Housing Prices

I. A. Karp

1 Introduction

Over the past decade King County Washington, home to the city of Seattle, has been on of the fastest growing regions in the country gaining over 300,000 people in that time, and has struggled to meet the subsequent housing demand[1]. This has fueled a thriving housing market, making the accurate prediction of housing costs crucial for both developers and buyers. I will use a linear model to investigate which attributes play the biggest role in a house's value.

2 Methods

2.1 Dataset

Our dataset includes over 20,000 sold houses in King County from 2014 to 2015[2] which we have randomly split into 10,000 points of training and 10,000 of test data. In addition to the price and location of the house, the set also contains detailed information on floor space, build quality, and view as well as some information on the surrounding houses. The set also includes the year in which renovations took place (with 0 as the default) which I transformed into a binary variable telling whether the property had been renovated since 1990. I also used k-means clustering on the testing data to transform the zip-code variable into three groups indicating a cheap, normal, or expensive neighborhood. The algorithm correctly identified the areas of Medina and Mercer Island as expensive which are where Bill Gates and Jeff Bezos live respectively and where I always went with my friends on Halloween to get the most candy.

2.2 Model Selection

2.2.1 Feature Selection

I began with a full model containing fifteen unique variables, six interaction terms selected based on personal interest, and a squared term for the Living Space variable. I then used BIC as my selection criteria, checking all possible subsets of the full model. I preferred BIC over other metrics like AIC as I would like to be more aggressive and get a better sense of which variables are truly important. The final model includes fourteen unique variables, and five interaction terms.

2.2.2 Outlier Handling

I flagged twenty points with a Cook's distance above $\frac{4}{n-2}$ and investigated each case-by-case. Only properties with either incorrect data entry, or highly abnormal circumstances were removed from the data set. Some of those removed include a small single-story home suspiciously claiming to have thirty-three bedrooms; a property purchased at a discounted rate for the construction of an elementary school; and a property whose view of Lake Washington and Mount Rainier brought a tear to my eye, yet was labeled as having no view at all. Many others had simply dropped a digit from their square feet of Living Area which can only be expected in such a large dataset. Those retained were generally either near a highway or airport, or simply ugly, variables which our model is unable to capture.

2.2.3 Weighted Residuals

Our diagnostics demonstrated a significant degree of heteroskedasticity. Since we would like to use the model for statistical inference, we opt for a weighted model, giving less weight to high residual

points. This means that the most expensive properties have little impact on the model and thus the following results will be most relevant to houses near the average house price of about $550,000.

3 Results

Table 1 Regression Results after weighting and BIC

	Price Coefficient (±95% CI)
Bedrooms	$-12,743.550^{***}$ (1,418)
Bathrooms	8,481.328 (5,372.148)
Lot Area (ft^2)	0.499^{***} (0.092)
Floors	$-1,447.787$ (7,873)
Waterfront	$-134,822.900^{***}$ (52,170)
View (1-4)	$36,037.570^{***}$ (2,794)
Condition (1-5)	$81,730.250^{***}$ (5,858)
Grade (1-13)	$75,975.980^{***}$ (3,915)
Year Built (After 1900)	$-1,321.564^{***}$ (48.680)
Nbrhd. Living Area (ft^2)	29.574^{***} (2.862)
Renovated Since 1990	$-55,993.620^{***}$ (16,300)
Expensive Zip-code	$374,273.900^{***}$ (21,190)
Cheap Zip-code	$-177,350.600^{***}$ (2,782)
Living Area (ft^2) x	-35.579^{***} (7.646)
Living Area (ft^2) x^2	0.034^{***} (0.002)
Lot Area × Living Area	-0.0001^{***} (0.00004)
Bathrooms × Floors	$9,991.040^{***}$ (3,628)
Waterfront × Living Space	225.964^{***} (32.61)
Renovated × Living Area	57.687^{***} (10.91)
Condition × Grade	$-10,558.550^{***}$ (927.8)
Constant	$-66,361.620^{**}$ (26,430)
N	9,986
R^2	0.693
Adjusted R^2	0.692
Residual Std. Error	1.443
F Statistic	$1,124.457^{***}$

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

3.1 Removed Terms

Through the model selection process, the square feet of basement space, and the average lot space of the surrounding houses were deemed unnecessary. This likely stems from the fact that the area of the basement can be approximated as a function of the living space and number of floors and thus it is unnecessary when these two terms are included. The information gained form including the neighborhood lot space is also likely sufficiently explained by the neighborhood living space and neighborhood price.

3.2 Model Diagnostics

From the residual and fitted values plot we still see a degree of heteroskedasticity, though within a reasonable bound. Further, the flatness of the fitted line indicates that our data is well-approximated by the model. We additionally notice from our Q-Q plot that our residuals have a heavy-tailed distribution, though again this is within reasonable bounds. Finally, the residuals and leverage plot demonstrates that the model contains few influential outliers as few points are outside an acceptable range of Cook's distance.

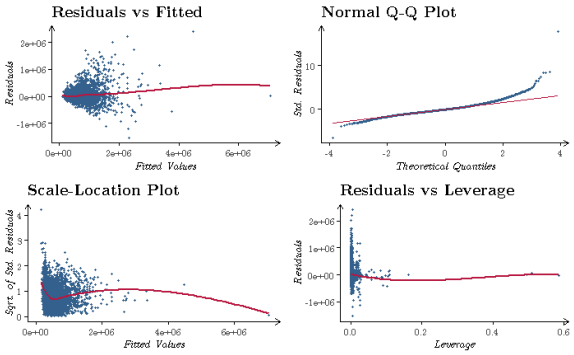

Fig. 1 Model diagnostics plots.

3.3 Coefficient Estimates

3.3.1 Living and Lot Area

We find the Living Area makes up a significant proportion of a house's value to the extent that all other variables can be considered as caveats to it. We additionally find no example of diminishing returns on higher square footage. However, as we see in the plot below, very few houses in the dataset reach a large enough size for this observation to be generalizable. Lot Area, on the other hand, has a much smaller positive impact on price, while the negative interaction term between lot and living area indicates that properties are penalized for having high values of both. From a developers perspective, this incentivizes using as much area as possible for living space.

3.3.2 Year Built and Renovated

Strangely, both the coefficients of the year built and whether the property has been renovated since 1990 are negative. However, as we see from the interaction term, a house having been renovated recently significantly increases the predicted value of each square foot of living area such that the negative term for renovation simply acts as a corrective constant. The fact that newer houses tend to be cheaper is also not surprising as recent growth trends have incentivized developers to use

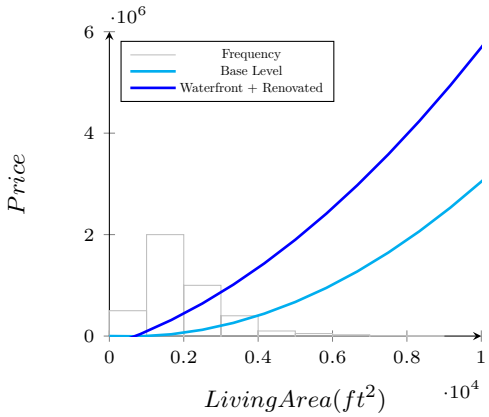

Fig. 2 Predicted price from Living Area with additional interaction terms and histogram of frequencies.

expensive land for large-scale apartment construction so that new houses are only built on cheap land.

3.3.3 Floors, Bedrooms, Bathrooms

From the combination of interaction and standalone terms we find a beneficial effect from having many bedrooms and bathrooms together, so long as the home has at least two bathrooms to begin with. In addition to this, homes with a large combined number of floors and bathrooms show significant increases in price. This may indicate that developers shouldn't waste too much living space on things like living rooms or home offices and must ensure that a home can sustain as many people living in it as possible.

3.3.4 Waterfront and Location

We find that a property being on the waterfront significantly increases the price per square foot of living area so long as the property is larger than 600ft^2 (which is almost always the case). Further, we find that the zip-code of a property can add almost four-hundred-thousand dollars to the price for an expensive area, or remove almost two-hundred-thousand for a cheap one compared to base level. In addition to this, the average area of the surrounding houses adds a considerable amount the expected price, as much as a million dollars in the most extreme case.

3.3.5 Condition and Grade

The Condition variable measures what state a house is in such as whether it needs repairs, while the grade measure the underlying quality of the structure and construction. The coefficient estimates indicate that grade can have a much higher impact on the housing price, as a maximum grade of 13 is estimated to add over a million dollars to the price while a maximum condition is estimated to add only half a million. However, the negative interaction terms between these two variables indicates a degree of diminishing returns, where a grade above 7 (which is coincidentally the standard) stops improving the price. Alternatively, we see that most homes in the training data which have a high grade also have a very large living area, such that this negative term may simply be correcting for an overestimation in the estimate of the living area coefficient. Either way, it seems that for mid-price homes the extra effort of improving the grade of a home is not worth the cost.

3.4 Prediction

For a test of the reasonableness of the model, we check to see which property in the county it deems most valuable. Out of the over 10,000 properties in the test set we find that 26408 NE 70th St in Redmond is the highest valued property by the model at a predicted $6,543,623. With eight bathrooms, seven bedrooms, 13,540 square feet of living space, and 307,752 of lot space this seems like a reasonable estimate. However, this property is actually in a very remote area adjacent to a luxury horse resort (a resort for the horses themselves, not for people to ride them) and was only sold for $2,280,000.

Fig. 3 A current photo of 26408 NE 70th St.

The actual most expensive house, 1137 Harvard Ave E in Seattle, was predicted to have a price of $5,623,573 while in reality sold for $7,700,000 due to its impressive architecture and amenities like a pool and basketball court, as well as a valuable location near downtown. Thus, it appears that our model is not unreasonable in its estimates, but is incapable of taking into account many common circumstances.

Fig. 4 A current photo of 1137 Harvard Ave E.

4 Discussion

4.1 Model Limitations

The key weakness of the model come from it's lack of location information, which is one of the most important factors in property evaluation. While we do have some information on a zip-code level, variables such as proximity to schools, highways, airports, grocery stores, and bus lines would be incredibly helpful, but also burdensome to acquire. We additionally lack any variables which measure how developed the properties lot space is. For example, a large lot on one property may simply be an open field, while in another it's a pool and a basketball court. Overall, the model gives some key insights into which attributes are valuable in a property at a base level, but would be quite insufficient for predicting the value of any specific house.

4.2 Alternative Models

The biggest potential improvement to the model would come from treating each discrete variable, like View Quality and Condition, as distinct levels instead of numeric values. However, this would greatly decrease the model interpretability and increase the complexity of what is already a rather cumbersome model. We could also consider adding more interaction terms, but this would result in similar issues. A more advanced non-linear model would likely perform better, but ours proves strikingly effective when it comes to useful interpretation.

5 Conclusion

Overall, our model performs rather well for a linear model on a highly qualitative data set. If R^2 is to be believed, the model explains as much as 69% of the observed variance in the data, which for a subject as potentially volatile as housing prices is not too bad. However, our model is missing many key details that while sufficient for discussing general housing trends makes it unsuitable for case-by-case prediction. In spite of this, we are able to conclude that living area and location are the most important factors for a house's price. We additionally find that developers should leverage the value of their waterfront properties by building them with as much living space as possible. Further, the actually quality of the house's construction is not important so long as it meets construction standards. Finally, it is important to include as many bedrooms as possible in a home while keeping the number of bathrooms as close to 1:1 as possible. In conclusion, retail developers today should adopt a similar strategy to New York tenement owners, cramming as many people into as little a space as possible while keeping quality as low as possible.

References

[1] Gutman, D., Shapiro, N.: Seattle grew by more than 100,000 people in past 10 years, king county population booms, diversifies, new census data shows. Seattle Times (2021)

[2] Laz, L.: House Sales in King County, USA. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction/data (2021)