**Project title: Predicting House Values in Ames, Iowa**

Project Description/Outline

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, our goal is to predict the final price of each home of the test data set using the training data set.

Research Question to Answer

What housing characteristics affect the price of a home most significantly?

H0: The following housing characteristics have no effect on housing prices in Ames, Iowa.

* Year remodeled
* Lot area
* Living area sq ft
* Total basement sq ft
* Overall quality

Ha: The above housing characteristics have an effect on housing prices in Ames, Iowa.

Datasets to be used

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

Rough Breakdown of Tasks:

1. Preprocessing data

Data Overlook

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Numerical Columns | | Categorical Columns | | # of categories |
| LotFrontage | Feet of street connected to house | MSSubClass | Type of dwelling | 16 |
| LotArea | Lot size sqft | MSZoning | Zoning classification | 8 |
| YearBuilt | Orig. construct date | Street | Type of road access | 2 |
| YearRemodAdd | Remodel date (no remodel = same as YearBuilt) | Alley | Type of alley access | 3 |
| MasVnrArea | Masonry veneer area sqft | LotShape | Shape of property | 4 |
| BsmtFinSF1 | Type 1 finished sqft | LandContour | Flatness of property | 4 |
| BsmtFinSF2 | Type 2 finished sqft | Utilities | Utilizes available | 4 |
| BmstUnfSF | Unfinished sqft | LotConfig | Lot configuration | 5 |
| TotalBsmtSF | Total area sqft | LandSlope | Slope of property | 3 |
| 1stFlrSF | 1st floor sqft | Neighborhood | Physical locations w/in Ames | 25 |
| 2ndFlrSF | 2nd floor sqft | Condition1 | Proximity to conditions | 9 |
| LowQualFinSF | Low qual finished sqft | Condition2 | Same as above if more than 1 | 9 |
| GrLivArea | Above grade (ground) living area sqft | BldgType | Type of dwelling (simple) | 5 |
| BsmtFullBath | Basement full bathrooms | HouseStyle | Style of dwelling | 8 |
| BsmtHalfBath | Basement half bathrooms | RoofStyle | Type of roof | 6 |
| FullBath | Full bathrooms above grade | RoofMatl | Roof material | 8 |
| HalfBath | Half bathrooms above grade | Exterior1st | Exterior covering | 17 |
| Bedroom | Bedrooms above grade | Exterior2nd | Same as above if more than 1 | 17 |
| Kitchen | Kitchens above grade | MasVnrType | Masonry veneer type | 5 |
| TotRmsAbvGrd | Total rooms above grade (not bathrm) | ExterQual | Quality of material exterior | 5 |
| Fireplace | # fireplaces | ExterCond | Condition of material exterior | 5 |
| GarageYrBlt | Year garage built | Foundation | Type of foundation | 6 |
| GarageCars | Car capacity garage | BsmtQual | Eval. Height of basement | 6 |
| GarageArea | Size of garage sqft | BsmtExposure | Walkout/garden lvl walls | 5 |
| WoodDeckSF | Wood deck area sqft | BsmtFinType1 | Rating basement finished area | 7 |
| OpenPorchSF | Open porch area sqft | BsmtFinType2 | Same as above if more than 1 | 7 |
| EnclosedPorch | Enclosed porch area sqft | Heating | Heating type | 6 |
| 3SsnPorch | 3 season porch area sqft | HeatingQC | Heating quality and condition | 5 |
| ScreenPorch | Screen porch area sqft | CentralAir | Central AC | 2 (Y/N) |
| PoolArea | Pool area sqft | Electrical | Electrical system | 5 |
| MiscVal | $val of misc feature | KitchenQual | Kitchen quality | 5 |
| MoSold | Month sold (MM) | Functional | Home functionality | 8 |
| YrSold | Year Sold (YYYY) | FireplaceQu | Fireplace quality | 6 |
| OverallQual | Rates material/finish of house (1-10) | GarageType | Garage location | 7 |
| OverallCond | Rates condition of house (1-10) | GarageFinish | Interior finish of garage | 4 |
|  |  | GarageQual | Garage quality | 6 |
|  |  | GarageCond | Garage condition | 6 |
|  |  | PavedDrive | Paved driveway | 3 |
|  |  | PoolQC | Pool quality | 5 |
|  |  | Fence | Fence quality | 5 |
|  |  | MiscFeature | Misc feature not covered | 6 |
|  |  | SaleType | Type of sale | 10 |
|  |  | SaleCondition | Condition of sale | 6 |

1. Adjusting categoricals/data munging (7/23/19 in-class)

* may need to change categorical data into numeric representations (looks like Chris’ ipynb does this already)
* NA / NaN / Missing value identification and cleansing

1. Data analysis / Regression analysis (Deadline: 7/25/19)

* Thank you Chris for the very in-depth preliminary but very descriptive dive into the data with 1) one-way graphs with all columns vs. sales price, 2) average sales price by selected field, 3) GLM analyses.
* Do we approach Machine Learning now? XGBoost, MLBox, regression techniques, lasso, sklearn, neural network, gradient boosting

**Data Observation**

* Histograms of Sales Price Frequency and log of Sales Price Frequency

스크린샷이(가) 표시된 사진

자동 생성된 설명스크린샷이(가) 표시된 사진

자동 생성된 설명

* Residual Deviance

텍스트이(가) 표시된 사진

자동 생성된 설명

* When considering deviance as a suitable measure of goodness of fit of a GLM, tt can be observed that 4 of our 5 independent variables are listed in the top 20 of 79 variables. Dev1(log sales price ~ columns) and Dev2 (sales price ~ columns) were calculated by subtracting residual deviance of the data from the null deviance.
* One-way plots of target features (LotArea, OverallQual, YearRemodAdd, GrLivArea, & TotalBsmtSF)

스크린샷이(가) 표시된 사진

자동 생성된 설명

Table 1. GLM (Gaussian)

텍스트이(가) 표시된 사진

자동 생성된 설명

Table 2. GLM (Poisson)

텍스트이(가) 표시된 사진

자동 생성된 설명

Table 3. GLM (Gamma)

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

1. Hypothesis testing (Deadline: 7/25/19)

* In consideration of the p-values <0.05, all three GLMs show statistical significance that the independent variables, LotArea, OverallQual, YearRemodAdd, GrLivArea, and TotalBsmtSF, affect the dependent variable, Sale Price. In other words, the null hypothesis, stating that these variables have no effect on Sale Price of Houses in Ames, can be rejected. (Further discussion?)

1. Creation of presentation and story (Deadline: 7/27/19)
2. Presentation layout/practice (Deadline: 7/27/19)
3. Presentation Date: 7/30/19 (next Tuesday)