An Investigation of Regression Techniques to Predict Housing Prices

## Team Members:

Brain Hitchcock and Brandon Lasher

## Description of Project:

Using the dataset provided by Kaggle found at <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview> we will build a model to predict the housing prices, when given specific features. We will do an analysis of features to determine which features are the most important. We will look at PCA and other feature simplification techniques to narrow down the feature set. Once the data has been cleaned and the analysis has been completed, we will apply two different regression techniques to attempt to predict the housing price. First, we will use polynomial regression. Second, we will use a random forest. We will use mean squared error, mean absolute error and r2 error to determine how well our models preformed. Once we have completed the creation and testing of our models, we will create a presentation to describe our thought process and approach as well as the results.

## Data Set

The data set for this project can be found <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview>. The dataset consists of the following data points:

|  |  |
| --- | --- |
| **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 |

## Techniques and Platforms

We will use two different ML techniques based on regression. First, we will use polynomial regression and second, we will use random forest regression. We will use mean squared error, mean absolute error and r2 error to determine how well our models preformed. The platform we will use will be Google Colab and their python/Jupiter notebook environment. Colab will have a shared workspace and be able to work and build this project together. Once the project is complete, we will also post it as well as corresponding documentation to github. To visualize and understand the data we will mainly be using the Pandas library. To build the models we will be using the SKlearn library and the respective modules for the regression models we chose.