Predicting House Prices in Ames, Iowa

(Using regression models for prediction)

Marvin Fong

Agenda

1. Problem Statement
2. Production Model
3. Production Features
4. Business Recommendations
5. Limitations of Model

Problem Statement

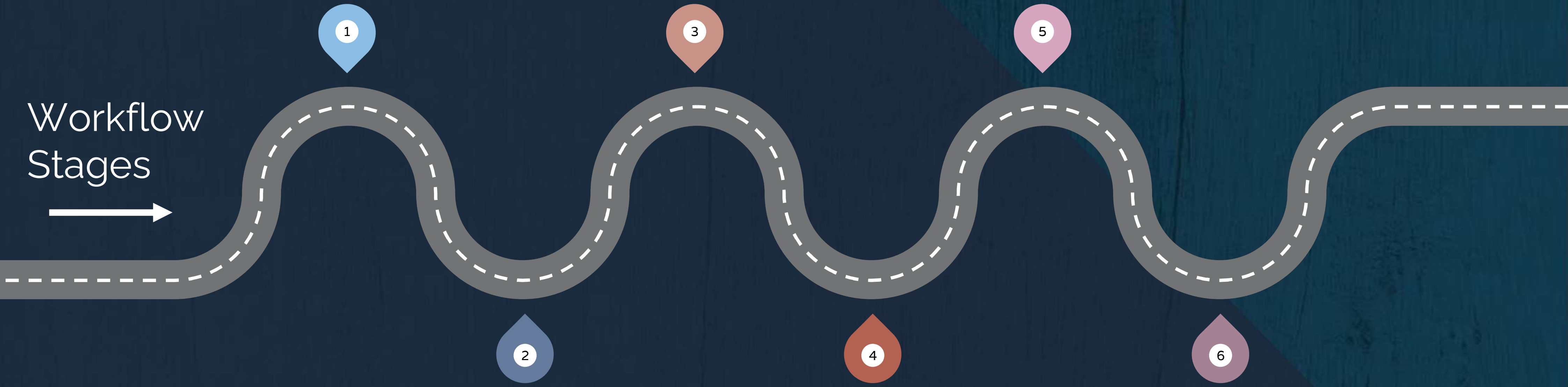
As a data team from a leading American online real estate marketplace, we are pitching to management on our model to predict housing sales prices. So that the company could identify undervalued properties and flip them for a profit.

Based on data from Ames, Iowa, the model should ideally have as low as possible Root Mean Square Error (RMSE) and answer business questions such as:

- Which features add the most value to a home
- Which neighborhood gives the highest return on interest
- Is the model applicable to other cities in U.s or countries

Production Model

Workflow Stages



1. Import data and clean
2. Exploratory Data Analysis
3. Feature Engineering and Preprocessing
4. Model with Ordinary, Ridge, Lasso & Elastic Net Regression
5. Evaluate and select best performing model based on RMSE and R^2
6. Select top 30 features for our production model to increase interpretability

Production Model

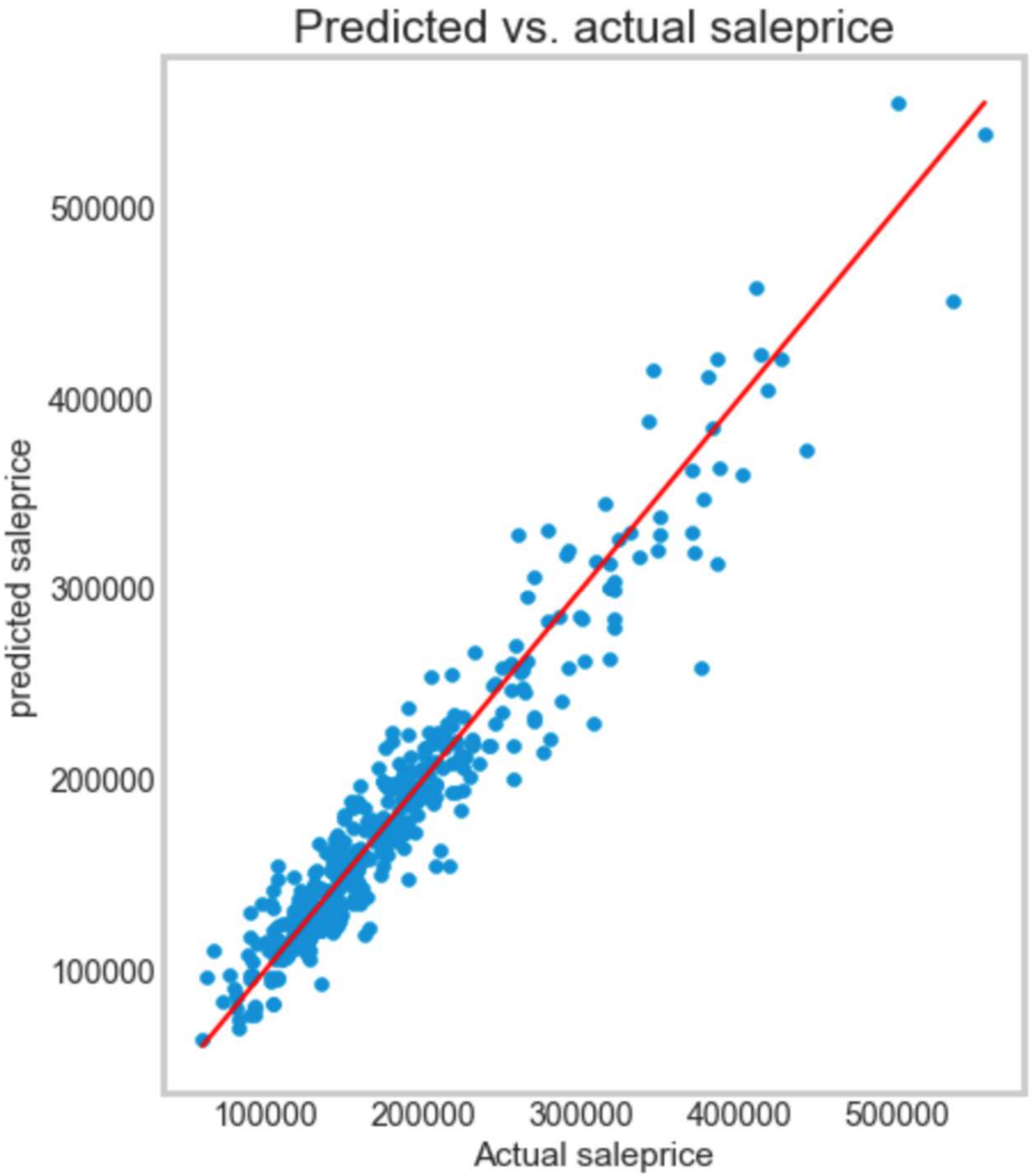
Model Selection

- After running different regression models (Ordinary, Ridge, Lasso, Elastic Net), our lasso model had the best predictive performance.
- At this point, we had trained our model using 88 independent variables (features) which made it a fairly complex model and hard to interpret.
- A key question we considered was: **how many features should our model take?**
- A model with more features may be more accurate but may be difficult to interpret without extensive domain knowledge, while a model with fewer features may be less accurate but simple to interpret.
- Ultimately, we prioritized interpretability for our lasso model by taking 30 features only. Since a **300%** decrease in total number of features resulted in a mere decrease of **0.7%** in our R^2 score.

Production Model

How good is our production model?

- We are able to account for approximately 90.4% of the variation in sale price of a property.
- Our model is also able to predict the sale price with deviation within $22,316.
- However, a point to note is that based on the scatter plot shown, our model does get less accurate at predicting higher sale price values.

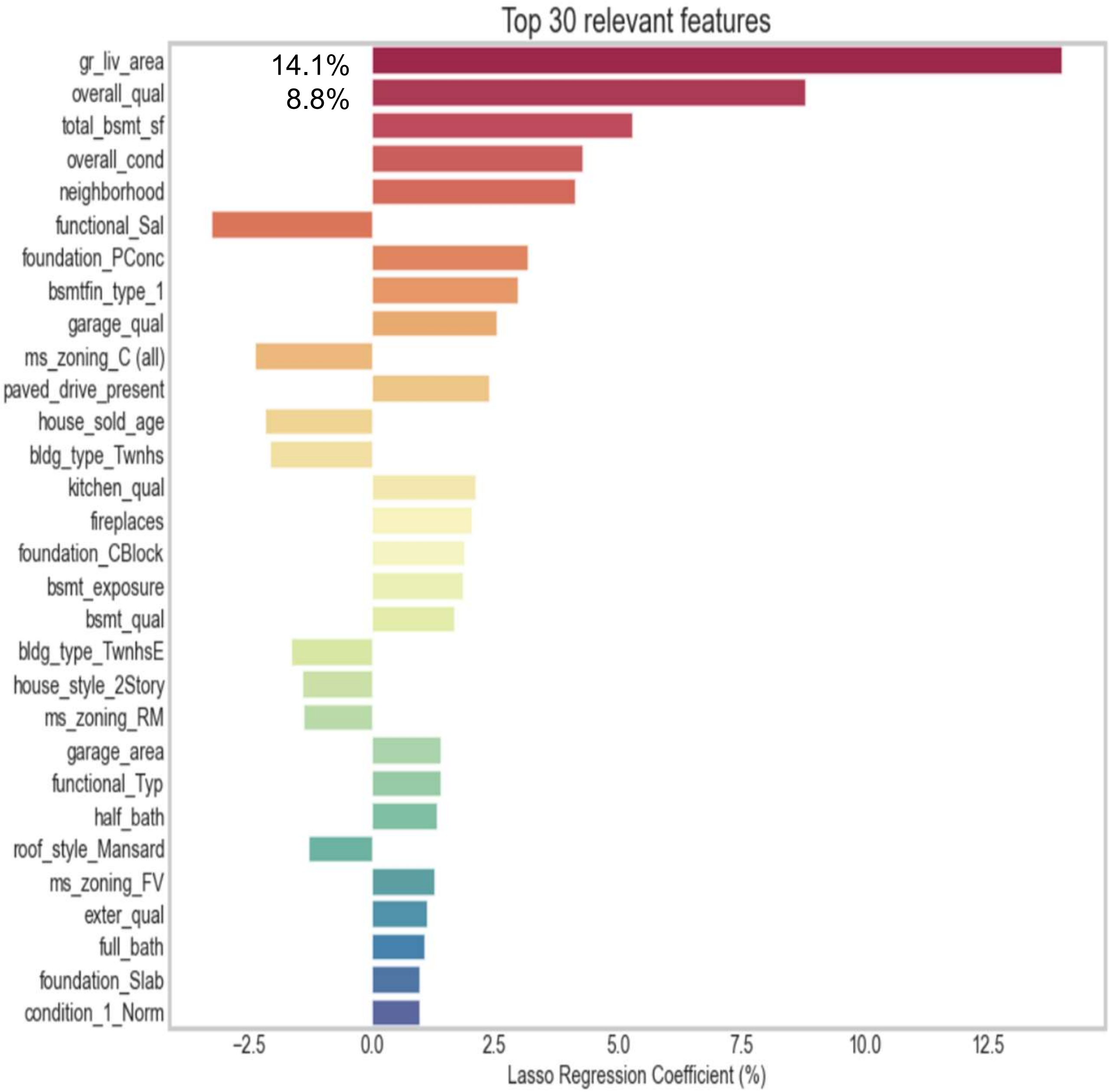


Production Features

Features are ranked based on lasso regression coefficient %, with the top feature having the highest absolute %.

Interpretation of coefficients:

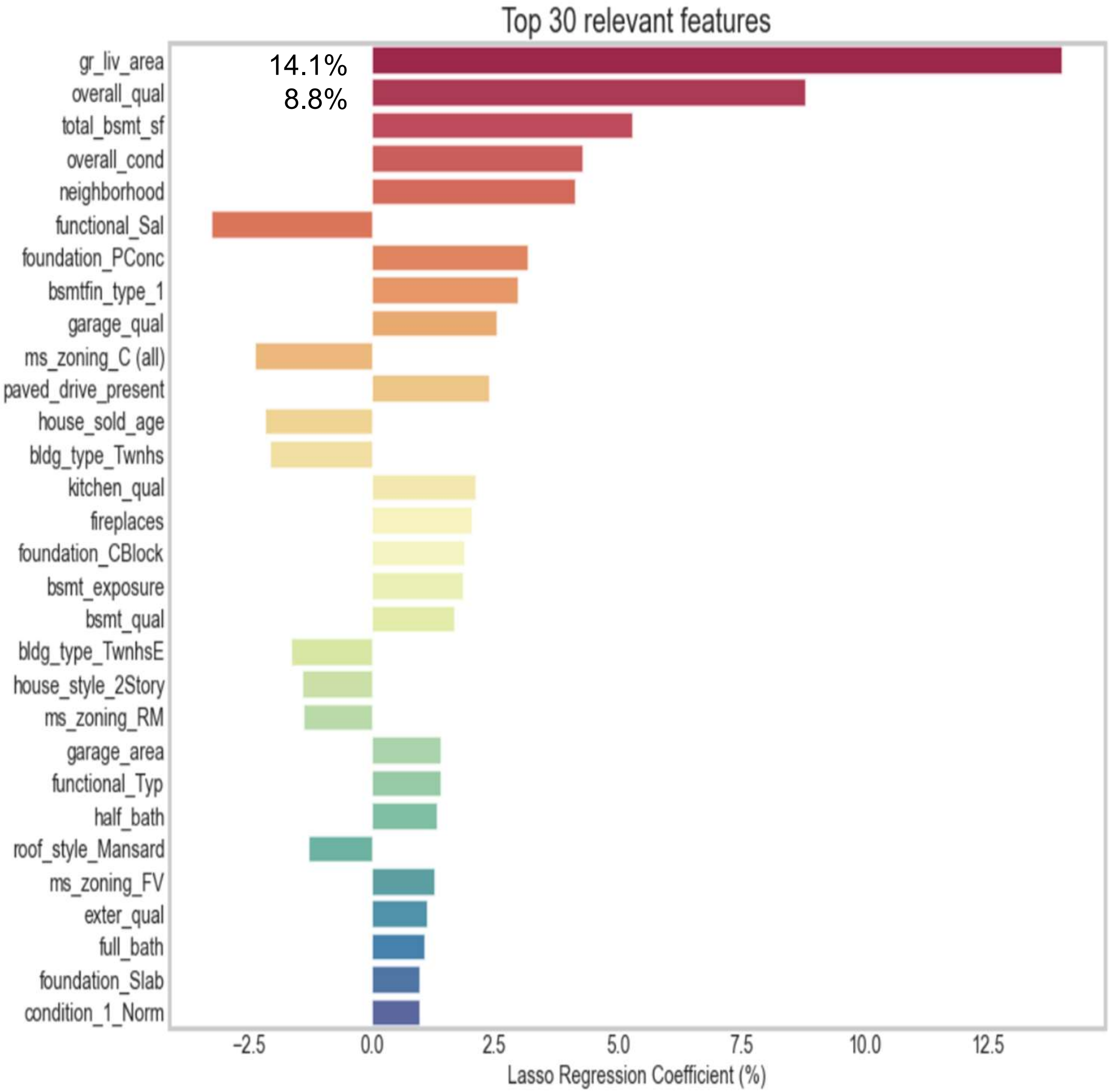
- For every 1-unit increase in above ground living area (sqft.), the house sale price will increase by 14%.
- For every 1-unit increase in overall material and finish of the house, the house sale price will increase by 9%.



Production Features

As seen on the bar chart, it is worth noting that area and quality of the house in general seems to be the biggest contributor to value of housing.

Furthermore, the neighborhood in which the house resides also play an important part in determining the sale price, seeing as it is our top 5 most relevant feature.



Production Features

We will be focusing on some of the features below when giving business recommendations. These features are deemed important and feasible in relation to house flipping.

Area features:

- Above ground living area
- Total square feet of basement area
- Presence of paved drive

Quality features:

- Overall material and finish of the house
- Overall condition of the house
- Poured concrete foundation
- Basement finished area
- Garage quality
- Number of fireplaces

Business Recommendations

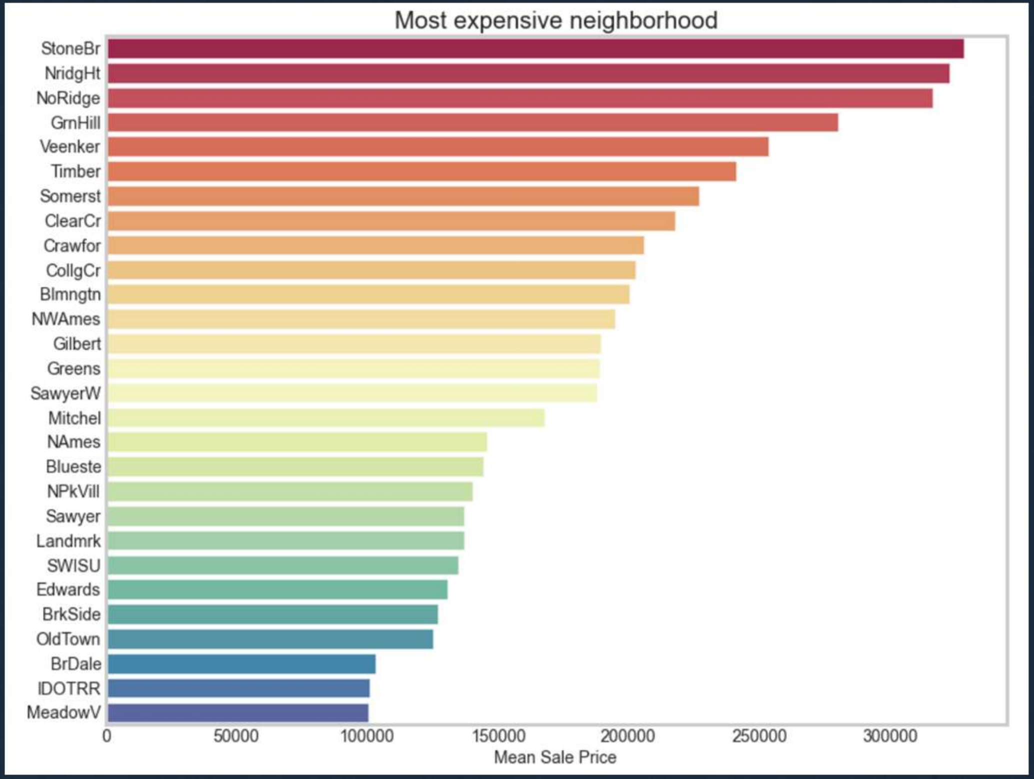
Which features add the most value to a home

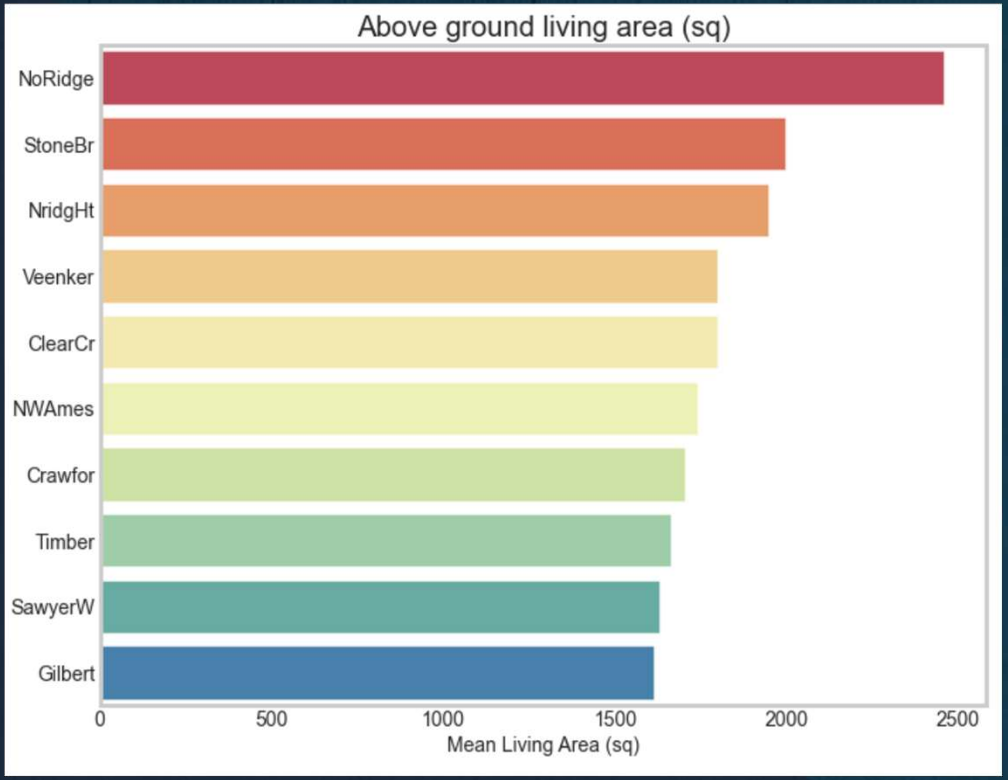
Based on our model, our company looking to flip undervalued houses and sell them for a profit could focus on the following during the renovation (i.e. flipping) process:

- Improve the overall quality of material and finish of the house
- Improve the overall condition of the house (e.g. repainting and upkeeping)
- The foundation of the house should be 'Poured Concrete'
- Improve the basement finished area (* do not let it go unfinished)
- Renovate the garage if it is in bad condition (* add a garage if there isn't one)
- Add a paved driveway
- Add a fireplace

Business Recommendations

Which neighborhood gives the highest return on interest?





Business Recommendations

Which neighborhood gives the highest return on interest?

The company should look into the 3 most expensive neighborhoods:
Stone Brook, Northridge Heights and Northridge.

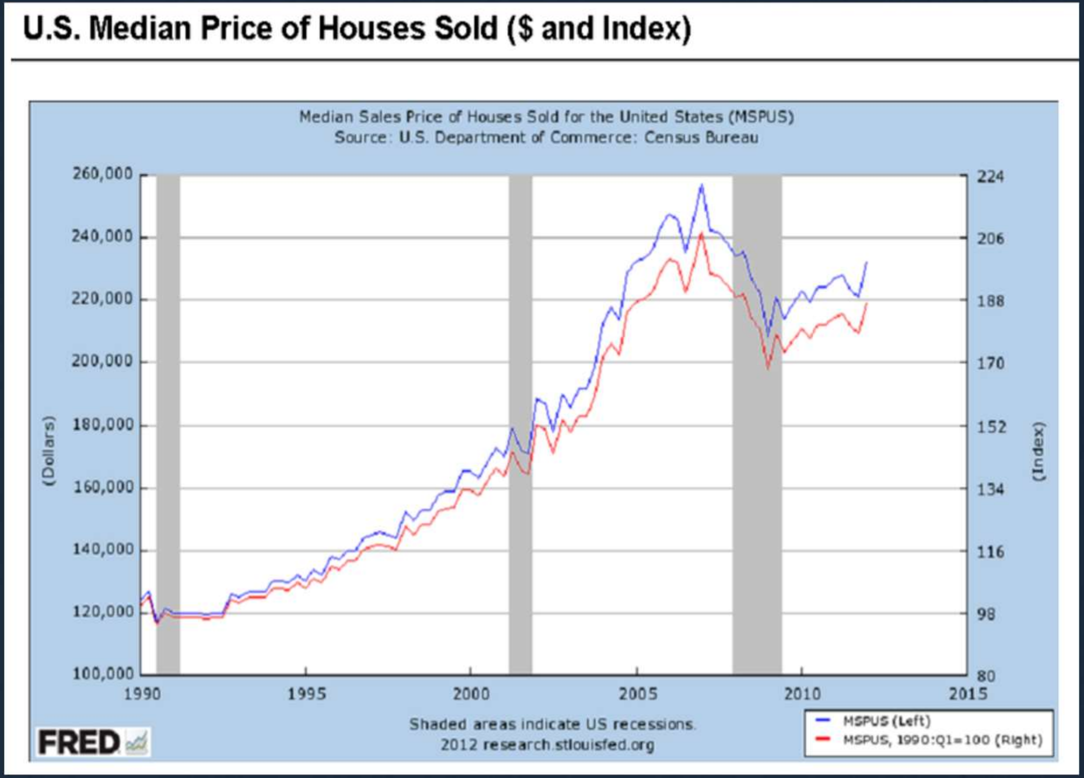
1. The company should avoid buying the property if its predicted sale price exceeds the respective mean sales price.

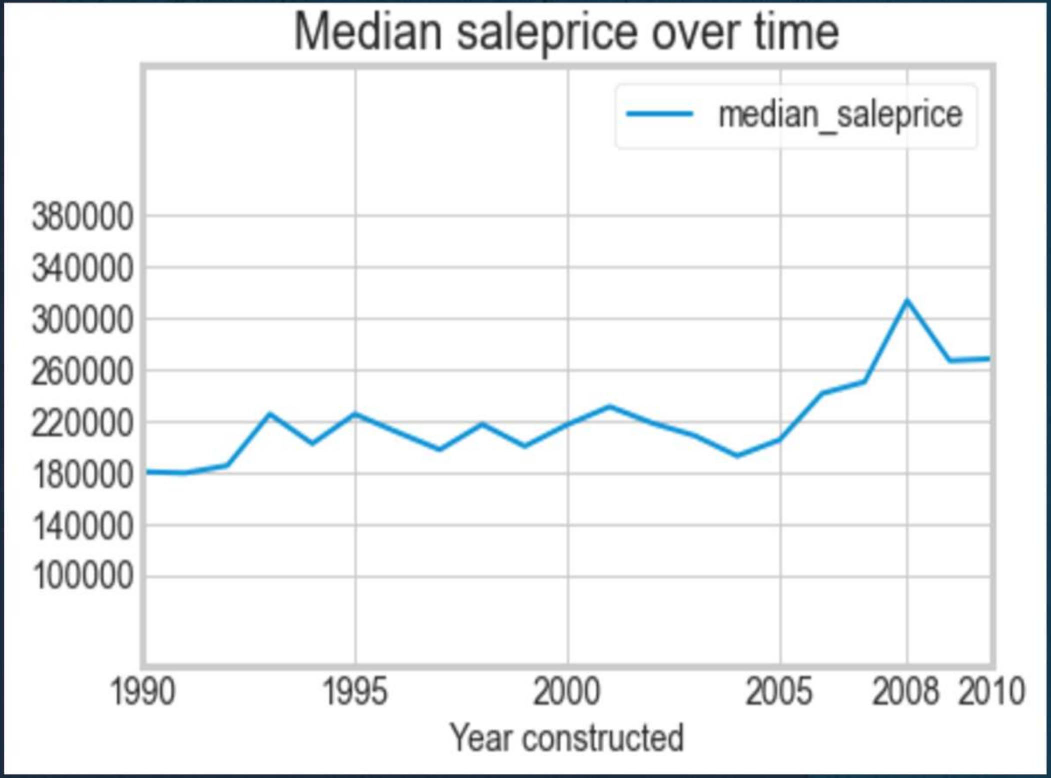
2. If predicted sales prices are approximately below $300,000, then the property is most likely undervalued. Management should look into the **area** and **quality** features as mentioned previously to determine which undervalued property would give the most ROI.

3. Focus attention into finding undervalued properties especially in Northridge as houses there are generally larger in size. Hence, management could purchase an undervalued house there, expand the house to be one of the larger houses in the neighborhood (as mentioned, 1 sqft increase in above ground living area results in a 14% increase in sale price) and sell the house for a much higher profit.

Business Recommendations

Is the model applicable to other cities in U.S or countries?





Business Recommendations

Is the model applicable to other cities in U.S?

Our team do feel that this model may be applicable to a certain extent when using on other cities in the U.S.

The city of Ames do display trends that are in line with the majority of the cities. One example shown is the trend of housing prices even during the **subprime mortgage crisis of 2008**.

However, one point to keep in mind that the last date recorded was in 2010 and there are not enough recent data to determine that Ames sale price trend is still the same as the major US cities' trend due to various factors such as:

- Housing policy changes
- Recession economy
- COVID'19.

Business Recommendations

Is the model applicable to other countries?

In order to revise the model to make it more universal, we could go about it through a few ways:

- Collect more recent data (2011-2021), preferably from other cities (e.g. Los Angeles, New York) and countries (e.g. Singapore, Germany, Japan)

- Create a model with a time period of 2000 - 2021.

- More universal features could also be added into the new data such as 'high rise buildings', 'elevators' and 'centralise carparks'.

Limitations of Model

Future improvements?

In order to revise the model to make it more universal, we could go about it through a few ways:

- Greater domain knowledge required when doing feature selection and feature engineering for the model to obtain greater accuracy.

- Model does not take into account buyers’ personal preferences. Hence, flipping houses purely based on the model’s feature recommendations might not yield the highest profits.

- Model does not take into account independent variables that relates to buyers. For example, income groups, race and age of buyer.

Any Questions?

If not... Thank you!