Capstone project 1: House Prices: Advanced Regression Techniques
Milestone report

1. Research questions

In this project I would like to solve the following research questions:

1) How does the pricing differ by houses' features?
2) Which feature impact the housing price the most, both negatively and positively?
3) Does home-style make a difference in housing price given other features of the houses the same?
4) Predicting the housing prices eventually.

2. Audience of the analysis

1) The home-buyers, home-builders and the home sales platforms will be benefited from this prediction. The reasons are as follows:

2) Home-buyers can use the information to understand whether the house will increase or decrease in value in the future.

3) Home-builders could choose to build more houses which have bigger market demand and potentially generate more profit.

4) House sales platforms and agencies can use the information to better estimate the house values for making decisions and promoting certain houses to boost their revenue.

3. Data

1) Data source:

Competition on Kaggle

2) Data description:

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

3) Data wrangling steps:

Step1. Understand the data: column types, data frame shape, correlation between the variables.

Step2. Remove outliers

Step3. Imputing missing values

Method for imputing missing values:

I have combined the training and test set for this step and I looked at categorical variables and numeric variables separately.

- Methods of imputation for categorical variables:

  -If number of missing values is less than 5, use mode to replace missing values

  -If number of missing values is more than 5, use "None" to replace missing values

- Methods of imputation for numeric variables:

  -If number of missing values is less than 25, use "mean" to replace missing values

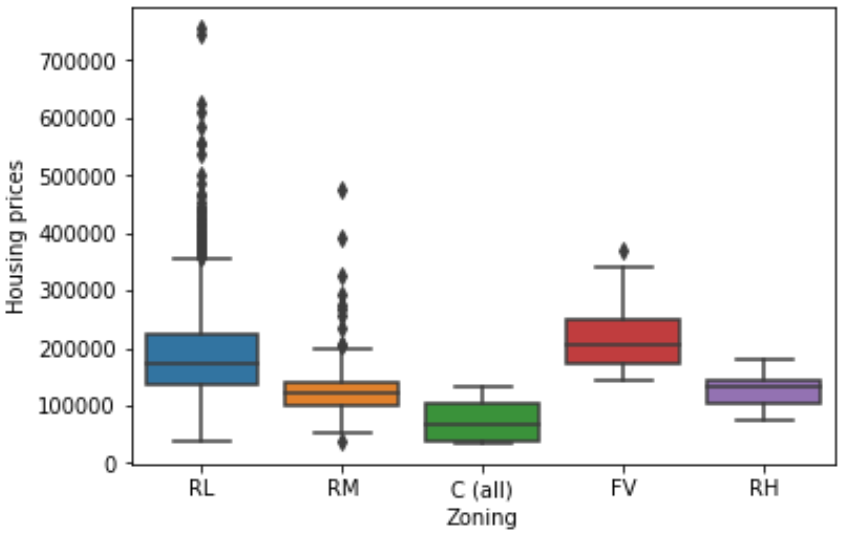
  -If number of missing values is more than 5, use "median" to replace missing values

Method for dealing with outliers:

I dealt with the outliers by mostly reading the graphs; I deleted the obvious outliers from the data.

4. Data exploration and storytelling with hypothesis testing

1) How Sales Price looks like by MSzoning? MSzoning contains "Agriculture", "Commercial", "Floating Village Residential", "Industrial", "Residential High Density", "Residential low density", "Residential low density park", "Residential Medium density".



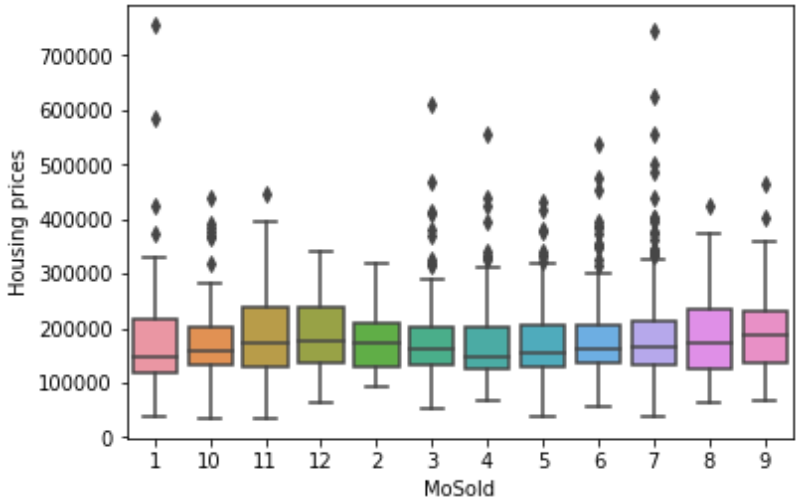
The above figure shows that the median housing price for "Floating Village Residential" is the highest. The next highest median housing price is "Residential low density", while the lowest median housing price is "Commercial". The box plot presents that the housing price for "Floating Village Residential" is more dispersed. The least dispersed housing prices is the "Residential Medium Density". Medium housing prices: FV>RL>RH>RM>C. Mean housing prices: FV>RL>RH>RM>C

2) Explore how does housing price look like by neighborhood.

	Mean Housing Price
NoRidge	335295.317073
NridgHt	317099.276316
StoneBr	310499.000000
Timber	242247.447368
Veenker	238772.727273
Somerst	225379.837209
ClearCr	212565.428571
Crawfor	210624.725490
CollgCr	197965.773333
Blmngtn	194870.882353
Gilbert	192854.506329
NWAmes	189050.068493
SawyerW	186555.796610
Mitchel	156025.750000
NAmes	145847.080000
NPkVill	142694.444444
SWISU	140199.333333
Blueste	137500.000000
Sawyer	136793.135135
OldTown	128225.300885
Edwards	127318.571429
BrkSide	124834.051724
BrDale	104493.750000
IDOTRR	100655.000000
MeadowV	98576.470588

The mean housing price by neighborhood shows that on average, the most expensive neighborhood is "NoRidge", followed by "NridgHt", "StoneBr", "Timber" and "Veenker".

3) Does "Month sold" show any pattern in housing prices? This aims to see whether there is some seasonality of the housing market.



I looked at the mean housing prices for houses sold in September and November versus the mean housing prices sold for other months. The method I used is the independent T-test. The p-value for the test is 0.04, this shows that the difference between the mean housing price for September November and the mean housing price for other months is statistically significant. So seasonality does exist.

4) Want to explore whether garage type makes a difference in housing prices

Build in and attached garages have very high mean housing price comparing to other types of garage. I did a hypothesis testing to see whether this is statistically significant. I used the independent T-test again to compare the mean housing prices with the building and attached garages and mean housing prices with other garages. The hypothesis test result shows that the difference between the mean housing prices with buildin and attached garage and the mean housing prices with other garage types is statistically significant.

5) Exploring some correlation between the features and the housing prices. I am specifically interested in the following features: exterior quality, exterior condition, kitchen quality, house square feet, total property square feet, number of total bathrooms, age of the house, overall quality and overall condition.

   Given the hypothesis test results, all the p-values of the feature shows the correlation between the feature and housing price is statistically significant. Thus, housing prices is highly correlated with (absolute r>0.5) exterior quality, exterior condition, kitchen quality, house square feet, property square feet, total number of bathrooms, and overall quality. Specially the most correlated three features are total property square feet, house square feet and overall quality.

5. Machine learning and Hyper-parameter tuning

1) Use PCA to reduce features
2) Supervised learning: Linear Regression, Elastic Net, Lasso, Ridge Regression, SVR and Gradient boosting

   The RMSE are as follows:

```
- Linear regression : mean : 0.009906, std : 0.000378
- Support vector regression : mean : 0.030622, std : 0.002065
- Gradient boosting tree : mean : 0.011265, std : 0.000391
- Lasso regression : mean : 0.033205, std : 0.001342
- Lasso regression 2 : mean : 0.014733, std : 0.000673
- Ridge regression : mean : 0.009899, std : 0.000379
- Bayesian ridge regression : mean : 0.009830, std : 0.000392
- Kernel ridge regression : mean : 0.808388, std : 0.044143
- Kernel ridge regression 2 : mean : 0.009751, std : 0.000386
- Elastic net regularization : mean : 0.033205, std : 0.001342
- Elastic net regularization 2 : mean : 0.010908, std : 0.000516
```

   By compiling the above code several times and observing the different scores each time, we can classify the models by accuracy: 1st: Kernel Ridge Regression 2 2nd: Bayesian

Ridge Regression 3rd: Ridge Regression 4th: Linear Regression 5th: Elastic Net Regularization 2 6th: Gradient Boosting Tree 7th: Lasso Regression 2

3) Hyper-parameter Tuning
   Used the GridSearchCV package
4) Used the stacked model of lasso , ridge, kenel regression, elastic net, gradient boosting, Linear regression and Bayesian regression.
   Combining the extracted features generated from stacking with original features.
   The RMSE score is 0.00932685.
   It is a good improvement from any model alone.