Predicting Housing Category

Abstract

The goal of this project is to build a classification model to predict the category of houses. A real estate company want to separate their products into two groups, basic and luxury. Therefore, the customers can easily find and look at their ideal housed based on the budges or preference. The XGBoost has the best performance in the prediction housing category among all models. The price of house per unit (meters) is the most important features and the size of house also play an important role.

Design

The real estate company has different kinds of products (houses) and their customers have different needs. Classifying the category of houses via machine learning models would enable the customer find their dream houses efficiently and let the real estate agent promote the houses easily based on customers' need. In this way, the sales of houses increase.

Data

The dataset contains 10,000 houses with 18 features. The features include the size of houses (squre meters, number of rooms, number of floors, basement and attic and garage size), the number of previous owner, the age and the prices of houses. The target of the model is the category in two group, luxury or basic.

Algorithms

Data cleaning- Drop duplicates. Checking the NaN value in dataset.

EDA and Feature Engineering- Convert category feature to integer. Calculate the price per square meter based on the size and price of houses. Calculate the age of houses based on the years the houses were built.

Models- k nearest neighbors, logistic regression, decision tree, random forest, naive bayes, XGBoost were used as the prediction models. Ramdom over sampler and under samplers are used to resample the data.

Model evaluation- the data was split into training and test set. Different matric: accuracy, precision, recall, f1, ROC curves and auc are calcuted to evaluate the performance of the model.

Tools

- Numpy and Pandas- data organizing and manipulation
- SK learn and xgboost- classification model
- Matplotlib and Seaborn- data visualization
- Tableau for interactive visualizations

Communication