Wells Fargo Analytical Challenge

Lynell Amanna, Proxima Das Mohapatra, Shirish Dhar

Master’s candidates, UC Berkeley School of Information

**September 27, 2016**

# Goals

1. **What drives growth in balance between month 0 and month 12?**
2. **What demographic types, if any, are more likely to increase (or reduce) their number of accounts and/or balance between month 0 and month 12?**
3. **What types of accounts, customer interactions, customer events, or Wells Fargo outreach, are more correlated with account and/or balance change?**

# 

# 

# 

# Deliverable 1

# Results

Our exploration of the data-set culminated in the following results:

1. Best predictors of growth between month 0 and month 12 are seen when the customers are divided by demographic type. Each demographic class has different indicators of growth which help predict the change in balance better. We tried to come up with indicators that work well for all demographic groups, but the accuracy of such predictions was low. This led us to believe that dividing the overly broad dataset into smaller chunks of demographics would be the most optimised approach.
2. This question was better answered by visual exploration of the data. We found that demographic combination types 1-1, 1-2 and 0-5(Customer Demographic ai - Customer Demographic aii ) are most likely to **increase**  their account balance between the 12 months, and the demographic types 1-5 and 3-4 (Customer Demographic ai - Customer Demographic aii) are most likely to **decrease** their account balance between the 12 months.
3. The Pearson coefficients were derived for each feature - output combination and those with the highest positive or negative correlations for each demographic combination were used to train the model. The Pearson coefficient method is a lesser-known but effective method to find hidden trends between input features and an output variable. This method helped make our approach much more focussed towards a filtered subset of features that have a deeper relation with the output.

# Data Wrangling

## Pre-processing data

The account balance data for each month was converted into a category - increase (boolean value 1) or decrease (boolean value 0). For this, we compared the account balance difference between the current month and the next month. This was done so that we don’t make any assumptions about month 0 (whose data wasn’t available). This also resulted in no balance delta for month 12 (since we don’t have month 13 data). Hence we excluded month 12 rows from the original dataset. We decided to convert our targets to coarse step increase or decrease because the questions asked us to find factors for increase/decrease, and not the amount of increase/decrease in account balance change.

Another approach that was initially considered was, to take the net difference on any customer account balance between month 1 and 12. But this would result in the loss of information between consecutive months.

## Splitting up data

Since each customer had 12 rows associated with their customer number, we decided to split up the dataset into training, development and test sets based on customer numbers rather than on individual rows. This ensured that all the rows for a customer remained in one set, and their account balance transitions from one month to the next could be captured and learned by the model. i.e. if cust\_num 1 appeared in the training dataset, then month 1 - month 11 rows were in the training dataset (as noted before, we removed month 12 from the dataset).

# Model methodologies

1. Since the data anonymization stopped us from using any intuition about features in the data, we decided to use correlation to determine the most promising features in the dataset (with respect to balance change).
2. The features selected in the above step that had the most statistically significant correlation to our target were used to train various machine learning models. We initially considered using recurrent neural networks, SVM and random forests.

# Final models used

The highest accuracy was achieved with Random Forests with a preliminary feature selection using Pearson coefficient.

The correlation coefficient for each feature in the data set was calculated, and those with the highest positive(+1) or negative correlation(-1) were selected to train the model. The accuracy was used as a metric to rank the success of the model.

Caveats: Using the pearson coefficient to rank the correlation of each feature with the output, assumes that each feature is independent of the other, which may not be the case. We overcame this by testing out combinations of features to train the model. The use of this holistic two-pronged approach helped us, firstly, choose the best features that individually contribute the most to the output, as well as the best combination of features that cumulatively affect the output the most.

With Random Forests, the model built on the training set, is robust but in many cases could overfit the data from the training set, reducing the likelihood of classifying a random data point (not encountered in the training set) correctly.

# 

# Deliverable 3

**i. What drives growth in balance between month 0 and month 12?**

The initial analysis was conducted on the entire training set, with the selected features that showed the highest positive or negative correlation. However, despite using a robust model(all three models were tested - Random Forests, RNN and SVM), the accuracy obtained was considerably low. (~55%)

On further analysis, we found that dividing the data into subsets for each demographic combination resulted in a significantly improved accuracy. (~ 82%)

From this behavior we understand that, each demographic type exhibits behavioral patterns unique to its own category. The delta in account balance between two consecutive months can be mined for patterns, but results are more promising within a demographic category. A deeper understanding of the data itself could provide more insight into answering why each demographic type combination had such a significant impact on the predictive power of the designed model.

Due to the above phenomenon, it is also seen that various feature combinations impact the predictive power of the model differently for each demographic category type. The following features are those that are common across most demographic types and have the highest correlation and predictive power:

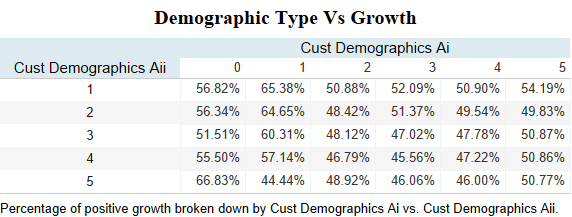
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| typeA\_ct | wf\_outreach\_flag\_chan\_i | wf\_outreach\_flag\_chan\_iv | cust\_outreach\_ai | typeB\_ct | cust\_outreach\_av | typeF\_flag |

Although the demographic type of the customer is a major factor in predicting ‘growth’, other drivers such as the type of account that a customer holds, the customer outreach channel type, the number of outreach can also be considered as key features that can help predict growth. (In this case, we have defined growth in terms of positive delta in account balances between consecutive months).

The model thus created can be used to access the delta for a customer for the coming month with accuracy as high as 82% within a demographic category.

**ii. What demographic types, if any, are more likely to increase (or reduce) their number of accounts and/or balance between month 0 and month 12?**

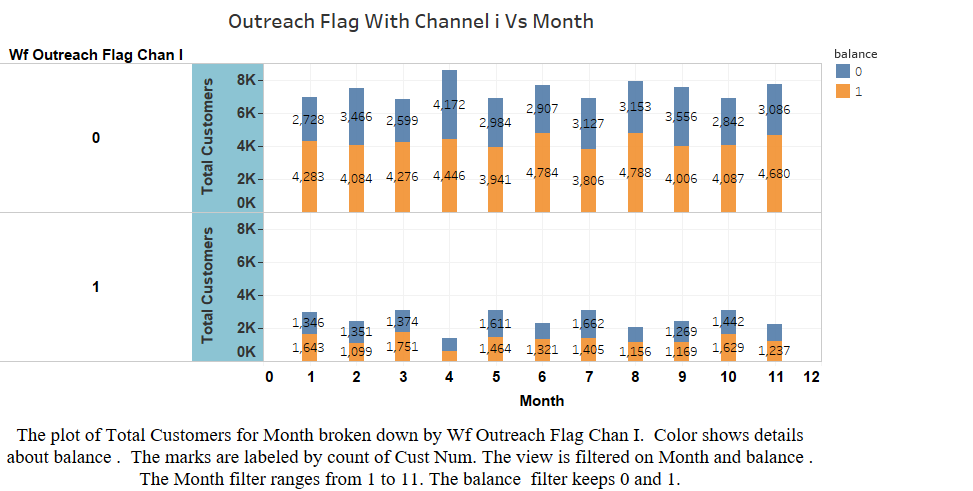
The graph below displays the percentage of positive delta in account balance over 12 months for each customer demographic category. Another way to look at this visual is, the probability that a customer that belongs to Customer Demographics Ai=1 and Customer Demographics Aii=1 would have positive growth in the next month is ~65%. It must be noted however, that the size of the delta has not been taken into consideration for this analysis and could be a possible area for further exploration.

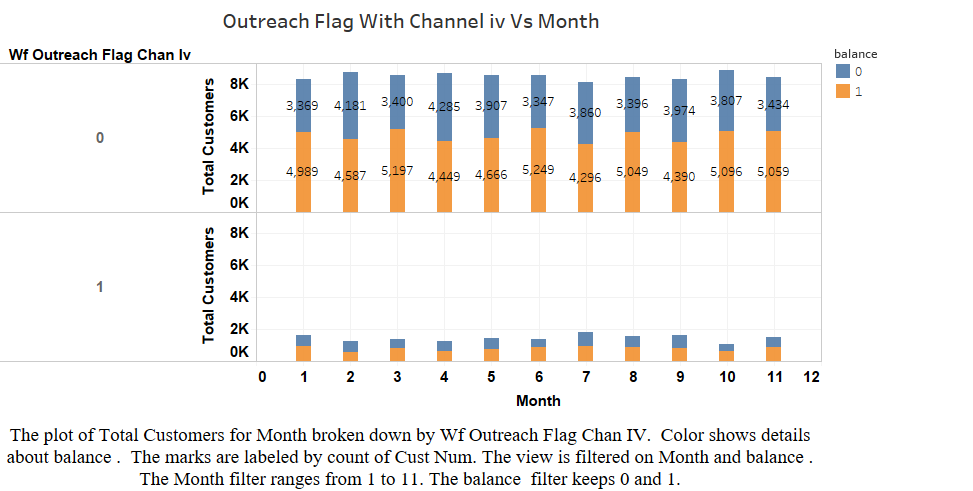


**iii. What types of accounts, customer interactions, customer events, or Wells Fargo outreach, are more correlated with account and/or balance change?**

According to our analysis it was found that the outreach flag for channel i and the outreach flag for channel iv resulted in boosting the accuracy of the model by ~3%. The following graphs display the number of customers in each month that had a positive or a negative change in balance, with the presence(1) or absence(0) of the flag for market outreach using channel i and channel iv.

In the graph below balance 0 indicates a negative change/no change in the account balance for the consecutive month and 1 indicates a positive change.





**Conclusion**

There exist clear indicators of account balance (growth) predictors, such as the demographic type, the method of outreach(which channels were used to reach out to the customer). As the quantity of the data increases, the model’s ability to accurately make a prediction increases. For the purpose of our analysis we define growth in terms of account balances and the change each month for a customer, the analysis and findings could be different if growth is measured in terms of number of accounts held by a customer.