**Analysis for the project – Q&A’s as per code:**

Summary:

Methods used: Data Preprocessing, Descriptive Statistics, Correlation Analysis, Regression Modeling, Statistical Analysis, Model Evaluation Metrics, Feature Engineering, Visualization, Variance Inflation Factor, Logistic Regression Models, Multicollinearity Detection

1. You have two interval variables here: *JoiningYear*, *Age.*Perform a 0/1 scaling of these features. Also, perform dummy encoding for the remaining variables, so that each variable with n levels creates n-1 variables. Use *get\_dummies* function form Pandas for this. What are the total number of columns available after encoding Including the target variable)? (Enter the integer value)

* 11

2. Now, if you found variables with high VIF, remove **iteratively**them if their VIF is greater than 10. When you can't remove any more variables, report, with 3 decimal digits, the value of the highest VIF

* 2.885

3. What is the proportion of 1's in the target variable? Give result with three decimals.

* 0.280

4. In view of this, is this a balanced dataset?

* No

5. What is the  adjusted value on the validation dataset rounding to three decimals?

* 0.484

6. In view of these results, can you say that you are overfitting your data?

* No

7. Check the collinearity between variables?

8. Perform a 70/30 training/validation split. Use *random\_state=0*for the split. The first model (Model 1) we will try will be a logistic regression with one variable. We will pick the variable with the highest correlation with the target variable. Then train this model with *statsmodels.*Plot the confusion matrix and explain if you would consider this model or not and why..

* We cannot consider this model. From the confusion matrix, it is evident that we cannot consider this model as there is a **Class imbalance** in the data, where most employees are falling into "Not Leaving" category and hence the model is not able to properly predict the ones who are actually leaving. It completely failed to predict the ones who are actually leaving.

9. For our second model, we will include all variables in the dataset. Perform a logistic regression model. In view of your results, categorize your variables into significant and non-significant variables.

10. Eliminate all variables that were not significant previously. If the intercept was not significant, eliminate it as well. Run another logistic regression (Model 2) with the remaining variables. The resulting model should now be significant with all its variables significant at the 0.05 level of significance.

We will not use this metric for assessment, but what is the value of the pseudo R-square value in the validation set of this new model logistic regression? (Enter the answer rounding to three decimal points)

* -1.039

11. Now that we have a model that is useful, let us see how it performs. Using *Scikit-Learn*, run the same logistic regression. Make sure you are familiar with the [argumentsLinks to an external site.](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html" \t "_blank) of this function. Bear in mind that you are not imposing any penalty and that you do not want an intercept. Check the values of the coefficients. They should be REALLY close to those found with *statsmodels*. The difference is because the numerical methods to attain them are slightly different in both cases. Write below the highest coefficient obtained (in absolute value), with three decimals.

* 1.669

12. Give the accuracy of your model in the validation set, with three decimals.

* 0.743

13. As far as accuracy is concerned, do you think this model overfits? Explain.

* Training Accuracy: 0.739 and Validation Accuracy: 0.743. The model is not likely overfitting as the values are close to each other and the performance is consistent across training and validation sets.

14. Do you think accuracy is a good metric for this exercise? Explain. If you think it is not, which do you think would be a good metric?

* Even though the model is not over fitting, accuracy may not be the best metric because there is an imbalance in the dataset. Even though in real world scenario the turnover rate maybe really low(less number of resignations), however in our case we are classifying this as an imbalanced data. If we have a better balanced data then the following will be useful accordingly:  
  Recall: This would be helpful us capture and understand how the model captures the actual "Leave" cases. Maybe it might be a tradeoff with Precision if we decrease the threshold, but with our imbalanced dataset, Recall would be a better measurement  
  ROC-AUC Score: Can use this as well to understand how well the model discriminates between "Leaving" and "Not Leaving."

15. Perform another multivariate logistic regression (Model 3) with same target and the variables *PaymentTier\_3*, *Age*, and *Gender\_Male*. What is the value of the maximum p-value for the significance of the variables that you obtained? Give three decimals.

* 0.084

16. Assuming that you want to select a model that provides the best true negative rate possible, among the three models (Model 1, Model 2, Model 3) which one would you keep and why?

* True Negative Rate measures proportion of actual negatives which are correctly identified, in our scenario "Not Leaving". However, this result might be biased as our data is severely imbalanced. The following are the TNR rates  
  Model 1 TNR: 1.000

Model 2 TNR: 0.891

Model 3 TNR: 0.963

The best model based on TNR is: Model 1  
If we do not care about how the model is performing and just want to take the highest TNR then Model 1 TNR would be selected, which means that we are not very much concerned about the "Leave" rates. In order to make a better inference we might want to go with Model 3, where TNR is good and the little tradeoff means that there is better recall for predicting actual "Leave" cases