Troy Kayne
Nov. 24, 2018
Github Repo: [kayne10/gaia-take-home-exercise]

Problem 2: Data Analysis

One goal we have at Gaia is reducing the frequency in which our members churn, meaning cancel their subscription service. Imagine that non-technical business stakeholders (product managers, marketers, executives, etc…) have come to you looking for insights in our data surrounding churn/retention. Assume that you pulled together the following random sample of data (column definitions below):
https://drive.google.com/file/d/1_rYzbmx5XFAaEo6TUwWsDf-4cwAb9jke/view?usp=sharing

- **user_id**: unique ID for a member/user/subscriber of our service
- **churn**: a 0 or 1 flag where 1 means that the user has canceled service in the past 30 days and 1 means that the user is still a paying subscriber
- **days_as_member**: the number of days the user has either been a paying member or the number of days the user was paying prior to canceling service
- **plan_name**: The name of the plan the user is paying for
- **plan_period**: The duration/cadence that user's plan grants access to our service and the the user is thus billed again
- **days_since_last_login**: The number of days since the user has last logged into our service (from their last day of being a paying subscriber)
- **days_since_last_video_view**: The number of days since the user has last viewed at least one second of video content from our service (from their last day of being a paying subscriber)
- **behavior_segment**: An internally derived user segment denoting the type of content the user is most interested in viewing
- **total_seconds_watched**: The total number seconds of content watched by the user in the over past 180 days
- **total_videos_watched**: The total number videos watched for at least 1 second by the user in the over past 180 days
- **total_videos_completed:** The total number videos watched for at least 80% of their duration by the user in the over past 180 days
- **total_videos_under_one_minute:** The total number videos watched for between 1 and 59 seconds by the user in the over past 180 days
- **total_series_watched:** The distinct series (think TV series) watched by the user in the over past 180 days

Questions

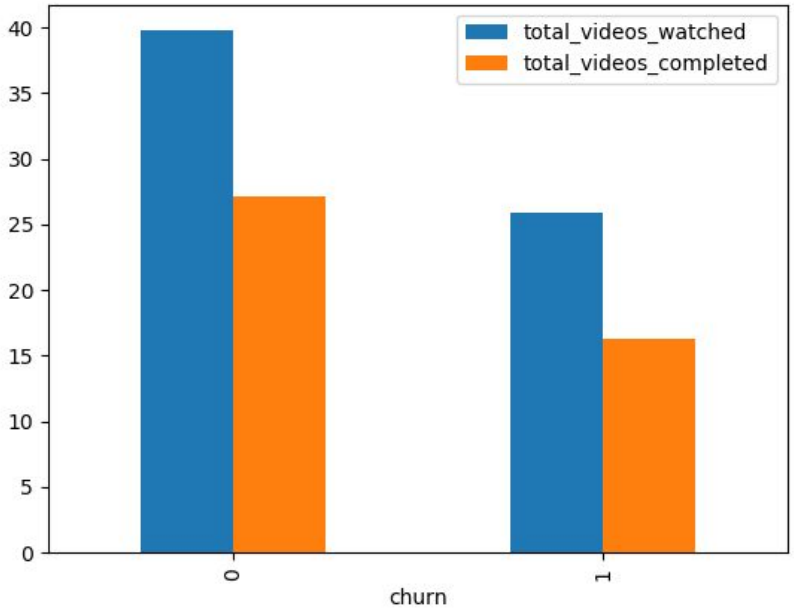
1. What do you notice about users who churn/do not churn? Are there specific types of users more or less likely to churn? Please provide a few sentences describing an insight or two that you see in the data.

   I have noticed that the means of most features for users who do not churn are regularly greater than the means of most features for users who do churn. The **total_videos_completed** is a significant feature regarding users who churn (almost half as much) as well with **days_as_member**. Interestingly, under the feature of **days_since_last_login,** users who churn have a mean of 151 days and users who do not churn have a mean of 138 days. That is fairly close and surprising to me. Also, features **total_videos_completed** and **total_videos_watched** have a correlation value of .97.

2. Form a hypothesis about why certain users are more or less likely to churn. Please create a data visualization (graph, table of data, etc…) to support this hypothesis and write a brief description of how those visualization support your hypothesis.

Hypothesis: Users with fewer videos completed/watched are more likely to churn than users with more videos completed/watched.

Data Visualization:



Description: This bar graph above represents the mean of total videos watched and total videos completed for users who churn and do not churn. A churn value of 0 (do not churn) has a higher average in total videos watched and total videos completed than users who do churn. This illustrates that users seem to cancel their subscription if they do not watch/complete videos. They may be not be interested in watching videos anymore.

3. How would you validate the hypothesis you proposed in (2)? Do not actually conduct the analysis, just describe in a sentence or two how you would conduct it if you had the time to do so.

I would conduct hypothesis testing with two mutually exclusive hypotheses. My null hypothesis would be users with less than 12 total videos completed will churn, and my alternate hypothesis would be users with more than 12 total videos completed will not churn. I would find the t-test statistics, and then I would compute the probability of my results assuming the null hypothesis is true. With a significance level of 0.05, I would compare my p-value and alpha to determine if I reject or fail to reject my null hypothesis.