**ABHINAV SACHDEVA**

**DATA SCIENCE INTERVIEW ASSIGNMENT:**

Q1: What software languages and libraries did you use to solve the problem?

A1: Software language used: R. The libraries used to solve this problem are listed below:  
dplyr, readr, ggplot2, caret, glmnet, rpart, randomForest, gbm, xgboost. Caret.  
  
Q2: What steps did you take to prepare the data for the project? Was any cleaning necessary?  
  
A2: To start off, I studied each and every variable closely by looking at its structure as well graphically. After that, I converted each variable into it’s correct data type. There wasn’t any kind of data cleaning which was required for this dataset because there wasn’t a single missing value nor was there any inconsistencies which any of the values provided in the data. For some of the variables, there were levels which had NONE but that made sense and gave us information. For eg: Degree- NONE absolutely makes sense.

Q3: What algorithmic method did you apply? Why? What other methods did you apply?  
  
A3: Because our dependent variable was a continuous one, I started off with Multiple Regression, to improve the performance, I applied regularization algorithms like Ridge, Lasso but that did not improve the performance at all. After that, I tried RandomForest and GBM to see how the tree based methods are performing on the data. Both of them worked better than Multiple Regression. Finally, my final model was based on ensemble methods.  
  
Q4: Describe how the algorithmic method that you chose works?  
  
A4: Multiple regression gave me a very good R-squared value (74.xx), a good start. Both ridge and lasso regression didn’t improve RMSE at all. Random forest model consumed excessive computing time due to large dataset. Boosting methods exhibit best prediction performance out of all single models. Ensemble method of the two aggregating algorithms produce optimal output. Hence deployed for final prediction.

Q5: What features did you use? Why?  
  
A5: I used all the features given in the dataset barring CompanyID because as shown in my analysis, companyID shows little variance in response variable.  
  
Q6: How did you train your model? During training, what issues concerned you?  
  
A6: Model is first trained by training set, then tested on hold-out set. During training, trade-off between accuracy, optimal parameter and speed concerned me.  
  
Q7: How did you assess the accuracy of your predictions? Why did you choose that method? Would you consider any alternative approaches for assessing accuracy?  
  
A7: Since this is a problem where we have to predict a continuous variable, RMSE is used as accuracy measure. Yes, I would consider other alternative approaches for assessing accuracy such as R squared and MSE.  
  
Q8: Which features had the greatest impact on salary? How did you identify these to be the most significant? Which features had the least impact on salary? How did you identify these?  
  
A8: Tree-based methods generate handy variable importance scores. JobType has the greatest impact on salary, followed by yearsExperience.   
CompanyId has the least impact on salary, followed by major, as shown from initial EDA and visulization, and can be confirmed by linear and tree models.