HOUSE PRICES

ADVANCED REGRESSION TECHNIQUES

Ararat Darren Ferahyan, 303498772

Shardul Dave, 306598622

Wesley Hwang, 301320557

California State University, Los Angeles

A Report on-

House Prices: Advanced Regression Techniques

Under course-

CS4661: Introduction to Data Science

Submitted To-

Dr. Mohammad Pourhomayoun

Contents

1. Project Description and Details

In this project we are trying to predict house prices depending on different variables using regression techniques. The main purpose of it is that the buyers can get an idea of the house prices in which they like or are trying to purchase.

2. Project Goals

The project goal is to get better pricing of houses with our developed methods and algorithms. Also, to try and get the accuracy of our developed methods to be as near to the data as possible.

3. Data Details

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

Data Sets: test.csv (1459 X 80)
train.csv (1460 X 81)

Note: The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

4. Developed Methods, Algorithms and Tools

We used methods from the libraries namely scikit learn, pandas, numpy.

We used pandas read_csv() method to read the train.csv and test.csv data files.

We used scikit's preprocessing.scale() method to normalize the data.

We used LogisticRegression() 's fit method to train the data, its predict method to predict the data and accuracy_score() method to get the accuracy.

We used DecisionTreeClassifier() 's fit method to train the data, its predict method to predict the data and accuracy_score() method to get the accuracy.

We used KNeighborsClassifier() with k=3 ‘s fit method to train the data, its predict method to predict the data and accuracy_score() method to get the accuracy.

We used RandomForestRegressor with max_depth=2, random_state=0, n_estimators=100 ‘s fit method to train the data, its predict method to predict the data and explained_variance_score() method to get the accuracy.

Also we created methods such as

1. createValues()- to handle NaN values.
2. createSaleCondition_Normal(a)-to convert categorical SaleCondition_Normal to numerical
3. createSaleCondition_Abnorml(a)- to convert categorical SaleCondition_Abnormal to numerical
4. createSaleCondition_Partial(a)- to convert categorical SaleCondition_Partial to numerical
5. createSaleCondition_AdjLand(a)- to convert categorical SaleCondition_AdjLand to numerical
6. createSaleCondition_Family(a)- to convert categorical SaleConditio_Family to numerical
7. createSaleCondition_Aloca(a)-to convert categorical SaleCondition_Aloc to numerical

5. Results

Accuracy using Logistic Regression is 0.005479452054794521

Accuracy using Decision Tree is 0.0136986301369863

Accuracy using KNN Classifier is 0.00821917808219178

Accuracy using Random Forest Regression is 0.40399259640650076

6. Performance Analysis

From these results, we came to the conclusion that Random Forest Regression is the best for our data and Logistic Regression is the worst.

7.Responsibility of Team Members

We split up the work equally so that everyone could get a chance to work on algorithms, functions, project paper, and the power point slides. This made the project go smoother.