**Background & Context**

The HR team in JMD company stored data of promotion cycle last year, which consists of details of all the employees in the company working last year and also  
if they got promoted or not, but every time this process gets delayed due to so many details available for each employee - it gets difficult to compare and decide.

So this time HR team wants to utilize the stored data to make a model, that will predict if a person is eligible for promotion or not.

You as a data scientist at JMD company, need to come up with a model that will help the HR team to predict if a person is eligible for promotion or not.

**Objective :**

1. Explore and visualize the dataset.
2. Build a classification model to predict if the customer has a higher probability of getting a promotion
3. Optimize the model using appropriate techniques
4. Generate a set of insights and recommendations that will help the company

**Data Dictionary:**

* employee\_id: Unique ID for the employee
* department: Department of employee
* region: Region of employment (unordered)
* education: Education Level
* gender: Gender of Employee
* recruitment\_channel: Channel of recruitment for employee
* no\_ of\_ trainings: no of other trainings completed in the previous year on soft skills, technical skills, etc.
* age: Age of Employee
* previous\_ year\_ rating: Employee Rating for the previous year
* length\_ of\_ service: Length of service in years
* awards\_ won: if awards won during the previous year then 1 else 0
* avg\_ training\_ score: Average score in current training evaluations
* is\_promoted: (Target) Recommended for promotion

**Best Practices for Notebook :**

* The notebook should be well-documented, with inline comments explaining the functionality of code and markdown cells containing comments on the observations and insights.
* The notebook should be run from start to finish sequentially before submission.
* It is preferable to remove all warnings and errors before submission.
* The notebook should be submitted as an HTML file (.html) and NOT as a notebook file (.ipynb)

**Best Practices for Presentation :**

Like in real-world projects, the ultimate destination of any project or work is generally an executive or decision-making meeting, where you are supposed to present your solution to the business problem, based on the project/work you have done. The purpose of this presentation is to simulate that kind of experience and to draw the attention of your audience (a business leader like CMO, COO, CFO, or CEO) to the key points of your project, which are

* Business Overview of the problem and solution approach
* Key findings and insights which can drive business decisions
* Model overview and performance summary
* Business recommendations

Please keep the following points in mind while making the presentation:

* Focus on explaining the takeaways in an easy-to-understand manner.
* Inclusion of the potential benefits of implementing the solution will give you the edge.
* Copying and pasting from the notebook is not a good idea, and it is better to avoid showing codes unless they are the focal point of your presentation.
* Please submit the presentation in PDF format only.

**Submission Guidelines :**

1. There are two parts to the submission:
   1. A well commented Jupyter notebook [format - .ipynb]
   2. A presentation as you would present to the top management/business leaders in PowerPoint

**Requirements :**

|  |  |  |
| --- | --- | --- |
| **Section** | **Requierement** | **Details / MUST HAVE** |
| Exploratory Data Analysis | * Problem definition, questions to be answered * Data background and contents * Univariate analysis * Bivariate analysis | 1. Definition of problem (as per given problem statement with additional views) 2. Observations on shape of data, data types of various attributes, missing values, statistical summary. 3. Univariate Analysis (boxplots, histograms, countplots, distribution plots for important variables ) 4. Bivariate Analysis (Correlation matrix, Pairplot, target vs other variables and any other important graph) |
| Illustrate the insights based on EDA | Key meaningful observations on individual variables and the relationship between variables | 1. Comments on range of attributes, outliers of various attributes 2. Comments on the distribution of the variables and potential fixes for better modeling 3. Comments for each univariate and bivariate plot |
| Data pre- processing | Prepare the data for analysis   * Feature Engineering * Missing value Treatment * Outlier Treatment | 1. Identification of missing values 2. Split data into train and test sets 3. Missing value treatment or a strong reason for not treating the missing values 4. Reason for dropping columns |
| Model building - Logistic Regression | * Create logistic regression model * LR with upsampling * LR with downsampling * Regularize model or provide reason why not regualrize | 1. Logistic regression model built using sklearn 2. Comment on model performance 3. Make logistic regression with upsampled data 4. Logistic regression with downsampled data 5. Regularize model, if needed and comment on performance |
| Model building - Bagging and Boosting | - Build Decision tree, random forest, bagging classifier , Xgboost, AdaBoost, and gradient boosting models | 1. All 6 models are build 2. Commented on model performance for each model |
| Hyperparameter tuning using grid search | * Tune all the models using grid search * Use pipelines in hyperparameter tuning | 1. Choose the best metric to evaluate the performance of the model and for scoring in GridSearchCV. Recall should be selected. 2. Tuned best 3 models using pipelines 3. Commented on why choosing those 3 models only 4. Commented on model performance |
| Hyperparameter tuning using random search | * Tune all the models using random search * Use pipelines in hyperparameter tuning | 1. Choose the best metric to evaluate the performance of the model and for scoring in GridSearchCV. Recall should be selected. 2. Tuned best 3 models using pipelines 3. Commented on why choosing those 3 models only 4. Commented on model performance |
| Model Performances | * Compare the model performance of all the models * Comment on the time taken by the grid and randomized search in optimization | 1. Compare model performances and comment 2. Comment on time taken by grid and random search |
| Actionable Insights & Recommendati ons | - Business recommendations and insights | 1) Actionable Recommendations - such that the business can take action upon them. |
| Presentation - Overall quality | * Structure and flow * Crispness * Visual appeal   All key insights and recommendations covered? | 1. Clear structure and flow - everything sits well in a story - (Problem - Data overview - Solution overview - Findings - Recommendations) 2. Crispness - not too many words - just enough to keep the focus on key things/points 3. Visual appeal - use of charts, colors, diagrams, format, symmetry - informative visualizations that are easy to interpret 4. All key insights and recommendations covered - all key ones from EDA are stressed upon & important insights are not left just in the notebook - final model is crisply presented - comparative tables for test-train-validation with rationale for choosing the final model - crisp summary of drivers of the model 5. clear points on what business should do with this model and how will that be beneficial   Comments on additional data sources for model improvement, model implementation in real world, potential business benefits from improving the model. (These are key to differentiating a good and an excellent solution) |
| Notebook - Overall | * Structure and flow * Well commented code | 1. Well structured notebook with a logical flow 2. Clean and well commented code |