MedTourEasy HealthCare organization

Data Analysis Report

Prepared by

Udesh Senevirathne

udeshchathumal22449@gmail.com

Introduction

Blood transfusion saves lives - from replacing lost blood during major surgery or a serious injury to treating various illnesses and blood disorders. Ensuring that there's enough blood in supply whenever needed is a serious challenge for the health professionals. According to WebMD, "about 5 million Americans need a blood transfusion every year".

Our dataset is from a mobile blood donation vehicle in Taiwan which belongs to the Blood Transfusion Service Center in Hsin-Chu City in Taiwan. The Blood Transfusion Service Center drives to different universities and collects blood as part of a blood drive. We want to predict whether or not a donor will give blood the next time the vehicle comes to campus.

The Driven Data Set

The data is stored according to RFMT marketing model which is a variation model of the RFM model

According to the data set RFM is stands for

R-(Recency) : Duration since the last blood donation

F-(Frequency):The total number of donations(or purchases)

M-(monetary) : the amount of blood donated

Then the RFMT- is stands for R-Recency

F-Frequency

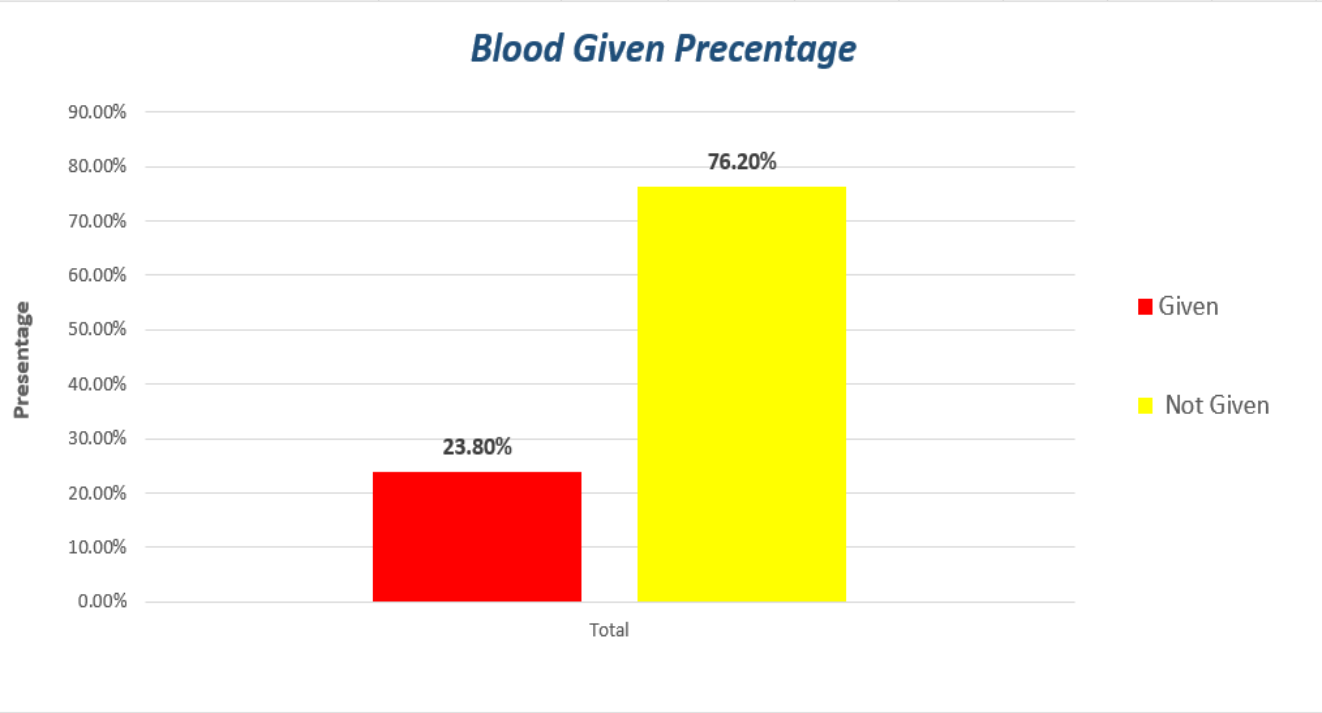
M-Monetary

T-Time

C-Churn Rate

We want to predict whether or not the same donor will give blood the next time the vehicle comes to campus. The model for this is a binary classifier, meaning that there are only 2 possible outcomes:

- 0 - the donor will not give blood
- 1 - the donor will give blood



This informed us that in our dataset `0`s appear(not given) 76% of the time. We want to keep the same structure in train and test datasets, i.e., both datasets must have 0 target incidence of 76%.

We use TPOT classifier for our dataset to choose the best classifier, and using sklearn library We can calculate the AUC score for this. And finally we got 0.7850 auc score .

Conclusion

The demand for blood fluctuates throughout the year. Blood donations slow down during busy holiday seasons. An accurate forecast for the future supply of blood allows for an appropriate action to be taken ahead of time and therefore saving more lives.

In this notebook, we explored automatic model selection using TPOT and AUC score we got was 0.7850. This is better than simply choosing `0` all the time (the target incidence suggests that such a model would have 76% success rate). We then log normalized our training data and improved the AUC score by 0.5%. In the field of machine learning, even small improvements in accuracy can be important, depending on the purpose.

Another benefit of using logistic regression model is that it is interpretable. We can analyze how much of the variance in the response variable (`target`) can be explained by other variables in our dataset.