1. How long did it take you to solve the problem?
   Higher level polynomial regression and grid searching on the best model have taken long time for me as I really do not know how to deal with memory issues on my laptop.

2. What software language and libraries did you use to solve the problem?
   I used python for this portfolio project, and it is my first project in python. I have done some programming in python before but most of my coursework I have done using R. I used the libraries: numpy, panda, scikit learn, stats models and matplotlib.

3. What steps did you take to prepare the data for the project? Was any cleaning necessary?
   In the data wrangling part, I do not have much to do. Because there are no missing values. I checked if there any negative values in the target variable since the salary cannot be a negative value.

4. What algorithmic method did you apply? Why? What other methods did you consider?
   I started using linear regression. Then I tried multiple regression, polynomial regression and the tree-based models such as random forest, gradient boosting and xgboost. I chose xgboost for further processing.
5. Describe how the algorithmic method that you chose works?
   XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. This algorithm recently dominates applied machine learning and Kaggle competitions.
6. What features did you use? Why?
   I used all the features plus a newly added feature for the model prediction. companyId seems to be useless on checking feature importance of the model. The newly added feature ended up as an important predictor of this model. On checking feature importance with random forest edu_level does not seem to be important.
7. How did you train your model? During training, what issues concerned you?
   Training models that need huge memory spacing created problems for me. I started my project on jupyter lab on ubuntu. Later I used google colab for grid searching. I even tried installing dask on to my laptop.

8. How did you assess the accuracy of your predictions? Why did you choose that method?
   I used mean squared error 'mse' as by the instruction.

Would you consider any alternative approaches for assessing accuracy?

I used rmse for assessing accuracy on a plot visualization.

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?

The features job_type and edu_level has got most impact on salary. Feature importance function of the model is used to identify the feature importance of the columns. The least impacted feature is companyId but still I included this for model prediction because logically thinking, financial status of the company does have an impact on the salary of the employees.

9. Please explain any additional work that you did as part of this project.
   For this project I tried installing dask and I did lots of research online since this is my first project that I did completely in Python.