A Comparison of Random Forest and Naïve Bayes in Classifying Credit Card Customers

Ethan Chew Wei Xun

1. Problem Description and Motivation

- With competitive and saturated markets, customer retention continues to be an integral part of cost management and revenue generation in any industry, product and portfolio
- This study, however, scopes in specifically on the classification on credit card customers (either as Existing Customer or Attrited Customer)
- Goal is to build two models, namely: Random Forest and Naïve Bayes
- Compare and analyse the performance metrics of both models in classifying credit card customers

2. Dataset, Data Preparation/Processing, and Exploratory Data Analysis

2.1. Context of Dataset

- Dataset: 'Credit Card customers' from Kaggle
- The original dataset is comprised of both existing and attrited customer samples, with 10,127 observations and 22 features, with the features being a mix of categorical (qualitative), discrete and continuous
- Missing values in the original dataset is tagged as 'Unknown'

2.2. Data Preparation/Processing

- 3,046 observations with missing values are removed
- No outliers are removed because either:
  a) All points are between the minimum and maximum of the boxplot; or
  b) The boxplots and histogram plots are mostly skewed and/or suggest a multimodal distribution
- Qualitative categorical features are relabeled with numerical discrete values and non-categorical features are subsequently standardised using its corresponding mean and standard deviation
- After splitting the resulting dataset of 7,081 observations into a training and testing set (80-20 split respectively), the training set is comprised of only 15.41% of attrited customers, as seen in Figure 1
- Synthetic Minority Oversampling Technique (**S**MOTE) will be applied to oversample the minority class (i.e. attrited customer), resulting in 33.33% of attrited customers in the training set, as seen in Figure 2

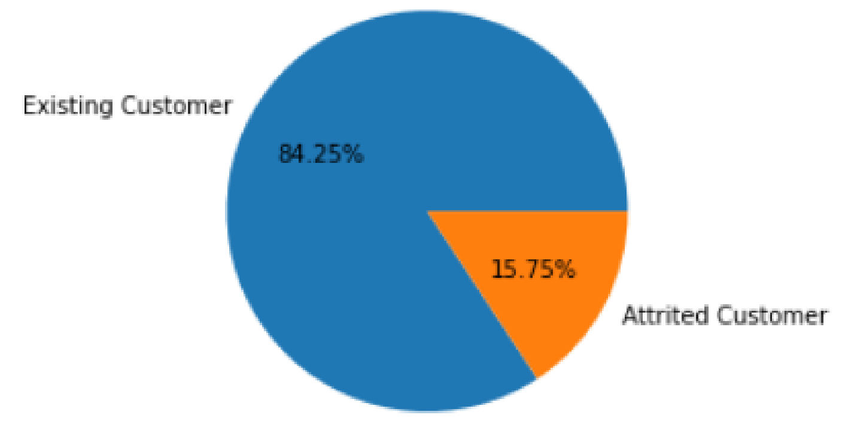


Figure 1: Distribution by Class (Pre-SMOTE)

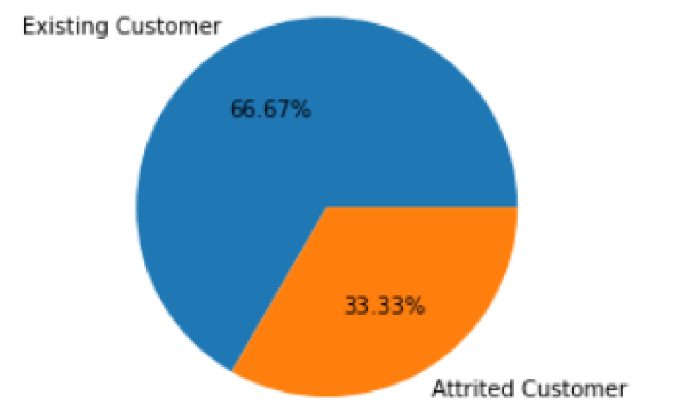


Figure 2: Distribution by Class (Post-SMOTE)

- Without applying SMOTE, the trained model will naturally be biased through over prediction of the majority class and predictive accuracy measures may not be appropriate[1]
- A correlation heatmap and variance inflation factor (VIF) table, as shown in Figure 3 and 4 respectively, is mapped out to assess for any correlation and relationship or dependencies between the features

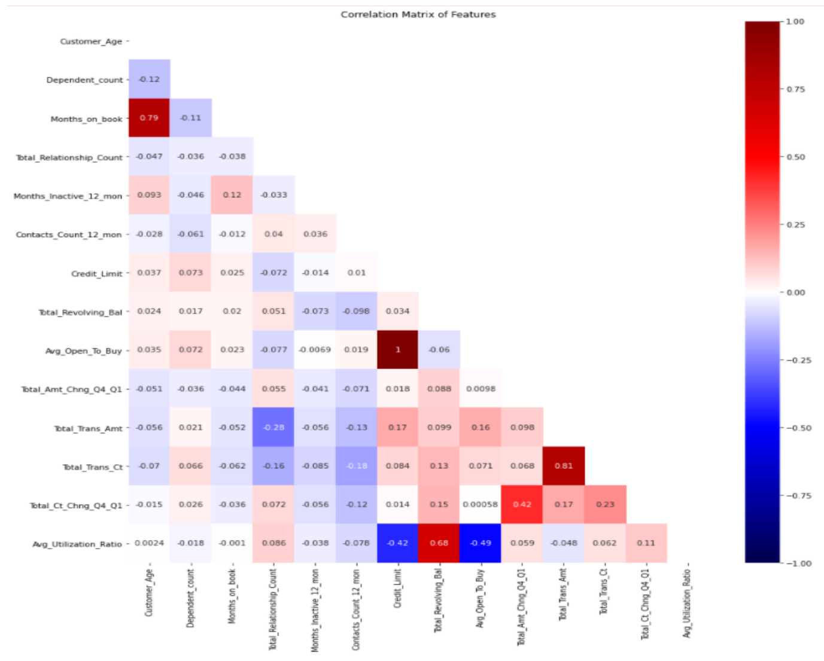
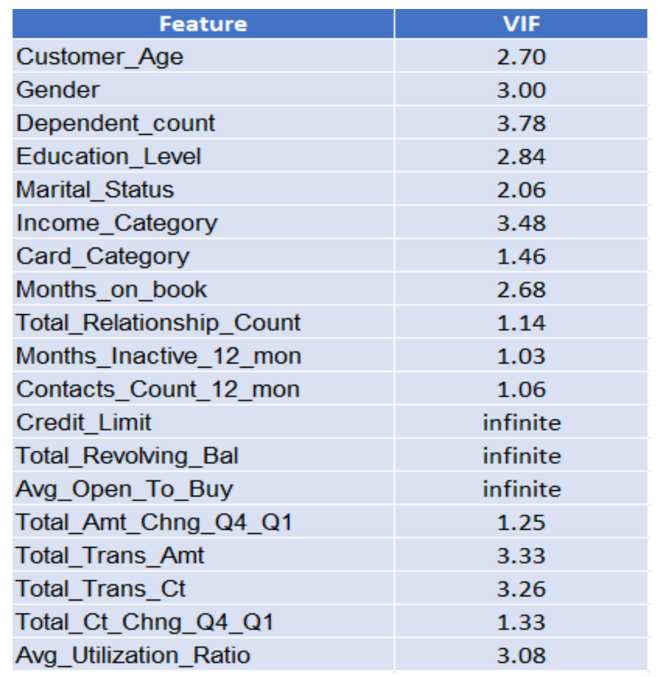


Figure 3: Correlation Heatmap



Feature	VIF
Customer_Age	2.70
Gender	3.00
Dependent_count	3.78
Education_Level	2.84
Marital_Status	2.06
Income_Category	3.48
Card_Category	1.46
Months_on_book	2.68
Total_Relationship_Count	1.14
Months_Inactive_12_mon	1.03
Contacts_Count_12_mon	1.06
Credit_Limit	infinite
Total_Revolving_Bal	infinite
Avg_Open_To_Buy	infinite
Total_Amt_Chng_Q4_Q1	1.25
Total_Trans_Amt	3.33
Total_Trans_Ct	3.26
Total_Ct_Chng_Q4_Q1	1.33
Avg_Utilization_Ratio	3.08

Figure 4: VIF by Features

- The credit limit can be expressed as a sum of the revolving balance and average open to buy, hence a correlation of 1 between the credit limit and average open to buy, and unbounded VIF for those 3 features
- Thus, the average open to buy feature is removed to ensure minimal dependencies between the features

3. Overview of Model Methodology

3.1. Random Forest (RF)

3.1.1. Summary

- RF is a supervised learning method for classification and regression which produces an ensemble of decision trees which are built on bootstrapped samples of the training dataset
- Bootstrapping here samples with replacement and for each tree, at each split, a single feature is selected from a subset of randomly sampled features from the original set of features
- The final prediction, classification in this case, is done through majority voting across the decision trees
- To reduce the risk of overfitting, model complexity, and computation time, hyperparameters (i.e. number of decision trees, tree depth, and number of features to consider at each split) can be tweaked

3.1.2. Advantages

- Data does not require a lot of pre-processing and features can be of different scales and types
- Combination of bootstrapping and ensemble of decision trees reduces the risk of overfitting and variance
- Does not make any underlying assumptions on distributions of features

3.1.2. Disadvantages

- Tradeoff between complexity and interpretability (e.g. through tree depth)
- Computationally complex and expensive in terms of computation time, power and resources required

3.2. Naïve Bayes

3.1.1. Summary

- NB is also a supervised learning method for classification which assumes that each feature are independent from each other, given the class or label
- It learns from the joint and conditional probability distribution over the features, X, and class, Y
- The conditional probability of the class given the observed features (i.e. p(Y|X)) determines what the predicted class will be tagged as (i.e. either existing or attrited customer) and in this study, the argument of the maximum (argmax) is used as the decision rule
- Depending on the type and distribution of the features, a gaussian or multinomial NB can be adopted and prior distributions can be set in accordance to the distribution of the classes

3.1.2. Advantages

- Performs well despite not strictly abiding to the independence assumption amongst features
- Simple to implement and aren't many hyperparameters to tune
- Not data intensive and can be trained effectively over a small training dataset

3.1.2. Disadvantages

- Makes strong assumptions on the independence of features and is not likely to hold in practice
- Outperformed by more complex algorithms and learning methodologies (e.g. Boosted Trees and RF)[2]
- Will assign zero probabilities to unseen categories that are represented within the training set[3]

4. Hypothesis Statement

- Expect RF to marginally outperform NB in terms of accuracy in classifying credit card customers but both RF and NB to perform significantly better than random guessing
- Also expect larger misclassification in the minority class (i.e. attrited customer) as compared to the majority class for both, RF and NB models

5. Training and Evaluation Methodology

5.1. Training Methodology

- The cleaned dataset of 7,081 observations are split into a training and testing using a 80:20 ratio, resulting in a testing set of 1,417 observations which remains unseen until model testing
- SMOTE is applied to the training set, as mentioned in 2.2, resulting in 1,494 additional observations for a total of 7,158 observations
- When training the NB model, k-fold cross validation is adopted to minimise the risk of overfitting, with 10 partitions, and the non-categorical features are binned so that a multinomial NB can be fitted
- K-fold cross validation was not used when training the RF model as bagging or bootstrapping accounts for the risk of overfitting and already adds computational complexity (e.g. time)
- The RF and NB model will be selected through hyperparameter tuning where both RF and NB are iteratively fitted with varying hyperparameters and the best performing one is selected

5.2. Evaluation Methodology and Metric

- The fitted models will be evaluated by the area under its Receiver Operating Characteristic (ROC) curve (AUC), which plots sensitivity (or True Positive Rate (TPR)) against fall-out (or False Positive Rate (FPR))
- Selected models will also be evaluated based on its sensitivity , specificity (or True Negative Rate (TNR)), and harmonic mean (F_1-Score) between precision (or Positive Predictive Value (PPV)) and recall
- The AUC provides a uniform way to compare ROCs while the F_1-score is useful for imbalanced data because false positives (FP) are incorporated with more importance[4], and sensitivity and specificity both provides a more granular view on the true positives(TP) and true negatives (TN)

6. Model Selection and Experimental Results

6.1. Random Forest Model Selection

- A set of hyperparameters, namely the number of decision trees, tree depth and features to consider at each split, are tuned and the best performing RF, based on its AUC, is selected
- The fitted RFs are iteratively fitted over integer values of the hyperparameters, and in particular, up to 10 trees, tree depth of 5 and 5 features to consider at each split are considered
- Figure 5 provides a plot of number of trees, tree depth, and number of features to consider, coloured by its corresponding AUC
- Figure 5 suggests that deeper and a larger number of trees are preferred, along with more features to consider at each split
- The selected model is a RF with 9 trees, tree depth of 4 and 5 features to consider at each split

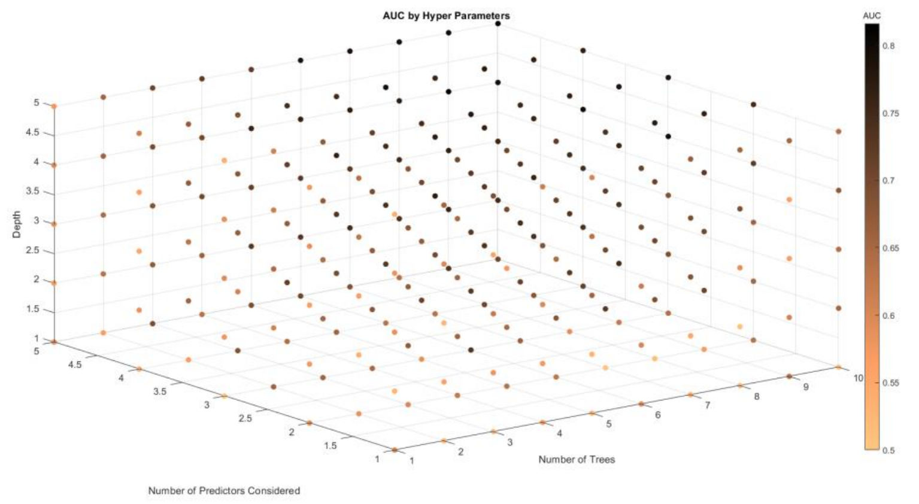


Figure 5: AUC by Hyperparameters - RF

6.2. Naïve Bayes Model Selection

- Similarly when fitting the multinomial NB model, a set of hyperparameters are tuned and the best performing NB model, based on its AUC, is selected
- Here, the non-categorical features are binned to get a view on the relationship between the number of bins and model performance
- The hyperparameters include the number of bins (up to 8 bins) and the prior distribution (empirical and uniform are considered)
- A multinomial NB is iteratively fitted over the hyperparameter values and Figure 6 shows the AUC of each NB, against the number of bins, by the prior distribution
- Figure 6 suggests that larger number of bins are preferred for the empirical distribution while there are diminishing returns from more bins for the uniform distribution
- The selected model is a multinomial NB with 8 bins and an empirical prior distribution

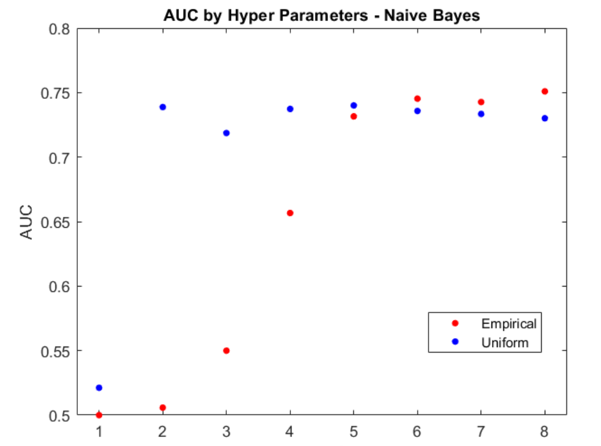
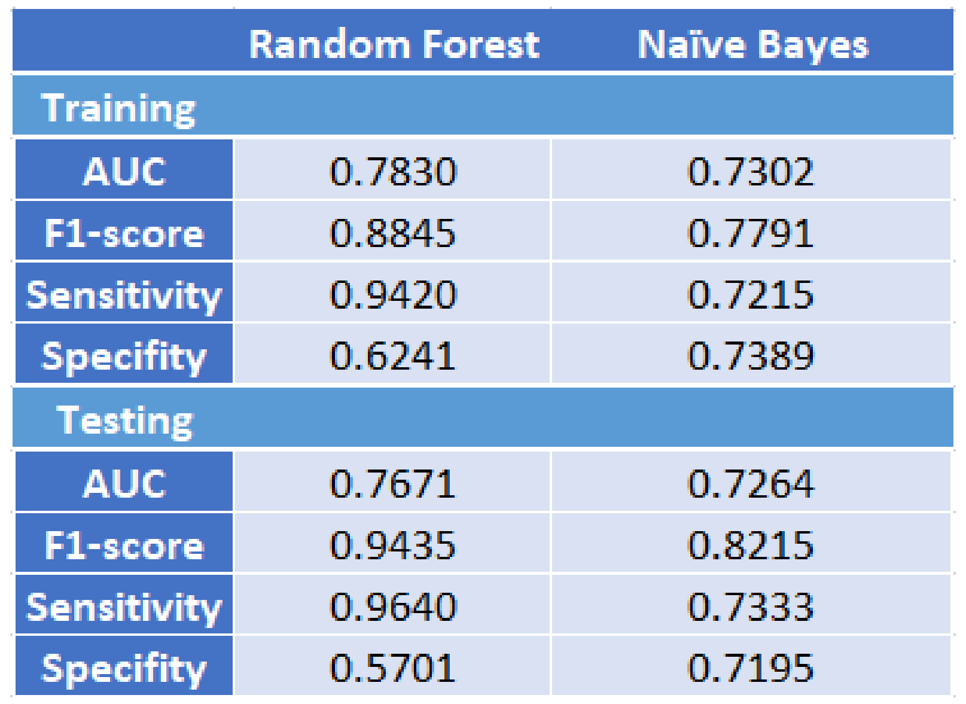


Figure 6: AUC by Hyperparameters - NB

6.3. Experimental Results

- Figures 7 and 8 shows results of the selected models over the training and testing set



	Random Forest	Naïve Bayes
Training		
AUC	0.7830	0.7302
F1-score	0.8845	0.7791
Sensitivity	0.9420	0.7215
Specifity	0.6241	0.7389
Testing		
AUC	0.7671	0.7264
F1-score	0.9435	0.8215
Sensitivity	0.9640	0.7333
Specifity	0.5701	0.7195

Figure 7: Performance of Each Model Across Datasets

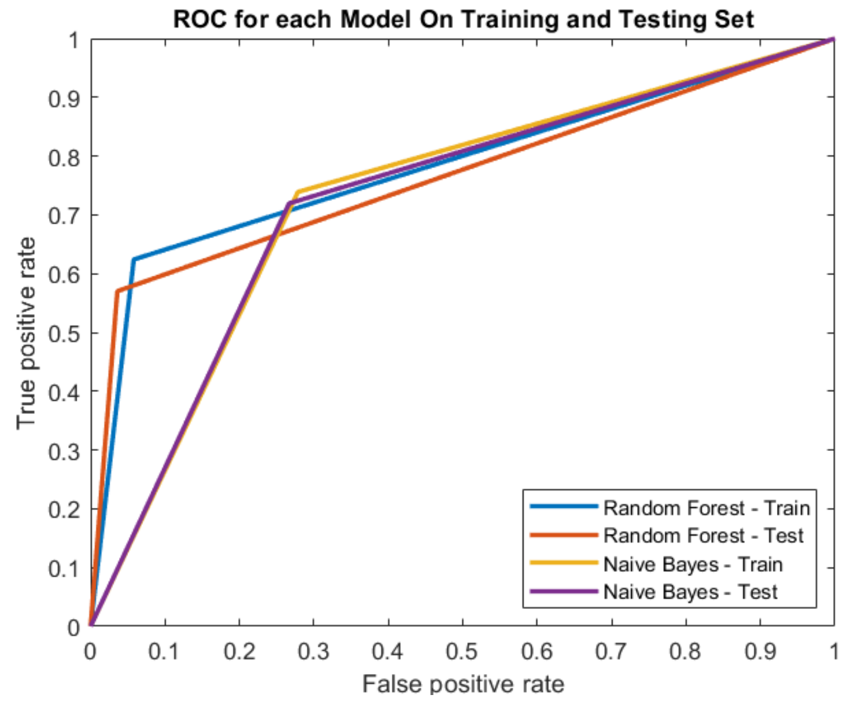


Figure 8: ROC for Each Model Across Datasets

7. Analysis and Evaluation of Results

- The ROC plots the relationship and displays the tradeoff between the TPR and FPR at different decision thresholds but by looking solely at the ROC curves for both models, it is not apparent which model performs better as neither are strictly closer to (0,1) of the ROC graph (i.e. the TPR and FPR is not higher and lower respectively over all possible decision thresholds)
- Thus, the AUC may serve as a better and more convenient comparative measure as it summarises the location of the ROC and is independent of the decision threshold[5], and from Figure 7, the AUC of the RF is larger than that of NB over training and testing – a more discriminatory model in RF as compared to NB
- The F_1-score differs from the AUC as it assesses the tradeoff between precision and recall (or sensitivity) but concurs with the above as the F_1-scores of the RF are again larger than that of NB over training and testing, indicating a more discriminatory model in RF as compared to NB
- This could be a result of either, or a combination of, a difference in complexity of model between RF and NB, inability to strictly abide to the independence assumption between features and/or information loss (e.g. of the probability distribution) from the binning of non-categorical features
- However, despite both AUC and F_1-score results coinciding, it should be noted that the difference in magnitude between the F1-scores of both models greatly outweighs the corresponding value between AUC's, suggesting that the NB model may tend closer towards making predictions of the negative class (i.e. attrited customer) as compared to RF
- Indeed, when looking at the sensitivity and specificity of RF, the sensitivity in training and testing significantly outweighed its specificity, implying there's more misclassification in the negative class (i.e. attrited customer) as compared to the positive class (i.e. existing customer)
- Conversely, NB struck more a middle ground by sacrificing accurate classification in the positive class (i.e. existing customer) for less misclassification in the negative class (i.e. attrited customer), with marginal differences between sensitivity and specificity
- Looking solely at AUC, F_1-score and sensitivity, one may be quick to conclude that RF is better than NB but when viewed under the lens of specificity, the conclusion may not be as clear cut
- With the context of the data and classification problem, a more prudent individual may enjoy results similar to that of the NB model to get a more conservative view of customers who are about to leave the credit card portfolio so that pre-emptive measures can be taken, while a more resource rich individual may prefer a results similar to that of the RF instead so as to not offer products to customers who are at no risk of leaving to begin with

8. Lessons Learned and Future Work

8.1. Lessons Learned

- Features and the dataset in general should be well understood so that expert judgement and manual overlay can be applied in tandem with statistical analysis
- Model performance can be and should be assessed through various metrics and model preferences (for the most part) is subjective and depends heavily on the intention of the model and context of the data

8.2. Future Work

- Further account for class imbalances in the training set through undersampling the majority class[1] (i.e. existing customer) along with oversampling the minority class (i.e. attrited customer)
- Employ Principal Component Analysis (PCA) to reduce the dimensionality of the features and explore feature importance to get a view of key features that drive the classification of credit card customers
- Refit a NB model but without binning non-categorical features and specifying a probability distribution for each feature to mitigate any information loss from the binning process
- Incorporate Laplace Smoothing into the NB model to account for any potential zero probabilities[3] and measure other performance metrics (e.g. training and testing time) that would be of use in practice

9. References

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", 2011.

[2] R. Caruana, A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms" *In Proceedings of the 23rd international conference on Machine learning*, 2006, pages 161–168.

[3] K. M. Al-Aidaroos, A. A. Bakar and Z. Othman, "Naïve bayes variants in classification learning," 2010 *International Conference on Information Retrieval & Knowledge Management (CAMP)*, 2010, pages 276-281.

[4] Advances in Intelligent Data Analysis XVIII: 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings. (2020). Germany: Springer International Publishing, pages 458-461.

[5] Hajian-Tilaki K., "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation." *Caspian J Intern Med*. 2013;4(2): pages 627-635.