Customer Retention Using Analytics At Bank of America

ISYE 6501 - INTRODUCTION TO ANALYTICS MODELING

COURSE PROJECT

Table of Contents

Summary:

For this course project, I try to implement an analytics-based solution for ***Customer Retention at Bank Of America***. This was one of the problems Bank Of America had tried to solve using analytics.

More info in the link below: https://www.informs.org/Impact/O.R.-Analytics-Success-Stories/Industry-Profiles/Bank-of-America

Financial institutions like Bank of America rely heavily on income earned from customer interest payments and purchases, along with customer investments in banking products like credit cards, mortgage loans, auto loans etc. Due to the wide array of banking products in the market today, it is easy to lose a customer just because of the simple fact that they found a cheaper product in a competitor bank. Hence the bank needs to not only make sure its products are competitive enough, but also to ensure that existing customers are given a stimulus to stay with the bank itself. This makes customer-retention one of the highest priorities for the bank in ensuring its financial growth and development. An analytics based-solution would help the bank greatly in targeting those customers who have a potential to leave the bank, along with ensuring that the bank comes up with market-leading products.

An analytics-based solution to ensuring customer retention at Bank Of America would look to solve the below questions

- What are the key customer portfolios the bank has?
- Who are the customers that are more likely to close accounts at the bank in the near future (next month, quarter)?
- Which type of customers should the bank try to retain the most?
- What methods of customer retentions have worked in the past?
- Given a budget, What pricing adjustments can be made to motivate the customer to stay with the bank?
- Given a budget, What new offers (new credit card? More cash-back?) can be extended to the customers to retain them?
- What new variation of products can be created so that the Bank remains the best-in-class for these product?

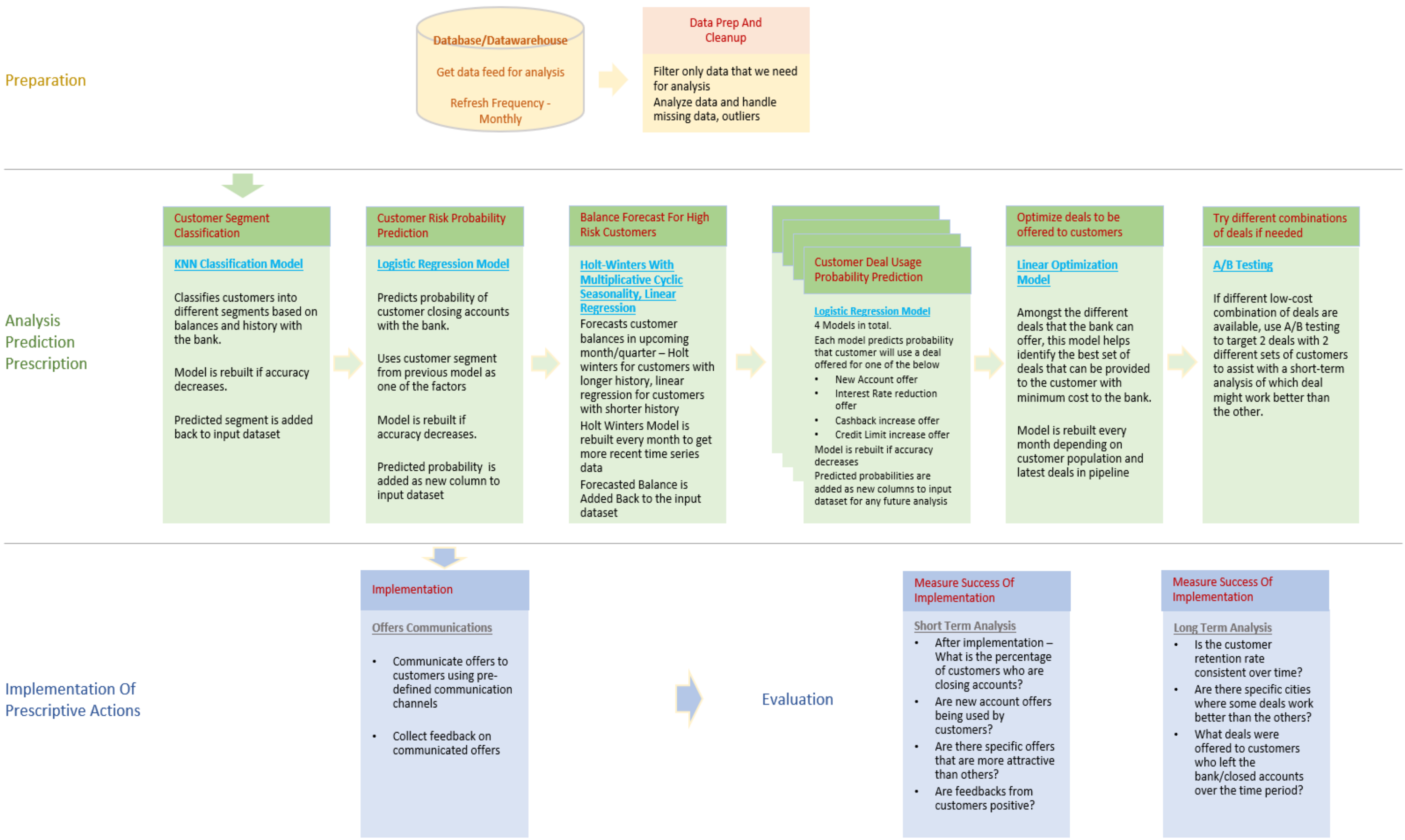
My submission below would help answer these questions using a combination of analytical modeling, forecasting and optimization.

Solution Architecture

Below is how the analytics-based solution would look like at a high-level. The analytical solution broadly comprises of the below steps most of which happen sequentially, but some of the models would also be run in parallel

1. Data-prep and cleanup
2. KNN Classification model to classify customer in risk-based segments
3. Python or R based program to generate a weighted satisfaction score for customers based on prior communications and feedback
4. Logistic regression model to predict probability of customers at risk of closing accounts with the bank
5. Holt-Winters exponentials smoothing model with multiplicative seasonality to predict customer balances in upcoming month and/or quarter
6. Logistic regression models to identify the probability that a customer will use a newer product or deal
7. Linear optimization model to optimize the best deals that can be offered to customers at lowest cost to the bank
8. A/B Testing of deals to finalize best set of deals that can be used in future
9. Taking Prescriptive actions based on recommendations from the analytical process
10. Evaluating the effectiveness of these analytical prescriptions over time and adjusting models/optimization variables as needed.

The illustration in the next page gives a birds-eye view of what the analytical process would look like. Further details on each of the steps are in the following pages. *(Please zoom in if the text is blurry or not clearly visible.)*


Preparation
Database/Datawarehouse
Get data feed for analysis
Refresh Frequency - Monthly
Data Prep And Cleanup
Filter only data that we need for analysis
Analyze data and handle missing data, outliers
Analysis
Prediction
Prescription
Customer Segment Classification
KNN Classification Model
Classifies customers into different segments based on balances and history with the bank.
Model is rebuilt if accuracy decreases.
Predicted segment is added back to input dataset
Customer Risk Probability Prediction
Logistic Regression Model
Predicts probability of customer closing accounts with the bank.
Uses customer segment from previous model as one of the factors
Model is rebuilt if accuracy decreases.
Predicted probability is added as new column to input dataset
Balance Forecast For High Risk Customers
Holt-Winters With Multiplicative Cyclic Seasonality, Linear Regression
Forecasts customer balances in upcoming month/quarter – Holt winters for customers with longer history, linear regression for customers with shorter history
Holt Winters Model is rebuilt every month to get more recent time series data
Forecasted Balance is Added Back to the input dataset
Customer Deal Usage Probability Prediction
Logistic Regression Model
4 Models in total.
Each model predicts probability that customer will use a deal offered for one of the below
• New Account offer
• Interest Rate reduction offer
• Cashback increase offer
• Credit Limit increase offer
Model is rebuilt if accuracy decreases
Predicted probabilities are added as new columns to input dataset for any future analysis
Optimize deals to be offered to customers
Linear Optimization Model
Amongst the different deals that the bank can offer, this model helps identify the best set of deals that can be provided to the customer with minimum cost to the bank.
Model is rebuilt every month depending on customer population and latest deals in pipeline
Try different combinations of deals if needed
A/B Testing
If different low-cost combination of deals are available, use A/B testing to target 2 deals with 2 different sets of customers to assist with a short-term analysis of which deal might work better than the other.
Implementation Of Prescriptive Actions
Implementation
Offers Communications
• Communicate offers to customers using pre-defined communication channels
• Collect feedback on communicated offers
Evaluation
Measure Success Of Implementation
Short Term Analysis
• After implementation – What is the percentage of customers who are closing accounts?
• Are new account offers being used by customers?
• Are there specific offers that are more attractive than others?
• Are feedbacks from customers positive?
Measure Success Of Implementation
Long Term Analysis
• Is the customer retention rate consistent over time?
• Are there specific cities where some deals work better than the others?
• What deals were offered to customers who left the bank/closed accounts over the time period?

Identifying Customer portfolios

Banks usually have customers spread across multiple buckets depending on the type of accounts and utilization of the bank's services. I have broadly classified the bank's customer portfolio into 3 buckets.

- *Individual Customers*
  - Customers with one or more accounts for personal use
  - Makes up the majority of customer population
  - Usually have more products catered towards these customers (checking account, savings acct, credit cards, auto loans etc)
  - Average balances might be low, but could cause a big impact if more customers leave the bank
  - More prone to moving to better products in the market
- *Small Business Customers*
  - Customers with more accounts, catered towards maintaining a small business
  - Makes up the second highest set of customer population
  - Average balances are higher, a small number of customers leaving the bank could have more impact
  - Less prone to moving to different products/banks
- *Institutional Customers*
  - Large organizations, corporations who have multi-million dollar portfolios with the bank
  - Even losing one customer can cause significant financial impact
  - Least prone to moving to different banks due to diligence these customers perform before deciding to be attached to Bank of America

Each of these customer portfolios have their own set of parameters within with the bank operates and comes up with products (eg: products intended for personal customers might not be equally efficient for small business customers or institutional customers). Hence, we cannot have a one-size-fits-all set of analytical approach that would satisfy all 3 classes of customers.

For the purposes of this project, I focus on how analytics can help retain "Individual Customers" at the bank. These customers would make up the majority of the bank's customer population. These customers also usually produce the biggest churn in account closures, simply due to the fact that there are more products out in the market that cater to these individual customers. These customers also tend to have more variability in the products they own (checking accounts, savings accounts, credit cards, auto loans, home loans etc). Hence, retaining these customers can benefit the bank the most due to the large volume of accounts and underlying balances.

Data Prep – Data collection and refresh

Customer Data

This is the core customer data, and includes summarized transactions and balances, along with all other attributes (age, married, homeowner, number of accounts, debts, number of loans etc) for the customer. This data set will be used heavily in all the models that are involved in the analytics. For this submission, I would be considering ONLY customers who have had ATLEAST 6 months of relationship with the bank for analysis purposes. It also would be too costly and difficult to come up with useful estimates based on a shorter history of data for customers who have not been with the bank for at least 6 months.

- **Data Source** – Bank's Data Warehouse/Operational Databases
- **Data Refresh Frequency** – Once at start of every month
- **Data Population** – All summarized customer activity until the end of the last month

A data feed from the bank's data warehouse or other databases would help us to get the customer data that we would need for analysis. The data would be refreshed at the beginning of each month, with summarized data until the end of the prior month. This will give us a clean cutoff of the customer portfolio and would account for any mid-month activities done by customers – Customers usually tend to close out accounts or rebalance their assets towards end of the month.

Customer Service Data

This data will be used to assign a customer satisfaction score to each customer, and will be used as a factor in some of the models used.

- **Data Source** – Bank's Customer Care Databases
- **Data Refresh Frequency** – Once at start of every month
- **Data Population** – All communications customer has made with the bank until the end of the last month

I am making the assumption that the customers' interactions with customer service are already available in the bank's database – these includes data from customer phone calls, complaints, feedback emails, and any social media related contacts where the customer has **AGREED** to share his account details and communications with the bank.

Banking Products Data

This data will be used to perform optimization on the current products to come up with newer products with least cost to the bank. This data will also help to identify which set of deals we can offer to customers on top of these products (eg: more credit limit on credit cards, more travel points on travel cards etc). I am assuming that the bank's pricing models have already calculated key metrics associated with each of the product types (like cost incurred to the bank for each product, what states some of the products can be used in etc)

- **Data Source** – Bank's Product Databases and Pricing models
- **Data Refresh Frequency** – Once at start of every month
- **Data Population** – All products currently live at the bank. We might not need historical data since those products are no longer active in the market.

Data Cleanup

Once the data has been made available, we will perform the below activities on the datasets to ensure we are having good data for analysis as well as predictions.

- **Handling Missing Data**
    - Although most of the data should be available for use, I would assume that there would be some missing values in the data fields. Without knowing the scope of missing data, it would be tough to have a one-size-fits-all approach for missing data. However, we could create an automated process or report that would scan the data, and then flag each factor which has missing data, and a recommended approach to data handling (Imputation vs Categorical variable addition vs Removing data point). This report/output can then be used to treat the missing values in the data before we train or predict using the models.
- **Outlier Removal**
    - This step can happen just before we build any model, and only for the factors involved in training/testing/validating the model.

Identifying Risk Factors For Customers Leaving The Bank

There are a few indicators that can be used to flag customers who might be at risk of exiting the bank – We can use these indicators to predict probability of customers who might be at risk of leaving the bank or closing one or more accounts with the bank.

1. ***Closure of accounts*** -> This is an important indicator and will play a big role in identifying customers who are at risk of leaving the bank. Customers who have closed one or more accounts recently are at a bigger risk of closing out their portfolios than customers who have not.
2. ***Number of recent customer care interactions and weighted score*** -> Usually customers close accounts if they have had bad experience with a bank product or customer service. A higher number of recent customer care interactions with a lower customer satisfaction weighted score could be used as a factor in predicting if the customer will close his account or not.
3. ***Unusual reduction in balances*** in the recent past. This is not a direct indicator as there could be some seasonality associated with reduced balances. However, a sudden reduction in balance could also be an indicator that customers are moving funds away from the bank to a different place.

While account status and balances can be obtained from the customer data feed, for customer interactions weighted score determination, I will use a python or R-based program that will scan all customer interactions with the bank, and assign a weighted score for each of the interactions based on customer sentiment. I will use a pre-determined weighted score list to assign scores to words (score being the response variable) - Typically, positive words are given a score of 1, negative words given a score of -1 and neutral words given a score of 0. We can then use the program to predict sentiment scores for each of the customer interactions with the bank and aggregate them up to come up with an overall weighted score for each customer. I use a threshold of past 6 months for the sentiment score aggregations so that older communications do not come in the way of misinterpreting recent sentiments of the customer.

Segmenting Customers Into Risk-Based Segments

I use a KNN classification model to classify customers into the below 4 different segments, based on their total balances, debts, history with the bank and number of accounts. Below are the 4 ways in which I choose to classify the customers, with a numeric score assigned to each customer segment.

1 -> Customers have higher balance, more number of accounts, and have lower number of years with the bank
2 -> Customers have higher balance, more number of accounts, and have higher number of years with the bank
3 -> Customers have lower balance, and lower number of years with the bank
4 -> Customers have lower balance, and higher number of years with the bank

Typically, a customer with more balances, and a longer history with the bank is more likely to stay with the bank, whereas a customer with lower history has a higher risk of leaving the bank. Hence, out of the classified segments, we would be more interested in customers who fall under either category = 1 or category = 3 as they are more prone to terminating their relationships with the bank. We will add the predicted segment classification as a column to the input dataset(Customer_Risk_Segment). This would help us to use this as a factor in some of the downstream models, and to perform any segment-oriented analysis needed.

Customer Segment Classification Model

```
Given
       Number of accounts
       Total Balances (all accounts)
       Total Debts (all accounts – credit card balances, auto loan balances etc)
       Number of years with the bank
Use
       KNN Classification
To
       Classify the customers into 4 different risk-based segments – 1,2,3,4
```

Building the model

- While building the model, we will use a pre-computed dataset which has customers that fit the criteria for each of these 4 segments.

Validation of Model

Cross-validation of the model will be done to ensure we have the highest prediction rate from the model. Since the volume of customers is large, we can have a bigger precomputed dataset with the 4 customer segments and use a higher k-fold validation like 20-fold cross validation to ensure the model's predictions are within acceptable limits.

Rebuilding of Model in subsequent months?

Once the model has been trained and validated, we **need not rebuild the model each month** when the process runs. Instead, each month, we can use a subset of the incoming data to test the model, and ensure its accuracy is still as good as it was the previous months. **Only if a sudden drop in accuracy is observed, we need to retrain the model**. One scenario that could happen to cause a drop in accuracy would be unexpected financial crises like the current COVID-induced financial impacts to customers, which would cause the average balances of customers to go down in general. In such cases the thresholds for these 4 segments would need to be re-calibrated so that we do not mis-classify customers into the wrong segments. Going forward, we can use the updated model until the time when the accuracy drops again, post which we rebuild the models with more recent balance trends at that time.

Identifying Customers With Higher Probability Of Closing Accounts

We will use a logistic regression model to predict the risk of each customer terminating his relations with the bank. One of the factors that would be used while training the model is the 'Customer_Risk_Segment' from the previous model. Note that I do not use the factors -> Total balances, Number of accounts, Total debts, Number of years with the bank, since these factors have already been used to determine the Customer_Risk_Segment. Using these factors again poses a risk of highly correlated factors on the regression model. I also use the customer satisfaction weighted score that was predicted using NLP/Deep Learning as a factor for this model.

To start with, a relatively lower threshold of 0.6 or higher probability can be used to target any customer who has more than 60% chance of leaving the bank. We can add the probability as a new column on the input dataset as well so that we can perform any additional analysis that might be needed to further divide these high-risk customers into most-probable-amongst-high-risk-customers, least-probable-amongst-high-risk-customers etc. This would also give us the flexibility to split the dataset into different smaller datasets based on Customer segments, and perform analysis on them separately if needed.

Customer Risk Identification Model

```
Given
        # Of Accounts closed in the past 6 months
        Customers Satisfaction Weighted Score
        Credit score
        Income
        Homeowner vs Not
        Married
        Number of Children
        # of Payment defaults in past 6 months
        City
        Occupation Type
        Customer_Risk_Segment
Use
        Logistic Regression
To
        Predict the probability that a customer is at risk of closing an account or not
```

Building the model

- While building the model, we will need historical data with a good mix of customers who are still with the bank, and those who are no longer with the bank – A dataset from the last 12 months of customer data should help to get a more peek into what the customer portfolios looked like for customers who left the bank.
- The factor "# of accounts closed in the past 6 months" need to be refactored for the customers who have left the bank or closed all accounts so that we get the number of accounts closed within the past 6 months of the last month they were with the bank. (Eg: if current month is July, but the customer terminated all his accounts with the bank in May, then we need to get the number of accounts he closed from 6 months prior to May)

Validation of Model

Cross-validation of the model will be done to ensure we have the highest prediction rate from the model. Since the volume of customers is large, we can have a bigger precomputed dataset with the 4 customer segments and use a higher k-fold validation like 20-fold cross validation to ensure the model's predictions are within acceptable limits.

Rebuilding of Model in subsequent months?

Once the model has been trained and validated, we **need not rebuild the model each month** when the process runs. Instead, each month, we can use a subset of the incoming data to test the model, and ensure its accuracy is still as good as it was the previous months. **Only if a sudden drop in accuracy is observed, we need to retrain the model**. As with the classification model, unusual drops in accuracy would be expected during abnormal financial situations, during which we might have to have a more stricter threshold to target customers who are the most probable to terminate their relations with the bank.

Forecasting Cost To Bank For Closed Accounts

Once we have identified the population of customers who are at a higher risk of closing accounts with the bank, we can predict what their balances will be for the next month (or quarter) so that we know what is the quanta of business the bank would lose in the upcoming month. Customer balances and debts are also a factor of time and have a seasonality component to it – typically customers tend to spend more during the holiday season and hence can have lesser balances/more debt on credit cards. March/April is a time when customers might have slightly elevated balances on account of tax returns, receiving bonuses etc. Hence, I will use an exponential smoothing model like the Holt Winters model with multiplicative seasonality to smoothen out some of these interim spikes and troughs in customer balances with the bank. We can potentially have 2 models here – one which would predict the total balances for the next month, and another one which would predict the total balances for the next quarter based on historic monthly and quarterly balances. I have used the Monthly balances for further analysis in this submission.

For customers who are relatively newer to the bank, we might not have historical data, and so I would use a linear regression-based model to predict the balances for these customers. Once we predict the balances from both these models, we will add the predicted balances as a new column to the input dataset so that we can perform any additional analysis needed (split by customer segment, type etc or build operational reports to management or financial planning teams showing what the loss to the bank would be for each customer segment in the upcoming month)

Customer Balance Forecast Model -> Customers with longer history

```
Given
        Month and date
        Total net balance with the bank (Total Balances minus Total debt)
Use
        Holt-Winters with multiplicative seasonality
To
        Forecast Total balances for next month
```

Customer Balance Forecast Model -> Customers with shorter history with the bank

```
Given
        Number of accounts
        Average 6-month balance
        Average 6-month debts (mortgage payments, auto loan payments, credit card payments etc)
        Married
        Number of kids
        Homeowner or not
        # of Payment defaults in past 6 months
        Income
        City
        Occupation Type
        Number of loans with the bank
Use
        Linear Regression
To
        Predict Total balances for next month
```

Building the model

Holt-Winters

While building the model, we will need historical data from at least the past 24 months to forecast what the future customer balances would look like.

Linear Regression

While building the model, we would need to have a collection of customers who have been with the bank for less than 6 months, and also customers who have been with the bank longer than 6 months, with predicted value of their 7th month balance (as an example) in the response variable. This would help us to train the model to real-scenarios and not just cater to customers who might have a shorter set of data and accounts and thus seem as though they might not have potential for growing their balances at the bank.

Validation of Models

For both the models, we should be able to perform cross-validation of the model will be done to ensure we have the highest prediction rate from the model. For the Holt-winters model, we can try different combinations of alpha, beta and gamma to identify the model that best predicts the balances. Since the volume of customers is large, we can have a bigger precomputed dataset and use a higher k-fold validation like 20-fold cross validation to ensure the model's predictions are within acceptable limits.

Rebuilding of Model in subsequent months?

Holt-Winters

The Holt-winters model will need to be **re-trained/rebuilt each month** so that we do not let any prior seasonality affect future predictions. Also, data trained on past trends might not be a good fit for more-recent trends in the balances within the bank's customer portfolio.

Linear Regression

Once the model has been trained and validated, we **need not rebuild the model each month** when the process runs. Instead, each month, we can use a subset of the incoming data to test the model, and ensure its accuracy is still as good as it was the previous months. **Only if a sudden drop in accuracy is observed, we need to retrain the model**.

Targeted Deals

With the forecasted balances in place, we would now know the below

- xi -> number of customers who have a high risk of closing an account with the bank the next period (month)
- bi -> total balance per customer if he retains his accounts with the bank
- ∑xi * bi -> Total balances bank might lose if all customers close their accounts with the bank

One of the ways the bank can retain customers is by having targeted deals/offers custom-made to these customers to motivate them to stay with the bank. Examples of these would be a higher cash-back offer on certain purchases, increase in credit limit on credit card, reducing the interest on credit card balances or insurance products etc. Each of these offers come at a price to the bank, and so we should weigh the benefit of offering these to the customers vs the cost to the bank when these offers are used to their maximum potential.

Targeted Deals Usage Pattern Prediction

We can further segment the high-risk customers by predicting the number of targeted products/offers that a customer would really end up using. For example, if a customer has not used any of his/her credit card benefits like cash back offers or restaurant discounts in the past, the cost to the bank in offering these products to these customers would be lower since the customers might not end up using these offers, but the prospect of having these offers could be motivation enough to make them stay with the bank. Identifying the probability whether a customer would use a targeted deal or not would then help the bank to decide if a customer should be given ONLY one offer (maybe a credit card balance increase) or a COMBINATION of offers (an interest rate reduction, more travel rewards etc).

I am assuming the bank has historically provided 4 different types of offers to customers

- Interest Reduction Offers
- Cash Back Increase offers
- Credit Card balance increase offers

- New Account offers

We can create 4 different logistic regression models to predict the probability that a customer would use each of these offers. From the outcome of these models, we can get to know how the usage of deals for the high risk customers for this month are spread across the different deal categories. We will add these 4 predicted probabilities as columns to the input dataset for any further analysis needed.

Customer Offers Usage Pattern Identification Model – Interest reduction offers

```
Given
        Customer_Risk_Segment
        # Of Accounts with the bank
        Credit score
        Income
        Homeowner vs Not
        Married
        Number of Children
        City
        # of Interest-Reduction Products offered in the past
Use
        Logistic Regression
To
        Predict the probability that the customer would use this offer
```

Customer Offers Usage Pattern Identification Model – Cash Back increase offers

```
Given
        Customer_Risk_Segment
        # Of Accounts
        Credit score
        Income
        Homeowner vs Not
        Married
        Number of Children
        City
        # of Cash-back increase offers extended in the past
Use
        Logistic Regression
To
        Predict the probability that the customer would use this offer
```

Customer Offers Usage Pattern Identification Model – Credit increase offers

```
Given
        Customer_Risk_Segment
        # Of Accounts
        Credit score
        Income
        Homeowner vs Not
        Married
        Number of Children
        City
        # of Credit increase offers extended in the past
```

```
Use
        Logistic Regression
To
        Predict the probability that the customer would use this offer
```

Customer Offers Usage Pattern Identification Model – New Account Opening Offers

```
Given
        Customer_Risk_Segment
        # Of Accounts
        Credit score
        Income
        Homeowner vs Not
        Married
        Number of Children
        City
        # of New Account Products offered in the past
Use
        Logistic Regression
To
        Predict the probability that the customer would use this offer
```

Building the model

We will use historical deal usage information to train the model, and ensure its accuracy is at an acceptable range. Considering that we might not have the deal usage information for ALL customers, we will need to pick a training dataset that has a good mix of customers who have used such deals in the past along with customers who have not used such deals in the past, so that the model can be fitted appropriately.

Validation of Models

We should be able to perform cross-validation of the model will be done to ensure we have the highest prediction rate from the model. Since the volume of customers is not too large, we can have a slightly smaller precomputed dataset with the data and use a mid-level k-fold validation like 10-fold cross validation to ensure the model's predictions are within acceptable limits.

Rebuilding of Model in subsequent months?

Once the model has been trained and validated, we **need not rebuild the model each month** when the process runs. Instead, each month, we can use a subset of the incoming data to test the model, and ensure its accuracy is still as good as it was the previous months. **Only if a sudden drop in accuracy is observed, we need to retrain/rebuild the model**.

Targeted Deals Optimization Models

Since we now know the probability that each of the customers would use one of the 4 type of deals, we can get the total number of customers who are likely to use each of the 4 deal types. At this point, we would have the below information handy

- Cij -> Cost to the bank for each of the offers i within each offer type j (j = Interest Reduction Offers, Cash Back Increase offers, Credit Card balance increase offers, New Account offers). I am assuming that the cost to the bank for each offer type has already been pre-calculated by pricing models already available at the bank. We will just reuse the information from the pricing and product models.

- Nj -> Maximum of customer for whom we can share an offer for one of the 4 offer types j (we can set a threshold 10% more than what the models predicted. Eg: if the logistic regression model for Cash Back offers predicted that 1000 of customers have a probability of 0.8 or more of using a new cash back increase offer, then we set Nj to be 1100)
- nj -> Minimum number of customers for whom we should share an offer for one of the 4 offer types j
- B -> Maximum cost the bank can incur from additional offers shared with customers (this amount will be set to an amount lower than the combined balances the bank might lose if the customers were to close their accounts with the bank)

The below examples would help conceptualize the different offers under each of the 4 offer types.

Offer Type (j)	**Interest Reduction Offers**	**Cash Back Increase offers**	**Credit Card balance increase offers**	**New Account offers**
Example of Offers (i)	• 0.1% reduction in credit card APR rate • 0.2% reduction in APR rate for credit card	• 2% more cashback on restaurant purchases • 1% more cashback on grocery purchases	• 5% increase in credit card balance • 2% increase in credit card cash limit	• New savings account with 0% interest rate • New credit card with 0% APR

We will now use a linear optimization model which will give the best combination of each of the 4 deals which will result in the least cost to the bank, assuming that all the customers who were predicted to use the offers uses it. We can also look at building a separate optimization model for each customer type or segment type as needed.

Targeted Deal-Selection Optimization Model

```
Given
        Different types of offers
        Cost of each offer to the bank
        Minimum and maximum number of customers each of the offers can be sent to
        Maximum cost that can be incurred by the bank
Use
        Linear Optimization Model
To
        Pick the best combination of these offers with the least cost to the bank
```

The parameters of the Linear Optimization model would look something like the below.

- Cij -> Cost to the bank for each of the offers i within each offer type j (j = Interest Reduction Offers, Cash Back Increase offers, Credit Card balance increase offers, New Account offers). I am assuming that the cost to the bank for each offer type has already been pre-calculated by pricing models already available at the bank. We will just reuse the information from the pricing and product models.
- Nj -> Maximum of customer for whom we can share an offer for one of the 4 offer types j
- nj -> Minimum number of customers for whom we should share an offer for one of the 4 offer types j
- B -> Maximum cost the bank can incur from additional offers shared with customers (this amount will be set to an amount lower than the combined balances the bank might lose if the customers were to close their accounts with the bank)

Variables

Xij -> number of each offer i selected for each offer type j (eg: 2% cashback on restaurant purchases, 4% cashback on hotel bookings for offer type of 'Cash Back offers')
Yj -> 1 if offer j is selected, 0 if not

Constraints

Total Cost of Offers selected <= Budget/Maximum Cost B
Number of Offers Xi selected for offer type j should be >= nj
Number of Offers Xi selected for offer type j should be <= Nj

Objective Function
Minimize total cost of All Offers

Rebuilding of Model in subsequent months?

Since the parameters like the minimum and maximum number of offers that can be selected for each offer type, and the cost to the bank for each offer type varies from month to month, we would have to get an updated optimization model each month to give us the best combination of offers that can be shared with the customers.

Prescriptive Action From Analytics Approach

The optimization methods used above would then help us to arrive at the below prescriptive course of action as an example

1. For High Risk customers with Higher net balances, the below offers would help in customer retention
    a. Increasing the credit limit by 5%
    b. Giving more location-based offers at a discount of 2%(eg: cash back for any purchases at www.<website>.com)
    c. Offering a new credit card with APR of 19.3% and free balance transfer
2. For High Risk customers with lower net balances, the below offers would help in customer retention
    a. Cash-back offer increase of 0.05% per transaction
    b. Increasing cash-back on restaurant purchases by 1%

This knowledge of deals to be used for each of these customers can then be used to communicate these deals to the customer, as well as help management in coming up with better products that come pre-packaged with similar offers or rates. ***This will be the prescriptive outcome of the analytical approach that we have followed for Customer Retention.*** Based on these recommendations, the bank's marketing or customer communications team can then send out communications for these offers to the different customers. Due to regulations, banks are often forced to communicate to the customers ONLY through the channel the customer has opted-in for communications (Email, SMS, Physical mail etc). Hence, these deals will be communicated to the customers using these channels. Once the offers have been sent out, a survey can also be established, targeted at these customers to get a feedback on whether they are interested in the new deals that we have offered them.

Evaluating Analytics Efficiency

We can evaluate the success of the analytics-based recommendations over a short-term period(3 months) by gauging the below factors

- What percentage of high-risk customers are still with the bank?
    - A higher percentage of customers retained would be a testament to the success of this approach
- Has the percentage of customers who are leaving the bank or closing accounts decreased after we have implemented the analytics-based targeted deals, when compared to before the implementation?
    - A lower percentage of lost customers is again a testament to the success of this approach
- If we had offered new accounts to the customers, are they accepting the offer and opening new accounts with the bank?
    - This is a good metric to follow since a rise in account openings is not only a success factor for the analytics-based approach, it also brings in new business to the bank, which is a double whammy so to speak.
- What does the customer feedback look like on the deals that we had offered them?

Over a longer time frame, we can also use the below metrics to help with improving our analytics on some of the deals being offered. This would help us to either come up with more offers, or assist management in decisions around creating newer products to help retain customers at the bank.

- What type of deals are more used by customers?
- Are there certain cities where certain deals are more used?
- For the customers who leave the bank or close accounts, what type of deals were offered?

(Alternative option) A/B Testing of offers

Based on the outcome of the optimization models, if we need to test potentially multiple combinations of the offers that result in the lowest cost to the bank, then we can also perform an A/B testing of these offers over a period of time to help trim down the different deals that the bank offers to the customers. Each month, based on the outcome of the combination of deals offered, we can trim down the list of deals that are offered to the optimal minimum – this would also help management to come up with newer types of deals which might play a better role in customer retention than the ones we have used so far.

The below are the options that can be evaluated as part of this A/B testing

1. Which methods were used on customers who severed relationship with the bank in these 3 months?
2. Is there a method which was mostly ineffective (if so we can remove this method from our analysis going further)

Conclusion

As I started thinking about how the analytics-based solution can help the bank in retaining customers, I understood that there are multiple approaches in which this can be done – the one I have thought of above being just one possible approach. The investment-to-return ratio is one of the biggest factors that I feel analytics should help maximize, so that the initial investment in analytics results in a higher return for the bank in the long run. The approaches I mentioned above would take a few months of feedback, and evaluation from customer patterns to tweak and optimize the parameters. However, in the long run, I feel that the bank would have more dollar savings going with analytics based approaches, owing to more customers staying with the bank.

Addendum – Retention Of Small Business and Institutional Customers

While the analytical approaches used in my submission above helps solve the problem of individual customers leaving the bank, we can also follow a similar approach to come up with best recommendations for Small business and Institutional customers. Below are some of the key changes that I feel we would have on the models that we build for each of these customer types, but I feel the way we approach the problem would remain the same.

A few differences

- Factors used would change for regression and classification models since we are now focusing on the business/institution rather than an individual. Hence, factors like Married, #of children etc will be replaced by other business-centric factors like number of employees, headquartered zip code, type of business etc.
- While we could predict the future balances for individual customers, there might be some complexity in predicting balances for the business, since businesses usually have ‘floating balances’ wherein they see both an influx of funds from the business itself and an outflow of funds for operational expenses like salaries, and other expenses.
- Institutional customers would be more difficult to model since the volume of data (number of customers) might not be large enough for us to do effective training of the models.
- Institutional customer data might also come with its own set of restrictions on the data itself like some fields being off-limits for analytics like client, location, balance information etc.