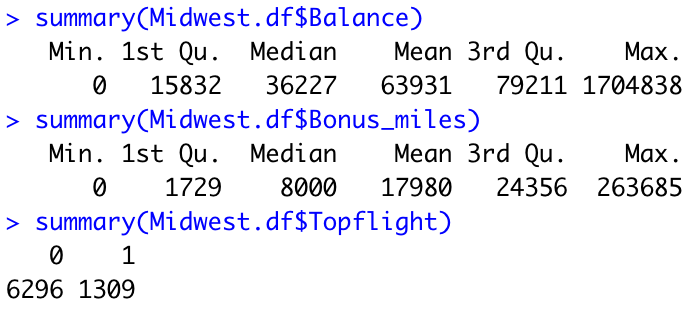
Matt Rechtien

Prescriptive Analytics – Term Paper

1. The CRISP-DM process is composed of six steps: business understanding, data understanding, data preparation, model building, testing / evaluation, and deployment. A data dictionary directly deals with the first four steps and sets up the final two steps by requiring the organization to learn about and become well-versed in the data they own. They are not just grabbing random sources to try and guess at the relationships.
   1. Business Understanding – creating a data dictionary requires that an organization understands what business problem they are trying to solve. Without knowing the goals for the project an organization cannot determine what data they will need to solve it. The descriptions provided help develop the business and data understanding.
   2. Data Understanding – the most direct impact of the data dictionary is the requirement to lay out the exact data and sources needed, and if the data that is available is sufficient enough to meet the needs of the project. If not, then you can supplement with created data. (This is the difference between raw or Telcom created fields).
   3. Data Preparation – a data dictionary explains the data type, max length and whether the data is raw or created. This makes the transformation of the data seamless.
   4. Model Building – knowing exactly what kind of data you are working with (types, dummy, etc) makes building the model easier because you don’t have to spend time finding out how to plot your variables.

The testing / evaluation and deployment aren’t directly impacted by the data dictionary, but by taking the time to build the dictionary, an organization has a better understanding of the model as a whole. Understanding that makes it easier to test and deploy the model as an end product.

1. After loading the csv, I set every variable in the data set to the data type listed on the Data Dictionary. Since we want to exclude the ID variables—as it is a unique identifier but has no impact on the resulting decision to purchase a phone service contract—I dropped the variable from the set. Here are the summary statistics for *Balance, Bonus\_miles,* and *Topflight* variables:



From an initial glance there are a lot of outliers in both the *Balance* and *Bonus\_miles* variables. (A boxplot verifies this hunch).

1. Lines 45 – 61 is where I calculate our first logistic regression. Our model output looks like:

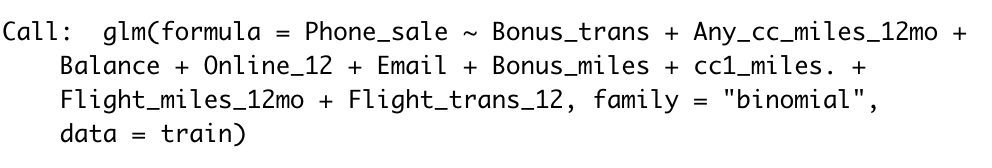
Table

Description automatically generated

And our odds ratio for 95% confidence intervals:

Text

Description automatically generated

1. We can run a step function to eliminate variables which leaves us with the following variables for our next logistic regression:

Then we can create a logistic regression with just this formula. This gives the following output and odds ratios:

Table

Description automatically generated

Text

Description automatically generated

1. Next, we’ll build our two neural networks which yield the following ROC curves.

Diagram

Description automatically generatedThis is for the 20-node network.

Diagram

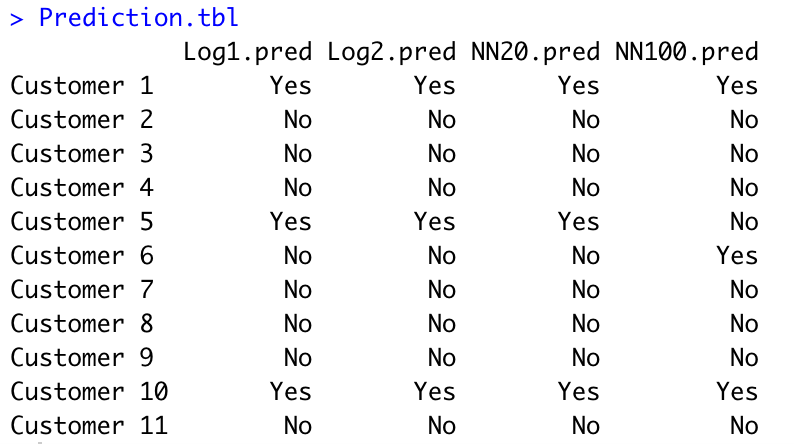
Description automatically generated with medium confidenceAnd this if for 100-node network.

1. After creating our four confusion matrices we can fill out the table with the following information:

Text, letter

Description automatically generated

1. For Midwest airlines I would the model to use depends on what they are most interested in achieving. If they need a model that will be mor precise, then they should choose the Neural Network with 100 nodes because it is the only one that predicted over 50% of the customers who purchased the contract correctly. However, the model with the best overall AUC score would be the logistic regression after the we used the step function to eliminate variables. The most accurate model is either our first logistic regression or the 100-node neural network. So, with all this information and the thought that a model that is better at predicting the people who would sign up from this campaign I would recommend our 100-node neural network for MidWest Airlines.
2. We can import the prediction data set, make sure that the variables are designated in the same way (char / num / factors) and then run the data through all of our models to determine if each customer would sign up for the card. This yields the following table:



The first three models have the same results, but the NN100 model swaps customer 5 and 6.