Project Report: Predicting Customer Attrition in Banking

[Noah Bateman/nb626] – *Models: LDA, CART, Boosting, Deep Learning*

[Tom Neame/tjjn2] – *Models: GLM, KNN, Random Forest, SVM*

Contents

Abstract

This project applies a range of supervised machine-learning models to predict customer churn using demographic, relationship, and behavioural data from a credit-card portfolio. After data inspection and multicollinearity checks, eight model classes were fitted: logistic regression, LDA, KNN, SVM, decision trees, random forests, gradient boosting, and deep learning. Models were evaluated using out-of-sample AUC, sensitivity, specificity, and threshold-optimised performance. Behavioural variables, especially transaction count and recent activity changes, emerged as the strongest predictors of churn. Linear models underperformed due to structural underfitting, while boosted trees achieved the highest discriminatory power. The analysis shows that retention-critical metrics, such as sensitivity, improve substantially through non-linear modelling and threshold tuning. Overall, the results highlight the value of behaviour-based features and ensemble learning when assessing churn risk in financial services.

1. Problem Formulation & Motivation

Motivation: Customer attrition (churn) is a typical business case challenge for service-based industries like banking. Identifying customers who are likely to leave allows the business to actively intervene, improving customer retention and reducing the costs associated with acquiring new customers.

Research Question: *Can machine learning algorithms effectively predict credit card customer attrition using historical transaction and demographic data? Which class of models; linear, tree-based, or deep-learning, provides the best predictive performance for this specific dataset?*

- **Objectives:**
    - To explore the dataset and identify key predictors of attrition.
    - To implement and optimise a variety of supervised learning models.
    - To evaluate model performance using metrics suitable for imbalanced classification tasks (e.g., Sensitivity, AUC).

2. Data Overview

Source: The "Credit Card Customers" dataset from Kaggle.

Characteristics: 10,127 observations with 21 initial features. The target variable is Attrition_Flag (approx. 16% attrition rate – low population, however the large number of observations counters the small percentage of churned individuals).

Data Preparation:

- **Cleaning:** Removed unique identifiers (CLIENTNUM) to prevent overfitting on IDs.
- **Multicollinearity:** Correlation analysis revealed high collinearity between [Credit_Limit and Avg_Open_To_Buy], [Total_Trans_Amt and Total_Trans_Ct], and [Customer_Age and Months_on_book].
- **Feature Engineering:** For linear models sensitive to multicollinearity (LDA, GLM), we removed highly correlated features and created the variable `dat_no_corr`. For non-linear models (Trees, Deep Learning), we retained features to maximize information gain.

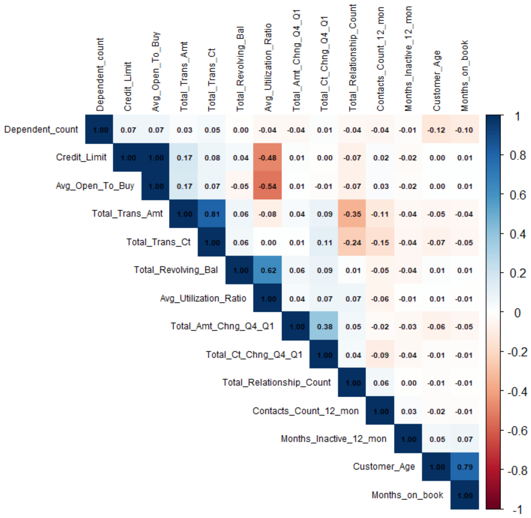


Figure 1: Correlation Heatmap with correlated variables

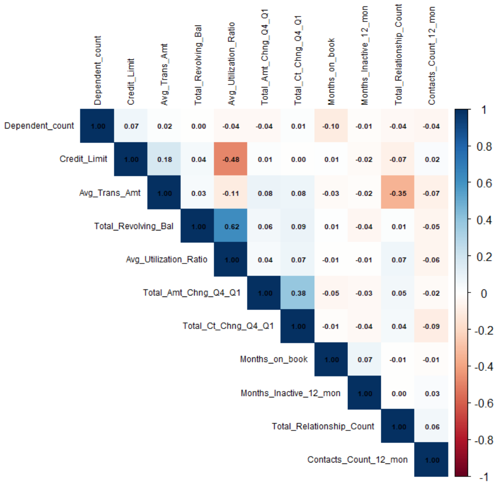


Figure 2: Without highly correlated variables

3. Methodology

We decided to move forward with 8 models to fully understand and predict the dataset.

- **Linear Baselines:** We selected **Logistic Regression (GLM)** and **Linear Discriminant Analysis (LDA)** to establish a baseline performance and test if the decision boundary between classes is linear.
- **Non-Parametric & Distance-Based: KNN** and **Support Vector Machines (SVM)** were chosen to capture non-linear boundaries without assuming a specific functional form for the data distribution.
- **Tree-Based Methods:** We progressed from a single **Classification Tree (CART)** for its interpretability to **Random Forests (RF)** and **Gradient Boosting (GBM)** to leverage ensemble learning for reducing variance and bias, maximising prediction capability.
- **Deep Learning:** A feed-forward **Neural Network** was implemented to investigate if a complex architecture could learn feature representations that outperform traditional machine learning methods on this tabular data.

4. Analysis & Interpretation

This section compares the performance of the models.

4.1 Linear Baselines (GLM & LDA)

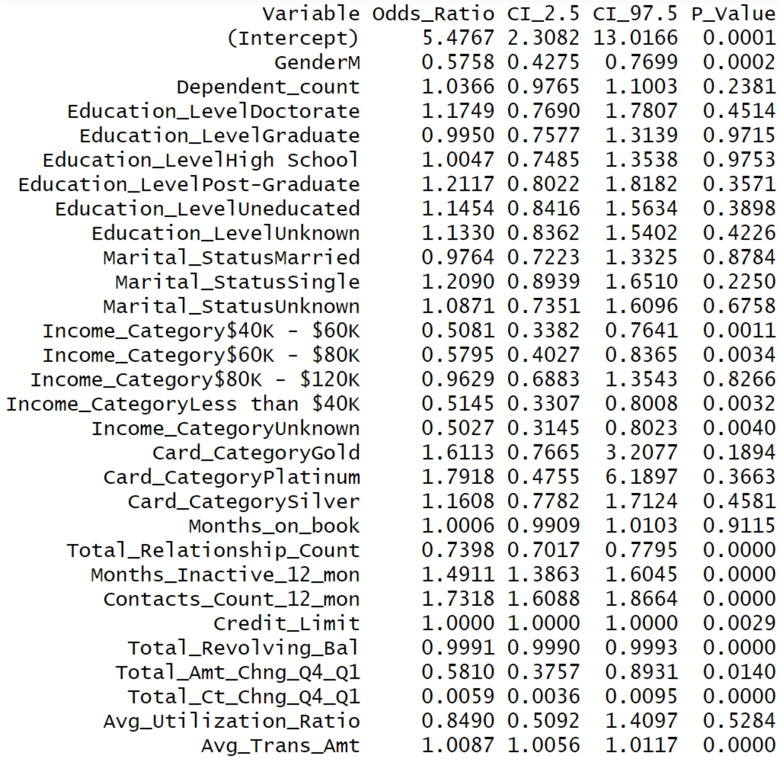
Logistic Regression (GLM):

- **Performance:** The GLM model achieved an AUC of **0.837**. However, the default classification threshold proved too conservative. By tuning the threshold down to **0.2**, we improved the Sensitivity to **68.0%**, but this came at the cost of overall Accuracy, which dropped to **80.6%** (below the No Information Rate).

- **Interpretation:** The trade-off observed here highlights the limitations of linear boundaries for this dataset. To achieve a viable detection rate (Sensitivity > 65%), the linear model is forced to sacrifice significant precision (Positive Prediction Value approx. 42%), resulting in a high rate of false alarms. This inefficiency suggests that the drivers of attrition are non-linear, pushing the use of more complex tree-based classifiers.

- **Inference and Odds ratios:** Beyond prediction, we analysed the model coefficients (Odds Ratios) and their P-values to understand the statistically significant drivers of churn:
  - Contacts Count (OR = 1.73, p < 0.001): This is a critical behavioural indicator. The model estimates that for every single additional time a customer contacts the bank, their odds of churning increase by **73%** (95% CI: [1.60, 1.86]). This confirms that frequent contact is a leading indicator of customer dissatisfaction.
  - Total Relationship Count (OR = 0.74, p < 0.001): Holding more products is a protective factor. For every additional product held, the odds of churning decrease by **26%**, validating the importance of cross-selling strategies.
  - Revolving Balance (OR = 0.999, p < 0.001): This variable acts as a "stickiness" factor. While the effect per dollar is small, the trend is highly significant: customers who carry a balance (debt) are less likely to leave than those who pay off their card in full every month.
  - Months Inactive (OR = 1.49, p < 0.001): Passive disengagement is a major risk; for every month a customer is inactive, their churn risk increases by nearly **50%**.
  - Gender (OR = 0.58, p < 0.001): The model identifies a significant demographic trend: Male customers are approximately **42% less likely** to churn than Female customers (the reference category), holding all other variables constant.

- Change in Transaction Count (OR = 0.006, p < 0.001): The variable Total_Ct_Chng_Q4_Q1 has an extremely low odds ratio. This implies that customers who increase their transaction activity (Q4 vs Q1) are essentially immune to churn. Engagement is the strongest protective factor.
- Insignificant Features: We found that Marital_Status (p > 0.20) and Education_Level (p > 0.35) were statistically insignificant. This suggests that while gender plays a role, a customer's family or education background does not meaningfully predict whether they will leave the bank; their account activity is far more important.



```
                       Variable Odds_Ratio CI_2.5 CI_97.5 P_Value
                    (Intercept)     5.4767 2.3082 13.0166  0.0001
                        GenderM     0.5758 0.4275  0.7699  0.0002
                Dependent_count     1.0366 0.9765  1.1003  0.2381
       Education_LevelDoctorate     1.1749 0.7690  1.7807  0.4514
        Education_LevelGraduate     0.9950 0.7577  1.3139  0.9715
     Education_LevelHigh School     1.0047 0.7485  1.3538  0.9753
   Education_LevelPost-Graduate     1.2117 0.8022  1.8182  0.3571
      Education_LevelUneducated     1.1454 0.8416  1.5634  0.3898
         Education_LevelUnknown     1.1330 0.8362  1.5402  0.4226
          Marital_StatusMarried     0.9764 0.7223  1.3325  0.8784
           Marital_StatusSingle     1.2090 0.8939  1.6510  0.2250
          Marital_StatusUnknown     1.0871 0.7351  1.6096  0.6758
     Income_Category$40K - $60K     0.5081 0.3382  0.7641  0.0011
     Income_Category$60K - $80K     0.5795 0.4027  0.8365  0.0034
    Income_Category$80K - $120K     0.9629 0.6883  1.3543  0.8266
Income_CategoryLess than $40K       0.5145 0.3307  0.8008  0.0032
         Income_CategoryUnknown     0.5027 0.3145  0.8023  0.0040
              Card_CategoryGold     1.6113 0.7665  3.2077  0.1894
          Card_CategoryPlatinum     1.7918 0.4755  6.1897  0.3663
            Card_CategorySilver     1.1608 0.7782  1.7124  0.4581
                 Months_on_book     1.0006 0.9909  1.0103  0.9115
       Total_Relationship_Count     0.7398 0.7017  0.7795  0.0000
         Months_Inactive_12_mon     1.4911 1.3863  1.6045  0.0000
          Contacts_Count_12_mon     1.7318 1.6088  1.8664  0.0000
                   Credit_Limit     1.0000 1.0000  1.0000  0.0029
            Total_Revolving_Bal     0.9991 0.9990  0.9993  0.0000
           Total_Amt_Chng_Q4_Q1     0.5810 0.3757  0.8931  0.0140
            Total_Ct_Chng_Q4_Q1     0.0059 0.0036  0.0095  0.0000
          Avg_Utilization_Ratio     0.8490 0.5092  1.4097  0.5284
                  Avg_Trans_Amt     1.0087 1.0056  1.0117  0.0000
```

Figure 3: GLM odds ratios and p-values output table

Linear Discriminant Analysis (LDA):

- **Performance:** The LDA model achieved an accuracy of **87.3%**, which is statistically significant compared to the No Information Rate (83.9%). However, its Sensitivity was very low at **33.2%**.
- **Interpretation:** The low sensitivity suggests that the classes are not linearly separable. The model struggled to identify the minority class (churners), likely because the relationship between transaction behaviours and attrition is non-linear. While the model effectively identified stable customers (Specificity 97.6%), it missed two-thirds of the actual attrition cases.
- **Conclusion:** LDA assumes that predictors are normally distributed and share a common covariance matrix. Our EDA showed skewed distributions (e.g., Credit_Limit) and potential non-linear relationships (e.g., Total_Trans_Ct). The poor sensitivity confirms that a simple linear boundary is not sufficient to capture the "high-risk" category of customers, further pushing the need for non-parametric approaches.

4.2 Non-Linear Classifiers (KNN & SVM)

K-Nearest Neighbours (KNN):

- **Performance**: KNN was the poorest performing model with an Accuracy of **81.7%**, which is below the No Information Rate (84.3%). Sensitivity was critically low at **24.5%** and the AUC was 0.667.

- **Interpretation:** This poor performance indicates that the model was overwhelmed by the majority class. Furthermore, the distance-based metric likely failed because of the differing scales of variables (e.g., Utilisation Ratio vs. Balance). Despite standardisation attempts, the high dimensionality of the data meant simple distance was ineffective for separating classes, and so the model failed to detect the minority class, incorrectly labelling 361 out of 478 churners as safe.

Support Vector Machines (SVM):

- **Performance**: The SVM model showed significant improvement over the linear baselines and KNN, achieving an Accuracy of **93.3%** and a strong AUC of **0.963**. Sensitivity increased to **67.4%**.

- **Interpretation:** The use of the Radial Basis Function (RBF) kernel allowed the model to capture non-linear boundaries better than GLM. While highly precise (Specificity 98.1%), it still missed approximately 33% of churners, making it a reliable but slightly conservative classifier.

- We performed a Grid Search to optimise the hyperparameters. The tuning process validated that the default **Cost (1.0)** and a **Gamma of 0.1** yielded the optimal balance of bias and variance. While more aggressive costs (10.0) were tested, they did not yield significant improvements in Sensitivity, confirming the stability of the RBF kernel on this dataset.

4.3 Tree-Based Ensembles (CART, RF, GBM)

Classification Tree (CART):

- **Structure:** The initial (fully branched) tree showed that the most important variables used in the tree construction were Total_Trans_Ct, Total_Revolving_Bal, Total_Relationship_Count, Total_Ct_Chng_Q4_Q1, and Total_Trans_Amt. This shows that besides the 'relationship count' variable, all the predictors were behavioural rather than demographic. We then used both deviance and graphical methods to find the optimal size for the tree model, balancing interpretability and model predictive accuracy. The pruned tree (size 9) identified Total_Trans_Ct as the primary splitting variable. Customers with fewer than 58 transactions were immediately flagged as high risk.

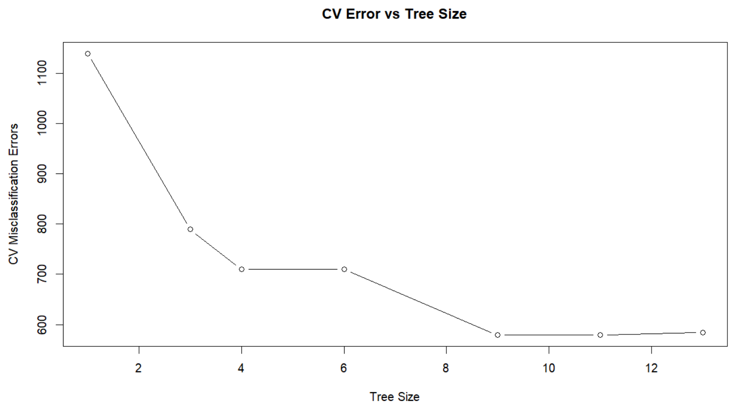


Figure 4: Tree size against CV Misclassification Errors (size 9 selected)

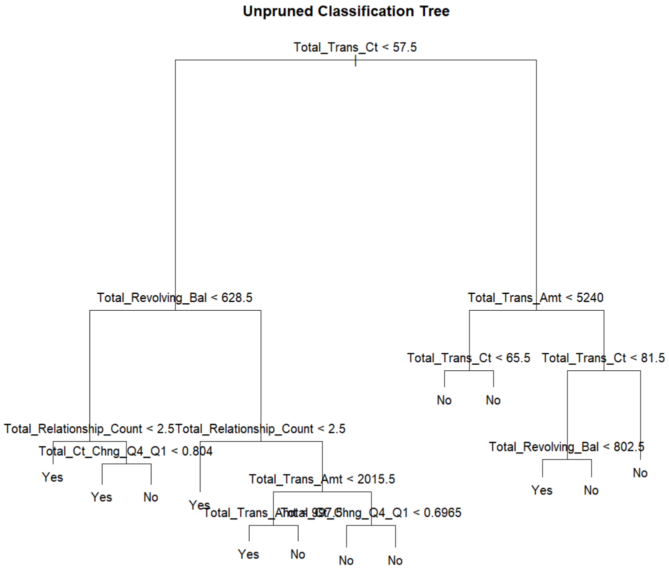

Unpruned Classification Tree
Total_Trans_Ct < 57.5
Total_Revolving_Bal < 628.5
Total_Trans_Amt < 5240
Total_Trans_Ct < 65.5
Total_Trans_Ct < 81.5
No
No
Total_Relationship_Count < 2.5
Total_Relationship_Count < 2.5
Total_Ct_Chng_Q4_Q1 < 0.804
Total_Revolving_Bal < 802.5
Yes
Yes
No
Yes
Total_Trans_Amt < 2015.5
No
Yes
No
Total_Trans_Amt < 997.5
Total_Ct_Chng_Q4_Q1 < 0.6965
Yes
No
No
No

Figure 5: Unpruned Classification Tree (all branches)

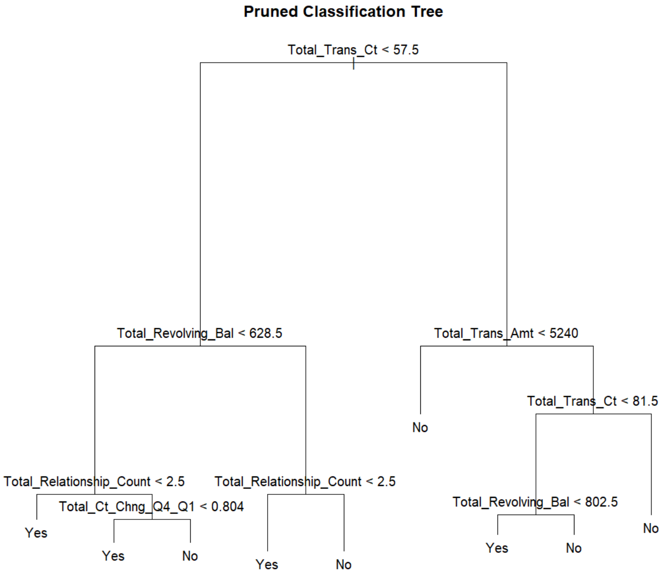

Pruned Classification Tree
Total_Trans_Ct < 57.5
Total_Revolving_Bal < 628.5
Total_Trans_Amt < 5240
No
Total_Trans_Ct < 81.5
Total_Relationship_Count < 2.5
Total_Relationship_Count < 2.5
Total_Ct_Chng_Q4_Q1 < 0.804
Total_Revolving_Bal < 802.5
Yes
No
Yes
No
Yes
No
Yes
No

Figure 6: Pruned Classification Tree (size 9)

- **Performance:** The tree model significantly outperformed the linear baseline, achieving a **Sensitivity of 72.5%** and an **AUC of 0.945**. This demonstrated further the value of non-linear segmentation for this dataset.
- **Handling Multicollinearity:** Despite the high correlation between Total_Trans_Ct and Total_Trans_Amt, the tree automatically handled this by selecting the stronger predictor for the primary split. This validated the decision to retain correlated features for tree-based models.

Random Forest (RF):

- **Performance**: Random Forest proved to be the second most robust model in our testing. It achieved a dominant Accuracy of 96.1% and an AUC of 0.987. Most importantly for the business case, it achieved a Sensitivity of 82.6% without needing manual threshold tuning (unlike GLM).
- **Interpretation:** By aggregating 500 decorrelated trees, the model successfully reduced variance and overfitting. It correctly identified 395 churners while maintaining a very high Specificity (98.6%), making it one of the most balanced models for deployment among those tested.
- We moved from a single Classification Tree to a Random Forest to reduce the variance associated with individual trees. We utilized 500 trees (ntree=500) to ensure the error rate stabilised. Crucially, we used mtry=4 (approximately the square root of predictors – 17). By forcing each split to consider only a random subset of 4 variables, the model decorrelates the trees, preventing strong predictors like Total_Trans_Ct from dominating every single tree and allowing the model to learn more subtle patterns.

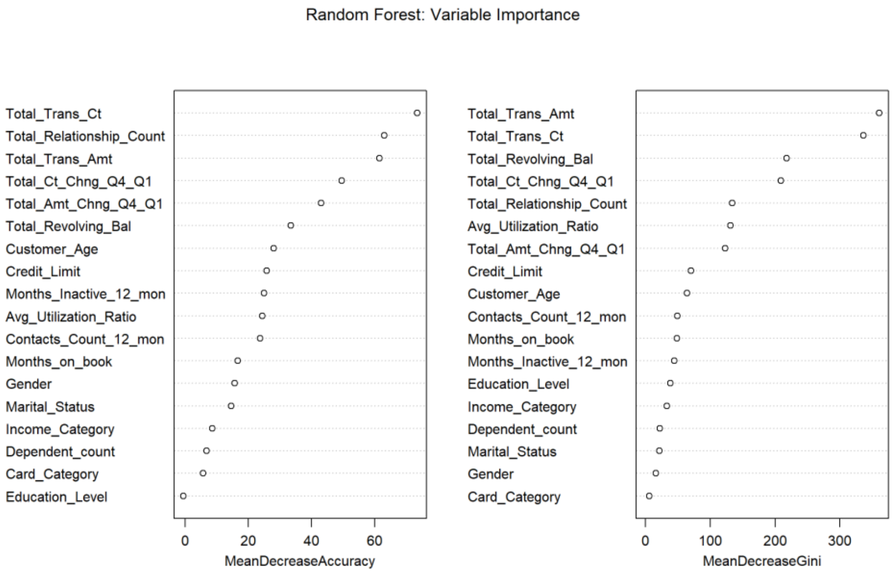


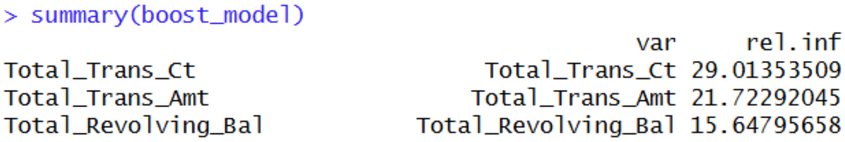
Figure 7: Random Forest variable importance plot using Gini index

- Variable Importance - while Random Forests are often considered "black boxes," we analysed the Mean Decrease in Gini Index to understand which variables were driving the decisions:
  - **Total_Trans_Ct (Transaction Count):** This was the single most important predictor. Customers with low transaction counts are the highest churn risk.
  - **Total_Revolving_Bal (Revolving Balance):** Consistent with the GLM findings, this confirms that customers who do not carry a balance (pay off fully or are inactive) are at higher risk of leaving.
  - **Total_Trans_Amt (Transaction Amount):** The interaction between how often a customer spends versus how much they spend provides a clear signal for attrition.

Gradient Boosting (GBM):

- **Performance:** GBM was the top-performing model overall, achieving a **Sensitivity of 88.1%** and a near-perfect **AUC of 0.993**.

- **Interpretation:** By correcting the errors of previous trees, GBM captured interactions that other models missed. The "Relative Influence" analysis confirmed that customer behaviour (Total_Trans_Ct, Total_Trans_Amt, Total_Revolving_Bal) accounts for **~67%** of predictive power, while demographics (Age, Gender) are negligible.



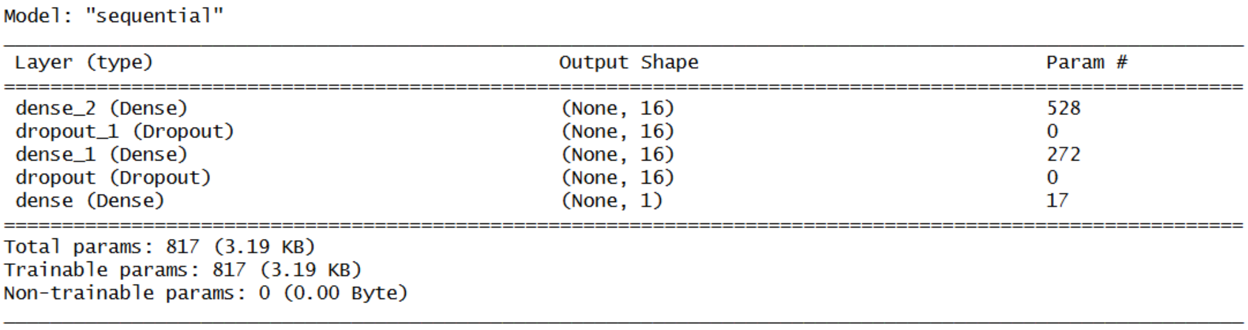
```
> summary(boost_model)
                                            var     rel.inf
Total_Trans_Ct                   Total_Trans_Ct 29.01353509
Total_Trans_Amt                 Total_Trans_Amt 21.72292045
Total_Revolving_Bal         Total_Revolving_Bal 15.64795658
```

Figure 8: Variable Importance Plot for Gradient Boosting Model (Top Predictors)

- **Methodological Note:** We increased model complexity (5,000 trees) to maximize performance. The substantial jump in sensitivity (from 72% in CART to 88% in GBM) proves that the relationship between transaction behaviour and churn is highly non-linear and benefits significantly from ensemble learning.

4.4 Deep Learning

- **Architecture:** We implemented a lightweight feed-forward neural network (817 parameters) with dropout regularisation to prevent overfitting on this relatively small dataset (approx. 10k rows).



```
Model: "sequential"
________________________________________________________________________________________________
 Layer (type)                            Output Shape                          Param #
================================================================================================
 dense_2 (Dense)                         (None, 16)                            528
 dropout_1 (Dropout)                     (None, 16)                            0
 dense_1 (Dense)                         (None, 16)                            272
 dropout (Dropout)                       (None, 16)                            0
 dense (Dense)                           (None, 1)                             17
================================================================================================
Total params: 817 (3.19 KB)
Trainable params: 817 (3.19 KB)
Non-trainable params: 0 (0.00 Byte)
________________________________________________________________________________________________
```

Figure 9: Deep Learning Model Architecture

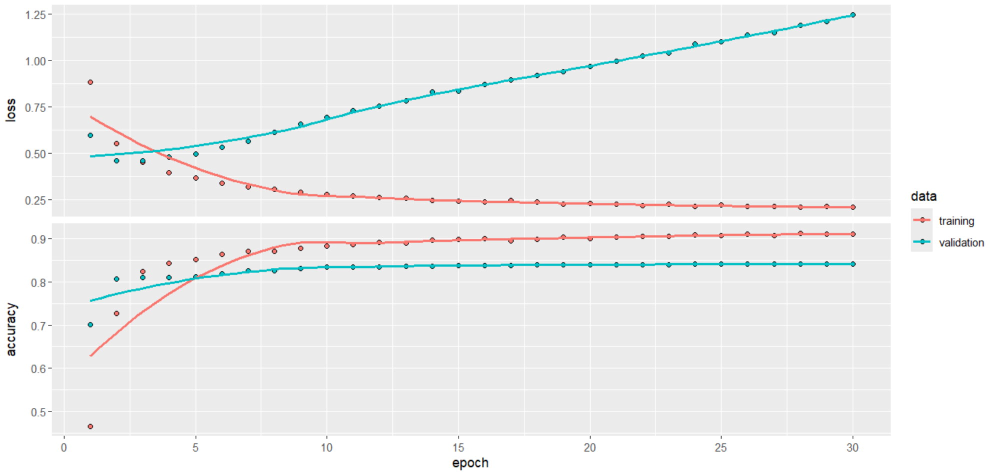


Figure 10: The training history plot reveals early divergence between training and validation loss, confirming that despite regularisation, the model began to overfit the training data after approximately 10 epochs

- **Optimisation:** The default classification threshold (0.5) resulted in low sensitivity (50%). By analysing the ROC curve and manually tuning the threshold to **0.2**, we improved Sensitivity to **70.9%**, making the model viable for identifying churners.

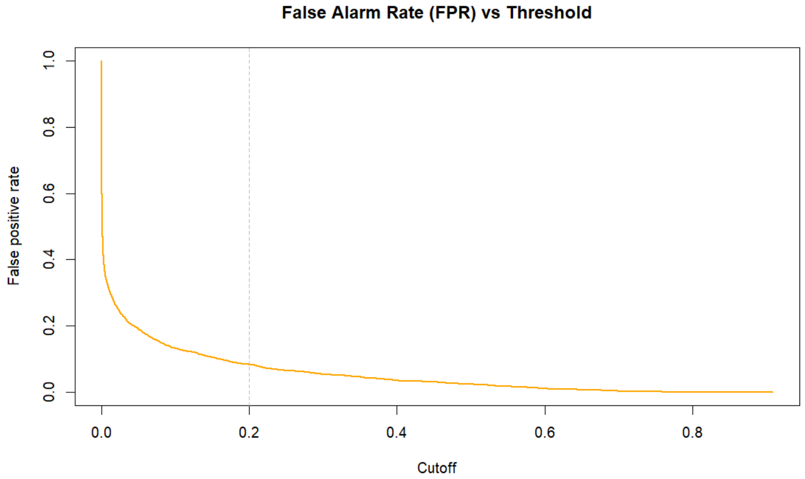


Figure 11: Cutoff against False positive rate; any lower than 0.2 increases FPR substantially

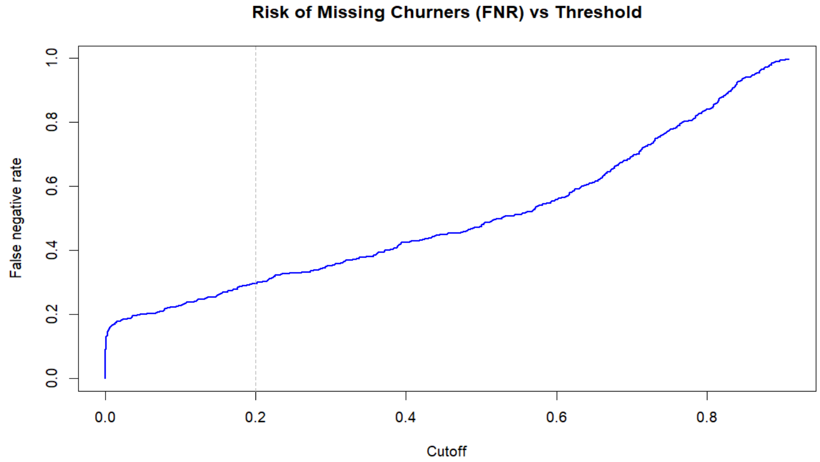
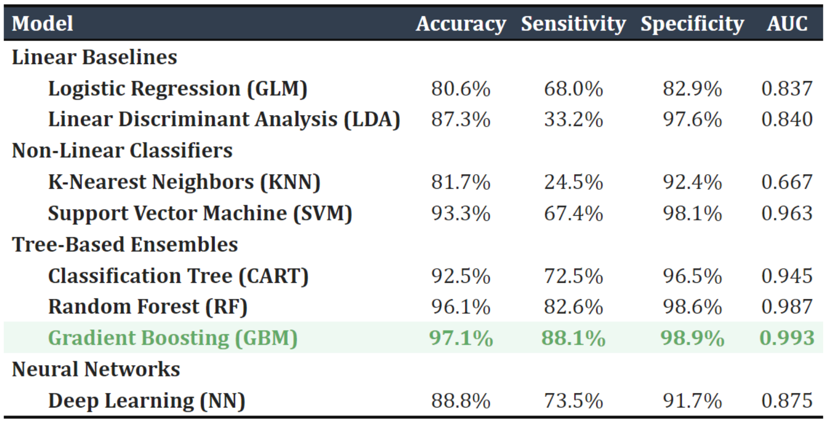


Figure 12: Cutoff against False negative rate; cutoff of 0.2 keeps rate down whilst increasing sensitivity

- **Result:** With a final **AUC of 0.875**, Deep Learning outperformed the linear models and non-linear classifiers but did not match the performance of the tree-based ensemble methods (CART, GBM, RF). This reinforces the finding that Neural Networks often struggle to beat Gradient Boosting on structured, tabular data where feature engineering is important for model success.

5. Critical Evaluation & Conclusion



Model	Accuracy	Sensitivity	Specificity	AUC
Linear Baselines				
Logistic Regression (GLM)	80.6%	68.0%	82.9%	0.837
Linear Discriminant Analysis (LDA)	87.3%	33.2%	97.6%	0.840
Non-Linear Classifiers				
K-Nearest Neighbors (KNN)	81.7%	24.5%	92.4%	0.667
Support Vector Machine (SVM)	93.3%	67.4%	98.1%	0.963
Tree-Based Ensembles				
Classification Tree (CART)	92.5%	72.5%	96.5%	0.945
Random Forest (RF)	96.1%	82.6%	98.6%	0.987
Gradient Boosting (GBM)	**97.1%**	**88.1%**	**98.9%**	**0.993**
Neural Networks				
Deep Learning (NN)	88.8%	73.5%	91.7%	0.875

Figure 13: Model Comparison Table of key metrics, highlighting most successful model

Champion Model:

- **Gradient Boosting (GBM)** is selected as the final model. It provided the highest discriminatory power (AUC 0.993) and the best balance of Sensitivity and Specificity. It successfully captured complex interactions between transaction frequency and transaction amount.

Alternative Recommendation:

- If model interpretability is a priority over raw performance, the **Classification Tree (CART)** is a strong alternative. It offers decent sensitivity (72.5%) while providing a clear, visual set of decision rules that can be easily explained to non-technical stakeholders.

Limitations:

- **Deep Learning:** While effective, it required significant preprocessing (scaling, one-hot encoding) and tuning for a result that was ultimately inferior to the out-of-the-box performance of tree ensembles.

- **Data Limitations:** The extremely high performance of GBM suggests the variables are highly predictive. Future work should validate this model on a different time-period dataset to ensure the transaction patterns are stable and reproducible over time.

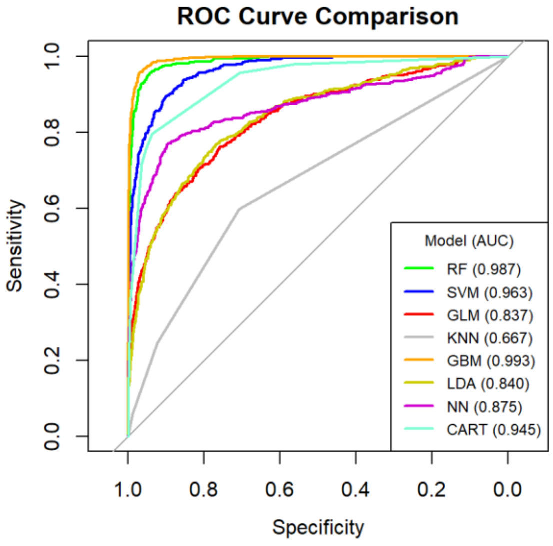


Figure 14: Plot of combined ROC curves for all models

Implication for a credit card issuer (Bank):

From a commercial banking perspective, model selection cannot be based on "Accuracy" alone due to the differing costs of errors. In the context of customer attrition, standard performance metrics translate directly to financial outcomes:

- **Sensitivity:** This represents the model's ability to catch customers before they leave. For the bank, this is the most critical metric. Low sensitivity means the bank is failing to identify at-risk accounts, resulting in the direct loss of customers.
- **Specificity:** This represents the model's ability to correctly ignore satisfied customers. A low specificity implies a "False Alarm," where the bank wastes marketing budget (e.g., retention bonuses) on customers who were never going to leave.
- **Accuracy:** While useful, accuracy can be misleading in imbalanced datasets. Since only 16% of customers churn, a naive model could achieve 84% accuracy by simply predicting "No Churn" for everyone; a result that provides zero business value.
- **AUC (Area Under the Curve):** This measures the model's overall ability to rank customers by risk. A high AUC means that if we pick a random churner and a random non-churner, the model successfully assigns a higher risk score to the churner.

Effects of Model Errors:

- The cost of false negatives (Missed Churners): If the model fails to identify a customer about to leave (as seen with KNN and LDA), the bank loses the customer, years of interest payments and fees. This is the most expensive error.
- The cost of false positives (False Alarms): If the model incorrectly flags a happy customer as "at-risk" (as seen when we tuned GLM), the bank might needlessly offer a retention bonus or discount. This marketing cost is negligible compared to the loss of a customer.

Therefore, our preference for **GBM** and **Random Forest** is driven by their high sensitivity (88.1% and 82.6%). While a linear model might have high "Accuracy" by simply guessing everyone is safe, a high-sensitivity model ensures the bank can intervene with the majority of at-risk accounts. Even if this results in some wasted retention offers (lower specificity) or confused customers who receive a letter incentivising them to stay when they are not going to leave; the return on investment of saving high-value customers outweighs the cost of the retention campaign.

6. Appendix

Target Variable

Attrition_Flag
Type: Factor (Yes/No)
Indicates whether the customer has closed their credit-card account ("Attrited") or remains active ("Existing").

Demographic Variables

Customer_Age
Age of the customer in years. Removed from linear models due to high correlation with Months_on_book.

Gender
Sex of the account holder (Male/Female).

Dependant_count
Number of dependants financially supported by the customer.

Education_Level
Highest education level attained (High School, Graduate, Post-Graduate, Doctorate, Uneducated, Unknown).

Marital_Status
Marital category (Married, Single, Divorced, Unknown).

Income_Category
Annual income bracket (e.g. <40K, 40–60K, 60–80K, 80–120K, 120K+).

Product & Relationship Variables

Card_Category
Type of credit card held (Blue, Silver, Gold, Platinum).

Months_on_book
Duration (in months) the customer has held the account.

Total_Relationship_Count
Number of products the customer holds with the bank.

Behavioural Variables (Last 12 Months)

Months_Inactive_12_mon
Number of inactive months within the past year.

Contacts_Count_12_mon
Customer-initiated service contacts over the last 12 months.

Credit_Limit
Maximum credit available to the customer.

Total_Revolving_Bal
Outstanding revolving balance (carried month-to-month).

Avg_Open_To_Buy
Average available credit (Credit_Limit − Total_Revolving_Bal). Removed due to perfect correlation with these components.

Total_Trans_Amt
Total transaction amount over the past 12 months.

Total_Trans_Ct
Total number of transactions over the past 12 months.

Avg_Utilization_Ratio
Average proportion of credit used relative to the limit.

Total_Amt_Chng_Q4_Q1
Ratio of spending amount in Q4 relative to Q1.

Total_Ct_Chng_Q4_Q1
Ratio of transaction count in Q4 relative to Q1.

Engineered Variables

Avg_Trans_Amt
Mean transaction value: Total_Trans_Amt ÷ Total_Trans_Ct. Used in the multicollinearity-free dataset.

GitHub Link

The full codebase, including all scripts used for model fitting, evaluation, and data preparation, is available at:

https://github.com/TomNeame/Credit-Card-Data.git