PH125.9x - Capstone: House Prices

Murilo Mendel Costa

Abstract

This project aim to predict the price of a house based it location, size, built year and other 76 features. This report will cover the process of loading the dataset, apply feature engineering methods and fit a model to predict the price of a house based on it characteristics. The dataset was obtained on Kaggle platform, and this content is available for knowledge improvement purpouse.

Keywords: house prices, predition, harvardX, capstone, r markdown

---

1 Introduction

House Prices is a Kaggle competition for knowledge improvement purpose which the main challenge is to develop a predictive model capable to estimate a house price based on it own characteristics. The data set was extracted from the competition web page (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/) and has 79 features to be explored in order to reach the best predictive model.

The train and test data set are composed by houses from Ames, Iowa and many of their characteristics.The Ames Housing dataset was compiled by Dean De Cock for use in data science education.

The project approach is to make a feature engineering on the available data and build a model to predict a house price using features from the train set that most minimizes the loss function when applied to the validation set. A residual mean squared error (`RMSE`) was used as the loss function (the typical error we make when predicting a movie rating) for performance measure.

2 Methodology

As the first step, we import the Ames House data set from the previous downloaded file on Kaggle web page. Once we had a train set, we analyzed some of its properties through descriptive statistics and histograms, in order to gain insights for the prediction model. Then, we looked for opportunities to fit the data and select the best predictors in order to train different models and evaluate each of them to reach the best approach.

Finally, the model was built using 25 most relevant features `House Price` set.

2.1 Data extraction

```
# Import libraries
library(tidyverse)
library(caret)
library(data.table)
library(ggplot2)
library(knitr)
library(stringr)
library(readr)
library(rpart)
library(leaps)
library(corrplot)
library(randomForest)

# The data was downloaded from: https://www.kaggle.com/c/house-prices-advanced-regression-techni
# The downloaded file is stored in the project current directory, inside the \data folder

# Read the train and test .csv file
# Read the train and test .csv file
HP <- as.data.frame(read_csv(file.path(getwd(),"/data/train.csv"), col_names = TRUE))
```

2.2 Data description

The `Ames House Price` train set contains `1312` rows and `81` columns, each variable meaning: > 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

We can analyze the data set using some descriptive statistics.

```
dim(HP)
```

```
## [1] 1460   81
```

```
summary(HP)
```

```
##        Id           MSSubClass      MSZoning          LotFrontage
##  Min.   :   1.0   Min.   : 20.0   Length:1460        Min.   : 21.00
##  1st Qu.: 365.8   1st Qu.: 20.0   Class :character   1st Qu.: 59.00
##  Median : 730.5   Median : 50.0   Mode  :character   Median : 69.00
##  Mean   : 730.5   Mean   : 56.9                      Mean   : 70.05
##  3rd Qu.:1095.2   3rd Qu.: 70.0                      3rd Qu.: 80.00
##  Max.   :1460.0   Max.   :190.0                      Max.   :313.00
##                                                      NA's   :259
##     LotArea          Street             Alley             LotShape
##  Min.   :  1300   Length:1460        Length:1460        Length:1460
##  1st Qu.:  7554   Class :character   Class :character   Class :character
##  Median :  9478   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 10517
##  3rd Qu.: 11602
##  Max.   :215245
##
##  LandContour         Utilities          LotConfig          LandSlope
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Neighborhood        Condition1         Condition2          BldgType
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
```

```
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   HouseStyle         OverallQual      OverallCond      YearBuilt
##  Length:1460        Min.   : 1.000   Min.   :1.000   Min.   :1872
##  Class :character   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954
##  Mode  :character   Median : 6.000   Median :5.000   Median :1973
##                     Mean   : 6.099   Mean   :5.575   Mean   :1971
##                     3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000
##                     Max.   :10.000   Max.   :9.000   Max.   :2010
##
##   YearRemodAdd   RoofStyle           RoofMatl         Exterior1st
##  Min.   :1950   Length:1460        Length:1460        Length:1460
##  1st Qu.:1967   Class :character   Class :character   Class :character
##  Median :1994   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1985
##  3rd Qu.:2004
##  Max.   :2010
##
##  Exterior2nd         MasVnrType          MasVnrArea      ExterQual
##  Length:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median :   0.0   Mode  :character
##                                        Mean   : 103.7
##                                        3rd Qu.: 166.0
##                                        Max.   :1600.0
##                                        NA's   :8
##   ExterCond          Foundation          BsmtQual           BsmtCond
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  BsmtExposure       BsmtFinType1         BsmtFinSF1     BsmtFinType2
##  Length:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median : 383.5   Mode  :character
##                                        Mean   : 443.6
##                                        3rd Qu.: 712.2
##                                        Max.   :5644.0
##
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF       Heating
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Length:1460
##  1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   Class :character
```

```
##  Median :   0.00   Median : 477.5   Median : 991.5   Mode  :character
##  Mean   :  46.55   Mean   : 567.2   Mean   :1057.4
##  3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2
##  Max.   :1474.00   Max.   :2336.0   Max.   :6110.0
##
##   HeatingQC          CentralAir         Electrical          1stFlrSF
##  Length:1460        Length:1460        Length:1460        Min.   : 334
##  Class :character   Class :character   Class :character   1st Qu.: 882
##  Mode  :character   Mode  :character   Mode  :character   Median :1087
##                                                           Mean   :1163
##                                                           3rd Qu.:1391
##                                                           Max.   :4692
##
##     2ndFlrSF     LowQualFinSF       GrLivArea     BsmtFullBath
##  Min.   :   0   Min.   :  0.000   Min.   : 334   Min.   :0.0000
##  1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130   1st Qu.:0.0000
##  Median :   0   Median :  0.000   Median :1464   Median :0.0000
##  Mean   : 347   Mean   :  5.845   Mean   :1515   Mean   :0.4253
##  3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777   3rd Qu.:1.0000
##  Max.   :2065   Max.   :572.000   Max.   :5642   Max.   :3.0000
##
##   BsmtHalfBath        FullBath        HalfBath       BedroomAbvGr
##  Min.   :0.00000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000
##  Median :0.00000   Median :2.000   Median :0.0000   Median :3.000
##  Mean   :0.05753   Mean   :1.565   Mean   :0.3829   Mean   :2.866
##  3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :2.00000   Max.   :3.000   Max.   :2.0000   Max.   :8.000
##
##   KitchenAbvGr   KitchenQual         TotRmsAbvGrd     Functional
##  Min.   :0.000   Length:1460        Min.   : 2.000   Length:1460
##  1st Qu.:1.000   Class :character   1st Qu.: 5.000   Class :character
##  Median :1.000   Mode  :character   Median : 6.000   Mode  :character
##  Mean   :1.047                      Mean   : 6.518
##  3rd Qu.:1.000                      3rd Qu.: 7.000
##  Max.   :3.000                      Max.   :14.000
##
##    Fireplaces    FireplaceQu         GarageType         GarageYrBlt
##  Min.   :0.000   Length:1460        Length:1460        Min.   :1900
##  1st Qu.:0.000   Class :character   Class :character   1st Qu.:1961
##  Median :1.000   Mode  :character   Mode  :character   Median :1980
##  Mean   :0.613                                         Mean   :1979
##  3rd Qu.:1.000                                         3rd Qu.:2002
##  Max.   :3.000                                         Max.   :2010
##                                                        NA's   :81
##  GarageFinish         GarageCars      GarageArea      GarageQual
##  Length:1460        Min.   :0.000   Min.   :   0.0   Length:1460
##  Class :character   1st Qu.:1.000   1st Qu.: 334.5   Class :character
```

```
##  Mode  :character   Median :2.000   Median : 480.0   Mode  :character
##                     Mean   :1.767   Mean   : 473.0
##                     3rd Qu.:2.000   3rd Qu.: 576.0
##                     Max.   :4.000   Max.   :1418.0
##
##   GarageCond         PavedDrive          WoodDeckSF      OpenPorchSF
##  Length:1460        Length:1460        Min.   :  0.00   Min.   :  0.00
##  Class :character   Class :character   1st Qu.:  0.00   1st Qu.:  0.00
##  Mode  :character   Mode  :character   Median :  0.00   Median : 25.00
##                                        Mean   : 94.24   Mean   : 46.66
##                                        3rd Qu.:168.00   3rd Qu.: 68.00
##                                        Max.   :857.00   Max.   :547.00
##
##  EnclosedPorch      3SsnPorch       ScreenPorch        PoolArea
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :  0.00   Median :  0.00   Median :  0.00   Median :  0.000
##  Mean   : 21.95   Mean   :  3.41   Mean   : 15.06   Mean   :  2.759
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :552.00   Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##     PoolQC             Fence           MiscFeature          MiscVal
##  Length:1460        Length:1460        Length:1460        Min.   :    0.00
##  Class :character   Class :character   Class :character   1st Qu.:    0.00
##  Mode  :character   Mode  :character   Mode  :character   Median :    0.00
##                                                           Mean   :   43.49
##                                                           3rd Qu.:    0.00
##                                                           Max.   :15500.00
##
##      MoSold           YrSold       SaleType         SaleCondition
##  Min.   : 1.000   Min.   :2006   Length:1460        Length:1460
##  1st Qu.: 5.000   1st Qu.:2007   Class :character   Class :character
##  Median : 6.000   Median :2008   Mode  :character   Mode  :character
##  Mean   : 6.322   Mean   :2008
##  3rd Qu.: 8.000   3rd Qu.:2009
##  Max.   :12.000   Max.   :2010
##
##    SalePrice
##  Min.   : 34900
##  1st Qu.:129975
##  Median :163000
##  Mean   :180921
##  3rd Qu.:214000
##  Max.   :755000
##
```

```
head(HP)
```

```
##   Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1  1         60       RL          65    8450   Pave  <NA>      Reg         Lvl
## 2  2         20       RL          80    9600   Pave  <NA>      Reg         Lvl
## 3  3         60       RL          68   11250   Pave  <NA>      IR1         Lvl
## 4  4         70       RL          60    9550   Pave  <NA>      IR1         Lvl
## 5  5         60       RL          84   14260   Pave  <NA>      IR1         Lvl
## 6  6         50       RL          85   14115   Pave  <NA>      IR1         Lvl
##   Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
## 1    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 2    AllPub       FR2       Gtl      Veenker      Feedr       Norm     1Fam
## 3    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 4    AllPub    Corner       Gtl      Crawfor       Norm       Norm     1Fam
## 5    AllPub       FR2       Gtl      NoRidge       Norm       Norm     1Fam
## 6    AllPub    Inside       Gtl      Mitchel       Norm       Norm     1Fam
##   HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
## 1     2Story           7           5      2003         2003     Gable  CompShg
## 2     1Story           6           8      1976         1976     Gable  CompShg
## 3     2Story           7           5      2001         2002     Gable  CompShg
## 4     2Story           7           5      1915         1970     Gable  CompShg
## 5     2Story           8           5      2000         2000     Gable  CompShg
## 6     1.5Fin           5           5      1993         1995     Gable  CompShg
##   Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
## 1     VinylSd     VinylSd    BrkFace        196        Gd        TA      PConc
## 2     MetalSd     MetalSd       None          0        TA        TA     CBlock
## 3     VinylSd     VinylSd    BrkFace        162        Gd        TA      PConc
## 4     Wd Sdng     Wd Shng       None          0        TA        TA     BrkTil
## 5     VinylSd     VinylSd    BrkFace        350        Gd        TA      PConc
## 6     VinylSd     VinylSd       None          0        TA        TA       Wood
##   BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
## 1       Gd       TA           No          GLQ        706          Unf
## 2       Gd       TA           Gd          ALQ        978          Unf
## 3       Gd       TA           Mn          GLQ        486          Unf
## 4       TA       Gd           No          ALQ        216          Unf
## 5       Gd       TA           Av          GLQ        655          Unf
## 6       Gd       TA           No          GLQ        732          Unf
##   BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
## 1          0       150         856    GasA        Ex          Y      SBrkr
## 2          0       284        1262    GasA        Ex          Y      SBrkr
## 3          0       434         920    GasA        Ex          Y      SBrkr
## 4          0       540         756    GasA        Gd          Y      SBrkr
## 5          0       490        1145    GasA        Ex          Y      SBrkr
## 6          0        64         796    GasA        Ex          Y      SBrkr
##   1stFlrSF 2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
## 1      856      854            0      1710            1            0        2
## 2     1262        0            0      1262            0            1        2
```

```
## 3      920      866           0      1786           1           0        2
## 4      961      756           0      1717           1           0        1
## 5     1145     1053           0      2198           1           0        2
## 6      796      566           0      1362           1           0        1
##   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
## 1        1            3           1          Gd            8        Typ
## 2        0            3           1          TA            6        Typ
## 3        1            3           1          Gd            6        Typ
## 4        0            3           1          Gd            7        Typ
## 5        1            4           1          Gd            9        Typ
## 6        1            1           1          TA            5        Typ
##   Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1          0        <NA>     Attchd        2003          RFn          2
## 2          1          TA     Attchd        1976          RFn          2
## 3          1          TA     Attchd        2001          RFn          2
## 4          1          Gd     Detchd        1998          Unf          3
## 5          1          TA     Attchd        2000          RFn          3
## 6          0        <NA>     Attchd        1993          Unf          2
##   GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
## 1        548         TA         TA          Y          0          61
## 2        460         TA         TA          Y        298           0
## 3        608         TA         TA          Y          0          42
## 4        642         TA         TA          Y          0          35
## 5        836         TA         TA          Y        192          84
## 6        480         TA         TA          Y         40          30
##   EnclosedPorch 3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature MiscVal
## 1             0         0           0        0   <NA>  <NA>        <NA>       0
## 2             0         0           0        0   <NA>  <NA>        <NA>       0
## 3             0         0           0        0   <NA>  <NA>        <NA>       0
## 4           272         0           0        0   <NA>  <NA>        <NA>       0
## 5             0         0           0        0   <NA>  <NA>        <NA>       0
## 6             0       320           0        0   <NA> MnPrv        Shed     700
##   MoSold YrSold SaleType SaleCondition SalePrice
## 1      2   2008       WD        Normal    208500
## 2      5   2007       WD        Normal    181500
## 3      9   2008       WD        Normal    223500
## 4      2   2006       WD       Abnorml    140000
## 5     12   2008       WD        Normal    250000
## 6     10   2009       WD        Normal    143000
```

The first step is removing features with much NAs values. The cut point is over 40% of the data composed by NA.

```
# Dealing with features with high null ratio
missing_rows <- HP[!complete.cases(HP),]
nrow(missing_rows) # At first, there are 1460 missing rows
```

```
## [1] 1460
```

```r
# Checking the quantity of NULL values for each column, and remove columns with more than 40% NU
for(col in colnames(HP)){
  null_percent = length(which(is.na(HP[[col]]))) / length(HP[[col]])
  if(null_percent > 0.4){
    HP[[col]] <- NULL
  }
}

# After the first clean, there are still 366 missing rows
missing_rows <- HP[!complete.cases(HP),]
nrow(missing_rows)
```

```
## [1] 366
```

```r
rm(missing_rows)
```

The next step is completing the missing information. First, we need to know the columns which has NAs values.

```r
# Dealing with null values
# Summary of NAs per columns
NAcol <- which(colSums(is.na(HP)) > 0)
sort(colSums(sapply(HP[NAcol], is.na)), decreasing = TRUE)
```

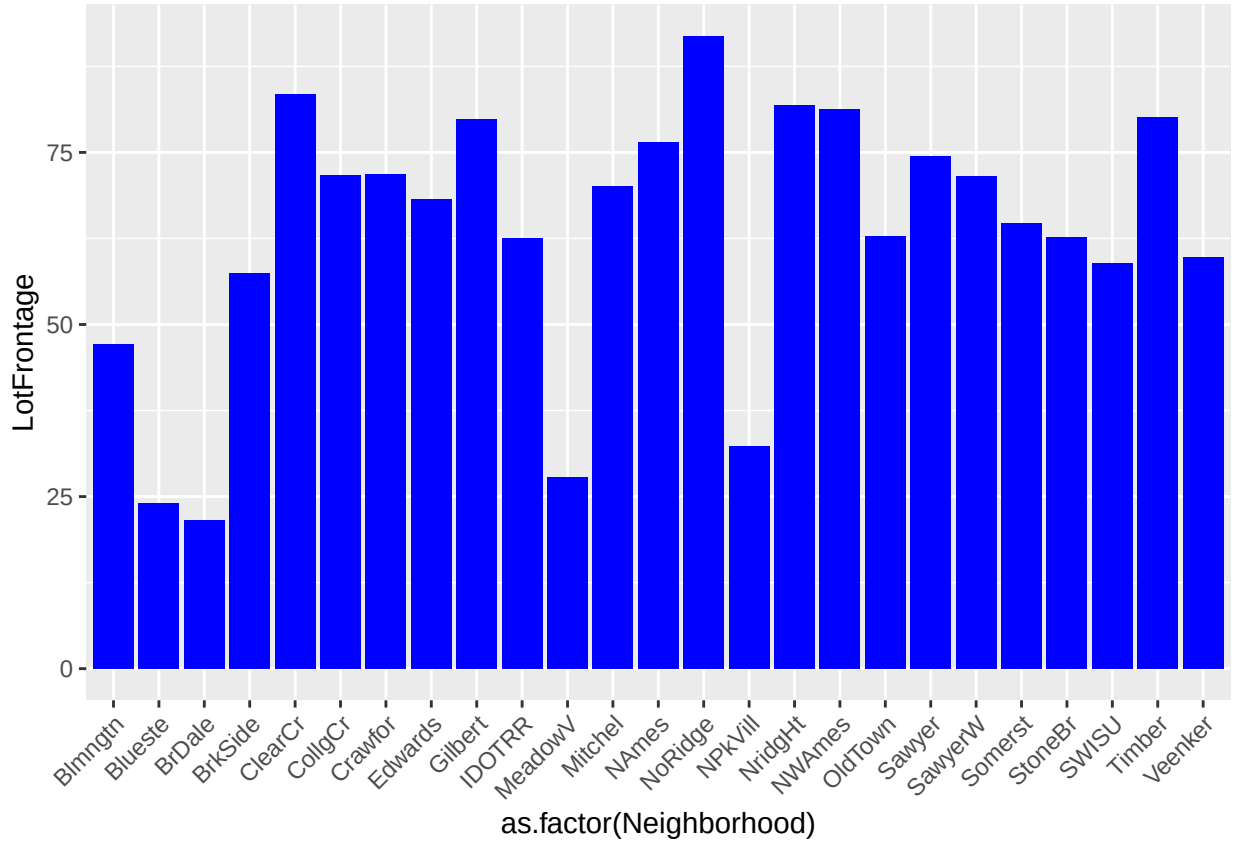
```
##  LotFrontage   GarageType  GarageYrBlt GarageFinish   GarageQual   GarageCond
##          259           81           81           81           81           81
## BsmtExposure BsmtFinType2     BsmtQual     BsmtCond BsmtFinType1   MasVnrType
##           38           38           37           37           37            8
##   MasVnrArea   Electrical
##            8            1
```

The first columns to be filled is the 'LotFrontage', which is related to the linear feet of street connected to property. To fill this values, we look for the mean value over the neighborhoods.

```r
# LotFrontaage Variable = 259 NAs
sum(is.na(HP$LotFrontage))
```

```
## [1] 259
```

```r
# Plotting the variable over the Neighborhoods possibilities
ggplot(HP[!is.na(HP$LotFrontage),], aes(x=as.factor(Neighborhood), y=LotFrontage)) +
  geom_bar(stat='summary', fun.y = "median", fill='blue') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Complete the NAs with the median of it neighborhood
for (i in 1:nrow(HP)){
  if(is.na(HP$LotFrontage[i])){
    HP$LotFrontage[i] <- as.integer(median(HP$LotFrontage[HP$Neighborhood==HP$Neighborhood[i]],
  }
}
```

The next columns to be filled is the GarageType. In this case, the garage type with no value will be filled with 'No Garage' information

```
# Garage Type Variable = 81 NAs
HP$GarageType[is.na(HP$GarageType)] <- 'No Garage'
HP$GarageFinish[is.na(HP$GarageFinish)] <- 'None'
HP$GarageQual[is.na(HP$GarageQual)] <- 'None'
HP$GarageCond[is.na(HP$GarageCond)] <- 'None'
HP$GarageYrBlt[is.na(HP$GarageYrBlt)] <- HP$YearBuilt[is.na(HP$GarageYrBlt)]
```

All other garage variable have the same NAs values, probably caused by the fact that these rows are related to houses that ha no garage.

Basement Variables has almost the same quantity of NAs values, related probably to the same houses. Looking for rows with only one NA to ensure the remaining are related to the same houses.

```
# Basement Variables = 37 NAs
HP[!is.na(HP$BsmtExposure) & (is.na(HP$BsmtFinType2)|is.na(HP$BsmtQual)|is.na(HP$BsmtCond)|is.na
```

```
##     BsmtExposure BsmtFinType1 BsmtQual BsmtCond BsmtFinType2
## 333           No          GLQ       Gd       TA         <NA>
```

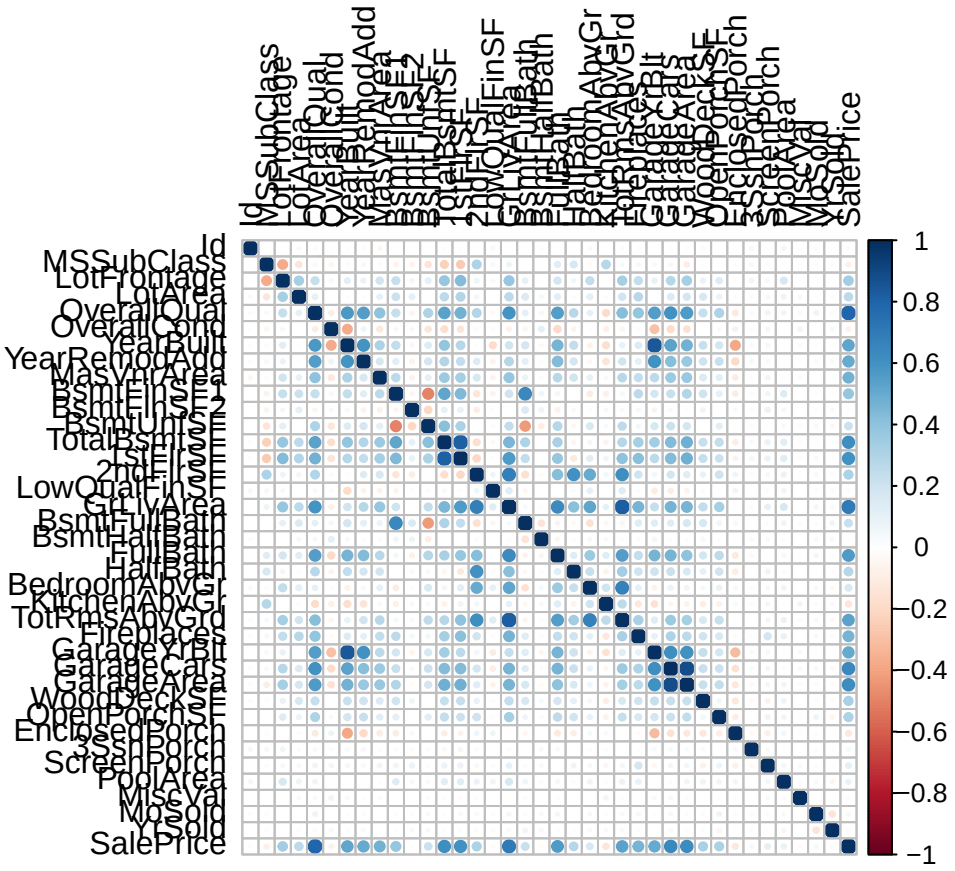
```
HP$BsmtFinType2[333] <- names(sort(-table(HP$BsmtFinType2)))[1]

HP$BsmtExposure[is.na(HP$BsmtExposure)] <- 'None'
HP$BsmtQual[is.na(HP$BsmtQual)] <- 'None'
HP$BsmtCond[is.na(HP$BsmtCond)] <- 'None'
HP$BsmtFinType1[is.na(HP$BsmtFinType1)] <- 'None'
HP$BsmtFinType2[is.na(HP$BsmtFinType2)] <- 'None'
```

Finishing, we fill the Masonery and Electrical features

```
# Masonry veneer features = 8 NAs
HP$MasVnrType[is.na(HP$MasVnrType)] <- 'None'
HP$MasVnrArea[is.na(HP$MasVnrArea)] <-0

# Electrical Information = 1 NA
HP$Electrical[is.na(HP$Electrical)] <- names(sort(-table(HP$Electrical)))[1]
```

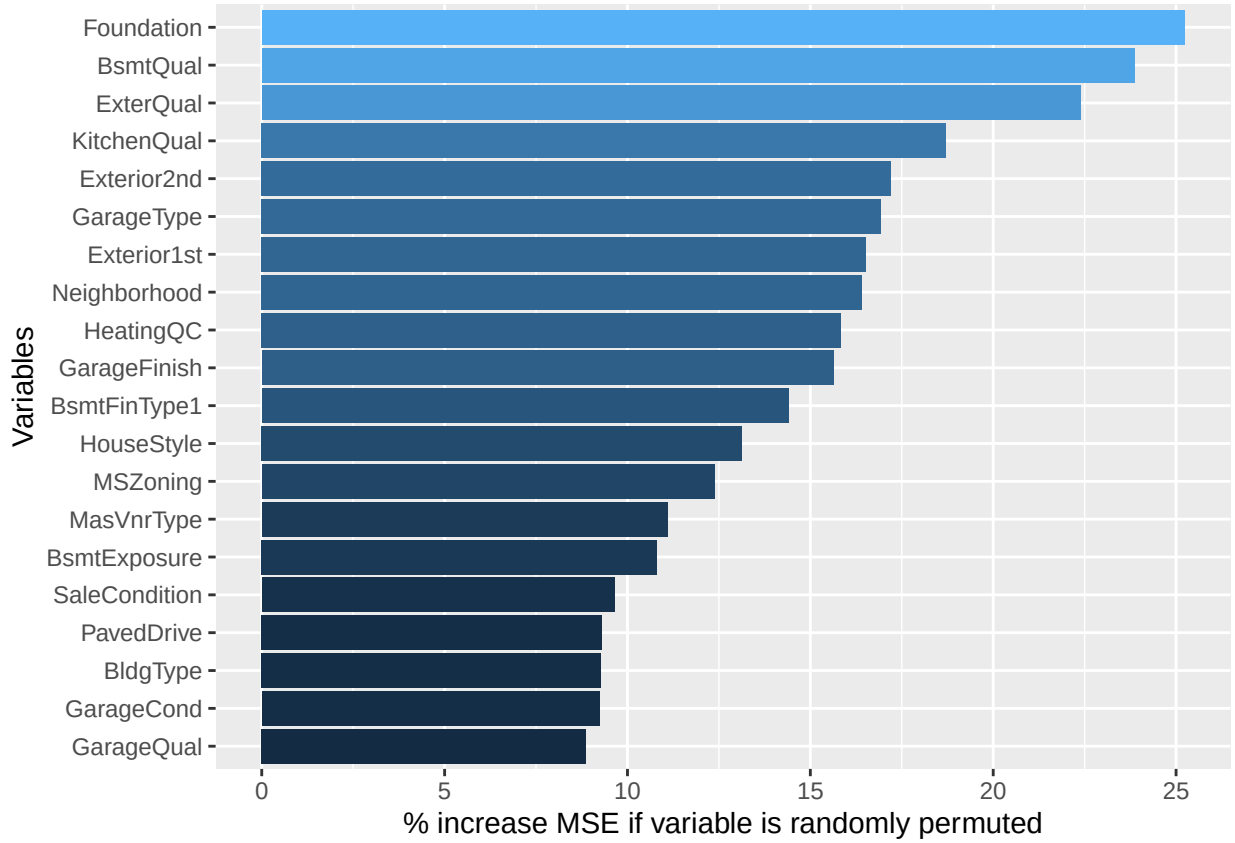
The next step, we will have a first look to the numerical data and select the most relevant ones based on the correlation between SalePrice feature.

```
HP.numerical <- HP[sapply(HP, is.numeric)]

# Numerical features first look
# Calculating features correlation
numericalFeatures.correlation <- cor(HP.numerical, use = "pairwise.complete.obs")

corrplot(numericalFeatures.correlation, tl.col="black", tl.pos = "lt")
```



Another step is to have a first look and choose the most important ones:

```
HP.categorical <- HP[sapply(HP, is.character)]

# Categorical feature first look
set.seed(1)
quick_RF <- randomForest(x=HP.categorical[1:1460,-76], y=HP.categorical$SalePrice[1:1460], ntree
imp_RF <- importance(quick_RF)
imp_DF <- data.frame(Variables = row.names(imp_RF), MSE = imp_RF[,1])
imp_DF <- imp_DF[order(imp_DF$MSE, decreasing = TRUE),]

# Features with most importance
ggplot(imp_DF[1:20,], aes(x=reorder(Variables, MSE), y=MSE, fill=MSE)) + geom_bar(stat = 'identi
```



From these two studies, we select the most important features and convert the categorical features to numeric(factor) to start modelling

```
# From these plots, we can select the numerical features with high correlation to the target var
selected_var <- c('LotFrontage','LotArea', 'OverallQual','YearBuilt','MasVnrArea',
                  'TotalBsmtSF','1stFlrSF','GrLivArea','FullBath','TotRmsAbvGrd',
                  'Fireplaces','GarageCars','GarageArea','Foundation','BsmtQual',
                  'ExterQual','KitchenQual','Exterior2nd','GarageType','Exterior1st',
                  'Neighborhood','HeatingQC','GarageFinish','BsmtFinType1','MSZoning',
                  'SalePrice')

HP <- HP[,selected_var]

HP[sapply(HP, is.character)] <- lapply(HP[sapply(HP, is.character)], as.factor)
HP[sapply(HP, is.factor)] <- lapply(HP[sapply(HP, is.factor)], as.numeric)
```

2.3 Modeling

```
# Validation set will be 10% of train data
set.seed(1, sample.kind = "Rounding")
validation_index <- createDataPartition(y = HP$SalePrice,
```

```
                                          times = 1,
                                          p = 0.1,
                                          list = F)

HP.validation <- HP[validation_index,]
HP.train <- HP[-validation_index,]

rm(validation_index)
```

Now we make some approaches to reach the best one to predict sales from the selected variables.

The first approach is a naive approach, predicting Sales Price as the mean of all houses in the dataset.

```
mean(HP$SalePrice) # Mean Sale Price is $180,921.2
```

```
## [1] 180921.2
```

```
sd(HP$SalePrice) # Standard Deviation is $79,442.5
```

```
## [1] 79442.5
```

```
mu <- mean(HP$SalePrice)

naive.rmse = RMSE(HP.validation$SalePrice, mu)
rmse.results <- tibble(Model = "01. Mean Sale Price Approach(Naive)",
                       RMSE = naive.rmse)
kable(rmse.results)
```

Model	RMSE
01. Mean Sale Price Approach(Naive)	65573.89

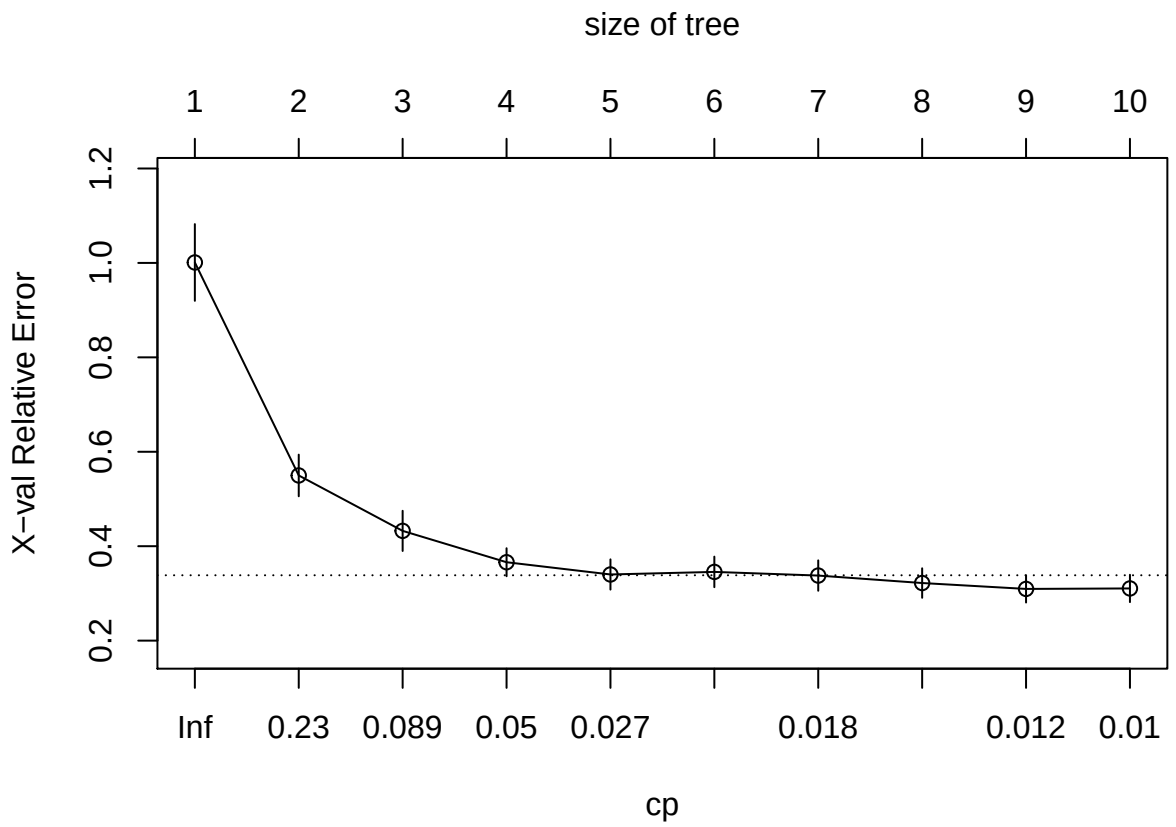
Second approach is a linear model fit.

```
# Linear Regression Fitting
linearRegression.fit <- lm(SalePrice ~ ., data = HP.train)

linearRegression.predict <- predict(linearRegression.fit, newdata = HP.validation)
linearRegression.rmse <- RMSE(linearRegression.predict, HP.validation$SalePrice)

rmse.results <- rmse.results %>%
  bind_rows(tibble(Model = "02. Linear Regression",
                   RMSE = linearRegression.rmse))
kable(rmse.results)
```

Model	RMSE
01. Mean Sale Price Approach(Naive)	65573.89
02. Linear Regression	26598.07

Fitting a decision tree model:

```
decisionTree.fit <- rpart(SalePrice~.,
                          data = HP.train,
                          control = rpart.control(cp = 0.01))

plotcp(decisionTree.fit)
```




```
decisionTree.predict <- predict(decisionTree.fit,
                                newdata = HP.validation)

decisionTree.rmse <- RMSE(decisionTree.predict, HP.validation$SalePrice)

rmse.results <- rmse.results %>%
  bind_rows(tibble(Model = "03. Decision Tree",
                   RMSE = decisionTree.rmse))
kable(rmse.results)
```

Model	RMSE
01. Mean Sale Price Approach(Naive)	65573.89
02. Linear Regression	26598.07
03. Decision Tree	34174.21

The last approach involves a Lasso regression:

```
set.seed(27)
lasso.fit <- train(x = HP.train[,-26], y = HP.train$SalePrice,
                   method ='glmnet',
                   trControl = trainControl(method="cv", number=5),
                   tuneGrid = expand.grid(alpha = 1, lambda = seq(0.001,0.1,by = 0.0005)))
lasso.fit$bestTune
```

```
##     alpha lambda
## 199     1    0.1
```

```
lasso.rmse <- min(lasso.fit$results$RMSE)

rmse.results <- rmse.results %>%
  bind_rows(tibble(Model = "04. Lasso Regression",
                   RMSE = lasso.rmse))
kable(rmse.results)
```

Model	RMSE
01. Mean Sale Price Approach(Naive)	65573.89
02. Linear Regression	26598.07
03. Decision Tree	34174.21
04. Lasso Regression	36449.98

3 Conclusion

The predictive model with the most accurate approach is the linear model.

Other approaches, involving more complex methods, or a combination of different models could reach better results, but the RMSE achieved is a good start point, and make the model really useful.

```
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
```

```
## Running under: Windows 10 x64 (build 19041)
##
## Matrix products: default
##
## Random number generation:
##  RNG:     Mersenne-Twister
##  Normal:  Inversion
##  Sample:  Rounding
##
## locale:
## [1] LC_COLLATE=Portuguese_Brazil.1252  LC_CTYPE=Portuguese_Brazil.1252
## [3] LC_MONETARY=Portuguese_Brazil.1252 LC_NUMERIC=C
## [5] LC_TIME=Portuguese_Brazil.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] randomForest_4.6-14 corrplot_0.84       leaps_3.1
##  [4] rpart_4.1-15        knitr_1.29          data.table_1.12.8
##  [7] caret_6.0-86        lattice_0.20-41     forcats_0.5.0
## [10] stringr_1.4.0       dplyr_1.0.0         purrr_0.3.4
## [13] readr_1.3.1         tidyr_1.1.0         tibble_3.0.1
## [16] ggplot2_3.3.2       tidyverse_1.3.0
##
## loaded via a namespace (and not attached):
##  [1] httr_1.4.2           jsonlite_1.7.0       splines_4.0.2
##  [4] foreach_1.5.0        prodlim_2019.11.13   modelr_0.1.8
##  [7] assertthat_0.2.1     highr_0.8            stats4_4.0.2
## [10] blob_1.2.1           cellranger_1.1.0     yaml_2.2.1
## [13] ipred_0.9-9          pillar_1.4.6         backports_1.1.7
## [16] glue_1.4.1           pROC_1.16.2          digest_0.6.25
## [19] rvest_0.3.6          colorspace_1.4-1     recipes_0.1.13
## [22] htmltools_0.5.0      Matrix_1.2-18        plyr_1.8.6
## [25] timeDate_3043.102    pkgconfig_2.0.3      broom_0.7.0
## [28] haven_2.3.1          scales_1.1.1         gower_0.2.2
## [31] lava_1.6.7           farver_2.0.3         generics_0.0.2
## [34] ellipsis_0.3.1       withr_2.2.0          nnet_7.3-14
## [37] cli_2.0.2            survival_3.1-12      magrittr_1.5
## [40] crayon_1.3.4         readxl_1.3.1         evaluate_0.14
## [43] fs_1.4.1             fansi_0.4.1          nlme_3.1-148
## [46] MASS_7.3-51.6        xml2_1.3.2           class_7.3-17
## [49] tools_4.0.2          hms_0.5.3            lifecycle_0.2.0
## [52] glmnet_4.0-2         munsell_0.5.0        reprex_0.3.0
## [55] compiler_4.0.2       rlang_0.4.6          grid_4.0.2
## [58] iterators_1.0.12     rstudioapi_0.11      labeling_0.3
## [61] rmarkdown_2.3        gtable_0.3.0         ModelMetrics_1.2.2.2
## [64] codetools_0.2-16     DBI_1.1.0            reshape2_1.4.4
```

```
## [67] R6_2.4.1             lubridate_1.7.9      shape_1.4.4
## [70] stringi_1.4.6        Rcpp_1.0.4.6         vctrs_0.3.1
## [73] dbplyr_1.4.4         tidyselect_1.1.0     xfun_0.16
```