**Predicting House Prices Using Machine Learning**

**Group Project:** Credit to Anthony Parrillo, Akshit Arora, Alan Lin and Antonio Salvador

**Objective:**  To employ machine learning techniques to build a model to predict sales prices based on dataset provided in Kaggle. The dataset included 1460 observations with 80 variables for the training dataset (similar size of dataset as test). The dataset was collected in Ames IA from 2006 to 2010 and includes most of the questions that a buyer or property assessor would ask to determine the value of a property.

**Context:** Real Estate is worth almost $30 Trillion in the US economy, and its dynamism is one of the indicators of a healthy economy. There are approximately 2 million active real estate agents in the US; they play an important role linking buyers and sellers. Traditionally, sales prices in the real estate market are estimates based on the “last” comparable sales on similar neighborhood and houses; therefore, finding a model that accurately predicts a house’s selling price can be of value to complement current empirical/limited approaches and strengthen objectivity in prediction.

**Methodology:**

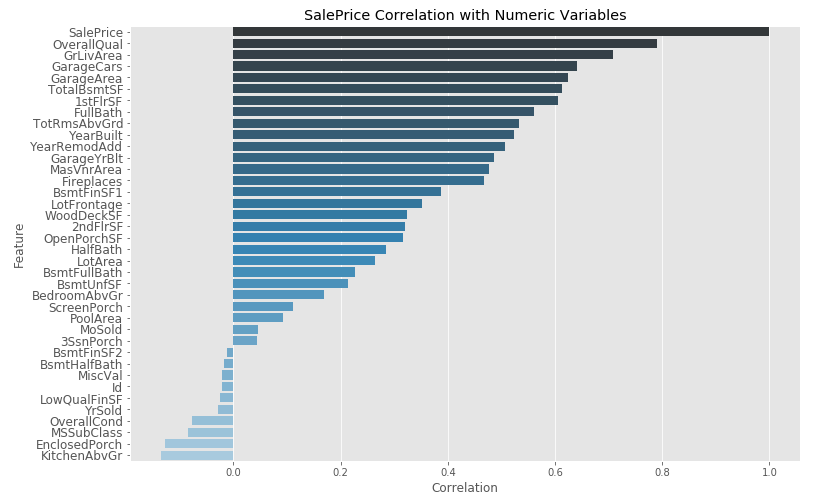
We implemented a 4 step approach:

1. **Understanding the dataset:**We analyzed basics demographics of the city of Ames , including population, economy, household income, geographic location, etc. The original dataset includes sales of almost 3000 houses in a period of 4 years. That represents approximately 10% of all the houses in Ames, IA. a population of about 60,000 with an close to 3 persons per household. Ames IA, could be categorized as a college town, in suburban Midwestern America, home to Iowa State University. This might signify lower volatility in house prices and a economy dependent in educational services (probably more stable that in an industrial town).

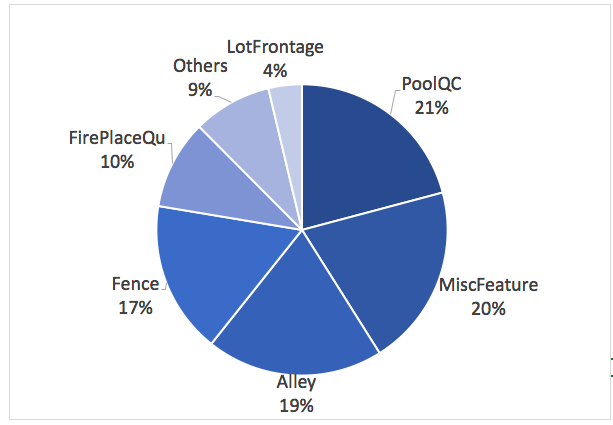
We evaluated the types of variables: continuous (mostly related to dimension of different parts of house); discrete variables (quantifying items occurring in house), nominal/categorical data (identifying different conditions) and ordinal variables (ratings to various items of the house).

We conducted EDA for numeric variables to understand their nature, such as standard deviation, skewness, missingness, etc.

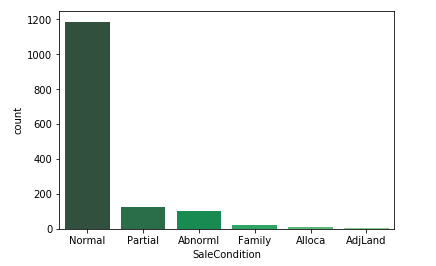
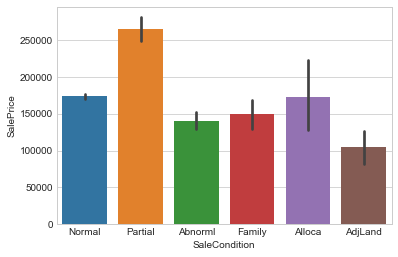
1. **Cleaning & Feature Engineering:**
2. Correlation: we evaluated the degree of correlation of numeric variables with sales price in order to have a quick view of potential feature importance and redundancies (similar variables with high degree of correlation).



1. Missingness: we identified features with a high percentage of missingness in order to determine impute methodology. In most cases missingness was related to natural absence of the factor itself (no fence, no kitchen, etc) which could be solved by imputation of zeros. In case of Lot Frontage which showed a significant percentage (20%) of missingness we assessed that median value imputation was most efficient method to use.



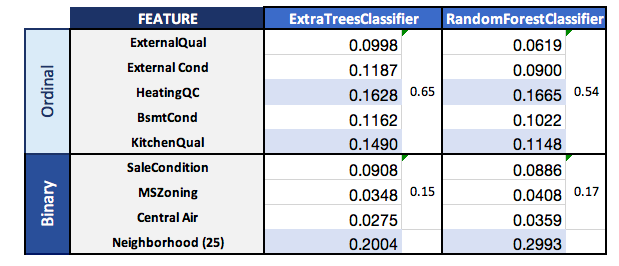
1. We combined redundant features that were also correlated, such as, 1st floor footage + 2nd floor footage + basement footage  as Total footage for house.
2. We evaluated categorical features against sales price to understand the impact of different classes and the count within each class. In some cases one class within the feature was predominant over others, for example “normal” class within “Sale Condition’ (this helped later on in “dummification” process).



1. We worked on several nominal/ordinal data to be transformed into numeric data. In the case of ordinal variables, the transformation was achieved by mirroring the “string” rating into a “numeric “ rating (most of times 1 to 5 / 1 for poor to 5 for excellent). In the case of categorical data,the approach was “dummification.”

It is important to mention that a high priority was placed on the Neighborhood feature, as we found it to be an important factor in sales price in any real estate market (location, location, location). We tried to transform this categorical data into groups of neighborhoods by household income, but the information found in this regard was not accurate or correlated to the sales prices in the dataset (i.e low household income was found in most expensive sales prices neighborhood). The specific neighborhood could be something to look into in future complements to original dataset. One idea could be to geographically located each neighborhood and find its respective US census track in order to access demographic information collected at census level.

1. Once categorical data was transformed into numeric form, we performed a random forest test (using two methodologies in scikit -learn) in order to assess their relative importance for sale prices. This was used later on in modeling part giving priority to those features with higher relative weight to assess their impact in the accuracy of model prediction.

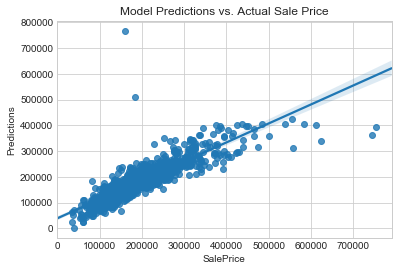


1. We performed one-dimensional cluster analysis for Total Square Footage and Neighborhood in order to explore potential clusters or outliers. Findings on potential outliers were tested in the model to assess impact in accuracy.
2. We normalized some of the features that were selected to be part of the model and showed a high degree of skewness (>0.75).
3. **Modeling**When looking at models, we determined that a linear model would be the best choice. So we compared Ridge, Lasso, and ElasticNet models.We had to tune the models setting the appropriate lambda values  for each of the models.

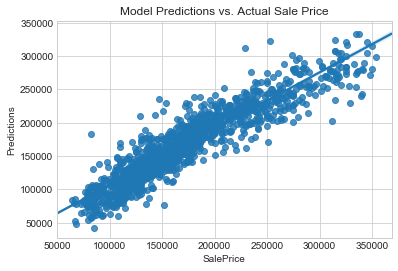
We used a 80-20 train-test split cross-validation to train our model.

Ridge and Lasso models were giving the best accuracy values, so we tested them across many different variables that were input. For our final input of 59 variables, Lasso gave us the best overall accuracy.

The following is the results of our predictions versus the actual sale prices:



1. **Refining Model**In refining the model, we were very careful to make sure that we did not "overfit" the model. As you can see in the graph above, there are some clear outliers that we needed to remove. Specifically, we removed all the sale prices below $63,000 and above $355,000, as well as a few clear outliers. Fortunately, our model did not overfit and produced results against Kaggle exactly consistent with our cross-validation accuracy.



For a real estate agent in reality, it will be hard to consider 59 factors, so we tried to reduce the number of variables used to a more manageable sox. However, the result of that reduction was 30% poorer accuracy, so we saw that at least with these linear models, it will be best to use more factors rather than less.

**5. Conclusions:**

* + The model with greater complexity delivers more accurate predictions; however, a more simplistic model could be considered for practical use (trade off in lower accuracy).
  + The model could improve if neighborhoods are standardized to US census track for more demographic information like household income.
  + The dataset could be complemented with other variables that seem to be important in house hunting/buying, such as crime, school, and transportation.
  + If would be important to validate the model with more recent data (years) as real estate seem cyclical with ups/downs.