Predicting User Retention in Subscription Based Services

The Data Incubator Finalist Interview

William Lewis

11/15/2017

Subscription Services: They're not just for magazines

- Wide range of products and services with subscription based consumption models
  - Music and Movies (Netflix, Spotify, Apple Music, Amazon Prime)
  - Online gaming (e.g. Xbox live, PlayStation Now)
  - Software (Adobe CS, Microsoft Office 365)
  - Consumer goods (Dollar shave club, BarkBox)

- Exhibit continued growth (below is about subscription box companies):

> In the month of April 2017, subscription company websites had about 37 million visitors. Since 2014, that number has grown by over 800%.[1]

  - Yet in the same article: "We are seeing a lot of volatility within various categories."

- Motivates: "Can user demographics and usage behavior be used to predict retention."

[1] *Forbes https://www.forbes.com/sites/richardkestenbaum/2017/08/10/subscription-businesses-are-exploding-with-growth/#4f4546c96678*

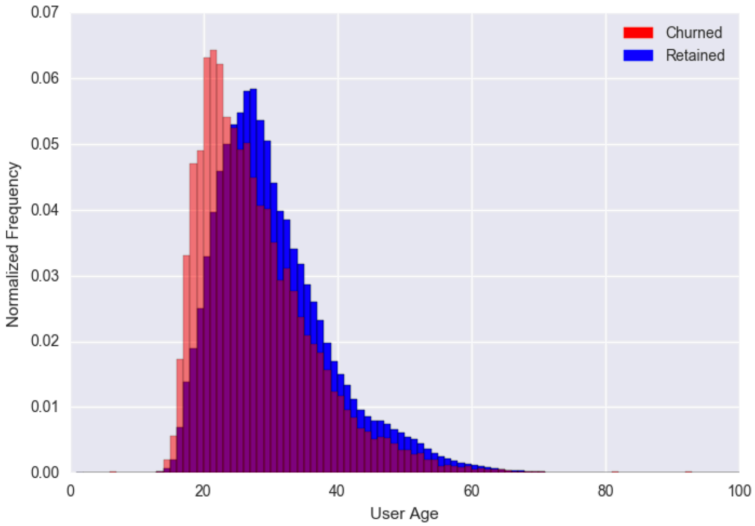
Predicting User Retention: Music Streaming Service

- Recently a music streaming service made 40GB (decompressed) of user demographics, transaction, and usage data available
  - https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data
- Project demonstrates skills with
  - Working with large data
  - Bottleneck Identification
  - Data cleaning
  - Feature engineering
  - Model refinement and selection
- Includes an ~30GB user log file
  - previously analyzed only a small portion of that data
  - Created a SQLite database which may be queried by groups of users
    - In order to achieve results on timescale available:
      - Index creation
      - Fewer queries

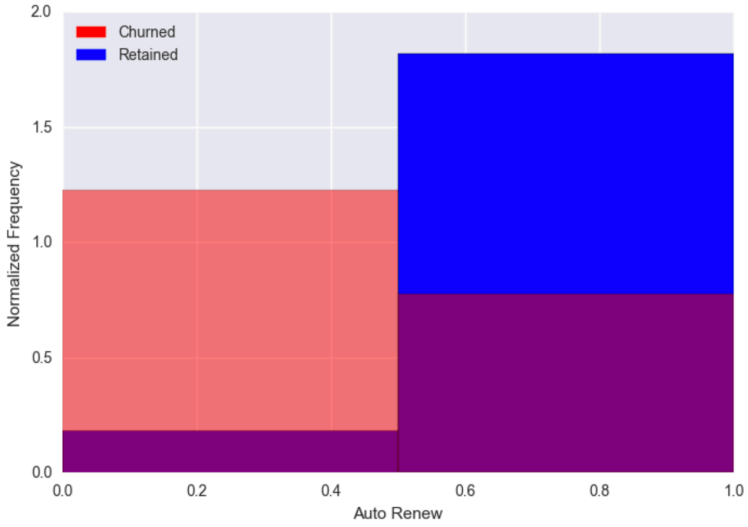
User relationships

Young users fleeing the nest?

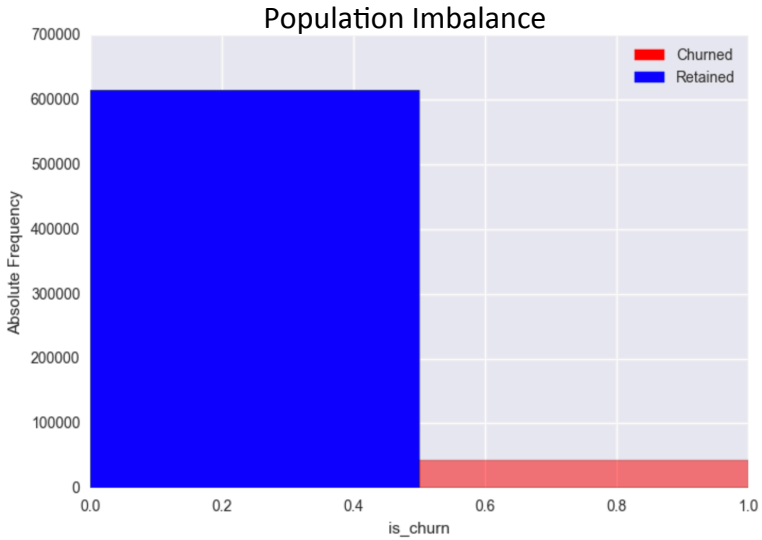
```
The t-test for difference in mean age between churn and retained users gives:
Ttest_indResult(statistic=-25.527175423464698, pvalue=2.1077667057649482e-142)
```

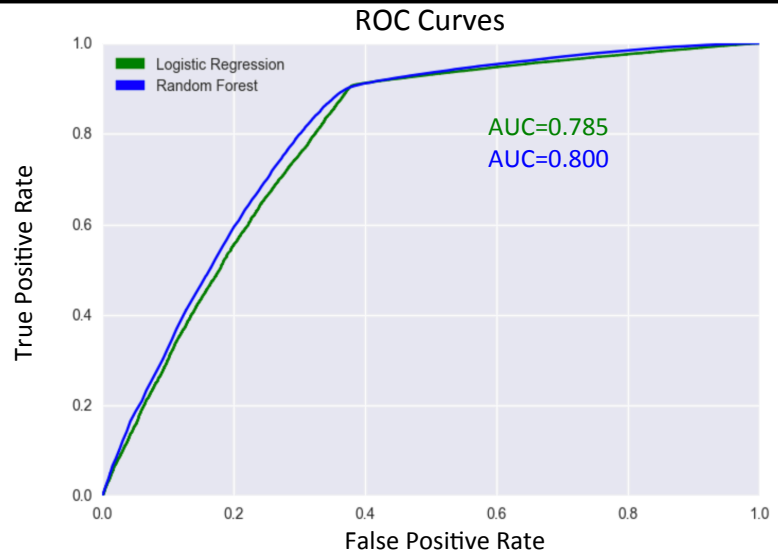


Auto renew is how they get you!



First pass with Logistic Regression and Random Forest

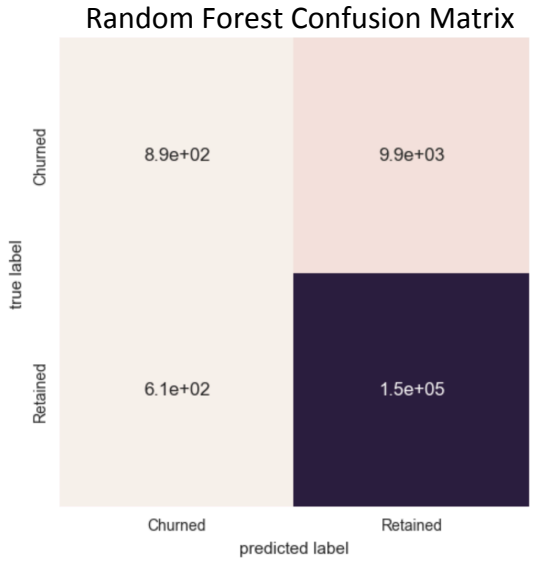




Random Forest 2-fold cross-validation estimates for:

```
Sensitivity is: 0.996036181218
Specificity is: 0.0827682303762
Positive Predictive Value is: 0.939481232685
Negative Predictive Value is: 0.593604263824
Accuracy is: 0.93632824599
```

Red flag?!



Additional Remarks and Next Steps

- I tried undersampling (oversampling) the majority (minority) populations
  - For logistic regression saw improvement in specificity but overall accuracy dropped
- Need to return to feature engineering step
  - Initial run with only 20 features.
    - ~50% demographics
    - ~50% aggregated features from log file
    - One feature is the most recent auto-renew status
  - Many interesting features are possible
    - measurements for distributions of usage for each user
      - e.g. standard deviation
    - weighted averages rather than unweighted
      - weight later data more
    - total membership duration
    - total amount spent
    - average discount