2nd SEMESTER

HR ANALYTICS – ASSIGNMENT: 2

Submitted on: **29-05-2022**

Submitted to: **Dr. Ashutosh Murti**

HR Analytics at SCALENEWORKS - Behavioral Modeling to predict renege

Submitted by:

ANUPAM SINHA – M2021ANLT005

Solutions to Questions:

Sol 1: The various activities performed in the analytics project lifecycle are:

1. **Identifying the end goal and motive of the project**.
   This involves defining the business problem, generating the first hypothesis to be tested with data later, and starting data learning.
2. **Preparation of Data**.
   This phase involves Data acquisition, Data entry and filtering, and rescaling.
3. **Planning and choosing the model:**
   This phase involves extracting, loading, and transforming the dataset.
4. **Building the model using tools**
   We can use any tool like python libraries, R, or Weka.
5. **Communication of the result**:
   This phase analyses whether our result is a success or failure.
6. **Documentation and Operationalize the model**.
   The data from the sandbox is moved and run in a live environment throughout this process.

Challenges pertinent to data faced while executing analytics project are:

1. Handling large volumes of data.
2. Fixing data-related issues like combining data collected from various sources.
3. Uncertainty in data due to error.
4. Data parameters need to be converted into the same scale
5. Use case understanding of the project should be simple.

Sol 2 ***Steps for Developing Logistic Regression:***

Step 1. Import libraries and packages:

```
#Importing librarires

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import train_test_split

from sklearn.metrics import confusion_matrix

from sklearn import datasets

from sklearn.metrics import accuracy_score

from sklearn.metrics import classification_report

from sklearn.metrics import confusion_matrix

from sklearn.linear_model import LogisticRegression
```

from sklearn.preprocessing import LabelEncoder

Step 2. Load datasets

attrdata = pd.read_csv("D:/ANUPAM/study/sem2/HR analytics/individual assign/hr_dataset.csv")

attrdata - DataFrame

Index	Candidate Ref	DOJ Extended	ion to accept	Notice period	Offered band	hike expected	hike offered	ent difference	Joining Bonus	date relocate	Gender	ndidate Sour	Rex in Yrs	LOB	Location	Age	Status
0	2110407	1	18	30	2	-20.79	13.16	42.86	0	0	0	0	7	6	9	34	0
1	2112635	0	22	30	2	50	320	180	0	0	1	2	8	9	2	34	0
2	2112838	0	7	45	2	42.84	42.84	0	0	0	1	0	4	9	9	27	0
3	2115021	0	30	30	2	42.84	42.84	0	0	0	1	2	4	9	9	34	0
4	2115125	1	5	120	2	42.59	42.59	0	0	1	1	2	6	9	9	34	0
5	2117167	1	21	30	1	42.83	42.83	0	0	0	1	2	2	9	9	34	0
6	2119124	1	41	30	2	31.58	31.58	0	0	0	1	2	7	9	9	32	0
7	2121918	0	13	45	2	40	208.64	120.45	0	0	1	2	4	9	9	34	1

Step 3. Checking for missing values.

```
[5 rows x 17 columns]
Candidate Ref                    0
DOJ Extended                     0
Duration to accept offer      2719
Notice period                    0
Offered band                     0
Pecent hike expected in CTC    747
Percent hike offered in CTC    596
Percent difference CTC         851
Joining Bonus                    0
Candidate relocate actual        0
Gender                           0
Candidate Source                 0
Rex in Yrs                       0
LOB                              0
Location                         0
Age                              0
Status                           0
```

Step 4. pre-processing of Data:

Replacing missing value with median values of that column.

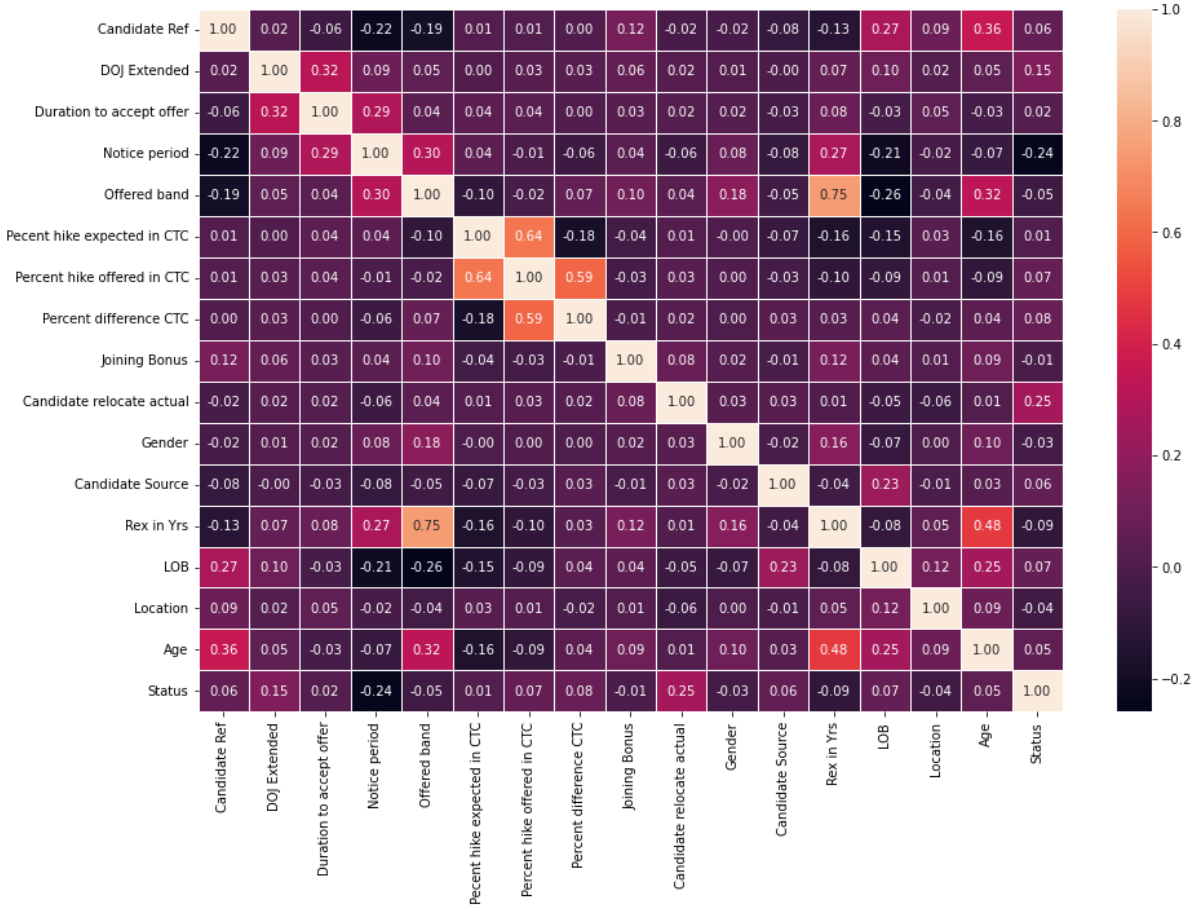
attrdata['Duration to accept offer'] = attrdata['Duration to accept offer'].fillna(attrdata['Duration to accept offer'].median())

attrdata['Pecent hike expected in CTC'] = attrdata['Pecent hike expected in CTC'].fillna(attrdata['Pecent hike expected in CTC'].median())

attrdata['Percent hike offered in CTC'] = attrdata['Percent hike offered in CTC'].fillna(attrdata['Percent hike offered in CTC'].median())

attrdata['Percent difference CTC'] = attrdata['Percent difference CTC'].fillna(attrdata['Percent difference CTC'].median())

Step 5. Generating correlation matrix:



It is used to determine the correlation between the attributes which are dependent on each other.

The dependent attributes add similar types of impact in calculations and so keeping both of them is not worth it. Therefore, we remove one of the two dependent attributes. It helps in saving space and time in calculations.

Here, Rex in years and offered band is highly correlated so, we will remove one of them.

A similar procedure is applied, is to other attributes.

Step 6. Removing non-relevant variables:

These columns are non-relevant- 'Percent hike offered in CTC', 'Rex in Yrs', 'Candidate Ref'.

7. Label Encoding:

```
# Encoding
attrdata['DOJ Extended'] = encoder.fit_transform(attrdata['DOJ Extended'])
attrdata['Offered band'] = encoder.fit_transform(attrdata['Offered band'])
attrdata['Joining Bonus'] = encoder.fit_transform(attrdata['Joining Bonus'])
attrdata['Candidate relocate actual'] = encoder.fit_transform(attrdata['Candidate relocate actual'])
attrdata['Gender'] = encoder.fit_transform(attrdata['Gender'])
attrdata['Candidate Source'] = encoder.fit_transform(attrdata['Candidate Source'])
```

attrdata['LOB'] = encoder.fit_transform(attrdata['LOB'])

attrdata['Location'] = encoder.fit_transform(attrdata['Location'])

attrdata['Status'] = encoder.fit_transform(attrdata['Status'])

attrdata['Duration to accept offer'] = encoder.fit_transform(attrdata['Duration to accept offer'])

The above encoded are also the important factors for renage.

8. Splitting data into train and test dataset:

#scaled_features_df = pd.read_csv("D:/ANUPAM/study/sem2/HR analytics/individual assign/processed table1.csv")

X = attrdata.drop(columns=['Status'], axis=1)

Y = attrdata['Status']

#Splitting data – Train test split

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3, random_state=12344)

9. Logistic Regression Model:

model = LogisticRegression()

print(model.fit(X_train, Y_train))

print(model.score(X_test, Y_test))

Model Score – R^2 value – 0.708

```
LogisticRegression()
0.7086486486486486
```

10. Confusion matrix

	0	1
0	2622	0
1	1078	0

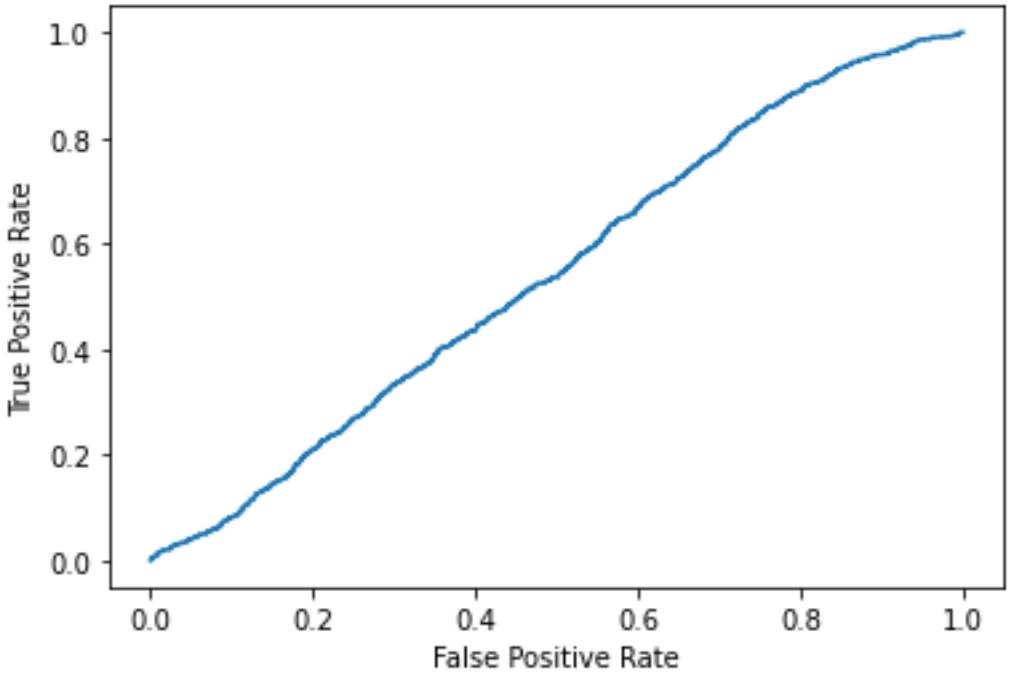
11: The probability of acceptance of an offer and joining the firm.

.Y predict probability 1denotes not joined and 0 – joined

Y_predict_prob - NumPy object array

	0	1
0	0.688949	0.311051
1	0.749958	0.250042
2	0.752504	0.247496
3	0.744675	0.255325
4	0.733078	0.266922
5	0.731315	0.268685
6	0.731492	0.268508
7	0.749336	0.250664
8	0.672766	0.327234

ROC Curve:



For choosing the threshold probability cutoff

```
Thresold aus:  0.541512943850309
```

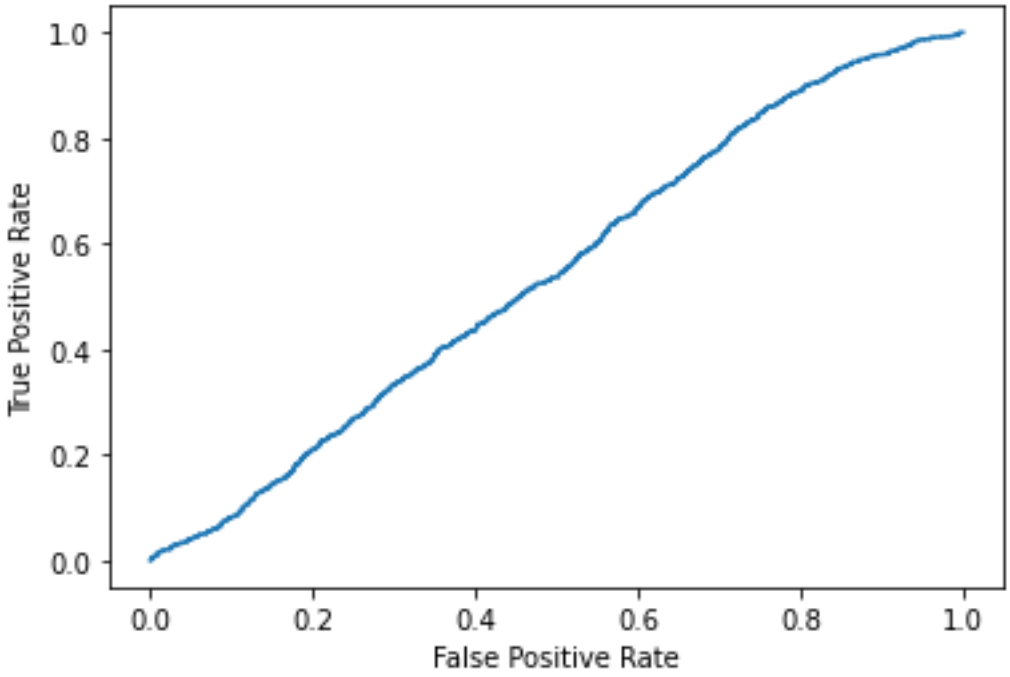
Sol 3: The probability of acceptance of an offer and joining the firm.

.Y predict probability 1denotes not joined and 0 – joined

Y_predict_prob - NumPy object array

	0	1
0	0.688949	0.311051
1	0.749958	0.250042
2	0.752504	0.247496
3	0.744675	0.255325
4	0.733078	0.266922
5	0.731315	0.268685
6	0.731492	0.268508
7	0.749336	0.250664
8	0.672766	0.327234

ROC Curve:



For choosing the threshold probability cutoff

```
Thresold aus:  0.541512943850309
```

Sol 4: The different parameters confusion matrix which determines any models efficiency:

Sensitivity:

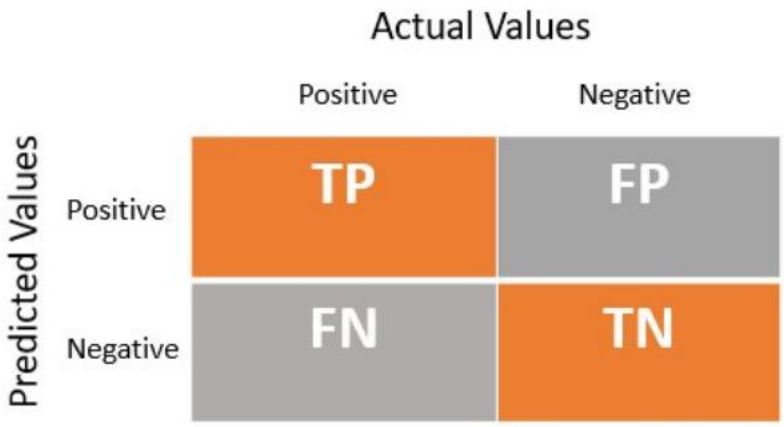
It determines how good the model is in detecting positive occurrences.

Specificity

It determines how precise is the positive class assignment.

Accuracy:

It is the ratio of correct predictions to made predictions.



$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

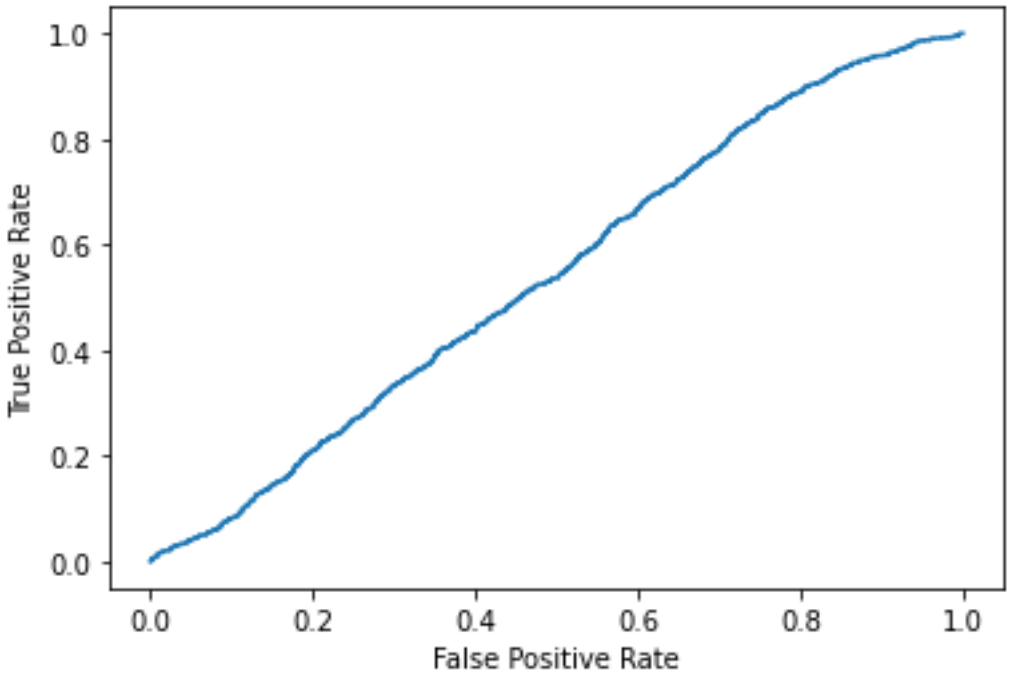
$$Specificity = \frac{TN}{TN + FP}$$

```
Sensitivity :  1.0
Specificity :  0.0
Accuracy :  0.7086486486486486
```

Sol 5: We use REC curve to calculate the threshold:

ROC Curve:

Given Below is the ROC curve:



For choosing the threshold probability cutoff

```
Thresold aus:  0.541512943850309
```

Scalene works should choose 0.54 as the cut-off probability to classify joining and not-joining the curve.

Sol 6:

Cook's distance, Di, is used to determine the influential outliers from the set of predicting variables in Regression analysis. It is used in regression analysis to identify points that negatively affect the regression models.

Cooks distance impact on Logistic regression model:

1. When it has a high x-value, an extreme y-value is likely to bear a lot of influence on a model.
2. We can consider such a point highly influential if removing the point would substantially impact the regression model.
3. It alerts us to potentially erroneous entries in a data set or makes us more aware of unusual values in our data.

Standardized Residual:

The standardized residual is a measure of the error in the AIC. The lesser the residual better it fits the data. With the addition of a more informative predictor, the residual will decrease more than one value in logistic model.