Exploratory data analysis:

The cleaning involves but not limited to the follow:
1. Regular expression cleaning up
2. Changing categorical data to numeric values
3. Changed the dataset to dummy set for better segregation of information

Some problems found during eda :
1. The dataset is relatively small which could affect the training result
2. The dataset is heavily skewed towards non churning at a 8:2 ratio.

Model design:

Three models were used to train and evaluate the dataset:
1. MLP
2. Random Forest
3. Logistic regression

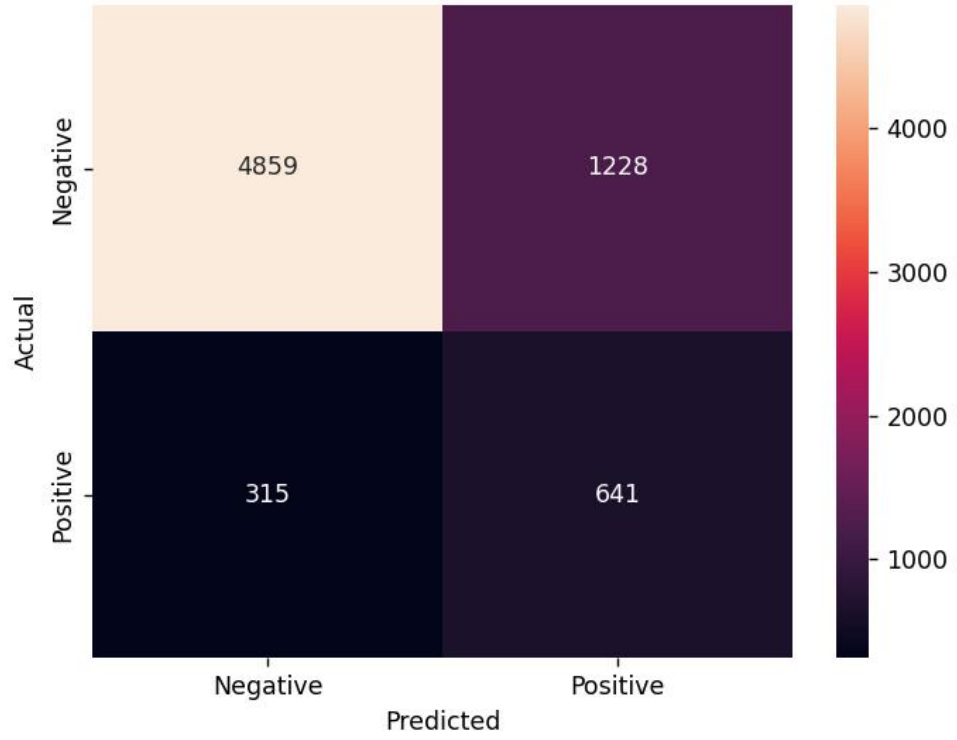
With MLP using 1 hidden layer with input dim of 512 and output dim of 512 as well.
The MLP was trained over a total of 30 epochs and using a learning rate of 0.001.

All three model were evaluated over 5 fold cross validations and the results are as below.

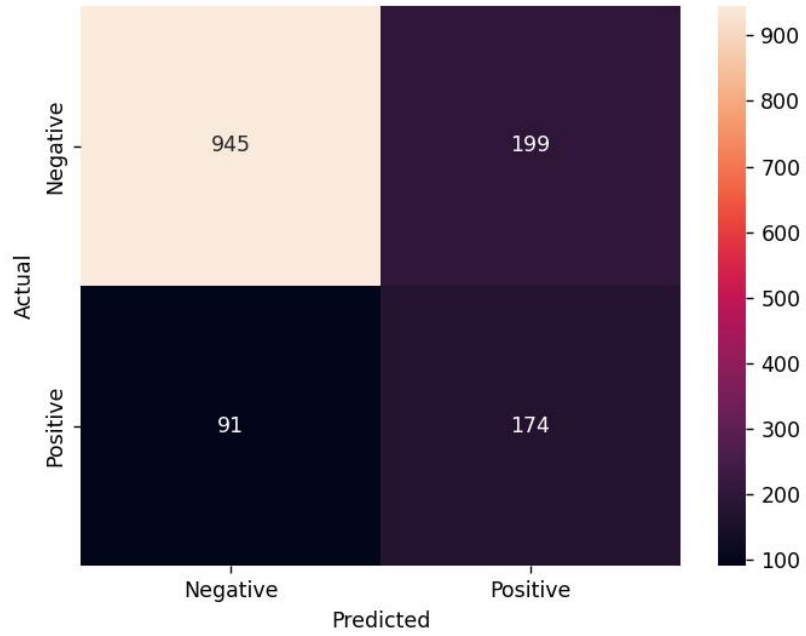
Avg result	MLP	RF	LR
Training time	3 Minutes	1 Minutes	10 seconds
Accuracy	78%	80%	80%
F1 Score	0.45	0.54	0.63

Confusion matrix:

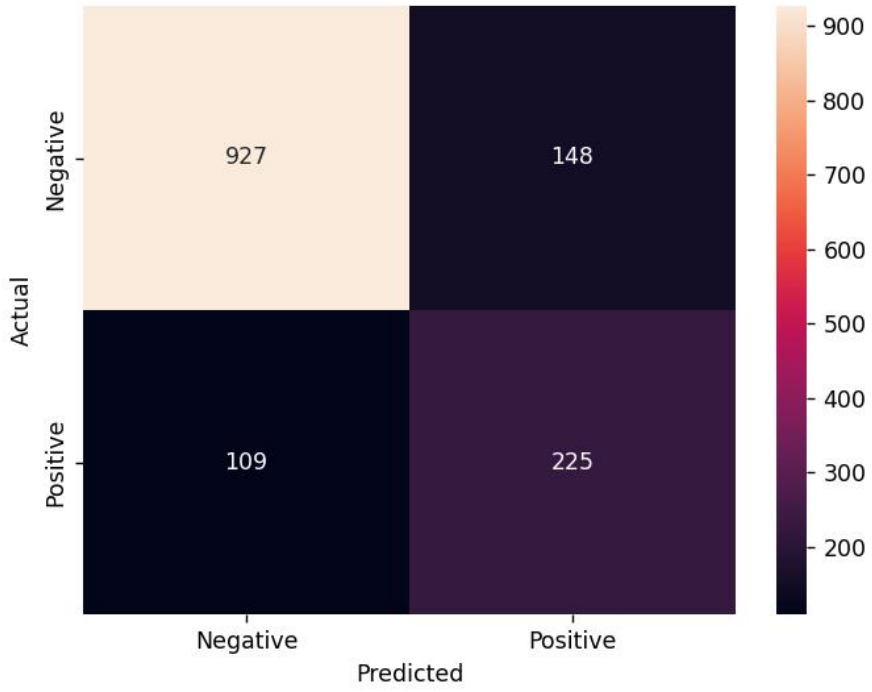
1. MLP



2. RF



3. LR



Conclusion:

None of the models performed exceptionally well as although all models' accuracy are high, that is mainly due to the highly skewed data samples in the dataset.

This is reflected in the f1 scores, where all three models scored relatively low compared to the accuracy.

The low f1 scores might be due to the rather small sample base and could be improved if more data are gathered.

However, out of the three models, logistic regression performed the best where the model outperformed the other two models in all aspects such as running time,accuracy and f1 score.