dropdown arrow. The "Housing Type" section has three radio button options: "Entire House", "Apartment", and "Single Room / Shared Room". Below these options is a blue button labeled "CALCULATE". At the bottom of the form is a section labeled "Result".

Conclusion

As the real estate market continues to progress and housing becomes more sought after, variations in pricing will further persist. In times of booms or crashes, home owners, overlooking the market, would need this tool to ensure their estate is priced at a competitive value. This tool will eventually include large cities such as Boston, Los Angeles, San Francisco, Houston, Dallas and more, improving the overall versatility. With the inclusion of more cities, More regression models and features will be added in hopes of further improving the accuracy for the end user.

References

- “3.3. Metrics and Scoring: Quantifying the Quality of Predictions.” Scikit Learn, scikit-learn.org/stable/modules/model_evaluation.html.
- Chauhan, Nagesh Singh. “A Beginner's Guide to Linear Regression in Python with Scikit-Learn.” Medium, Towards Data Science, 25 Feb. 2019, towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f.
- Girgin, Samet. “Support Vector Regression in 6 Steps with Python.” Medium, PursuitData, 22 May 2019, medium.com/pursuitnotes/support-vector-regression-in-6-steps-with-python-c4569acd062d.
- Gomonov, Denis. “New York City Airbnb Open Data.” Kaggle, 12 Aug. 2019, www.kaggle.com/dgomonov/new-york-city-airbnb-open-data.
- Khan, Ruman. “Understanding Optimization in ML with Gradient Descent & Implement SGD Regressor from Scratch.” Medium, Medium, 3 Nov. 2019, medium.com/@rumankhan1/understanding-optimization-in-ml-with-gradient-descent-implement-sgd-regressor-from-scratch-4e11dac74c9.
- “Sklearn.linear_model.LinearRegression.” Scikit Learn, scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.
- “Sklearn.linear_model.SGDRegressor.” Scikit Learn, scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html.
- “Sklearn.svm.SVR.” Scikit Learn, scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html.
- Turgut, Duygu. “Airbnb NYC Price Prediction.” Kaggle, 22 Oct. 2019, www.kaggle.com/duygut/airbnb-nyc-price-prediction.