Targeted Marketing Efforts in Blood Bank Donors

Beth Hilbert, Kyle Idoine, Jeremy Davis, and Richard King

Our goal is to improve the marketing efforts of a blood bank donation center, reducing their marketing cost while also improving their donor return rate. We want to be able to reduce their marketing costs by predicting the individuals who are most likely to donate in the following month, based on their past donation behaviors. With this model, non-profits will be able to better allocate their resources and target people for marketing who are more likely to show up, giving them the greatest return on investment. To develop our model, we used a data set gathered by Professor I-Cheng Yeh in Hsin Chu Taiwan which listed the Recency, Frequency, Monetary, and Time behaviors of 748 randomly selected donors. We used SPSS Modeler to build 4 Classifications algorithms. Then we measured the results and selected the best algorithm for the blood bank to use to predict future donors. With our results, the blood bank can focus on maintaining the most profitable donors.

1. Introduction

In recent years, there has not been a more energizing field of study than that of medicine. The study of the human body, how it works, and how to fix it has become an obsession across the world. From the Mediterranean diet to the study of Cancer, every week seems to have a new buzz article about the latest trend or discovery in health. From the beginning of time, man has been fascinated with the inner workings of the human body, so much so that humans have created a market around medicine and invested millions upon millions of dollars into finding cures for all types of various diseases. It is this investment that has allowed the average life span of a human to increase to 71.5 years in 2014 compared to that of 48 years in 1950 (1). Hundreds of procedures and vaccines have been invented to help keep this number climbing, one of which are blood transfusions. Blood transfusions take blood from one person's body and pump it into the body of another person who has suffered serious blood loss from a surgery or severe injury (2).

For a blood transfusion to be successful, the donor blood needs to meet certain requirements based on the blood type of the recipient. There are four blood types that someone can have, along with a positive or negative connotation for each. The types of blood combinations are as follows: A positive, A negative, B positive, B negative, AB positive, AB negative, O positive, and O negative. When receiving a blood transfusion, the blood type of the donor must match the blood type of the recipient, or else the recipient's body will endure an autoimmune attack of the foreign (donated) blood cells. There is one exception to the rule however. The O blood type can be donated to any person and is often referred to as the "universal donor" as the blood can be received and not cause an autoimmune response. Since the O blood type universally is accepted as a donor, it is used in extreme emergencies when the patient is losing blood too quickly and a test to determine the patient's blood type would take too much time (2). Along with saving people who have had large amounts of blood loss, blood transfusions can help relieve chronic or genetic blood diseases. There are three types of blood transfusions that are most commonly used in these circumstances: red blood cell transfusions, platelet transfusions and plasma transfusions. A red blood cell transfusion is typically used when someone has a disorder that restricts the amount of iron in their blood, thus causing the red blood cell count in a person's blood to be lower than normal. This is known as an iron deficiency or anemia. Platelet transfusions are often used for people who have blood cancer, or leukemia. A side effect of chemotherapy is a decrease in platelets, which is a serious issue because a platelet is the part of the blood that helps a person coagulate. If they have low platelet counts, it is possible for them to never stop bleeding. Finally, plasma transfusions are typically received by people who have liver failure, as plasma contains nutrients that are essential for a person's overall health (3).

Blood transfusion is an essential part of hospital medicine and patient care. In order to have enough blood to provide to every patient that needs it, hospitals rely of donations. Non-profits around the world go from town to town, soliciting and collection blood or plasma from any person willing to donate, given their medical history checks out. Arguably, the largest and most well-known of these non-profits is the American Red Cross. The Red Cross was founded in 1881 by Clara Barton as she was inspired by her visits to Europe after the First World War. It was

there that the original Red Cross had arose in Paris, bringing blood and resources to those affected during the war (4). Today the American Red Cross has grown into an entity that is known and respected across the world. The American Red Cross goes across the country collecting blood donations that save the lives of millions. One of their most challenging tasks, however, is getting people out to donate. According to the American Red Cross, the total population in America that is eligible to donate blood is 38%, but less than 10% of the population donates annually (5). Those are staggering numbers, especially with the growing population and the need for donor blood increasing continually. Marketing has since become an integral part of the process. Since these organizations are in towns for such a short amount of time, not to mention their status as non-profits, efficiency and cost-effectiveness are vital when advertising the location and time of the next blood donation opportunity. Helping to derive whether a person who has given blood before will be likely to donate again will be the focus of our model.

2. Literature Review

Customer Relationship Management (CRM) is an area of marketing that seeks to acquire and retain the most profitable customers. RFM ranks customers according to their historic behaviors of Recency, Frequency, and Monetary value to make predictions for future behavior. RFM is considered to be one of the most powerful models used in the CRM aspect of marketing. There is a lot of discussion in literature of the need for RFM, but there are few studies based on the sound mathematical theory used in RFM. Our dataset was initially developed by Professor I-Cheng Yeh to specifically explore this mathematical aspect of RFM. (6) Professor Yeh attempted to derive statistical models to estimate the probability that one customer will buy the next time and the expected value of the total number of times the customer will buy in the future. For his data, he used a random sample of donors from a Blood Bank in Taiwan. That data set was also used as the input to our model. His conclusion was that RFM was not a precise quantitative prediction model that could be implemented universally regardless of domain. He could not create a universal mathematical RFM model that would work in all situations. For example, some industries have greater churn so Recency might have greater weight. In other industries the Frequency is a larger predictor. Even though he could not create a universal model,

Professor Yeh was able to determine that there is an advantage to using a RFM model when it is weighted for the specific industry and when the customers were segmented into groups. In that case, the response rate of marketing campaigns can be greatly increased using RFM (Fig. 1).

Some of blood bank studies have evaluated the association between involuntary blood donors (those who donated for a specific friend in need) and voluntary donors with the aim of increasing the voluntary rate of donations in developing countries. (8) Other studies have focused on the safety of the blood donation process. (9)

For our study, we used Professor Yeh's data set to predict whether individuals who had donated in the past were likely to return during the following month.

3. Problem Analysis

We noticed this business problem is very much like a churn data set where we were predicting who will and who won't donate. Because of this relationship we decided to focus on predictive categorical models to represent our data.

We followed the Cross-Industry Standard Process for Data Mining (CRISP-DM). The first step is determining the business opportunity we are addressing. The CRISP-DM process needs a clearly identified goal to provide business value. Our focus is to predict a group of people likely to donate who can then be targeted for marketing purposes. Most of our time was spent on steps two and three, understanding and preparing the data, which is described in our data source section. We had to continue returning to this step while building our models and while this is a small focus of our report, it was the largest focus of our time. The fourth CRISP-DM step is model building. The bulk of our report focuses on measuring and comparing the algorithms we selected. The fifth step of CRISP-DM is Evaluation. During evaluation, we used a different set of data than what was used to create the algorithms to determine how accurate the model is. Because we selected the decision tree algorithm, our evaluation is shown on that stream. Once each of those steps has been finished, the final step in the CRISP-DM, deployment, is carried out. During deployment, the blood bank will use more current data with target field unknown to predict who will be most likely to donate. Then the marketing department will communicate with those specific donors encouraging them to return to the blood bank.

4. Data

Data Source. The data selected for this study was the Blood Transfusion Service Center Data Set and is publicly accessible from the UCI Machine Learning Repository. This data set was collected by Professor I-Cheng Yeh from the Chung-Hua University. The set contains 748 real donors randomly selected from a donor database at a Blood Transfusion Service Center in Hsin-Chu City in Taiwan. For each donor, there are 5 attributes: Recency (months since last donation), Frequency (total number of donations), Monetary (total cc of blood donated), Time (months since first donation) and a binary value representing whether the person donated blood in March 2007 (1=donated blood, 0=did not donate). Our prediction target is the field titled "whether he/she donated in March 2007", which will help the business most effectively market and distribute information to the correct group of people, giving them the largest turn out per those people solicited.

Understanding the Data. We used the Data Audit node to get a better picture of our data set (Fig. 2). During the audit we found there were no missing values, which is great news. However there were some outliers, which can influence the data mining process. We then needed to determine if these outliers contributed value or should be deleted by looking at additional data on the specific fields (Fig. 3). The graphs show our set is positively skewed (outliers are higher than the average) as well as the standard deviation being very large in the Monetary (cc blood) attribute. We decided the best option would be grouping each attribute into categories instead of evaluating each instance individually. How we did this grouping is referred to as binning, and is discussed in our Data Preparation section.

On evaluation of attributes, we see the response rate curve decrease sharply (Fig. 4). We receive most of the return on investment (people who return to donate blood) on less than 20% of the instances from the data set. This is great news and shows us that focusing our marketing towards a narrowly selected group could greatly increase results while reducing our expenses.

Preparing the Data. Upon observing the data in a table format, it became clear that there would need to be some manual cleaning of the data to produce a viable model. These changes

involved splitting values into meaningful groups (known as binning), renaming values so they could be understood by business users (known as reclassifying), and sampling in a way that the small percentage of positive results of our target would increase to a reasonable Sensitivity level (known as stratifying the sample). These changes would give balance to the data and make the reports more easily readable.

Our first step focused on our target field. Whether a person donated in March of 2007 (our target attribute) was represented as a series of ones and zeros, with one referencing that the individual donated blood and a zero referencing that the individual did not. If this were to be presented to the line of business in its original form of ones and zeros, it would have no significance to them. To fix this, a reclassification was done using the reclassify node in IBM SPSS. A number one was given the description of "Will Donate" and a zero was given the description of "Won't donate". For the output of all the models, the target data will be represented in terminology that the line of business will understand and holds value to them.

The next steps involved splitting the other four attributes into meaningful categories using a separate binning node for each attribute. For each attribute, they were separated into three sections using the binning function and represented the top, middle, and bottom 33% of the data. This binning method also eliminated the weight of any outliers present in the data by using a count and not an average to split the values could be separated and weighted equally. It was important that we not just erase outliers because in our sample, those who had given the most quantity of blood were also likely to also return the following month. The final splits for each attribute are as follows: for Recency, any value between 0 and 4 months were pulled together and represented as Bin 1. Bin 2 contained all people with a Recency of more than 4 months and less than or equal to 13 months. Finally, bin 3 contained anyone who had a Recency of more than 13 months. For Frequency, any value between 1 and 2 donations were pulled together and represented as Bin 1. Bin 2 contained all people with a Frequency of more than 4 months and less than or equal to 6 donations. Finally, Bin 3 contained anyone who had a Frequency of more than 6 donations. For Monetary, any value between 250 and 500 ccs were pulled together and represented as Bin 1. Bin 2 contained all people with a Monetary amount of more than 500 ccs and less than or equal to 1500 ccs. Finally, Bin 3 contained anyone who

had a Monetary amount of more than 1500 ccs. For time, any value between 2 and 21 months were pulled together and represented as Bin 1. Bin 2 contained all people with a time of more than 21 months and less than or equal to 40 months. Finally, Bin 3 contained anyone who had a Monetary amount of more than 40 months.

After binning, we renamed these four bins with more meaningful names, just as we had done for the target attribute in the first step to improve readability. After the four attributes were binned and split into categories, they were all represented a by single number, being one, two or three. If this were to be presented to the line of business, it would have no value to them. To give these bin numbers line of business value, the reclassify node was once again used. For Recency, a value of 1 was given a description of “0-4 months”, the bin value of 2 was reclassified as “4-13 months” and a bin value of 3 was reclassified as “14+ months”. For Frequency, a value of 1 was given a description of “frequent”, the bin value of 2 was reclassified as “moderately frequent” and a bin value of 3 was reclassified as “isolated”. For Monetary, a value of 1 was given a description of “low”, the bin value of 2 was reclassified as “medium” and a bin value of 3 was reclassified as “high”. For time, a value of 1 was given a description of “recent”, the bin value of 2 was reclassified as “moderate” and a bin value of 3 was reclassified as “long”.

The final piece of cleaning the data used the functionality of the selection node within IBM SPSS. Upon observing the data, it became clear that the data was skewed; of the total data set of over 700 participants, only 178 of those were determined to donate blood in March 2007. This would have caused our models to lose a great deal of predicative power. To fix this, the selection node was used to select a random sample of 160 participants that did donate in March 2007 and 160 participants that did not donate. With a data set that had an equal distribution, or Frequency of yes or no, a more accurate predicative model could be built. This is known as stratifying the data.

Now that everything has been separated into equal categories, the categories have been defined with terms that the line of business can understand and derive value from, and a random but equal sample of both outcomes for the target attribute were determined, the data

is ready to use to develop a model. This process encompassed the business understanding, data understanding and data preparation piece of the CRISP-DM methodology.

5. Models/Algorithms/Major Results

Overview. We selected 4 classification algorithms to develop our models. Classification algorithms look through historic data to make a prediction and then test that prediction against a known field in the historic data.

Model 1 Support Vector Machines

Support Vector Machines are among the most powerful classification algorithms. SVMs use mathematical functions to map non-linear relationships into linearly separable feature spaces. These mathematical functions are called Kernels. The boundary lines between the spaces are called vectors, which is where the model gets its name. Once mathematically transformed, the SVM model tries to find the best lines (called hyperplanes) that maximizes the separation (called margin) between the vector lines created by the Kernel. The wider the margin in between the two categories (referred to as margin hyperplane) the better the model will be at predicting the category for new records. Narrow margins result from overfitting the data and can result in misclassification.

There are a few parameters to adjust when building a SVM model (Fig. 5). One parameter we adjusted was trying different Kernels (math functions). There are four Kernels (linear, polynomial, RBF, and sigmoid). Another parameter we adjusted was the regularization parameter (C). The regulation parameter controls the trade-off between maximizing the margin (which aids in correctly predicting new records) and minimizing the training error (small number of misclassified data points improving accuracy of the model with training data). A higher C means you will have higher accuracy for training data but it could be at the expense of overfitting. The default for C is 10 but we found the best results between 3 and 5 depending on the Kernel chosen. Gamma (for polynomial and sigmoid Kernels) and RBF gamma (for RBF) is another parameter to adjust. The gamma should generally be set between 3/k and 6/k. A higher gamma can result in better accuracy with training data at expense of overfitting. We had 4 variables, so we set gamma to ¾ but also tested it at 1.

We ran confidence matrix to compare the parameters we had adjusted (Fig. 6). We found the best results of SVM algorithms using the RBF Kernel. RBF Kernel ranked highest at 81% in Sensitivity (which means it accurately predicted the positives as positive). In our case The RBF Kernel predicted 130 would donate and missed 30 who would donate (the lost business opportunity). Sigmoid was only 60% so would immediately be eliminated based on our business problem because it misses too many business opportunities. In the RBF Kernel, Precision was average at 68%. This represents the proportion who are going to respond positively out of the donors we are marketing to. Polynomial at 64% and Linear at 68% had similar Precision numbers so this wasn't a deciding factor in model selection. Accuracy was highest in RBF at 71%. The weighted F-score was a lot higher in the RBF at 74%, versus Linear at 70%, Polynomial at 71% and Sigmoid at 56%. Given our business problem, the RBF is the best SVM solution because it correctly identifies the most actual positives and is comparable to other SVM Kernels in incorrectly predicting positives.

In most of the Kernels, the Monetary and Frequency seemed to have highest Predictor Importance (Fig. 7). In SVMs, this is not used as a weight, but if we had a model with a lot of fields, we could eliminate some of the fields with lower importance and rerun the SVM model to see if results improved.

Model 2 Artificial Neural Network

The neural net can be compared to the way the brain processes information with a much simpler approach. The nodes in the neural net act like neurons that have a larger number of connections to other processing units. These processing units are layered and arranged accordingly. The neural network usually has three base layers known as input, hidden and an output. The units represent a target field and passes through the layers. The units connect to one another with different weights. The input data is usually shown in the first layer and values are assigned to each neuron and will pass through the next layer of neurons and so on. Sooner or later a result is delivered to the last output layer.

The network considers every tuple and record and generates a prediction for each one. The network will adjust when needed to and will do this by adding weight to "tip the scale". At the

beginning, all the weights are random. This could lead to weird results. The network will learn from its mistake through repetition. Information will flow through the network and will slowly change how the weights are distributed. As the network learns it becomes more and more accurate at simulating the known outcomes. Once trained the network can be used to predict future use cases where we do not know the data that is to come. For example, if someone will donate blood on the month of March 2007.

The neural net was used in our example for in-depth exploration and a prediction model that was based on weighted units. The neural network became a powerful general function estimator and helped me come up with results. The only requirements of a neural net are that the net needs to have at least one target and one input field. Using a neural net has a wide range of predictive uses. The net also allows for minimal model structure. The relationship between nodes is determined and that is how the neural net learns. On the downside, neural networks are not easily interpretable. It may be difficult to understand the underlying process that makes up the relationships.

When working within a neural net model of my own. I noticed how relatively easily it was to get the model set up and running. First, I decided on what units were the most important in determining the target unit, which was if someone will donate blood on the month of March 2007. In (Fig. 8) you could see where the weights have been distributed by importance to each unit. Recency seemed the most important which thus yields the most weight and holds the most importance. I believe this to be true because if you recent something more often; there is more likely a chance of them being there in march 2007.

Per (Fig. 9) we show that the model is a multilayer perceptron which has 5 hidden layer neurons. The model was able to produce a 69.7% Accuracy. In (Fig. 10) Our Coincidence matrix shows great results with wrong predictions at 19.79%. In (Fig. 11) you can see the input layers are set as the networks Bias, Frequency, Recency, Time, and Monetary. The next layer is the hidden layer where the network uses Bias, and five hidden neurons called Neuron1 through Neuron5. This all points to the output layer where we determine whether the tuple, person, will donate blood on the month of March 2007.

Model 3 Bayesian

The Bayesian classification method, named for the 18th century English clergyman who worked with probability and decision theory, Thomas Bayes, predicts the probability that a selected tuple belongs to a targeted class. In essence, the model works by analyzing the independencies between select variables in the dataset and detects causal relationships between them.

The Bayesian model is useful in that it provides a means to analyze the effects of several different variables and to predict how they could potentially affect your target variable. These networks can also allow for prediction even in the case of missing information as it makes use of the data that is presently available. Additional value from the use of the Bayesian method is found in its ability to assist in avoiding overfitting and gives the audience a very simplistic diagram that is easy to follow.

The method itself is based on Bayes' Theorem, which operates under the understanding that for tuple *X*, for example, the conditional probability that *X* belongs to class *C* can be determined given that we in fact are aware of the description of attributes of *X* itself. It analyzes the target variable, inspects the relationships between that variable and other input variables, and determines which stands as a valuable predictor of the target variable. With its relatively simplistic design, the method's accuracy, within specific domains, is often ranked in comparability to that of other similar classification methods such as decision trees or neural networks.

When using the Bayesian method within the SPSS modeler, the probabilities between the relationships are depicted by the arrow links between each specific node. As you can see in (Fig. 12), the model's target of "whether he/she donated blood in March 2007" was most closely linked with the Recency predictor. This is further elaborated in (Fig. 13) which displays a bar chart marking the importance of each separate predictor. According to this figure, Recency was the most important predictor associated with our target at nearly 80% against the next nearest predictor which was set at around 15%. The predictors labeled Monetary and Frequency were both equally measured at less than 15%.

Additionally, a coincidence matrix was generated in (Fig. 14) which shows our correct predictions at 69.71% and our incorrect predictions at 30.29%. Conjunctly, our second coincidence matrix (Fig. 15) displays the actual number of those who will/don't donate blood. In considering the results of this model alone, we can reasonably conclude that whether a person had donated blood in March 2007 strongly relies on how recently they had actually given blood.

Model 4 Decision Tree

Decision trees are a very popular model to use when trying to describe a relationship to the line of business. There are clear leaves and branches that show how each split was separated and by what characteristics. This form of model is very is for the line of business to understand. When reading a tree, the first split represents the relationship with the greatest relationship when trying to predict the outcome of the target attribute. There are many types of decision trees out there, and none of them are significantly better than the other, so to figure out which model to use, the situation one is trying to solve needs to be considered. For this business problem, three decision tree models were used to determine which model would most accurately predict the best way to market to blood donors. The three decision tree models of interest were C5.0, CHAID and CART.

C5.0 is a decision tree that runs using a gain ration calculation, and this type of calculation is biased towards unbalanced splits, were one split is usually much larger than the other. When the model was run on the training selection for the data, Recency was determined to be the attribute with the greatest predictive importance (Fig. 16). In fact, it was such a strong predictor that is where the splitting stopped. Two leaves were formed, the first leaf containing all participants that had a Recency of 0-4 months and the other leaf containing any participant that had a Recency of greater than 4 months (Fig. 17). A coincidence matrix was run to determine the performance of the model (Fig. 18). The metrics that were focused on were Accuracy, Sensitivity (Recall), Specificity, Precision and F1. The values for the C5.0 carts are as follows and can be seen in (Fig. 19). The Accuracy was 70%, the Sensitivity was 74%, the Specificity was 65%, the Precision was 68% and the F1 value was 71%.

CHAID is a decision tree that runs using a Chi-squared algorithm to determine independence. When the model was run on the training selection for the data, Recency was determined to be the attribute with the greatest predictive importance (Fig. 20). The next split was based on the second most important predictive variable, Frequency. Four leaves were formed, and a total of two splits. The first split was based on Recency and created three leaves, one for each of the Recency values, 0-4, 4-13 and 14+. The second split was based on Frequency and created two leaves. The first leaf was all the remaining participants from the first split labeled as "frequent" and the second leaf was the combination of the "moderate Frequency" and "isolated" participants (Fig. 21). A coincidence matrix was run to determine the performance of the model (Fig. 22). The metrics that were focused on were Accuracy, Sensitivity (Recall), Specificity, Precision and F1. The values for the C5.0 carts are as follows and can be seen in (Fig. 23). The Accuracy was 66%, the Sensitivity was 58%, the Specificity was 74%, the Precision was 69% and the F1 value was 63%.

CART is a decision tree that runs using the GINI index to determine the best predictive attribute. When the model was run on the training selection for the data, Recency was determined to be the attribute with the greatest predictive importance (Fig. 24). In fact, it was such a strong predictor that is where the splitting stopped, just like in C5.0. Two leaves were formed, the first leaf containing all participants that had a Recency of 0-4 months and the other leaf containing any participant that had a Recency of greater than 4 months (Fig. 25). A coincidence matrix was run to determine the performance of the model (Fig. 26). The metrics that were focused on were Accuracy, Sensitivity (Recall), Specificity, Precision and F1. The values for the C5.0 carts are as follows and can be seen in (Fig. 27). The Accuracy was 68%, the Sensitivity was 74%, the Specificity was 61%, the Precision was 66% and the F1 value was 70%.

So, which model provides the best predictive measure for this data set? When reviewing the business problem, the donor banks want to get the highest number of people who they predict to be donors, and thus solicit, turn into actual donors. The Precision measure is defined as the total people who show up to donate and were predicted to show up, as a ratio to the total predicted donors. This is the metric that these models should maximize. When reviewing the three model types, the CHAID has a Precision value of 69%, C5.0 is 68% and CART is 66%. These

values are not different enough to truly rule one as better than the other. The next metric to look at in this business problem would be the Sensitivity, which is defined as the number of people who donate, who were predicted to donate, in a ratio to the number total number of people who donate. Once again, this metric should be maximized. When reviewing this metric for the three decision tree models, CHAID's value (58%) is significantly lower than the other two (74%), this would rule CHAID out as a model to be used. As the other metrics are reviewed, the difference between C5.0 and CART is slight. As the determining factor, we look to the F1 value, which is a calculation that compares the Sensitivity and the Precision metrics into one, thus giving us the most powerful metric when determining which decision tree model would be the best. Since C5.0 has a slightly higher F1 value (71%) thank CART does (.70), C5.0 will be the decision tree model that will be used to represent this data.

Comparing Algorithms

The implementation of our model will involve marketing focused toward those donors we predict to return next month. We considered a positive response to be a donor who returns during March. There are five measures we could use to compare our models. If we incorrectly predict a donor as not returning who would actually return, then we miss a business opportunity. If we incorrectly predict a donor as returning who would actually not return, we waste time and money marketing to them. Because our objective is to increase marketing to those who would respond positively, Accuracy and Selectivity will be the most important metrics to focus on.

Overall Accuracy is the number of instances correctly predicted positive plus number of instances correctly predicted negative as a proportion of all instances.

While overall Accuracy can be an important indicator of the model's performance, we determined that for our business objective, Selectivity was a better indicator of model performance. Our objective is to target donors who are likely to return. Selectivity is the proportion of correctly identified positives out of all who actually are positives. If Selectivity is large, that means we are correctly identifying the donors who are likely to return. If Selectivity is small, that means there are a lot of donors who would return but we are incorrectly

classifying them as non-returners. If we are not predicting them to return, we will not be marketing to them. A low Selectivity is the largest risk in our business objective because it would result in missing a large segment of people who would respond positively to our marketing efforts.

In addition to Selectivity, Precision could be considered important because our marketing campaign will target those we predict to be positive. Precision is the number of correctly identified positives over both correctly and incorrectly predicted positives. This means Precision predicts how effective our marketing will be. We want to spend most of our marketing on those who are likely to respond. So if we have a large number of those predicted to be positive who aren't actually likely to donate in March, we would be wasting money in marketing to them. While this is important, we considered it a lower risk than Selectivity because the cost of marketing was not as important to us as the lost business opportunity of missing actual positive. In addition, none of the models had an unusually low Precision, so this did not end up being an issue.

F score is another measure to consider that balances Selectivity and Precision.

We did not consider Specificity to be an important measure. Specificity is the number of correctly identified negatives as a proportion of all actual negatives. Because we do not have an action identified to perform with predicted negatives, this was not a big issue in our business problem.

After training data had been run in every model to train them, it became clear that there was no significant statistical difference between the metrics used to analyze predictive accuracy of each model. SVM had that high accuracy, but this model is very difficult to explain to the business user. The ANN is met with similar resistance as most of the work happens within the hidden layers creating a "black box" effect. Since there was no significant statistical difference between the models, we chose to use the decision tree model as the line of business representation. The decision tree is an incredibly simple model to understand, and will provide the business with significant results to help drive their marketing agendas.

Validating Results Using Testing Data

Once the final model had been decided upon, another random sample of 160 people was selected using the sample node in IBM SPSS and with this sample the C5.0 decision tree was run. We ran an evaluation node comparing the training and testing data (Fig.28). This confirmed our conclusion that targeting our marketing to those most likely to respond would be effective because of the sharp decrease in the lift chart.

The results are as follows. The Accuracy for the test model came out to 69%, Sensitivity/Recall was 76%, the Specificity was 62%, the Precision was 66% and the F1 value was 71% (Fig. 29). The statistic that were generated from our test model were very similar to the ones generated with the training data. The Accuracy was fairly equal, 70% for the training set and 69% for the testing set, but the Sensitivity for the testing data was slightly higher, with a 74% Sensitivity value for the training data and a 76% Sensitivity value for the testing data. As stated before, Sensitivity was the metric of focus for us as we focused on the business problem at hand. Seeing that value increase with the training data was encouraging.

The predictive indicators for the testing data and training data were identical, once again showing Recency as the greatest predicting attribute (Fig. 30). Two spits were generated that mirrored the splits that were observed with the training data, on spit containing all people with a Recency value of 0-4 years and the other split containing people with a Recency value of 4+ years (Fig. 31). Based on these test results, we would make the recommendation for the line of business to focus their marketing strategy on a person's Recency; ideally targeting those individuals that have a Recency value of 0-4 months. If the budget allows, that net can be cast further out, but the lower the Recency value, the more likely that person is to donate in any given month.

6. Limitations

The biggest limitations are the ones set on us from the data. In predicting future events the importance of using up-to-date information will help retrieve better results. The data set we used in our models was collected in 2007. That was ten years ago, and was pre-recession. This could account for a much different way of living. Data that is ten years old has a stale

personification. Limitations can exist in many forms the first being that the data was old. The second limitation that existed was that the data was skewed. Skewed as in there were some anomalies that may have thrown off the results. Skewed data can be caused by having a few outliers that don't fit the regular mold. This leads us to the third limitation set on us by our data, a small sample size. Our data set only had 748 tuples. That is relatively small when the tuples were people entries. Certain cities have above a half a million people and we were working with a data set of 748 people. Stale, skewed, and small data all limit our model's prediction capabilities.

When talking about stale data it is important to say that you are dealing with the quality of information of the data. This can vary from data set to data set. In our case, we had a set of data that was ten years old. If we wanted to predict what would happen in 2017. We would be very limited by this set. Consumer, or in our case donator, behavior can very generation to generation. People's schedules change and trends move. Making predictions and claims based on stale data could become costly.

Anomalies occur but how you deal with them could cause limitations. If you limit the anomaly a certain way you may be setting your model up for failure. Skewed data can throw a curve ball at your prediction model and may offset the results. This can become a major issue if not addressed. Skewness is the measure of the distribution. If you have an outlier in the distribution, it may very well move the whole distribution one way or another. This effect is doubled when you are working within a small data set.

Having a small amount of entries can throw a wrench in your prediction model and small data sets are often tricky to handle and require finesse. When dealing with small data sets it may be best to just stick to the basics or use what is known as the K.I.S.S properties. To help when working with a small set of data, keep the models simpler, use a limited number of hypotheses, and employ a Bayesian model.

Limitations and restrictions almost always influence results. The three S's of our limitations were stale, skewed, and small; all of which applied to our data set and were transferred over to limiting our models.

7. Conclusion

Donating blood has become a pinnacle of modern medicine. Countless injuries and diseases can be remedied by the generous donation of individual's time and resources, in this case their very own blood. There are many non-profits that have joined the noble effort to collect these resources from donors, none of which are larger, or more well-known, than the American Red Cross. The American Red Cross will travel around the country, setting up donor banks for specific periods of time to try and gather as much blood as they can from donors. Once their time is reached, they will move on to the next city. Since these companies are non-profits, and do their work in the cities they visit during a limited window, marketing efficiently becomes vital for donation success. This is the problem that this study looks to solve. What attributes can the marketing teams of these non-profits focus in on to help drive the best donor turn out?

A C5.0 decision tree model was used to develop a business model that was easy to understand, and gave accurate information of the predictive factors for each attribute within our data set. Based on the findings of the report, Recency (how long it has been since their most recent donation) was determined that have the greatest predictive impact. A decision tree was created that had two leaves. The first leaf contained a heterogeneous node of mostly people who will donate and had a Recency value of 0-4 months. The second leaf contained a heterogeneous node of mostly people who will not donate and had a Recency value of 4+ months. Based on these findings, we would recommend that the marketing team focus their efforts on people who have a Recency value of 0-4 months. If the budget allows for a wider net, then the target Recency range can be expanded to satisfy the budget, but only Recency should be focused on as it was determined to be by far the strongest predictive indicator.

To take this project further, we would recommend updating the data set. Now that we have determined that Recency is strongest predictor, a current data set should be collected with a larger sample size, within the United States. Then another predictive analysis should be run with this update data set to confirm the validity of our conclusion that those predicted most likely to donate are those donors who have given within the last four months.

8. Bibliography

Data Set provided by Yeh, I-Cheng, Yang, King-Jang, and Ting, Tao-Ming, "Knowledge discovery on RFM model using Bernoulli sequence," Expert Systems with Applications, 2008. Accessed on February 15, 2017 from http://archive.ics.uci.edu/ml/machine-learning-databases/blood-transfusion/

1. Life Expectancy (n.d.). In Wikipedia. Retrieved February 15, 2017, from https://en.wikipedia.org/wiki/Life_expectancy

2. What Is a Blood Transfusion? (n.d.). In National Heart, Lung, and Blood Institute. Retrieved February 15, 2017, from https://www.nhlbi.nih.gov/health/health-topics/topics/bt

3. Learn About Blood (n.d.). In Red Cross Blood. Retrieved February 15, 2017, from http://www.redcrossblood.org/learn-about-blood/blood

4. Types of Blood Transfusions. (n.d.). In National Heart, Lung, and Blood Institute. Retrieved February 15, 2017, from https://www.nhlbi.nih.gov/health/health-topics/topics/bt/types

5. A Brief History of the American Red Cross. (n.d.). In Red Cross Blood. Retrieved February 15, 2017, from http://www.redcross.org/about-us/who-we-are/history

6. Blood Facts and Statistics. (n.d.). In Red Cross Blood. Retrieved February 15, 2017, from http://www.redcrossblood.org/learn-about-blood/blood-facts-and-statistics

7. Yeh, I., Yang, K., & Ting, T. (2008). Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*.

8. Asharani, S., Ganesh, S.Hari, (2014) A Survey On Blood Transfusion Based On Data Mining Techniques, International Journal of Scientific & Engineering Research, Vol5, Issue6. Retrieved February 22, 2017 from http://www.ijser.org/researchpaper%5CA-survey-on-blood-transfusion-based-on-data-mining-techniques.pdf

9. Erraguntla, M., Kamel, Hany, Whitaker, B., and Mayer, R. (2014) Data Mining to Improve Safety of Blood Donation Process. Conference Paper from the Proceedings of The Hawaii International Conference on System Sciences. Retried February 22, 2017 from file:///C:/Users/MandD/Downloads/HICSS%2047.pdf

10. Bayesian Network Node. (n.d.). In IBM Knowledge Center. Retrieved February 19, 2017, from https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/bayesian_networks_node_general.htm

11. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

9. Appendix

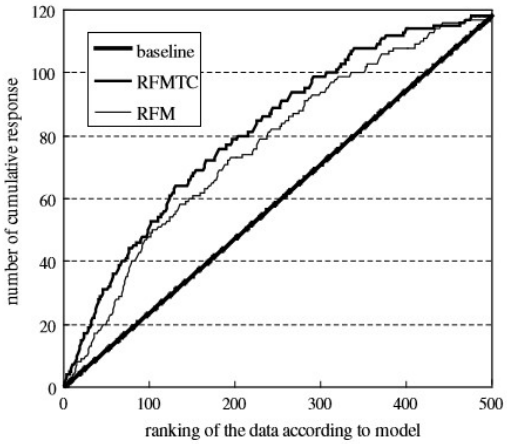


Fig. 1. Lift chart of RFMTC model and RFM model for training set.

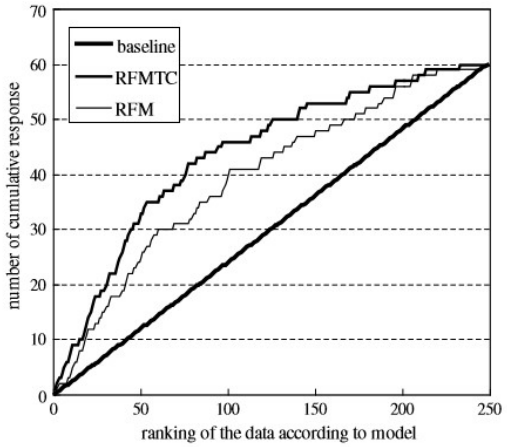
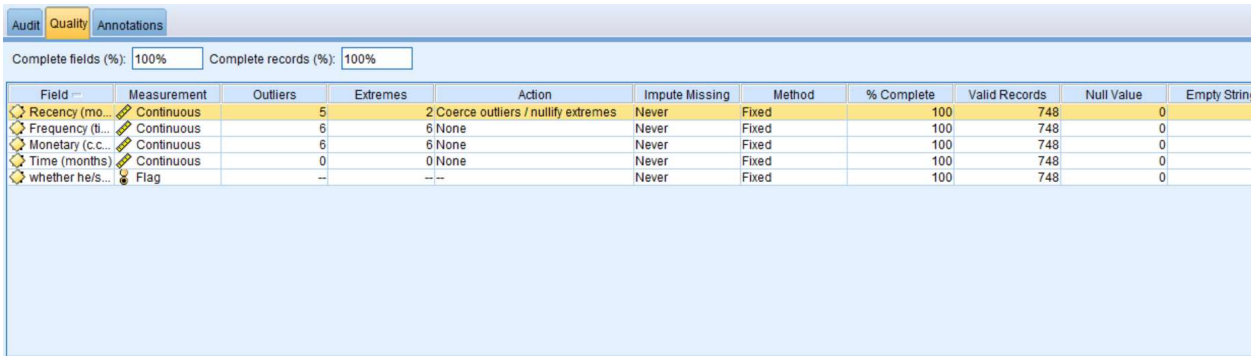


Fig. 2. Lift chart of RFMTC model and RFM model for testing set.

Figure 1 - ROC Curve Showing Lift Benefit of Weighting RFM (source: Professor Yeh's Report)



Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty Strin
Recency (mo...	Continuous	5	2	Coerce outliers / nullify extremes	Never	Fixed	100	748	0	
Frequency (ti...	Continuous	6	6	None	Never	Fixed	100	748	0	
Monetary (c.c...	Continuous	6	6	None	Never	Fixed	100	748	0	
Time (months)	Continuous	0	0	None	Never	Fixed	100	748	0	
whether he/s...	Flag	--	--	--	Never	Fixed	100	748	0	

Figure 2 - Data Audit of Full Data Set Shows No Missing Value But the Presence of Outliers

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Recency (months)		Continuous	0	74	9.507	8.095	1.880	--	748
Frequency (times)		Continuous	1	50	5.515	5.839	3.211	--	748
Monetary (c.c. blood)		Continuous	250	12500	1378.676	1459.827	3.211	--	748
Time (months)		Continuous	2	98	34.282	24.377	0.749	--	748
whether he/she donated blood in March 2007		Flag	0	1	--	--	--	2	748

[1] Indicates a multimode result [2] Indicates a sampled result

OK

Figure 3 - Data Audit of Entire Data Set Reveals Positively Skewed Results Need Evaluating

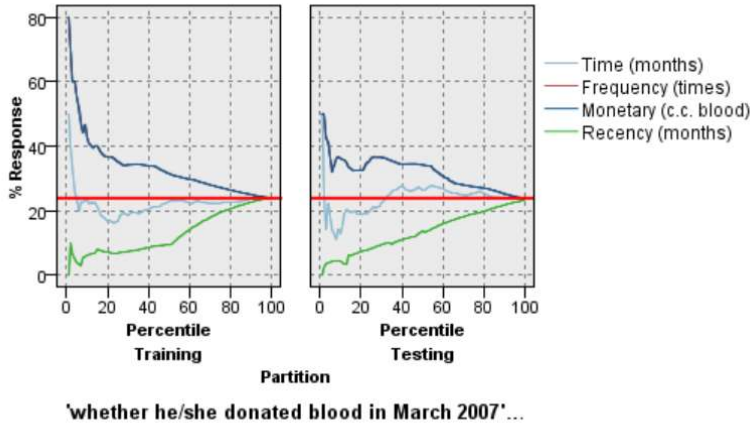


Figure 4 – Evaluation of Attributes

Model Type	Overall Accuracy	Sensitivity (predicted positive over those actually positive)	Specificity (predicted negative over those actually negative	Precision (actual positive over those predicted positive)	F1
Linear C=3	67%	76%	58%	64%	70%
Linear C=10	70%	75%	65%	68%	71%
Polynomial (C=3 gamma=1 degree=2)	67%	79%	55%	64%	70%
Polynomial C=5 gamma=.75 degree=7)	69%	74%	65%	68%	71%
RBF (C=3 RBF gamma=.75)	71%	81%	61%	68%	74%
RBF (C=3 RBF gamma=1)	69%	77%	61%	66%	71%
RBF (C=5 RBF gamma=.75)	69%	71%	67%	68%	70%
Sigmoid (C=3 gamma=.75)	52%	60%	44%	52%	56%
Sigmoid (C=10 Gamma=.75)	55%	54%	56%	55%	54%

Figure 5 - SVM Measurements

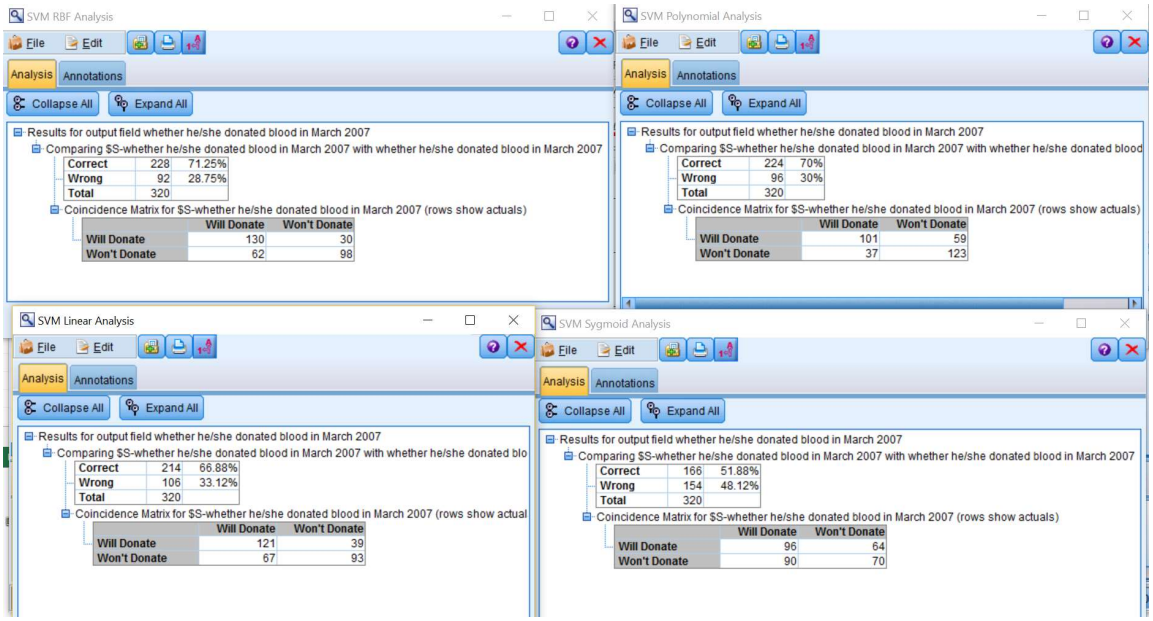


Figure 6 - SVM Confidence Matrix Comparison

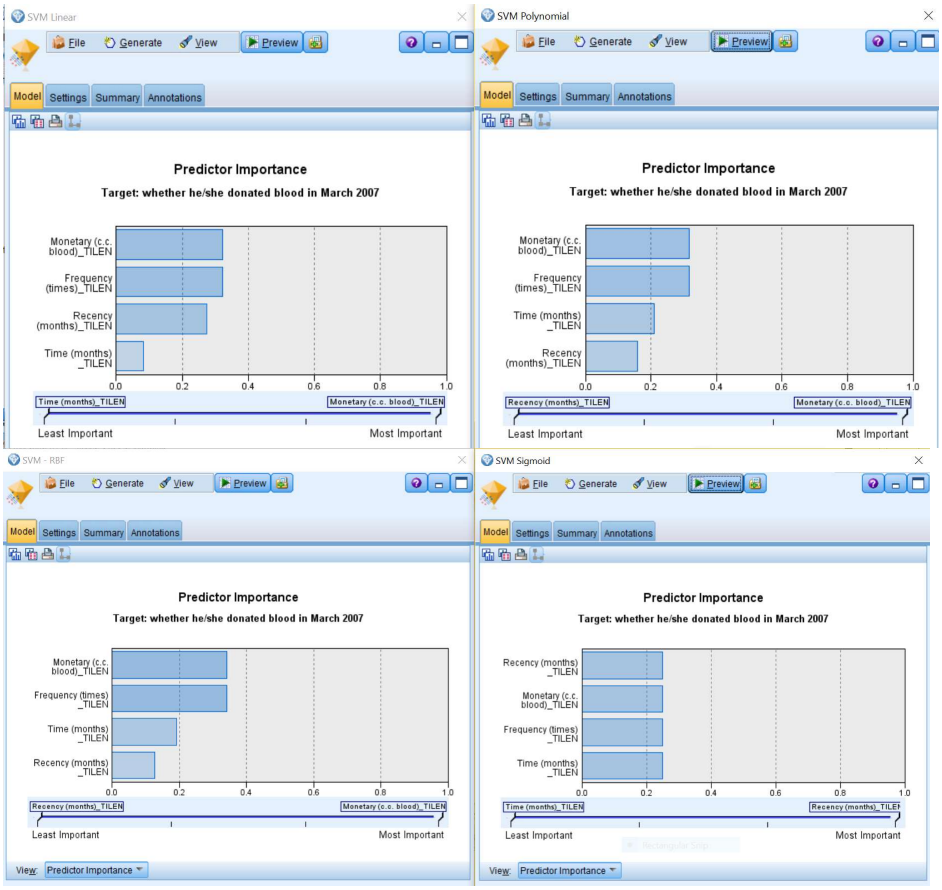


Figure 7 - SVM Model - Predictor Importance

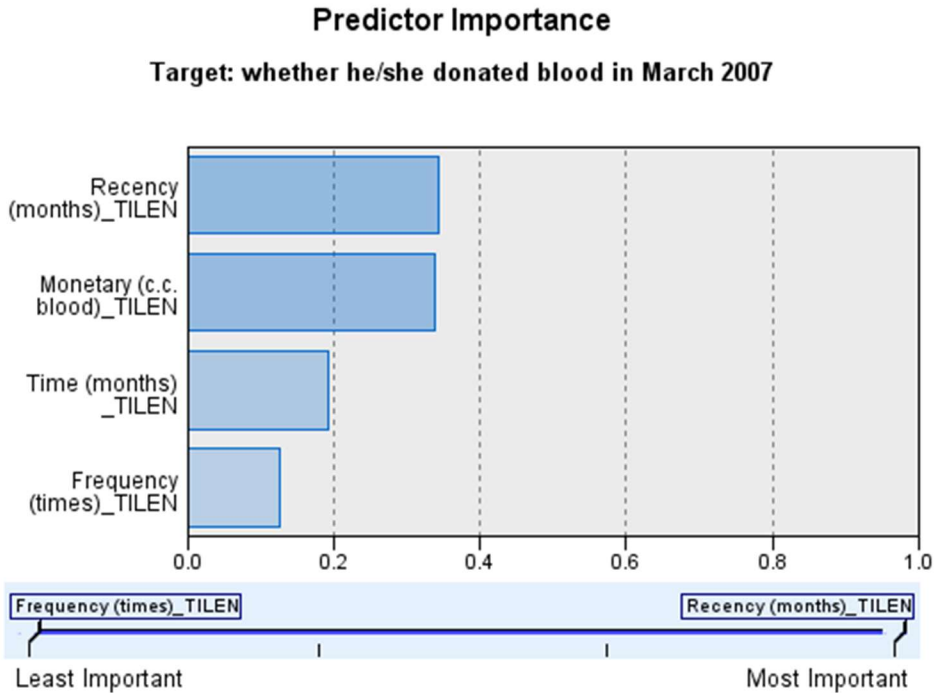


Figure 8 - ANN Predictor Importance

Model Summary

Target	whether he/she donated blood in March 2007
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	5

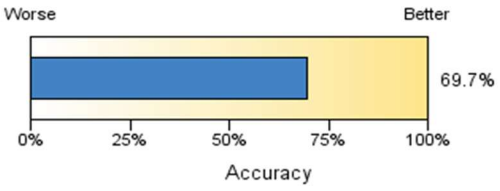


Figure 9– ANN Model Summary

Results for output field whether he/she donated blood in March 2007

Comparing $N-whether he/she donated blood in March 2007 with whether he/she donated blood in March 2007

Correct	232	68.24%
Wrong	108	31.76%
Total	340	

Coincidence Matrix for $N-whether he/she donated blood in March 2007 (rows show actuals)

	Will Donate	Won't Donate
Will Donate	99	71
Won't Donate	37	133

Figure 10 - ANN Coincidence Matrix

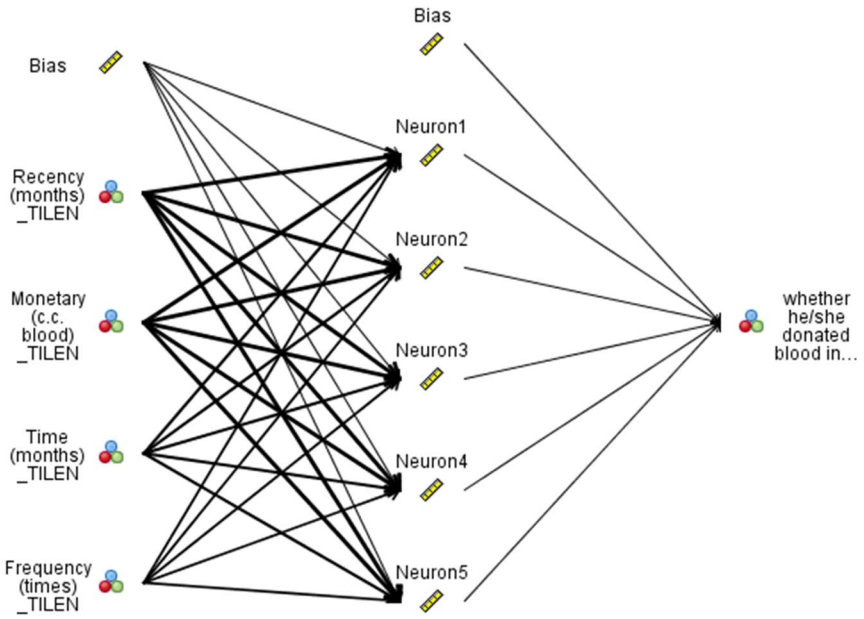


Figure 11 - ANN Network Connections

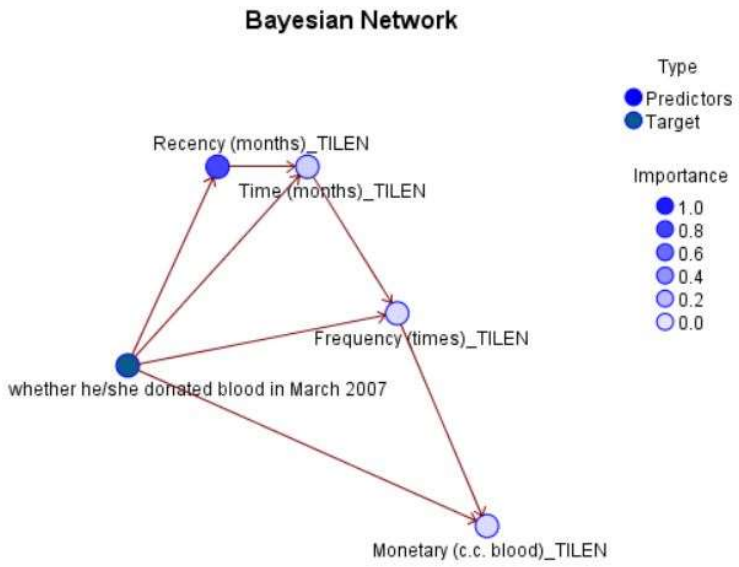


Figure 12 - Bayesian Network

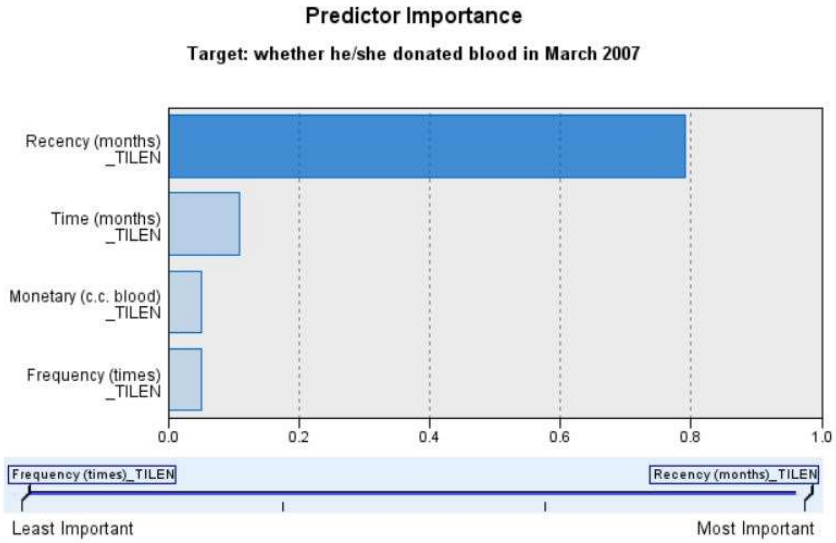


Figure 13 – Bayesian Predictor Importance

Comparing $B-whether he/she donated blood in March 2007 with whether he/she donated blood in March 2007

Correct	237	69.71%
Wrong	103	30.29%
Total	340	

Figure 14 – Bayesian Confidence Matrix 1

Coincidence Matrix for $B-whether he/she donated blood in March 2007 (rows show actuals)

	Will Donate	Won't Donate
Will Donate	116	54
Won't Donate	49	121

Figure 15 - Bayesian Confidence Matrix 2

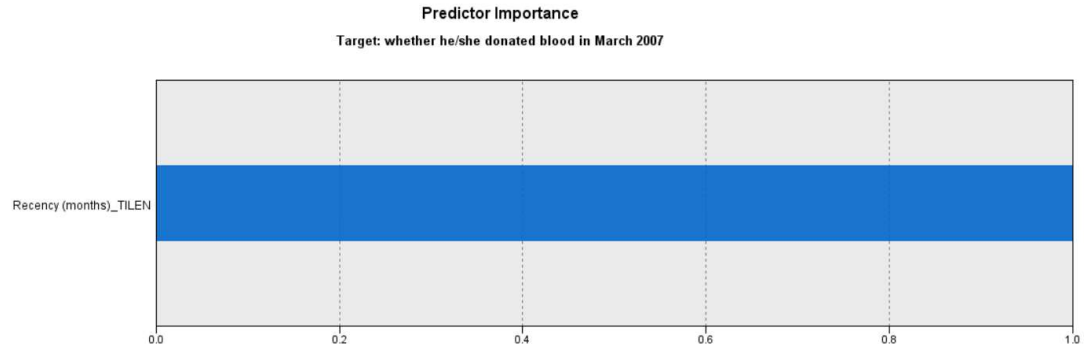


Figure 16 - C5.0 Decision Tree Predictor Importance

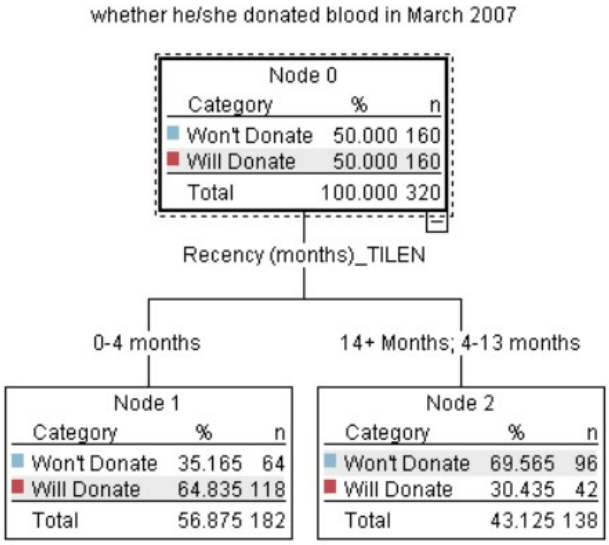


Figure 17 – C5.0 Decision Tree Leaves

Correct	223	69.69%
Wrong	97	30.31%
Total	320	

Coincidence Matrix for $C-whether he/she donated blood in March 2007 (rows show actuals)

	Will Donate	Won't Donate
Will Donate	119	41
Won't Donate	56	104

Figure 18 – C5.0 Coincidence Matrix

Metric	C5.0
Accuracy	0.70
Sensitivity/Recall	0.74
Specificity	0.65
Precision	0.68
F1	0.71

Figure 19 – C5.0 Calculations

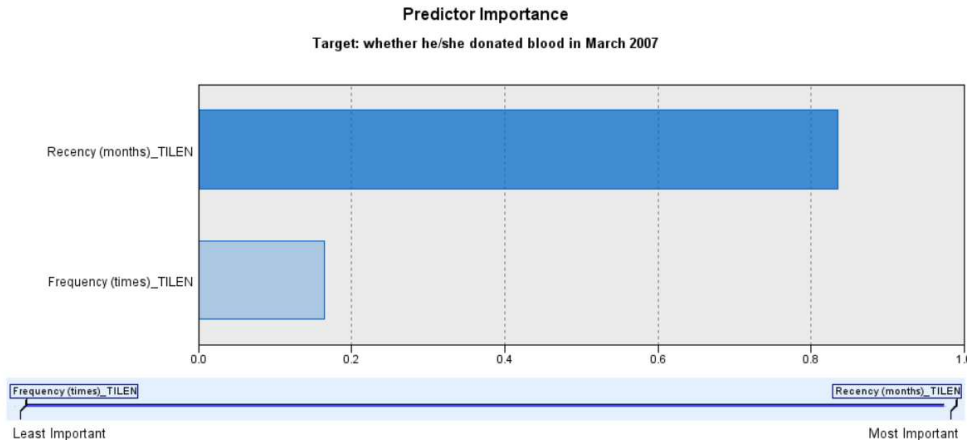


Figure 20 - CHAID Decision Tree Predictor Importance

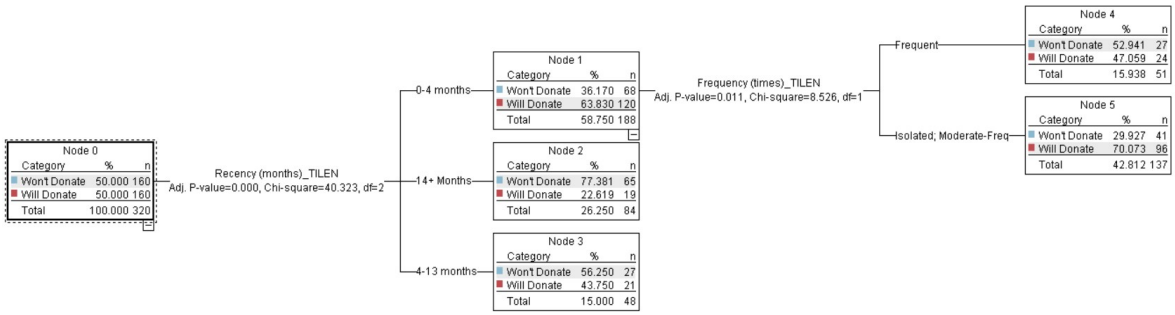


Figure 21 – CHAID Decision Tree Leaves

Correct	212	66.25%
Wrong	108	33.75%
Total	320	

Coincidence Matrix for $R-whether he/she donated blood in March 2007 (rows show actuals)

	Will Donate	Won't Donate
Will Donate	93	67
Won't Donate	41	119

Figure 22 – CHAID Coincidence Matrix

Metric	CHAID
Accuracy	0.66
Sensitivity/Recall	0.58
Specificity	0.74
Precision	0.69
F1	0.63

Figure 23 – CHAID Calculations

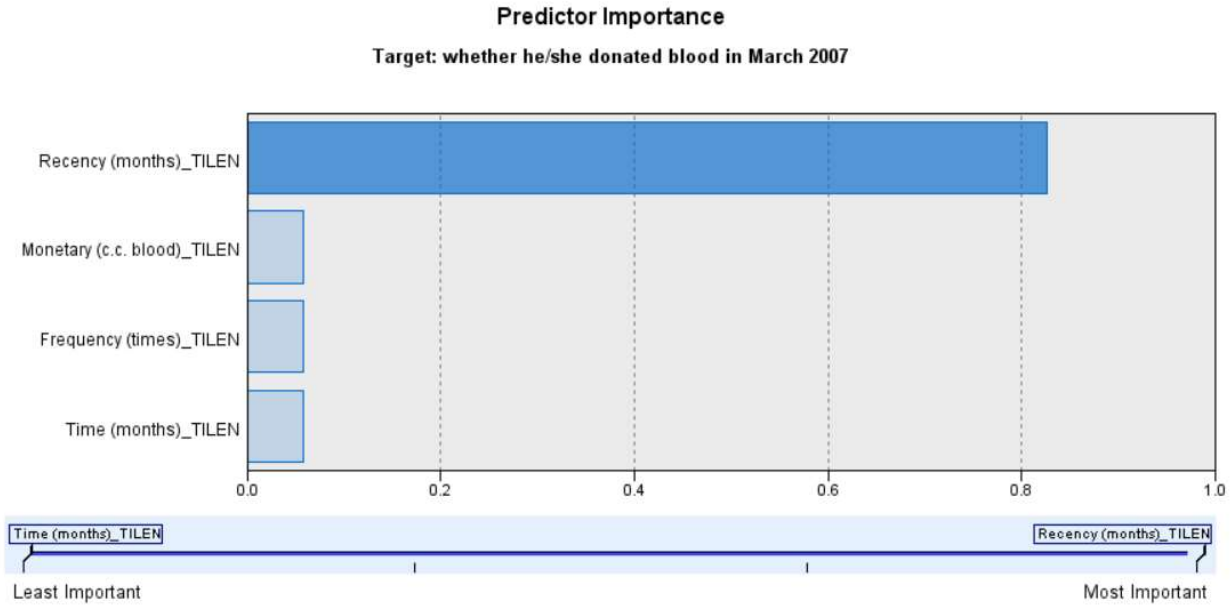


Figure 24 - CART Decision Tree Predictor Importance

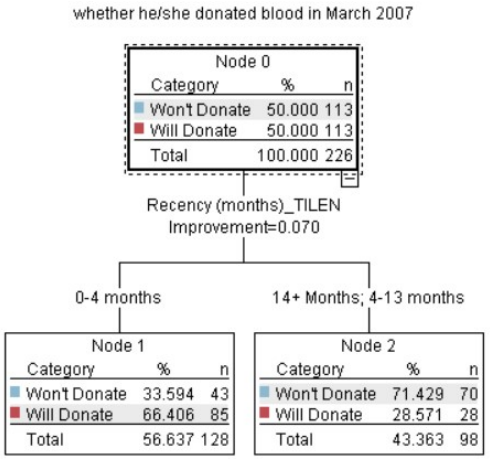


Figure 25 – CART Decision Tree Leaves

Correct	217	67.81%
Wrong	103	32.19%
Total	320	

Coincidence Matrix for $R-whether he/she donated blood in March 2007 (rows show actuals)

	Will Donate	**Won't Donate**
Will Donate	119	41
Won't Donate	62	98

Figure 26 – CART Coincidence Matrix

Metric	**CART**
Accuracy	0.68
Sensitivity/Recall	0.74
Specificity	0.61
Precision	0.66
F1	0.70

Figure 27 – CART Calculations

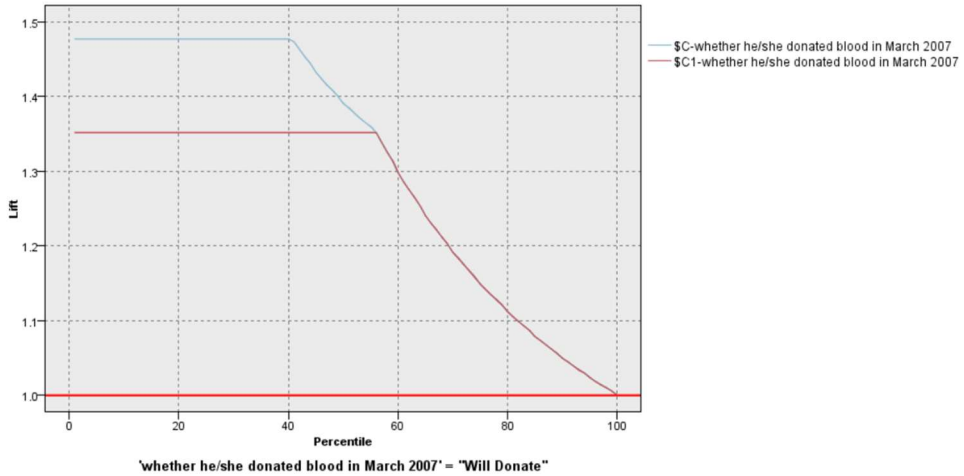


Figure 28 – Evaluation Node Lift Chart Showing Training versus Testing Data

Metric	C5.0
Accuracy	0.69
Sensitivity/Recall	0.76
Specificity	0.62
Precision	0.66
F1	0.71

Figure 29 – C5.0 Testing Calculations on Testing Data

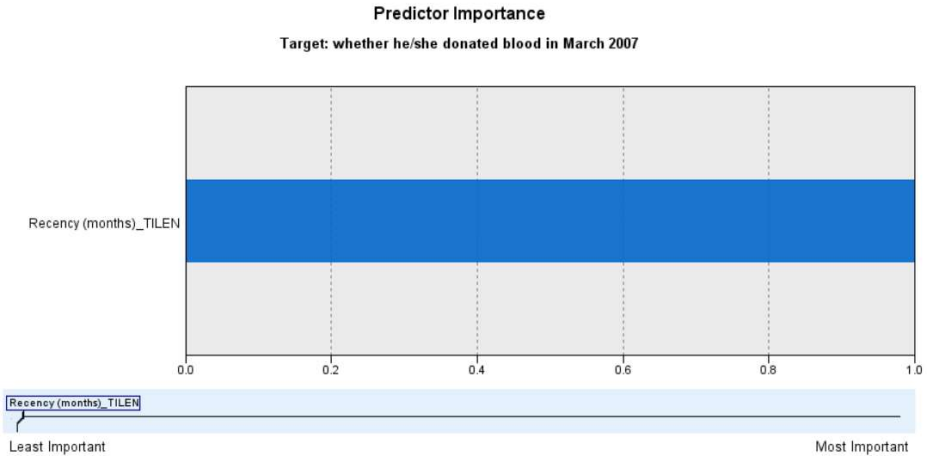


Figure 30 – C5.0 Testing Decision Tree Predictor Importance On Testing Data

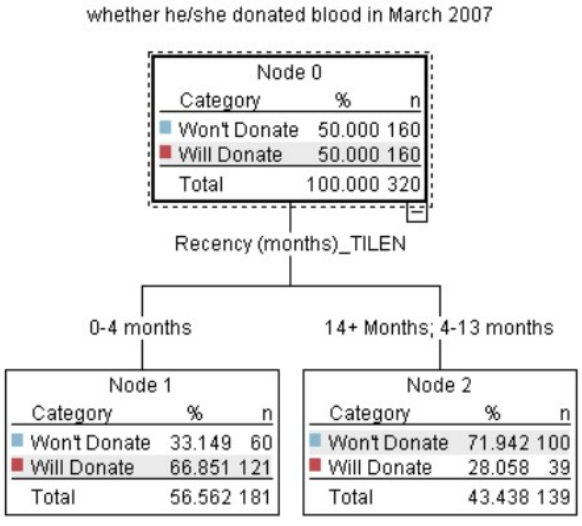


Figure 31 – C5.0 Test Decision Tree Leaves on Testing Data