House price prediction using linear regression

Sarthak Shrivastava
Machine learning intern
AI tech systems
sarthak051999@gmail.com
www.ai-techsystems.com

Abstract - Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project, house prices will be predicted given explanatory variables that cover many aspects of residential houses. The goal of this project is to create a regression model that are able to accurately estimate the price of the house given the features. The heart of the problem lies within to develop an efficient 3 layer and a 5 layer neural network architecture that can make house prices predictions and to compare there percentage errors.

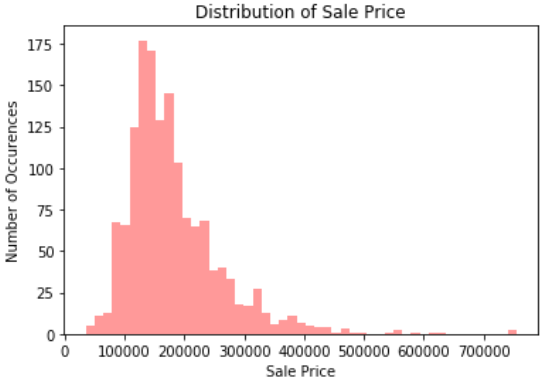
Keywords - exploratory data analysis, data vizualisation, handling missing data, outlier treatment, model training, linear regression, gradient boosting.

I. INTRODUCTION

Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project. house prices will be predicted given explanatory variables that cover many aspects of residential houses. As continuous house prices, they can be predicted with various regression techniques including Lasso, Ridge, SVM regression, and Random Forest regression; as individual price ranges, they can be predicted with classification methods including Naive Bayes, logistic regression, SVM classification, and Random Forest classification. We can also perform PCA to improve the prediction accuracy. The goal of this project is to create a linear regression model that is able to accurately estimate the price of the house given the features.

II. METHODOLOGY

A. Dataset



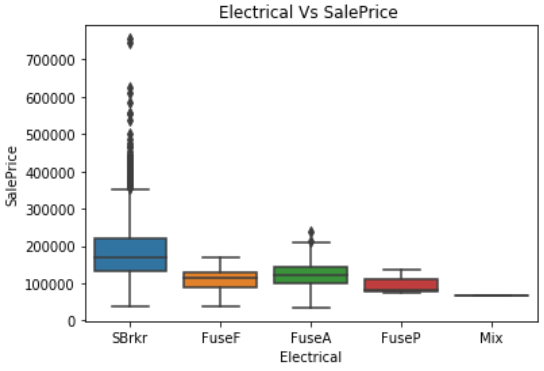
Here is brief description of all data fields.

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet

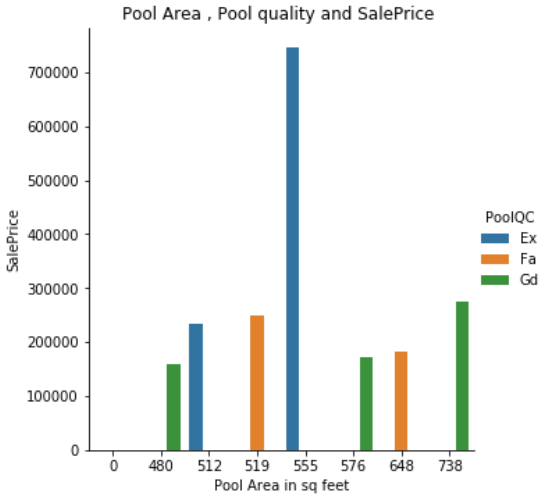
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- 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

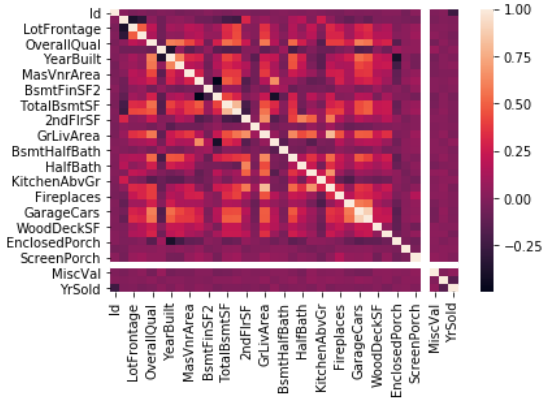
B. Exploratory Data Analysis



Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. We performed some bivariate analysis on the data to get a better overview of the data and to find outliers in our dataset. Outliers can occur due to some kind of errors while collecting the data and need to be removed so that it don't affect the performance of our model.

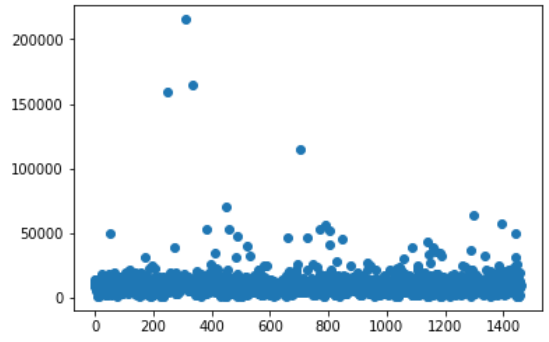


C. Data Preprocessing



The dataset is the prices and features of residential houses sold from 2006 to 2010 in Ames, Iowa, obtained from the Ames Assessor's Office. This dataset consists of 79 house features and 1460 houses with sold prices. Although the dataset is relatively small with only 1460 examples, it contains 79 features such as areas of the houses, types of the floors, and numbers of bathrooms. Such large amounts of features enable us to explore various techniques to predict the house prices. The dataset consists of features in various formats. It has numerical data such as prices and numbers of bathrooms/bedrooms/living rooms, as well as categorical features such as zone classifications for sale, which can be 'Agricultural', 'Residential High Density',

'Residential Low Density', 'Residential Low Density Park', etc. In order to make this data with different format usable for our algorithms, categorical data was converted into separated indicator data, which expands the number of features in this dataset. Besides, there were some features that had values of N/A; we replaced them with the mean/median/0 of their columns so that they don't influence the distribution.



D. Model

We performed supervised learning algorithm 'regression' as it seems more reasonable to perform regression since house prices are continuous.We fitted our model with linear regression algorithm on train data and predicted house prices on the test data. In order to further improve our models, we also performed gradient boosting.We made an instance of gradient boosting regressor and fitted it with our data.For best results and more optimized Model we changed the parameters of gradient boosting regressor.

n_estimators = 15000,
max_depth = 2,
min_samples_split = 2,
learning_rate = 0.005,
loss = 'huber

III. CONCLUSION

For regression problem, the model performed with an rmse 0.14056 without outlier treatment.After treating outliers the rmse reduced to 0.13729 and after removing three columns that were the most prone to outlies the rmse dropped to 0.13747.Finally,improving parameters of gradient boosting resulted in an rmse of 0.13223 which is the best for now with learning rate = 0.005 and n_estimaters =15000, loss= 'huber' instead of ls and depth =2.

IV. REFERENCE

[1] De Cook, Dean. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project." Journal of Statistics Education, vol. 19, no. 3, 2011.