|  |
| --- |
| Harbin Institute of Technology, Shenzhen |
| **Machine Learning Report** |
| House Prices Prediction |
| **Name 刘洋麟**  **StuId 19S151086**  **Date 2019.12.21** |
|  |

# **Introduction**

In this project, we use the house price data set (train data) from Kaggle.com to predict the final sale prices of each home in the test data set. Observing the train data set, each home involves a large number of features, in total 79 explanatory variables, describing almost every aspect of the resident homes.

The models and the method of data preprocessing is based on Python 3.7. And we also implement some Python packages, such as matplotlib, pandas, sklearn and scipy, to have better data visualization and built our models.

Firstly, we should understand our data set and the meaning behind all features through both univariate and multivariate approaches. Then, we apply data cleaning techniques to reduced noisy data and check for basic assumption of regression. Through data visualization, we can delete some obvious outliers. And using package pandas to check the missing values of all features. After checking, we apply different value filling approaches to different kind of features.

Second, we conduct feature engineering. Although data have been elementary processed, some of them cannot be directly used is models. And sometimes, adding some new features that are related to prediction can influence the models to have better results. Therefore, we convert some values into right form and add some new features.

Lastly, we build models, score them and get the prediction results from models. After comparing a few different models, we choose Ridge regression, Lasso regression, SVR regression, Bayesian Ridge, Elastic Net regression and Kernel Ridge regression to predict the final sale prices of the test data set.