Fulton Bank Datathon

Team 3
Jenny Chen, Katie Lee, Mark Rauschkolb, Kevin Sun, Hanson Wang
Project Description

- **Task:** 1) Segment Fulton Bank’s customers, 2) Identify customers who are most likely to churn, and 3) Predicts when customers are likely to leave
- To segment Fulton Bank’s customers and understand which customers were most likely to churn based on provided features, we implemented 3 classification approaches: decision tree analysis, logistic regression, and XGBoost
- Following classification, we created churn datasets and identified probabilities of churn for the customer segment most likely to churn using a shifted-Beta-Geometric model

### Customer Segmentation & Least likely to churn

- To segment customers based on their likelihood of churn, we used three distinct models: decision tree analysis, logistic regression and XGBoost
- XGBoost does not need feature engineering and scales well, however it is a bit-more time consuming to process than the other models. Decision tree, on the other hand, offers us an easy way to interpret our analysis
- Each model has their pros and cons, and we arrived at slightly different results and accuracy rate from those three models and will discuss our findings in the presentation

### When they will churn?

- The sBG model ultimately answers questions relating to customer retention and is applied to a churn dataset.
- Individual customer behavior story: at the end of each period, a customer renews his contract with probability $1 - \theta$ *(geometric distribution)*
- Churn propensities ($\theta$) varies across customers based on observable and unobservable characteristics
- To model unobserved heterogeneity, assume: $\theta \sim \text{Beta}(\alpha, \beta)$
Agenda

Member Introduction & Projection Description

Data Processing and Cleaning

Classification Methods

Decision Tree Analysis
Logistic Regression
XGBoost

Modeling Time to Churn

Takeaways & Conclusion
Data Preprocessing

Data Filtering

- Just like in our previous model, we only kept the most recent data of consumers (only those that were added after Dec. 2018.)
- This dataset includes 154 variables in total, and I dropped a few repetitive variables, and other variables like loan balance
- However, as shown in the missing data visualization below (msno.matrix), many variables still contain large amounts of missing data

Dummies Variables

- We believe there are a couple of non-numeric variables (Bank Branch, low / middle / moderate income, customer’s lifetime value categorization) that could be significant to predicting a customer likelihood of churn

Data Cleaning

- I dropped the columns that have more than 200,000 null values, and then dropped rows that have any null values at all
- The result is a dataset with 84 columns and 258,593 rows

Train Test Split

- We split the data into train test with a test size of 0.2 in order to evaluate the model
Agenda

Member Introduction & Projection Description

Data Processing and Cleaning

**Classification Methods**

- Decision Tree Analysis
- Logistic Regression
- XGBoost

Modeling Time to Churn

Takeaways & Conclusion
Decision Tree Analysis

No Checking Activity
MSE = 0.084
Samples = 228714
Value = 0.093

MSE = 0.048
Samples = 211025
Value = 0.051

No Presence of Direct Deposit Service
MSE = 0.241
Samples = 17689
Value = 0.594

MSE = 0.0
Samples = 4392
Value = 1.0

No ACH credits
MSE = 0.248
Samples = 13297
Value = 0.46

MSE = 0.166
Samples = 5152
Value = 0.791

MSE = 0.188
Samples = 8145
Value = 0.251

Method and Interpretation

- We ran a decision tree regressor on the y-train and x-train data with a maximum leaf nodes of 4, for simplicity’s sake.
- Our model yields an R-squared value of 0.48, while this is not necessarily an ideal accuracy score, it does paint a basic picture of the kinds of customers that would be likely to churn.
- As shown in the decision tree on the left, the key determining factor for customer churn, according to the decision tree regressor, is whether a customer has had any checking activity.
Logistic Regression

Method and Interpretation

- Because we are looking to standardize across *when* people opened their account, a logistic regression was the first way to start.
- After fitting the logistic regression model, we see that the accuracy of logistic regression classifier on test set is 0.90.
- (ROC) curve is a common tool used with binary classifiers. The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner).
Gradient Boosting

Extreme Gradient Boosting

- Speed and performance on top of traditional Random Forests
- Can handle sparse data (good for our sparse dataset for our one-hot encoded categorical features)
- Boosting allows weak learners to learn from past mistakes into a strong learner

Results

- Feature attribution should have **consistency and accuracy**
- By plotting the impact of a feature on every sample we can also see important outlier effects. For example, while DIRDEPSRV(Binary field showing presence (1) or absence (0) of direct deposit service in the household) is not the most important feature globally, it is by far the most important feature for a subset of customers
- Most product related features were not impact in our model, but the top ones included: STARTPROD_FTAC002 (Simply Checking) and NEWPROD_FCCSIG (Signature Credit Card)
Agenda

Member Introduction & Projection Description

Data Processing and Cleaning

Classification Methods
  - Decision Tree Analysis
  - Logistic Regression
  - XGBoost

Modeling Time to Churn

Takeaways & Conclusion
sBG Methodology (credit to Peter Fader)

Churn Model

- The sBG model ultimately answers questions relating to customer retention and is applied to a churn dataset.
- Individual customer behavior story: at the end of each period, a customer renews his contract with probability $1 - \theta$ (geometric distribution)
- Churn propensities ($\theta$) varies across customers based on observable and unobservable characteristics
- To model unobserved heterogeneity, assume: $\theta \sim \text{Beta}(\alpha, \beta)$

Continuous Mixture Model: Shifted-Beta-Geometric (sBG) Distribution

$$P(T = t \mid \alpha, \beta) = \int_0^1 P(T = t \mid \theta) f(\theta \mid \alpha, \beta) \, d\theta$$
$$= \frac{B(\alpha + 1, \beta + t - 1)}{B(\alpha, \beta)}.$$  

$$P(T > t \mid \alpha, \beta) = \int_0^1 P(T > t \mid \theta) f(\theta \mid \alpha, \beta) \, d\theta$$
$$= \frac{B(\alpha, \beta + t)}{B(\alpha, \beta)}.$$  

Churn Dataset Example

<table>
<thead>
<tr>
<th>Year</th>
<th># Customers</th>
<th>% &quot;Alive&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1000</td>
<td>100%</td>
</tr>
<tr>
<td>1</td>
<td>800</td>
<td>80%</td>
</tr>
<tr>
<td>2</td>
<td>700</td>
<td>70%</td>
</tr>
<tr>
<td>3</td>
<td>550</td>
<td>55%</td>
</tr>
<tr>
<td>4</td>
<td>510</td>
<td>51%</td>
</tr>
<tr>
<td>5</td>
<td>500</td>
<td>50%</td>
</tr>
<tr>
<td>6</td>
<td>490</td>
<td>49%</td>
</tr>
<tr>
<td>7</td>
<td>471</td>
<td>47%</td>
</tr>
</tbody>
</table>

Churn Dataset Example
sBG Methodology (credit to Peter Fader)

### Methodology

1. **Construct churn datasets for each account start date, using the feature relmonths**

2. **Identify customers within the two identified segments from the decision tree classification**

3. **For the segment of customers most likely to churn, apply an sBG model to understand probabilities of churn in the months following account opening**

4. **Interpret the Alpha Beta hyperparameters to understand when current customers will churn**

<table>
<thead>
<tr>
<th>RELMONTHS</th>
<th>Count</th>
<th>% Survived</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>0.98473282</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>0.98091603</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>0.97328244</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>0.95801527</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>0.93129771</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>0.91984733</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>0.90839695</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>0.90076336</td>
</tr>
<tr>
<td>8</td>
<td>19</td>
<td>0.89694656</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>0.88931298</td>
</tr>
<tr>
<td>10</td>
<td>22</td>
<td>0.88167939</td>
</tr>
<tr>
<td>11</td>
<td>23</td>
<td>0.8740458</td>
</tr>
<tr>
<td>12</td>
<td>24</td>
<td>0.87022901</td>
</tr>
<tr>
<td>13</td>
<td>26</td>
<td>0.86641221</td>
</tr>
<tr>
<td>14</td>
<td>28</td>
<td>0.85114504</td>
</tr>
<tr>
<td>15</td>
<td>29</td>
<td>0.84351145</td>
</tr>
<tr>
<td>16</td>
<td>30</td>
<td>0.83587786</td>
</tr>
<tr>
<td>17</td>
<td>32</td>
<td>0.83587786</td>
</tr>
</tbody>
</table>

*Churn Dataset for Accounts Started March 2018*
sBG Model

Takeaways

- Given the distribution of theta, which represents propensity to churn, we see that there is very little *unobserved heterogeneity* amongst the of customers in the most likely to churn segment.
- This implies that most of the customers who are most likely to churn have similar motivations for opening/closing accounts.

![Graph showing actual vs. predicted churn rates](image)

**Beta**

- $\beta$:
  - 0.0 0.5 1.0
- 1

**Alpha**

- $\alpha$:
  - 25454410.88
- 185360316.1

Average Theta: 12%
## Summary

### Gradient Boost
- On a randomly sampled test set, our XGBoost model could predict whether they churn (closed_HH=1) with 98% accuracy.
- Additionally, we identified the feature importance of features such as RELMONTHS and product features such as the presence of the Signature Credit Card. We also found that the higher their Mobile Usage, Online Usage the less the likelihood of churn.
- Generally, customers with smaller months with Fulton and do not have a Signature Credit Card are more likely to churn.

### Decision Tree
- For the decision-tree analysis, we refiltered the data by dropping columns that have more than 300,000 null or missing values.
- We also dropped columns with significant collinearity that would impact the accuracy of our model, and dropped any rows with NA.
- A lack of checking account or direct deposit is a clear warning that a customer is likely to churn.

### Logistic Regression
- Because we are looking to standardize across when people opened their account, a logistic regression was the first way to start.
- After fitting the logistic regression model, we see that the accuracy of logistic regression classifier on test set is 0.90.
- (ROC) curve is a common tool used with binary classifiers. The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner).

### sBG Model
- From the sBG methodology, we gain a deeper understanding of the segment of customers who are most likely to churn and can calculate approximate probabilities of churn at each time period following account opening.
- We expect all customers classified as least loyal to churn after ~2 years.

### Suggestions
- Improve mobile app interface and website, convince customers to join mobile app as part of marketing.
- Product-wise, the signature credit card program has the best performance in reducing churn, Fulton could have more similar programs to increase customer loyalty to reduce churn.
- Overall, be wary of customer churn in the initial two years because eventually customers will become loyal and generate high value.
Q&A