eJam

Data Scientist Test Assignment

Congratulations! Due to your expertise in Data Science, you have been shortlisted for the next step in the application process for this role. In order to be considered for the next round (interview with the manager), you will have to complete this assignment within three (3) days from the date it was shared with you.

As you are aware, eJam is a performance marketing company, specialising in direct to consumer eCommerce for its own portfolio of company owned, company-operated brands. Carefully placed advertisements on social media channels and powerful analytics are the backbone of the business.

As a result, you have been provided with the Seattle AirBnb dataset of listings data. You will need to predict the prices of listings shared in the test dataset. A natural use case for this would be in helping people price their listings. In your assignment, feel free to use either Python or R, with any libraries of your choice. You will be evaluated on being able to justify your solution during the interview, explain the underlying mathematics and statistical phenomena, and make accurate predictions.

Task Description

1. Download a sample of the Seattle AirBnb listings dataset linked here (original data can be found here).
2. The goal of the assignment is to predict the prices of AirBnb listings from the test set, using a model carefully selected by you, trained, tested, and explained.
3. Conduct some exploratory data analysis and understand the relationships between potential predictors. Document your EDA in a notebook.
4. Note that this is not a particularly large dataset. You will be partially scored based on your ability to perform ETL on the dataset. Describe what you have done for ETL in 3-4 sentences.
5. Try out a few different models (use your judgement after doing the EDA), and note down why you have tried each one (2-3 sentences describing the "why" is enough).
6. Pick your final model, and explain why this model is better than the others. Train it, test it, and list out your analyses (4-5 sentences, or more if required). Finally, run your predictions on the real test set provided above.

Submission

Your final submission should be emailed to the hiring manager and have two files, as follows:

1. Notebook with the following components and partial scores for each component:
    a. **EDA**- documented in the notebook, with graphs describing correlations between variables, potential predictors, initial analyses on the data, and feature engineering (if any)
    b. **Data Engineering** - documented in the notebook, in a few sentences describing the ETL process and any data engineering that was performed
    c. **Initial Modelling** - a few models run on smaller folds of the dataset, with explanations for why each model was experimented with
    d. **Model Selection** - analyses around output from each of the models initially selected, and justification for selecting one model over the others you had initially contemplated
2. CSV file of listing prices for the test set:
    a. **Final Predictions** - each listing from the data set and the model-predicted price (2 columns: id, price)

If you need further clarification on this assignment, reach out to the hiring manager. Your submission will be evaluated within one day, after which you will be notified about an interview. Best of luck!