PREDICTING HOUSE PRICES

Kalmer Keerup, Mehis Taevere, Arlo Tammekun

Introduction

Due to legislative reasons the real estate business is dominated by realtors. A number of them have created their own schemas for determining the sales prices for different houses, which they are keen on protecting, as this enables them to estimate the optimal market prices very operatively. This project aims to find out the most outstanding factors of house sales price within the used data and create a model to predict the possible sales prices based on the factors given.

Data

The Ames Housing Dataset was compiled by Dean De Cock for data science education. It is a part of an on-going Kaggle competition. The dataset is based on sales of individual residential property in Ames, Iowa, USA from 2006 to 2010. The dataset contains 3970 rows of data and 80 features that are directly related to property sales. The data was very well-formed so there was little need to normalize any data.

Methods

Due to the nature of the prediction task at hand, we needed to use regression models. For this task, we chose to use random forest with the mean square error criterion and gradient boosting with the least squares loss method. In our cross-validation optimizing, both methods reached same levels of accuracy, around 80 to 85 percent, gradient boosting and random forest respectively.

Fig. 1

Feature Importance (MDI)

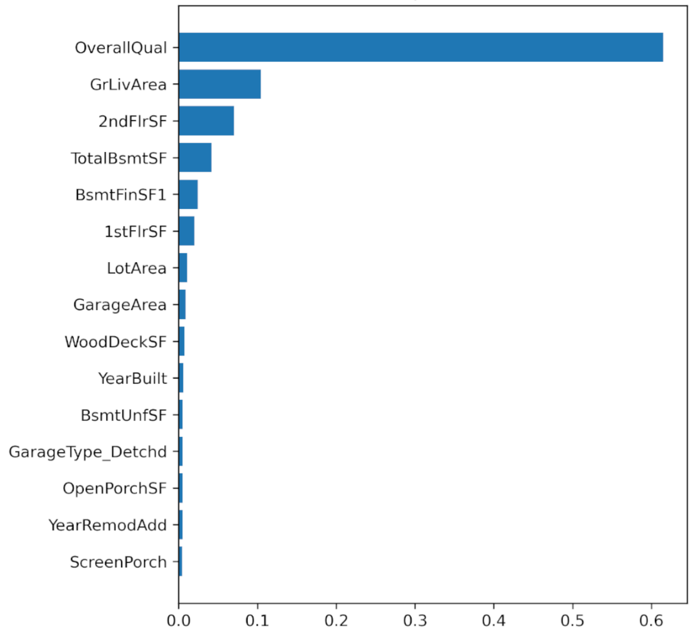
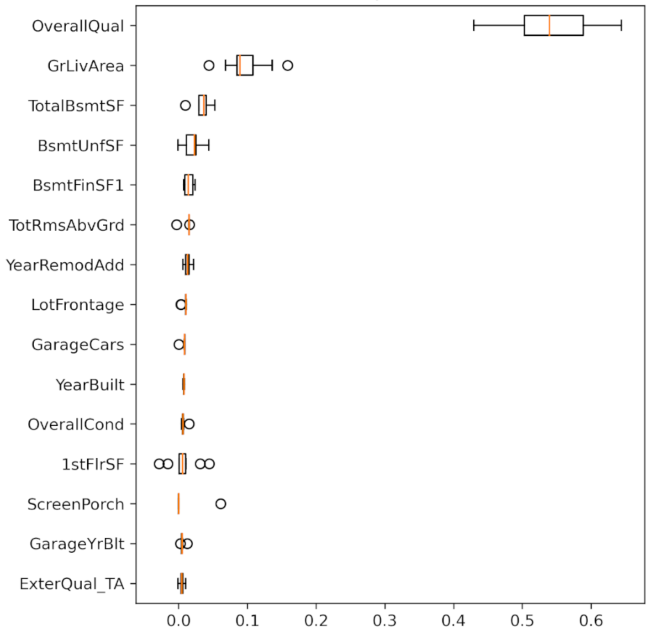


Fig. 2

Permutation Importance (test set)

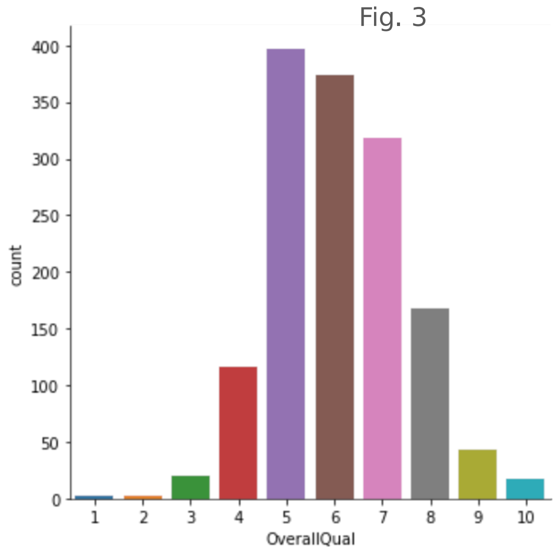


Findings/Results

Accuracy scores of regression models:
•Random Forest Regressor: cross-val score ~0.86709, Kaggle score: 0.14547
•Gradient Boosting Regressor: cross-val score ~0.81809, Kaggle score: 0.17250

We found that the features that contibute most towards a houses final price are overall quality, area of above ground living area, basement area, nr. of rooms, year built etc. (Fig. 1)

Fig. 3



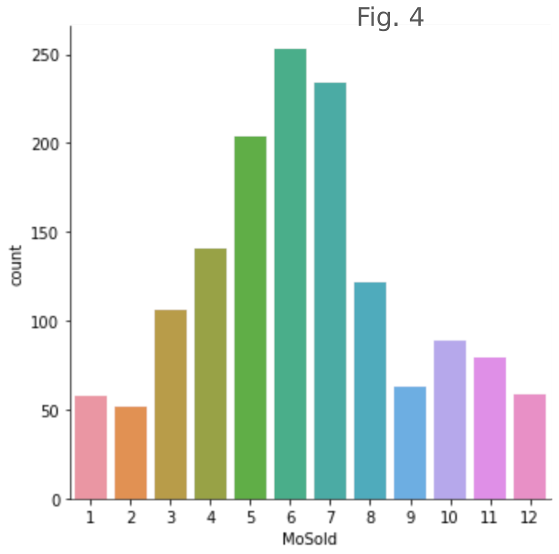
* Neighbourhood describes phisycal location within Ames city limits

** Overall quality rates a houses material and finish 1-10

Observations

Some interesting patterns can be observed in this dataset:
- The frequency of overall quality ratings approximates normal distribution (Fig. 3)
- The volume of house sales in the summer is four times the volume of the winter (Fig. 4)
- Increasing regression models estimators past 200 shows very little improvement toward it's performance

Fig. 4



Project code

https://github.com/kk-boop/IDS-House-Prices

UNIVERSITY OF TARTU
1632
IT Akadeemia toetab Skype™