HR Case

Instructions:

- You have to use Python for the complete analysis. Use Jupyter notebook for documentation and python code.
- In case of any issue with the code, please mail to rahul@awesomestats.com

----

We will be using HR data in this exercise. Refer the **Exhibit 1** to understand the feature list. Use the HR data and answer the below questions.

1. Load the dataset in Jupyter Notebook using pandas
2. Build a correlation matrix between all the numeric features in the dataset. Report the features, which are correlated at a cut-off of 0.70. What actions will you take on the features, which are highly correlated?
3. Build a new feature named LOB_Hike_Offered using LOB and percentage hike offered. Include this as a part of the data frame created in step 1. What assumption are you trying to test with such variables?
4. Create a new data frame with the numeric features and categorical features as dummy variable coded features. Which features will you include for model building and why?
5. Split the data into training set and test set. Use 80% of data for model training and 20% for model testing.
6. Build a model using Gender and Age as independent variable and Status as dependent variable.

    - Are Gender and Age a significant feature in this model?
    - What inferences can be drawn from this model?

7. Build a model with statsmodel.api to predict the probability of Not Joining. How do you interpret the model outcome? Report the model performance on the test set.
8. Build a model with statsmodel.formula.api to predict the probability of Not Joining and report the model performance on the test set. What difference do you observe in the model built here and the one built in step 7.
9. Build a model using sklearn package to predict the probability of Not Joining. What difference do you observe in this model compared to model built in step 7 and 8.
10. Fine-tune the cut-off value using cost of misclassification as a strategy. The cut-off should help classify maximum number of Not Joining cases correctly.
11. Fine-tune the cut-off value using youdens index as a strategy. The cut-off should help balance the classification of Joined and Not Joined cases.
12. Apply the cut-off values obtained in step 10 and step 11 on the test set. What inference can be deduced from it?

Exhibit 1

Sl. No.	Name of Variable	Variable Description
1	Candidate reference number	Unique number to identify the candidate
2	DOJ extended	Binary variable identifying whether candidate asked for date of joining extension (Yes/No)
3	Duration to accept the offer	Number of days taken by the candidate to accept the offer (continuous variable)
4	Notice period	Notice period to be served in the parting company before candidate can join this company (continuous variable)
5	Offered band	Band offered to the candidate based on experience and performance in interview rounds (categorical variable labelled C0/C1/C2/C3/C4/C5/C6)
6	Percentage hike (CTC) expected	Percentage hike expected by the candidate (continuous variable)
7	Percentage hike offered (CTC)	Percentage hike offered by the company (continuous variable)
8	Percent difference CTC	Percentage difference between offered and expected CTC (continuous variable)
9	Joining bonus	Binary variable indicating if joining bonus was given or not (Yes/No)
10	Gender	Gender of the candidate (Male/Female)
11	Candidate source	Source from which resume of the candidate was obtained (categorical variables with categories Employee referral/Agency/Direct)
12	REX (in years)	Relevant years of experience of the candidate for the position offered (continuous variable)
13	LOB	Line of business for which offer was rolled out (categorical variable)
14	DOB	Date of birth of the candidate
15	Joining location	Company location for which offer was rolled out for candidate to join (categorical variable)
16	Candidate relocation status	Binary variable indicating whether candidate has to relocate from one city to another city for joining (Yes/No)
17	HR status	Final joining status of candidate (Joined/Not-Joined)