ANLY 533 – Regression Module Assignment

Predicting Home Prices Using the Zillow API and Multiple Regression

Analyst: Adam Sampson

Contents

[Introduction and Business Problem 2](#_Toc497343126)

[Assumptions and Methodology 2](#_Toc497343127)

[Description of the data: 2](#_Toc497343128)

[Data Preparation: 2](#_Toc497343129)

[Modeling Process: 3](#_Toc497343130)

[Results 3](#_Toc497343131)

[Results without lastSoldPrice: 3](#_Toc497343132)

[Results with last sold price: 4](#_Toc497343133)

[Interpretation: 4](#_Toc497343134)

[How well the model works: 4](#_Toc497343135)

[Improvements needed: 4](#_Toc497343136)

# Introduction and Business Problem

The problem presented was to find a fair market price for a target property, in this case 1420 Christy Ave, 40204. For a startup real-estate firm this is very useful data. It can be used to determine whether to buy a property to flip, what to advise a client to set their price at, or even which markets to service (since earnings are based on % commission).

This problem has been limited to one address. However, the code to come to this conclusion is documented and ready to run future analyses.

# Assumptions and Methodology

## Description of the data:

Data has been collected using the ZillowAPI. In particular the data has come from the GetDeepComps and GetUpdatedPropertyDetails APIs with more details at https://www.zillow.com/howto/api/GetDeepSearchResults.htm.

The data was cleaned so that only numerical or categorical data was left. In other words, any information that was a link or a text block with owner description was discarded for this analysis due to complexity.

Since zip code was the same for all the comps, this value was not included in the analysis. If an analysis was done of the whole city this value would need to be put back into the analysis.

Data that was used for the analysis includes: Zestimate (as a target), last sold price, zindex value (a measure of the property values in the total neighborhood), tax assessment, lot size, finished interior size, number of bathrooms, number of bedrooms, total number of rooms, total number of floors, exterior material, architecture, basement type (if any), roof type, parking type, heating sources, heating system, and floor covering.

## Data Preparation:

In order to perform a regression, it was necessary to determine which variables were numeric (interval) data and which were categorical. All text answers were categorical.

The number of bathrooms, bedrooms, total rooms, floors, and covered parking spaces was used as categorical. This was done primarily because each of these was not continuous (you can’t have 1.1 bathrooms). Additionally, these predictors were not normally distributed which is necessary for continuous numerical values.

The remaining variables were analyzed for normality. A number of the variables were skewed and were adjusted using a log transformation. These were tax assessment, lot size, finished interior size.

Additionally, an attempt was made to use data from the US FED (in St Louis) to capture the change in median US home price from the time of the assessment and from the time of the last sale. The model was analyzed with the non-adjusted values and with the adjusted values.

Finally, an attempt was made to impute (predict) the missing values of total rooms and number of floors using the number of bathrooms and bedrooms to prevent problems in the model from missing values.

## Modeling Process:

The modeling was performed for two different cases. The first case was that the value of the last sale was not known. This was done in an attempt to handle future cases where the real estate firm may be interested in a property was sold a long time ago, or where no records exist (such as in a new sale). The second case was where the value of the last sale is known.

The data was divided into two data sets. 80% went into a training set and the remaining 20% went into a test set.

The data from the training set was all put into a linear regression calculation and checked for problems. The first problem was that there were too many missing values in many of the categorical values and they had to be removed. Next, variables that had high collinearity (they related with each other too highly) were removed. The Cook’s cutoff was calculated and the data was plotted to see which variables were outliers or had too much leverage on the model. These data points (rows) were all removed. Finally, the model was checked for collinearity again and the final model was produced.

Once the model was created it was used to predict the Zestimate. The Zestimate was selected because the last sold price may have been some time in the past and no longer apply in today’s market. Also, the Zestimate and the last sold price are highly correlated in this data set so not much will be different. However, because of this high correlation it was necessary to do one model without the last sale price. Since this data set has all recently sold properties any properties that haven’t sold in a long time should consider using the model without sale price.

# Results

## Results without lastSoldPrice:

The model for calculating Zestimate (without using last sold price) is:

Zestimate = 16880\*ln(lot size) + 68240\*ln(finished interior size) +

0.3437\*zindex value (aka neighborhood index) +

0 if exactly 1 bathroom + 20650 if exactly 1.5 bathrooms +

9992 if exactly 2 bathrooms + 55460 if exactly 2.5 bathrooms +

18200 if exactly 3 bathrooms – 0 if exactly 1 bedroom –

10950 if exactly 2 bedrooms – 27.29 if exactly 3 bedrooms –

43700 if exactly 4 bedrooms + 98550 if exactly 5 bedrooms - 459500

This model explains 70.83% of the variation in the training set (adjusted R squared).

Predictions made with this model against the training data have an 86.92% correlation and have a mean absolute error of $19,928.

Predictions made with this model against the test data have an 74.93% correlation and have a mean absolute error of $28,601.

The Zestimate (actual) for 1420 Christy Ave was $167,547. Or prediction with the first model was $174,466.

## Results with last sold price:

The model for calculating Zestimate (using last sold price) is:

Zestimate = 0.8115\*last sold price + 12560\*ln(finished interior size) +

0.1478\*zindex value (aka neighborhood index) – 71570

This model explains 97.4% of the variation in the training set (adjusted R squared).

Predictions made with this model against the training data have an 98.7% correlation and have a mean absolute error of $6,714.

Predictions made with this model against the test data have an 96.3% correlation and have a mean absolute error of $10,746.

The Zestimate (actual) for 1420 Christy Ave was $167,547. Or prediction with the second model was $174,005.

# Interpretation:

## How well the model works:

The first model had a MAE of $28,601. This is an enormous amount of error possible. However, for our property of interest the error was only $7100. It is possible that this may be due to the fact that the property being evaluated cannot truly be compared to the most expensive house in the data set which is $364,216. The mismatch here is likely due to taking too many comps recursively.

The second model had a MAE of $10,746. This is a much more comfortable model. Once again there is a large disparity in property prices.

## Improvements needed:

For the future, it would be excellent to filter comps which are way out of the price range of the target property. Without more data sources this will decrease the size of the data set substantially. One major improvement that could be made would be to pull similar valued properties from all over the city and then account for the difference in demand for each neighborhood. But significantly more data from other sources would be necessary to do this. Additionally, there are significant limits on the Zillow API.

Another improvement that could be made would be to calculate the neighborhood index from other data sources so that it is not necessary to rely as much on Zillow. This index can also be used to make better adjustments along with the housing market data from the FED in order to create a stronger model with less reliance on Zillow.

Ensuring that data is captured for all of the variables with missing data (exterior type, flooring type, etc.) would also help the model significantly. Unfortunately the first model is heavily limited by the variables that are missing too much data to be useful.