MGMTMSA 403: Optimization

Assignment 3: Predicting Airbnb Prices

Background

The file `AirbnbTrain.csv` contains data on 1700 Airbnb listings in Hollywood, CA. The dataset contains features such as the location (by latitude and longitude), number of guests accomodated, number of beds, and other variables. The dataset also contains the price per night of each Airbnb listing. Your assignment will be to formulate an optimization model to predict the price of Airbnb listings using this dataset.

Note: Given a set of model coefficients $\beta_1, \beta_2, \ldots, \beta_d$, the average prediction error of the model $\boldsymbol{\beta}$ for a data set $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$ is given by

$$\text{Error} = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{d} \beta_j x_{ij} \right|$$

Questions

1. **Model 1.** Formulate the least absolute deviations regression problem as a linear program. Solve the linear program using the data given in the file `AirbnbTrain.csv.` What is the prediction error, in $/night, of your model on the test set (provided in `AirbnbTest.csv`)?

2. **Model 2.** Suppose that to improve interpretability, you wish to build a model that predicts Airbnb prices using only the three most important variables. Modify **Model 1** by including a constraint that allows at most three variables to have non-zero coefficients.

   a) List the names and coefficients of the three variables selected by the optimization model.

   b) What is the new prediction error, in $/night, of **Model 2**?

3. **Model 3.** Suppose now you wish to build a model that predicts Airbnb listing price using only three variables, where one of the variables is the number of beds.

   a) List the names and coefficients of the two other variables selected by the optimization model.

   b) Which variable was in **Model 2** but is no longer in **Model 3**? Briefly explain in 1-2 sentences why this variable might have been dropped.

   c) What is the new prediction error, in $/night, of **Model 3**?