HarvardX - Data Science: Capstone Course - Blood Transfusion Service Center Project

James Melanson

2021-07-28

Contents

1 Introduction

1.1 Purpose

Marketing departments do not have unlimited budgets to attract sales. Further, certain customers are more likely to convert into a sale based on information that a business already possesses. The Recency, Frequency, and Monetary Value model allows a business to target customers based on how recently a customer made a purchase, how often they purchase, and how much they generally spend. However, can this approach be used to predict repeated engagement with blood donation services?

The aim of this project is to produce a machine learning-based model to predict whether a blood donor donated blood to the Blood Transfusion Service Center in Hsinchu, Taiwan in March 2007. Models were characterized by comparing their predictions of donation status to true donation statuses contained in a hold-out test set, and computing indicators of accuracy.

1.2 Dataset

The Blood Transfusion Service Center (BTSC) data set contains information related to blood donation at the BTSC in Hsinchu, Taiwan. Donated to the University of California, Irvine (UCI) Machine Learning Repository in October 2008, the BTSC data set contains data from 748 blood donors.

This data set contains one data-containing file, “transfusion.data’, which contains and a file containing a description of the data set,”transfusion.names".

1.2.1 Transfusion.data File Structure

The transfusion.data file contains information related to individuals who donated blood to the BTSC in Hsinchu, Taiwan. There are 748 donors within the data set and four variables that may be used for predicting the outcome of whether a donor donated blood in March 2007.

The predictor variables include:

- Recency (months), the number of months since a donor’s last blood donation
- Frequency (times), the total number of blood donations a donor has given
- Monetary (c.c. blood), the total amount of blood donated in cm^3
- Time (months), the number of months since a donor’s first blood donation

The outcome variable includes the following variable:

- Whether he/she donated blood in March 2007, which is self-explanatory

1.2.2 Transfusion.names File Structure

The Transfusion.names file was included with the data set downloaded from the UCI Machine Learning Repository and serves as metadata. Included within the file, are the following characteristics of the data set:

- number of records
- number of attributes
- economic sector the data set is taken from
- data set source and owner
- general description of the data set
- attribute information
- citation information

1.3 Goal of the project

The goal of this project was to produce a machine learning based model to predict whether a blood donor that previously donated to the BTSC would donate again in March 2007. As part of this project, I trained machine learning algorithms on a portion of the BTSC dataset to create a prediction system. The model with the highest F1 score during training, excluding the regression-based approach, was validated on a test set that was held out during algorithm development.

The algorithms that were developed during this project predict whether a blood donor donated again in March 2007 using blood donation data from other users. Model performance was measured by comparing their predictions of whether a donor donated blood in March 2007 to whether a particular donor actually donated blood in March 2007; using the root mean squared error (RMSE), accuracy, and F1 score.

1.4 Key steps that were performed

Key steps that were performed as part of this analysis include:

1. Data retrieval from the UCI Machine Learning Repository from the following link: https://archive.ics.uci.edu/ml/machine-learning-databases/blood-transfusion/transfusion.data
2. Renaming of columns for improved interpretability
3. Partitioning the data into a training set, `training`, and a hold-out test set, `validation`, to test the accuracy of the final model
4. Partitioning the training set `training` into two subsets: a training set, `train`, and a test set, `test_set`, to test the accuracy of algorithms during model development
5. Development of machine learning-based algorithms using the RMSE, accuracy, and F1 score as metrics to compare the performance of different models

2 Methods

2.1 Data Retrieval

The data used for this analysis was retrieved using R from the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/machine-learning-databases/blood-transfusion/transfusion.data. A temporary file name was created using the `tempfile` function and the `download.file` function was used to download the transfusion.data file to the file name created with `tempfile`.

The transfusion.data file was inspected using Windows Notepad as the .data file extension was not recognized as having a particular file structure. Upon file inspection in Windows Notepad, transfusion.data was recognized as being a comma-separated value file. Thus, transfusion.data was read into R using the `read.csv` function using default settings.

2.2 Data Cleaning

The transfusion.data file did not require extensive cleaning. The transfusion.data file was read into R using the `read.csv` function with default settings to produce a data set wherein each row represented an observation and each column represented a variable. Further, there was no missing data within the data set.

Columns in the data set were renamed to improve interpretability and efficiency in writing concise code, as follows (new name = old name):

- Time since last donation (months) = Recency..months.
- Total number of blood donations = Frequency..times.
- Amount of blood donated (mL) = Monetary..c.c..blood.
- Time since first donation (months) = Time..months.
- Donated Blood in March 2007? = whether.he.she.donated.blood.in.March.2007

2.3 Partitioning of the Blood Transfusion Service Center's Data into Training and Validation Sets

The `caret` package was used to partition the BSTC data into training and validation sets. The training set, `training`, consisted of 90% of the transfusion.data file and was used to train machine learning-based algorithms to predict whether a blood donor donated blood in March 2007. The validation set, `validation`, consisted of 10% of the transfusion.data file and was used to test the accuracy of the final model.

A 90%/10% split of the BTSC data was performed due to the relatively low number of observations compared to other data sets which use machine learning. Further, the low prevalence of blood donors who donated again in March 2007 was low compared to the number of donors who did not donate again in March 2007.

2.4 Root mean squared error (RMSE)

The RMSE was defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

where:

- N = number of observations
- $\hat{y}$ = machine learning algorithm's prediction of whether a donor donated blood in March 2007
- y = the true status of whether a donor donated blood in March 2007

2.5 Accuracy

The accuracy was defined as follows:

$$Accuracy = mean(\hat{y} = y)$$

where:

- $\hat{y}$ = machine learning algorithm's prediction of whether a donor donated blood in March 2007
- y = the true status of whether a donor donated blood in March 2007

2.6 F1 score

The F1 score was defined as follows:

$$F_1 = \frac{1}{\frac{1}{2}(\frac{1}{recall} + \frac{1}{precision})}$$

where:

$$recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

2.7 Exploratory Data Analysis

2.7.1 Training Data Set

The training data set consists of observations from 673 donors who donated blood to the BTSC in Hsinchu, Taiwan.

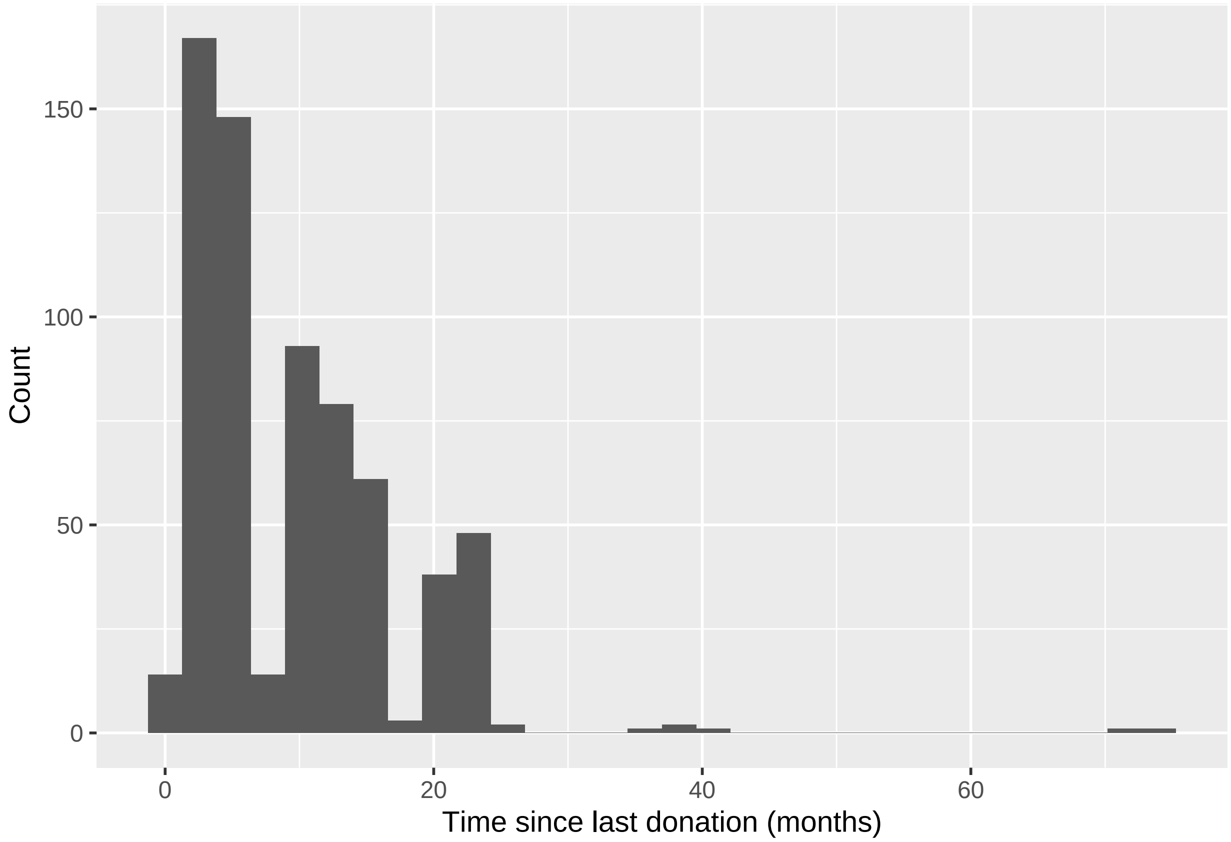

Figure 1: Histogram of Time Since Last Blood Donation - Training Set

Table 1: Summary Statistics of "Time Since Last Donation (months)" in the Training Data Set

Mean	Median	Min	Max	Range
9.539376	7	0	74	74

Table 2: Summary Statistics of "Total Number of Blood Donations" in the Training Data Set

Mean	Median	Min	Max	Range
5.456166	4	1	50	49

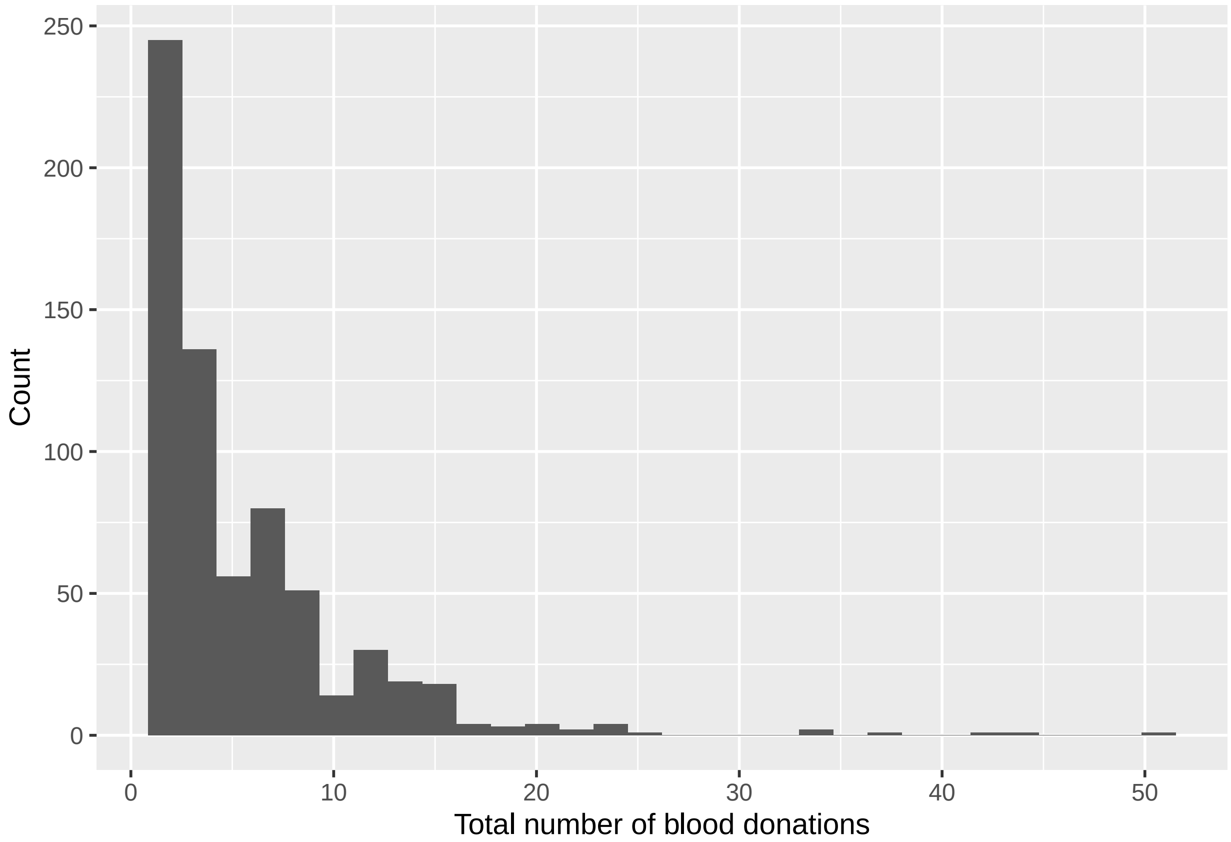

Figure 2: Histogram of Total Number of Blood Donations - Training Set

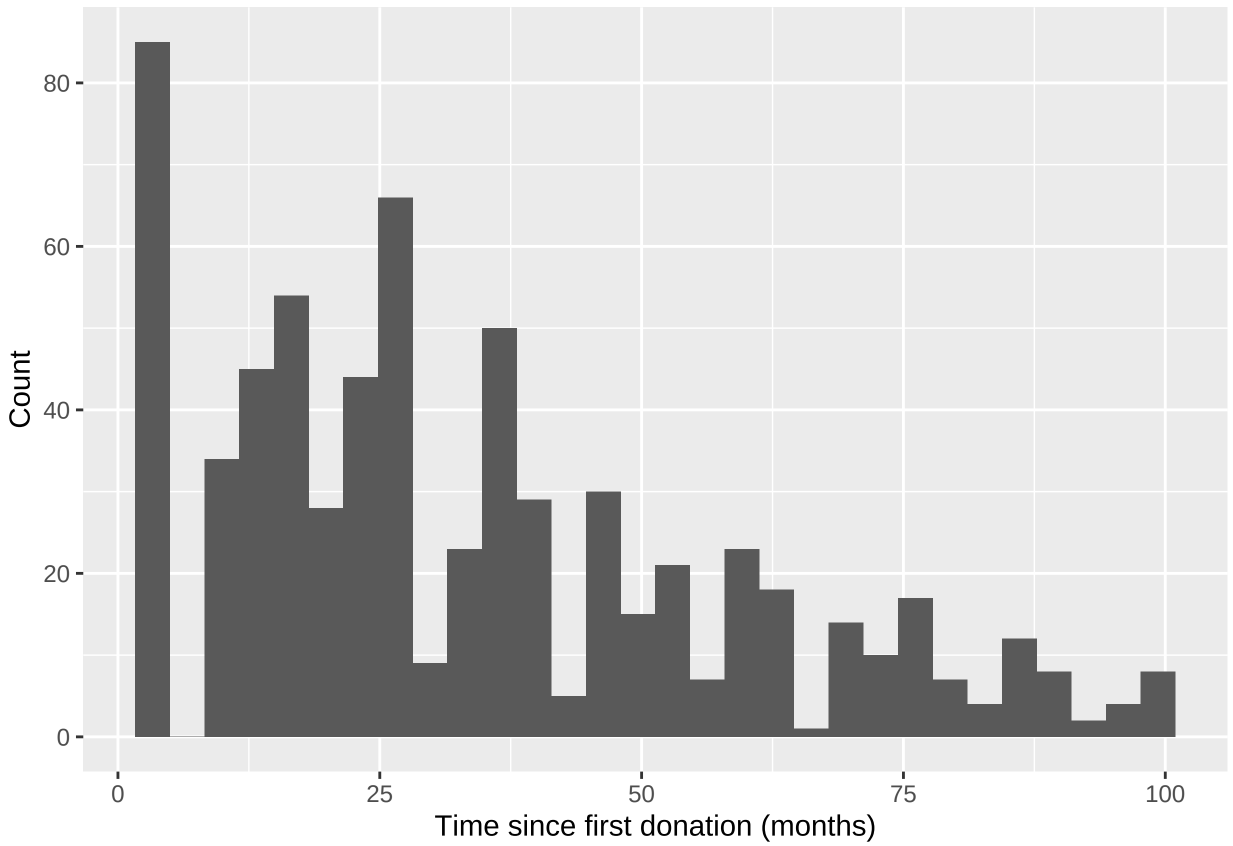

Figure 3: Histogram of Time Since First Blood Donation - Training Set

Table 3: Summary Statistics of "Time Since First Donation (months)" in the Training Data Set

Mean	Median	Min	Max	Range
34.02377	28	2	98	96

Table 4: Summary Statistics of Blood Donation Status in March 2007 in the Training Data Set

Donated Blood in March 2007?	n
0	463
1	142

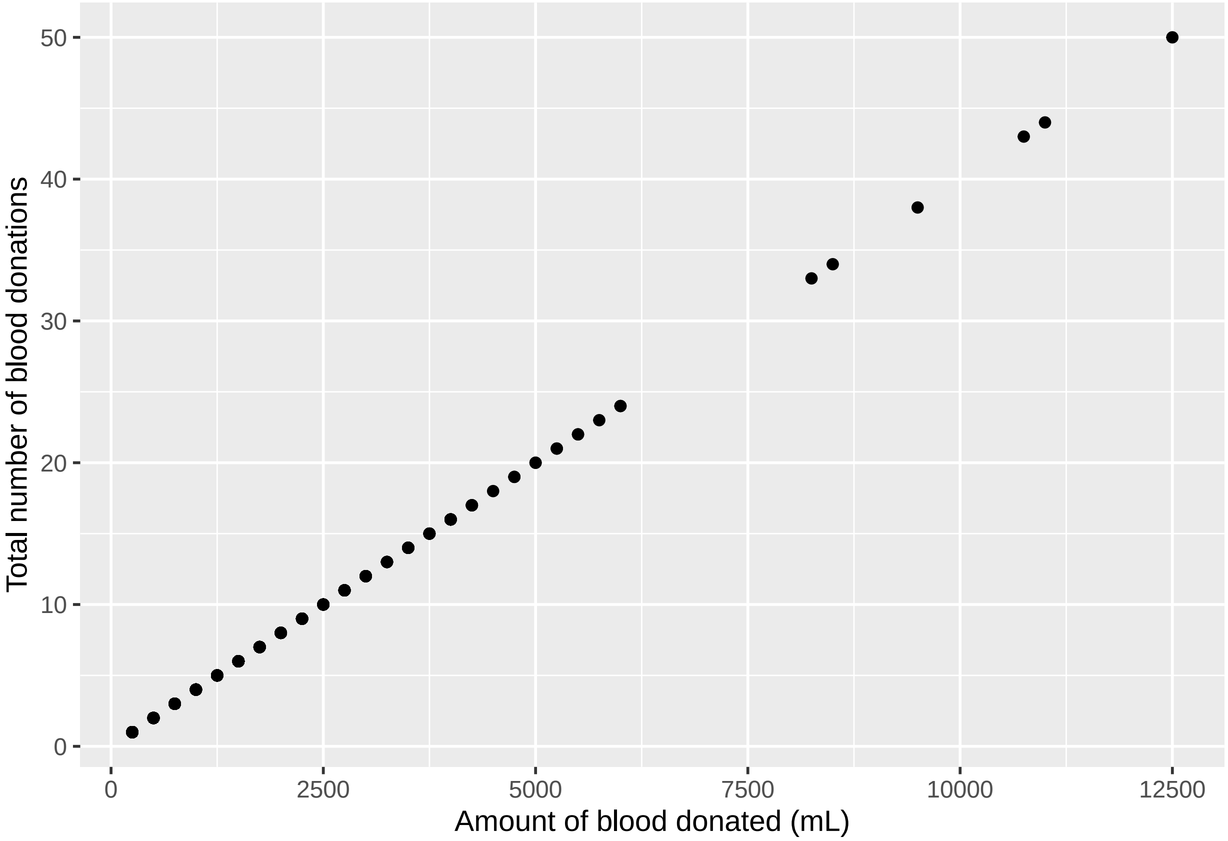

Figure 4: Total number of blood donations versus Amount of blood donated (mL)

Due to the collinearity between the "Total number of blood donations" and "Amount of blood donated" variables, the "Amount of blood donated" variable was excluded from further analysis.

2.7.2 Validation Data Set

The `validation` data set consists of 75 observations of whether a donor donated blood in March 2007. Further exploration of this data was not performed so as to preserve the utility of this data as a hold-out test set.

2.8 Modelling Approaches

2.8.1 Model 1: Flip a coin

This model does not use any of the information contained within the BTSC dataset, and simply guesses whether a donor donated in March 2007. This was done by sampling from a population of 1s and 0s with a 50% probability of selecting either number; with 1 corresponding to the person having donated blood in March 2007, and 0 corresponding to a person not donating blood in March 2007.

This would not be a candidate for a final model in predicting blood donation status in the `validation` data set. However, it is useful as a baseline measure for comparing different algorithms' accuracies.

2.8.2 Model 2: Logistic regression

This model uses regression to determine the following probability:

$$Pr(Y = 1|X_1 = x_1, X_2 = x_2, X_3 = x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where:

- Y is the status of whether an individual donated blood in March 2007, with 1 corresponding to a donor having donated and 0 corresponding to a donor not having donated
- X is a predictor
- x is the value of a particular predictor

Predictions are made by this model by determining if the probability of a donor having donated, given the values of the predictors, is greater than 0.5. If the probability of having donated according to the logistic model is greater than 0.5, it will predict that the donor donated blood in March 2007.

2.8.3 Model 3: k-Nearest Neighbours

This model uses the "k-Nearest Neighbours" algorithm to generate outcome predictions for the `test_set` dataset. This model queries the blood donation status of the `k` nearest points to a test point and assigns an outcome status based on majority rule. The parameter `k` was varied during model development to empirically determine its optimal value.

2.8.4 Model 4: Random Forest

This model uses the "Random Forest" algorithm to generate outcome predictions for the `test_set` data set.

To generate outcome value predictions:

1. A bootstrap sample containing 605 observations is generated by sampling with replacement from the `train` dataset and randomly selecting `x` predictors
2. A decision tree is generated and outcome value is predicted
3. Steps 1 and 2 are repeated for a total of 25 trees
4. The average prediction for a particular `test_set` observation across all trees is taken as the final prediction

   Note: `x` is a tunable parameter whose value can be varied during model development to empirically determine an optimal value with respect to model accuracy

2.8.5 Model 5: Ensemble of k-Nearest Neighbours and Random Forest

This model uses both the “k-Nearest Neighbours” and “Random Forest” algorithms to generate outcome predictions for the `test_set` data set. To generate a prediction of whether a donor donated blood in March 2007, the probability for a particular outcome status was calculated with each algorithm separately and then these two probabilities were averaged. If the averaged probability was greater than 0.5, this model predicts that the donor donated blood in March 2007.

3 Results

3.1 Model 1: Flip a coin

In Model 1, predictions of whether a blood donor donated blood to the BTSC in March 2007 were generated by flipping a coin. The RMSE of Model 1 was 0.7174301. Further, the accuracy and F1 score were 0.4852941 and 0.5679012, respectively.

Model	RMSE	Accuracy	F1 score
1: Flip a coin	0.7174301	0.4852941	0.5679012

3.2 Model 2: Logistic regression

In Model 2, logistic regression was performed using three of the predictors present in the BTSC data set to produce predictions on the outcome variable: whether a donor donated blood to the BTSC in March 2007. The RMSE of Model 2 was 0.8911328. Further, the accuracy and F1 score were 0.7352941 and 0.8421053, respectively.

Model	RMSE	Accuracy	F1 score
1: Flip a coin	0.7174301	0.4852941	0.5679012
2: Logistic Regression	0.8911328	0.7352941	0.8421053

3.3 Model 3: k-Nearest Neighbours

Model 3 utilized the “k-Nearest Neighbours” algorithm to generate predictions of whether a donor donated blood to the BTSC in March 2007. The RMSE of Model 3 was 0.9235481. Further, the accuracy and F1 score were 0.7352941 and 0.8392857, respectively.

Model	RMSE	Accuracy	F1 score
1: Flip a coin	0.7174301	0.4852941	0.5679012
2: Logistic Regression	0.8911328	0.7352941	0.8421053
3: k-Nearest Neighbours	0.9235481	0.7352941	0.8392857

3.4 Model 4: Random Forest

Model 4 utilized the “Random Forest” algorithm to generate predictions of whether a blood donor donated blood to the BTSC in March 2007. The RMSE of Model 4 was 0.9851844. Further, the accuracy and F1 score were 0.6764706 and 0.8, respectively.

Model	RMSE	Accuracy	F1 score
1: Flip a coin	0.7174301	0.4852941	0.5679012
2: Logistic Regression	0.8911328	0.7352941	0.8421053
3: k-Nearest Neighbours	0.9235481	0.7352941	0.8392857
4: Random Forest	0.9851844	0.6764706	0.8000000

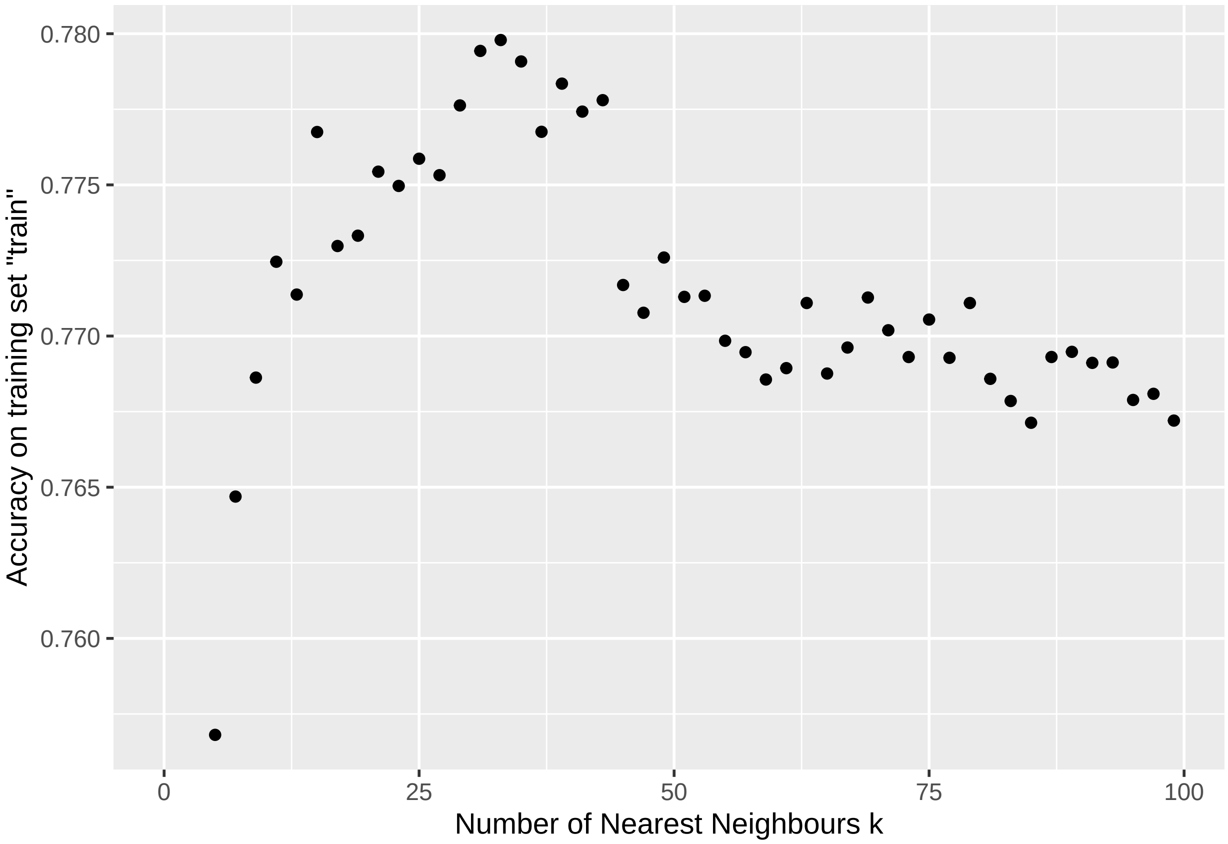

Figure 5: Model 3 - Accuracy on training set “train” versus Number of Nearest Neighbours k

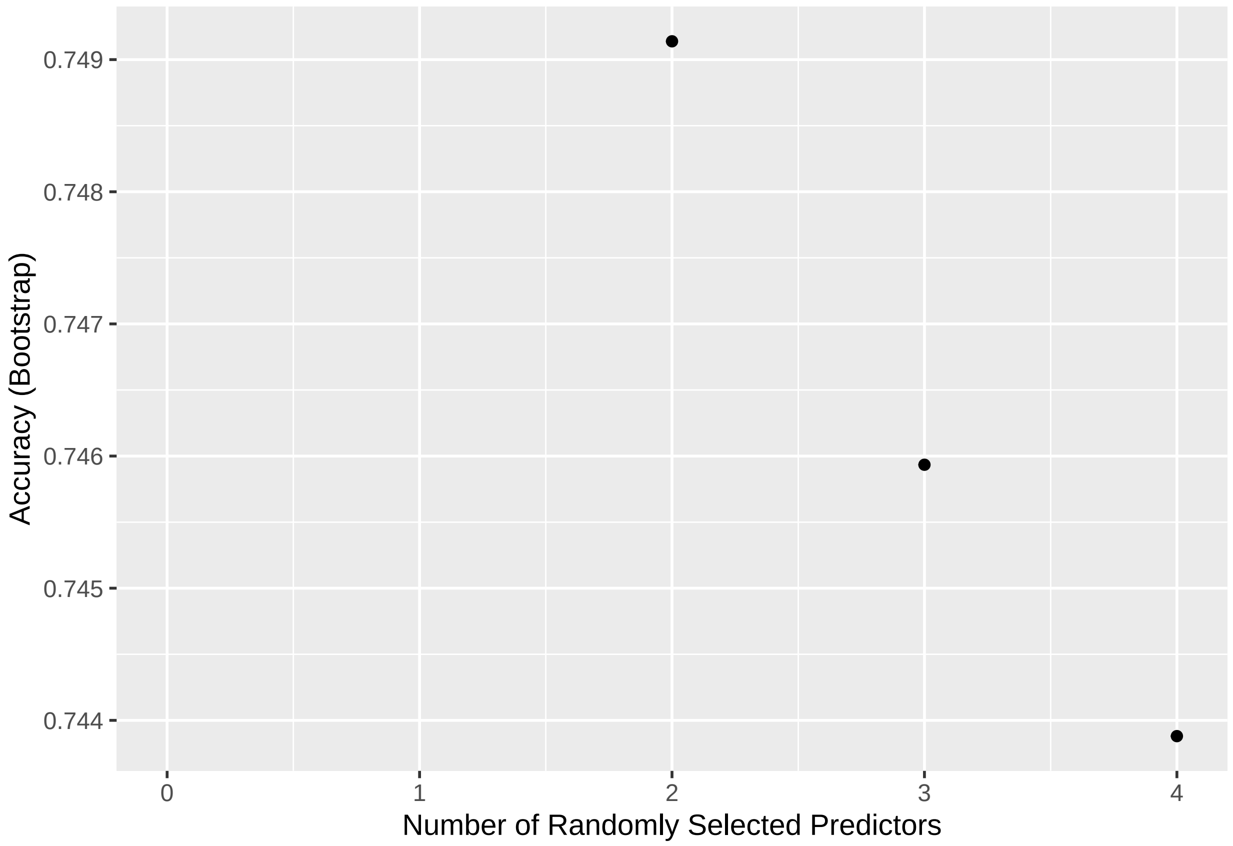

Figure 6: Model 4 - Accuracy (Bootstrap) versus Number of Randomly Selected Predictors

3.5 Model 5: Ensemble of k-Nearest Neighbours and Random Forest

Model 5 used both the “k-Nearest Neighbours” and “Random Forest” algorithms to generate predictions on whether a donor donated blood to the BTSC in March 2007. The RMSE of Model 5 was 0.9155519. Further, the accuracy and F1 score were 0.7205882 and 0.8318584, respectively.

Model	RMSE	Accuracy	F1 score
1: Flip a coin	0.7174301	0.4852941	0.5679012
2: Logistic Regression	0.8911328	0.7352941	0.8421053
3: k-Nearest Neighbours	0.9235481	0.7352941	0.8392857
4: Random Forest	0.9851844	0.6764706	0.8000000
5: Ensemble of kNN and Random Forest	0.9155519	0.7205882	0.8318584

3.6 Validation of the final model’s accuracy against the hold-out test set

Validation of the final model, Model 3, was performed on the hold-out test set `validation`. Model 3 was chosen as the final model as it had the highest F1 score out of the models that went beyond logistic regression. The RMSE of Model 3 with respect to the `validation` test set was 1.0132456. Further, the accuracy and F1 score were 0.76 and 0.8548387, respectively.

Model	RMSE	Accuracy	F1 score
3: k-Nearest Neighbours	0.9235481	0.7352941	0.8392857
Validation using hold-out test set `validation`	1.0132456	0.7600000	0.8548387

4 Conclusion

The aim of this project was to develop a machine learning-based model that could predict whether a blood donor donated blood to the BTSC in March 2007; using the time since last donation in months, time since first donation in months, and total number of blood donations as predictors. The final model that was developed utilized a k-Nearest Neighbours algorithm and achieved an accuracy and F1 score of 0.76 and 0.8548387, respectively.

One of the limitations of this project is its generalization. Although 30.7% of donors in the `training` data set donated again in March 2007, which may be considered good by blood donation standards, the number of `n`'s are significantly reduced when split out by conditions. Further, these models may or may not be extrapolated to blood transfusion centres elsewhere; depending on the demographics of the blood donors in this data set and logistics specific to the BTSC in Hsinchu.

Future work in modelling this data could utilize techniques such as principal component analysis to improve the accuracy and F1 score of the model. However, this may come at the cost of model interpretability.

5 References

Irizarry, R. A. (2021, July 3). Introduction to Data Science. Rafalab.Github.Io. https://rafalab.github.io/dsbook/

Kuhn, M. (2021). caret: Classification and Regression Training. R package version 6.0-88. https://CRAN.R-project.org/package=caret

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., . . . & Yutani, H. (2019). Welcome to the Tidyverse. Journal of open source software, 4(43), 1686.

Yeh, I., Yang, K., Ting, T. (2008, October 3). UCI Machine Learning Repository: Blood Transfusion Service Center Data Set. University of California, Irvine Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center