Customer Churn Analysis

A data-driven approach to proactively managing customer churn

1.0 - Executive Summary:

The ability to retain a customer in the service industry is a top priority for many organizations as it serves as a reoccurring/compounding revenue stream – the longer a customer is retained, the more profit is earned. Studies also show that the cost to acquire a new customer is anywhere from 5 to 25 times greater than the costs of retaining current customers (HBR, 2014). However, the traditional churn measures are lagging in nature – that is to say the customer has already made the decision to leave before the company knows they are severing ties. Most organizations opt to take a proactive approach toward reducing customer churn, and build teams of customer advocates whose main goal is to forge relationships with their customers and, hopefully, figure out customer's needs and pain points before they close the door. However, this becomes challenging as the business begins to scale and each advocate's book of business increases. Wading through the sea of customers in search of the ones who are in need of help can become downright impossible, and the need for effective customer management becomes clear.

In this study we take a data-driven approach to proactively managing customer churn. We use a real-life data set from the financial industry, customer data from a bank, and model the outcome of if the customer left the bank or stayed (in the specific period of time). The data set was downloaded from Kaggle.com (located here), which is a popular data science platform. The outcome response is the column labeled "Exited", and a value of 0 indicates the client stayed with the bank, where a value of 1 indicates they left.

When choosing the methods to apply to this classification problem, we make two arguments: first, we are ok with "over identifying" customers as likely to churn since more touch points with customers is never a bad thing (however, we hope to filter it down to something manageable by a customer advocate department). Secondly we don't make model interpretability (i.e. "understanding the inner workings") a top priority, which allows us a wider selection of methods to choose from - this study is focused on finding an efficient way to identify customer churn from a large list of customers. Therefore, we opt for the powerful neural network as our leading candidate model since it works well with larger datasets and are more flexible to nonlinear problems. Secondly, we choose the logistic regression since there appears to be a lot of overlap between the outcome classes for many of the predictor variables. Both of these methods will be described in further detail in the analysis.

2.0 - Data Exploration:

In this section we review the dataset provided. An initial look reveals that there are 10,000 customers with 14 total variables in which there are two prevailing categories – those that relate to customer demographic information such as country of residence, age, gender, estimated salary & credit score, and variables that relate to the customer's business relationship with the bank such as tenure, their account balance, how many products they use, if they have a credit card, and if they are an active member. We are also provided row number, customer id, and surname; however, these are more identification type variables and provide no value to this study and are therefore dropped from the analysis. Accounting for our response variable “Exited”, this leaves us with a remaining 10 variables to use in our modeling process.

With the remaining variables, there is a good mix of data types – some of the variables are categorical such as country of residence, gender, having a credit card, being an active member. We also have count type variables such as the number of products a customer has, and the tenure (in years) with the bank. Finally, we have continuous values such as age, credit score, and estimated salary. It’s important to point out these variable types, as we’ll need to not only treat these types differently within the data preparation process, but also with any interpretation we do after the modeling process.

Our next step is to analyze the distributions of the variables and check for any outliers or missing data. Surprisingly, there are no missing values within this dataset and no imputation or dropping of observations is needed! Next we check the outcome variable “Exited”, and we notice a class imbalance where we have 1 record of customer churn for every 4 records of customers who do not churn. This will impact our choice of performance metrics, which will be further detailed in the modeling process section.

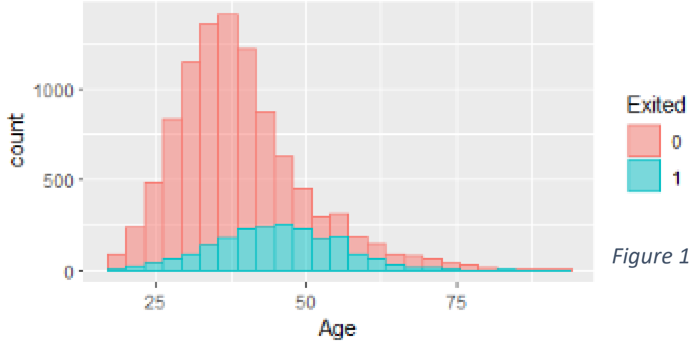

Figure 1

When reviewing the distribution of the Age variable, we see it is skewed to the right (fig. 1). Therefore, we choose to apply a log transformation to the Age variable to make it more normally distributed before beginning our modeling process.

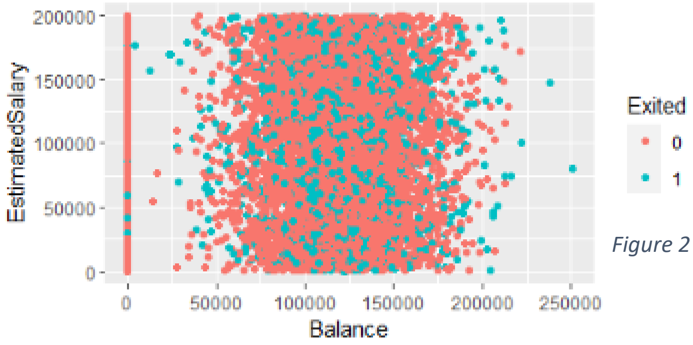

Figure 2

Our last step in the exploration section is to check relationships among variables. We are specifically looking for any patterns or relationships between variable pairs (potential of collinearity), and relationships between variables and the outcome. As we hinted in the executive summary section, most variables tend to have relationships like the one seen here between estimated salary and balance (fig. 2) – i.e. there is no relationship between the variables themselves or with the outcome. Because of this, we opt for non-linear methods as the leading candidate.

3.0 – Data Modeling

Now that we’ve conducted our initial exploration of the dataset and completed the preparation on our variables, we can move into the modeling process. However, before we begin, we need to choose our performance metric that will be used to compare the different models and will allow us to select the “best” model as output from this process. We choose the area under the curve (AUC) measure on the receiver operation characteristics (ROC) curve as our performance metric of choice for the following reasons:

1. It measures a model's sensitivity vs. its specificity - in other words, the model's ability to correctly pick out customers that will churn vs. correctly pick out customers who will not churn. It provides an intuitive way visualize the trade-off between the two.
2. It's more robust than just a raw accuracy measure since we identified that the outcome in this dataset is skewed toward customers who stay with the bank, a 4:1 ratio. If we just used raw accuracy, it'd be easy to get a high naïve accuracy by just assuming every customer is staying – this isn't how we should be optimizing our model and would defeat the purpose of the study!

Our first candidate model is the neural network. We choose a neural net since there is a lot of data available to us, which is where these methods perform well. Secondly, there doesn't appear to be linearity among our variables (as described above), which is also where a neural net can provide utility.

The neural networks we are working with use two types of tuning parameters – the node size, and the decay rate (to prevent overfitting). To find the "best" values for these parameters (among a selection of values), we use 10-fold cross-validation. At each "fold", we use 10 different values for node size, and 10 different values for decay to fit the model and make predictions – the outcome is 100 different models. We use the predicted outputs for each of these models, and calculate the AUC score respectively. Here we identify the model with 7 nodes and a decay of 0.1 as the best neural net model with an AUC score of 0.87 (fig. 3).

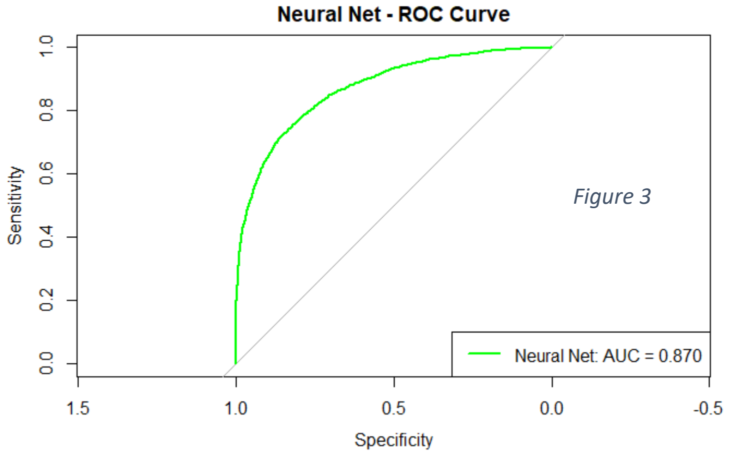

Figure 3

Next we will fit the logistic regression model using an exhaustive stepwise technique within the same 10-fold cross-validation process as implemented for the neural network. For each fold of the cross-validation process, we evaluate all possible combinations of variables and determine the "best" model (some variables may be excluded). We also calculate the AUC score (0.771), and add it to the ROC graph (fig. 4).

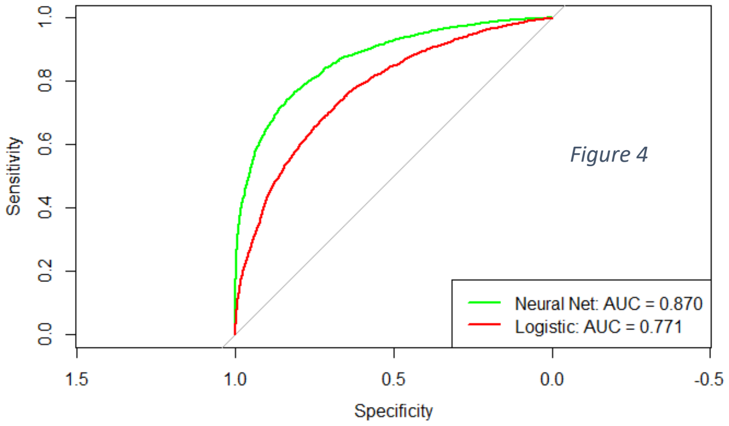

Figure 4

In comparing the two models , we see clear evidence that the neural network is outperforming the logistic regression model and therefore is selected as the choice model.

As a final step, we need to assess the model on truly "unseen" data to give an honest assessment of performance. We wrap an 80/20% (train/validation) split around the model fitting process (our sample size is large enough) where 80% goes into the fitting process, and 20% is reserved for the true assessment. We calculate an expected honest AUC of 0.848 using the neural network.

4.0 – Model Interpretation & Recommendation

Although we've made the priority in this study to be more focused on maximizing model utility over interpretability, we are able to glean some information as to how the predictor variables impact the outcome of a customer's decision to stay. The graphic below (fig. 5) is known as an Olden plot (Olden and Jackson), and displays the relative importance of each variable toward the outcome.

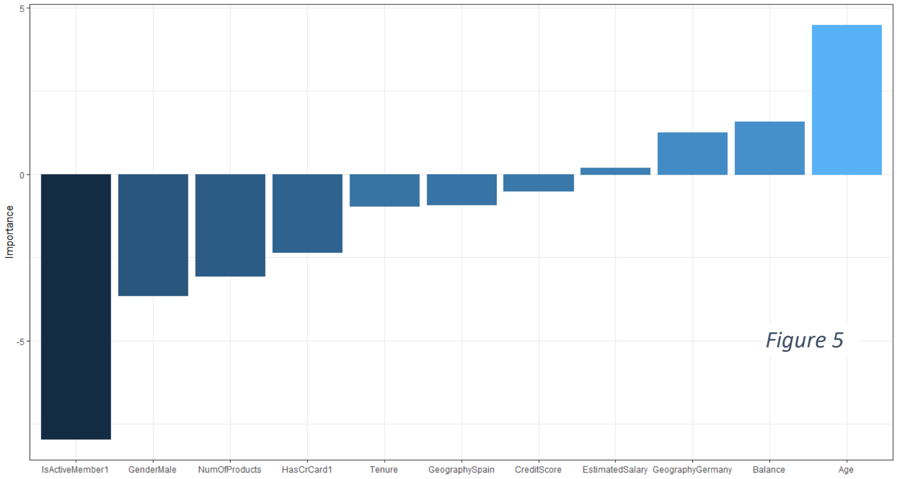

Figure 5

We can see clear evidence that the variables age, account balance, and customers located in Germany are more associated with churning, whereas the variables being an active member, male, and number of products is more associated with not churning (which makes logical sense – the more active and engaged, then less likely to churn!).

We also store the predictors and coefficients from each fold in the logistic regression model fitting process, and analyze these. To no surprise there is quite amount of overlap with the Olden plot, and we can identify an increase in the number of products owned means lesser odds of churning, and consequently an increase in age and balance means even greater odds of churning.

Overall, we'd suggest targeting marketing campaigns for those older, German customers with a high account balance as a means to proactively tackle the issue of customer churn from a different (sales and marketing) perspective.

5.0– Summary and Conclusion

In this study, we took a data-driven approach to the very important strategic effort of customer retention. We did a deep analysis of the dataset provided, and made the case for the implementation of a neural network model as a mechanism to proactively identify customers, with a certain level of sensitivity, who are likely to leave the organization.

We also interpret our candidate model, and although we cannot directly quantify their impact with the neural net, we are provided with some indications of those variables that influence the specific outcomes greater than the others. This interpretation provides opportunities for action, and should be considered for further action and research. Using this model in production to identify customers likely to churn will give a "leg up" to customer advocates and help them maximize their return on retention efforts.

References:

Amy Gallo. Harvard Business Review. (2014, October). *The Value of Keeping the Right Customers.* https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

Olden, Julian D, and Donald A Jackson. “Illuminating the ‘Black Box’: a Randomization Approach for Understanding Variable Contributions in Artificial Neural Networks.” *Ecological Modelling*, Elsevier, 18 Apr. 2002, www.sciencedirect.com/science/article/abs/pii/S0304380002000649?via=ihub.