Data Science Interview Assignment – Salary Predictions

Overview

Submission Deadline: Please try to complete this assignment within 10 days after receiving it.

As a guideline, you should expect to spend around 4-12 hours to complete this exercise. The assignment does not have to be completed all at once.

This is an opportunity to showcase your unique skill set. You may choose to highlight any areas that you believe are unique strengths:

- Advanced predictive modeling
- Feature engineering
- Data engineering and pipelining
- Software engineering, e.g. object-oriented design, unit testing, etc.
- Data visualization
- Front-end development

Please do not feel obligated to conquer all the above bullet points – they are simply a list of areas that you can focus on to showcase your skills as part of this exercise.

Feel free to choose whatever language you are most comfortable using to complete this assignment.

Once completed, submit your code, results, and answers to:

joe.smith@big-tech-analytics.com

Assignment Description

Your job as a data scientist is in this assignment is to examine a set of job postings with salaries and then predict salaries for a new set of job postings.

Data supplied

You are given three CSV 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 salary data 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.

Deliverables

The following deliverables must be submitted:

• Your test_salaries.csv file containing the salary predictions for the test data set (.zip, .7z, or .gz compression is allowed).

• The code that you wrote to solve the problem (.zip, .7z, or .gz compression is allowed).

• Answers to the questions below.

Questions

Please **briefly** answer the following questions. Try to be as concise as possible while still giving complete answers – you will have the opportunity to add more detail and explanation regarding your approach via a follow-up discussion.

1. How long did it take you to solve the problem?

2. What software language and libraries did you use to solve the problem?

3. What steps did you take to prepare the data for the project? Was any cleaning necessary?

4. What algorithmic method did you apply? Why? What other methods did you consider?

5. Describe how the algorithmic method that you chose works?

6. What features did you use? Why?

7. How did you train your model? During training, what issues concerned you?

8. How did you assess the accuracy of your predictions? Why did you choose that method? Would you consider any alternative approaches for assessing accuracy?

9. 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?

10. Please explain any additional work that you did as part of this project.