Lisbon Real Estate Investment

By: Miho Hunter

ABSTRACT: The goal of this project is to use classification models to identify properties that are good for investment.

DESIGN: Lisbon is an attractive city for buying property for various reasons. I am interested in building a model to categorize good and bad investment properties.

DATA: The data used for this model is from Re/MAX Portugal. Data for houses and apartments listed for rent and for sale were obtained by web scraping. 1270 properties were listed for rent, 6340 properties were listed for sale at the time of web scraping. 7 features were used for both regression and classification models including the categorical features.

ALGORITHMS: The first step was to build and optimize a linear regression model to predict rent from "for rent" properties. The model was applied to obtain potential rent for "for sale" properties. Monthly profit/loss was calculated from the sale and rent price. After data preparation and EDA, I built baseline models using various classification algorithms. I then chose the most promising algorithm, which was XGBoost. Finally, the best performance model was chosen by tuning hyperparameters by grid search.

TOOLS: Selenium and Beautiful Soup were used for scraping data from Re/MAX Portugal site and also for obtaining coordinates data from wikipedia for cities and divisions. Pandas was used for exploratory data analysis, and matplotlib and folium were used for visualization. Regression and classification models were built using sklearn.

COMMUNICATION: All slides, visuals, codes, and data are available on my github site.

CONCLUSIONS: XGBoost had the highest scores out of the 7 algorithms I tested. Test Accuracy was 0.794, and F1 score was 0.789. By optimizing the parameters by grid search, I was able to improve the test accuracy slightly to 0.788 and F1 score t0 0.810.