Analysis of the Bank of America “Prefered Rewards” Program enrollment

Christopher Botica, Betül Çam, Nalin Gupta, Anushka Tak

DS 5220 Supervised Machine Learning , Fall 2018

I. Program Overview and Objective

Preferred Rewards is a form of loyalty membership program where clients get rewarded for deepening their relationship with the bank. It is designed to deepen the relationships by consolidating savings, checking, investments, and retirement funds into a “one-stop” banking with Bank of America. Benefits include higher credit card rewards, advantageous interest rates, no brokerage trading fee and others.

The project goal is to predict the enrollment based on client behavior. Given the sensitive nature of client relationship, we prioritized prediction over explanation. Furthermore, in searching for the optimal model, we build our criteria of “optimal” by taking into account the bank's goal: to use this model for targeted marketing. We will develop this concept in section 5.3. Examples of potential uses for the bank include: client outreach via direct mail or email, developing accurate, actionable leads lists for Financial Advisors, targeted advertisements in post-authentication zones, (e.g., pop-up reminder boxes while client is using online banking or ATM)

II. Data Collection

Client data gets captured at the end of each transaction and stored in multiple databases. After detailed due diligence and various level senior executive approvals, we were granted access to randomly selected 127,240 fully anonymized client data. The clients were offered Prefered Rewards Program privileges; some enrolled and some declined the offer.

Data set includes individual observations between October 2015 and October 2017 with over 150+ different variables collected in 7 different time periods to observe continuous behavior.

Data captures wide range of detailed information at about a client's behavior, including the types of products heavily used, prefered channels of communication, client's wealth segment, client's preferred level of engagement with the bank associates, whether client had any complaints, how many inbound and outbound transactions completed, how much revenue the bank makes from each of the products owned by a client, how much a client costs to the bank, what kind of punitive fees paid by the clients, and more.

III. Data Preparation & Feature Engineering

We were given a rich data set which came with unique set of challenges. The analyst initially pulled the data created multiple nested subtotals in the data but had limited availability to explain the details and variable names. As a team, we have pulled a sample of 50 records for all 1200+ available features and examined each one to determine which ones were the building blocks to the subtotals. For example, checking, savings and CDs were added to create a column called “Non_Brokerage Deposit Account Balance”. Then, newly created subtotal became a part of another subtotal. Once we stripped the data to its building blocks, then we have leveraged our domain knowledge to identify the relevant ones. For example, we have excluded revenue from client and cost per client fields because they were highly correlated to the balances and usage frequency. Also, clients who enroll in the program tend to have deeper relationship with the bank which makes them unlikely to pay punitive fees. Final variables included are 3 month rolling average balances for deposit, investment, credit and other product categories, open account totals for each of those categories, segment indicators as dummy variables, channel behavior (online banking, ATM,mobile banking, credit card, call centers, branch visits, etc) and client complaints.

All of data categories were repeated for 7 time periods: at the time of enrollment, one- three-six months prior enrollment, and one-three-six months post enrollment. We have used the prior time periods.

IV. Data Exploration

A. Data distribution and Transformation

Our dataset is well balance with respect to our target variable: a total of **53,359** cases of enrolled clients and **73,881** cases of non-enrolled.

One major setback in our dataset is that some predictors change based on the temporal development. In this case, our model would likely misinterpret the external changes for signal and could overfit when applied to future data. Fortunately, certain variables such as credit card balance and deposit balance have very similar distribution across each time frame. Analyzing these distributions, we find the variables are skewed. We remediate this issue by applying log transformations to the rolling month average balances.

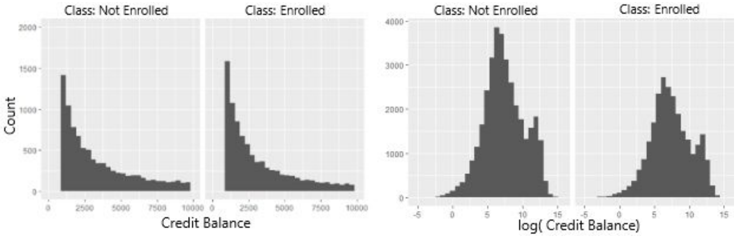


Fig 1. Log transforming credit balance

On the other hand, skewness is expected in other variables such as number of deposit accounts . We expect the results to be skewed towards the left. For example, the majority of clients have less than 5 accounts but there are very few clients that have more then 20 open accounts. This high number of accounts may be indicative of an unusually high user participation. Therefore, we leave such variables untransformed.

B. Variable Selection and Trend Analysis

Analyzing the trends and behavior of each predictor in relationship to the dependent variable provides us the much needed insight into choosing and fine-tuning our model. We completed a thorough analysis of the predictors and used this analysis in our a priori model comparison in section 5(A). For the interest of space, we will document the analysis on a smaller subset of predictors: “client product summary”, “client channel preference”, and “total number of complaints”, “rolling month average balance”, “money transfer”, and “online banking”.

Client Product Summary An analysis of the “client product summary” suggests that the few cases of clients who have only a credit card with the bank are more likely to join the program. Another important trend among the underrepresented categories is is that customers that have only investment accounts show no promise of enrollment. However the vast majority of clients have deposit accounts or multiple accounts (under the category “multi product”), which show no significant preference towards enrolling or not enrolling. Given that there is limited information about the more informative categories (i.e., “credit card only”, and “ investments only”), we find that there is very little signal and more noise for this problem

Client Channel Preference This variable displays a better separation between the classes compared to other predictors. Specifically, we find that clients with mixed channel preference enroll more often, whereas clients with no preferences do not. The clients with no channel preference seem to indicate a lower customer engagement, which in turn, could translate to a lower likelihood of enrollment.

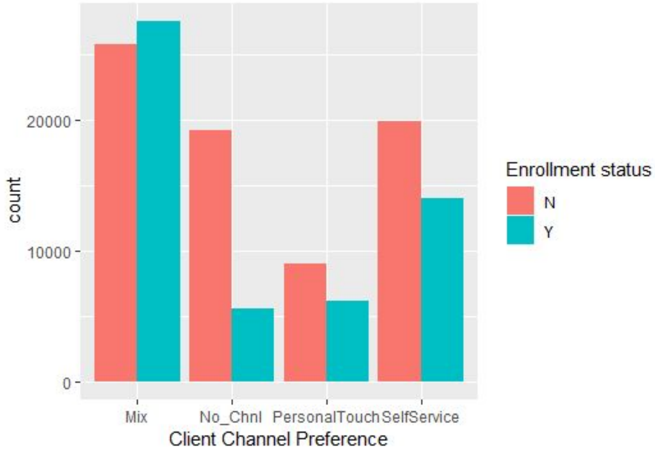
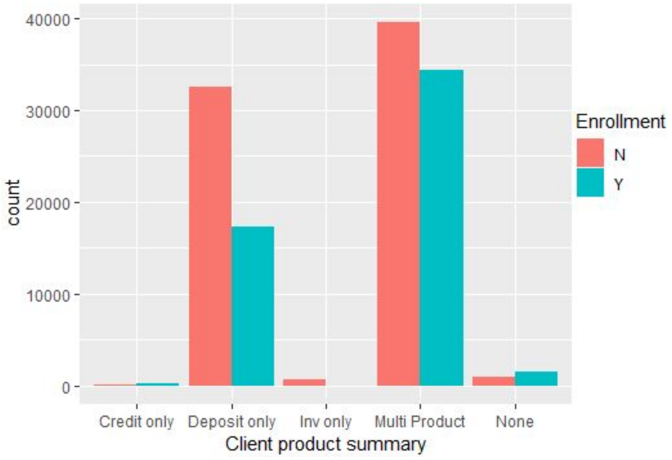


Fig 3. Client channel preference distributions

Deposit Balance We see a clear trend that clients who are enrolled in the rewards program usually have a higher account balance than clients who are not enrolled. However, we also observe a large number of outliers, we could be possible since there be a number of clients who have deposit balances much higher than most of the other clients.



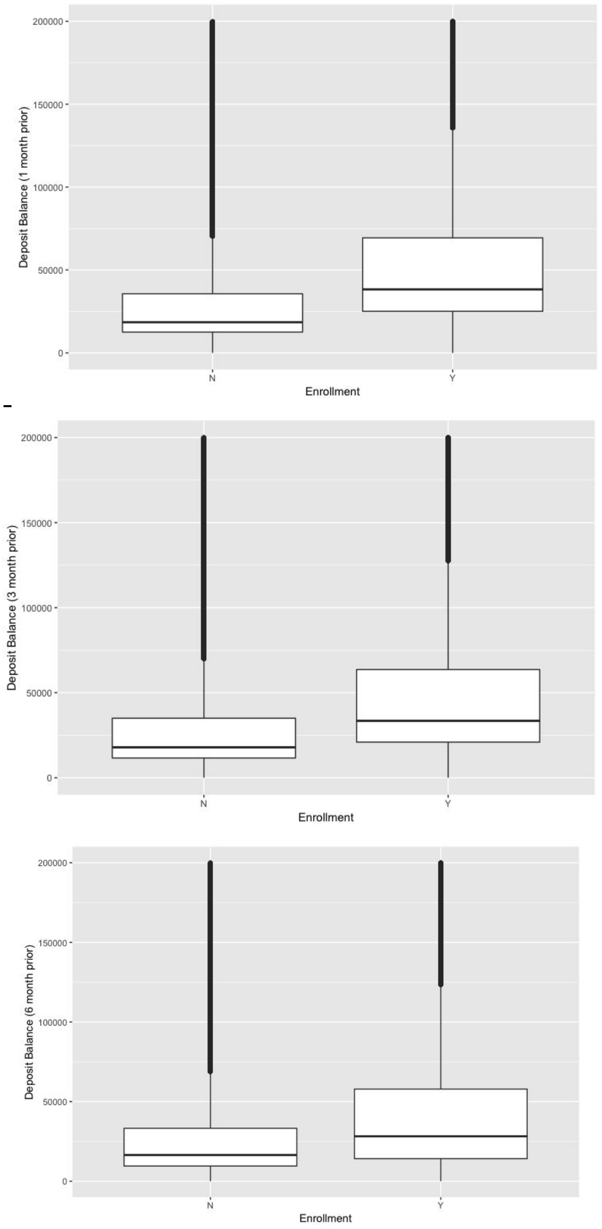


Fig 4. Distribution of deposit balances for each enrollment class

Money Transfer We see as the number of money transfer increases, we find more distribution in the 'Yes' enrollment category than 'No'. Therefore, the likelihood of one enrolling in the program increases as well.

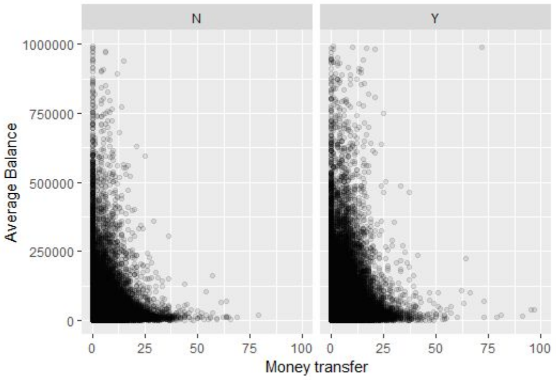


Fig 5. Analyzing money transfers according to enrollment class

Online/Mobile Banking Trend Analysis Fetching from the graph, we can tell the distribution is similar in both class categories in mobile banking whereas there is higher likelihood of enrolling in the program for the customer who engages in more online transactions. Though, there exists some outliers (as would be in real world) that corresponds to be in favor of enrollment.

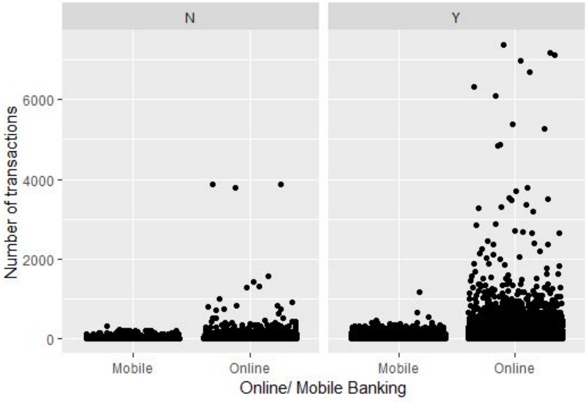


Fig 6. Distribution of banking trends facetted by enrollment class

V. Data Modeling

A. A Priori Modeling Comparison

Given our objective of predicting the client enrollment, we want to compare a wider range of models and techniques and select those that maximize our prediction. Therefore, we use our a priori knowledge of each candidate model, specifically its strengths and weaknesses in accordance to our problem, to select the best candidates.

SVM and **Decision Trees** are widely used models for classification. They both make few assumption about the data and can handle non-linear features very well. However, both of these models yield direct classifications, instead of the probability of individuals enrolling in the benefits program. Unlike computer vision classification problems (for example, cat / no cat), human behavior is not a clear yes or no. Thus, obtaining the likelihood of enrollment is more desirable for our project, specifically due to the possibility of fine-tuning the performance criteria and selecting the optimal cutoff for classification discussed in Section 5C. Therefore, we consider the following four models that yield the likelihood of enrollment, rather than a direct classification.

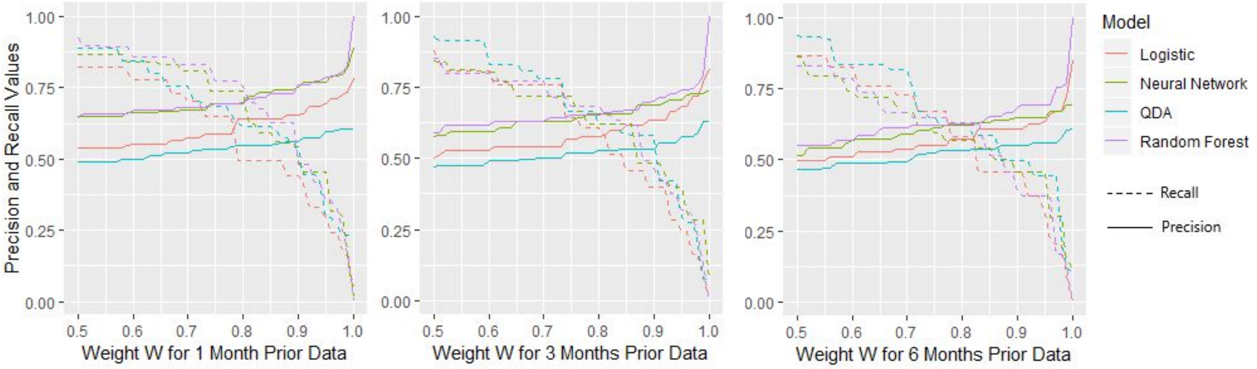


Figure 7. Distribution of Precision and Recall for range of weights *w*. Each value corresponds to the cutoff that maximizes the weighted F_1 score for *w*. Note that an increase in *w* corresponds to an increase in the cutoff.

Logistic Regression is our first candidate and perhaps the most baseline model for classification. It derives its strengths in its simplicity: inherently linear with few parameters to train. Its decision boundary is a hyperplane, reducing the overall bias of our model (compared to more complex and flexible models), however making it deficient in accounting for non-linearity in the data. Furthermore, we expect the logistic model to be unstable in handling highly correlated predictors and well-separated classes mentioned in Section IV.

Even though it is not our primary purpose, the logistic model allows us to interpret the odds of a given class. More importantly, the model yields the probabilities of success (i.e., enrollment), rather than a direct classification. Literature on the topic of logistic regression[1] suggests that the data should contain at least ten events for each predictor in the model. Otherwise, problems such as increased variance and also bias in the regression coefficients could occur[1]. Given that we have 120,000+ events and 22 predictors, we avoid this problem.

Neural Networks impose fewer assumptions about the data and are more complex, compared to their linear model counterparts. This allows the neural network to better account for non-linearity in the data, at the cost of a significant increase in number of parameters to estimate and hyperparameters to choose from. Thus, this model is more sensitive to the signal of the data and allows for many more options in the modeling process, all of which open more possibilities for overfitting, especially when the signal-to-noise ratio is small.

QDA (Quadratic Discriminant Analysis) is an excellent candidate due to its flexibility by incorporating a quadratic decision boundary and its ability to handle well-separated classes. We expected QDA to handle the bias-variance tradeoff well, given that it is a happy medium between the more inflexible logistic regression and flexible neural network.

Random Forests use multiple decision trees to make a decision for a given observation. Each tree is given subset of the data and a subset of the features and a max vote of the trees is taken. This reduces variance and is able to capture complex signals. The depth of each tree and the number of trees are the crucial hyperparameters which need to be tuned. If they aren’t chosen carefully, random forests can easily lead to overfitting. Random forests also lack interpretability.

B. Model Architecture

Based on our results from our a priori model comparison in section 5(A), we decided to employ logistic regression, QDA, neural networks, and random forests on each of our three temporal datasets independently: one month prior data, three months prior data, and six months prior data. We performed a 80-10-10 split on each of these datasets to obtain training, development and test sets and trained the 4 models on the independent training sets. We used our development set to tune the hyperparameters for each of the models. Next, we applied our trained models to the test set, and obtained the likelihood of customers enrolling in the program.

The distribution of these probabilities from the development set for each model and each time frame, is shown in figure 8.

C. Performance Criteria

Even though traditional classification metrics, such as selecting the class with the highest probability, or computing area under the ROC curve, would provide us with a classification, this may not be the optimal. For example, the ROC summarizes model performance over all thresholds, however some of the regions (such as in the extremes) are useless to our performance measure as they correspond to high false positive and false negative rates. The key here is defining what “optimal” means. Since the bank has limited resources and wants to concentrate on marketing for potential members, rather than non-members, we want to minimize the false-positive rate and increase the true-positive rate. This means maximizing the precision and minimizing the recall.

Maximizing the precision alone would be possible by simply selecting a very large cutoff *C* and classifying every customer according to the cutoff (those with likelihood greater than *C* will be classified as enrolled, and vice-versa). However, maximizing precision would lead to an increase of *C* to very close to 1. In other words, this process wouldn’t allow the model to risk incorrectly misclassifying a customer as enrolling (i.e., the false positive), so instead it would set every observation as a negative (i.e., not enrolled), even though some customers enroll. In order to combat this problem, we take into account the precision-recall tradeoff and use a weighted harmonic average of precision and recall. In other words, a weighted F_1 score.

$$F_1 = \frac{1}{\frac{0.5}{\text{recall}} + \frac{0.5}{\text{precision}}}, \quad \text{Weighted } F_1 = \frac{1}{\frac{(1-w)}{\text{recall}} + \frac{w}{\text{precision}}}$$

where *w* is the weight applied to precision. Note that the regular F_1 score has equal weights for precision and recall. Using this new metric we can apply a higher weight to precision, favoring it over recall. Therefore, the optimal decision criteria for the bank is choosing a cutoff value that maximizes the weighted F_1 score.

Note that the weight w is a hyperparameter. In fact, we intentionally allowed for this hyperparameter in order to let the bank tailor it according to its need. For example, if the bank has a tighter budget and is more averse to risk, it would increase the value of *w*, in order restrict the output of the model to only those customers with a very high likelihood of enrollment. Conversely, if the bank is willing to invest more in advertisement, it would relax *w* and allow for more customers to be classified as potential members (even though some might not enroll).

Since the purpose of our project is classification, we decide to tune *w* according to the distribution of precision and recall (across all values of *w*). Note that we want to favor precision over recall. Thus, we will apply a weight *w* greater than 0.5.
For each value of w from 0.5 to 1, we calculated the cutoff that maximizes the weighted F_1 score, with it associated precision and recall, and plotted these in Figure 8.

With no specific guidance from the bank on the level of risk to take, we make a subjective decision based on the distributions of precision and recall. We consider that the weight *w* = 0.9 maximizes precision, while also accounting for recall. Thus, our weighted F_1 score becomes

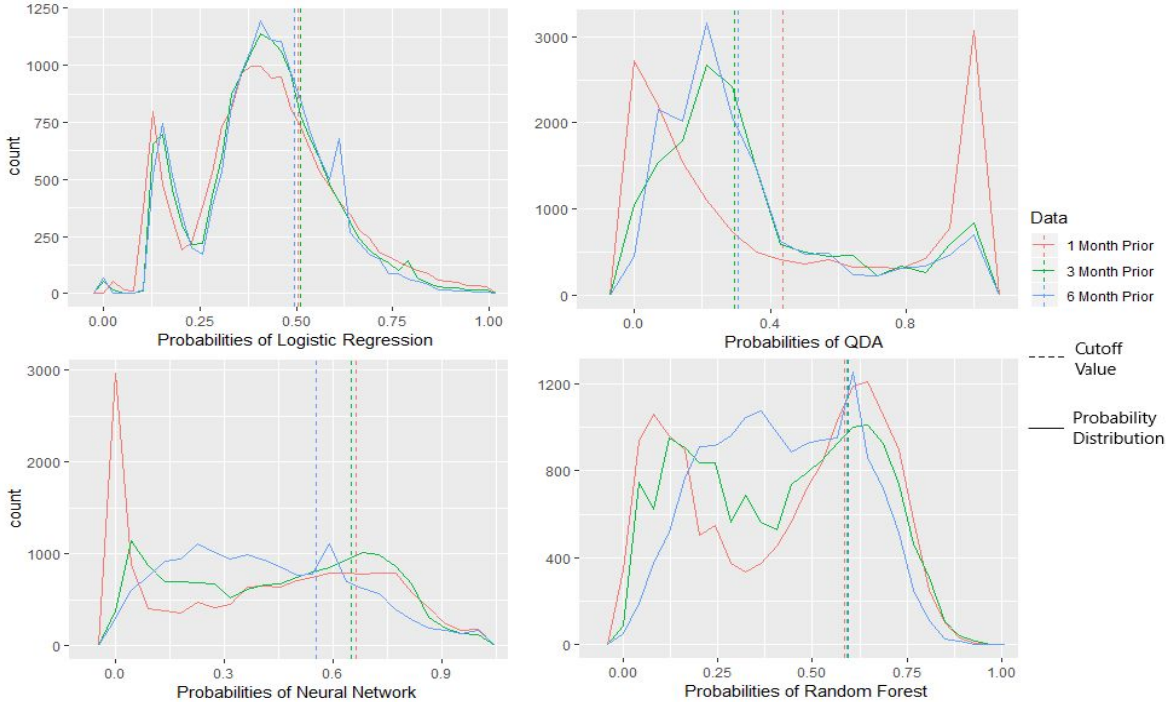


Figure 8. Distribution of the likelihood of enrollment for each model and time data.

$$\text{Weighted } F_1 = \frac{1}{\frac{0.1}{\text{recall}} + \frac{0.9}{\text{precision}}}$$

D. Probability distribution

For each model and each time-series data, we find the cutoff that maximizes the weighted weighted F_1 defined in figure 9. We find the distribution of these probabilities on the development set, along with the optimal cutoff (found on the training set) and plot these in Figure 8.

We can observe from the figures that there is clear distinction between the plots for different models. In the case of a random coin toss, we would expect the probability of client enrolling to be 0.5 and hence the graph of probability distribution would just be a straight line through 0.5. Keeping this base case in mind, we would expect our models to have a larger spread (flat peaks) in the center and peaks at the edges, indicating a good decision amongst the data points.

Logistic Regression For logistic regression we see a peak at 0.45, which stays consistent even if we make predictions further in the future. The peak in the middle indicates most observations lie near the decision boundary and hence a linear decision boundary might not be able to capture the complexity of our data.
QDA provides different graphs for predictions in the near future (one month) and for distant future (three months and six months). For one month prior data we observe a flat graph in the middle and peaks at the edges, which indicates a good decision boundary. However, in the case of three month prior and six month prior, we observe peaks skewed towards the left, which indicates the inability of QDA to model predictions further into the future.
Neural Network provides us with a similar distinction as QDA. It also has different probability distributions for near future (one month) and distant future (three month and six month) predictions. However, in the case of neural network, we observe for one month prior that the graph has a peak on the left extreme and then is flat throughout, which indicates that it would be extremely precise in the classifying clients who aren't going to enroll. In the case of predictions in the distant future, the graph is flat throughout showing a good spread in the distributions, however it does not indicate an effective decision boundary, as there are no peaks on the edges.
Random Forest provides us the the best distributions compared to all the other models. If we look at one month prior and three month prior predictions, we can clearly see peaks at the edges and a dip in the middle. This indicates a good decision boundary at least for predictions within the next 3 months. However if we look at predictions made from six month prior information we see a graph with a flat peak in the middle, which indicates a non-effective decision boundary, indicating to us that as we make prediction further in the future, we would require more complex decision boundaries to model the data.

E. Results

We report the precision of enrollment for each model on the development set in Table **1**. Each of these values is calculated as $\frac{TP}{TP+FP}$, using the cutoff that maximizes the weighted F_1. To exemplify the increase in performance by using the weighted F_1, we also include the precision calculated using the regular, unweighted F_1 score

From Table 1, we observe that there is a clear increase in precision of enrollment by using weighted F1 score as a performance measure as compared to unweighted F1. This motivated us to choose weighted F1 as our performance criteria, since the cutoff it selects for classification sacrifices recall to improve precision. Looking at precision provided by the different with varying complexity we see that Logistic Regression and QDA provide a lower precision but stay consistent through different time periods. This indicates that they generalize over the data and are not able to model the data correctly. However in the case of Neural network there is a sharp decrease in accuracy as we make prediction further in the future, which indicates some temporal features or interactions between predictors which were not being captured by more complex methods. Random forest finds a good balance between the simpler methods such as logistic and QDA and complex methods such as neural network. It provided us with high precision while only having a small drop in precision as we make predictions further in the future. With this we can conclude that while, neural network is the most effective model for predictions in the near future, random forest is effective for predictions in the distant future.

VI. Models Limitations and failure Points

Logistic Regression: Although logistic regression is a linear decision model, it performed better than QDA, which allows more flexibility. This was because of the regularization parameter which was introduced into the model. It was also observed that logistic regression has a sharp peak for its probability distribution graph (at 0.5).

Precision Maximizing Weighted F_1 Precision Maximizing F_1		Logistic	QDA	Neural Network	Random Forest
Precision of Enrollment	1 Month Prior	0.653 0.539	0.559 0.486	0.766 0.640	0.733 0.643
	3 Months Prior	0.633 0.499	0.535 0.469	0.688 0.580	0.698 0.589
	6 Months Prior	0.610 0.498	0.552 0.472	0.648 0.524	0.686 0.551

Table 1. Precision of enrollment for each model and time frame. In order to show the increase in precision by the weighted F_1, both precisions from maximizing F_1, and the weighted F_1 are shown

This implies that it doesn’t have a good spread of the probabilities and most of its decisions don’t have a good separation.

QDA: This model did not provide us with good results, however it was able to capture the outliers in the data which were in some cases not captured by even the neural network.

Neural Network: A two layer neural network with 100 and 80 neurons respectively was used. In the probability distribution graph we noticed no peaks, which indicates that the probabilities are distributed evenly, which indicates a good decision boundary. However, the neural network suffers from a high training time and computational complexity. The data suffers from a low signal to noise ratio. We believe the underlying signal can be a modelled by a fairly simple function but due to the high level of noise in the data, even complex functions are not being able to provide high accuracy. Even when neural networks with larger number of hidden layers were modelled on the data, they provided almost the same accuracy as simpler neural networks, hence confirming our hypothesis.

Random Forest: This model provides us with the best result out of the four models. The precision suffers as we make predictions further in the future, but it doesn't have a significant drop as we see in neural networks. Random forest was able to learn interactions between predictors and complex decision boundary even for predictions in the future, where the other models failed.

Based on our results from the tests on the validation set, for each time frame we select the model with the highest precision: NN for 1 month prior data, and Random Forest for 3 and 6 months prior data.

We report the precision on the test set:

Model	Neural Network [2 Hidden layers (100,80)]	Random Forest [Depth: 14]	Random Forest [Depth: 12]
Time Frame	1 Month Prior Data	3 Months Prior Data	6 Months Prior Data
Precision	0.771	0.701	0.686

VII. Conclusions

We observed a clear distinction in terms of predictors which were important for predictions made at different time frames (one month, three month and six month into the future). For predictions in the near future (one month/three month), predictors indicating customer engagement such as number of visits to the bank, number of credit card uses and number of complaints along with a customers communication channels are important predictors. However when we are making predictions further into the future (six months), the predictors indicating the revenue a customer generates for the bank such as deposit balance, average amount per card transaction and the customer segment become important predictors. We saw a shift in the importance of customer engagement as a predictor as we make predictions further in the future because customer engagement in the past might not be a good representation of their relationship with the bank in the present, whereas the revenue they generate is highly likely to stay similar in a six months period.

We also found that customers with only investment accounts have a very low probability of joining the rewards program. This follows from the fact that the rewards program benefits are mainly for customers with banking accounts.

It was also observed that having a large number of preferred communication channel with the bank was not a strong positive predictor, but have a no communication channel was a strong negative predictor.

VIII. Future Scope

The goal of the project was making accurate predictions for the data, however, we would like to be able to interpret the models for a better understanding of the data. This was missing for the more complex models such as neural network. In the future it would be beneficial to implement post processing techniques such as LIME to find out, which predictors/combination of predictors lead to a good decision boundary as provided by the neural network.

Since we are modelling human decision have temporal data of the customers which indicates the sequence of actions a customer took leading up to joining the rewards program would be beneficial for modelling the data.

We also believe ensemble methods such as a GBM's would be worthwhile to try on the data due to its low signal to noise ratio.

References

[1] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373–1379 https://www.ncbi.nlm.nih.gov/pubmed/8970487/

Acknowledgements

Data provided by Bank of America Consumer Analytics Department

IX. Statement of Contributions

Authors: Christopher Botica, Betül Çam, Nalin Gupta, Anushka Tak

Each Team member had inputs and overlaps in each of the areas of our project. We worked together on the majority of the project (as a whole). However, each of us took responsibility of completing several specifics:

Christopher Botica- A priori models and distribution comparisons, Performance Criteria, Data Modelling

Nalin Gupta - Data Modelling, Model Comparison and Conclusions, Future Scope

Betül Çam - Sourcing data, including legal, risk compliance approvals. Initial variable selection.

Anushka Tak - Data Exploration and Analysis using various Visualization techniques, Trend Analysis