Explorative Data Analysis Outcomes:

1.Officer Year of Birth --> Range Between 1946 to 1987 (~41 years)

2.Officer Years of Experience --> Having -ve Values but experiance must + ve

3.Most of the Reprted_Year and Occurred_Year comes under same time span (2018 to 2022) Highest in 2018 and lowerst in 2022

4.All the -ve years of experience comes under Gender=N

5.Male having highest years of experience then compare to female

6.Officer Gender =N having Highest Reporting Months then compare to Male and Female

7.CIT Certified Indicator is Active after 2016 (Reported Year Range b/w 1900 -2020)

8.April month highest CIT Certified Indicator values occurred

Machine Learning Outcomes

Evaluation Metric: F-Score

1) The metrics such as Precision, Sensitivity, Recall, Specificity, F-Score, ROC-AUC Curve is used to evaluate the Classification problem.Here, we could observe that the output data is slightly imbalanced, here we could use the F-1 Score. F-1 score is generally useful when working with the imbalanced dataset and it also combines precision with recall into a single metric.

2) f1_score is 58% with Logistic Regression with L2 penalty

3) f1_score is 58% with Logistic Regression with L2 penalty with 10 folds Cross Validation and liblinear solver.

4) f1_score is 58% with Logistic Regression with L2 penalty and saga solver.

5) By Using KNN, F1_Score is 82% which is better then Logistic regression.

6) By Using RandomForest, F1_Score is 88% which is better than all the previous models.

7) By Using Xgboost, F1_Score is 87% which is less slight than Random Forest Model.

8) By Using SVM, F1_Score is 72% which is less than xgboost model.

9) By comparing all the above implemented models, we can conclude that Random forest model has improved the performance of the model.

Conclusion : Classification Problem is overcome by evaluating various factors such as Precision, Sensitivity, Recall, Specificity, F-Score, ROC-AUC Curve. Among all the algorithms between KNN,RandomForest,Xgboost,SVM ,"Random forest model has improved the performance of the model"

Further Improvements and outcomes:

Further, the results can be improved by having vast knowledge on the business domain, which could be useful in understanding and pre-processing the data. Thus, the outliers and any random noise can be removed.

Neural networks along with Tensor-flow models can improve the performance of the model

References :

1.https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

2.https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

3.https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

4.https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

5.https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/