

DSC 190: Intro to Data Mining – Data Mining Challenge

HDSI, UCSD, Spring 2020

Challenge Objective

Given the partial dataset of the Airbnb listings at New York City, you are asked to design data mining models to build a relationship between the price (dollars per day) and the other observed variables. That is, you are asked to predict the price of the listing, given all its information.

Due date: May 17, 4:59:59 PM. Note that the **time displayed on Kaggle is in UTC**, not PT.

Data and Baselines

train.csv: It contains all the training data that you can use in this challenge. The first column “id” provides you the unique key to identify the listings. Different columns show different features/attributes of a listing, including free texts, numerical features, and categorical features. There are also many missing values. So please conduct some exploratory data analyses (EDAs) first before you work on feature engineering.

test.csv: It contains all the listings that you need to predict their prices. The format is the same as the train.csv, except that the price column has been removed.

simple_baseline.ipynb: It contains simple baseline methods there. By running this notebook, you will be able to get **simple_linear_regression_baseline.csv** (leaderboard score: 115.03166) and **mean_value_baseline.csv** (leaderboard score: 134.98579) as output. The first one is produced by the linear regression model + very simple features. The second one is blindly predicting the mean price based on training data for all listings.

You can download all the datasets and baselines [here](#).

Evaluation Metrics

Your predictions will be evaluated against the groundtruth prices. The RMSE metric is adopted. For each test listing, we will calculate the squared error between the groundtruth and your prediction. We will take an average of all listings and then get the square root. For more information, please check the [Lecture 0 Slides](#) or [this webpage](#).

Scoring in Course Grade

If you can achieve a RMSE strictly smaller than the “simple-linear-regression” benchmark, you will be able to get 50% of the credits. If you can achieve a RMSE smaller than 100, you will be able to get 30% more of the credits. The remaining 20% will be decided based on your ranking.

Submission Format

You are asked to run your models locally and upload your final prediction file. It is a csv file with headers of two columns: **Id** and **Predicted**. The first character must be capitalized. The first column corresponds to the **id** in the **test.csv** file and the second column contains the predicted price.

Once submitted, the system will evaluate on a fixed portion (50%, randomly chosen) of the test set and compute RMSE accordingly. Then your score will be displayed on the leaderboard.

Please note that the leaderboard during the challenge is NOT final. The final leaderboard will be refreshed once the challenge ends. A new RMSE score will be calculated based on the other 50% portion which has not been tested yet.

Every day, you can make at most 20 submissions. So please start early and make sure you have enough time to tweak your models and hyperparameters. You will be able to choose 2 submissions for the final evaluation and the system will pick the best score you have.

General Rules

- No external data.
- No teaming.
- No cheating.
- Have fun!