Predicting Rent Prices for US Apartments

Final Report — Team 82

Walker Stevens, Jonathan Eaton, Jasmine Camacho, Graham Ellis, and Dylan Leonard

Overview of Project - Introduction

Rents are a central figure to measuring commercial real estate. Being able to predict rents has long been considered a non-trivial problem in the industry, as many factors need to be considered such as geography, proxies for desirability, and the pure quantifiable measurements of the unit. In this project we aim to predict rent prices for a host of rental properties in the US, which is crucial in the decision making of multiple players in the industry, including owners, tenants, developers, governments, and researchers.

Predicting rents is of critical importance to several key stakeholders in commercial real estate. Firstly, landlords benefit from pricing prediction as they directly profit from renting out their units to tenants. Owners would decide where to buy properties based on the predicted rents, as this indicates what areas would be more lucrative on a faster timeline. In the same vein, tenants themselves will also benefit from this knowledge by seeing how rents may be less in some areas and buildings despite having comparable features to more expensive predicted units.

Secondly, developers also benefit from knowing what the rent will be set at. Given that rent has a strong correlation to demand, developers would be able to have more targeted information and likely recoup more from where they invest. For example, they might see that an older building in a fast growing area is lagging and may decide to accept a contract to renovate or build a new structure in place, by being able to calculate how much the agreement change would affect the new rent figure.

In addition, studies like this directly help housing nonprofits and public research firms, because understanding how rents evolve and are priced helps determine optimal housing policy. Noticing that rents are ballooning can indicate a failure in zoning policy, public housing, etc. Researchers who do this analysis can and will lobby their local government based on this information.

Overview of Project - Literature Survey

In “Apartment rent prediction using spatial modeling” (Valente, J., Wu, S., Gelfand, A., & Sirmans, 2005), they describe using a novel approach for including location in a model for apartment rents. In this approach they explicitly specified spatial association as the distance between pairs of locations. We seek to use a similar approach by utilizing the longitude and latitude records in our dataset to create a number of discrete clusters by manhattan or euclidean distance between property locations. In determining which approaches to apply to modeling apartment rents, we found “Ensemble learning based rental apartment price prediction model by categorical features factoring” (Neloy, A. A., Haque, H. S., & Ul Islam, M. M., 2019) helpful in illustrating the merits and performance of a variety of different algorithms, including ensemble methods. In terms of individual algorithms they found Random Forest had the lowest RMSE compared to Linear Regression models with various regularizations, SVM, and Neural Networks. However, they achieved the best overall accuracy with an Ensemble Gradient Boosting Approach. We will be comparing similar models, and this paper validated focusing linear regression and random forest, particularly given our data sets are similar in their high number of categorical attributes.

Overview of Project - Initial Hypotheses

We hypothesize that the main factors in predicting rental prices will be location, number of bedrooms, number of bathrooms, and square feet. Since the dataset includes rental prices from multiple states throughout the US, we need to fit a regression model across several cities in particular states in order to determine which specific factors have the greatest impact on rental costs. Although we suspect that the square footage will be a major factor in helping us predict prices, we are also aware that prices will vary depending on the city and region, regardless of the square footage. An apartment located in a large city may cost substantially more than an apartment with similar dimensions in a small city so we will keep this in mind while selecting our variables.

In order to decide which variables should be included in our regression model based on associated factors to the price, we will need to create dummy variables and split our dataset. We can examine how geography affects pricing by examining trends in price variations between various cities and states. We believe stepwise regression will be a good method in selecting appropriate variables for our model and will in turn help us reliably forecast rental prices in certain areas.

Overview of Data - Dataset and Cleaning

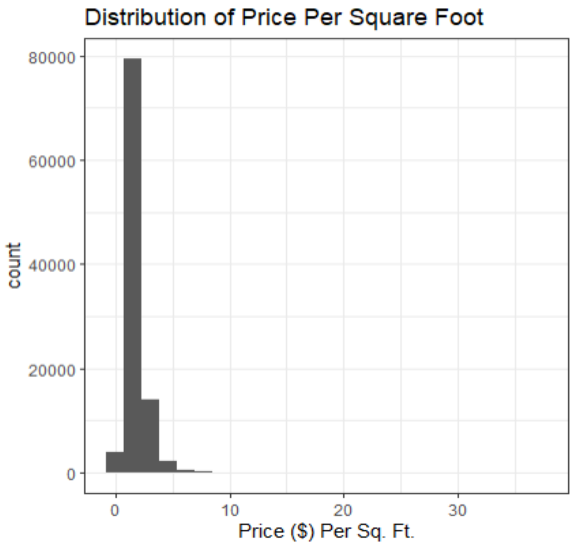
Our dataset contains 100,000 rows of rental property data from 2019. Features include property ID, description, amenities, and other essential property descriptors (e.g., internal area, number of bedrooms and bathrooms, etc.). There is also information about the city along with latitude and longitude. Importantly, each property also lists a rental price, which will be the focus of our analysis. The data comes from the UCI Machine Learning Repository, labeled as “Apartment for rent classified data set”. A condensed version of the data (10,000 rows) is available, but our team contacted the data owner Fredrick Nilsson for the expanded data.

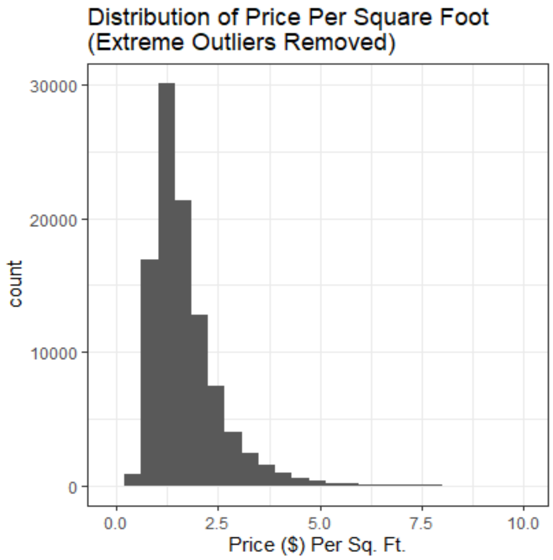
To more effectively utilize our data, we began by creating columns containing an indicator variable for each of the 28 possible amenities listed. This enables us to use the amenities as factors in our regression and for our random forest. Similarly, we created indicator variables for whether cats or dogs are allowed.

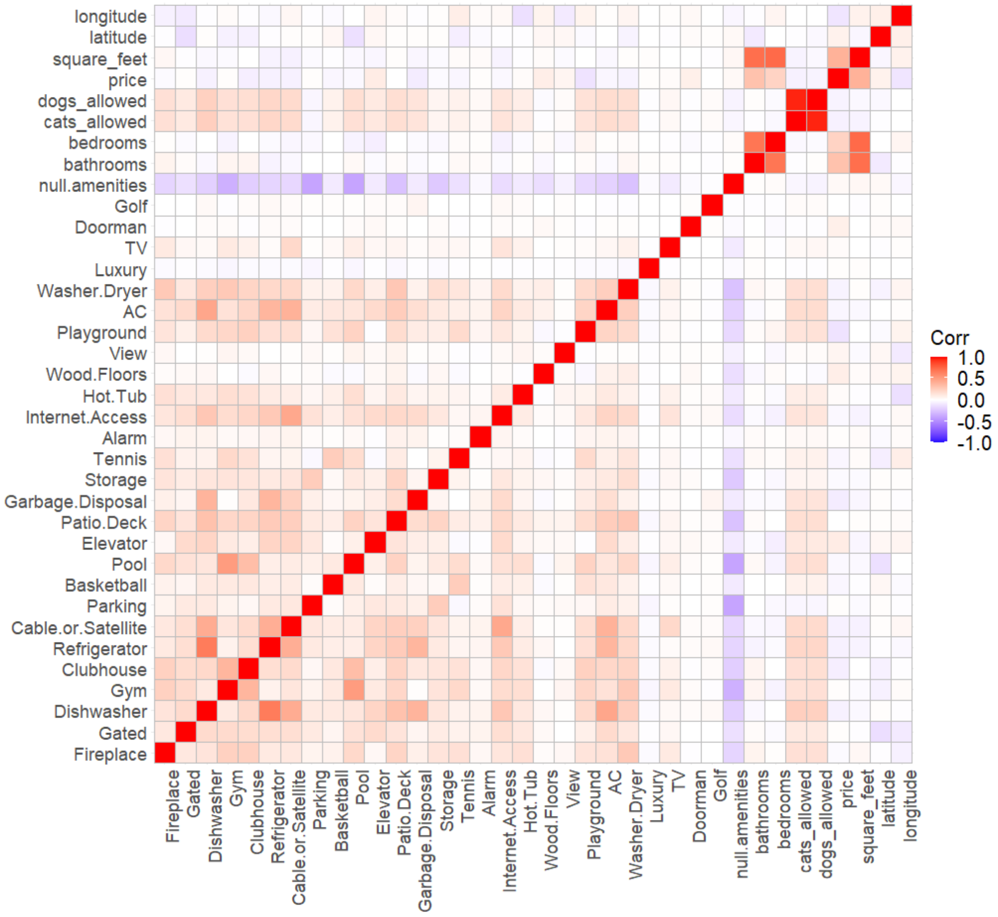
To clean our data, we removed any rows with null or NaN values in the price column. Additionally, 63 rows had null bathrooms count and 123 rows had null bedrooms count. Since most if not all building codes require a bathroom in a residence, we imputed the missing bathrooms data as 1 bathroom. We also imputed the missing bedrooms data as 0 bedrooms (i.e., studio apartment). Four properties were priced on a weekly basis, so we converted this price to a monthly cost by multiplying by 52 weeks and dividing by 12 months, which is a more accurate conversion than simply multiplying the weekly price by 4 weeks.

Several hundred listings clearly described the property as a studio apartment, but the number of bedrooms was listed as greater than 0. We believe this is because some descriptions included statements such as, “We have studio - 2 beds units available for rent,” even though the listing itself was for a studio apartment. The scraping algorithm appears to have not recognized that studio is equivalent to 0 beds, so in this example it would interpret the number of beds as 2 rather than 0. We corrected this by changing the number of bedrooms for all studio apartments to 0.

We also removed extreme outliers in price, square feet, and price per square foot. Given the nature of housing data, each of these fields is right-skewed. Since they are bound by 0 at the bottom and have no theoretical upper bound, there are several outliers on the high end of the range. Some of these may be due to data entry errors, and some may be luxury properties or corporate rental spaces that are not representative of the population of properties we wish to analyze. After careful analysis, we removed the 47 most extreme outliers. The histograms below show the distribution of the price per square foot data before and after outlier removal. From these charts, we can see that removing these outliers yields a dataset that is more representative of our target population.







In addition to examining outliers, we explored the correlation between the variables. The plot to the left demonstrates the correlation all the features have to one another, and to the target variable “price”. Intuitively, we can understand why bathrooms, bedrooms, and square feet all have positive correlations with price. Simple logic would imply that the larger the place is, generally the more expensive it will be. In addition, we can see which categorical variables are highly correlated, such as cats_allowed and dogs_allowed or

Refrigerator and Dishwasher. Although there is correlation between some of the variables, multicollinearity does not appear to be a significant problem.

Overview of Data - Feature Engineering, Geographical Clustering

In order to account for apartment location in our models of apartment rental prices, we used k-means clustering, based upon the longitude and latitude records for each geographic area: West Coast, Northeast, and Southeast. The number of clusters used in k-means was determined by an "elbow" plot of total within-cluster sum of squares versus cluster number and gap statistic versus number of clusters. We could then build models of apartment rental prices for each geographic area using the clusters as a factor variable. This improved the performance of our models versus baseline models that did not include location variables, and was more interpretable than including each individual city and state as a factor variable. The attached exhibit shows west coast apartment listing locations by the 13 clusters used for that data set.

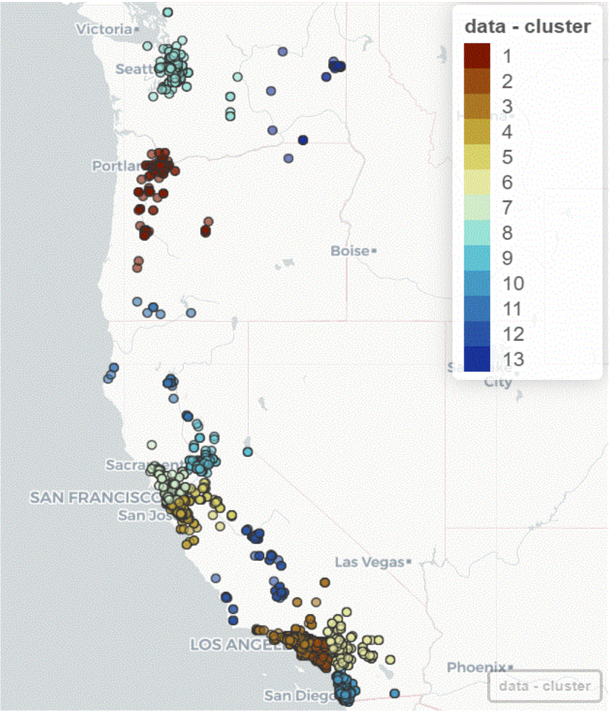


Exhibit: West Coast Location Clusters

Overview of Models - Models and Approach

We chose to use linear regression and a random forest for our models in order to determine which model will have the most accurate results. The data was randomly split into training and test datasets to create our models and determine which are most effective. We experimented with linear and stepwise regression for variable selection, as well as testing different numbers of trees for our random forest model. We also tested several log transforms of our linear model, and we found that the log-linear model provided the most accurate predictions. This is likely because rent prices have a large spread and are right skewed, so using a logarithmic transform brings the values closer together.

Linear models have the advantage of being easily interpreted by breaking down the values associated with each predictor. On the other hand, random forest models offer no specific insight into variable coefficients. However, they are often robust and can outperform linear models. We will compare the outcomes of these models later on.

Our models provide a significant advantage over existing rent estimation tools. Current publicly available tools do not provide the user significant insight into the details of the analysis. Most simply have the user input the address of the property, and then the tool spits out a rent estimate. This does not allow the user the ability to apply more granular detail in their pricing analysis. For example, a landlord may be interested to see what amenities they could offer to increase the value of their offering. Would adding on-site laundry or allowing pets give them an edge in their pricing? Our models allow fine-level control for property owners to see where the most value lies, while also letting tenants see what the cost drivers are in their monthly rent. Our linear model would be most useful in providing specific insight into this data, but even with the black-box random forest approach, landlords could still get a predicted rent price using specific property details far beyond what other tools allow.

Another way our models exceed current tools is that landlords could use this for a break-even analysis on renovation investment. For example, if they were considering spending $500 on a dishwasher, they could use our models to check the expected rent with and without a dishwasher. If adding a dishwasher increased expected rent by $50 per month, for instance, then they could see that they would break even within 10 months. This type of tool is novel and not widely available currently.

Overview of Models - Variable Selection

The variables that most significantly affect rental pricing must be chosen in order to develop an appropriate model. Our initial presumption was that the square footage, number of bedrooms, number of bathrooms, and location would all be significantly connected with the cost of an apartment. By using the information we have gathered from those variables, we can provide predicted rental prices. We used decision trees and random forest models to identify the variables that are most significantly correlated with price in order to verify this hypothesis. We constructed binary dummy variables to divide up some of the data because the dataset originally had 22 variables, 11 of which were factor variables. We chose not to include the city and state variables in the model since they are factor based with several levels. Instead, we used the cluster, latitude, and longitude variables to represent location in the selection process. Three separate datasets were created for each of the three main US regions (northeast, south, and west) and a cluster variable was added based on latitude and longitude of apartment listings in each independent geographic area: West Coast, Northeast, and Southeast. Many binary variables were formed during the data cleansing process and were extracted from the amenities column; as a result, we decided to use a number of these binary variables in the decision tree and random forest models. Based on the suggested variables from stepwise regression, we eliminated several of the amenities from our end model.

Overview of Models - Linear Regression

South Region Linear Regression

For our regression on the South region, the output suggests that various amenities, as well as the number of bathrooms, bedrooms, and square feet, have a significant impact on the logarithm of the price of the apartment. Specifically, Gym, Refrigerator, Parking, Pool, Elevator, Garbage.Disposal, Internet Access, Hot Tub, Wood Floors, View, Playground, AC, Washer/Dryer, TV, Doorman, and cats allowed are positively correlated with the logarithm of the price of the apartment. The coefficients of the remaining amenities, Gated and Basketball, are not statistically significant, implying that these amenities do not have a significant impact on the price of the apartment. The coefficients for the different clusters are also significant and vary across clusters, which suggests that the location of the apartment has a significant impact on its price.

The model yielded an R-squared value of 0.69 which indicates that the model explains 69.4% of the variation in the dependent variable, log price. This means that the independent variables in the model can explain about 69.4% of the variation in the log of the rent prices of the apartments. It also yielded an RMSE of 505.6, meaning the average prediction of the model was off by an average of $505.60.

Northeast Region Linear Regression

The linear regression model shows that the apartment price is positively associated with the number of bathrooms, bedrooms, square footage, doorman, golf, dogs allowed, and the presence of amenities such as gym, elevator, patio/deck, storage, pool, and internet access. The coefficients for some variables were not statistically significant at the 5% level, including Luxury.

The R-squared value for the model is 0.694, which means that the predictor variables explain 64.4% of the variability in the logarithm of price in the dataset. The residual standard error is 0.2269, which is the estimated standard deviation of the residuals, or the average distance between the observed and predicted logarithm of price. It also yielded an RMSE of 513, meaning the average prediction of the model was off by an average of $513.

West Coast Regression

Several predictor variables were found to have a statistically significant relationship with the log-transformed rent prices at a 95% confidence level. The predictor variables that had a positive impact on the rent prices were Gym, Parking, Elevator, Storage, Tennis, Internet Access, View, Washer/Dryer, Doorman, null amenities, bathrooms, bedrooms, and dogs allowed. The predictor variables that had a negative impact on the rent prices were Gated, Cable or Satellite, Basketball, Garbage Disposal, Wood Floors, AC, and cats allowed. Some of these don't make logical sense considering AC would not make a house's value go down.

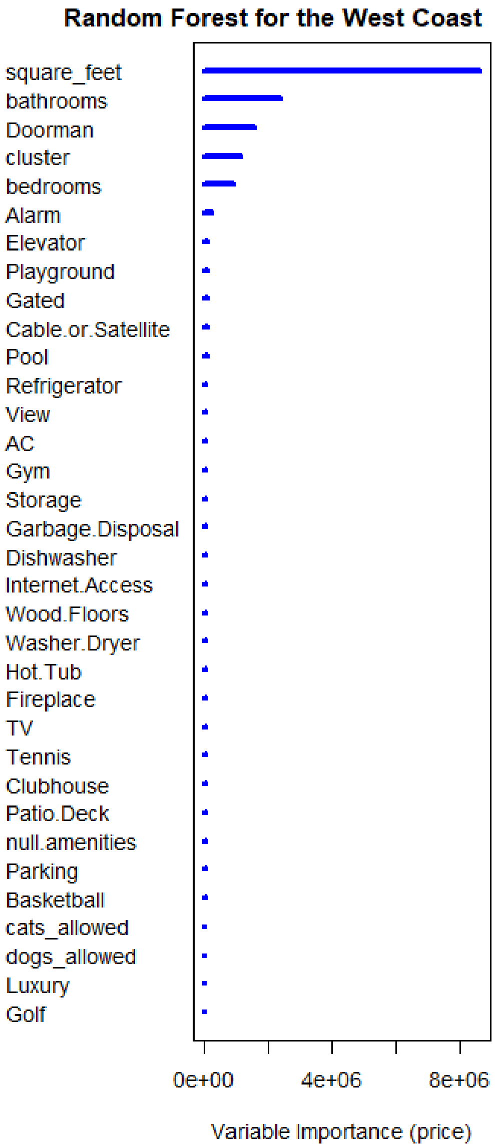
The R-squared value for the model is 0.6598, which means that the predictor variables explain 65.98% of the variability in the logarithm of price in the dataset. The residual standard error is 0.236, which is the estimated standard deviation of the residuals, or the average distance between the observed and predicted logarithm of price. It also yielded an RMSE of 740.91, meaning the average prediction of the model was off by an average of $740.91, which is the worst of the three tests.

Overview of Models - Random Forest and Decision Trees

We first started by creating a decision tree using various variables including: bathrooms, bedrooms, square feet, cluster, and amenities. After testing several different variables (out of the 55 present in the dataset), we found that the decision tree only selected square feet, cluster, and bathrooms to construct the tree. Before conducting a random forest model, we split 70% of the data into a training set and 30% into a test set. The same variables were applied to the random forest model and we found similar results in regards to the relative importance. The random forest model identified that bathrooms, square feet, and bedrooms had the most relative importance in relation to the price. We noticed that most amenities held very little significance to the price with the exception of a doorman. It appears that amenities do not have a significant effect on rental prices.

The same random forest model and prediction method were applied to each of the three regional datasets. The variables tested in the linear regression model were also used to test the decision tree and random forest models. The variables that showed the most relative importance to the price were essentially the same for each model. One decision tree and one random forest model were conducted for each dataset for comparison. The top five variables the model predicted had the most significance to the price are square feet, bathrooms, bedrooms, cluster, and doorman. The random forest model did help us confirm our prediction about which variables will affect the price, however, we were surprised to see the doorman be such a big predictor in the model since this variable was not included in our hypothesis. The final models consisted of only the five most significant variables in order to give us the most accurate predictions.

All three decision trees chose square feet, bathrooms, cluster, and bedrooms as decision points with the square feet being the largest factor for the model. This model shows that larger square footage generally increases the rental prediction prices, which aligns with our hypothesis. The cluster value represents the geological location within a region and this variable was included in the model. We assumed location would have a larger impact on the price, however, the cluster value did not show a clear correlation of increase or decrease in prices. The decision trees help with interpretation but are not the most reliable or accurate models for prediction. The random forest models performed better than the decision trees and displayed the predictors with the highest relative importance to the price. Similarly to the decision tree models, all three random forest models identified square feet as one of the highest importance values, which was to be expected. This confirms that square feet, number of bedrooms, and number of bathrooms are good predictors to help us forecast rental prices.



The R-squared value for the random forest model was typically around 0.69, meaning the independent variables are able to explain about 69% of the variability of the price variable in the random forest model. We calculated the mean absolute error (MAE) to determine the accuracy of the model. MAE is a measure that takes the average error between the actual and predicted values. The MAE ranged between 230 and 360, indicating that the model's predictions were typically off by a few hundred values on average. The table below shows a small sample of values from each dataset that were tested to predict prices and compare the results to observed values. By observing the differences, we can see the variability in accuracy of its predictions.

	South			Northeast			West		
Actual Prices	$1740	$1160	$1649	$4500	$1950	$999	$2150	$4950	$2548
Predicted Prices	$2429	$1355	$1374	$4400	$1326	$1232	$3701	$4928	$3799
Difference	**-$689**	**$305**	**$275**	**$100**	**$624**	**-$233**	**-$1551**	**$22**	**-$1251**

Overview of Models - Model Performance Comparison

To compare our linear and random forest models, we utilize mean absolute error (MAE), mean absolute percent error (MAPE), and R-squared. The results are displayed in the tables below.

West

Model	**MAE**	**MAPE**	**R^2**
Linear	423.6748	0.1789012	0.6598
Random Forest	351.1774	0.1520619	0.6856

Northeast

Model	**MAE**	**MAPE**	**R^2**
Linear	362.013	0.1816963	0.694
Random Forest	325.3006	0.1727545	0.6942

South

Model	**MAE**	**MAPE**	**R^2**
Linear	278.6035	0.1916694	0.69
Random Forest	230.1543	0.165017	0.6942

While the R-squared values are very similar for both models in each region, the random forest outperformed the linear model when compared using the MAE and MAPE. The random forest's better performance can be attributed to its ability to capture non-linear relationships and feature interactions, which are often not accounted for in linear models. Although the linear model has the advantage of ease of interpretation, the random forest's higher accuracy makes it more useful in achieving the end goal.

Conclusions

Using our random forest model, on average we can predict US rent prices within about 15 - 17%, depending on what region of the country we are analyzing. A key takeaway from our analysis is that regional differences in rent pricing make it extremely difficult to create a generalized national model to predict rent prices across the US. For example, Los Angeles obviously has higher rent prices than rural Texas, but even within the city of LA, pricing can vary dramatically. A hyper-localized model analyzing a single city or area within a city could be much more accurate because the effect of other external variables would be minimized.

Additionally, the dataset was fairly large (about 100,000 rows), but with each data point spread across the country, some areas had less robust data than others. That means the model may have been better fit to areas that had more rent data available. Rural areas or other regions with less data represented in our dataset likely had less accurate predictions as a result.

Overall, we are able to predict rent prices with reasonable accuracy. Further improvements could be made by expanding our dataset or narrowing the scope of our analysis to a specific region. Possibilities include localizing the model to census blocks or multifamily submarkets to negate some of the effects of features which contain considerable minutia, such as traffic, school availability, and nearest public transit. Alternatively, we could scrape additional data to account for these features and expand the geographical scope.

Works Cited

Neloy, A. A., Haque, H. S., & Ul Islam, M. M. (2019, February). Ensemble learning based rental apartment price prediction model by categorical features factoring. In Proceedings of the 2019 11th International conference on machine learning and computing (pp. 350-356).

Valente, J., Wu, S., Gelfand, A., & Sirmans, C. F. (2005). Apartment rent prediction using spatial modeling. Journal of Real Estate Research, 27(1), 105-136.