DSCI401 Project Proposal

Nisi Mohan Kuniyil

11/11/2021

Research Question

Exploring the effect of different features on house prices using multiple machine learning algorithms, and analyzing how accurate a model can predict the sales price of houses in Illinois metropolitan area.

The real estate market is a highly variable market where property prices are driven by a lot of factors. This project will analyze the Illinois real estate data and identify meaningful features from a large set of potential factors that has a major impact on property prices. Multiple machine learning algorithms will be applied to the dataset to compare and contrast the importance of different features in predicting the market prices of properties.

Since the target label that needs to be predicted is a numeric value, we will look at different kinds of regression models. We will start with a decision tree regressor model which will not only predict the housing price value from a set of features but will also provide a tree of explanation the model used to predict the price. This will make it easier to interpret the model's prediction and what features it used primarily to arrive at the predicted value. We will then explore a more powerful machine learning model called random forest, which uses a collection of decision trees to predict the property price. The random forest algorithm should produce a much more accurate model and will give us insights into the importance of different features that affects property prices. Lastly, we will also explore the Gradient Boosting algorithm to check if we can get an even better predictive model than the random forest model.

Project Plan

1. HousesSoldData.csv

There are 82 variables and 2930 observations in this dataset.

```
## 'data.frame':    2930 obs. of  82 variables:
##  $ Order          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ PID            : int  526301100 526350040 526351010 526353030 527105010 527105030 527127150 527145
##  $ MS.SubClass    : int  20 20 20 20 60 60 120 120 120 60 ...
##  $ MS.Zoning      : chr  "RL" "RH" "RL" "RL" ...
##  $ Lot.Frontage   : int  141 80 81 93 74 78 41 43 39 60 ...
##  $ Lot.Area       : int  31770 11622 14267 11160 13830 9978 4920 5005 5389 7500 ...
##  $ Street         : chr  "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley          : chr  NA NA NA NA ...
##  $ Lot.Shape      : chr  "IR1" "Reg" "IR1" "Reg" ...
##  $ Land.Contour   : chr  "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities      : chr  "AllPub" "AllPub" "AllPub" "AllPub" ...
##  $ Lot.Config     : chr  "Corner" "Inside" "Corner" "Corner" ...
##  $ Land.Slope     : chr  "Gtl" "Gtl" "Gtl" "Gtl" ...
##  $ Neighborhood   : chr  "NAmes" "NAmes" "NAmes" "NAmes" ...
##  $ Condition.1    : chr  "Norm" "Feedr" "Norm" "Norm" ...
##  $ Condition.2    : chr  "Norm" "Norm" "Norm" "Norm" ...
##  $ Bldg.Type      : chr  "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ House.Style    : chr  "1Story" "1Story" "1Story" "1Story" ...
##  $ Overall.Qual   : int  6 5 6 7 5 6 8 8 8 7 ...
##  $ Overall.Cond   : int  5 6 6 5 5 6 5 5 5 5 ...
##  $ Year.Built     : int  1960 1961 1958 1968 1997 1998 2001 1992 1995 1999 ...
##  $ Year.Remod.Add : int  1960 1961 1958 1968 1998 1998 2001 1992 1996 1999 ...
##  $ Roof.Style     : chr  "Hip" "Gable" "Hip" "Hip" ...
##  $ Roof.Matl      : chr  "CompShg" "CompShg" "CompShg" "CompShg" ...
##  $ Exterior.1st   : chr  "BrkFace" "VinylSd" "Wd Sdng" "BrkFace" ...
##  $ Exterior.2nd   : chr  "Plywood" "VinylSd" "Wd Sdng" "BrkFace" ...
##  $ Mas.Vnr.Type   : chr  "Stone" "None" "BrkFace" "None" ...
##  $ Mas.Vnr.Area   : int  112 0 108 0 0 20 0 0 0 0 ...
##  $ Exter.Qual     : chr  "TA" "TA" "TA" "Gd" ...
##  $ Exter.Cond     : chr  "TA" "TA" "TA" "TA" ...
##  $ Foundation     : chr  "CBlock" "CBlock" "CBlock" "CBlock" ...
##  $ Bsmt.Qual      : chr  "TA" "TA" "TA" "TA" ...
##  $ Bsmt.Cond      : chr  "Gd" "TA" "TA" "TA" ...
##  $ Bsmt.Exposure  : chr  "Gd" "No" "No" "No" ...
##  $ BsmtFin.Type.1 : chr  "BLQ" "Rec" "ALQ" "ALQ" ...
##  $ BsmtFin.SF.1   : int  639 468 923 1065 791 602 616 263 1180 0 ...
##  $ BsmtFin.Type.2 : chr  "Unf" "LwQ" "Unf" "Unf" ...
##  $ BsmtFin.SF.2   : int  0 144 0 0 0 0 0 0 0 0 ...
##  $ Bsmt.Unf.SF    : int  441 270 406 1045 137 324 722 1017 415 994 ...
##  $ Total.Bsmt.SF  : int  1080 882 1329 2110 928 926 1338 1280 1595 994 ...
##  $ Heating        : chr  "GasA" "GasA" "GasA" "GasA" ...
##  $ Heating.QC     : chr  "Fa" "TA" "TA" "Ex" ...
##  $ Central.Air    : chr  "Y" "Y" "Y" "Y" ...
##  $ Electrical     : chr  "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
##  $ X1st.Flr.SF    : int  1602 898 1315 2125 855 885 1387 1334 1807 1138 ...
##  $ X2nd.Flr.SF    : int  0 0 0 0 701 678 0 0 0 776 ...
##  $ Low.Qual.Fin.SF: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Gr.Liv.Area    : int  1788 888 1327 2061 1611 1553 1165 1320 1792 1842 ...
##  $ Bsmt.Full.Bath : int  1 0 0 1 0 0 1 0 1 0 ...
##  $ Bsmt.Half.Bath : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Full.Bath      : int  1 1 1 2 2 2 2 2 2 2 ...
##  $ Half.Bath      : int  0 0 1 1 1 1 0 0 0 1 ...
##  $ Bedroom.AbvGr  : int  3 2 3 3 3 3 2 2 2 3 ...
##  $ Kitchen.AbvGr  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Kitchen.Qual   : chr  "TA" "TA" "Gd" "Ex" ...
##  $ TotRms.AbvGrd  : int  7 5 6 8 6 7 6 5 5 7 ...
##  $ Functional     : chr  "Typ" "Typ" "Typ" "Typ" ...
##  $ Fireplaces     : int  2 0 0 2 1 1 0 0 1 1 ...
##  $ Fireplace.Qu   : chr  "Gd" NA NA "TA" ...
##  $ Garage.Type    : chr  "Attchd" "Attchd" "Attchd" "Attchd" ...
##  $ Garage.Yr.Blt  : int  1960 1961 1958 1968 1997 1998 2001 1992 1995 1999 ...
##  $ Garage.Finish  : chr  "Fin" "Unf" "Unf" "Fin" ...
##  $ Garage.Cars    : int  2 1 1 2 2 2 2 2 2 2 ...
##  $ Garage.Area    : int  528 730 312 522 482 470 582 506 608 442 ...
##  $ Garage.Qual    : chr  "TA" "TA" "TA" "TA" ...
##  $ Garage.Cond    : chr  "TA" "TA" "TA" "TA" ...
##  $ Paved.Drive    : chr  "P" "Y" "Y" "Y" ...
##  $ Wood.Deck.SF   : int  210 140 393 0 212 360 0 0 237 140 ...
##  $ Open.Porch.SF  : int  62 0 36 0 34 36 0 82 152 60 ...
##  $ Enclosed.Porch : int  0 0 0 0 0 0 170 0 0 0 ...
##  $ X3Ssn.Porch    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Screen.Porch   : int  0 120 0 0 0 0 0 144 0 0 ...
##  $ Pool.Area      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Pool.QC        : chr  NA NA NA NA ...
##  $ Fence          : chr  NA "MnPrv" NA NA ...
##  $ Misc.Feature   : chr  NA NA "Gar2" NA ...
##  $ Misc.Val       : int  0 0 12500 0 0 0 0 0 0 0 ...
##  $ Mo.Sold        : int  5 6 6 4 3 6 4 1 3 6 ...
##  $ Yr.Sold        : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ Sale.Type      : chr  "WD " "WD " "WD " "WD " ...
##  $ Sale.Condition : chr  "Normal" "Normal" "Normal" "Normal" ...
##  $ SalePrice      : int  209200 107700 163600 257900 184400 208700 213700 193600 241700 191200 ...
```

2. Data Tidying.

3. Visualization for exploratory data analysis.

4. Building and Training a Decision Tree Model

5. Building and Training a Random Forest Model.

6. Building and Training a Gradient Boosting Model.

References

https://medium.com/analytics-vidhya/kaggle-house-prices-prediction-with-linear-regression-and-gradient-boosting-c5694d9c6df4

https://towardsdatascience.com/house-prices-prediction-using-deep-learning-dea265cc3154