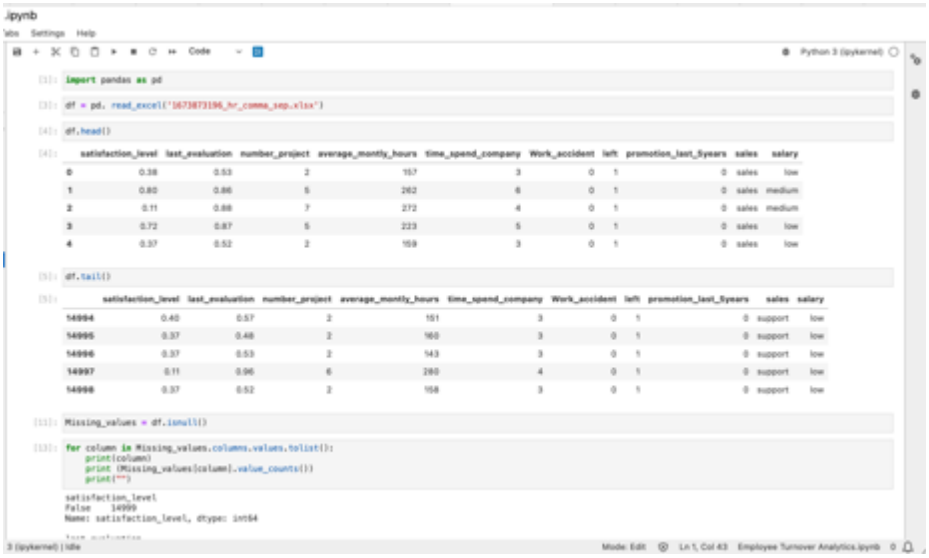
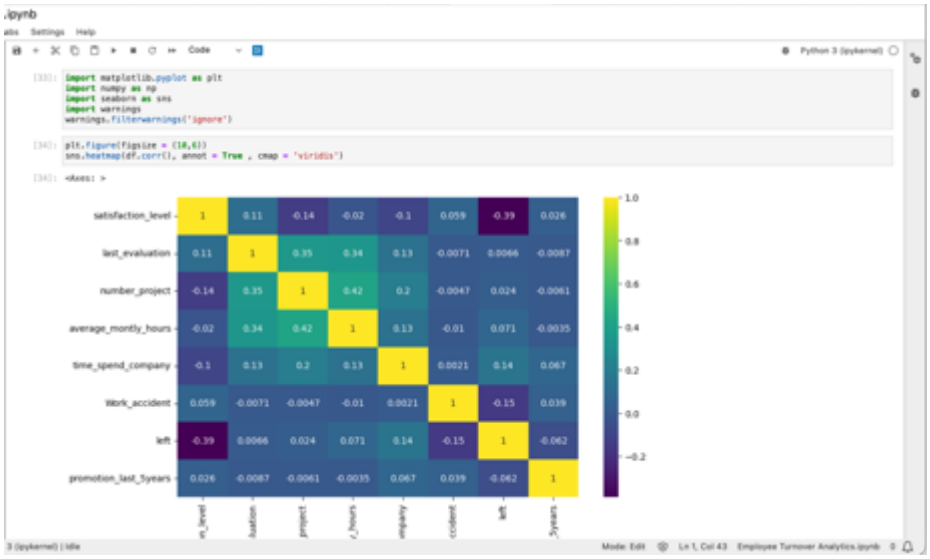
Perform the following steps:

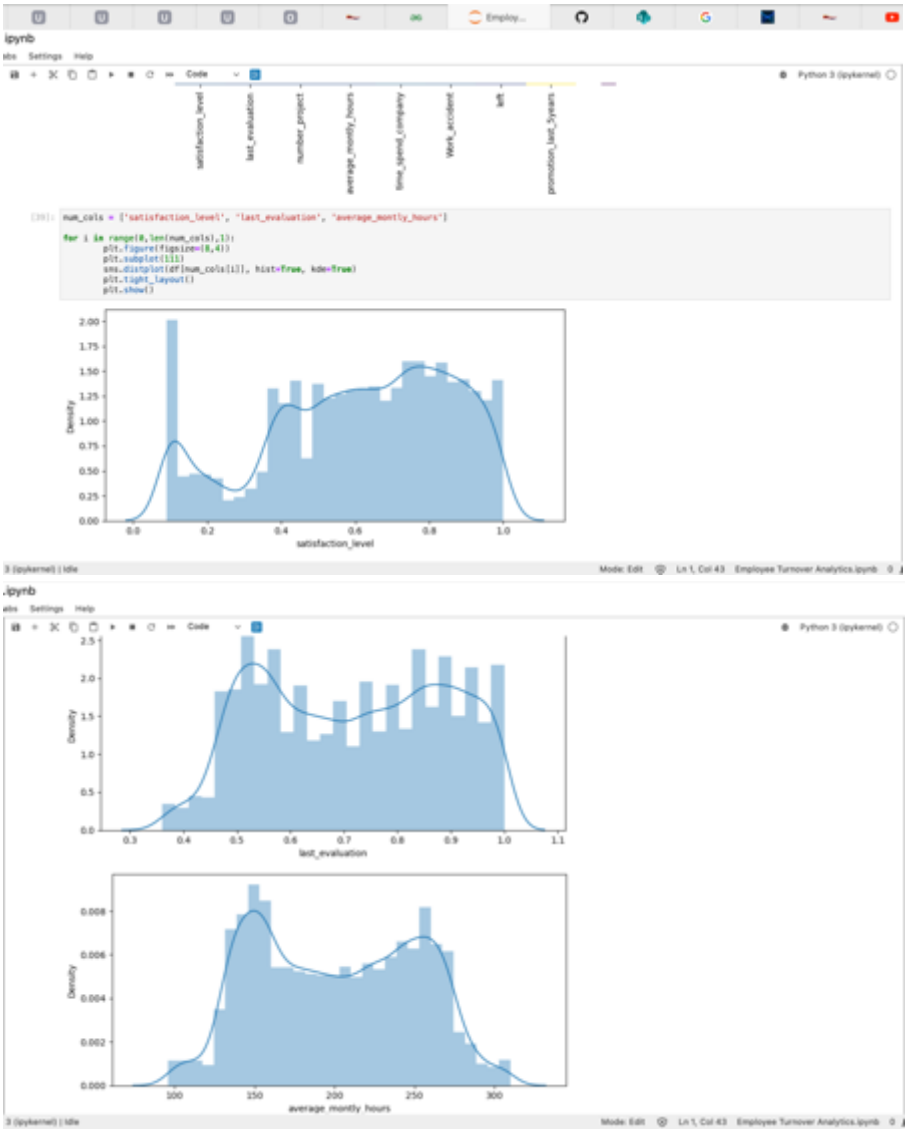
1. Perform data quality check by checking for missing values if any.



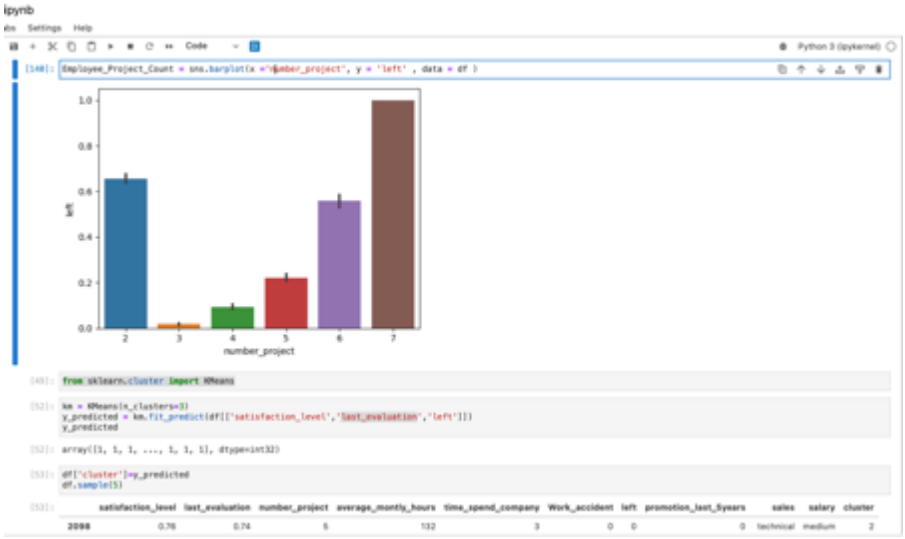
2. Understand what factors contributed most to employee turnover by EDA.
   2.1. Draw a heatmap of the Correlation Matrix between all numerical features/columns in the data.



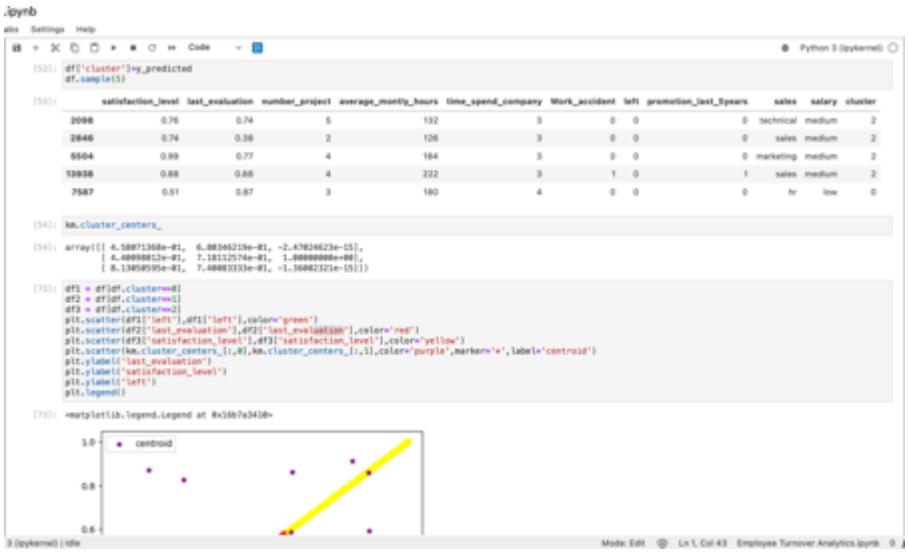
   2.2. Draw the distribution plot of
   - Employee Satisfaction (use column satisfaction_level)
   - Employee Evaluation (use column last_evaluation)
   - Employee Average Monthly Hours (use column average_montly_hours)

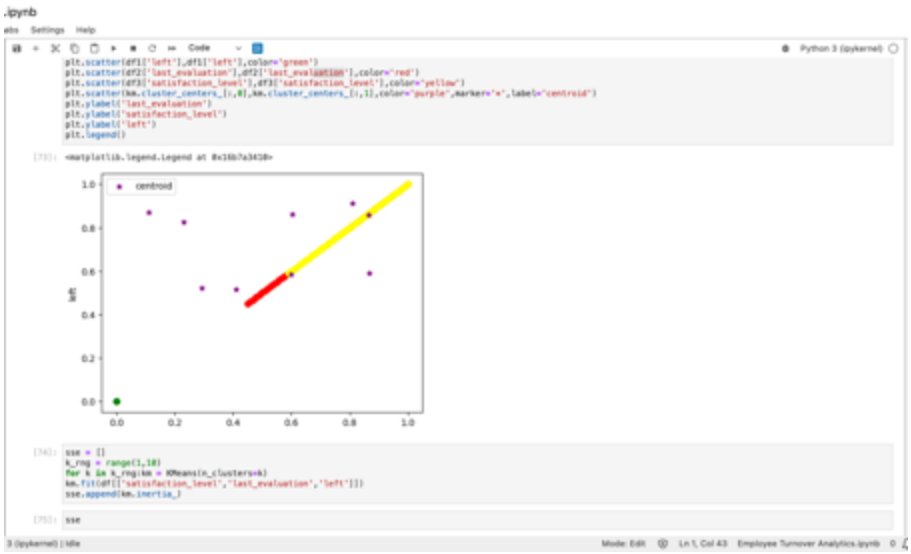


2.3. Draw the bar plot of Employee Project Count of both employees who left and who stayed in the organization (use column number_project and hue column left) and give your inferences from the plot.

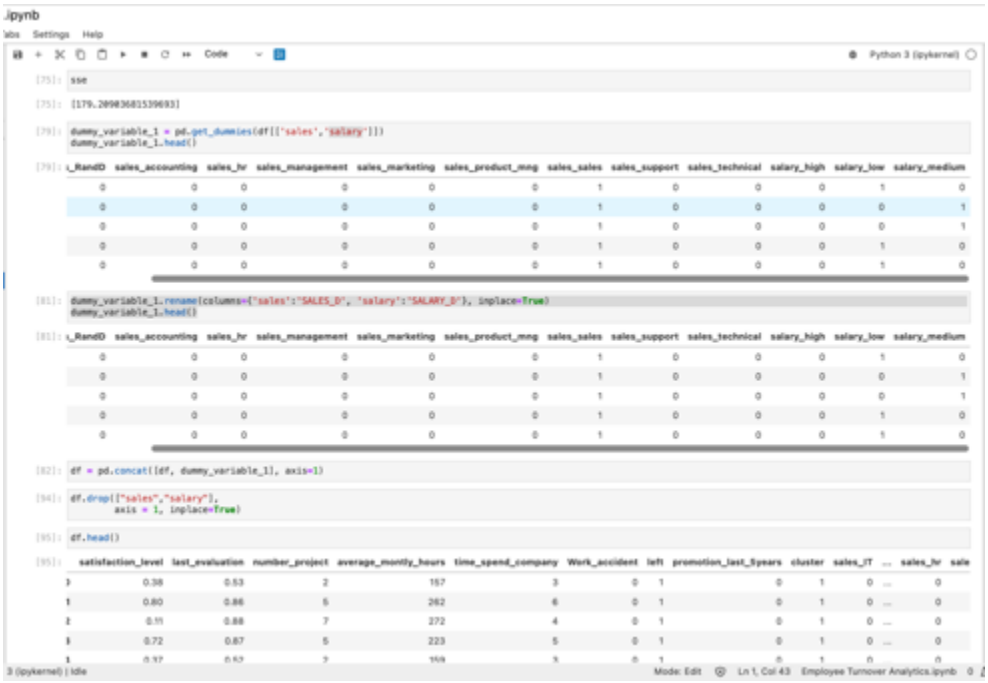


3. Perform clustering of Employees who left based on their satisfaction and evaluation.
    3.1. Choose columns satisfaction_level, last_evaluation and left.
    3.2. Do KMeans clustering of employees who left the company into 3 clusters.
    3.3. Based on the satisfaction and evaluation factors, give your thoughts on the employee clusters.

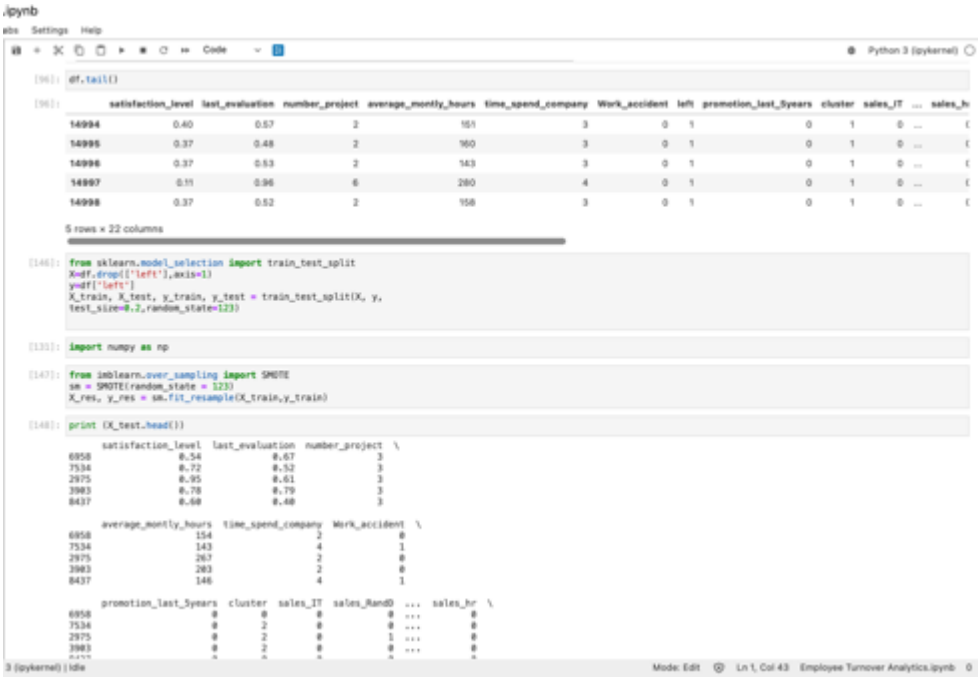




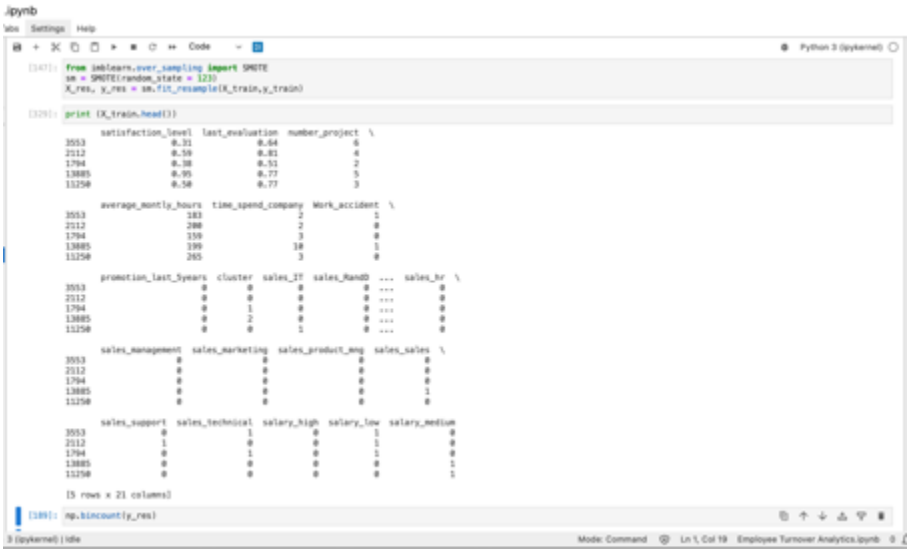
4. Handle the left Class Imbalance using SMOTE technique.
    4.1. Pre-Process the data by converting categorical columns to numerical columns by
        - Separating categorical variables and numeric variables.
        - Applying get_dummies() to the categorical variables.
        - Combining categorical variables and numeric variables.



4.2. Do the stratified split of the dataset to train and test in the ratio 80:20 with random_state=123.

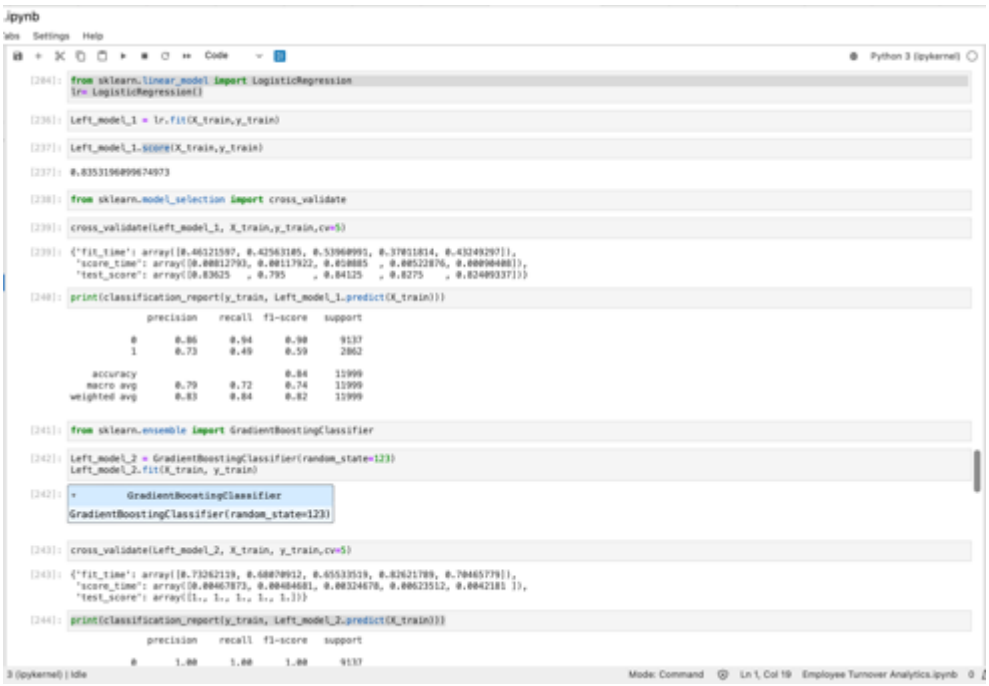


4.3. Upsample the train dataset using SMOTE technique from the imblearn module.

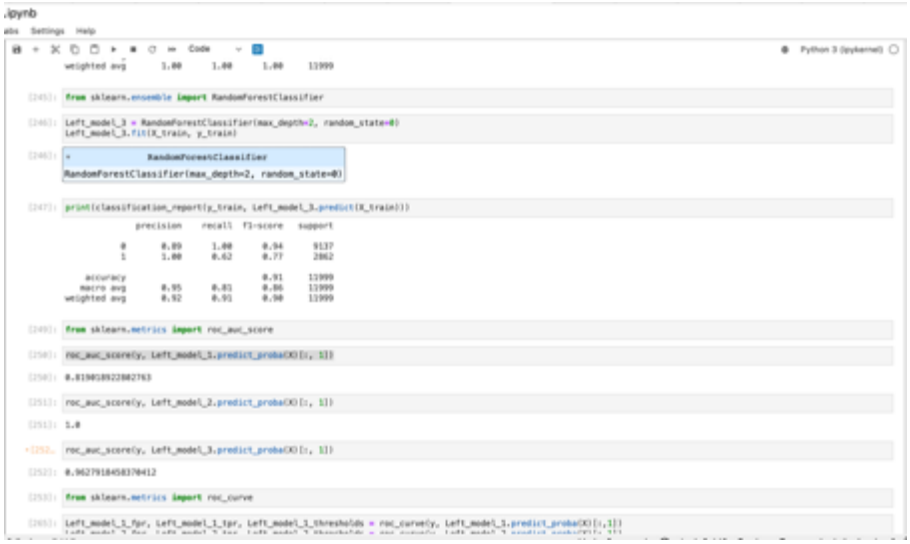


5. Perform 5-Fold cross-validation model training and evaluate performance.
   5.1. Train a Logistic Regression model and apply a 5-Fold CV and plot the classification report.

   5.2. Train a Random Forest Classifier model and apply the 5-Fold CV and plot the classification report.
   5.3. Train a Gradient Boosting Classifier model and apply the 5-Fold CV and plot the classification report.

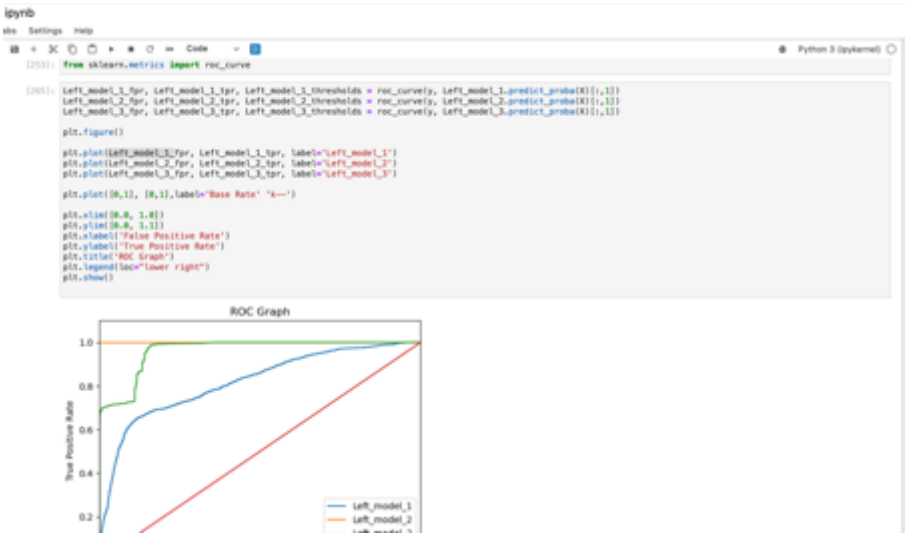


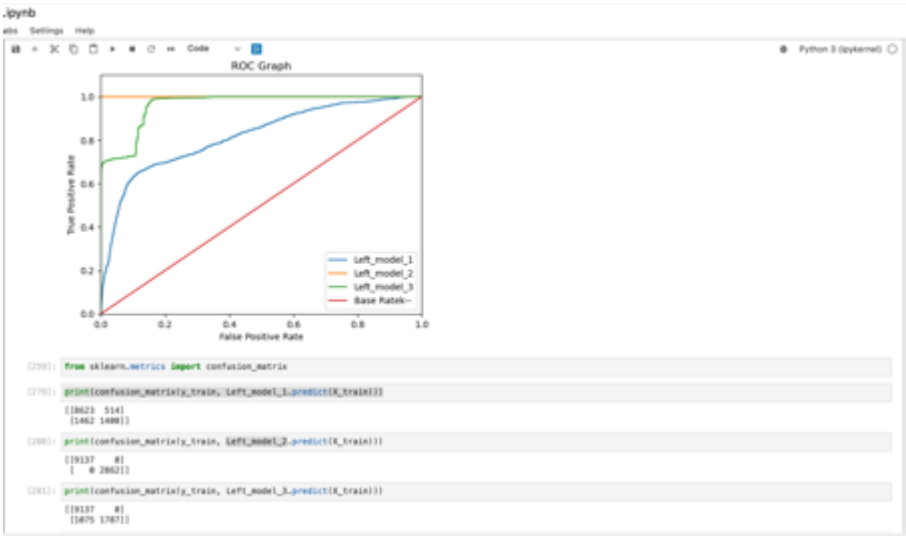
6. Identify the best model and justify the evaluation metrics used.
   6.1. Find the ROC/AUC for each model and plot the ROC curve.



6.2. Find the confusion matrix for each of the models.

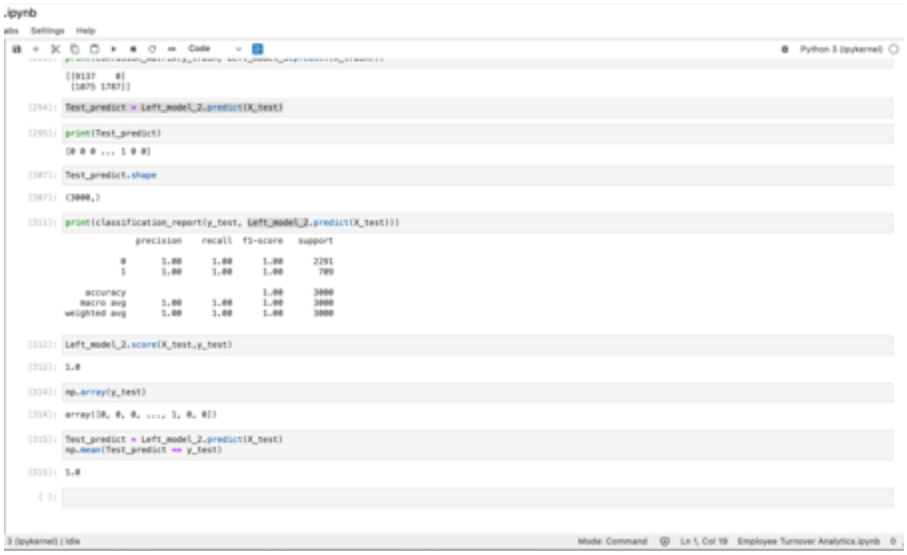
6.3. From the confusion matrix, explain which metric needs to be used- Recall or Precision?





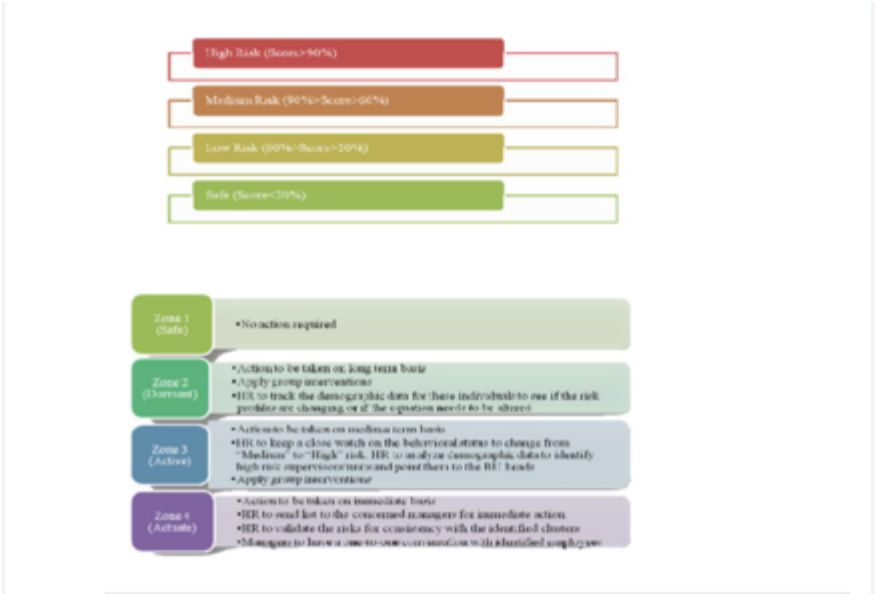
7. Suggest various retention strategies for targeted employees.

7.1. Using the best model, predict the probability of employee turnover in the test data.



7.2. Based on the below probability score range, categorize the employees into four zones and suggest your thoughts on the retention strategies for each zone.

- Safe Zone (Green) (Score < 20%)
- Low Risk Zone (Yellow) (20% < Score < 60%)
- Medium Risk Zone (Orange) (60% < Score < 90%)
- High Risk Zone (Red) (Score > 90%).



The rationale behind designing the problem statement on Talent Management:

**Note to Simplilearn:*