

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

① Numeric $\rightarrow 3$? Actual 6

Categorical $\rightarrow 9$ 6

City develop: More people from city index > 0.9

Gender $\rightarrow 23\%$ Missing Values
90% Male in observation

relevant experience $\rightarrow 72\%$ has relevant experience

enrolled uni $\rightarrow 2\%$ Missing OK

Major discipline $\rightarrow 14\%$ Missing
80-90% from STEM

experience \rightarrow OK

Company size $\rightarrow 31\%$ Missing
But?

Let's say City dev, enrolled university, education level, exp, company size, last new job, training hours, relevant exp.

Company type $\rightarrow 32.0\%$ Missing

last-new-job $\rightarrow 2.2\%$ Missing OK

② Target high correlation with development index.

③ Drop columns: City, Gender, enroll-id, company-type.

④ ~~Drop rows with null values less than 10%.~~
Drop rows with col having null values less than 10%.

⑤ Replace the company size with mean

⑥ Check if the Major-discipline play role in target.

⑦ Plot relevant experience vs target

⑧ Plot City development index, prob vs target

Last new Job \rightarrow Diff betwⁿ previous job and current job.

Models Score

Model	Score	PCT
Random Forest	82% 87%	99
	AOC: 0.626	64
Naïve Bayes	77% 76%	
	0.674 65	
Ada	82% 40%	
	0.638 65	
SVC	75%	0.5
XG	81%	0.64

Plots

- City dev ✓
- rel exp ✓
- enrolled uni ✓
- education - level ✓
- Major-disc ✓
- experience ✓
- Company-size ✓
- last new job ✓
- training hrs. ✓
- target

Less Accuracy in Models 20.

- Cluster the City index in 4 clusters (3D plot shows 4 clusters)
- Unsupervised? Clustering \rightarrow K-means.
- Calculate probability for target and each City index.
- Cluster using the above probability.
- Good Scores? \rightarrow Random forest.

Data \rightarrow K-means \rightarrow Random

Data \rightarrow Train Auto ML \rightarrow ? Scores.

Test

References

Anon., n.d. *Welcome to cuDF's documentation!*. [Online]

Available at: <https://docs.rapids.ai/api/cudf/stable/#welcome-to-cudf-s-documentation>

[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: <https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists>