FINAL REPORT

CENG 3521, DATA MINING

Edanur Pişkin - Yiming Aijiaerguli
edanurpiskin@posta.mu.edu.tr - 170709509@posta.mu.edu.tr

Friday 22nd January, 2021

Abstract

House prices were estimated in this study with AmesHouse dataset. Data had many missing values. These missing values were filled by various methods. Then, the data set that gave us the best result was prepared. Different models were used, namely linear regression and MLP. Results were compared according to the RMSE and R-squared values. The results obtained at the end of the study were satisfying.

1 Introduction

In this project, it is aimed to predict house prices by using house features. Prediction performance has been tried to be increased by using different methods such as feature selection, outlier detection and scaling. The missing data were filled by examining the categorical and numerical variables in detail. Various visualization methods have been used to select the features which will give best results. Features were examined in detail and their effects on the price were observed. Linear regression is used as the estimation algorithm. As the last step, all results are compared with each other. Algorithm and data that give the best result are determined.

2 Project Details

2.1 Data

Ames House Dataset was selected for the house price estimation project. The data includes 2930 different houses and 82 features in total. The columns are as follows.

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

2.2 Handling Missing Values

27 columns have missing values. Details of missing value are shown in Figure 1. Since some of the features have many missing values, deleting missing values will not be the right method. Different methods are used for categorical and numerical missing values. For categorical data, missing values show that the house does not have that feature. Therefore, the missing values are filled in as "does not have feature". Two ways can be used to fill in numerical data. Missing values can be filled by mean and median. However, it is necessary to decide which method should be used for each feature. To decide which method will be used, it is necessary to examine the distribution of columns with missing values. Two ways can be used to fill in numerical data. Missing values can be filled by mean and median. However, it is necessary to decide which method should be used for each feature. To decide which method will be used, distribution of columns with missing values needs to be examined. See Figure 3. Distributions which are similar to the normal distribution can be filled with mean value. But if the distribution is skewed, it should be filled with median value. When the plot is examined, it is seen that the LotFrontage feature is similar to the normal distribution, while the MasVnrArea and GarageYtBlt features are skewed. The missing values in LotFrontage will be imputed by the mean value and missing values in MasVnrArea and GarageYrBlt features will be imputed by the median value.

2.3 Binarization

Categorical data should be binarized to use data in regression analysis. After binarization, the data increased from 82 columns to 275 columns. One hot encoding is used.

Missing data counts of categorical attributes

	Columns	MissingValues	Percent
0	PoolQC	2917	99.556
1	MiscFeature	2824	96.382
2	Alley	2732	93.242
3	Fence	2358	80.478
4	FireplaceQu	1422	48.532
5	LotFrontage	490	16.724
6	GarageCond	159	5.427
7	GarageFinish	159	5.427
8	GarageYrBlt	159	5.427
9	GarageQual	159	5.427
10	GarageType	157	5.358
11	BsmtExposure	83	2.833
12	BsmtFinType2	81	2.765
13	BsmtQual	80	2.730
14	BsmtCond	80	2.730
15	BsmtFinType1	80	2.730
16	MasVnrArea	23	0.785
17	MasVnrType	23	0.785
18	BsmtHalfBath	2	0.068
19	BsmtFullBath	2	0.068
20	BsmtFinSF1	1	0.034
21	GarageCars	1	0.034
22	Electrical	1	0.034
23	TotalBsmtSF	1	0.034
24	BsmtUnfSF	1	0.034
25	BsmtFinSF2	1	0.034
26	GarageArea	1	0.034

Figure 1: Missing Values

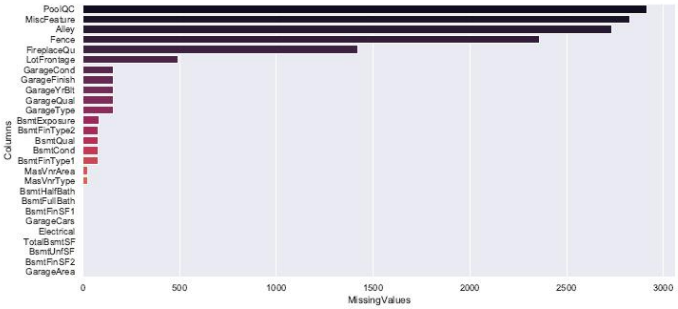

Figure 2: Missing Values Plot

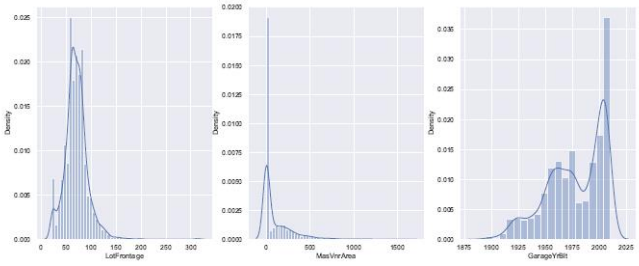

Figure 3: Distribution Of Attributes With Missing Values

2.4 First Regression Analysis

First of all, regression analysis was applied without any improvement of dataset. This result is taken as the base result. The results of other regressions will be compared with this base result. Before applying regression, it is necessary to fit the attribute which will be predicted to the normal distribution. For this, the log of the sales price is used. The graphs of the distributions are also shown below. See Figure 4. Then regression analysis is applied. R-squared and RMSE are used as performance measurement metrics. R-squared is a statistical measure of how close the data are to the fitted regression line. Root mean squared error (RMSE) is the square root of the mean of the square of all of the error. The results are shown below.
R-squared : 0.82
RMSE: 0.17
Results are not bad but can be improved. The plot in Figure 5 shows the scatter plot of actual values and predicted values.

Figure 4: SalePrice – Log Transformed SalePrice

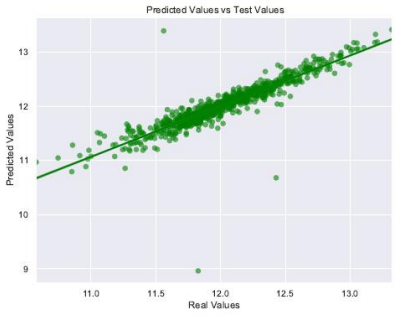

Figure 5: Non-featured Regression Plot

2.5 Outliers

Boxplot of sale price attribute are examined to detect outliers. Values outside 1.5 IQR value are assumed to be outlier. See Figure 6. Outlier values were filtered and regression was applied again. Results are given below.
R-squared: 0.77
RMSE: 0.19
R-squared value decreased and RMSE value is increased. These results are worse than our base results. For these reasons, it is decided to keep the outliers.

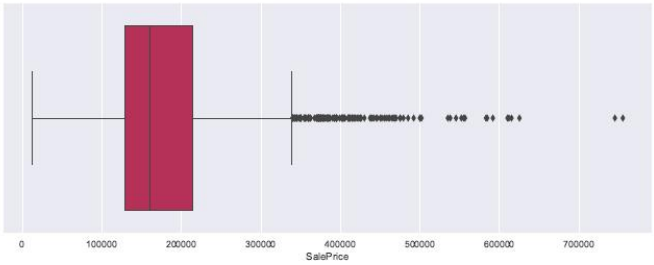

Figure 6: Box plot of SalePrice

2.6 Feature Selection

In order to determine the most important numerical features for prediction, the correlation between the features with each other and with the sales price was examined. See Figure 7. Boxplots of some of the most important features are shown in Figure 11. Boxplot was used to determine the most important square features. When the plot is examined, it is seen that some features are more correlated with the sales price than others. First 10 numeric properties with the highest correlation were used. Plot shown in Figure 9 shows the correlation between the selected most important features. When the results were examined, it can be seen that the overall quality, living area square feet, garage area, basement square feet, first floor square feet, built year, bathroom count, remodel date and garage built year were the most effective features on the price of the house. Boxplots were examined to select the most important categorical features. Those with significant differences between values identified as the most important categorical

features.
Regression was applied again with these selected data. Results is shown below and graph is shown in Figure 10.
R-squared : 0.92
RMSE: 0.12
When the results were examined, it was seen that a better result was obtained compared to the result selected as the base. It has been decided to continue with only important features.

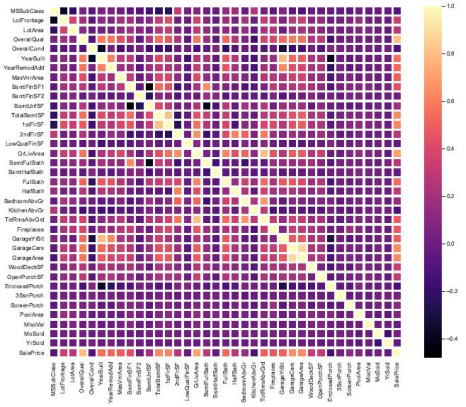

Figure 7: Correlations

	Attributes	Correlation
1	OverallQual	0.799
2	GrLivArea	0.707
3	GarageCars	0.648
4	GarageArea	0.640
5	TotalBsmtSF	0.633
6	1stFlrSF	0.622
7	YearBuilt	0.558
8	FullBath	0.546
9	YearRemodAdd	0.533
10	GarageYrBlt	0.509
11	MasVnrArea	0.502
12	TotRmsAbvGrd	0.495
13	Fireplaces	0.475
14	BsmtFinSF1	0.433
15	LotFrontage	0.341

Figure 8: Correlation

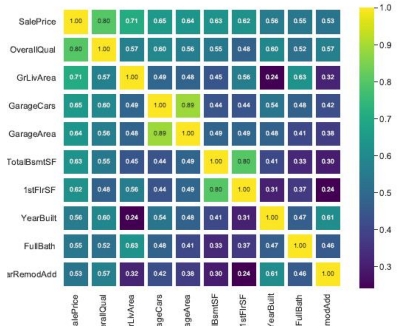

Figure 9: Correlation Between Most Important Features

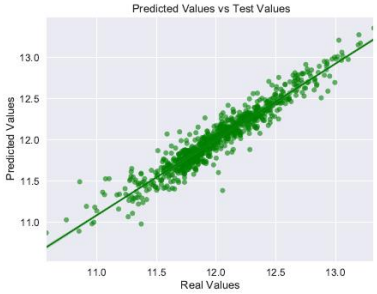

Figure 10: Feature Selection Regression Plot

2.7 Scaling

Standard scaler is used to scale the data. Then regression was applied with scaled data. Because of randomness result may differ. The results are shown below.
R-squared : -227
RMSE: 6.04
After feature selecting, most of the data are categorical. For this reason, scaling did not have a positive effect on regression results. R-squared values is decreased and RMSE is increased.

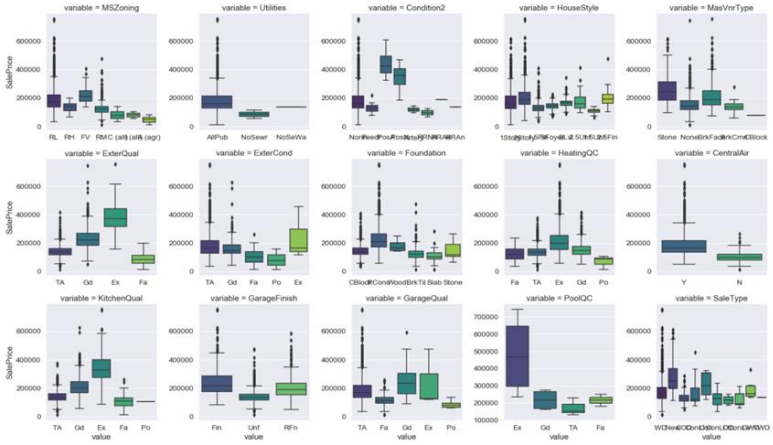

Figure 11: Box Plots of Most Important Categorical Attributes

2.8 MLP Regression

Finally, MLP regression is applied. It has been tested whether it will give better results than linear regression. The result is stated below.
R-squared : -0.59
RMSE: 0.5
The results obtained are worse than linear regression.

3 Results

4 different prediction models have been established. In the first model, all features were used without making any change in the data. Then outliers were determined and removed from the data. Since our data has many different features, the important ones among these features were selected and the regression model was established. Finally, MLP regression model is established. The R-squared and RMSE values of these models were calculated. The results obtained is shown below. When the results are examined, it is seen that the best result is linear regression. Outlier detection did not affect performance positively. The data that gives the best results is the data filtered with feature selection methods.

	R-squared	RMSE
NonFeatured	0.822	0.169
Without Outliers	0.768	0.193
Feature Selection	0.916	0.116
MLP Regression	-0.585	0.503

Figure 12: Result

4 Conclusion

Many data preparation methods and different prediction models are applied. When the results were examined, it was decided to use linear regression. The data set that gives the best result is the data prepared with the feature selection method. The features that affect house prices the most were determined. When the results are examined, it is seen that model can explain the variation of the data by 92%.