Problem statement

To predict which group of credit card customer tends to end their services with the bank and aims to reduce high attrition rate due to those groups of clients.

Findings from EDA:

1. In general, most clients belong to the income group of less than $40K.
2. In general, most clients belong to the education background of graduate.
3. As the customer are open to buy credit line, their credit limit increases.
4. As the age of customer increases, their months on book are longer as well.
5. As the credit limit increases, the average utilization decreases, showing that those with higher credit limit does not use their credit card as often.
6. As average open to buy credit line increases the average utilization decreases, showing that as the client purchase more credit line, their average utilization of their credit line decreases.
7. As the average utilization ratio increases, the credit limit decreases, showing that the client who utilises their credit card most often has the lower credit limit.
8. As the average utilization ratio increases, the average open to buy decreases, showing that client who utilises more of their credit line tend to open lesser credit line thereafter.
9. Based on the heatmap, we can see that dependent count and total relationship count is highly correlated, thus we can determine that as the dependent count increases or decreases, the client will be less or more willing to buy more credit line.
10. Most clients belong to the blue card category which is possibly the entry level of credit card for banks, meaning to say most clients are with lower income group and has lower spending power.
11. Most clients belong to either married or single with 'low usage, long term users' having more than 50% married compared to the rest of the groups. Meaning to say married and single would most likely apply and use the credit card of banks.
12. Most clients have either 2 or 3 dependents for all group types, meaning to say this group of clients will most likely be open to purchase credit line from banks.
13. The higher credit limit group are largely male while the lower credit, high balance are largely female. Meaning to say male has higher purchasing power but does not use

much of their credit card while female with lower purchasing power would most likely utilise their credit card often.

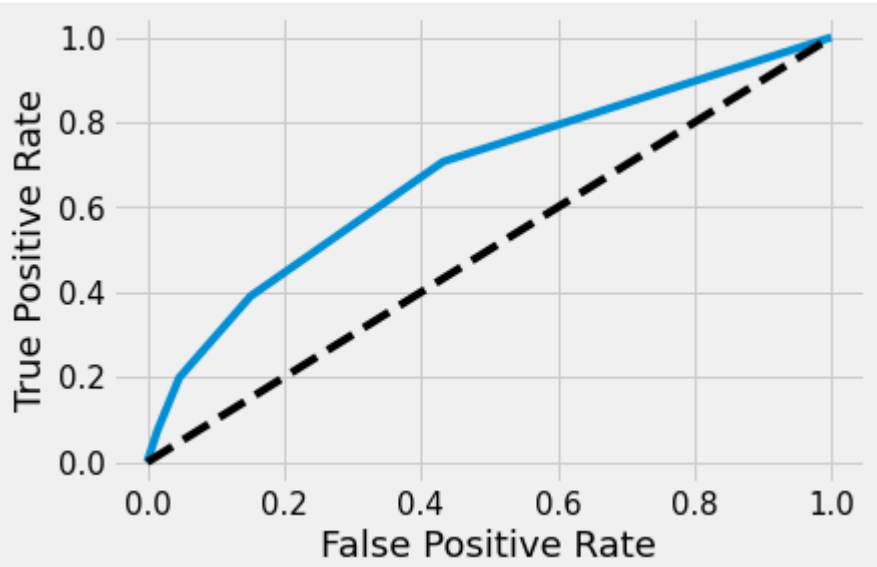
14. Most clients are from the age group of 40s and 50s, making them the most possible clients to be willing to open credit line with banks.
15. Low usage, long term users and low usage high credit long term users appears to have the most attrited customer compared to the rest of the group. This goes to show that low usage user has higher tendency to leave the bank.

About the Model

For this model, the classification model used will be K-Nearest. The KNN algorithm calculates the probability of the test data belonging to the classes of 'K' training data, and the class holding the highest probability will be selected. In this case, the likelihood of customer ending their services with the bank based on past data of the similarities of their usage behaviour to the group of customers that tends to end their services with the bank (described in the training data).

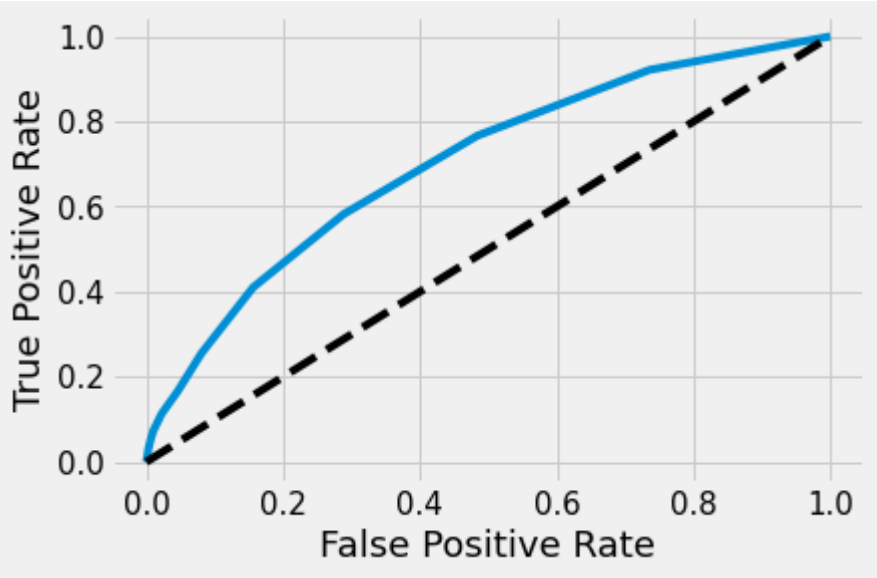
Model performance

In the initial model setup, the model produced results with 0.8304 accuracy and 0.576 AUC score.



After hyperparameter tuning:

After tuning our model using RandomSearchCV on n_nearestneighbours and leaf size, the model produced results with 0.8404 accuracy and 0.532 AUC score.



Also, attempted to use Random Forest Classifier(RFC) in an attempt to get better predictive scoring. Using RFC has given a better accuracy score of 1.0 compared to KNN which even after tuning having a score of 0.8404.

Limitation of Dataset

1. When trying to split the data into distinct groups to identify which credit card client's group has the highest attrition rate using K Means Clustering, there are some overlapping clustering which the data sets could be improve further for better accuracy.
2. The datasets are not distinct enough to identify which customer type may have highest potential attrition and the reasons for it to happen, the datasets could include things like their most spend on category such as luxury goods, restaurants, etc. Hence, the banks can better understand their customer and apply necessary perks to retain their credit card customer.

Suggestion and Future improvements

Since the accuracy score is rather low, a more robust machine learning algorithm should be used and explored to get better prediction results. Just like I did by cross checking with another type of algorithm such as RFC, which provided better results, this proof to show that other predictive algorithm could be used to achieve a better predictive result. Also with better distinct dataset, the result will be better improved.