LearnZillion

User Churn

Huey Tran

10/06/2014

Initial problem

At a startup, there is a focus on finding a leading indicator for a user who would turn into an engaged user. Work streams would then optimize around this metric.

7 friends in 10 days

1-day retention

Follow ≥ 30 people

Put ≥ one file in one Dropbox folder on one device

X connections in Y days

Source: Lean Analytics (2013), Alistair Croll and Benjamin Yoskovitz

LearnZillion

Question: What is LearnZillion's leading indicator for a user who would turn into an engaged user?

List of possible *levers*

- Viewed a lesson page
- Played a video
- Viewed Common Core Navigator
- Viewed Courses
- Downloaded a document
- Used search
- Created a class
- Added students
- Created an assignment

Timeframe

- One day
- 7 days
- 14 days
- 30 days

LearnZillion

Initial thoughts and ideas for what LearnZillion's leading engagement metric could be

- Viewed 3 videos within 14 days since account created
- Created a class within 14 days since account created
- Downloaded 2 documents in 14 days since account created
- Some combination of *levers* within 14 days since account created

Acquiring and processing data

Acquiring the data proved to be difficult

- Mixpanel
- API call needs to be executed to get hundreds of thousands of event data
- Original script contained errors
- Wanted to convert JSON into CSV
- Next iteration of script contained more errors
- Success (finally)!

Processing the data revealed non-trivial inconsistencies in our data

- Renaming of properties
- Inactive properties
- Inconsistent formatting
- Users were double-counted
- Pre-processing the data was A LOT of fun!

These were the parameters for the dataset

- **Users**
  - Churned Users
    - Account created within 6 month timeframe
    - Last visited was ≥ 30 days ago
  - Active users
    - Account created within same 6 month timeframe
    - Last visited was ≤ 30 days ago
  - Relevant properties
    - Account created date
    - Number of classes created
    - Number of videos played
    - Number of documents downloaded
    - Number of times visited
- **Events**
  - Visited
  - Video played
  - Document downloaded
  - Class created
  - Timeframe
    - Same timeframe as on the left

But before I did anything too crazy, I ran simple correlation tests

- Ran correlation between X variable event (e.g. played video, downloaded resource, created class) and Y variable event (i.e. visited)
  - Based on Pearson's r correlation, there was a weak positive relationship for
    - (played video, visited)
    - (downloaded resource, visited)
    - This is somewhat to be expected since the variables are orthogonal.
  - Based on Pearson's r correlation, there was no or negligible relationship for
    - (create class, visited)
  - In summary, weak correlations all around!
    - Should I waste time putting in effort to move forward if correlations are weak?
      - NAH!

Reframing the problem

LearnZillion may not have a leading engagement metric

- Leading engagement indicators fit in three categories
  - Network density
  - Content added
  - Visit frequency

- LearnZillion
  - No built-in virality factor(s) or element of community to enhance this
  - No ability for users to participate in any form beyond consuming content
  - Visit frequency entirely dependent, then, on user experience and academic content (finite)

Enter: New problem

- How does LearnZillion identify those users who, within the first 14 days of account creation, exhibit behavior consistent with “churned” users?
- Enter data science!
  - Classification problem
  - Use three modeling techniques
    - Support vector machine
    - Random forest
    - K-Nearest-Neighbors
- Helps business with:
  - Retargeting campaigns
  - Inbound / Outbound marketing or demand generation
  - Content marketing

LearnZillion

“Some combination of levers within 14 days of account created”

- Created dataset for all users who created an account within specified timeframe with the following properties
    - User ID
    - Account created
    - Videos played
    - Resources downloaded
    - Classes created
    - Visited
    - Churn
        - where value = true if user belongs to group that has not been on the site ≥ 30 days
        - where value = false if user belongs to group that has been on the site ≤ 30 days

The actual data science part

sklearn, cv, svm, rf, knn, oh my!

- Import __Future__, Pandas
- Read in CSV and created dataframe
- Defined X and y (where y = 1 if true and y = 0 if false for “Churn”) for feature space
- From sklearn:
    - Imported StandardScaler
    - Imported KFold
    - Imported SVC for SVM
    - Imported RF for RF
    - Imported KNN for KNN
- Import numpy
- Defined accuracy (where y = y) for SVM, RF, and KNN
- Profit!

Results

- Support vector machine
  - 87.5%
- Random forest
  - 86.4 %
- K-Nearest-Neighbor
  - 85.5%

Implications and Conclusions

What do I do with this information and knowledge? How can I improve it?

- Information can help business with:
  - Retargeting campaigns
  - Inbound / Outbound marketing or demand generation
  - Content marketing
- Clean up our data! This will help save lots of time for future projects
- Product: consider feature ideas that can help our viral coefficient
- Get probability of churn vs. classifier churn or no churn
- Approach algorithm differently to get ≥ 90% accuracy