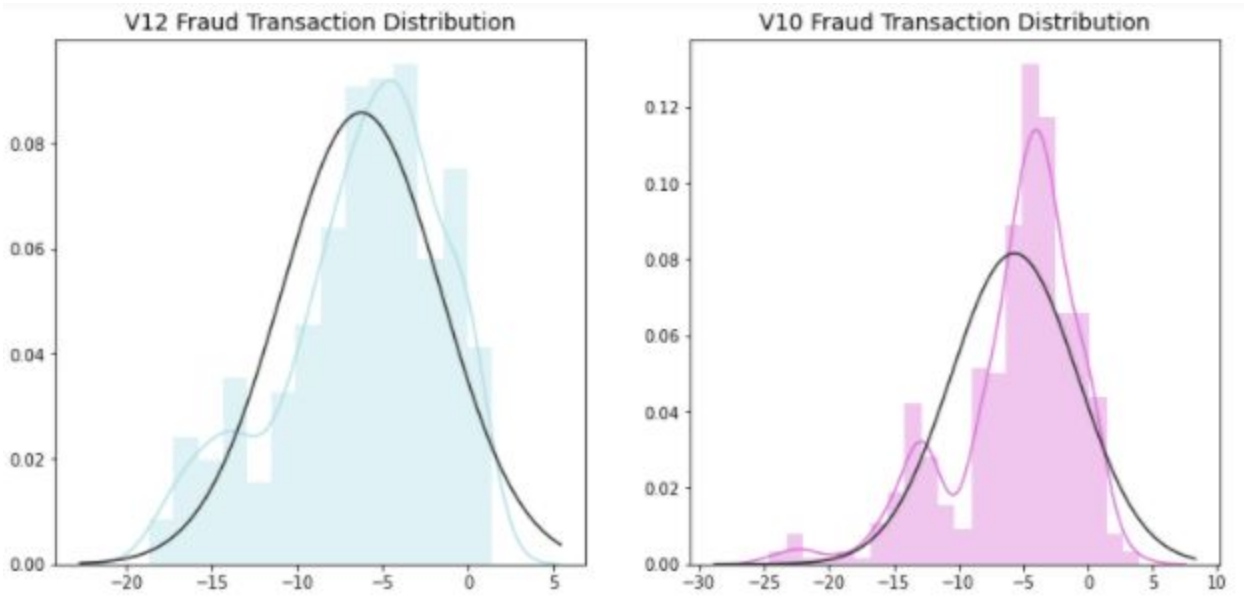
PROJECT DELIVERABLE 3

1. Final Training Results

For this deliverable, my goal is to improve my model next time so that my accuracy score and f_1 score will increase. I have learnt more about my variables and their correlation with my Class. Based on what I observed, the V17, V12 have negative correlation and it will be more likely to be a fraud transaction. Whereas the V11, V19 have a positive correlation. With that knowledge, I also did anomaly detection and my goal is to remove the outliers from the extremely correlated variables.



As I mentioned in the last deliverable, my dataset is very imbalanced and my TPM has suggested me to use SMOTE and XGBoost. The idea is that Synthetic Minority Over-Sampling Technique (SMOTE) is used to create new synthetic points in order to equal out the balance of class '0' and '1'. The technique will pick the distance between the closest neighbours of the minority class, in between these distance to create its synthetic points. This result in a higher f_1 score for my value in Naive Bayes and Logistics Regression as more information is kept and I am not losing any useful feature. However, the trade-off is that the training time was much longer, I estimate about 4 hours to complete for my dataset.

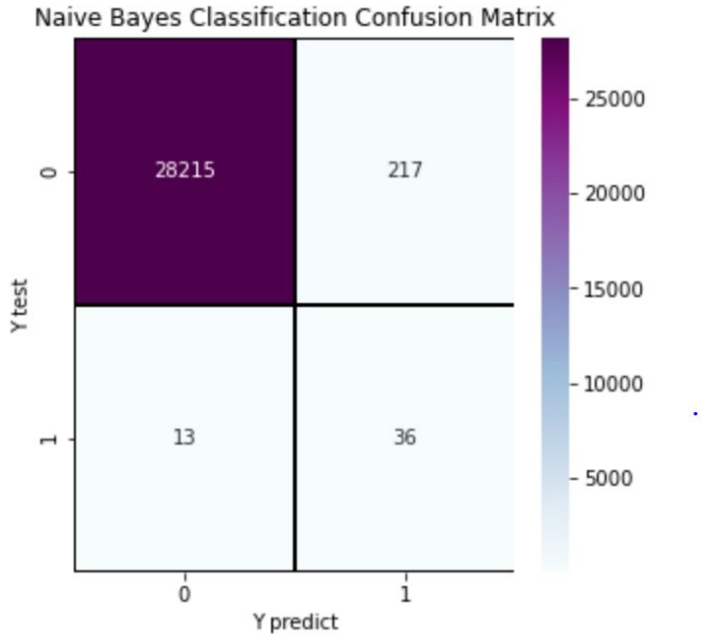
Apart from SMOTE, I also work on Grid Search to find the best parameter and to tune in the different hyperparameters for my models more efficient. I have tuned in variables for Logistics Regression with best parameters of c = 0.01, Supporting Vector with kernel = 'sigmoid', c = 0.7.

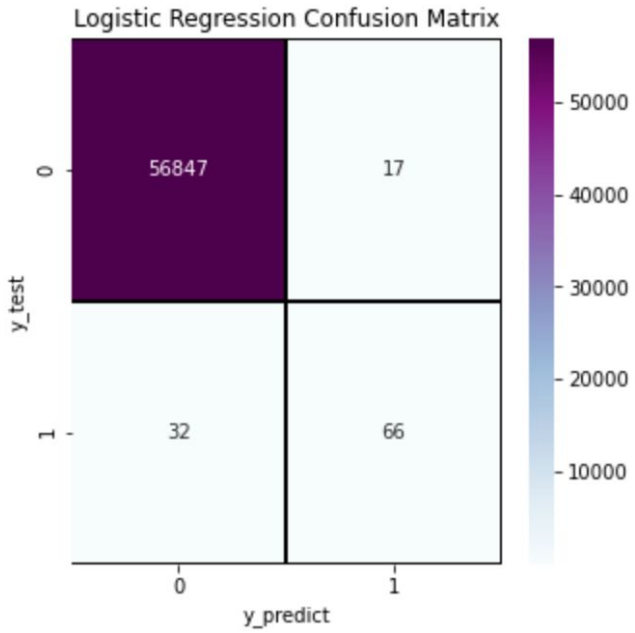
For Random Forest, I did not perform Grid Search on since the training time is already high for the model and I don't think my GPU can handle tuning hyperparameters for random Forest. For Naive Bayes Gaussian, it does not have available hyperparameters for me to perform Grid Search on.

In conclusion, the best model so far for my dataset is Logistics Regression with the f_1 score of 73% and an accuracy of 97%.

My confusion matrix with :

- True Negative (top left): correctly classify the genuine class
- False Negative (top right): incorrectly classify the genuine class (this is something I want to avoid)
- False Positive (bottom left): incorrectly classify the fraud class (this is not good but it is better to give careful prediction)
- True Positive (bottom right): correctly classify the fraud class i.e correctly detect the fraud transaction, which is the goal of my project.





As we can see that the for the False Negative matrix, our Logistics Regression does it much better than the Naive Bayes Gaussian method, with significant lower amount, 17 vs 217. Also, since the amount of fraud transaction is only 0.72%, the prediction of 66 cases are reasonable value.