Data Science Interview Assignment

Assignment Description

Being a job search engine, it's helpful if we can suggest an approximate salary to job seekers for a given job post. Unfortunately, not all job postings include the salary. This is where you come in: your first task as an Indeed Data Scientist is to develop a salary prediction system. The goal: provide estimated salaries for a new job posting.

Data Supplied

You are given three CSV (comma-separated) data files:

- `train_features.csv`: Each row represents metadata for an individual job posting. The "jobId" column represents a unique identifier for the job posting. The remaining columns describe features of the job posting.
- `train_salaries.csv`: Each row associates a "jobId" with a "salary".
- `test_features.csv`: Similar to `train_features.csv`, each row represents metadata for an individual job posting

The first row of each file contains headers for the columns. Keep in mind that the metadata and salaries have been extracted by our aggregation and parsing systems. As such, it's possible that the data is dirty (may contain errors).

The Task

You must build a model to predict the salaries for the job postings contained in `test_features.csv`. The output of your system should be a CSV file entitled `test_salaries.csv` where each row has the following format:
`jobId,salary`
As a reference, your output should mirror the format of `train_salaries.csv`.

To judge the accuracy, we will compare your salary predictions to a ground-truth using the root-mean-square error (RMSE).

As a guideline, you should expect to spend around 4 hours to complete this exercise (including model training time). The assignment does not have to be completed all at once. Please do not share the assignment, the data, or your solutions with anyone other than your recruiter.

Deliverables

The following deliverables must be submitted to Indeed:

- Your `test_salaries.csv` file containing the salary predictions and job Ids for the test data set (please use .zip or .gz compression).
- The code that you wrote to solve the problem
- Answers to the questions below [in `.pdf` or `.txt`].
- Any related files such as figures, etc…

Please do not include your name in any of the deliverables, since evaluation is double blind.

Questions

Please answer the following questions.

1. How long did it take you to solve the problem?
2. What software language and libraries did you use to solve the problem? Why did you choose these languages/libraries?
3. What steps did you take to prepare the data for the project? Was any cleaning necessary?
4. a) What machine learning method did you apply?
   b) Why did you choose this method?
   c) What other methods did you consider?
5. Describe how the machine learning algorithm that you chose works.
6. Was any encoding or transformation of features necessary? If so, what encoding/transformation did you use?
7. Which features had the greatest impact on salary? How did you identify these to be most significant? Which features had the least impact on salary? How did you identify these?
8. How did you train your model? During training, what issues concerned you?
9. a) Please estimate the RMSE that your model will achieve on the test dataset.
   b) How did you create this estimate?
10. What metrics, other than RMSE, would be useful for assessing the accuracy of salary estimates? Why?