Predicting Home Prices

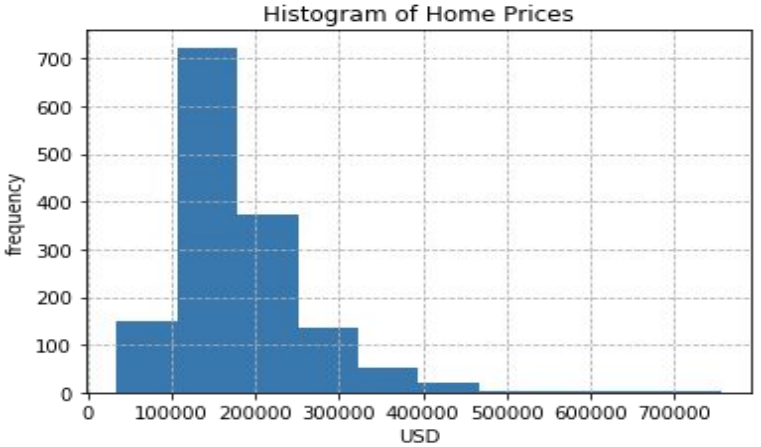
Jason Kang, Marcial Nava, Andres Zamora

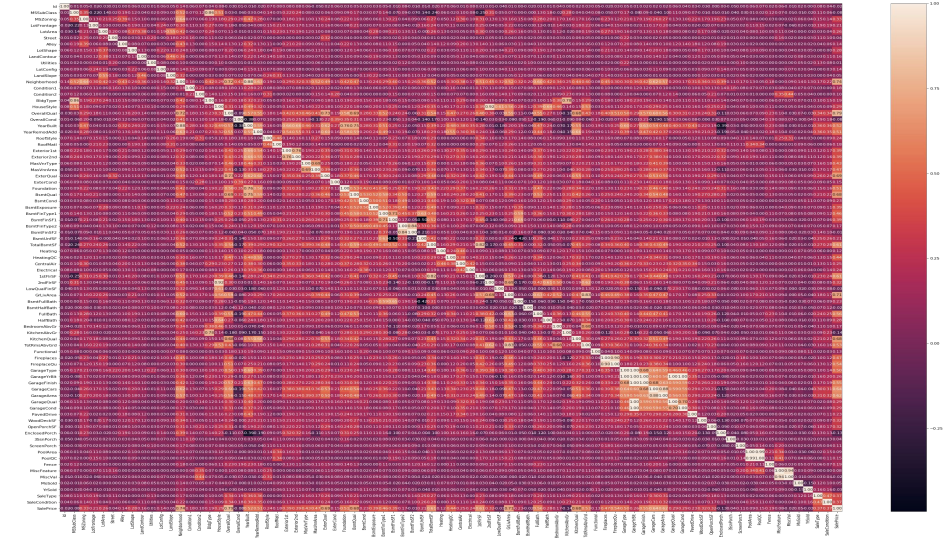
W207 Applied Machine Learning

Section 5

Understanding the data

- Numeric features with different units of measure (e.g, Sq Ft, units)
- Categorical features that need to be processed for machine learning to work
- The distribution of home prices is skewed





- 38 numerical features
- 43 categorical features
- Correlation among features (e.g. GarageCars, GarageAreas)
- Outliers

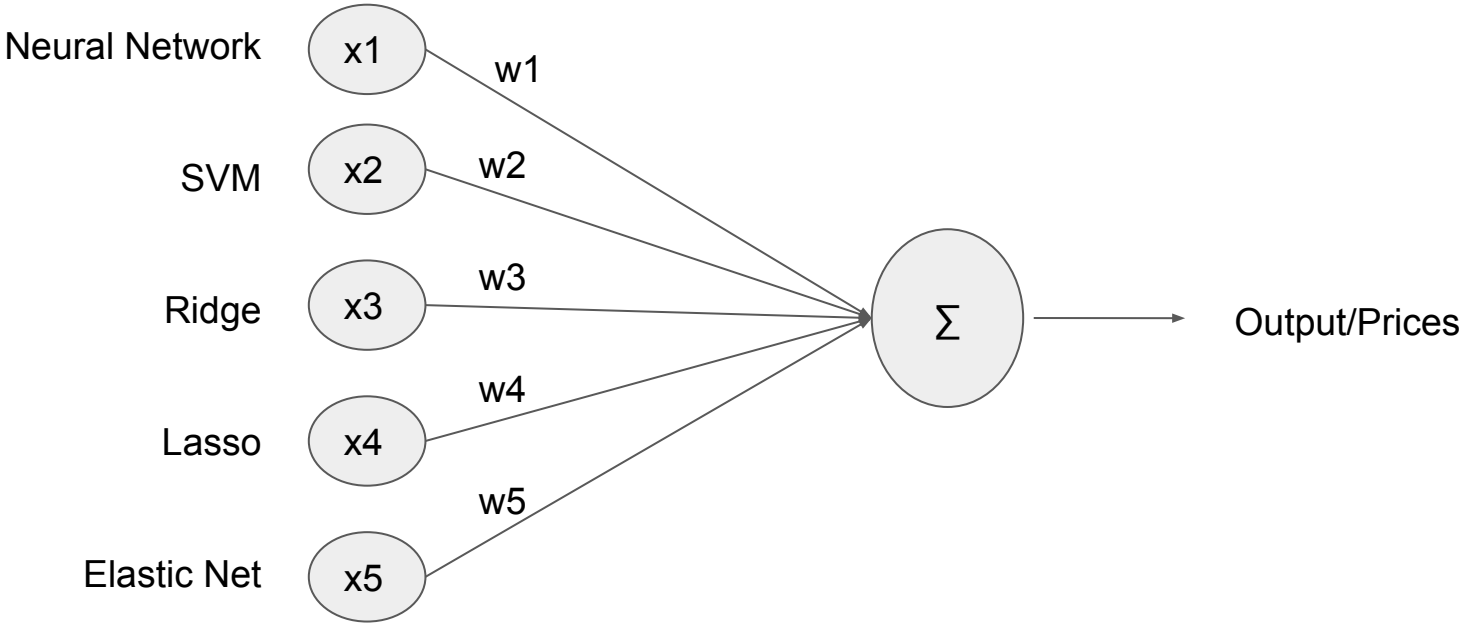
Feature engineering

- **Missing values**
  - Observations were missing for various reasons, and had to be treated accordingly
- **Dummification**
  - Categorical variables were dummified to make them compatible with machine learning
- **Normalization**
  - Both numeric and categorical figures were normalized to homogenize units of measure
- **Perfect Multicollinearity**
  - Basement Floor 1 + Basement Floor 2 = Total Basement Floor Area
- **Inconsistencies**
  - Correcting seemingly inconsistent entries (e.g. ‘PoolArea’ != 0 but ‘PoolQC’ = n.a.)
- **Irrelevancy**
  - Eliminated or ignored features that don’t help the prediction
- **Dimensionality**
  - Apply PCA to improve computational speed
- **Reclassification**
  - For example, Months sold was turned into a categorical variable
- **Feature transformation**
  - For instance, full bathroom and one bathroom were turned into one variable

Modeling

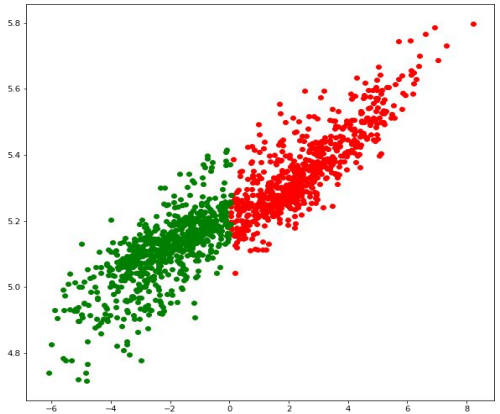
Model	Pros	Cons
Neural Networks	Flexibility, can model complex and non-linear relationships	Computationally intensive. “Black box.” Difficult to communicate. It often ends up in a local minimum.
Support Vector Machine	It can model complex non-linear relationships. It works well with limited observations	Results depending on the kernel you choose
Linear Regression		
Ridge	Effective when most variables are useful	Does not discriminate among variables
Lasso	Good at reducing useless variables	Not good if not we are not sure about the usefulness of variables
Elastic Net Regression	Good at selecting groups of highly correlated variables	Higher computational cost of cross validation. Higher risk of overfitting
Gradient Boosting Regression	Flexibility. Minimum pre-processing required	Computationally expensive. It may overemphasize outliers and cause overfitting
Decision Tree	Intuitive. Easy to find the importance of different features. Fairly robust to outliers.	Prone to overfitting, dataset has to be balanced. Can turned very complex.

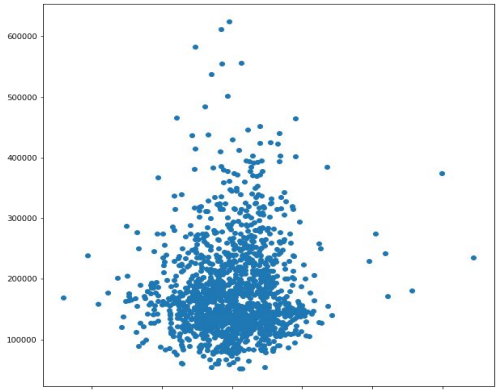
Ensembled model

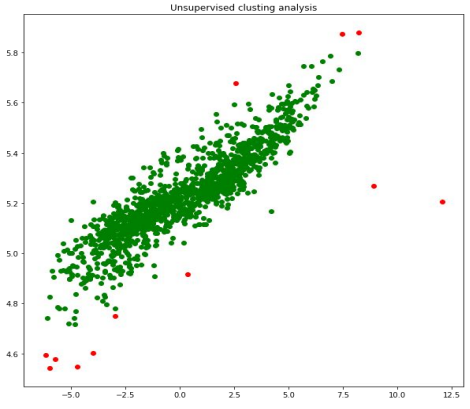


Lesson learnt - *what we tried but didn't work well*

- Over trimming of features
- Selecting top PCAs
- Using unsupervised learning to classify data and train model for each category
- Using unsupervised learning to identify outlier and delete them from training data

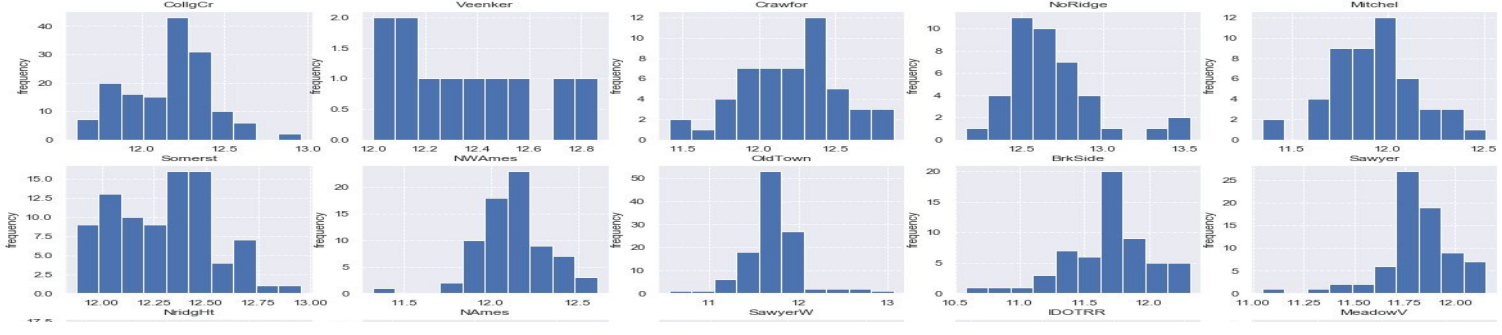






Lesson learnt - *what we tried but didn't work well*

- Recursive Neural network that throws out most significant predicted outliers from dev data each item it loops, and with each loop the dev data will become the training data for the next loop, and part of the training data will become dev data.
- Training ensemble model with top PCA component included.
- Using statistical method to reduce impact of extreme predictions.
    - Adjusted Value= prediction - P(>prediction)/P(>max) * (prediction - max)



Business case - on $37M test portfolio consisting 200 properties

- **Kaggle competition**
  - 0.1197 score put us in the first quintile
- **Investors can make money out of our predictions**
  - Even with the highest RMSE of 0.1215 (Ridge Regression), predicted prices are underestimated by 3%, which is also a fairly low rate.
  - The lower the error score, the more accurate the prediction. We can thus expect the ensemble model to do better.
  - Assuming a real estate investor that bids 10% below the predicted price:
    - Profit_Margin = (Market Price - Predicted Price * 90%) / Market Price = **12.8%**
    - Profit = **$4.7 million!**
  - Our models can save thousands of dollars to individual buyers by lowering the transaction costs associated with “asymmetric information.”

Other applications

- Helping home buyers to estimate the housing price for properties without recent transaction history.
- Helping home buyers narrow down selections.
- Helping tax authorities to determine appropriate tax rates.
- Helping real estate agents to better serve their customers.

Improvements

- Loop through different combinations of hyperparameters (computationally intensive).
- Run different combinations of models to see which gives the highest combined score for the ensemble
- By adding more ML model with each have different model but each helped, and that is why we used ensemble
- Our models can be replicated for other cities
- Information needs to be updated on a regular basis