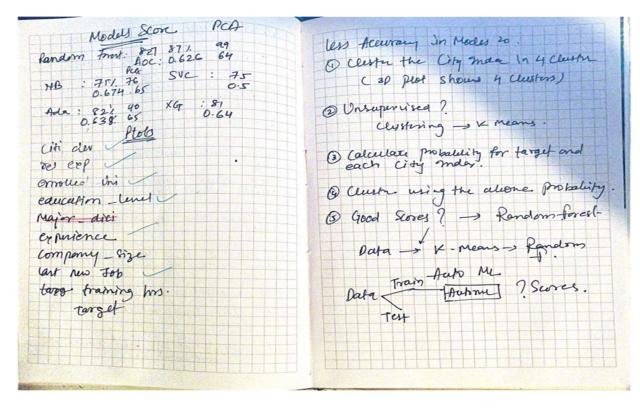
## Logbook

The following logs were maintained for the assignment.

- Python has detected three numeric and nine categorical features. However, the actual number of categorical features are 6, and numeric features are 6.
- More people in the dataset are from the cities with a development index greater than 0.9
- The gender column has 23% percent of missing values. However, 90% of the observations are male.
- The relevant experience column has 72% of people with relevant experience.
- The enrolled universities column has only 2 % missing values.
- Major discipline has 14% missing values, and more than 80% are from STEM
- The company size column has 31% missing values.
- Company type has 31% missing values.
- Last\_new\_job column has 2.2% missing values.
- Drop columns city, Gender, enrollment id and company type.
- Drop rows for missing values in a column less than 10%
- Replace the null values in company\_size with the mean value
- Plot all rest features with the target
- Split the data into train test split.
- Create a data pipeline structure One hot encoding ->PCA->Classifier
- Train Random Forest Classifier and calculate accuracy and AUC and plot ROC
- Train Gaussian Naive Bayes and calculate accuracy and AUC and plot ROC
- Train AdaBoostClassifier and calculate accuracy and AUC and plot ROC
- Train Support Vector Classifier and calculate accuracy and AUC and plot ROC
- Train XGBoost and calculate accuracy and AUC and plot ROC
- As the accuracy of the models is not satisfactory, cluster the data into four groups.
- Add the output of k-Means in the training dataset.
- As the performance of the Random Forest Classifier was good, use it to train modified training dataset.
- Calculate accuracy and AUC and plot ROC
- As the performance parameters are satisfactory. Use K- Fold cross-validation to check for over or underfitting.
- Use Grid Search cross-validation to find the best set of hyperparameters for PCA and Random Forest classifier.
- Calculate accuracy and AUC and plot ROC for the newly trained model.
- As this newly trained model is satisfactory on every test, use it to predict the test dataset.
- Calculate the confusion matrix and accuracy score.

As the Original logbook was maintained in a Notebook. So scanned copy is included below.

Data Science Actual	Company type - 32.04 Missing
O Numeric → 3? Actual	east-new-job -> 2.27. hissing on
catagorical -> 9 6	o larget high coreation with deallopment index.
city denting ! Mere people from	Odrop columns: City, Gender, emoli-id,
Grender -> 23%. Wissing Values  9'b'. Made in observation	Than to !.
Relevent = 72% has relevent on experience	brok rows with col having new values than 10%.
erround this > 2). Mirthly OK	Repeace the Company lize with
Major discipline - 141. Missing 80-90), from & TEM	O Check it the Major - discipline play evol in target.
experience - on	9 plot relevant enperionce us target
Company size - 31'1 wissey	O plot relevant enquionce us farget O plot and deulopment index prob is
ed Col City der enroued university education Juni, col company size (ast are jot, training heres when the	last now Job -> Off Behut premies 300



## References

Anon., n.d. *Welcome to cuDF's documentation!*. [Online] Available at: <a href="https://docs.rapids.ai/api/cudf/stable/#welcome-to-cudf-s-documentation">https://docs.rapids.ai/api/cudf/stable/#welcome-to-cudf-s-documentation</a> [Accessed 2021].

Bruce, P., Bruce, A. & Gedeck, . P., 2020. Practical Statistics for Data Scientists. 2 edition ed. s.l.:O'Reilly.

Géron, A., 2019. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. In: N. Tache, ed. Sebastopol: O'Reilly Media.

AutoKeras [WWW Document], n.d. URL https://autokeras.com/ (accessed 3.13.21).

Möbius, 2021. HR Analytics: Job Change of Data Scientists. [Online]

Available at: <a href="https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists">https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists</a>