Indeed Questions

Please answer the following questions.

1. How long did it take you to solve the problem?
   Approximately 3.5 hours

2. What software language and libraries did you use to solve the problem? Why did you choose these languages/libraries?

   Python – pandas, numpy, scipy, sci-kit-learn.
   I used these packages because they have great machine learning packages I'm familiar with.

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

   Data cleaning:

   Step 1: Drop these "id" column as its useless

   Step 2: Convert categorical data into "dummy" columns with binary values, dropping one of the columns to prevent artificial correlation

   Step 3: Put all numerical onto the same scale using standardization.

4. a) What machine learning method did you apply?

   Stochastic Gradient Descent and RandomForestRegressor

   b) Why did you choose this method?

   SGD is a baseline linear model that can demonstrate predictability of a linear model. Random Forest was attempted to deal with any possible non-linear relationships in the data.

   c) What other methods did you consider?

   GridsearchCV hyperparameter tuning of these techniques. If my laptop was a little faster I would have used it.

5. Describe how the machine learning algorithm that you chose works.

Stochastic gradient descent combines the gradient descent algorithm with ordinary least squares. It attempts to learn the regression coefficients based upon randomly selected data points. As it goes through the data, the coefficients get closer to values that (at least locally) minimize the error.

6. Was any encoding or transformation of features necessary? If so, what encoding/transformation did you use?

   I used "dummification" of categorical variables.

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?

   # **Using the Pearson correlation coefficient, we can find the most relevant features. Features with an absolute value of high correlation with the target variable, salary, will probably be good predictors.**

   Most impact – years of experience, some job types, some degrees
   Least impact: the company you work for

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

   Looping through a list of possible models, and then evaluating the metrics on the train and validation sets for each model. The issue was the processing time, the data set had a lot of rows.

9. a) Please estimate the RMSE that your model will achieve on the test dataset.

   20.1k

   b) How did you create this estimate?

   By evaluating the rmse on a validation set.

10. What metrics, other than RMSE, would be useful for assessing the accuracy of salary

estimates? Why?

R2 score is good because it gives a range between 0 and 1.