Analyzing San Diego's Housing Auction Data

Report by: David Gonzales
Date: 3/10/2021

1. Introduction

1.1. Background

Real estate in San Diego has historically seen large year-over-year growth in property values. Over the past five years, typical San Diego homes have seen their values increase about 40% and an amazing 14% in just the past year[1]. It's no wonder that investors flock to San Diego to invest time and money into all types of properties. San Diego is an ideal place for owning property as an investment due to the high rental rates, ideal climate, and limited new construction just to name a few reasons.

However, with high property valuations comes higher property taxes. Unfortunately, at times, an owner of a property may not be able to pay the property taxes anymore. When the owner of a property has not paid taxes for an extended amount of time, then the county tax assessor has the power to auction off the property to collect the unpaid taxes. The tax assessor does not seek to make a profit from these auctions or hope to sell at market value. The tax assessor only wishes to recoup the unpaid property taxes. This creates an opportunity for investors to bid on properties at steep bargains. The investor only needs to outbid other investors at auction and then they can take ownership of a property.

1.2. Problem

In all auction postings, there is always a “buyer beware” notice. This is because not all properties are bargains. Some properties may be financial pitfalls or time-consuming to actually take physical ownership of a property. It is wise for an investor to physically pass by a property before buying to get as much information as possible. This is key in any investment, information. The more information an investor can know about a particular asset the easier it becomes to assess the risk of investment. For upcoming auctions, how can investors quickly gather as much information as possible on a particular home? Even further, how can an investor predict the potential profit of each property on auction?

In this project, we will focus on answering these questions and more. This project is to serve as the first iteration of a tool that investors can use to optimize their time and money when shopping for auctioned properties.

2. Data acquisition and cleaning

2.1. Data Sources

This analysis will utilize the following data sources to build a model for winning bids and to calculate potential profits of a property: upcoming and historical auction data, Foursquare nearby venue data, tax assessor parcel data, and Zillow Zestimate data.

The auction datasets will be downloaded from the San Diego County Treasurer-Tax Collector website[2]. For our convenience, the website has provided Excel workbooks that list all properties that were sold at auction and which properties are coming up for auction. This dataset has key information such as assessor parcel number (APN), address, assessed value, and opening bid. Additionally, the historical data contains the winning bid. This will be our target variable to estimate in our analysis. A sample record is provided below:

APN	Address	Assessed Value	Opening Bid	Winning Bid
6333812600	01370 Green Bay St	$367,031	$183,600	$378,900

The auction data provides some key information about our properties of interest, but they are lacking details that may give more insight into the value of a property. Some of this missing information includes the number of bedrooms, number of bathrooms, livable square footage, etc. These details and more can be found in the tax assessor parcel data. This dataset was downloaded through the SANDAG/SanGIS Regional Data Warehouse Open Data Portal[3]. This is public information that can be linked to the auction data via the APN. This data resource is rich with information. It is critical to choose the most important variables from this data source to create a robust model for our winning bids. More details of this dataset will be given in section 2.3 (Data Summary) of this report.

Another data source to complement the auction data will be the nearby venue calls from Foursquare[4]. The nearby venues for each property will be quantified and used as a feature in our model building process. This information provides a summary of what the surrounding area of a property looks like through the point of view of venues (e.g. restaurants, entertainment, schools, transportation).

The last data source we will be using is the Zillow Zestimates[5] data of each property in the upcoming auction. The Zillow Zestimate is an estimated market value of a property based on Zillow's proprietary formula. This estimate is widely trusted in the real estate industry to gauge what a house may potentially sell for. We will use this data along with estimated winning bids to calculate the potential profit of each property.

2.2. Data Preparation

As to be expected, the data downloaded (or called upon) is not ready to be used for analysis as-is. This section will describe the data manipulation and wrangling that needed to be done before we perform our analysis.

Before we began to explore the data, we needed to load the data into our Python environment. The auction data was simple to read into our environment using the Pandas' *read_csv* and *read_excel* functions. Further, thanks to our *requests* library and the Foursquare API documentation, making calls to Foursquare was relatively straightforward. However, for our assessor parcel data, this took many attempts and troubleshooting to load the data. The original format of this data was a .dbf file. This file is typically found with GIS files and is used for mapping shapes onto a map. For our project, we only wanted to use the table-formatted information on the properties. Through much research and trial and error, the *simpledbf* library proved capable of reading in our .dbf file in its entirety.

The next step taken in our data process was to filter the auction data for improved property. Improved property typically means houses and condos. There were also unimproved properties which are mostly land-only properties. We ignored this subset of the data mainly due to interest. Investors typically want to do as little extra work as possible to be able to flip a property. Land property typically takes longer to sell than homes. San Diego homes can go into a pending sale in as little as a week. This is one of the primary reasons why this project only focused on single-family homes and condos.

Now that we have our scope defined, we needed to merge the appropriate data from the tax assessor parcel data, Foursquare, and Zillow. Luckily the auction data contained the APN which allowed for easy merging to the assessor parcel data via Pandas' *join* function. To utilize the Foursquare data, we could have used address, but the results were mixed. Depending on the address, the Foursquare data sometimes returned no results when results were expected. It was then decided to use the *geopy* library to add latitude and longitude coordinates to our auction data. This proved to be a wise decision and the Foursquare result returned venues as expected. Our last dataset to merge on was the Zillow data. Through many (free) API resources and many different libraries, none proved useful in getting the Zestimates. Since there were a small number of properties to predict winning bids for it was decided to manually look up each address on Zillow and record it in a workbook. The key used to then merge the Zestimates to the auction data was the APN.

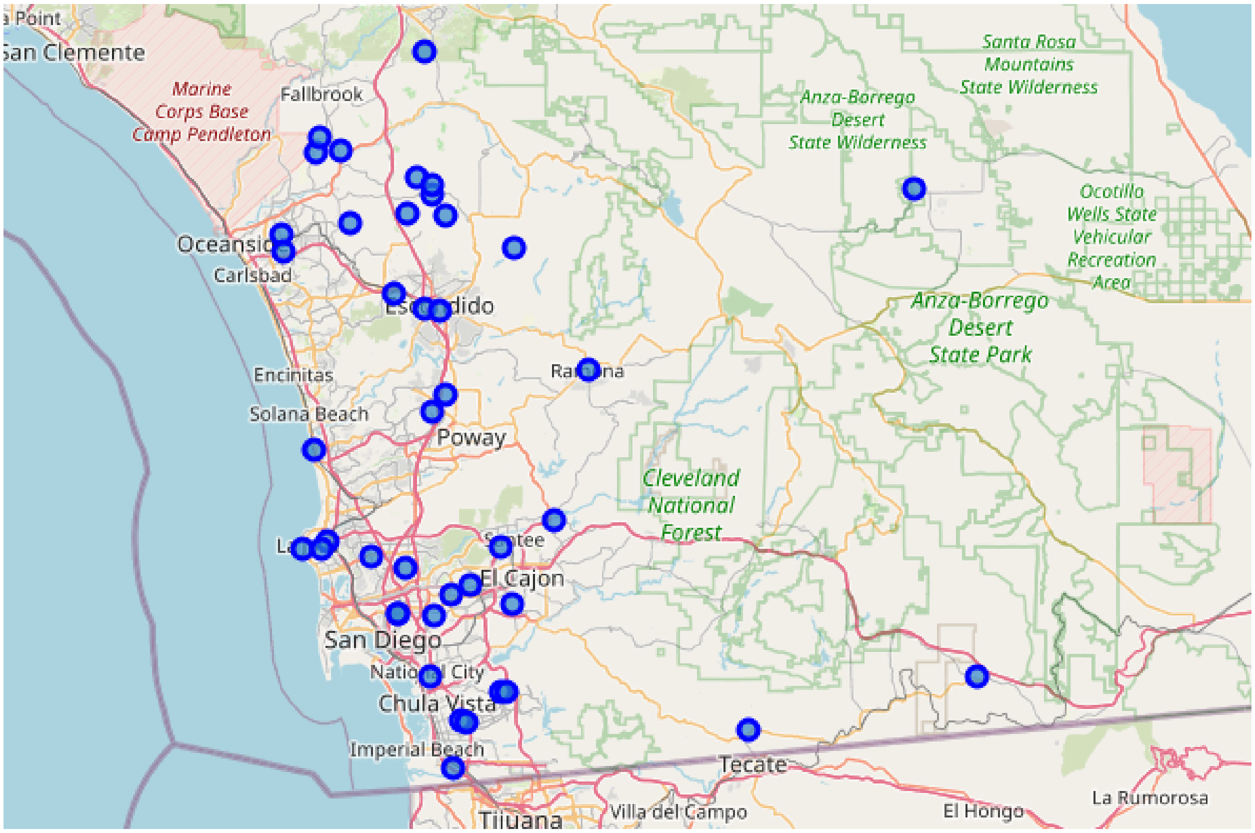
At this point, we have our data combined and are ready for data cleaning. There were not many major changes to the data other than conversion from string to numeric using different regular expressions and manipulations. For more details see the accompanying Jupyter notebook.

With a filtered, cleaned and combined dataset, we are now ready for analysis.

2.3. Data Summary

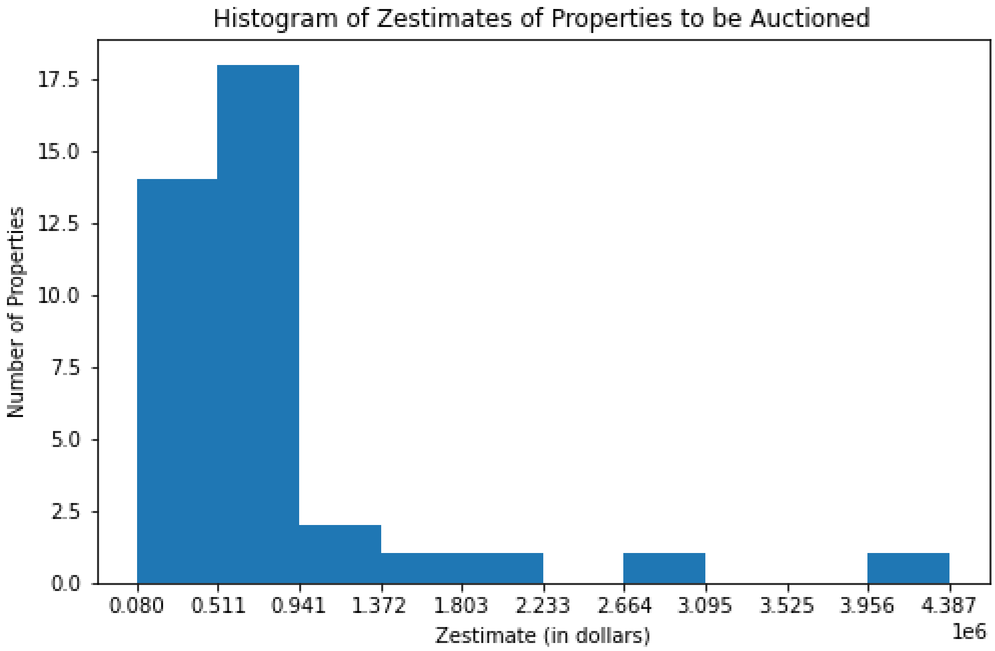
Before we begin the analysis we'll summarize the data sources to get a better understanding of our inputs.

The upcoming auction data filtered down based on our scope resulted in 38 records. These are the 38 properties that we are interested in predicting a winning bid for to further estimate potential profits. Below is a map of our potential properties that we could bid on:

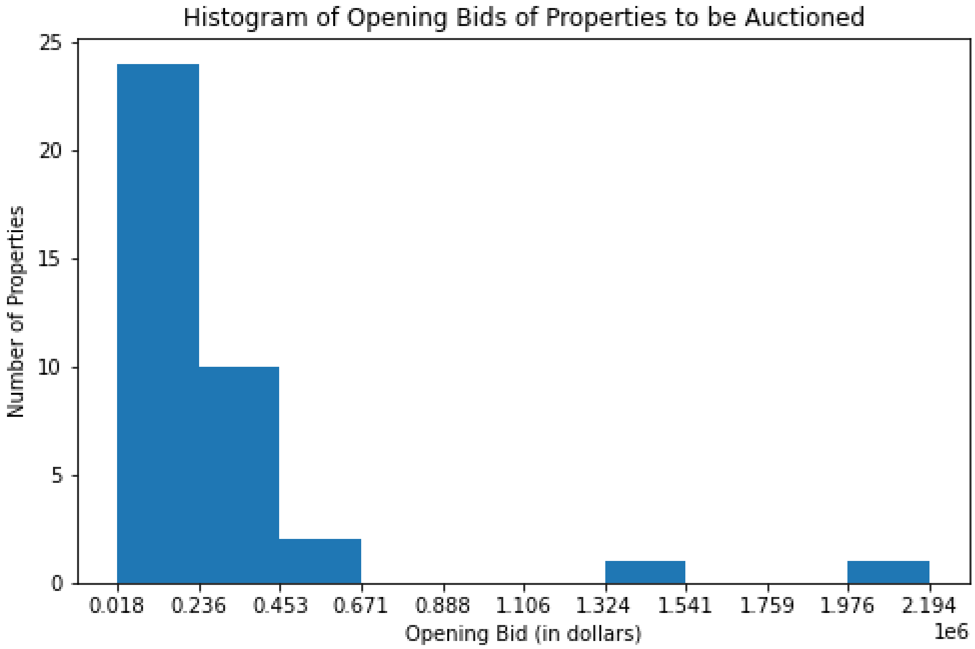


This map was created using the Folium library in Python.

As can be seen from the map, our properties span all parts of San Diego. We have properties in urban, suburban, and rural areas. Due to the diverse location of each property, we would expect a large range of Zestimate values. The following histogram displays the range of Zestimates and the number of properties in each bin.



We see that most of our properties lie between the $80,000 to $941,000 range. However, we do see several properties over $1M. Now we compare this to the distribution of opening bids and we see that the data is even more right-skewed than the Zestimate distribution. This type of skewness is good for investors because this means that the tax collector is willing to accept bids far below market value. This auction may have many potential bargains.



Now that we understand where our prospective properties are located, where the opening bids will start at, and what the potential market values might be, we are ready to model winning bids. The winning bid estimates will provide a critical piece in determining whether a property is worth taking a risk and buying.

3. Modeling and Analysis

3.1. Feature Selection

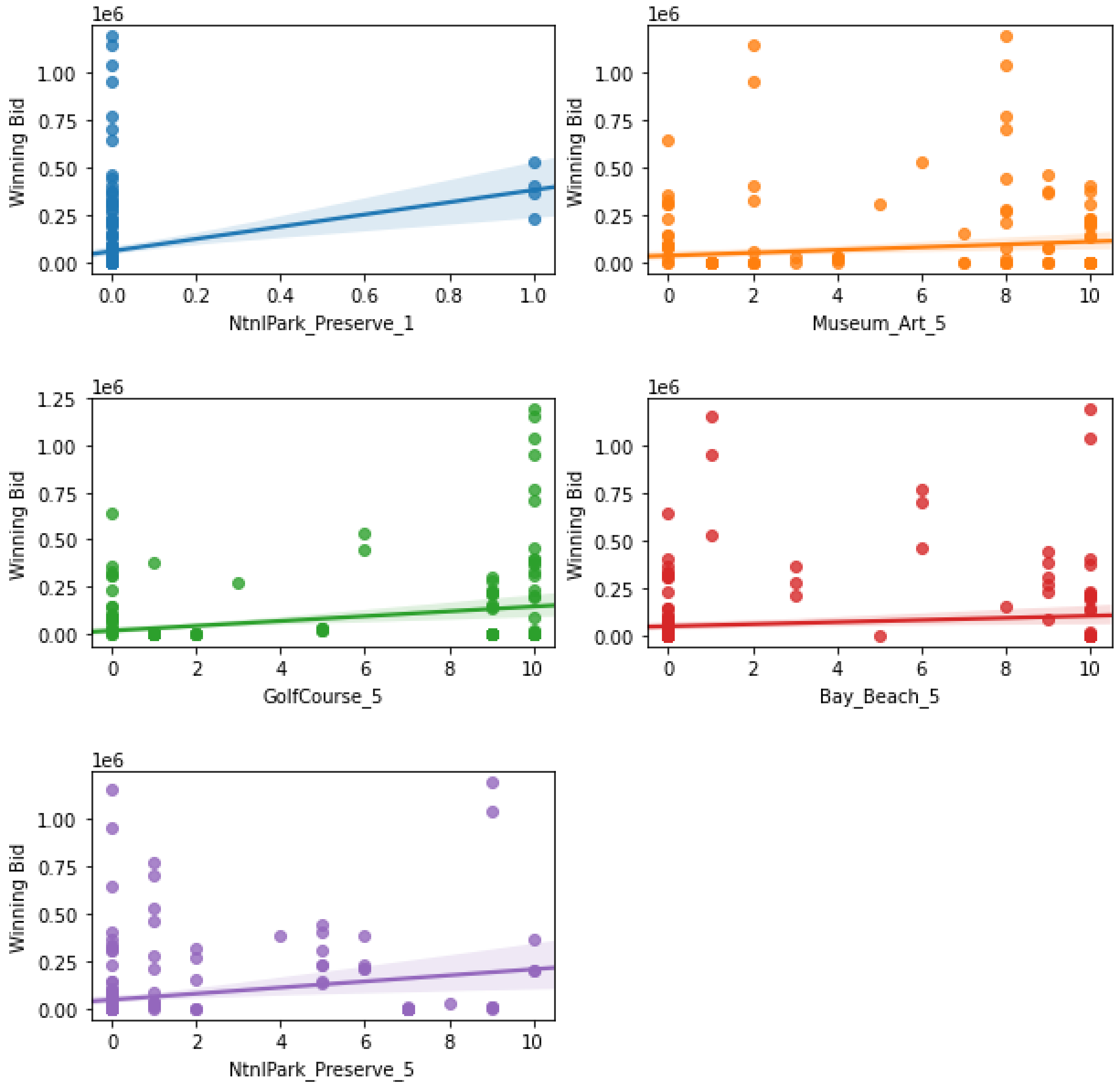
At the beginning of our analysis, we have to determine which type of model will we build to estimate winning bids. From machine learning techniques, we have a diverse set of options to use. The core of our analysis seeks to utilize many independent variables to predict a dependent variable that lies in a continuous numerical range. For this reason, this project chose to focus on creating a multiple linear regression model. Other options could have been taken, but this project chose to focus on this particular type of model.

Our first step in creating a multiple linear regression model is what is called feature selection. Even though we have more than twenty variables in our dataset that we could use for prediction, this might cause overfitting and might not generalize well to our test data. However, using too few variables in modeling might lead to underfitting. Underfitting our data would mean we are not capturing key relationships between our independent variables and winning bids. This is why we need to test each independent variable on its correlation with the dependent variable. We will use the Pearson Correlation Coefficient to determine the strength of the relationship between each of our prediction variables and our target variable (i.e. Winning Bid). The following table shows the results of running the *scipy.stats.pearsonr* function to calculate the Pearson Correlation Coefficients of our chosen set of independent variables compared to the Winning Bid variable.

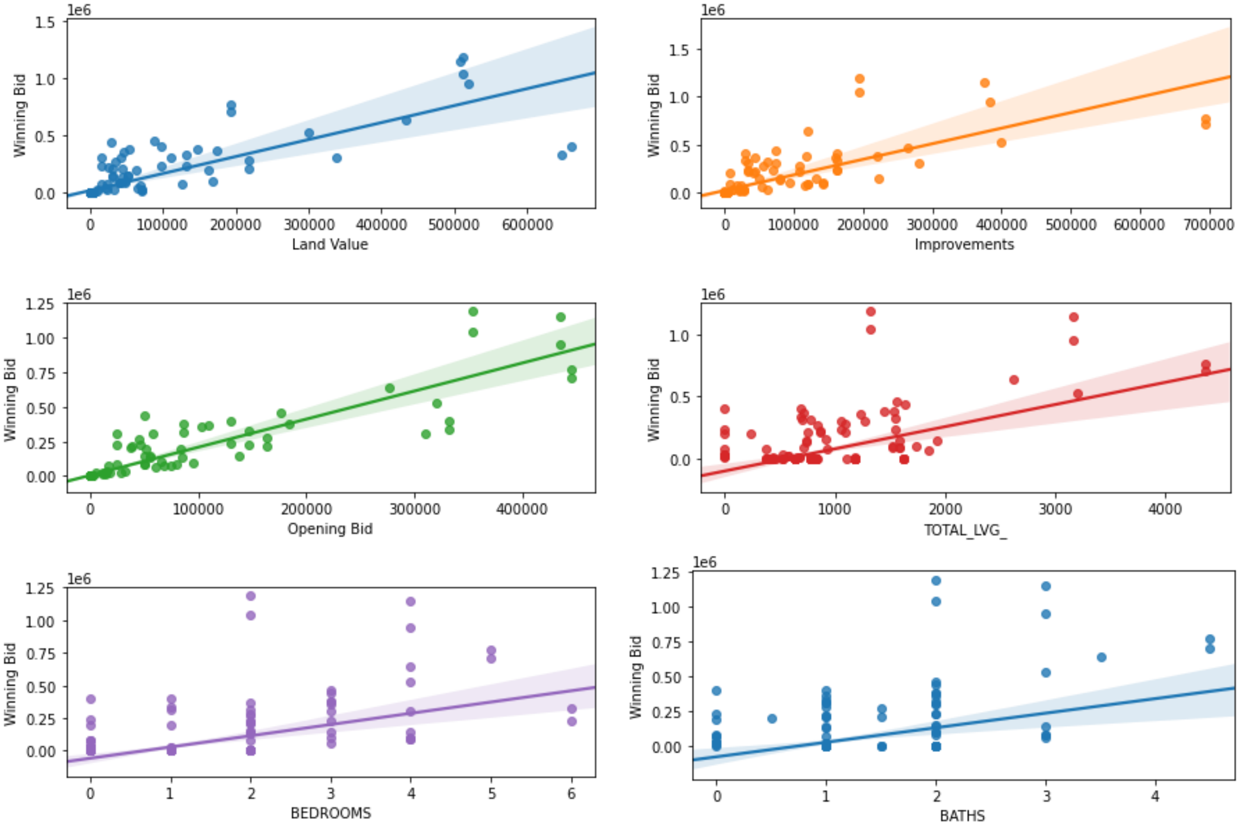
Independent Variable	Pearson Correlation Coefficient	P-value
Land Value	0.82	< 0.001
Improvements	0.78	< 0.001
Opening Bid	0.91	< 0.001
Museum_Art_1	0.08	0.18
Racecourse_1	0.05	0.47
Zoo_1	0.07	0.26
GolfCourse_1	0.09	0.14
Bay_Beach_1	0.06	0.36
NtnlPark_Preserve_1	0.22	< 0.001
ER_1	0.00	1.00
Museum_Art_5	0.17	0.01
Racecourse_5	0.09	0.15

Zoo_5	0.09	0.17
GolfCourse_5	0.29	< 0.001
Bay_Beach_5	0.14	0.03
NtnlPark_Preserve_5	0.24	< 0.001
ER_5	0.12	0.07
TOTAL_LVG_	0.55	< 0.001
BEDROOMS	0.50	< 0.001
BATHS	0.40	< 0.001

We can see that the variables associated with the green highlighted cells show moderate to strong evidence of linear correlation and are statistically significant. There were five variables (highlighted yellow) that were statistically significant but showed weak linear correlation. To further investigate this weak relationship, we can graph the relationship between the independent variable and the target variable. The following graphs show the weak linear relationships between the independent variables and Winning Bids.



Compare the previous graphs with the independent variables that had strong linear relationships and you can see the difference in the strength of relationships with the target variable.



From this analysis, it becomes apparent that the independent variables to use in our model are the following:

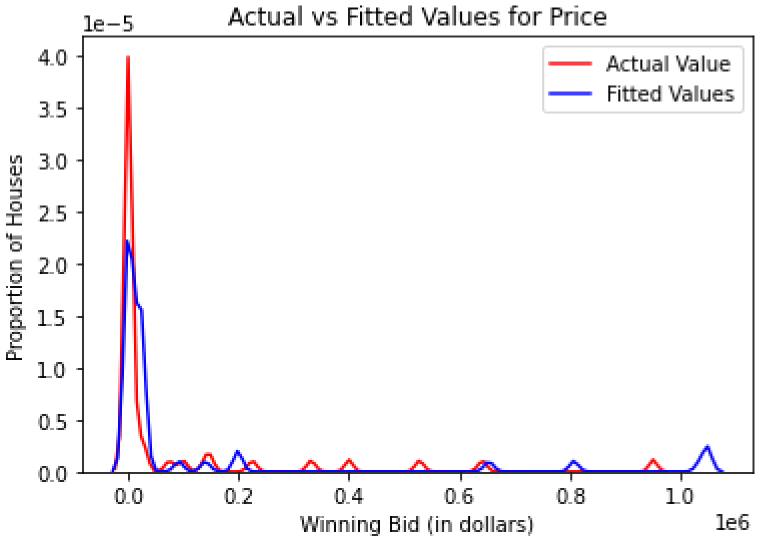
- Land Value: tax assessed value of the land that the property is on
- Improvements: tax assessed value of the structures that have been built on top of said land (i.e. house, utilities, etc.)
- Opening Bid: the price that bidding will start at
- TOTAL_LVG_: the total living area of the property (i.e. usually square footage of home or condo)
- BEDROOMS: the number of bedrooms the property contains
- BATHS: the number of bedrooms the property contains

With the independent variables selected, we are now ready to build our model that will predict the Winning Bids of our upcoming auction.

3.2. Model Creation and Evaluation

With independent variables in hand and training and testing data cleaned and configured, we are set to create our model. We will leverage the *sklearn* library to build our model. Specifically, we will use *linear_model*, *pipeline*, and *preprocessing* of *sklearn*. These packages are able to perform the necessary scaling, standardization, and linear regression to create our model. For more details on the exact parameters used, please refer to the accompanying notebook.

After the creation of our model, we need to evaluate whether it is accurate in explaining the variation of our target variable. In particular, we used the R^2 score to measure the strength of our model. The score was calculated to be 0.73. From this result, we can say that about 73% of the variation of the Winning Bid variable is explained by our model. For its application and its constraints, this score is received very well. To further investigate how our model performed we can plot the distribution of actual Winning Bids versus fitted by our model.



We can see that the fitted values are reasonably close to the actual values since there is much overlap between the two distributions. We do acknowledge that there is room for improvement, but overall this model has been determined to be a good tool to predict Winning Bids in future property auction for improved properties.

For further calculations of profit on each specific property, please refer to the Python notebook. In the notebook, each property uses this model to predict the Winning Bid and in turn predicts the potential profit given the Zestimate of the house.

4. Conclusion

From this analysis, we can see the beginnings of a model that can be used before an auction to help an investor determine which properties they should target for bidding. Depending on the amount of money an investor has, this model can help pinpoint which properties allow the most profit for a set amount of investment. Of course, this model should be used as a tool and not the absolute determining factor for investing. There are other factors that are not in any dataset we can access. One big piece of information that can help an investor is by knowing the current state of a house. Some questions to consider are: is the house in ruins, does it look recently upgraded, and are there current tenants that you will need to potentially evict? However, investors usually don't have enough time to physically go to each address and visually inspect properties. This is where the value of this model comes in. It can help an investor focus on which properties provide the most upside for investment. From there, the investor can further investigate a handful of properties instead of all available properties.

5. Limitations and Recommendations

This model is useful, but not perfect. It does have its limitations. We will go through some of these now.

- Having resources to access premium calls on Foursquare, would have allowed us to test our hypothesis of valuable properties being near "upscale" venues.
- Originally, this project was going to use housing sales data to model the price of each property. This could have served as an additional feature to estimate "Winning Bid". We did receive data, but the data was only for single-family homes and was missing all condo data. Further, this data was capped at houses that sold for $1.5M and less. We had many properties in our auction data that we would have needed to exclude due to a lack of housing sales data.
- The real estate data that we were going to use was only for the city of San Diego and not for the county. Since there were already a small number of properties to choose from in our auction list, limiting ourselves to only properties in the city of San Diego would not have allowed as robust of an analysis as we had in this project.
- This tool is not scalable for all people to have. The more people know what the forecasted winning bid might be, the more potential for influence in their own decision to bid or not. This model does not take into account what the winning bid might be if multiple people are using this tool. Knowing what another investor might bid, could bias your decisions on bidding.

Releasing the limitation on any one of these issues would most likely improve our model.

Now that we have gone through some of the limitations, we'll go through some of the recommendations that one can take to improve on this model.

- We could expand this model by introducing new features that may help better estimate "Winning Bid". One such feature could be to see how the amount of people participating in an auction influences the winning bid price. The hypothesis could be that the more people that participate in an auction, the more competitive an auction might be. This competition would drive up the winning bid price. For future iterations of this model, it would be beneficial to track the number of people participating in these auctions.
- Another improvement on this process could be to team up with an actual investor and ask if they could give resources to build a better model. With these resources, we could release the limit of regular and premium calls we make to Foursquare.

- In the next iteration of this model, we should include improved land properties. There are great investment opportunities in land-only properties. The next iteration of this analysis should highly consider including land-only properties to allow for more possibilities of finding a good bargain.

Overall, more actions can be taken to improve our model of winning bids. This project shows though that there is a systematic and quantitative way to classify what a "good" bid might be on a property. This tool shows its value to an investor by filtering down which properties have the potential for high profits and which properties could prove to be pitfalls in just a few minutes of runtime. Through multiple linear regression, we have created a tool that can predict the winning bid of a property. This tool helps investors budget for each property and saves time by knowing which properties to further research.

6. References

1. https://www.zillow.com/san-diego-county-ca/home-values/
2. https://sdttc.mytaxsale.com/
3. https://sdgis-sandag.opendata.arcgis.com/datasets/address-apn-1?geometry=-119.450%2C32.618%2C-114.224%2C33.424
4. https://developer.foursquare.com/docs/api-reference/venues/search/
5. https://www.zillow.com/howto/api/APIOverview.htm