San Diego Real Estate

By Susanna Han

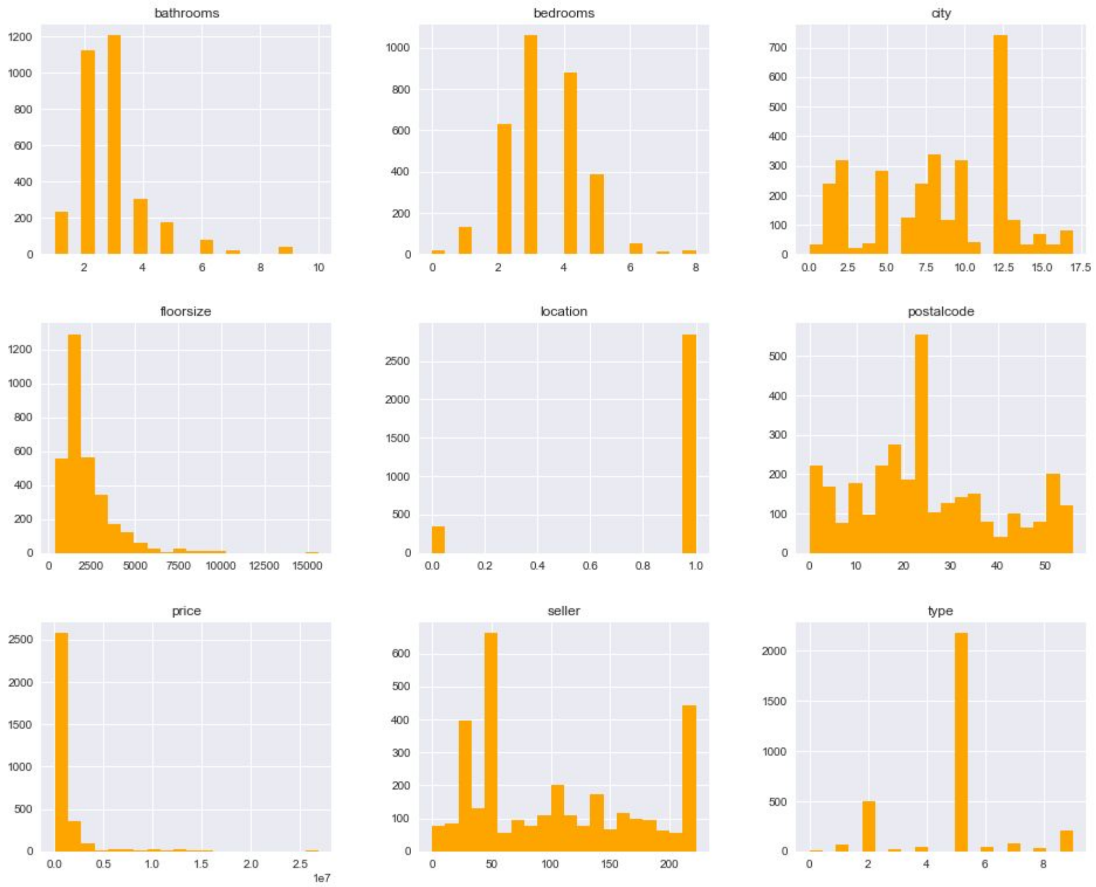
Purpose of Analysis:

The purpose of this analysis is to obtain information on different properties listed on Zillow in San Diego, CA. To compare the price of the properties to the other independent variables.

There are a total of **10 different variables** including the target variable, **price**, providing information on whether or not each variable has a correlation to the price.

3,200 properties are used in this data set:

- price of property
- type of property
- postalcode
- city
- url to listing
- floorsize in sqft
- number of bedrooms
- number of bathrooms
- agent/company (selling the property)
- location: suburb or city
- price point



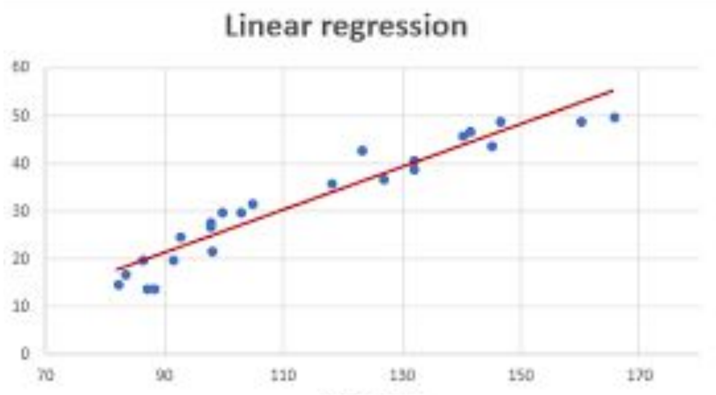
Check Distribution

Allows us to see where the majority of our data is gathered using histogram for each variable affecting the price of the home.

Two types of models use:

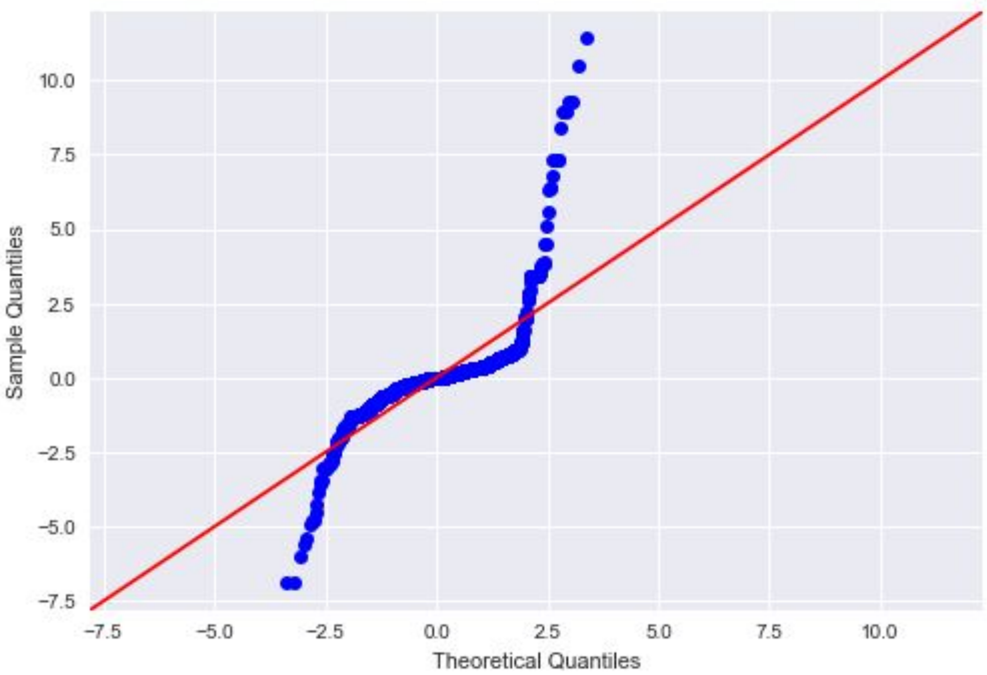
Linear Regression and **Random Forest**

LINEAR REGRESSION



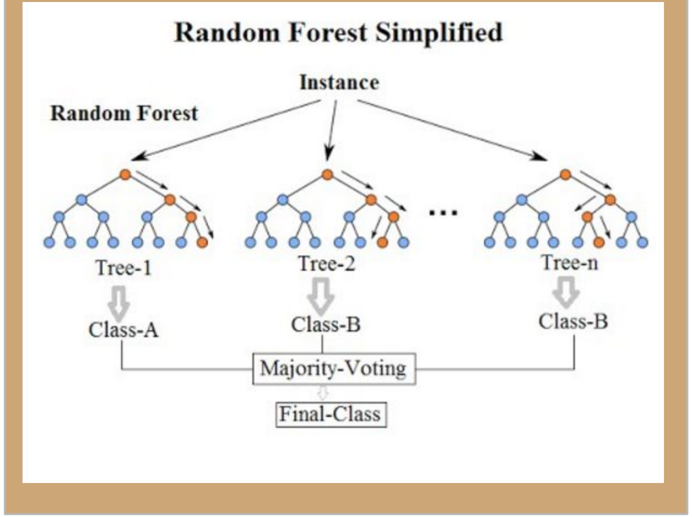
Ordinary Least Squares is a method that estimates the relationship between every variable with the independent variable. It minimizes the sum of the squares between the observed and predicted values often shown as a linear line.

LINEAR REGRESSION RESULTS



The r squared value in our linear regression model increased as we took away the location and seller variables, resulting in a r squared value of 0.799. Which can be interpreted as a **79% correlation.**

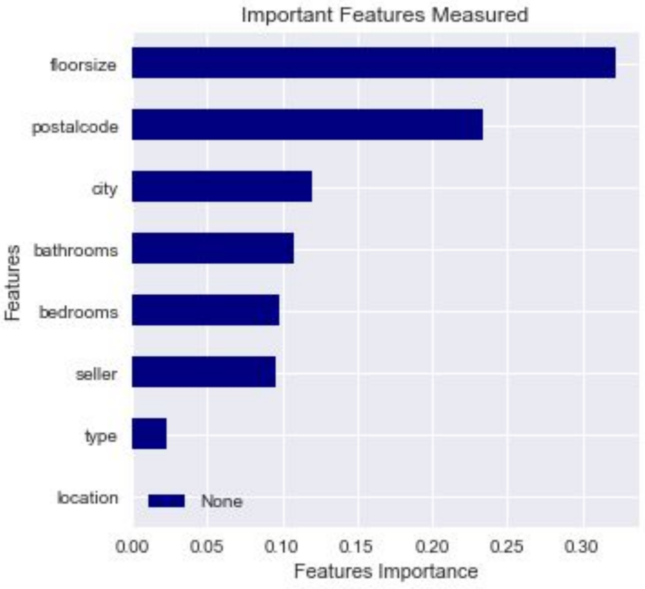
RANDOM FOREST



Random Forest is an algorithm that consist of many decision trees using random subsets of features that averages the outcome to make predictions.

RandomizedSearchCV is an algorithm that takes the input of parameters and attempts every combination and outputs the best hyper parameters for the model.

RANDOM FOREST RESULTS

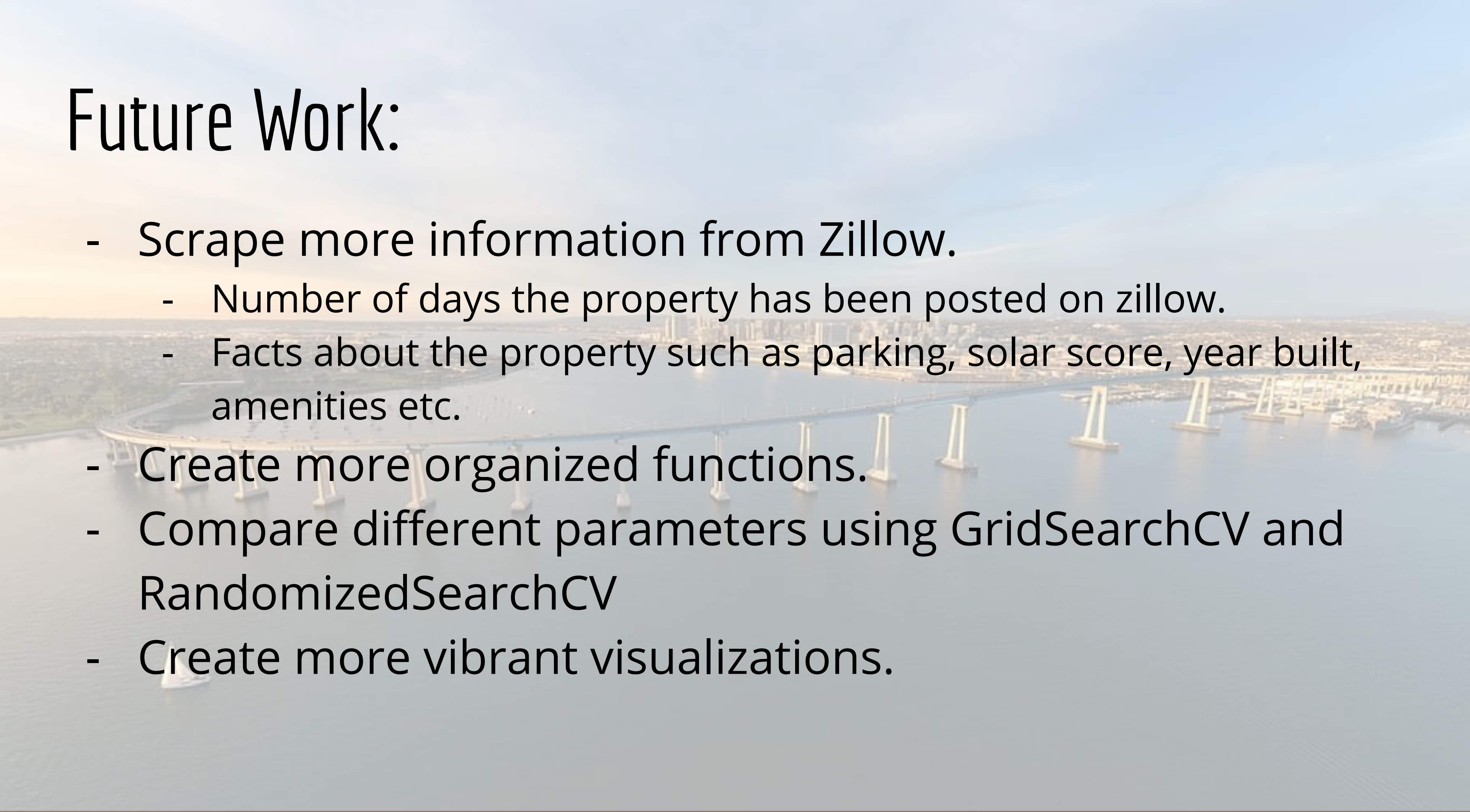


The random forest model performed well with the top three correlated features being floorsize, postalcode, and city.

The model was able to predict the price point of a property based on the features 99% accurately.

Things to keep in mind:

1. The postal code area of the property has more of an affect on the price than whether or not the property is in the city or suburb.
2. Different agents/companies have an impact on the price range of the homes they sell.
3. The floorsize of the property will have the biggest impact on the price.



Future Work:

- Scrape more information from Zillow.
    - Number of days the property has been posted on zillow.
    - Facts about the property such as parking, solar score, year built, amenities etc.
- Create more organized functions.
- Compare different parameters using GridSearchCV and RandomizedSearchCV
- Create more vibrant visualizations.

Thank you!