**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.