Info : In this project I have used “WE” rather than ‘I’ in explanations.

PROJECT NAME: Employee Performance Analysis (Code: 10281)

Client Info Ansd Requirement:

INX Future Inc, (referred as INX), is one of the leading data analytics and automation solutions provider with over 15 years of global business presence. INX is consistently rated as top 20 best employers past 5 years. Recent years, the employee performance indexes are not healthy and this is becoming a growing concern among the top management. There has been increased escalations on service delivery and client satisfaction levels came down by 8 percentage points. CEO, Mr. Brain, knows the issues but concerned to take any actions in penalizing non-performing employees as this would affect the employee morale of all the employees in general and may further reduce the performance. Mr. Brain decided to initiate a data science project, which analyses the current employee data and find the core underlying causes of this performance issues. Mr. Brain, being a data scientist himself, expects the findings of this project will help him to take right course of actions. He also expects a clear indicator of non-performing employees, so that any penalization of non-performing employee, if required, may not significantly affect other employee morals.

The following insights are expected from this project:

- Department wise performances.
- Top 3 important factors effecting employee performance.
- A trained model which can predict the employee performance based on factors as inputs.
- Recommendations to improve the employee performance based on insights from analysis.

CREATING ML MODEL

1. IMPORTING LIBRARIES

Libraries used: Pandas, Numpy, Matplotlib, Seaborn, LabelEncoder, BinaryEncoder, scipy, pylab, SMOTE, XGBClassifier, LogisticRegression, RandomForestClassifier, DecisionTreeClassifier, CatBoostClassifier, RandomizedSearchCV.

2. IMPORTING DATASET

Importing required dataset using pd.read_excel .

3. EDA

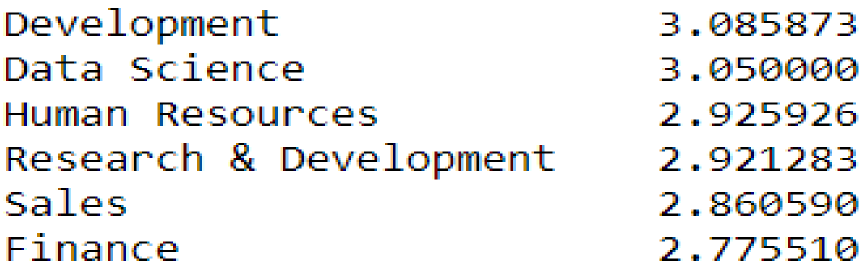
-> Overview of the data using Pandas Profiling.

-> Count plot of Age with respect to performance rating. From this it is clear that majority of the people have given performance rating as 3. People with age 34 having higher performance rating.

-> Count plot of Gender with respect to performance rating. 72.9% of people is having performance rating as 3. In that 43.8% are males and 29.1% are females. The number of males in the company is higher than females.

-> Analysis of **Department wise performance analysis.**

We have obtained department wise performance by grouping EmpDepartment with respect to PerformanceRating.



```
Development                 3.085873
Data Science                3.050000
Human Resources             2.925926
Research & Development      2.921283
Sales                       2.860590
Finance                     2.775510
```

-> Count plot of EmpLastSalaryHikePercent with respect to PerformanceRating.

-> Count plot of EmpEnvironmentSatisfaction with respect to PerformanceRating.

4. ENCODING CATEGORICAL COLUMNS

We need to convert all the features with object type values. The features which should encode are EmpNumber, Gender, EducationBackground, MartialStatus, EmpDepartment, EmpJobRole, BusinessTravelFrequency, OverTime, Attrition.

We have used One Hot Encoding Technique for nominal feature and Target Guided Encoding for Ordinal data.

5. FEATURE ENGINEEING

In this stage we will be removing features which are not important for modelling.

1. We will be dropping EmpNumber as it is an id.
2. By using VARIENCE THRESHOLD we will be removing those values which are having very low threshold between the values present in the same feature.
3. Checking for Multicollinearity. Here we will be removing the independent features which are highly correlated to each other. The feature which is dropped using this method is EmpJobRole.
4. Finding correlation between the target variable and the independent variable. We can remove those features which is having low correlation with the target variable. Because those features with low correlation will not contribute for prediction. So, we can remove that. The features with low correlations are:

   Age, Attrition, DistanceFromHome, EmpEducationLevel, EmpHourlyRate, EmpJobInvolvement,EmpJobLevel,EmpJobSatisfaction,EmpRelationshipSatisfaction, Gender,LifeSciences, Marketing, Married, Medical, NumCompaniesWorked, OverTime,Single,TotalWorkExperienceInYears,TrainingTimesLastYear, Travel_Frequently, Travel_Rarely

6. DATA NORMALIZATION

In this step we check whether the given data lies under the gaussian distribution. If not, we have to make it as a gaussian distribution. While looking into QQ plot and distplot we can classify continuous and discrete feature. We only need to normalize the continuous features, if it is not normally distributed.

Here the continuous features are ExperienceYearsAtThisCompany, ExperienceYearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager. After plotting we found that ExperienceAtThisCompany and YearsSinceLastPromotion is not uniformly distributed.

We have used Log () transformation method for normalising these 2 features.

7. OUTLIER REMOVAL

In this step we will be removing the outliers present in the continuous features. Outliers are identified by using the boxplot. The features with outliers are ExperienceYearsInCurrentRole, YearsWithCurrentManager.

Here we are using Inter Quartile Range (IQR) to remove the Outliers.

8. DEFINING INDEPENDENT AND DEPENDENT FEATURES

Here we are defining independent features as x and dependent feature(target) as y.

9. SPLITTING THE DATA

In this step we split the data into test and train using Train_Test_Split from sklearn.

10. BALANCING THE IMBALANCED DATASET

By using value_counts() function we found that our model is not balanced. We have to balance this data before deploying into the model. Here we are using over sampling technique for balancing the data. SMOTE is an over sampling technique from imblearn.

11. MODEL SELECTION

-> Modelling using **Logistic_Regession** gives an accuracy of 73.3%. On closer look at the precision and f1-score this model is poor in classification. So, we cannot proceed with this model hence we need to introduce a new model.

-> Modelling using **Decision Tree Classifier** gives an accuracy of 89.7%. This model is far better than the previous model in case of precision and f1-score. By using a single tree, we are getting a better result. So, by trying advanced models we will be able to attain more accuracy.

-> Modelling with **XGBoost Classifier** gives an accuracy of 93.3%. It is also having good precision and f1-score. Hyper parameter tuning is done so that we can improve the accuracy along with the precision and f1-score. K-fold cross validation is done so that we can confirm that out model is not overfitting. After hyper parameter tuning we can see an increased accuracy of 94.12%.

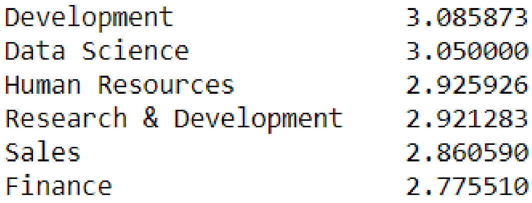
-> Modelling using **Random Forest Classifier** gives an initial accuracy of 94.7%. On hyper parameter tuning it gives an increased accuracy of 95.3% and also better precision and f1-score. K-fold cross validation is done in order to find that whether the model is overfitting or not. The k-fold values are close to the predicted value. Hence, we can conclude that our model is not over fitting. And this model can be considered because it is having good accuracy along with good precision and f1-score.

-> Modelling with **Cat Boost Classifier** gives an accuracy of 93.05%. Hence it is lower than that of XgBoost and Random Forest this model can be neglected.

12. CONCLUSION

Random Forest Classifier gives an accuracy of 95.3%. So according to perspective this model is good while considering the client requirements.

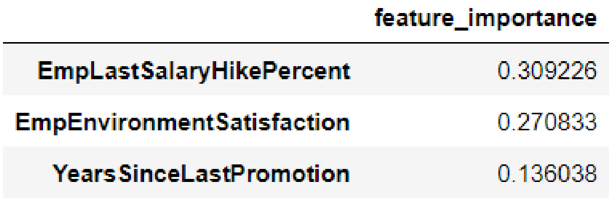
-> DEPARTMENT WISE PERFORMANCE:



```
Development               3.085873
Data Science              3.050000
Human Resources           2.925926
Research & Development    2.921283
Sales                     2.860590
Finance                   2.775510
```

This obtained by grouping EmpDepartment with the PerformanceRating. By doing this we found the department with weak performance rating. Finance is the weakest among the other departments. R&D and sales are also having poor performance. The company is requested to take necessary actions so that these departments can increase their performance.

-> TOP 3 FACTORS EFFECTING EMPLOYEE PERFORMANCE



	feature_importance
EmpLastSalaryHikePercent	0.309226
EmpEnvironmentSatisfaction	0.270833
YearsSinceLastPromotion	0.136038

* Random Forest gives an accuracy of 95.3%. By using feature importance, we have found top 3 features that affect the employee performance. The top 3 features are mentioned above.

* EmpLastSalaryHikePercent is positively correlated to the performance rating. That is when the when the salary hike percentage increases this will gradually leads to an increase in the performance rating.

* Employee environment satisfaction also affects the performance rating. When the company provides a good working environment the employees will be able to work properly without any distraction, this will lead to an increased performance rating.

* Years since last promotion is negatively correlated with the performance rating. When the number of years since last promotion increases that will definitely leads to a decrease in the performance rating.

-> ANALYSIS AND INSIGHTS

* We have tried 5 different models. The accuracies obtained are:

1. Accuracy of Logistic Regression = 73.3%
2. Accuracy of Decision Tree = 89.7%
3. Accuracy of XgBoost = 94.2%
4. Accuracy of Random Forest = 95.3%
5. Accuracy of CatBoost = 90.05%

Random forest and Xgboost is performing well. But on closer observation in precision and f1 score we can find that Random forest is good with classification than xgboost and also has an improved accuracy than the XgBoost. So Random Forest can be considered for client requirement.

* The important features that are positively correlated are Last salary hike percent and Employee Environment Satisfaction. This means when these factors increase, performance rating also increases. On the other hand, Years Since Last Promotion, Experience Years in current role, Experience Years at this company and Years with Current Manager are negatively correlated. That means when these factors increase, Performance rating decreases.

-> SUGGESTIONS

* The company need to provide better environment condition so that the employees can work efficiently without any distractions, which will gradually lead to an increased performance rating.

* The company should increase the salary of the employee according to their experience, area of work etc... without any failure.

* The company can create a separate department to monitor the functioning of all other departments. This monitoring department should contain at least one member from each other departments. So, they can monitor the performance rating of each departments separately and can deal with the departments with poor performance rating.

* Shuffling the managers from time to time will affect the performance. So, such kind of situations should be avoided.

* Promotions are required for employee in a department for better performance.

* Data Science department can hire more employee with necessary skill.

Questions from pdf PROJECT SUBMISSION GUIDELINES:

-> FEATURE SELECTION / ENGINEERING

1. What were the most important features selected for analysis and why?

EmpDepartment,

EmpEnvironmentSatisfaction,

EmpLastSalaryHikePercent, EmpWorkLifeBalance,

ExperienceYearsAtThisCompany,

ExperienceYearsInCurrentRole,

YearsSinceLastPromotion,

YearsWithCurrManager,

PerformanceRating

These are the most important feature selected for analysis because these features are highly correlated with the target variable.

2. Did you make any important feature transformation?

Yes, I have used log transformation for normalizing 2 continuous features.

-> ExperienceYearsAtThisCompany

-> YearsSinceLastPromotion

3.Correlation or interactions among the features selected and how it is considered?

While considering the correlation first I take the correlation between the independent features and consider a cut-off of 0.8. If the features exceed the cut-off then that feature can be considered as duplicate value hence we can drop that feature. Similarly, in the case of correlation between the target variable and the input features the cut-off is 0.1. Here the features which is having less correlation than the cut-off can be removed as those features are not important for model preparation.

-> RESULTS, ANALYSIS AND INSIGHTS

1. Did you find any interesting relationships in the data that don't fit in the sections above?

Yes, our target is to find the employee performance and our data set consist of physical data but I think that while considering the performance of employee we need to find his/her mental performance. Here, there is no features which explains the employee mental strength.

2. What is the most important technique you used in this project?

Target guided encoding, SMOTE, Correlation, IQR, Randomized CV search

3. Provide clear answers to the business problem mentioned in the project on basis of analysis?

* The company need to provide a better environment for the employee. they want to advance their technology to give a better environment and a suitable condition for the working of employees. Provide insurance, accommodations, food, basic needs etc...

* The company need to take care of their employee salary. needs to give salary hike according to their experience, area of work etc. without any failure.

* company need to ensure the department wise performance and ensure that there is no lacking in the requirements of the employee in department wise. need to provide some curriculum and courses to encourage the employees as well as it will increase the performance.

* The company can create a separate department to monitor the functioning of all other departments. This monitoring department should contain at least one member from each other departments. So, they can monitor the performance rating of each departments separately and can deal with the departments with poor performance rating.

* Shuffling the managers from time to time will affect the performance. So, such kind of situations should be avoided.