Algorithmic Machine Learning Project

Submitted by: Radhika Rajeevan & Sunit Nair

Title: Price prediction for listings on AirBnB (New York)

Summary: The project aims to analyze the AirBnB data set available here and apply techniques learned as part of the course and some additional methods to predict the price that a new property should expect to charge based on its features. The data for New York City was extracted and used for the purpose of this project. Analysis, filtering, and extraction of categorical variables from the data set was done using Python. The combined, cleansed data file was loaded into a database (MySQL) from where it is read for regression. Extraction of derived columns and running the regression models was performed in R and the following methods were used: Linear Regression, Tree, Ridge/Lasso Regression, and XGBoost. Validation set and k-fold cross validation techniques are used. The output of the predicted prices (for data without prices in data set) is written to a CSV file at the end.

Project phases and contributions:

Data set selection: Radhika Rajeevan & Sunit Nair

Data analysis: Radhika Rajeevan & Sunit Nair

Data cleansing and extraction of derived features (Python): Radhika Rajeevan

Database (MySQL) loading and connection to database from R: Sunit Nair

R code (Derived data/Regression/PCA): Radhika Rajeevan & Sunit Nair