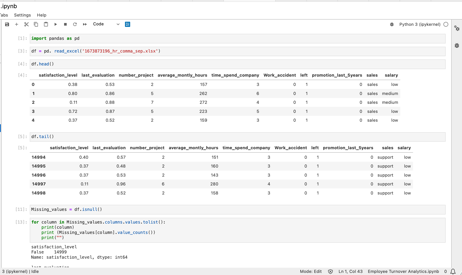
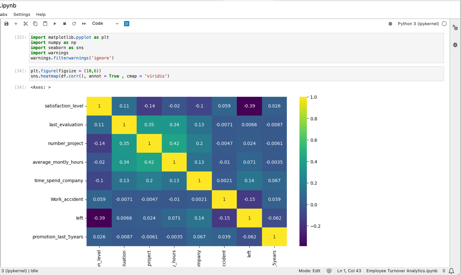
**Perform the following steps:**

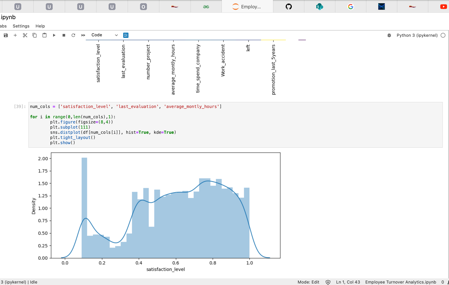
1. Perform data quality check by checking for missing values if any.

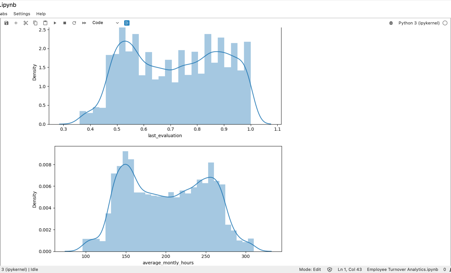


1. Understand what factors contributed most to employee turnover by EDA.
   1. Draw a heatmap of the Correlation Matrix between all numerical features/columns in the data.

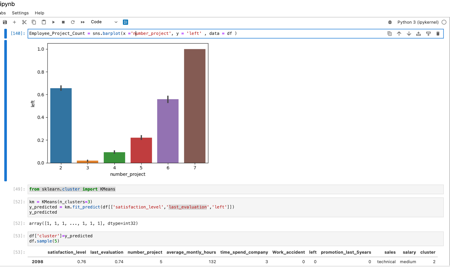


* 1. Draw the distribution plot of
     + Employee Satisfaction (use column satisfaction\_level)
     + Employee Evaluation (use column last\_evaluation)
     + Employee Average Monthly Hours (use column average\_montly\_hours)

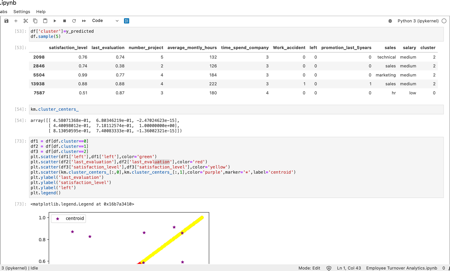


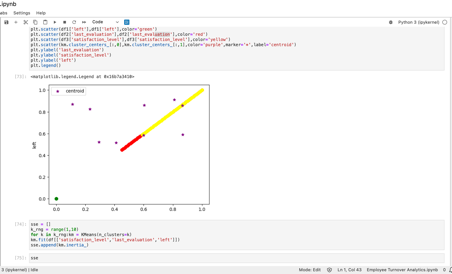


* 1. Draw the bar plot of Employee Project Count of both employees who left and who stayed in the organization (use column number\_project and hue column left) and give your inferences from the plot.

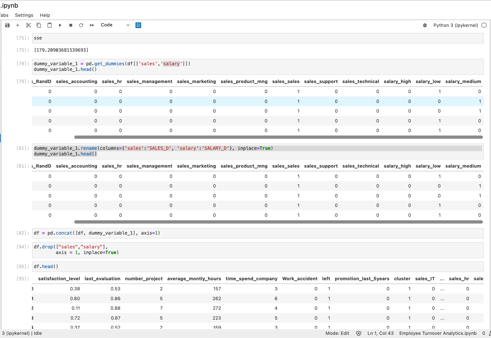


1. Perform clustering of Employees who left based on their satisfaction and evaluation.
   1. Choose columns satisfaction\_level, last\_evaluation and left.
   2. Do KMeans clustering of employees who left the company into 3 clusters.
   3. Based on the satisfaction and evaluation factors, give your thoughts on the employee clusters.

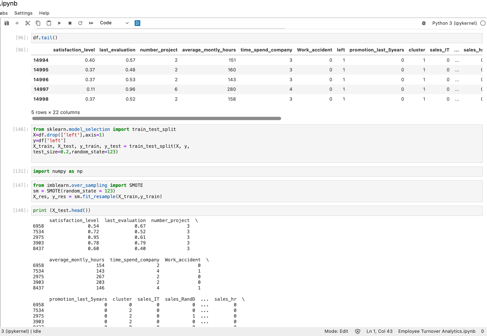




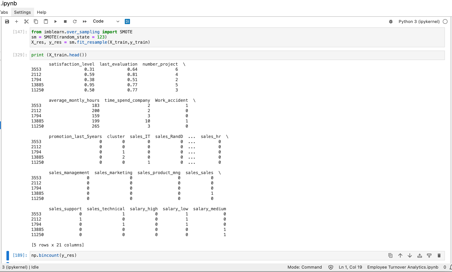
1. Handle the left Class Imbalance using SMOTE technique.
   1. Pre-Process the data by converting categorical columns to numerical columns by
      * Separating categorical variables and numeric variables.
      * Applying get\_dummies() to the categorical variables.
      * Combining categorical variables and numeric variables.



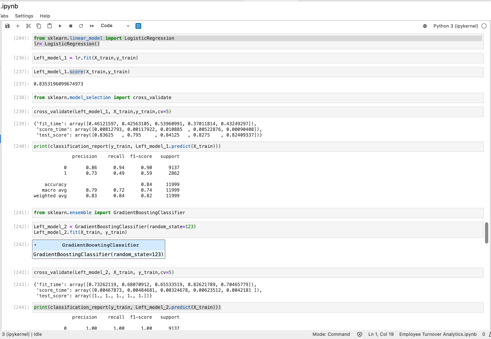
* 1. Do the stratified split of the dataset to train and test in the ratio 80:20 with random\_state=123.



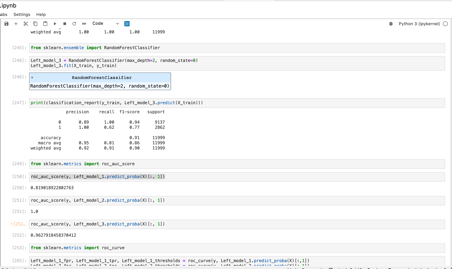
* 1. Upsample the train dataset using SMOTE technique from the imblearn module.



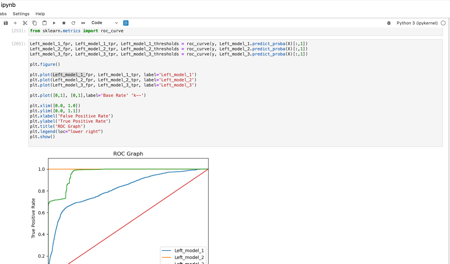
1. Perform 5-Fold cross-validation model training and evaluate performance.
   1. Train a Logistic Regression model and apply a 5-Fold CV and plot the classification report.
   2. Train a Random Forest Classifier model and apply the 5-Fold CV and plot the classification report.
   3. Train a Gradient Boosting Classifier model and apply the 5-Fold CV and plot the classification report.

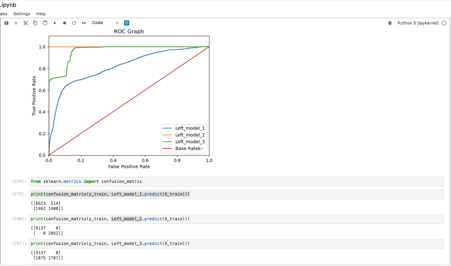


1. Identify the best model and justify the evaluation metrics used.
   1. Find the ROC/AUC for each model and plot the ROC curve.

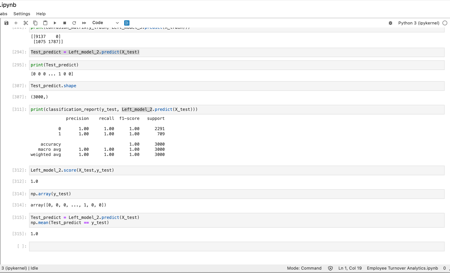


* 1. Find the confusion matrix for each of the models.
  2. From the confusion matrix, explain which metric needs to be used- Recall or Precision?

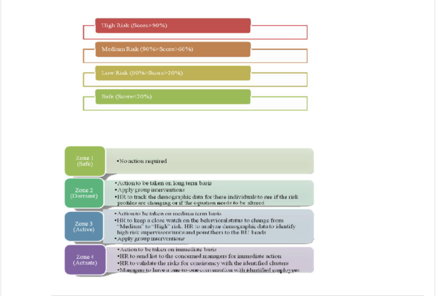




1. Suggest various retention strategies for targeted employees.
   1. Using the best model, predict the probability of employee turnover in the test data.



* 1. 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:*