**Assignment**

Objective is to create a model to predict remuneration of the vacancy/position in a data analytics firm.

In the train dataset, we have the input variables:

1. Experience
2. Job description
3. Job designation
4. Job Type
5. Key skills
6. Location
7. Salary
8. Company code

In the test dataset, we have all above variables except **Salary** (which we need to predict)

Apart from the output, also answer briefly, the following points:

1. Final variable list used in the model

**Final variable selected are: 'company\_name\_encoded', 'exp\_upper', 'key\_skills'**

**Exp\_upper is created from the ‘experience’ column by taking the higher value from the experience range.**

**Salary has also been converted to salaray\_final by taking both lower and upper range of salary.**

**MinMax scaler can also be used to fine tune.**

1. Pre-processing variables, e.g. cleaning variables, re-classification of categories, creation of dummy variables, outlier treatment, imputed variables, changed variable datatype/ structure, etc.

**Cleaning, data preparation is included in jupyter notebook**

**Job desig also included city names, that has also been cleaned.**

1. Model technique used

**For model building, Random forest regressor is used**

1. Variable importance/ strength/ contribution (if possible)

**SelectFromModel from feature\_selection is used for feature importance calculation**

1. Any other information

**Text columns can be refined using other packages like glove, currently due to lack of time have used pre trained google’s word2vec model.**

**I have attached the nlp\_ops.py file which is kind of pipeline for NLP.**

**Note:** Please put visualizations wherever possible. Jupyter notebook to be used for assignment.

There is an HTML report attached. Please note, that package is created by me and hence might not be able to run it. But I have attached the report for your reference.