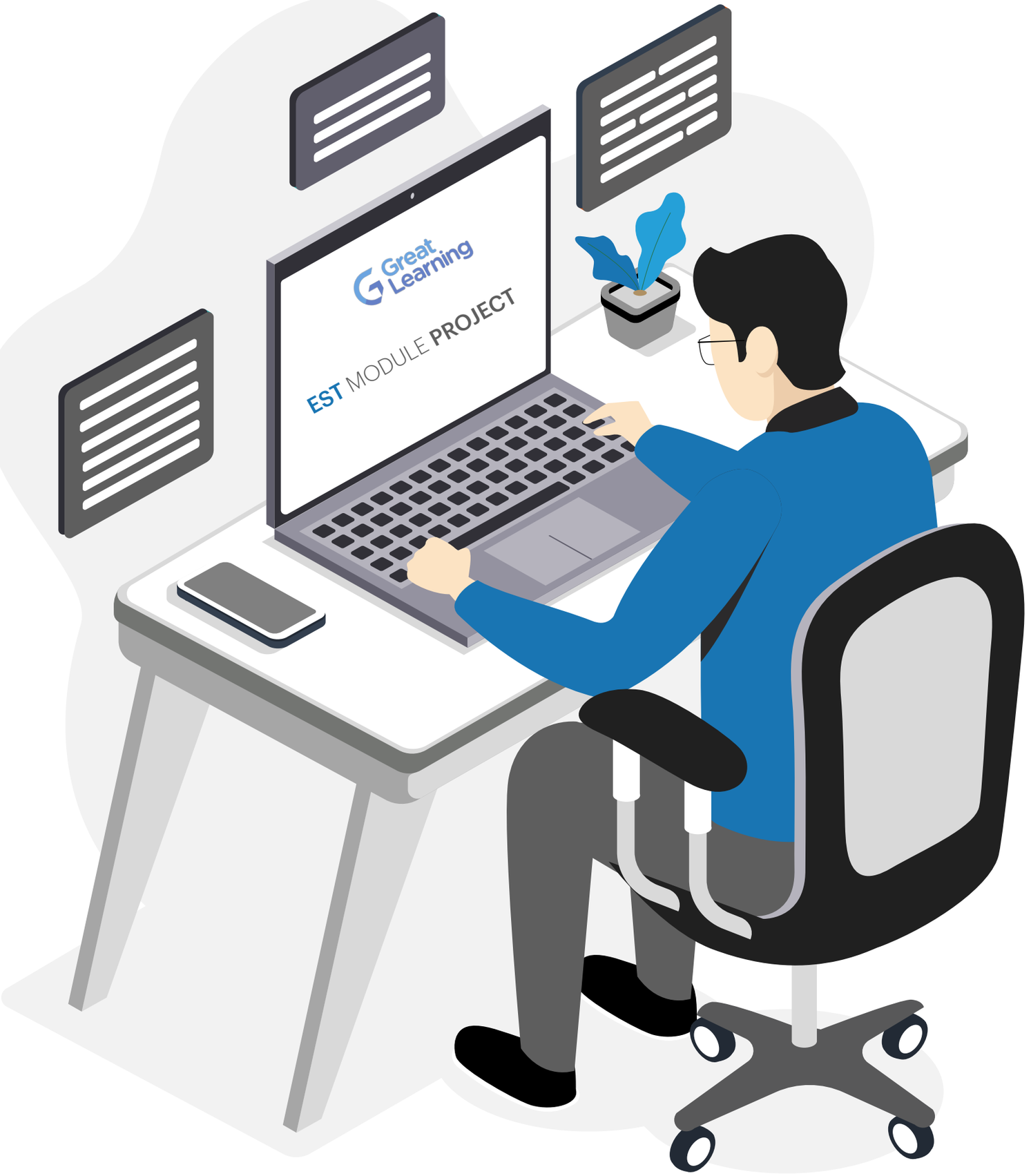

AIML | MODULE PROJECT

Ensemble Techniques

TOTAL SCORE 60

- **DOMAIN :** Travel & Tourism
- **CONTEXT:** You are a Data Scientist for a tourism company named "Visit with us". The Policy Maker of the company wants to enable and establish a viable business model to expand the customer base. A viable business model is a central concept that helps you to understand the existing ways of doing the business and how to change the ways for the benefit of the tourism sector. One of the ways to expand the customer base is to introduce a new offering of packages. Currently, there are 5 types of packages the company is offering - Basic, Standard, Deluxe, Super Deluxe, King. Looking at the data of the last year, we observed that 18% of the customers purchased the packages. However, it was difficult to identify the potential customers because customers were contacted at random without looking at the available information. The company is now planning to launch a new product i.e. Wellness Tourism Package. Wellness Tourism is defined as Travel that allows the traveller to maintain, enhance or kick-start a healthy lifestyle, and support or increase one's sense of well-being. This time the company wants to harness the available data of existing and potential customers to target the right customers. You as a Data Scientist at "Visit with us" travel company have to analyse the customers' data and information to provide recommendations to the Policy Maker and build a model to predict the potential customer who is going to purchase the newly introduced travel package. The model will be built to make predictions before a customer is contacted.

- **DATA DESCRIPTION :**

1. **Customer details**: CustomerID: Unique customer ID
2. **ProdTaken**: Whether the customer has purchased a package or not (0: No, 1: Yes)
3. **Age**: Age of customer
4. **TypeofContact**: How customer was contacted (Company Invited or Self Inquiry)
5. **CityTier**: City tier depends on the development of a city, population, facilities, and living standards. The categories are ordered i.e. Tier 1 > Tier 2 > Tier 3. It's the city the customer lives in.
6. **Occupation**: Occupation of customer
7. **Gender**: Gender of customer
8. **NumberOfPersonVisiting**: Total number of persons planning to take the trip with the customer
9. **PreferredPropertyStar**: Preferred hotel property rating by customer
10. **MaritalStatus**: Marital status of customer
11. **NumberOfTrips**: Average number of trips in a year by customer
12. **Passport**: The customer has a passport or not (0: No, 1: Yes)
13. **OwnCar**: Whether the customers own a car or not (0: No, 1: Yes)
14. **NumberOfChildrenVisiting**: Total number of children with age less than 5 planning to take the trip with the customer
15. **Designation**: Designation of the customer in the current organisation
16. **MonthlyIncome**: Gross monthly income of the customer

Customer interaction data:

1. **PitchSatisfactionScore**: Sales pitch satisfaction score
2. **ProductPitched**: Product pitched by the salesperson
3. **NumberOfFollowups**: Total number of follow-ups has been done by the salesperson after the sales pitch
4. **DurationOfPitch**: Duration of the pitch by a salesperson to the customer

- **PROJECT OBJECTIVE :** To predict which customer is more likely to purchase the newly introduced travel package
- **STEPS AND TASK [60 Marks]:**

1. **Data Understanding & Preparation: [10 Marks]**

   A. Read the 'Tourism.xlsx' dataset and print the first 5 rows. **[1 Mark]**
   B. Report the dimension, datatypes and missing values in the dataframe. **[1 Mark]**
   C. Print the average of monthly income for each Designation. **[1 Mark]**
   D. Check the percentage of missing values in each column and write your observations. **[1 Mark]**
   E. Check the number of unique values in each column. **[1 Mark]**
   F. Drop the 'CustomerID'. **[1 Mark]**
   G. Print a 5-point summary of the dataframe and share your observation. **[1 Mark]**
   H. Print the count of each unique category in each of the categorical variables. **[1 Mark]**
   I. Observe the unexpected values/categories in the categorical variables carefully and impute them with the best approach. **[1 Mark]**
   J. Convert the data type of each categorical variable to 'category'. **[1 Mark]**

2. **Data Exploration & Analysis: [10 Marks]**

   A. Create a copy of the prepared data to perform EDA. **[1 Mark]**
   B. Perform Univariate Analysis on numerical and Categorical data. Share your insights. **[2 Marks]**
   C. Examine the outliers and impute them. **[1 Mark]**
   D. Perform detailed Bivariate and Multivariate Analysis on the data and share your insights. **[2 Marks]**
   E. Group the data w.r.t to packages/products taken by the customers to build customer profiles. Write your observation on their statistical characteristics. [For Basic, Standard, Deluxe, Super Deluxe and King]. **[3 Marks]**
   F. Plot a correlation heatmap on the data and write your observations. **[1 Mark]**

3. **Model building: [35 Marks]**

   A. Split the data into X and Y and Drop the columns of customer interaction data from the dataset. **[1 Mark]**
   (As we aim to predict customers who are more likely to buy the newly introduced products, the new data for prediction will not have the customer interaction details)
   B. Impute the missing values with suitable approach. **[1 Mark]**
   C. Create dummy variables for string type variables and convert other column types back to float. **[1 Mark]**
   D. Split the dataset into train and test sets with 70:30 proportion. **[1 Mark]**
   E. Carry out the following steps for different classifiers such as Decision Tree, Random Forest, Bagging, AdaBoost, Gradient Boosting and Stacking
   [24 Marks, 4 marks for each model]
      i. Build a model and print the possible Performance Metric
      ii. Perform Cost Complexity Pruning for different alpha values and visualize its performance for train and test sets. Choose the best model and share the performance metrics (to be performed only for Decision Tree Classifier).
      iii. Tune the Hyper Parameters and fit the model with best parameters
      iv. Print the classification metrics and write your observations on the performance improvement

   F. Build a XGBoost Classifier and print its performance report. **[5 Marks]**
   G. Compare the performance of all models for train & test set and provide your insights. **[2 Marks]**

4. **Business Recommendation: [5 Marks]**

   A. Provide a detailed and useful business recommendation based on your observations and analysis.