KAGGLE HOUSE PRICES

Sandra Lannes, Sarah Lannes, Danver Hans Värv

Competition page.
Project repository

UNIVERSITY OF TARTU
Institute of Computer Science

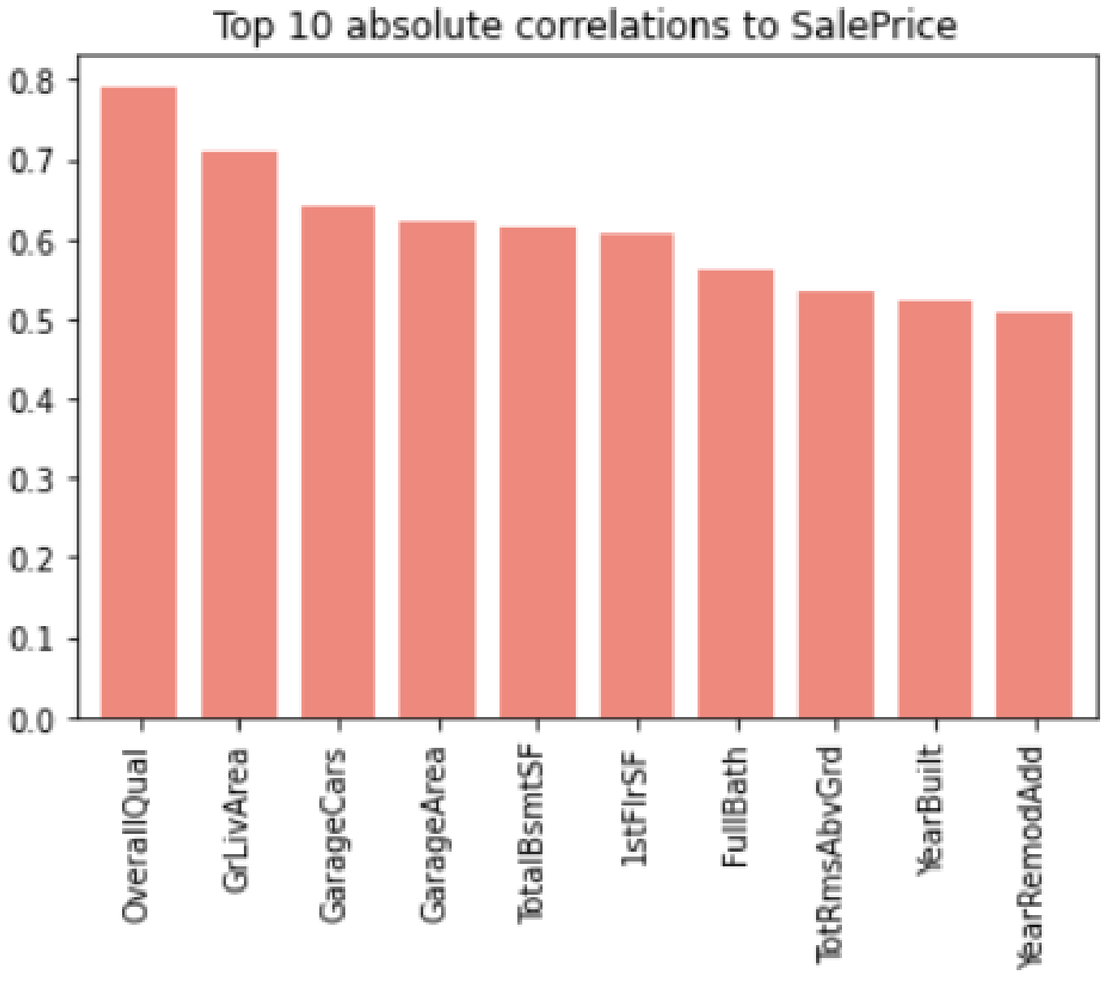
D
1

INTRODUCTION

For our project, we chose a Kaggle competition on the topic of predicting house prices. We were given a dataset with a number of different variables describing almost every aspect of residential homes in Ames, Iowa. It was said that this is a perfect competition for data science students who are looking to expand their skill set before trying a featured competition which seemed like a great match for our skill level.

THE PROCESS

We started with analyzing and cleaning up the data: filling in the NA values, removing outliers and one-hot encoding our data. After that, we started training predictive models starting from the ones that we learned during the course and finally experimenting with some new ones like XGBRegressor and GradientBoostingRegressor.



RESULTS

The main goal of this project was just to learn and consolidate our knowledge of data science. Our goal was to be in the top 50% of the most accurate teams.

Our success was evaluated on Root Mean Square Error (RMSE) which is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are. In other words, it tells you how concentrated the data is around the line of best fit.

In total, we submitted 16 models. During the work, we found out that regressors work best on our data and therefore we tried parameter tuning on different regressors. We got the best result with XGBRegressor, which gave us a score of 0.12851 and placed us in the top 28% of the leaderboard.

In conclusion, we met our goal of finishing in the top 50% of the leaderboard. In addition, we discovered and learned to use some new models, also found out the features that affect the price the most which can be found from the graph.