DataRobot

community.datarobot.com
The Art of Creating a Target Variable in Predictive Modeling

Lukas Innig
Rajiv Shah
Target creation is underrepresented

Most articles jump right into modelling with a prepared dataset.
What is the question?
How to define Churn event

- Subscription services
  - Insurances, Banks, Video streaming, Utilities, …
- High frequency of transactions
  - Grocery retailer
  - eCommerce
- Intermittent demand
  - Home retailer
  - Car dealership

Use Cancellation Date
Define period of inactivity
Consider other use cases

Easy
Hard
What to do with churners

• Need intervention:
  • Phone call/email, Discounts, Free delivery, ...

• Need this to be effective!
  • No uptake -> can’t prevent churn

• Need this to be cost effective!!
  • Keep customer only through heavy discounts
    -> Have unprofitable customers
Consider cost of interventions

- Low Risk of Churn
  - Low Success rate: Leave alone
  - High Success rate: Focus here

- High Risk of Churn
  - Low Success rate: Leave alone
  - High Success rate: Focus here
Ask a well-defined question to answer...

Early Churners

• What is the likelihood customer X will churn in the first 3 months of onboarding?

Normal Churners

• What is the likelihood customer X will churn in the next 6 months?

Repeat Churners

• What is the likelihood customer X will repeat churn?
Prediction: Will the churn occur in the first N months?

Target is if the customer churned in the first N months

**PROS**

- VERY SIMPLE and straightforward
- Will be able to use all observations of historical data
- Can deal with complex customer situations

**CONS**

- Not taking into account activity, only using onboarding data
- Can elongate the window to take more data into account
- Will be problematic if customer behaviors changed over time
Prediction: Will the churn occur in the **first 3 months**?

Note: Can/should experiment with multiple forward looking models
Prediction: Will the churn occur in the next N months?

Target is if customer churned in the last N months

**PROS**

- Will use more recent data

**CONS**

- More complex data prep
- Will have to exclude data if observation did not occur within the window of time that we are interested in

**TIPS**

- The further the look-back window used, the better the model can be (the tradeoff is there will be less lead time in production)
- Very common to use multiple models with different look-back windows
  - 3 month window
  - 12 month window

Note: Need to use customers that have been active as of a defined start time
Will the churn occur in the next 6 months?

<table>
<thead>
<tr>
<th>Unique Cust.</th>
<th>Last 36 Months</th>
<th>Last 6 Months</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>C</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>C</td>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>C</td>
<td>1</td>
</tr>
</tbody>
</table>

C : Churn Event

Exclude

Today’s Data Snapshot
Prediction: What is the likelihood Customer X will churn next month? (Person-Period form)

**PROS**

- Can use all data
- Can capture seasonalities

**CONS**

- Can be difficult ETL for some
- Need to be careful during training that churn windows don’t overlap
- ^ There’s a workaround though

<table>
<thead>
<tr>
<th>Id_name</th>
<th>Month</th>
<th>churn_by_month</th>
<th>emp_title</th>
<th>emp_length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jim</td>
<td>January</td>
<td>0</td>
<td>Time Warner</td>
<td>10</td>
</tr>
<tr>
<td>Jim</td>
<td>February</td>
<td>0</td>
<td>Time Warner</td>
<td>10.1</td>
</tr>
<tr>
<td>Jim</td>
<td>March</td>
<td>0</td>
<td>Time Warner</td>
<td>10.2</td>
</tr>
<tr>
<td>Jim</td>
<td>April</td>
<td>0</td>
<td>Time Warner</td>
<td>10.3</td>
</tr>
<tr>
<td>Jim</td>
<td>May</td>
<td>0</td>
<td>Time Warner</td>
<td>10.4</td>
</tr>
<tr>
<td>Jim</td>
<td>June</td>
<td>0</td>
<td>Time Warner</td>
<td>10.5</td>
</tr>
<tr>
<td>Jim</td>
<td>July</td>
<td>0</td>
<td>Time Warner</td>
<td>10.6</td>
</tr>
<tr>
<td>Jim</td>
<td>August</td>
<td>0</td>
<td>Time Warner</td>
<td>10.7</td>
</tr>
<tr>
<td>Jim</td>
<td>September</td>
<td>0</td>
<td>Time Warner</td>
<td>10.8</td>
</tr>
<tr>
<td>Jim</td>
<td>October</td>
<td>0</td>
<td>Time Warner</td>
<td>10.9</td>
</tr>
<tr>
<td>Jim</td>
<td>November</td>
<td>1</td>
<td>Time Warner</td>
<td>10.9</td>
</tr>
<tr>
<td>Joan</td>
<td>January</td>
<td>0</td>
<td>Photos By Fo</td>
<td>6</td>
</tr>
<tr>
<td>Joan</td>
<td>February</td>
<td>0</td>
<td>Photos By Fo</td>
<td>6.1</td>
</tr>
<tr>
<td>Joan</td>
<td>March</td>
<td>1</td>
<td>Photos By Fo</td>
<td>6.2</td>
</tr>
</tbody>
</table>
Will the churn occur in the next 6 months?

<table>
<thead>
<tr>
<th>Unique Cust.</th>
<th>Previous month</th>
<th>Last month</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>7</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>C</td>
</tr>
</tbody>
</table>

C : Churn Event
Question Time
DataRobot Community

- Engage, learn, and accelerate your AI/ML journey
- Connect with peers to find solutions to AI challenges
- Explore helpful content to take your AI to the next level
- Build your brand as an AI expert & thought leader
- Join your peers today at community.datarobot.com

Questions: aisuccess-webinars@datarobot.com