**ML Final Project Report**

1. Introduction

A bank based in Europe is worried that existing customers are no longer using them as a main bank. Therefore, we would try to find out the reasons for the change (customer churn) and to be able to predict this in Data Science.

Inputs are:  
(i) Credit Rate

(ii) Geography

(iii) Gender

(iv) Age

(v) Tenure

(vi) Balance

(vii) Product Number

(ix) Has Credit Card

(x) Active Member

(xi) Estimated Salary

Output is:

Exited (Customer churn)

1. Data Understanding / Pre-processing

Data is made up 10k records and 12 features. The feature of customer ID is dropped from consideration since it is arguably not related to customer churn.

Little correlation is found except for a weak correlation between age and exited.

Data is fairly well split among the different combinations of each feature as shown in the powerpoint.

Very few missing information (null data) is found and therefore the respective rows of data are dropped from the dataset.

Some features observed to have significant outliers, as defined by the default interquartile range, shown in the boxplots.

Ranges of features are found to differ widely among themselves and might exert unintended influence on the model interpretation especially for models that depended on ‘distance’. Some of the features are also observed to have the usual Gaussian like distribution. However, all the features are put through the more common test for Gaussian and turn up to be false. As a result, features are standardised instead of normalised in case of prejudicing the models with Gaussian assumptions.

Finally the output is examined whether it might be imbalanced which will also affect the results and the model evaluation metric used.

1. Machine learning model training

All the classification models are used to compare before deciding on the final one to use.

KNN – Works by clustering K number of points into 1 cluster with the aim of separating each cluster at the maximum distance into separate discernible classification.

SVM – Works by drawing arbitrary lines or planes dividing points into spaces at the maximum possible separation between those lines or planes henceforth creating spaces or classification.

Gaussian Naïve Bayes – Works by calculating the probability of each feature based on the normal distribution and then assigning the class respectively.

Decision Tree – Works by splitting the features including the range of it such that it establish pathways by which the highest probability of either class could exist.

Random Forest – Works based on a collection of decision trees, each unrelated and then collating the output of each tree for its final output.

It is evident from the non-Gaussian behaviour that several of them are unfit for use but we test it nevertheless for their performance as required.

Parameters used are all default until the model is chosen.

Parameters used in Random Forest to try boost its performance are:

1. n\_estimators – the higher the number (of trees), the more generalised the result until it tapers off.
2. Criterion – let computer decide whether to give better weight in lower probabilities in entropy or emphasis higher probability events in Gini
3. Max Depth – reduces the number of branches from the root in a decision tree to reduce overfitting
4. Max\_leaf\_nodes – parameter to prevent underfitting (smaller number) and overfitting (bigger number)
5. Evaluation / Results

Evaluation metric used to evalutate the model is primarily the recall rate of the positive class in the prediction outcome. This is due to the fact that the bank is most interested in finding out customer churn.

Random forest is used finally as:

1. It did not depend on the data as being Gaussian
2. Data is very much un – correlated which works into the advantage of random forest
3. Less afflicted by outliers especially since we noticed quite a few
4. Highest recall score

The model is unlikely overfitted due to average recall score of 0.62. Parameter controlling model from overfitting is also added in during the selection of parameters.

1. Conclusion

The model worked with acceptable recall score and could be used. Perhaps more could be done to enhance the model.

If there is more time, we could find out more about the relationship between age and exited and whether it can be suitably explained by any higher orders of regression, other than the simple linear relationship. This could be done with a scatterplot and if one exists, to use stacked models or ensembling methods for a better model.

Features with significant outliers could also be further analysed and put thru model testing again, in case these models are previously prejudiced by the presence of these outliers.

Data could also be segregated by geography to isolate factors not captured by the data.