Project 8 - Decision Trees and Neural Networks

Samuel McMillan and Eric Johnson

Slides Link: Project 8 - Decision Trees and Neural Networks
Github link: https://github.com/Samuel8Gold/cs5830_project8

Introduction

In this study, we focus on analyzing previous marketing campaigns run by a bank to find trends within the demographic of users that have opened an account. We aim to create a model that will allow for banks to identify key characteristics in clients that will allow them to predict future clients. With banks earning money based on the number of clients, their goal is to get as many people as possible to open an account and make deposits. Our models will attempt to predict whether or not a current client will make a deposit with the bank. Using neural networks and decision trees with different depths and parameters, we were able to create models with precision scores as high as 87%.

Dataset

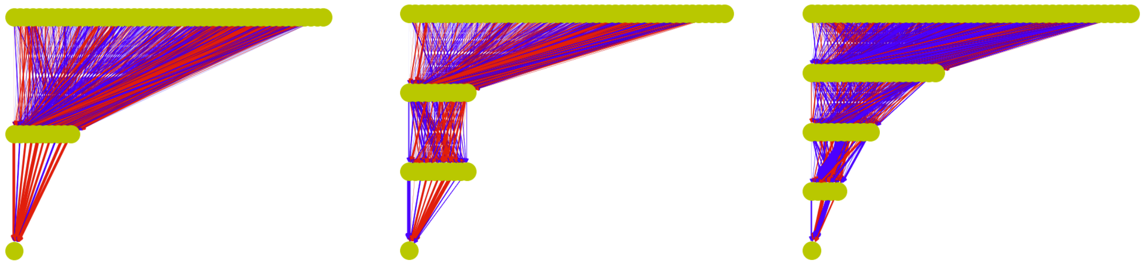
This dataset provides information regarding a bank's current customers and certain attributes that each customer has, as well as whether or not they made a deposit during the specified marketing campaign. Attributes include marital status, education, housing, loans, balance etc.

Analysis Technique

The analysis techniques we used included decision trees and neural networks. In order to find the best possible model for the dataset, we experimented with different depths in the decision trees, and different numbers of hidden layers with different amounts of nodes per layer. These models were appropriate for the dataset, as decision trees and neural networks both interact well with both continuous and discrete variables, especially categorical ones.

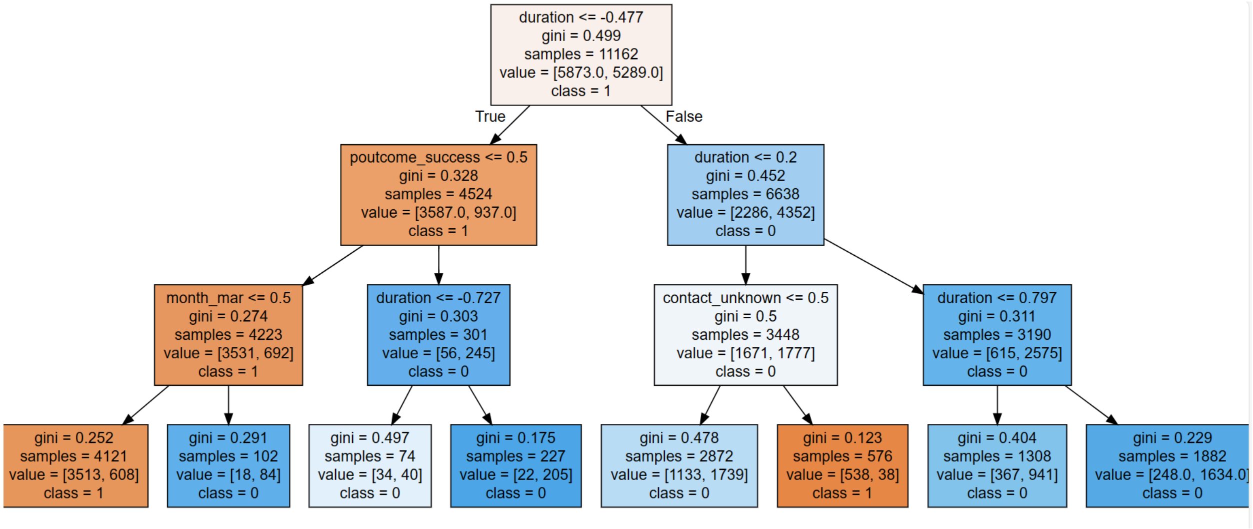
Results

For the neural network, we decided to implement 3 different architectures to investigate the effectiveness of the network. We tried a 10 node layer, a 10-10 node layer, and a 20-10-5 node layer. The visualizations for these networks can be seen below:



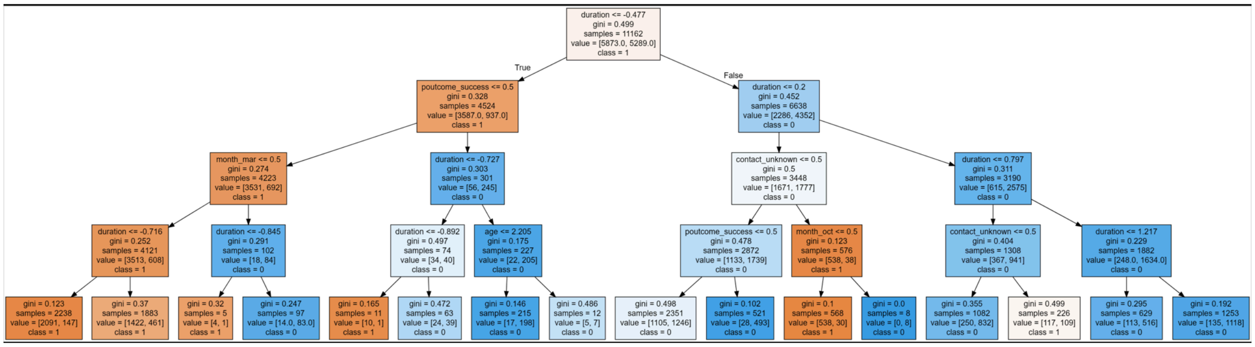
With the amount of input fields in the network, we have reason to believe that changing the amount of nodes in a layer and the number of layers does not impact our model 's accuracy, precision and recall. In fact both the single layer and double layer networks had the same F1 score of (85, 84), while the 3 layer network had scores of (84, 84), so a very nonconsequential effect.

Depth of 3 decision tree with all features:



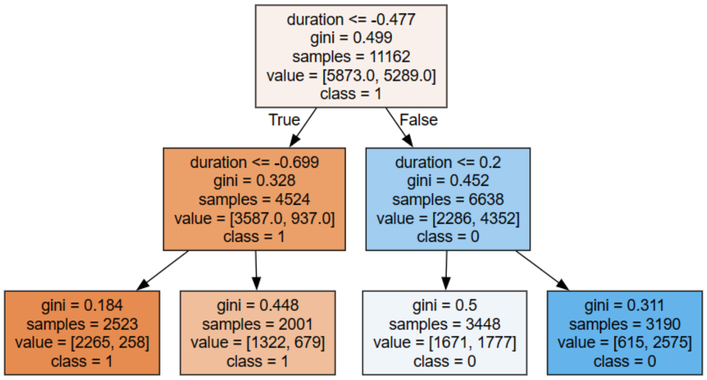
In the above, we can see there are reasonable correlations with it appearing the gini score to be .2-.4 on average. This didn't lead to the model being efficient, so we decided to modify it a bit as shown below.

Depth of 4 decision tree with all features:



With the depth of 4 we got Gini's of a bit better variety ranging from .25-.35 on average. Due to the many features the table created, it is reasonable to conclude that there are many factors that go into a successful marketing campaign. This means one should look into many factors one should look into. Some of the most prominent ones are poutcome_success, duration, and month (i.e., month_oct and month_mar). It is clear to see that time of last contact (i.e., the month) and the length of the chat of the last contant (i.e., duration) played a big role in determining whether

someone would make an deposit, so Bank Manager or whoever is running the campaign, I would start looking at how long and when you reach out to potential depositors first!
Just Duration Decision Tree:



The gini scores here averaged out around .35 which is the worst model, so far. That is surprising since it was seen that duration (i.e., duration of the last contact) was the main thing to look at when it came to predicting whether an individual would make a deposit. This means that in order to have a successful marketing campaign, focus on more than one area to work on and improve. Perhaps stick to a couple instead of just one thing to look at for a successful marketing campaign.

Technical

Upon initial examination, the data was rather clean for use. It did however require one hot encoding, as a significant amount of fields were categorical fields, and as such, yielded very large dimensionality in the dataset. One of these fields, “campaign”, was a categorical field with a significant amount of unique values. Because of this, we decided to remove the field, as its values alone were almost that of the entire rest of the dataset combined, and did not seem to yield a significant effect to the accuracy and precision of the model.

Neural networks and decision trees worked great for this dataset, as they do well using categorical and continuous variables, as long as the categorical variables are prepared properly. For the neural networks, we used one hot encoding, while for the decision trees, we used dummy variables.

Overall, there weren’t many hiccups we ran into. Aside from removing the campaign variable, the modeling went smoothly, and we just played around with different parameters to see what patterns we could find.