Zeth-Abney / **Flatiron-school** Public

Code Issues Pull requests Actions Projects Wiki Security

master Flatiron-school / phase-2 / PROJECT /

Zeth-Abney still fixing photos … 1 hour ago History

..	
LEARNCO	2 hours ago
data	2 hours ago
media	1 hour ago
.canvas	2 months ago
.gitignore	2 months ago
EDA.ipynb	2 hours ago
README.md	1 hour ago
model_dictionary.md	2 hours ago
student.ipynb	2 hours ago

README.md

Fliphouse, LLC. Regression analysis to inform purchasing decisions.

Author: Zeth Abney

Final Project Submission

Please fill out:

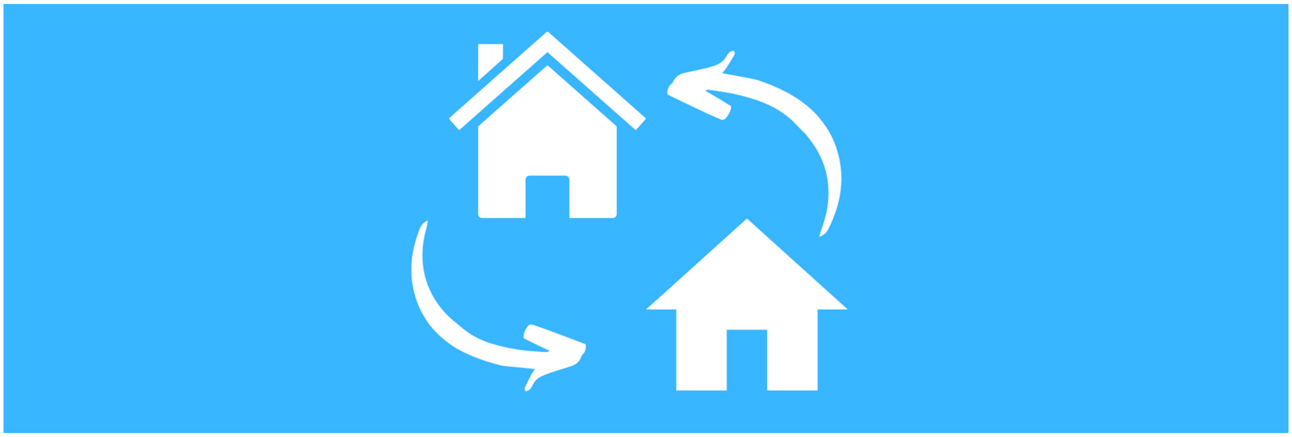
- Student name: Zeth Abney

- Student pace: full time
- Scheduled project review date/time: !!!TBA!!!
- Instructor name: Matt Carr
- Blog post URL: !!!TBA!!!

Overview

Fliphouse, LLC, a family owned business seeking to expand its market share is in the business of "flipping" houses. Fliphouse will strategically purchase a residential property, renovated it and then rent or sell it with the hopes of generating positive revenue by either renting or re-selling the property. Fliphouse seeks to develop a new competitive edge throughout this phase of expansion by becomming a data-driven company. For these reasons Fliphouse has decided to hire the consulting services of a data scientist to provide data-driven recommendations on how to enter and behave within the target market.

Business understanding



In this phase of expansion, Fliphouse seeks to enter the realestate market of the pacific northwest, and specifically Seattle Washington and the surrounding area (i.e. Kings County). Before begining any projects in the area, FlipHouse decision makers need to better understand how to determine the opportunity cost, and potential returns for any investments made in Kings County, as well as how to maximize those returns. The oportunity cost and potential returns can indeed be determined by understanding how time, physical location, and physical condition and attributes all affect the price of a real estate property.

Therefore, this anaylisis will seek to build a statistical model that is informative as far as specifiying what metrics to use and how strong each metric may be in terms of predicting the market value of a real estate property.

Data understanding

The data set used in this analysis is open-source data available directly from Kings County's website (https://kingcounty.gov/services/data.aspx). This particular data set covers various aspects of realestate transactions including date of sale, square footage of house and lot, proximity to recreational and natural resources, etc.

The initial data set used contains 21597 total records, aproximately 7% of which is eventually thrown out as a result of either data cleaning or model fitting. The dataset starts with 20 total features (i.e. columns), only 8 of which are ultimately included in the final model aslo with 6 additional features inferred from the original data (e.g. one-hot encoding).

For the purposes of regression modeling the data is manipulated so that every datapoint is encoded as either and integer or a decimal and there are no null or missing values. Also, the target variable 'price' is eventually log-transomed as part of the modelfitting process; keep in mind that because of the log-transformation the model coefficients should be interpreted as the percentages rather than the metric's own units. This is explained further in the regrssion results section of this notebook.

For more details on understanding the data and statistics of the final model see the model data dictionary

Data preparation

Throughout model development process about 7% of the original 21597 of the data is discarded.

Immediately upon import recards 453 are immediately eliminated due to some '?' values that could not be justifiably replaced with any sort of filler such as 0 or 'NONE'. By the end of the model development process there were 19982 records remaining eliminating about 7% of the initial data overall.

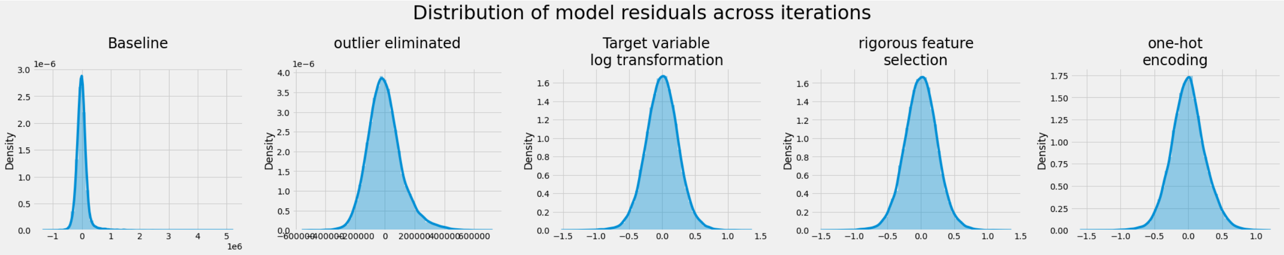
Then null values and non-numerical values are re-encoded so that all values in the data set are numerical, and there are no missing values. The details of these transformations are available in the Analysis notebook.

Next, in order to strengthen the model, outliers are eliminated. For this analysis an outlier is considered any record where the target variable (price) is greater than 3 standard deviations from the target variable mean. Using a statistical maximum (inter-quartile range * 1.5) instead was also investigated but it eliminated roughly half of the data which is liable to cause sampling issues.

In order to increase normality in both the target variable as well as the model itself the target variable is log transformed. It is important to note that because of this transformation any model coefficients should be interpreted is a change in percentage of the target variable, not in the fundamental units of the target variable (i.e. dollars).

At this point the model was strong, the set of features was somewhat limited. The 'zipcode' feature also still remained and was essentially uninterpretable due to the fact that it is not only categorical but nominal in nature. The next succesfuly better iteration was to extrapolate the proximity of a particular waterfront from the zipcode feature, and then one-hot encoding each waterfront location. The details of how this was performed is outlined below.

Modeling



An extensive iterative process was conducted to achieve the final model, to see this process in detail please review the exploratory analysis notebook .

The iterative process included the above described data engineering techniques, but also performing feature selecting and elimination at various points in the iterative process using a stepwise forward-bacward selection using a function found on stack exchange, recursive feature ranking and selection with cross validation based on coefficient strength from Scikit Learn, see the docs here, and finally elimination based on the variance inflation factors of the various features using the statsmodels outliers_influence module, see the docs here.

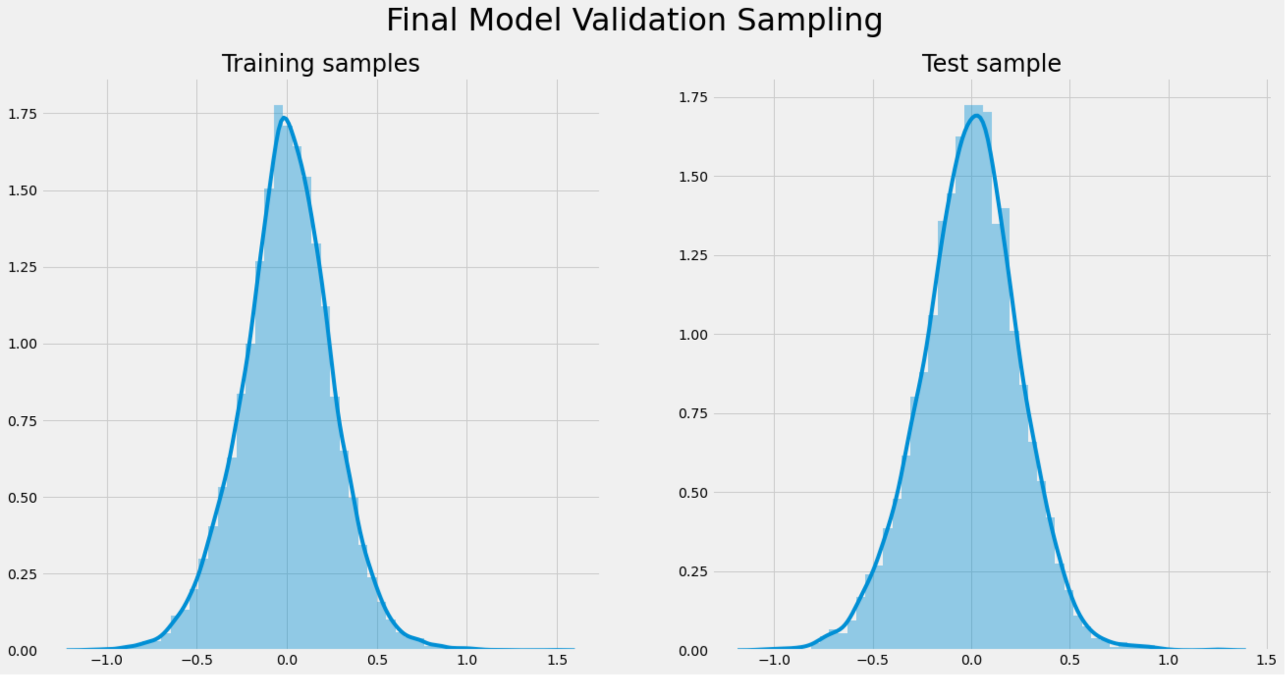
A small number of additional features are finally manually selected to be eliminated depending on how their absence affects aspects of the model most in need of improvement.

The final model appears to be fairly strong, although it is slightly leptokurtic and despite extremely low variance inflation factors throughout all of the model features, the condition number remains high indicating there may be some multi-colinearity present. If there is, its been extremely hard to locate.

I have confidence that this is a strong model because as best I can tell it satisfies the four assumptions of linear regression as best as it can with the data availble.

- When plotted the model residuals show strong resemblence to the t-distribution satisfying the normality assumption, save only for slightly long tails.
- The Drubin-Watson score is nearly 2.0 exactly, clearly demonstrating that the model satisfies the homoscedasticity assumption.
- The linearit assumption is difficult to validate with a single model alone, but I have been incresingly convinced that this model satisfies the linearity assumption thorughout the development of this model.
- The condition number is heavy which can be interpeted as an indaction of multi-colinearity. However there may be a scaling issue here, when the condion numebrs of each iteration of the model from the baseline up to the one in this notebook are plotted, the condition number of this model is 0 relative to its prototypical counterparts. Likewise for the Jarque Bera, and indication of normality.
- The kurtosis of this model is slightly outside the normally accepted range, throught the iterative process any attempts to address this issue drasticaly worsened all other indicators of model strenght. Additionally and kurtosis of 3.664 is the lowest of any iteration of the model so I have to run with it.

Further more after several iterations of test sampling the mean sqaured error of test samples are nearly exactly the same as their training sampel counterparts, indicating that this model not only fits well the sample of data used in this analysis but also the population of analogous data that really exists. This is illustated in the figure below.



Regression Results

The intention of this analysis to help Fliphouse, LLC. understand where to purchase properties as well as what physical aspects of the properties to consider before buying.

In regard to where to purchase properties, the latitude that a property lies on seems to be the strongest predictor of price. For every addition degree north, a properties price will increase by 1.47%. However the properties in this data set cover a range less than 1 degree. One degree of latittude is roughly 69 miles, and the range covered by the model data is only abouty 43 miles. So we can consider this to mean that aproximately for every additional mile north, the price of a property increases by 0.034%

Wether or not a property lies on a waterfront is a strong predictor as well, if it in fact does the price increases by 0.35%. Additionally if a property is even in the same zipcode as Lake Union the price of the property will increase by 0.31%. However property prices decrease if the proeprty is in the same zipcodes as Lake Washington, Duwamish, or Puget Sound.

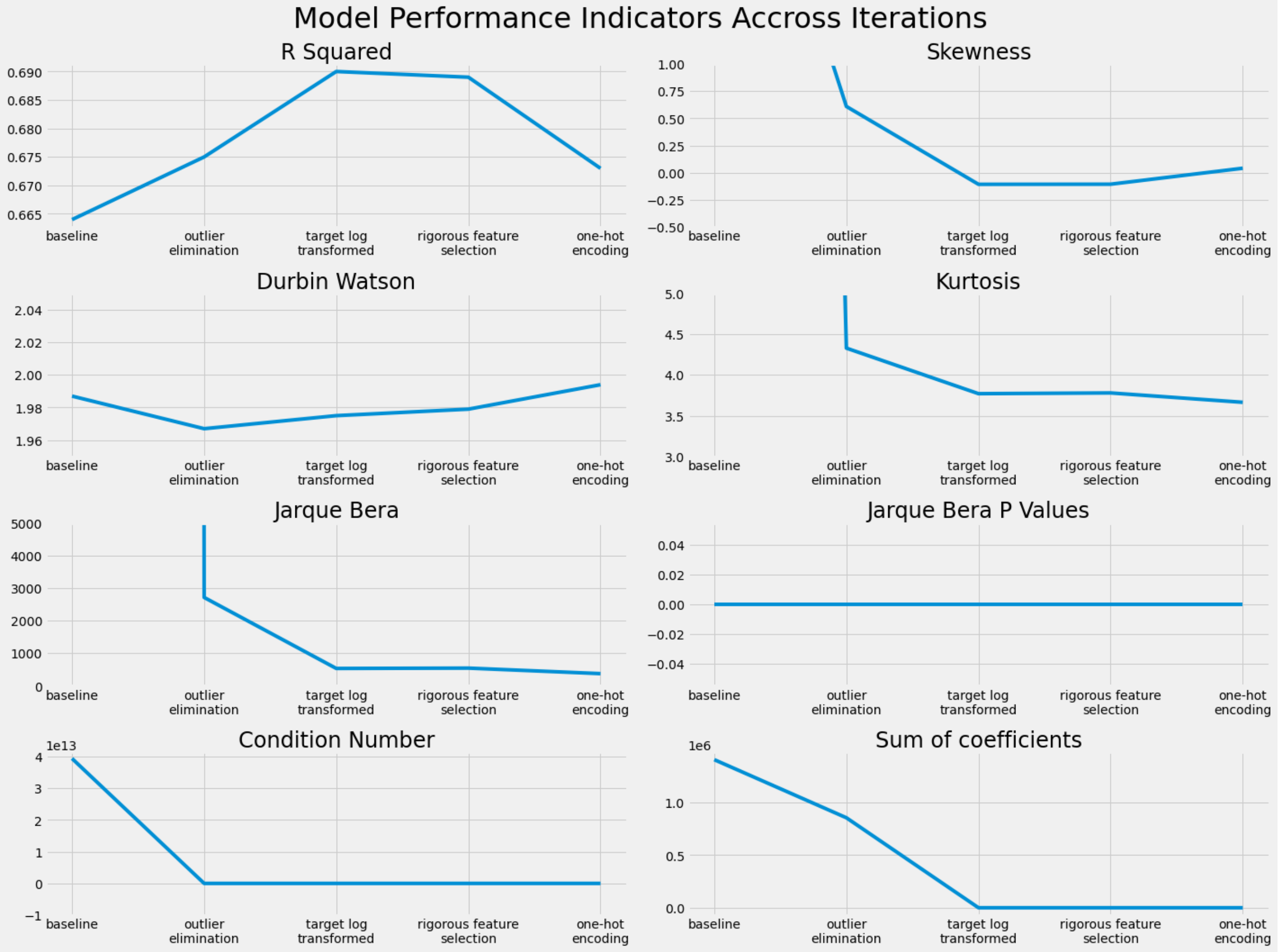
In regard to what physical aspects of a property to consider, grade is the strongest predictor of price. Grade is on a scale of 1-11 and indicates how well the property satsifies the building codes as well as the quality and level of luxury of the property; for every additional point on the 1-11 scale the price of a property increases by 0.15% meaning that rufurbishing a home that is not even up to code into a luxury home could increase the price by more than an entire percentage point.

Condition is another strong predictor and is similar to grade, it is a more qualitative version and does not refer to its satisfaction of building codes, it is on a scale of 1-5 and according to the model for every additional point here the price of the property will increase by 0.1%.

View is worth considering as well, the quality of the view from a property is also on a scale of 1-5 and for every additional point the price of a property increases by 0.07%. It is also worth noting that the price of a property increases by 0.09% for every additional bathroom in the house.

When is the best time to by is somewhat unclear. The model suggests that prices decrease slightly month-to-month if you begin in January (i.e. sale_month, -0.0025). However prices tend to increase the more recent the sale was (i.e. sale_date, 0.0002). Considering that the dataset covers a range of dates from May 02, 2104 to May 24, 2015, my intuition is that spring/early summer is the when prices are highest, but I'm not confident that I can empirically support that claim with this model.

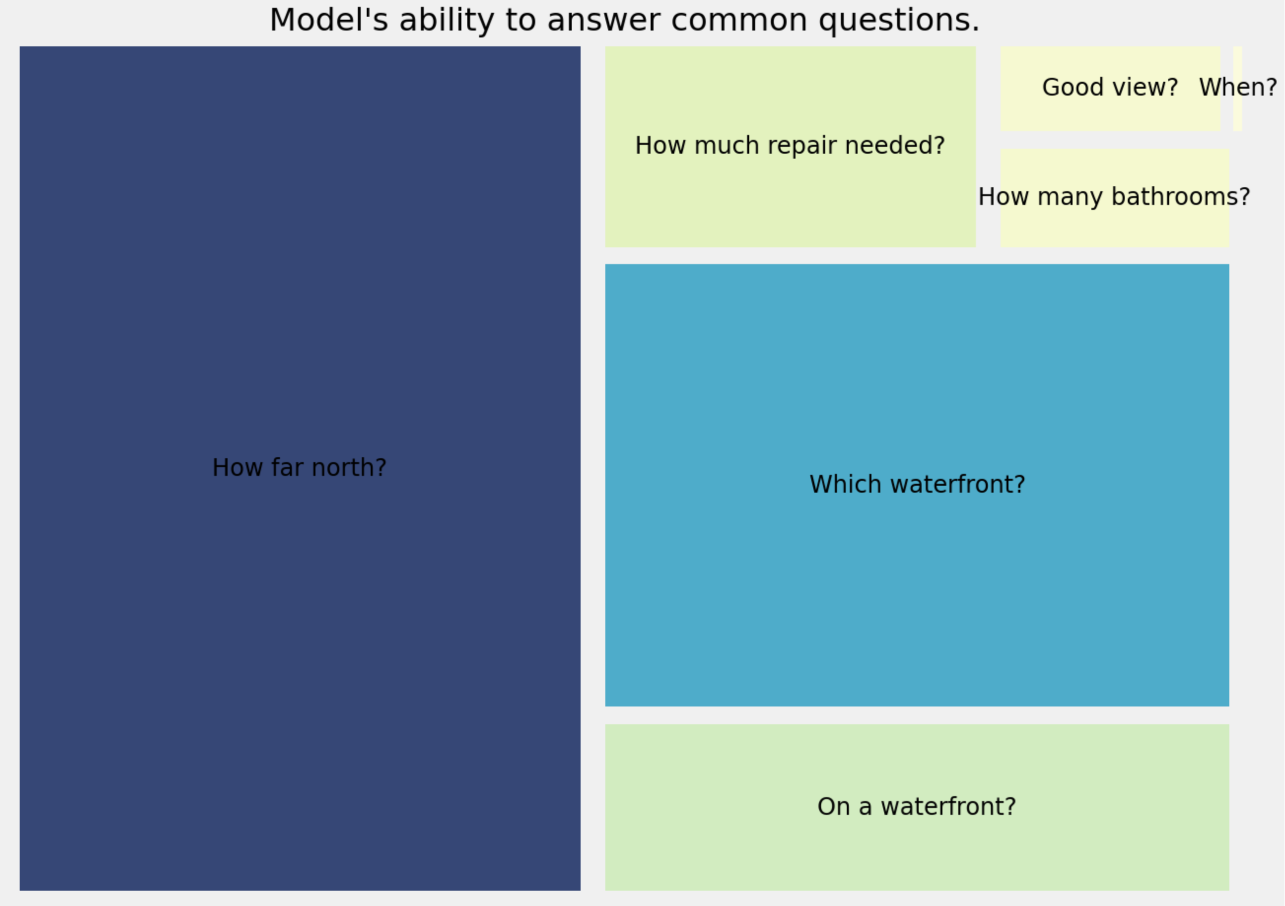
The figure below illustrates the change in select model performance metrics across iterations.



Conclusions

For the purposes of the client, Fliphouse, LLC. I recommend looking at properties on Lake Union, or north of downtown Seattle and on a waterfront, with a view of said waterfront. Furthermore I recommend finding properties that satisfty the aformentioned geogrpahic stipulations that are in need of repair and potentially do not satisfy city building codes. The best course of action according to this analysis would be to bring the building and property up to code and beyond and perhaps add one or two bathrooms to the structure.

The figure below is meant to hint at how much each answer to the various questions contributes to the price of a realestate listing in Kings County, Washington.



Next Steps

It is worth one-hot encoding someof the other categorical features such as view and grade. Also, investigating eliminating outliers based on category (e.g. there are only 27 data points with a grade of 4, and only 1 with a grade of 3 or less.). Some of the leptokurtosis may be address by continuing to eliminate outliers using this method.

This dataset covers a timespan of roughly one year, it would certainly strengthen the model to include analogous data from other years both before and after.

For More Information

See the full analysis in the Jupyter Notebook or review this presentation. For additional info, contact Zeth Abney at zethusabney@gmail.com

Repository Structure

```
├── data
├── LEARNCO
├── media
├── EDA.ipynb
├── model_dictionary.md
├── README.md
└── student.ipynb
```