CS235 Homework 2 (Kaggle Challenge)

Fall 2024

1 Description

Given the partial dataset of Airbnb listings in 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. Note, you may use off the shelf libraries to train the models.

Join the Kaggle competition here: `https://www.kaggle.com/t/0a12c45d0c6b409bb717f5a4222195bf`. Note, you will have to join with your <netid>@ucr.edu email address. If you have issues please let me know right away.

Note, this is an individual assignment and all submissions are individual. Kaggle might use the term teams, but teams can be size of 1 or more, but in our case its just a team of 1 :)

2 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 and mean_value_baseline.csv 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.

3 Evaluation Metrics

Your predictions will be evaluated against the ground-truth price using the RMSE metric. For each test listing, we will calculate the squared error between the ground truth and your prediction. We will take an average of all listings and then get the square root.

3.1 What is Root Mean Square Error?

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

The formula is :

RMSE $= \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}}$, where:

- variable i, denotes an item in dataset
- N number of data points / objects
- x_i is the actual value of object i
- y_i is the predicted value of object i

4 Registering your Kaggle Team Name

You must join Kaggle with your netID account and you will be asked to give your team a name. For the team name, enter your full name (Firstname$_Lastname$).$Folksthatdon't usetheirrealname/fullname$

5 Scoring

If you can achieve an RMSE strictly smaller than the "simple-linear-regression" benchmark, you will be able to get 50% of the credits.

Scoring Points (100 total)

- 30 points : if RMSE < simple-linear-regression benchmark
- 20 points : if RMSE < 100
- 10 points : dependent on ranking (of your top three submissions)
- 20 points : training/tuning and evaluation of three different models
- 10 points : plots to display results between different methods
- 10 points : discussion section

6 Submission Format

There are two parts to the submission. 1) Submit predictions to Kaggle (be sure to add a description that includes Firstname Lastname), 2) Submit code-base (ipynb) to Gradescope. Your approach should include:

1. Testing 3 different methods (and tuning hyper-parameters for these models). This also means you should have at least 3 submissions on Kaggle.
2. Data pre-processing / cleaning
3. Plots to visualize results from each method / model
4. An approach to validating your model (use your own approach, validation dataset, or cross-validation)
5. A section at the end of your notebook summarizing the results and what you have attempted and what worked or did not work. This should be at least a half page (embedded into your notebook).

You are asked to run your models locally and upload your final prediction file to Kaggle. It is a CSV file with headers of two columns: 'id' and 'price'. 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 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. Please start early and make sure you have enough time to tweak your models and hyper-parameters. The system will pick the best score you have submitted so far.

7 General Rules

- No external data.
- No teaming.
- No cheating.