IST 707 Project Proposal

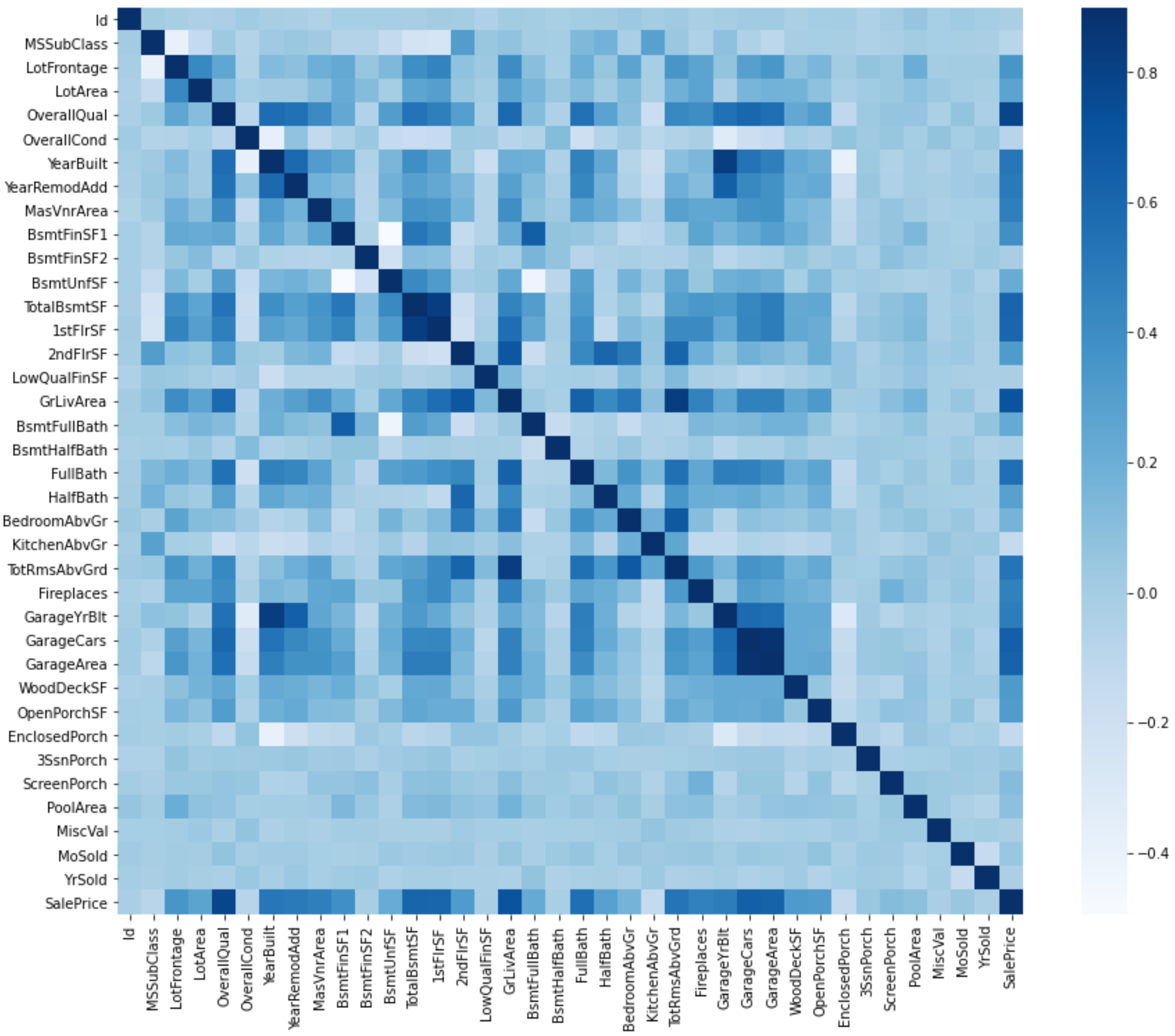
Zijun Yi, Xiaoben Yin

Project overview

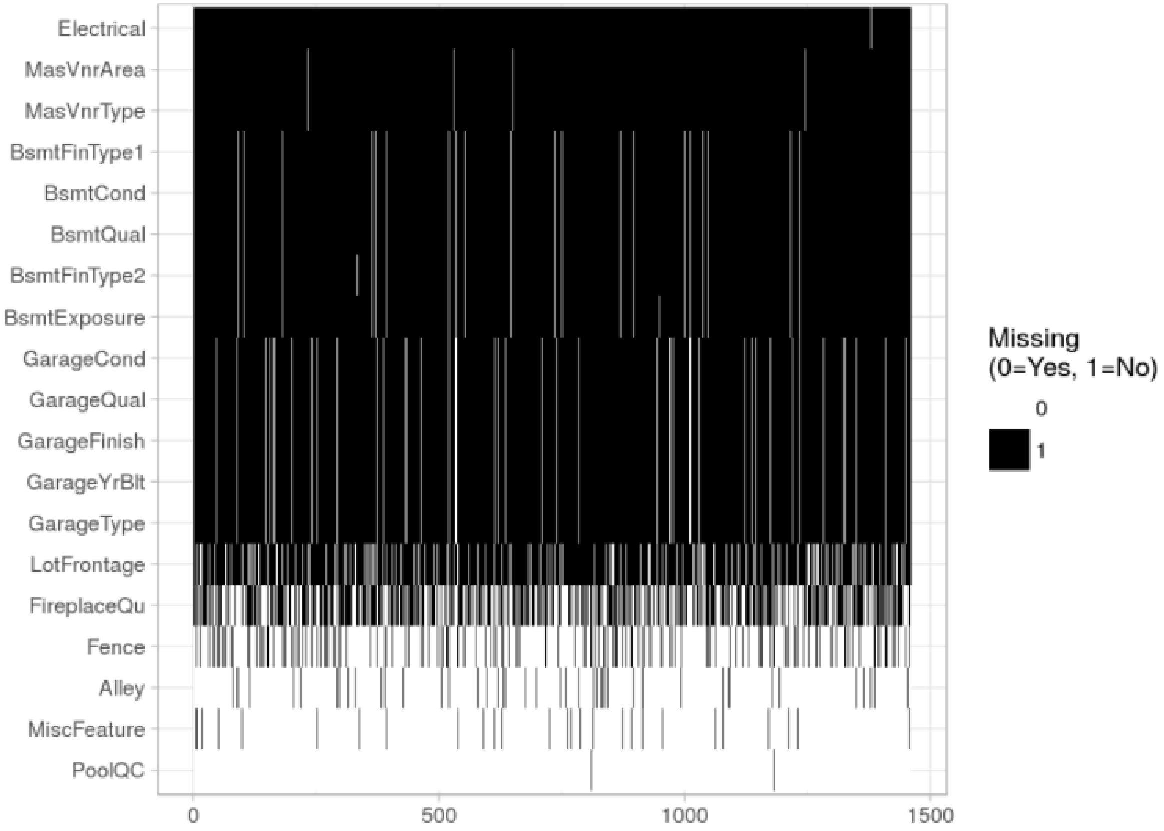
Owning a house is a dream of most people, but an ideal house with appropriate price and conditions is hard to find as the price is affected by many attributes: location, area, building type, number of bathrooms and bedroom, etc. In this project, we want to study the house sales prices and help customers find better options.

Dataset description

The data set we are going to use is from Kaggle, **Ames Housing Dataset**[1] by Dean De Cock. The dataset contains 1460 rows and 81 columns. It contains a detailed description of a house, some important columns are `GrLivArea`, the size of the living area, `OvreallQua`l, the overall quality, `GarageArea`, garage size, etc.



To summarize the missing values, in the categorical variables, most of them indicate that the majority of the houses do not have alley access, no pool, no fence and no elevator, 2nd garage, shed or tennis court that is covered by the `MiscFeature`, whereas the numeric variables do not have as many missing values. There are 259 in the `LotFrontage`, 8 in the `MasVnrArea` and 81 missing values in the `GarageYrBlt`. Besides, we also checked that the duplicate rows are 0 in the dataset



The dependent column `SalsePrice` is in dollars. We are going to predict the sales price with all the features in the linear regression models.

Data mining task

We plan to use KNN to deal with the missing values in the dataset as it's a good model for filling NA in both categorical and numeric variables. To predict the house prices and understand how each feature is related to the price, we're going to build two models: linear regression and random forest regression. Inside the random forest model, we intend to build random forest and GBT, and grid search cross validation. We'll check the performance of each model using metrics mentioned in the model evaluation and then use the best performed model to predict the price.

Model evaluation

Some metrics we are going to use to evaluate our model will be the Mean squared Error, R squared Error. This will examine the degree of overfitting in our model. To ensure an accurate result, we will be using a cross-evolution method to examine each model.

[1] https://www.kaggle.com/c/house-prices-advanced-regression-techniques ↩