08.03.2019

Supervised Learning Capstone

Predicting the Sale Price of a House

Leeor Nehardea

Thinkful Data Science Bootcamp

Overview

Over the course of about a year, I got more and more interested in real estate, especially of houses. As a future data scientist, I perceive the capstone as a great opportunity to explore this area a little further, and use the tools that I've acquired to predict the sale price of a house.

Research question

Can the price of a house be predicted using its "dry" data?

Steps

Exploratory data analysis - EDA

- Understanding the data
- Data cleaning

Data Exploration

- Using plots and analytics to see correlations, trends, and behavior of the data

Feature Engineering

- Changing and removing features to get

  1.better understanding, and less dependency between the features themselves

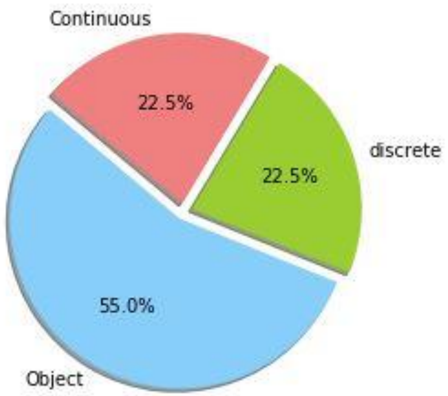
  2. better explanation of the target variable.

About The Data

★ **Data source:**
https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

★ **Details:**

- Location: Ames, Iowa, USA
- What: sale information of individual residential properties
- Years: 2006 to 2010
- Target variable: The sale price of a house
- Train dataset shape: 1460 rows, 80 columns
- Test dataset shape: 1459 rows, 80 columns

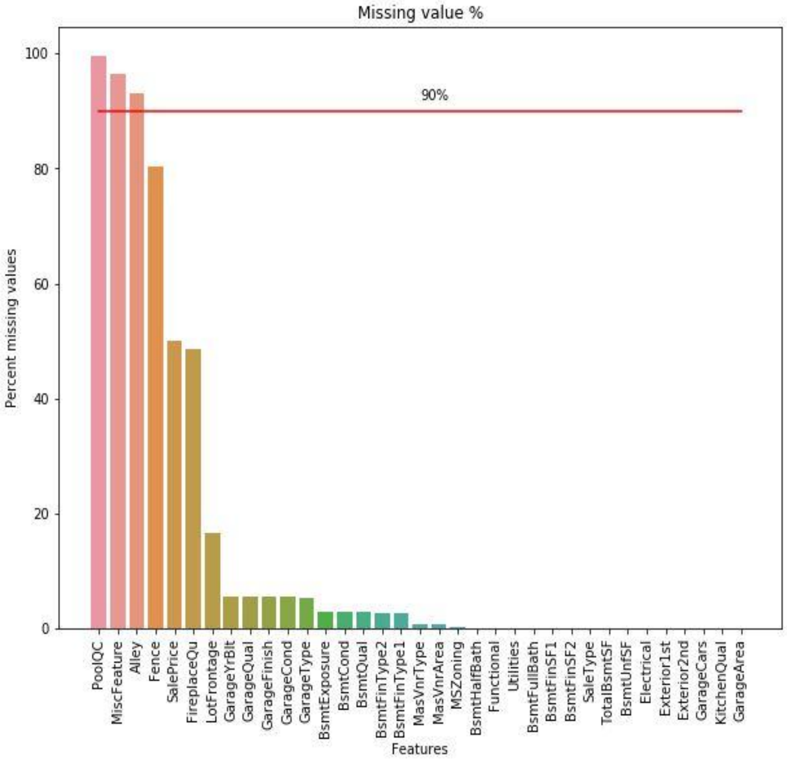


Exploratory Data Analysis

Exploratory Data Analysis

★ Business Decisions:

- Columns with 90% or more missing values will be dropped
- Trying to avoid dropping rows
- Missing values in object features will be replaced by the most basic/ not existing value.
- Missing values in numeric columns will be checked and compare to similar or same category features, and be replaced with the value that makes sense to that feature.
- Discrete numeric features will be changed to object features
- Discrete is defined as a feature with < 20 unique values; continues > 40 unique values



★ Main Issues

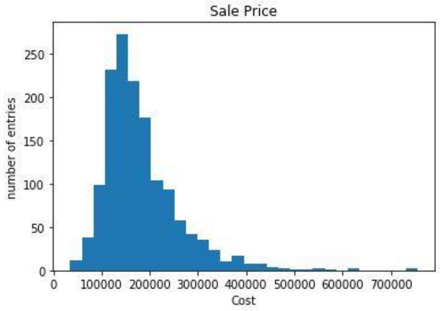
- Many categorical features with several values will have to be transformed to dummies
- Plenty of outliers at each continue feature
- Zero is given as an indicator for ‘does not exist’ for numeric features, causing clustering
- Skewed target variable

★ Handling

- Drop object columns that add no value
- Remove extreme outliers
- When making sense, replace with different values or remove the feature
- Take transformation of the target variable

Data Exploration

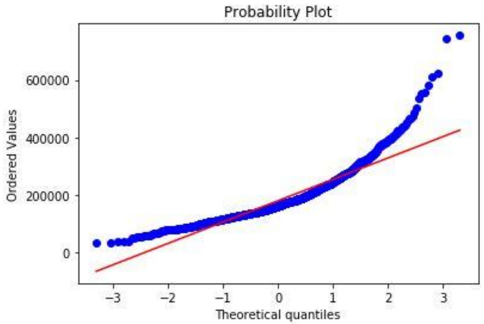
Target variable



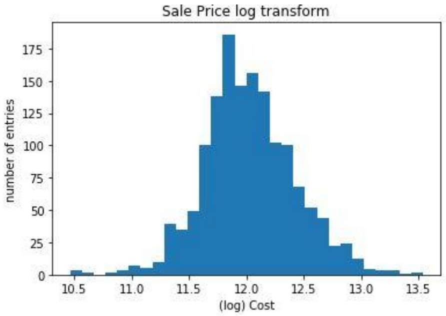
Fun fact

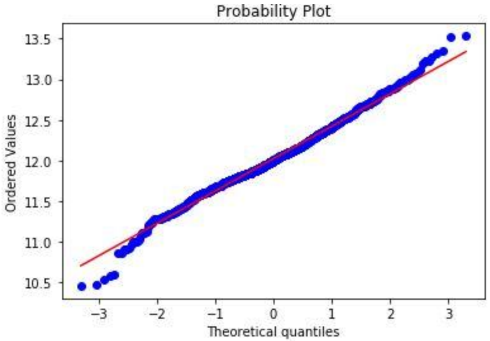
Iowa median income is $58,570 a year

According to Google



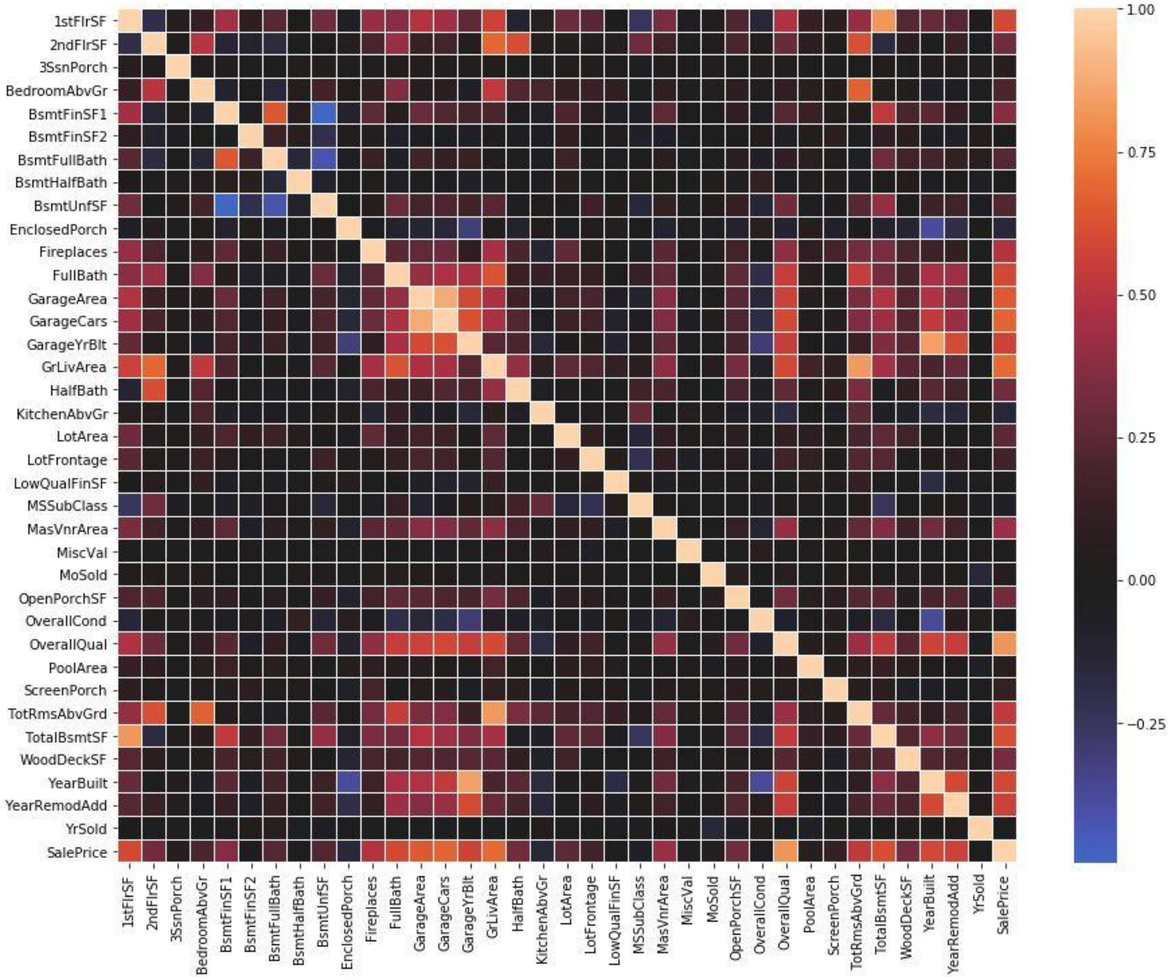
Log transform



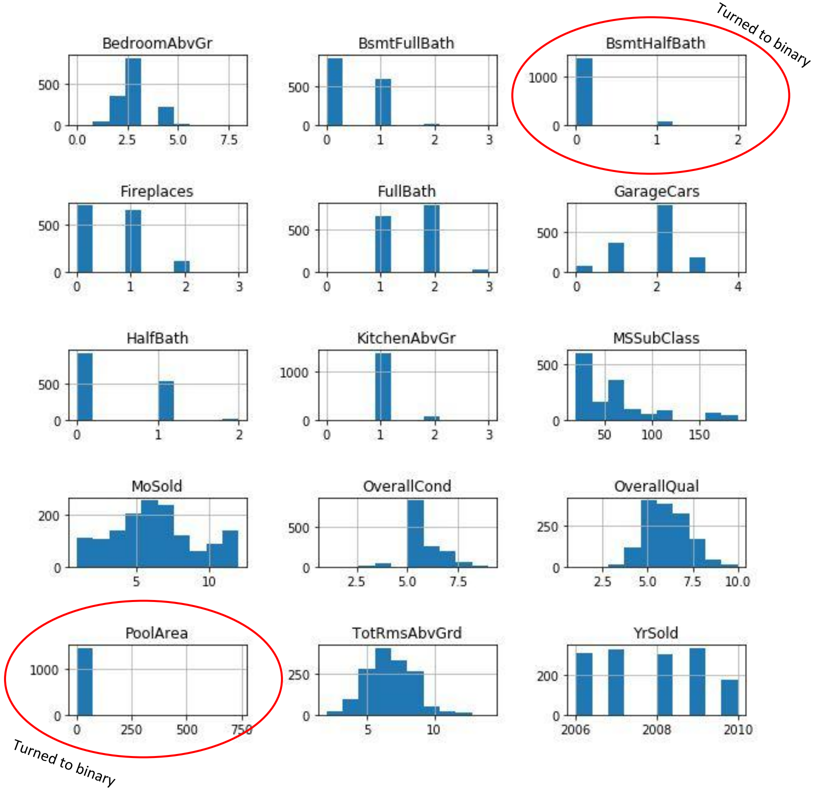


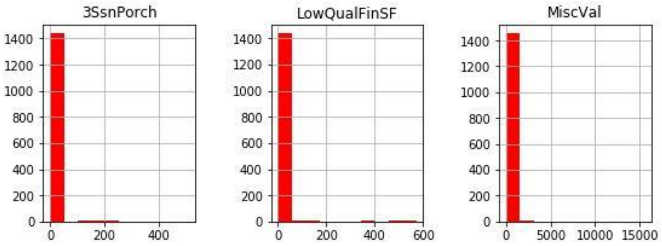
Correlation between all the numeric features and the target variable.

The lighter color the more correlation

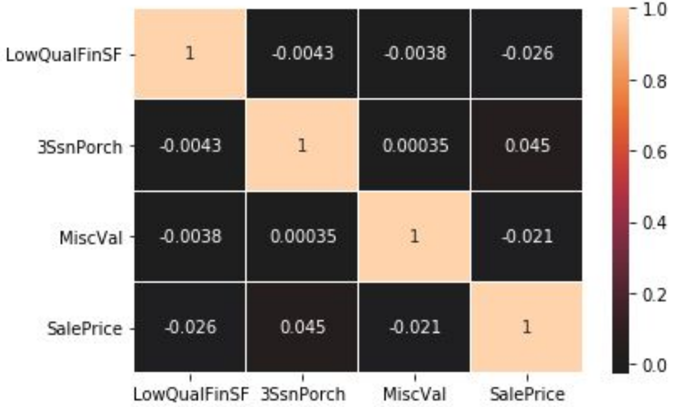


Discrete variables histogram plot
(2 - 20 unique features)



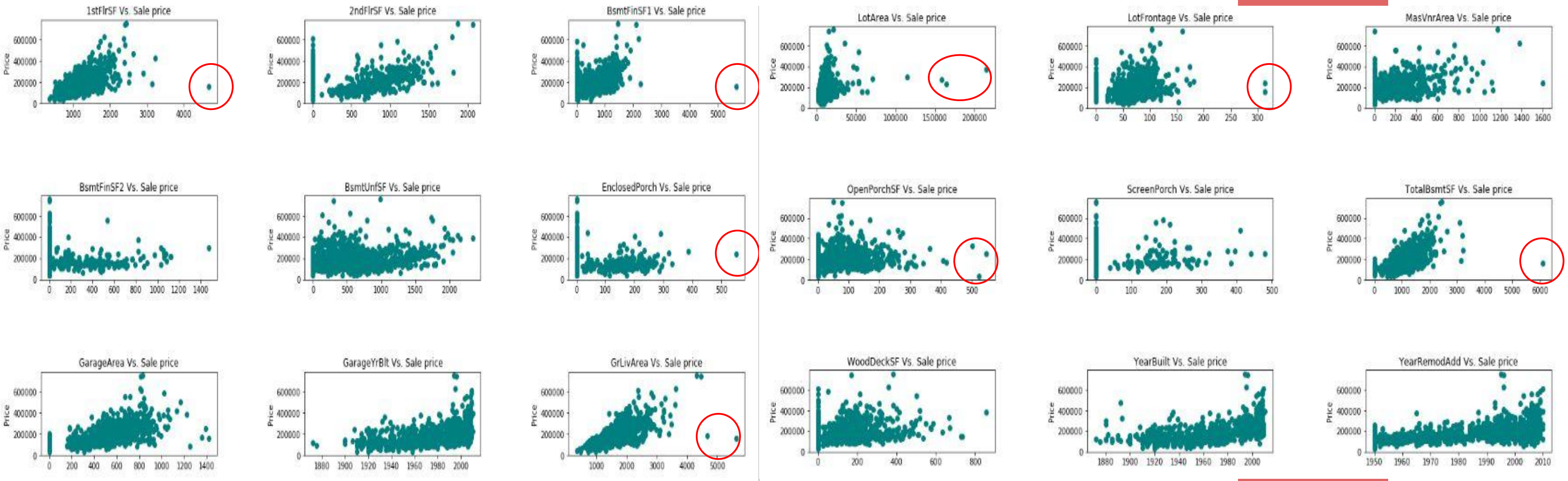


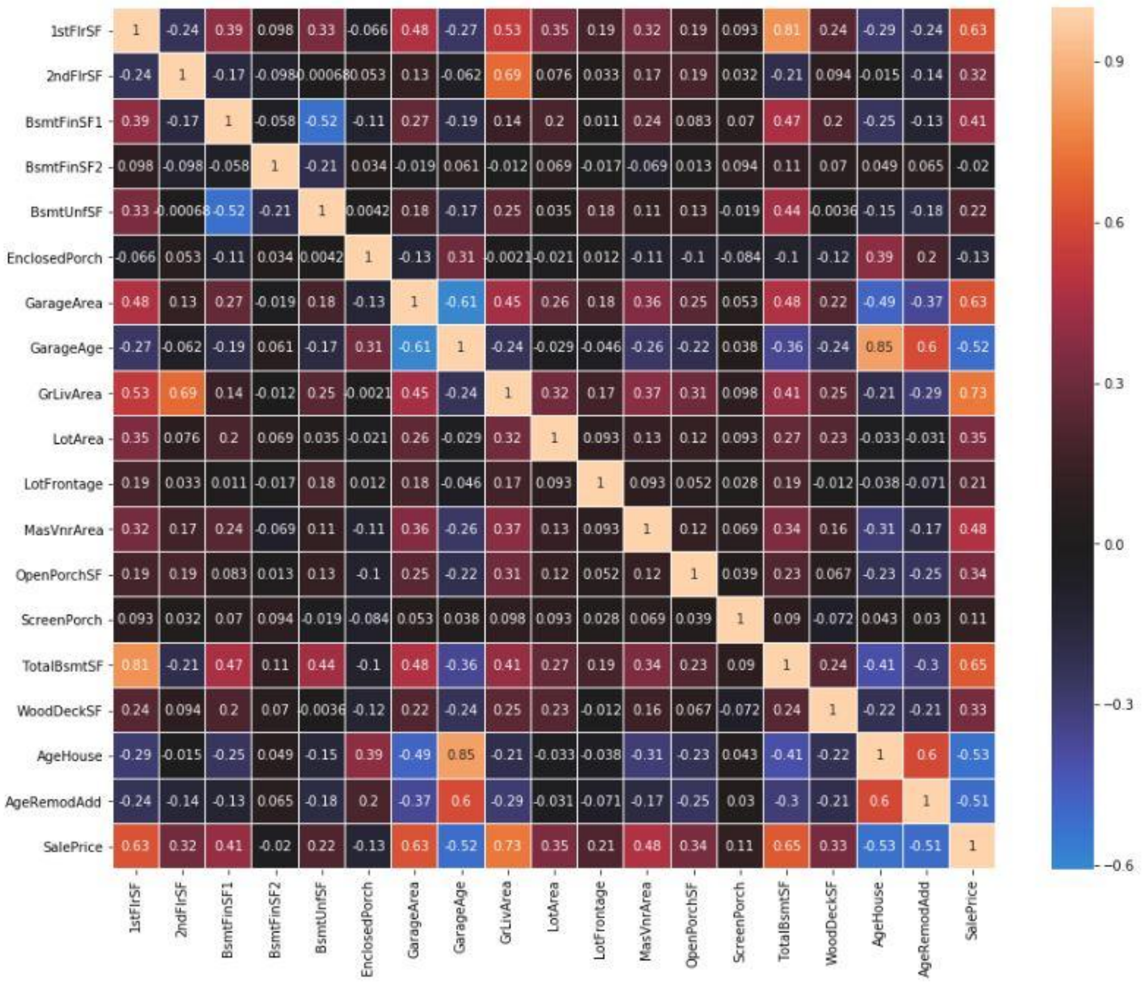
Features with 20 - 40 unique values
And their correlation to target variable



Numeric features
VS.
Sale Price

Outliers removal





After

- Removing extreme outliers
- Changing years to age
- Replacing discrete numeric features to objects
- Removing several columns

This heatmap is the result

Notice the year features; they turned to negative which makes sense. Older houses are expected to be less expensive than newer ones

Before training the model

make object features to dummies.

The data frame has 1450 rows and 347 features

It started with 1460 and 80

Models to be used

Ordinary least squares

Ridge

regression

Lasso

Random forest

R-squared Results

Ordinary Least Squares Statsmodels.regression model	R-squared: 0.935
Lasso Regression sklearn.linear_model.Lasso	Mean R-squared: 0.894 STD: 0.018
Ridge Regression sklearn.linear_model.Ridge	Mean R-squared: 0.904 STD: 0.015
Random Forest sklearn.ensemble.RandomForestRegressor	Mean R-squared: 0.882 STD: 0.012

Dimensionality reduction

Method 1:
Remove features with a p-value higher than certain value

Largest 20 P-values

Neighborhood_Blmngtn	0.995313
Exterior2nd_MetalSd	0.990442
BldgType_Twnhs	0.987516
Exterior1st_VinylSd	0.983987
LotShape_Reg	0.982954
MoSold_10	0.982687
Electrical_FuseF	0.981981
Exterior1st_Wd Sdng	0.981870
MSSubClass_45	0.972396
Electrical_FuseA	0.958610
SaleCondition_Normal	0.956355
ExterCond_Gd	0.955065
SaleType_Oth	0.953010
OverallQual_2	0.942749
BsmtFinType2_BLQ	0.939711
BsmtFinType1_GLQ	0.932458
Exterior1st_Stone	0.930592
BsmtFinType2_Rec	0.926463
MasVnrType_BrkFace	0.918655
Functional_Maj2	0.915529

Removing features with a p-value > 0.1 resulted in dimensions of (1450,96)

R-squared

Lasso
Avg: 0.805
Std: 0.045

Ridge
Avg: 0.809
Std: 0.045

Random forest
Avg: 0.822
Std: 0.025

OLS:
0.837

Method 2:
Usa PCA with n-components

100 components;
shape is (1450,100)

Lasso
Avg: 0.858
Std: 0.016

Ridge
Avg: 0.858
Std: 0.016

Random forest
Avg: 0.793
Std: 0.050

OLS
0.878

Conclusions and Future Iterations

Conclusions:

Overpaying on a property has an enormous effect when thinking of real estate. It could make a profitable property to an unprofitable one. Additionally, looking at every aspect of a house and compare it to others is not feasible for a human.

In the slides above, I have shown that leveraging machine learning technique could reduce the uncertainty concerning a value of a house. The models could help a business or an individual to make smarter decisions when evaluating the price of a house and therefore - its profitability.

I'm confident that using the above steps could make a property analysis be done faster, safer, and more comfortably.

Future actions:

In order to improve on my results I believe that these steps can be taken.

- Using boosting techniques.
- Looking closer into outliers and ways to handle them
- Getting more data to get more flexibility with data cleaning
- Having more "real estate oriented" mind and understanding

Questions?
Notes?
Concerns?