|  |
| --- |
| **Elmhurst College** |
| Classifying the Price of a House in Ames, IA |
| MSD 570 Fall A 2019 |

|  |
| --- |
| Megan Cusey  12-9-2019 |

# Overview

The following case study seeks to identify the strongest predictive variables in order to predict house sales prices in Ames, IA. A CRISP approach was applied in order to accomplish this task. The following technical documentation breaks down each component of this analysis into 1.) Business Understanding, 2.) Data Understanding, 3.) Data Preparation, 4.) Modeling, 5.) Evaluation, and 6.) Next Steps (what would typically be deployment). A compilation of all the files used in this project can be found on my GitHub repository titled [DataScienceProjects/MDS 556 – House Prices Regression/](https://github.com/megancusey/DataScienceProjects/tree/master/MDS%20556%20-%20House%20Prices%20Regression).

# Business/Data Understanding

The data set for the Ames, IA house market obtains data that describes the features of each house. The data and business problem originates from a Kaggle Competition data set that you can find [here](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview). For the purpose of this assignment, the decision was made to tweak the suggested business problem from Kaggle which was to predict the sales price of a home given a set of features to, identifying houses that are in the top 25% in sales price. Identifying houses that sell in the 25% of house prices is useful for two reasons:

1. If you are a realtor that focuses on only the most luxurious homes, this model can quickly advise you if listing a certain home may be worth his or her time.
2. If you are a real estate developer, you would want to identify homes that have a high potential for profit that may be listed at a price that under values the home.
3. As a real estate developer, knowing what features of the home to invest in can be useful to maximize profit.

In both cases, the cost of inaccurately classifying a home as a “Top Dollar” home can be significant. A real estate agent may experience an opportunity cost of focusing resources on a less profitable home as opposed to a real “Top Dollar” home. Similarly, a real estate developer may apply more labor and material into a home that will not produce the expected sale price. For these reasons, a successful model will have a high precision rate. The best models will produce results that minimizes false positives.

A [data dictionary](https://github.com/megancusey/DataScienceProjects/blob/master/MDS%20556%20-%20House%20Prices%20Regression/data_description.txt) was supplied along with the data set. While most of the features were relatively self-explanatory, there was a few that didn’t seem to be common knowledge to me, a 1-time homeowner. The following are definitions I researched in order to complete this analysis and solidify my business understanding:

* Artery Street, a highway
* Feeder Street, an important road that feeds into a highly trafficked area
* Masonry Veneer, an external layer to the home that is non-structural (example: brick or stone)

# Data Preparation

## Data Cleaning

The train.csv dataset contains 1,460 rows of data with 81 features. A combination of numeric, categorical, and ordinal data types where present in the data. The first step in data preparation was to create a method to input the raw data directly from the csv file, investigate each feature, build validations for each feature according to the data definitions provided, and ensure that no missing values were presented. The following provides general guidelines for this process:

* If the feature had an ordinal nature, convert the variable to a factor. Specify the parameter as ordered = TRUE and apply the levels in the appropriate order.
  + NA was used to specify that a feature was not present such as “No Fireplace” or “No Garage”. Missing values were mapped to the appropriate level that indicate the absence of a characteristic of a home.
* If the feature was categorical, convert the variable to a factor.
  + NA was used to specify that a feature was not present such as “No Fireplace” or “No Garage”. Missing values were mapped to the appropriate level that indicate the absence of a characteristic of a home.
* If the feature was numeric, I mostly let R default the data type to what it was when it loaded the data set.
  + In some features, a missing value indicates that the feature didn’t exist. In this case, I supplied 0 for those observations.
  + In other features, the mode was used to replace a missing value.

## Data Exploration

About the Target

A screenshot of a cell phone

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| summary(data$SalePrice) | | | | | |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| 34900 | 129975 | 163000 | 180921 | 214000 | 755000 |

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 34900 129975 163000 180921 214000 755000

Numeric Features Distribution

A screenshot of a cell phone

Description automatically generated

Numeric variables correlations

A screenshot of a cell phone

Description automatically generated

### Feature Engineering

* Domain Knowledge
* Interactive features – combine 2 or more features (products, sums, or differences)
* Combine Sparse Classes

Overall Condition – most of the observations where Average so I made Average its own category and combined categories beyond Average as Above Average and less than Average as Below Average

Overall Quality is a bit different because Very Good-Average contain the bulk of the data points. Will probably separate differently.

Condition1 and 2 I did right away b/c the columns meant the same features, the extra column was just used if there was another feature of the house that it applied to. Instead, I split out the conditions into separate columns. Had to look up what feeder and arterial street was

Masonry Veneer?

Turned remodel into a Boolean though I think there’s some room for the algorithm to mistunderstand this. If the YearBuilt is newer, then it wouldn’t have a remodel on it.

### Principal Component Analysis

# Modeling

## Linear/Logistic Regression

## Decision Tree with Bagging/Random Forest

## General Additive Model

## Kernel Method

## Support Vector Machines

# Evaluation

# Appendix

## Data Dictionary & Thoughts

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Description | Feature Engineering Type | Initial Thoughts |
| MSSubClass | Identifies the type of dwelling involved in the sale   |  |  |  | | --- | --- | --- | | 20 | 1-STORY 1946 & NEWER ALL STYLES | | | 30 | 1-STORY 1945 & OLDER | | | 40 | 1-STORY W/FINISHED ATTIC ALL AGES | | | 45 | 1-1/2 STORY - UNFINISHED ALL AGES | | 50 | 1-1/2 STORY FINISHED ALL AGES | | 60 | 2-STORY 1946 & NEWER | | 70 | 2-STORY 1945 & OLDER | | 75 | 2-1/2 STORY ALL AGES | | 80 | SPLIT OR MULTI-LEVEL | | 85 | SPLIT FOYER | | 90 | DUPLEX - ALL STYLES AND AGES | | 120 | 1-STORY PUD (Planned Unit Development) - 1946 & NEWER | | 150 | 1-1/2 STORY PUD - ALL AGES | | 160 | 2-STORY PUD - 1946 & NEWER | | 180 | PUD - MULTILEVEL - INCL SPLIT LEV/FOYER | | 190 | 2 FAMILY CONVERSION - ALL STYLES AND AGES | |  | Maybe break out into different features?   * # of stories * Finished attic * Age? * Duplex |
| MSZoning | Identifies the general zoning classification of the sale.   |  |  | | --- | --- | | A | Agriculture | | C | Commercial | | FV | Floating Village Residential | | I | Industrial | | RH | Residential High Density | | RL | Residential Low Density | | RP | Residential Low Density Park | | RM | Residential Medium Density | |  |  |
| LotFrontage | Linear feet of street connected to property |  |  |
| LotArea | Lot size in square feet |  |  |
| Street | Type of road access to property   |  |  | | --- | --- | | Grvl | Gravel | | Pave | Paved | |  |  |
| Alley | Type of alley access to property   |  |  | | --- | --- | | Grvl | Gravel | | Pave | Paved | | NA | No alley access | |  |  |
| LotShape | General shape of property   |  |  | | --- | --- | | Reg | Regular | | IR1 | Slightly Irregular | | IR2 | Moderately Irregular | | IR3 | Irregular | |  |  |
| LandContour | Flatness of the property   |  |  | | --- | --- | | Lvl | Near Flat/Level | | Bnk | Banked – Quick and significant rise from street to building. | | HLS | Hillside – Significant slope from side to side | | Low | Depression | |  |  |
| Utilities | Type of utilities available |  |  |
| LotConfig | Lot configuration |  |  |
| LandSlope | Slope of property |  |  |
| Neighborhood | Physical locations within Ames city limits |  |  |
| Condition1 | Proximity to various conditions |  |  |
| Condition2 | Proximity to various conditions (if more than one is present) |  |  |
| BldgType | Type of dwelling |  |  |
| HouseStyle | Style of dwelling |  |  |
| OverallQual | Rates the overall material and finish of the house |  |  |
| OverallCond | Rates the overall condition of the house |  |  |
| YearBuilt | Original construction date |  |  |
| YearRemodelAdd | Remodel date (same as construction date if no remodeling or additions) |  |  |
| 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 | Evaluates the quality of the material on the exterior |  |  |
| ExterCond | Evaluates the present condition of the material on the exterior |  |  |
| Foundation | Type of foundation |  |  |
| BsmtQual | Evaluates the height of the basement |  |  |
| BsmtCond | Evaluates the general condition of the basement |  |  |
| BsmtExposure | Refers to walkout or garden level walls |  |  |
| BsmtFinType1 | Rating of the basement finished area |  |  |
| BsmtFinSF1 | Type 1 finished square feet |  |  |
| BsmtFinType2 | Rating of basement finished area (if multiple types) |  |  |
| BsmtFinSF2 | Type 2 finished square feet |  |  |
| BsmtUnfSF | Unfinished square feet of basement area |  |  |
| TotalBsmtSF | Total square feet of basement area |  |  |
| Heating | Type of hearing |  |  |
| HearingQC | 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 bathrroms |  |  |
| FullBath | Bull bathrooms about grade |  |  |
| HalfBath | Half baths above grade |  |  |
| Bedroom | Bedrooms above grade (does not include basement bedrooms) |  |  |
| Kitchen | Kitchens above grade |  |  |
| KitchenQual | Kitchen Quality |  |  |
| TotRmsAbvGrd | Total rooms above grade (does not include bathrooms) |  |  |
| Functional | Home functionality ( Assume typical unless deductions are warranted) |  |  |
| 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 are in square feet |  |  |
| OpenPorchSF | Open porch area in square feet |  |  |
| Enclosed Porch | Enclosed porch area in square feet |  |  |
| 3SsnPorch | Three season porch area in square feet |  |  |
| ScreenPorcch | 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 (MM) |  |  |
| YrSold | Year Sold (YYYY) |  |  |
| SaleType | Type of sale |  |  |
| SaleCondition | Condition of sale |  |  |
| Sale price |  |  |  |