**Kaggle Project – 3: Playground Series 3 Episode 3**

**Classification with Employee Attrition Dataset**

**Prepared by: Dwaipayan Mukherjee (2211569)**

**About the Project:**

This project uses data synthetically generated by a deep learning model trained on the IBM HR Analytics Employee Attrition & Performance dataset. The aim is to predict employee attrition (i.e., whether an employee will quit or not) based on 33 variables:

**Age, BusinessTravel, DailyRate, Department, DistanceFromHome, Education, EducationField, EmployeeCount, EnvironmentalSatisfaction, Gender, HourlyRate, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, MonthlyRate, NumCompaniesWorked, Over18, OverTime, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StrandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager**.

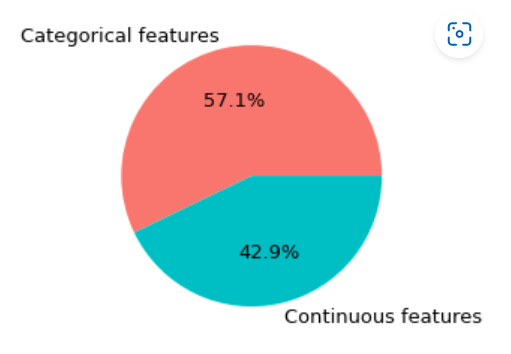
The **id** column is non-overlapping between train and test set, and is meant for data organization purposes. Thus, it likely carries no signal.

The target **Attrition** is a binary column.

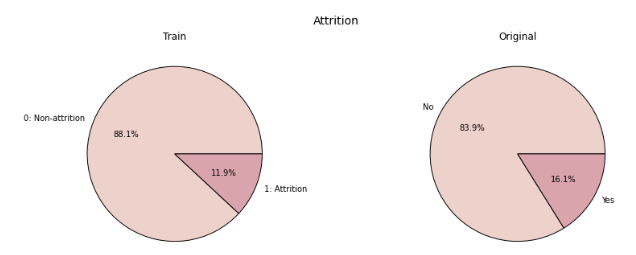
The evaluation metric for this project has been selected to be the **area under the ROC curve** between the predicted probability and the observed target.

**Steps that I have followed through the project:**

1. First I have imported the necessary libraries.
2. I have loaded the **train**, **test**, **submission** and **original** data files as Pandas dataframes.
3. Next, I have tried to understand the dataset:
4. I have checked the heads of the datasets.
5. Next, I have checked for missing values in any of the columns in the **train**, **test** and **original** datasets. There are no missing values in any of the datasets.
6. Next, I have gone for the summary statistics of the columns.
7. I have printed the Unique value counts of all the columns for all three datasets.
8. I have separated out the categorical and continuous columns in the dataset for Exploratory Data Analysis. I found that the dataset contains 57.1% categorical features and 42.9% continuous features:



1. I have printed the Unique Values of the categorical variables.
2. Lastly, I saw that the distribution of the target variable **Attrition** was highly imbalanced in the data. This is further shown in the EDA section.
3. **Exploratory Data Analysis:**
4. First, I have explored the distribution of the target variable **Attrition** in the **train**  and **original** set. In the train set, percentage of **Attrition** is 11.9%, while that in the original set is 16.1%.So, the train data is more or less representative of the original dataset. Moreover, I find that the dataset is highly imbalanced. To take care of this imbalance in the data, I will use Stratified K-fold CV while applying any model on the training data:



1. Next, I have explored the distribution of the categorical variables in the datasets and the likelihood of attrition based on the train data.
2. A major inference from this EDA is that all the categorical variables except three, namely, **EmployeeCount**, **Over18** and **StandardHours** influences the likelihood of attrition. Thus, these three variables provide no signal and hence can be dropped.
3. Furthermore, some of the classes of some categorical variables show high likelihood of attrition but high variance as well since there are only a few observations pertaining to that particular class.
4. Some categorical variables like **BusinessTravel**, **Education**, **EnvironmentSatisfaction**, **Gender**, **JobInvolvement**, **MaritalStatus** and **OverTime** shows marked variance in likelihood of attrition across their classes and so, these variables might be mighty important in determining the attrition probability of an employee.
5. Next, I have tried to find the distribution of the continuous variables in the datasets and the distribution of **Attrition** with respect to them.
6. The inferences that can be drawn from this EDA is firstly that the distribution of the continuous variables in the train and test data follow the distribution of the same variables in the original dataset.
7. Next, none of the continuous variables show much difference in distribution with respect to the attrition variable. Feature engineering might bring about discernible patterns in the distributions with respect to attrition.
8. **Feature Engineering:**
9. The following new features have been created:
10. `**MonthlyIncome/Age**`: a continuous variable measuring the monthly income over age.
11. `**Age\_risk**`: a binary variable which is 1 if age is less than 34. There is supposed to be a difference in the likelihood of attrition between younger and older folks.
12. `**HourlyRate\_risk**`: a binary variable which is 1 if hourly rate is less than 60. An hourly rate less than 60 may be a sufficient reason for employees to leave.
13. `**Distance\_risk**`: a binary variable which is 1 if distance from home is greater than or equal to 20. A distance greater than 20 miles between home and office may be a reasonable reason for employees to leave and seek employment somewhere nearer.
14. `**YearsAtCo\_risk**`: a binary variable which is 1 if years in company is less than 4 years. An employee who is a newcomer to a company might be more likely to leave than an experienced veteran of the company.
15. `**AverageTenure**`: a continuous variable obtained from dividing total working years by the number of companies worked.
16. `**JobHopper**`: a binary variable which is 1 if an employee has worked in more than 2 companies and if his/her average tenure in a company has been less than 2 years. It is a reasonable assumption to think such a person might change jobs frequently.
17. `**Attrition\_risk**`: a combined risk measure made of `Age\_risk`, `HourlyRate\_risk`, `Distance\_risk`, `YearsAtCo\_risk` and `JobHopper`.
18. I have created a function called **feature\_eng()** to apply the following feature engineering onto the train and the test set.
19. I redid the EDA on the newly created features:
20. For the categorical (newly engineered) features, the patterns are much more discernible and follow the initial assumptions on which they were engineered.
21. However, in case of the continuous variables, the median values for **Attrition** = 1 and **Attrition** = 0 are a bit different, though not much significant.
22. **Data Preparation:**
23. First, I have encoded the categorical features which were in the string format. I have used Ordinal Encoder from sci-kit learn library. I have used Ordinal Encoder rather than one-hot encoding because that would lead to a huge number of features and hence might have led to overfitting.
24. Next, I dropped the unnecessary variables from the train and test data. These variables were **id**, **EmployeeCount**, **Over18** and **StandardHours**. The three of the variables other than **id** had observation only for one class and hence would not provide a signal for Attrition.
25. Next, I have separated the features and the target variables.
26. Lastly, I have scaled the features using standard scaler.
27. **Modelling:**
28. First, I have created a function called **cross\_validate()** that takes in the features data as a Pandas dataframe, the series of labels and the model that I would like to fit. Then the function fits the model with a 10-fold Stratified Cross Validation, predicts and gives the fold-wise as well as average AUC score for the model. It also prints the feature importance.
29. **XGBoost Classifier**:
30. First, I have tried to tune and find out the best set of hyperparameters for an XGBoost Classifier model using GridSearchCV.
31. I have trained the model using the parameters *n\_estimators* as 100, *max\_depth* as 3, *learning\_rate* as 0.1, *min\_child\_weight* as 5, *subsample* as 0.7 and *colsample\_bytree* as 0.25, as advised by the GridSearchCV algorithm.
32. The **Average AUC** is found to be **0.8159**. The features with highest importance as **StockOptionLevel**, **OverTime**, **JobInvolvement**, **AttritionRisk** and **Age\_risk**.
33. **LightGBM Classifier**:
34. I have used the LGBM Classifier next with hyperparameters: *n\_estimators* as 400, *num\_rounds* as 275, *learning\_rate* as 0.1, *num\_leaves* as 200, *max\_depth* as 9, *min\_data\_in\_leaf* as 45, *lambda\_l1* as 0.01, *lambda\_l2* as 0.6, *min\_gain\_to\_split* as 1.40, *bagging\_fraction* as 0.45 and *feature\_fraction* as 0.3.
35. The **Average AUC** came out to be **0.8280,** with the most important features as **MonthlyRate**, **MonthlyIncome/Age**, **Age**, **MonthlyIncome**, **AverageTenure** and **PercentSalaryHike** to name a few.
36. **CatBoost Classifier**:
37. I have used a CatBoost Classifier with the hyperparameters: *loss\_function* as CrossEntropy, *learning\_rate* as 0.1, *l2\_leaf\_reg* as 0.01, *colsample\_bylevel* as 0.05, *depth* as 1, *boosting\_type* as Plain, *bootstrap\_type* as Bernoulli, *min\_data\_in\_leaf* as 20, *one\_hot\_max\_size* as 15 and *subsample* as 0.7.
38. The **Average AUC** came out to be **0.8488** and the most important features came out as **StockOptionLevel**, **OverTime**, **BusinessTravel**, **YearsWithCurrManager** and **MaritalStatus**.
39. **Random Forest Classifier**:
40. I have used a Random Forest Classifier with the hyperparameters: *n\_estimators* as 250, *max\_depth* as 9, *min\_samples\_split* as 25 and *bootstrap* as True.
41. The **Average AUC** came out to be **0.8225**, and the features with most importance were **StockOptionLevel**, **MonthlyIncome**, **MonthlyIncome/Age**, **AverageTenure** and **OverTime**.
42. **Final Submission**: Since the CatBoost Classifier gave the best results in training, I have chosen it for the final predictions on the **test** dataset.