1. What software language and libraries did you use to solve the problem?

Used R. Following libraries were used:

a) mlbench
b) caret
c) earth
d) rpart
e) e1071
f) randomForest

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

Following steps were taken:

a) Joined train_feature_date with train_salaries_date on jobID
b) Removed rows where salary was 0

3. What algorithmic method did you apply? Why? What other methods did you consider?

Applied Generalized Linear Models (GLM) as it is computationally less intensive compared to other algorithms and gives good prediction accuracy.

Following methods were also considered:

a) Linear regression
b) Decision Tree
c) Random Forest

4. Describe how the algorithmic method tha t you chose works?

In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

The GLM consists of three elements:

a) A probability distribution from the exponential family.
b) A linear predictor $\eta = \mathbf{X}\boldsymbol{\beta}$.
c) A link function g such that $\mathrm{E}(\mathbf{Y}) = \mu = g^{-1}(\eta)$.

The maximum likelihood estimates can be found using an iteratively reweighted least squares algorithm.
(source: Wikipedia)

5. What features did you use? Why?

Following features were used:

a) jobType

b) degree
c) major
d) industry
e) yearsExperience
f) milesFromMetropolis

CompanyId was not used as it did not have high correlation with the dependent variables.

Boxplots were used to determine the influence of categorical variables on the dependent variable.

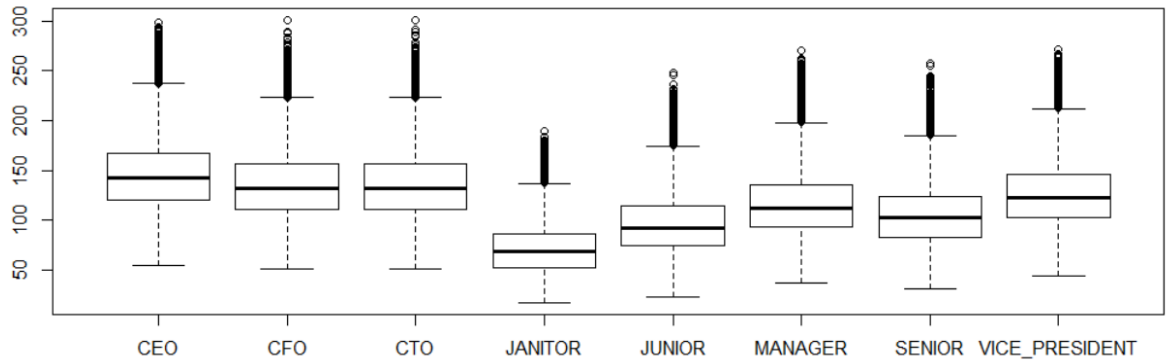

Figure 1: Boxplot of jobType and salary. It is observed that jobType has an effect on salary

Whereas, correlation was evaluated for numerical independent variables to gauge the influence on dependent variable

Variable	Correlation with Salary
YearsExperience	0.38
MilesFromMetropolis	-0.30

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

Training phase comprised the following steps:

a) Create random subsets for training (75% of data) and test data (25% of data)
b) Checked for model overfitting by cross-validation. Following are the error results for 10 different training subsets

MAE	MAPE	RMSE
15.84896	14.23772	19.61738
15.88646	14.26517	19.64318
15.83769	14.23755	19.60168
15.85477	14.24938	19.61528
15.83343	14.22597	19.60588
15.8455	14.24898	19.60254
15.81905	14.22237	19.5804
15.85065	14.25901	19.61833
15.87654	14.26545	19.65043
15.81947	14.22104	19.57801

7. How did you assess the accuracy of your predictions? Why did you choose that method? Would you consider any alternative approaches for assessing accuracy?

Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root mean squared error (RMSE) are three of the most common metrics used to measure accuracy for continuous variables. These metrics express average model prediction error in units of the variable of interest. They are negatively-oriented scores, which means lower values are better.

Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE should be more useful when large errors are particularly undesirable.

Accuracy is evaluated by subtracting MAPE from 100: **85%** accuracy

8. 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?

R function varImp was used to evaluate variable importance. It was observed that CompanyId had least impact on salary. Following is the snippet of important variables, higher the Overall value, more impact it has on salary.

	Overall
jobTypeCFO	109.130180
jobTypeCTO	109.233560
jobTypeJANITOR	653.941891
jobTypeJUNIOR	549.727976
jobTypeMANAGER	329.671322
jobTypeSENIOR	440.770869
jobTypeVICE_PRESIDENT	220.653860
degreeDOCTORAL	130.989320
degreeHIGH_SCHOOL	50.541727
degreeMASTERS	64.699859
degreeNONE	82.323349
majorBUSINESS	57.755327
majorCHEMISTRY	7.423283
majorCOMPSCI	29.774255
majorENGINEERING	79.599858
majorLITERATURE	28.376060
majorMATH	38.101452
majorNONE	37.525485
majorPHYSICS	17.122821
industryEDUCATION	117.511015
industryFINANCE	249.735676
industryHEALTH	74.340053
industryOIL	251.760345
industrySERVICE	58.383536
industryWEB	143.301370
yearsExperience	640.563851
milesFromMetropolis	508.495010