information :- in this project I have used "WE" rather than "I" in comments because it helps better to uderstand.

PROJECT :

Employee Performance Analysis(Code: 10281)

Client Info and 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 concerns 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 indicators of non performing employees, so that any penalization of non-performing employee, if required, may not significantly affect other employee morals.

Insights :

1. we need to find out department wise perfomance it will help in modelling and feature selection

2. we need to find out Top 3 Important Factors effecting employee performance by using corrillation (its a Client Requirement)

3. A trained model which can predict the employee performance based on factors as inputs. This will be used to hire employees

4. Recommendations to improve the employee performance based on insights from analysis(Client Requirement)

Creating The ML Model

STEP 1: Import Necessary Packages

Packages Used: pandas, Numpy,pylab,Sklearn,Matplotlib,Seaborn,Scipy,Imblearn,

Collections,Pickle

STEP 2: Import DataSet

importing required dataset from the current directry using pandas read_csv function

STEP 3: Exploratory Data Analysis

-> Checking weather the data is imported correctly or not

-> Checking the shape of the data

-> Checking is there any Nan value present in the data_set

Encoding The Feature:

We need to encode those features which has object type values for checking which are those features we use select_dtype function

the list of features which we need to encode are:

EmpNumber,Gender,EducationBackground,MaritalStatus,EmpDepartment,EmpJobRole, BusinessTravelFrequency,OverTime,Attrition.

we have used One Hot Encoding Technique for nominal features and Target Guided Encoding Technique for Ordinal Data.

STEP 4: Feature Reduction

By using this technique we remove the features which are not important/required for modelling:

1. Dropping features with no impact in target. in our data set "EmpNumber" is an id hence it is different for every data points so we can drop the feature directly.

2. Finding VARIENCE THRESHOLD.This method is used to remove those value's which are having a very low threshold between the values present in the same feature. here considering the threshold value as 0.1 the features we dropped using this technique from our data set are 'Other', 'Technical Degree'.

3. Finding correlation to drop features with highly corrilated to each other(Independent features).here we are considering 0.8 as our cutoff and if any features exeeds the cutoff then we can say that they are duplicates of other features so we can remove those duplicates from our data. the feature we have dropped using this method are 'EmpJobRole'.

4. Finding correlation to drop features who have low corrilation with Target (Dependent features).here we are considering 0.1 as our cutoff and if any features below the cutoff then we can say that they are having a low corrilation with the Target feature. so we can remove those duplicates from our data. the features we have dropped using this method are:

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

STEP 5: Data Normalization

To find weather the given data points lies under the gausiion distribution . If not we need to make make it as a guassian distribution.while considering QQ_plot and distplot we can classify both continues and discreate feature. We need to normalise the cotinues features if it is not under normal_distribution.

here the continues features are :

EmpLastSalaryHikePercent,ExperienceYearsAtThisCompany,ExperienceYearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager

From the above figure after analysing we can say that only 2 feature are not satisfing normal_distribution so we need to handle those features:

ExperienceYearsAtThisCompany, YearsSinceLastPromotion

for normalizing the two feature we use Log() Transformation Technique

STEP 6: Handling Outliers

If there is any outliers present while considering distance based model it will creates problems. so, we need to remove those outliers. Outliers need to be removed from continues features.By using boxplot we can easily find out weather outliers present in the feature or not. While considering the Boxplot its clear that there are Outliers present in our feature. hence we need to remove/handle those outliers.

The features having Outliers are :

ExperienceYearsAtThisCompany,YearsSinceLastPromotion,YearsWithCurrManager.

Here we are using Inter Quantile Range Method(IQR) to remove the Outliers

STEP 7: Defining Indipendent and Dependent Features

Here varaible X denotes the Independent Features and Y denotes the Target/Dependent variable. In our data set all the values in the features lies under 0 to 25 hence we doesn't need scailing we can directly go to train test split.

STEP 8: Splitting the Data

For splitting the data set as train and test we use Train_Test_Split from sklearn

STEP 9: Balancing the Data

we can use count plot to find weather the target variable is balaced or not. Our data is not balanced hence we need to balance the data before modelling. Here we are using upsampling technique for balancing the data. SMOTE is an upsampling technique from inbalanced learn mainly used for balancing the data here the technique creates new data points by taking the cetroid of k number of points.

STEP 10: Modelling

-> Modelling using Logistic_Regression gets 75 percent accuracy.Here while considering category 2 : accuracy is only around 50% so we cannot go with this model hence we need to indroduce new model.

-> Modelling using K-Nearest Neibour(KNN) gets 75percent accuracy. to find better K value we can use error curve graph.we get that global minima for k value is 2, hence we remodel the knn using k value as 2. Hence we apply k value as 2 and gets accuracy 77 percent.

-> Modelling using Decision Tree.while comparing the logistic and knn model decision tree works better and gives 89% accuracy so using one single tree we gets good percentage if we use ensemble technique we surely get 90% + accuracy so considering one more model.

-> Modelling using Random Forest gets 95 percent accuracy. we are getting a high accuracy while using random forest. so we need to check weather the model is overfitted or not. for that we use cross_validation_score as well as we need check the classification report.

-> Hyper Parameter Tuning Using GRID Search CV and finded the best parameters and remodelled the Randomforest using those values and we got 96 percent accuracy.

STEP 11: Saving The Model As Pickle File

For exporting the model in future we need a pickle file so for that we are saving the current model in a pickle file in the Binary Format

Conclussion:

Hence finished the project with an accuracy score of approximate 96% I think its a better accuracy while considering the client requiremnts.

->This is the department wise perfomance rating of the company.so by using this data the company can easily identify the week department and give them some guidence:

Development : 3.085873

Data Science : 3.050000

Human Resources : 2.925926

Research & Development : 2.921283

Sales : 2.860590

Finance : 2.775510

->Hence we find out the Top 3 Important Factors effecting employee performance are :

Here we have used the corrilation technique we find out the corrilation of the independent features with the target feature(depend) and taken the three features which have high corrilation with the target feature.

1. EmpEnvironmentSatisfaction

2. EmpLastSalaryHikePercent

3. EmpDepartment

Suggesions:

-> 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,accomadation,food,basic needs etc..

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

-> company need to ensure the department wise perfomance and ensure that there is no lacking in the requirents of the employee in department wise. need to provide some carriculam and courses to encorage the employees as well as it will increase the perfomance.

Questions from pdf PROJECT SUBMISSION GUIDELINES:

-> FEATURES 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 importent features that selected for analysis because these features are highly corrilated with target feature hence we cannot avoid these.

2. Did you make any important feature transformations?

* yes, I have used LOG Transformation for making two features into a normal distribution curve.

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

* while considering corrilation I take first corrilation between the indepedent features and consider a cutoff of 0.8 if the features exeeds the cuttoff then that feature is a duplicate of another one hence we can drop the feature. simularly in the case of corrilation between the target variable and input features the cut of is 0.1 here if the features which are having corrilation less than the cuttoff then those features are low corrilated with target hence we drop the feature.

-> 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 finding the employee perfomance and our data set consist of physical data but i think that while considering the perfomance of employee we need to find his/her mental perfomance here we no features which explains the employee

mental stength.

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

* Target Guided Encoding,SMOTE,Corrilation,IQR

3. Provide clear answers to the business problems 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,accomadation,food,basic needs etc..

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

-> company need to ensure the department wise perfomance and ensure that there is no lacking in the requirents of the employee in department wise. need to provide some carriculam and courses to encorage the employees as well as it will increase the perfomance.