Data Science Interview Assignment

Submission Deadline:

As a guideline, you should expect to spend around 4 hours to complete this exercise. The assignment does not have to be completed all at once. Once completed, submit your questions and results to:
rfurlong-ce@indeed.com

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 designate 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_DATE.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_DATE.csv: Each row associates a "jobId" with a "salary".
• test_features_DATE.csv: Similar to train_features_DATE.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 has 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_DATE.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_DATE.csv.

Deliverables:

Please Keep ALL Deliverables in Separate Files. (There should be 1 document for the csv file, 1 document for code, and 1 document for answers to the questions below)

The following deliverables must be submitted to Indeed:

- Your test_salaries.csv file containing the salary predictions for the test data set (.zip or .gz compression is allowed).
- The code that you wrote to solve the problem (.zip or .gz compression is allowed).
- Answers to the questions below.

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?
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?