Machine Learning Engineer Nanodegree

Capstone Proposal

Dirk Honda February 9th, 2018

Domain Background

For most individuals, home ownership is the largest and most important financial decision they will make. With approximately 64% of Americans owning a home [*] (https://www.census.gov/housing/hvs/files/currenthvspress.pdf) determining the fair market value of a home is critical. Historically, the value of homes has been predicated on the sale of homes in the area with similar features (e.g., beds, baths, living square footage). This approach is susceptible to outliers and variances that could skew the value of homes. The Kaggle competition, House Prices: Advanced Regression Techniques (https://www.kaggle.com/c/house-prices-advanced-regression-techniques) provides a dataset to which feature engineering, feature selection, and machine learning techniques can be applied to gain intuition and develop a model capable of producing future fair market values with greater accuracy. With over 12 years of residential real estate software experience, most dealing with this type of data, I have always been fascinated with the pricing of real estate. It is my intention to gain more real-world experience with the analysis of datasets, specifically those in real estate, as this will be invaluable to future employers.

Zillow, Inc. a leader in real estate analytics, has demonstrated that modeling with machine learning techniques can successfully be applied to the real estate domain. Zillow's Chief Analytics Officer, Stan Humphries outlined the approach and metrics in an article written at the end of July of 2017 [*] (https://www.zdnet.com/article/zillow-machine-learning-and-data-in-real-estate/). The article states, "...2006 when we launched, we were at about 14% median absolute percent error on ... homes [prices]." Mr. Humphries is quoted as saying,"So maybe it's a decision tree, thinking about it from what you may call a 'hedonic' or housing characteristics approach, or maybe it's a support vector machine looking at prior sale prices." when speaking about the types of machine learning algorithms used to predict home prices.

Problem Statement

As stated in the domain background, house prices are generally modeled on recently sold comparable homes (sales comps). This approach is based on the opinions and experience of brokers and realtors. As such it is susceptible to fluctuations, outliers, and human bias. By applying machine learning algorithms, a more robust and accurate regression model can be developed which can be evaluated by making sale price predictions on data prior to sale and then comparing them to their respective future purchase price.

Datasets and Inputs

The dataset includes four text files which can be downloaded from Kaggle (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data)

File Descriptions:

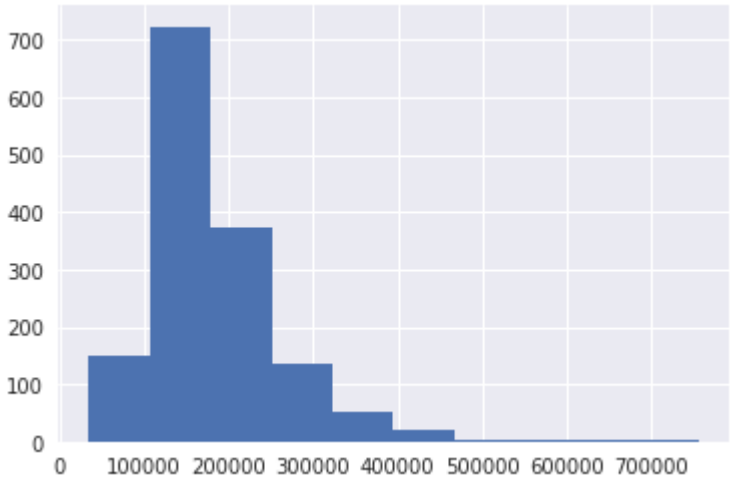
- train.csv - the training set
- test.csv - the test set
- data_description.txt - full description of each column, originally prepared by Dean De Cock but lightly edited to match the column names used here
- sample_submission.csv - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms

Data Columns:

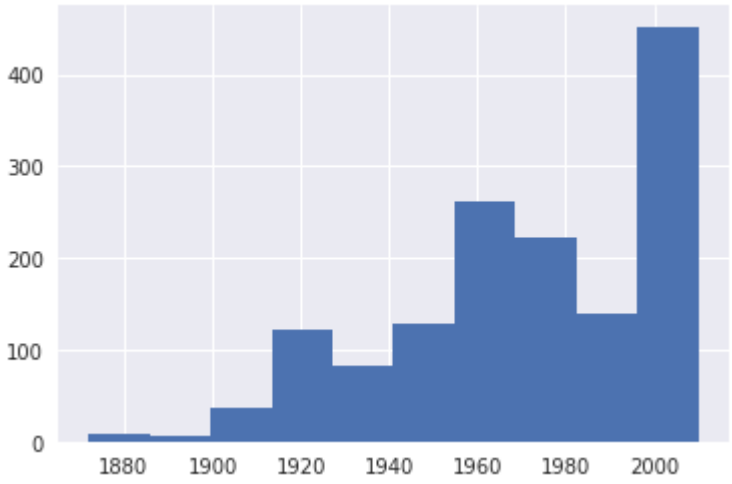
- SalePrice: the property's sale price in dollars. (This is our y-hat prediction)
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of 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

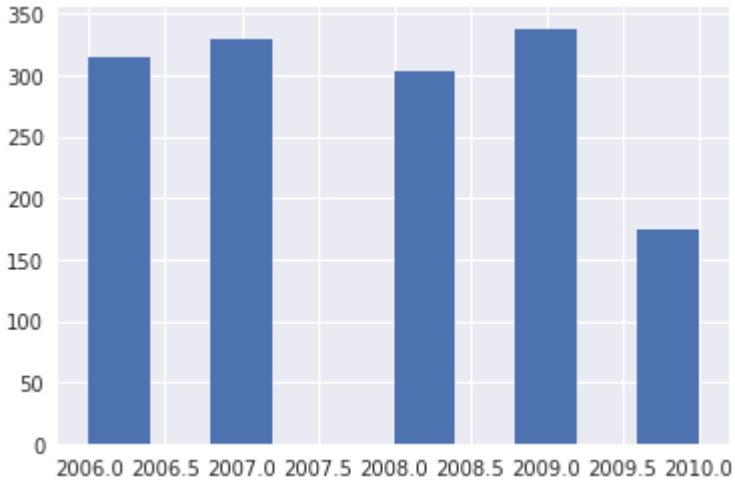
In [62]:
```
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt

data = pd.read_csv('housing/train.csv')
prices = data['SalePrice']
year_built = data['YearBuilt']
year_sold = data['YrSold']
month_sold = data['MoSold']

out = plt.hist(prices)
```



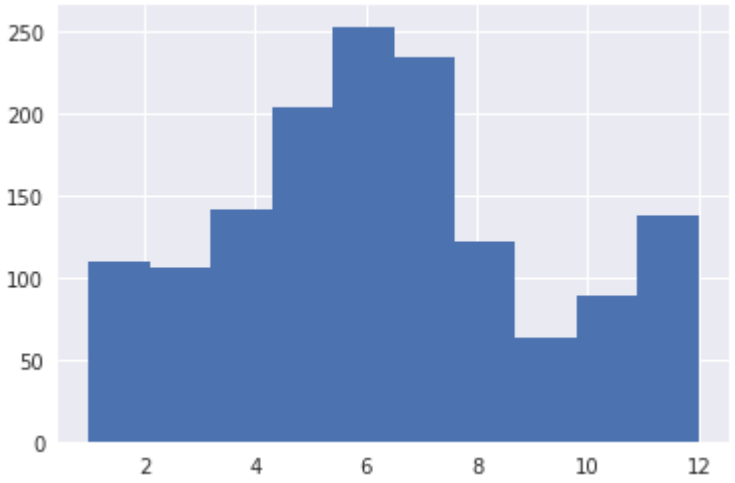
In [59]:
```
out = plt.hist(year_built)
```



In [65]:
```
out = plt.hist(year_sold)
```



In [64]:
```
out = plt.hist(month_sold)
```



Solution Statement

The solution will begin with data exploration to gain intuition and understanding of the attributes and features available. Next, data cleansing will be utilized to remove outliers, correct missing or null values from the training set. Additional analysis will be performed with statistical data visualizations with seaborn and or matplotlib. These visualizations should aid in the selection of features which will be used as dimensions of the following machine learning models: linear regression, SVM, Decision Tree, Random Forest, and XGBoost. After training each model, the model will be evaluated by RMSE between the logarithm of the predicted value and the logarithm of the observed sales price. The algorithms will then be tuned with grid search of hyperparameters. After repeating and refining the scores, a csv file will be submitted for official evaluation against an unknown dataset on Kaggle (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/submit). Submissions will be evaluated by RMSE between the logarithm of the predicted value and the logarithm of the observed sales price.

Benchmark Model

Intially, it was my intent to use the Zillow benchmark in evaluating the effectiveness of my model. Unfortunately, the test.csv will not include the SalePrice information. This would be necessary calculate the median absolute percent error. Instead the benchmark will be a RMSE of 0.43045 which was obtained by taking the mean of all the properties in the dataset and using that value for all records. Ultimately, the goal is to improve this score and rank in the top 20% of the public leaderboard on Kaggle (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/leaderboard). In order to rank successfully in the top 20% the submission would require a RMSE of 0.12049.

Evaluation Metrics

The objective of this Kaggle competition is to predict the sales price for each house in the dataset and submit the results for evaluation on the Kaggle.com website. Submissions are evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)[*] (https://www.kaggle.com/c/house-prices-advanced-regression-techniques#evaluation)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Project Design

An iterative approach will be taken with the steps listed below:

- Data Exploration: Familiarize myself with the data through an explorative process including statistical analysis and data visualization.
- Data Cleansing: Remove outliers, fill null values with mean, one hot encode categorical values, check relevance of every feature to the target feature.
- Train Model: Train several supervised machine learning models and tune with techniques such as cross validation and GridSearchCV.
- Test Model And Optimize: Optimize the model offline with the test dataset until satisifactory RMSE is achieved.
- Submit: Finally submit results to Kaggle for evaluation.
- Repeat as needed to meet appropriate benchmark.