Machine Learning Project 3
Develop a Staffing Promotion Algorithm

The data for this project is courtesy of the Data Science Nigeria Kaggle Competition. https://www.kaggle.com/c/intercampusai2019

Koffi Essien, 42, is the heir apparent to the highly revered Essien business dynasty. The enterprise has spanned decades with vast investment interest in all the various sectors of the economy.

Koffi has worked for 18 years in Europe and America after his first and second degrees at the Massachusetts Institute of Technology and Harvard University where he studied Electrical Engineering and Business Management respectively. Koffi is a very experienced technocrat and a global business leader who rose through the ranks to become a Senior Vice President at a leading US business conglomerate. His dad is now 70 and has invited him to take over the company with a mandate to take it to the next level of growth as a sustainable legacy. Koffi is trusted by his father and his siblings to lead this mandate.

On resumption, he had an open house with the staff to share his vision and to listen to them on how to take the business to the next level. Beyond the general operational issues and increasing need for regulatory compliance, one of the issues raised by the staff was a general concern on the process of staff promotion. Many of the staff allege that it is skewed and biased. Koffi understood the concern and promised to address it in a most scientific way.

You have been called in by Koffi to use your machine learning skills to study the pattern of promotion. With this insight, he can understand the important features that can be used to predict promotion eligibility. The dataset contains these variables as explained below:

• Employee No : System-generated unique staff ID

• Division: Operational department where each employee works

• Qualification: Highest qualification received by the staff

• Gender: Male or Female

• ChannelofRecruitment: How the staff was recruited – this is via internal process, use of an agent or special referral

• Trainings_Attended : Unique paid and unpaid trainings attended by each staff in the previous business cycle

• Yearofbirth: Year that the employee was born

• LastPerformanceScore Previous year overall performance HR score and rated on a scale of 0-14

• Yearofrecruitment : The year that each staff was recruited into the company

• Targets_met: A measure of employees who meet the annual set target. If met, the staff scores 1 but if not, it is a 0.

• Previous_Award : An indicator of previous award won. If yes, it is a 1 and if No it is a 0.

• Trainingscoreaverage: Feedback score on training attended based on evaluation

• Foreign_schooled: An indicator of staff who had any of their post-secondary education outside the country. Responses are in Yes or No

• PastDisciplinaryAction : An indicator if a staff has been summoned to a disciplinary panel in the past. This is indicated as Yes or No

• PreviousIntraDepartmentalMovement : This is an indicator to identify staff who have moved between departments in the past. Yes and No are the responses.

• Noofprevious_employers : A list of the number of companies that an employee worked with before joining the organisation. This is recorded as counts

Question 1

- Prepare a comprehensive exploratory data analysis (EDA) using the training dataset. The EDA should include exploring the shape of the data, checking correlation matrix, bar charts, simple and complex plots like heat map and KDE, summary statistics etc.

Question 2

- What new features did you engineer with the data? Explain the different steps used to engineer and encode the following features: categorical and numerical variables? Are there missing values in the data? If yes, how did you impute them?

Question 3

- Check for class imbalance in the dataset. Are they imbalance? What is the percentage of the minority class to the majority class. How did you solve the imbalance class problem? What metric would you use for test evaluation.

Question 4

- Take the data given and split it into 80% train set and 20% test set. Use random seed = 42.
- Train a Binary Logistic Regression, Decision Tree Classifier, Random Forest Classifier, a Support Vector Machine Classifier and an xgBoost classifier on the training data.
- What parameters and hyperparameters did you use for this model?
- Use a 5 fold Cross Validation to tune your model.

Question 6

- Present the Accuracy Score, Precision, Recall and Confusion Matrix for evaluating each model with the hold-out test data. Plot the ROC Curve for the five models. What's the accuracy paradox and explain while the accuracy is not a suitable evaluation metric when you have an imbalanced class problem. What other metrics can you use?

Question 7

- What are the feature importance generated from the classifiers above? Create a scatter plot or bar chart for them. Show a chart for the mean feature importance from the five models above.
- Use a Maximum Hard Voting Ensemble Model of the five base models above to present a final model.