Problem Statement:

Your client is a multi-national financial company, which offers multiple products to the consumers. There are multiple channels to offer these products to consumers although major contribution is coming from offline distribution channel. Offline channel sells Financial products to consumers via their agent network and as per government regulation these agents must be certified to sell financial products. There are multiple certification programs against different categories of financial products.

As this offline channel shares major contribution to total company sales, company **focuses on recruitment and certify them to build large agent network**. Here, **major challenge is training them to get the certifications** to sell various type of products.

You are given a dataset of **trainee performance for the training curriculum test wise within multiple programs**. Your task is to **predict the performance on such tests given the demographic information and training program/test details**. This will enable your client to strengthen its training problem by figuring out the most important factors that lead to a better engagement and performance for a trainee.

Data Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Cleaning Requirements** | **Remarks** |
| id | Unique ID |  | To be removed from the training data |
| program\_id | ID for program | Duplicate source. | Concatenate with test id. |
| program\_type | Type of program | Encoding required | Encode it by mean encoding method |
| program\_duration | Program duration in days |  | Scaling is not required as we use decision trees |
| test\_id | test ID | Duplicate source. | Drop the variable |
| test\_type | Type of test (offline/online) | Encode |  |
| difficulty\_level | Difficulty level of test | Encode by dummies or experiment with ordinal encoding | Dummies |
| trainee\_id | ID for trainee | Duplicate source | Drop the variable as it is part of ID |
| gender | Gender of trainee | Encode |  |
| education | Education Level of trainee | Ordinal Encode | Data given in the form of words, so we know the hierarchy. |
| city\_tier | Tier of city of residence for  trainee | So proceed with dummy variables. | Don’t know the hierarchy of the city tier. |
| age | Age of trainee | Fill null values by -1 | Type of model can be chosen as per the cleaning action |
| total\_programs\_enrolled | Total Programs Enrolled by trainee |  |  |
| is\_handicapped | Does trainee suffer from a disability? | Encode by replace |  |
| trainee\_engagement\_rating | Instructer/teaching assistant provided trainee engagement rating for the course |  | Null values available for 77 records |
| is\_pass | 0 - test failed, 1 -  test passed | No further cleaning required |  |

Basic Observations from exploratory analysis:

1. ID column is a concatenation of trainee\_id and test\_id.
2. Program\_type is the first character of the program\_id. So it has no significant information.
3. All Handicapped candidates have failed.
4. Each program has different tests mapped.

Tasks:

1. Verify if first char of program ID == program type.
2. Verify if trainee ID is first few digits of ID.

Feature Eng Ideas that can be explored on the top of the existing data:

1. Combine gender with handicapped as a single variable.
2. Impute trainee engagement rating based on the average of the test\_type, difficulty\_level , gender, education.
3. Impute age based on the average of the program\_test\_id