**Kaggle House Price Prediction Notes**

**12/20/17**

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

Plan for regression:

* Thorough study of factors that influence housing prices—should spend at least one full night studying this
* EDA
* Craft features based on subject matter knowledge
  + Interaction terms, etc.
* Use automated methods to select best models since we have so many variables (try multiple methods)
* Transform data as necessary to ensure the model assumptions are not violated
* Start with OLS regression
* Then, try GLM
* Then try support vector regression
* Then explore additional techniques

Start by exploring the data available so you can refine your subject matter study to focus no what information is available

Descriptions of variables from Kaggle page describing the dataset:

- 1 ID field, 1 DV, and 79 IVs

Data fields

Here's a brief version of what you'll find in the data description file.

• SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.

• MSSubClass: The building class

Multi-level categorical

• MSZoning: The general zoning classification

Multi-level categorical

• LotFrontage: Linear feet of street connected to property

• LotArea: Lot size in square feet

• Street: Type of road access

• Alley: Type of alley access

• LotShape: General shape of property

• LandContour: Flatness of the property

• Utilities: Type of utilities available

• LotConfig: Lot configuration

• LandSlope: Slope of property

• Neighborhood: Physical locations within Ames city limits

• Condition1: Proximity to main road or railroad

• Condition2: Proximity to main road or railroad (if a second is present)

• BldgType: Type of dwelling

• HouseStyle: Style of dwelling

• OverallQual: Overall material and finish quality

• OverallCond: Overall condition rating

• YearBuilt: Original construction date

• YearRemodAdd: Remodel date

• RoofStyle: Type of roof

• RoofMatl: Roof material

• Exterior1st: Exterior covering on house

• Exterior2nd: Exterior covering on house (if more than one material)

• MasVnrType: Masonry veneer type

• MasVnrArea: Masonry veneer area in square feet

• ExterQual: Exterior material quality

• ExterCond: Present condition of the material on the exterior

• Foundation: Type of foundation

• BsmtQual: Height of the basement

• BsmtCond: General condition of the basement

• BsmtExposure: Walkout or garden level basement walls

• BsmtFinType1: Quality of basement finished area

• BsmtFinSF1: Type 1 finished square feet

• BsmtFinType2: Quality of second finished area (if present)

• BsmtFinSF2: Type 2 finished square feet

• BsmtUnfSF: Unfinished square feet of basement area

• TotalBsmtSF: Total square feet of basement area

• Heating: Type of heating

• HeatingQC: Heating quality and condition

• CentralAir: Central air conditioning

• Electrical: Electrical system

• 1stFlrSF: First Floor square feet

• 2ndFlrSF: Second floor square feet

• LowQualFinSF: Low quality finished square feet (all floors)

• GrLivArea: Above grade (ground) living area square feet

• BsmtFullBath: Basement full bathrooms

• BsmtHalfBath: Basement half bathrooms

• FullBath: Full bathrooms above grade

• HalfBath: Half baths above grade

• Bedroom: Number of bedrooms above basement level

• Kitchen: Number of kitchens

• KitchenQual: Kitchen quality

• TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

• Functional: Home functionality rating

• Fireplaces: Number of fireplaces

• FireplaceQu: Fireplace quality

• GarageType: Garage location

• GarageYrBlt: Year garage was built

• GarageFinish: Interior finish of the garage

• GarageCars: Size of garage in car capacity

• GarageArea: Size of garage in square feet

• GarageQual: Garage quality

• GarageCond: Garage condition

• PavedDrive: Paved driveway

• WoodDeckSF: Wood deck area in square feet

• OpenPorchSF: Open porch area in square feet

• EnclosedPorch: Enclosed porch area in square feet

• 3SsnPorch: Three season porch area in square feet

• ScreenPorch: Screen porch area in square feet

• PoolArea: Pool area in square feet

• PoolQC: Pool quality

• Fence: Fence quality

• MiscFeature: Miscellaneous feature not covered in other categories

• MiscVal: $Value of miscellaneous feature

• MoSold: Month Sold

• YrSold: Year Sold

• SaleType: Type of sale

• SaleCondition: Condition of sale

Do a quick look-through for all of these variables and categorize them into continuous and categorical variables

Link to original description of the dataset:

<https://ww2.amstat.org/publications/jse/v19n3/decock.pdf>

* Data is for the sale prices and associated info for individual residential properties in Ames, Iowa from 2006 to 2010
* The training set provided by Kaggle contains 1,460 observations
* Part of the research must include the area of Ames, IA
  + Use map and Google to understand the different Neighborhoods

I have gone through all of the variables

* There is a lot of detail about the nature of the house itself:
  + What it is made out of
  + What the lot is like
  + The style (including # rooms and floors, types of rooms, etc.)
  + The quality of workmanship
  + The current condition
  + Size of various parts
* How it was sold
* Year/month
* Neighborhood
* Type of property

These comprise the general categories of variables

* For getting the best combinations of the variables as they are, I could simply use automated means, like forward or backward stepwise selection
* Of course, for prediction purposes, why not just include all of the variables
* There are so many variables that this will make inference difficult in a time-efficient manner
* The goal is not inference, the goal is prediction, so stick to that
* Only use inference insofar as it serves to forward the goal of predictive performance
  + This will be different for the violent crime regression project—I’ll be more interested in inference for that project, so I’ll make sure to include an inferential modeling section
  + I could even post it in the comments section of an ASP video once I have it complete and have a user-friendly summary set up
* So, I could just simply use a saturated model as the initial basis for prediction
  + For all categorical variables, simply create dummies for each category
  + Select the category that has the least association with the response as the reference category
  + For low-variety continuous and Likert-scale variables, collapse into binary variables
  + Transform continuous variables to make sure model assumptions are met

Of course, this initial method isn’t likely to get me extremely high performance

* The real difference comes in feature engineering
* Given the large number of variables, the best feature engineering approach to me seems to be to gain some subject matter knowledge and use it to engineer better features
* I don’t even necessarily need a lot of background knowledge—I could also engineer features by looking at the different variable combinations
  + However, given the number of variables and combinations possible, it makes more sense to develop some subject matter expertise and make educated guesses as to the best engineered features to include

Plan:

* Submit the following for the first round:
  + Saturated OLS model without transformations
  + Saturated OLS model with transformations
  + Most parsimonious model with transformations (exclude all non-significant terms)
* For this first round, create a method to transform the data, which will not be used later on (although I can use parts of this method to transform the data for subsequent submissions)
* For time considerations, I could randomly select one dummy variable from each set to be the reference category
  + Why don’t you use Pearson correlation to select—select the variable with the lowest Pearson correlation value to be the reference category
  + However, is it good policy to use Pearson correlation with binary variables?
  + In regards to imputation, for continuous variables, use the most frequently-occurring category as the imputation category and for categorical variables, use the median as the imputation value
    - For imputation purposes, treat Likert-scale variables and low-variety continuous variables as categorical

Thing to implement: instead of using the variable with the lowest Pearson correlation with the response as the reference category, use the variable with the highest p-value in a t-test to compare means (of each given category against all of the other categories)