Beverly Rice
Springboard Capstone 1 Project: Data Wrangling
19 Dec 2017

Overview:

Steps taken
Summary of changes made
Specific findings

Steps (see 'Data Wrangling.ipynb' on github profile):

Step 1: Acquire Data

The data was acquired through the kaggle.com KKBox Challenge page: https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data. Each file was downloaded and saved to a folder in my home directory

Step 2: Initial Exploratory Data Analysis

Applied info(), describe(), columns, min(), max(), isnull(), duplicated(), unique(), sum(), size, groupby, agg, to_datetime(), and transform(). Also plotted initial counts for each variable

Step 3: Clean Data

Identified/handled missing variables and outliers. Also narrowed down the timeframe based on significant changes in member count and usage

Step 4: Decide which original and new variables to factor in in further analysis

Summary of Changes Made:

1. Removed 'bd' variable – ranged from -7000 to +2000 and over 67% reported an age of '0'
2. Removed 'gender' variable – 65% did not report a gender; analysis shows that reported males and females churned at a very similar rate, 17.6% and 16.9% respectively.
3. Removed members who registered before 2010 – KKBox went international in 2009 and also received investments from at least 2 big-name companies; there was a significant increase in membership starting in 2010
4. Added Variable - registration day and month
5. Added Boolean – whether member used multiple payment methods
6. Added Boolean – whether member used multiple payment plans
7. Added Boolean – whether member received a discount
8. Added Variable – total subscriptions per member
9. Added Variable – total membership time
10. Added Variable – total days used
11. Added Variable – total songs listened to throughout membership
12. Added Variable – total songs listened to more than once

13. Added Variables – percentage of songs listened to for only 25%, 50%, 75%, 98.5% and 100% of the song length throughout membership

Specific Findings:

- members_v3.csv.7z
  - Given description: user information as of 11/13/2017
  - Zipped: 231.08 MB, Unzipped: 427.9 MB
  - 6769473 entries, 6 columns
    - msno: KKBox user IDs
      - 6769473 non-null objects; none missing
      - All unique/no duplicates
      - Made of combo of 44 characters/nums/symbols
      - All end with '='
    - city:
      - 6769473 non-null integers; none missing
      - 21 values 1:22; No '2'
      - Over 70% came from city 1
    - bd: age of members
      - 6769473 non-null integers; none missing
      - 386 values -7168:2016
      - Outliers present; wide range; likely falsified
      - 67% reported an age of '0'; most other users are between the age of 15 and 55
      - Dirty and therefore will likely be thrown out
    - gender
      - 2339968 non-null objects; 4429505 null objects
      - 2 values female and male
      - Analysis shows females and males churn at almost the same rate (varied by 1.3%)
      - 2/3 did not specify gender and considering the identified males and females churned at almost the same rate, variable will likely be thrown out
    - registered_via: registration method
      - 6769473 non-null integers; none missing
      - 18 values -1:19; No 0, 12, 15
      - >97% registered via method 3, 4, 7 & 9
    - registration_init_time
      - 6769473 non-null integers; none missing
      - will need to convert to date
      - Range: 2004-03-26 to 2017-04-29
      - Data shows registration significantly increased starting in 2010… Research shows that the company expanded beyond it's home country (Taiwan) and into the

international market in 2009. Further, KKBox received investments from KDDI Corporation and HTC Corporation in 2011. It also established a new live feature in 2014. Based on this info, we will likely focus on 2010 and newer

  - New Variables Created
    - reg_day: extracted registration day of month
    - reg_month: extracted registration month
- sample_submission_v2.csv.7z
  - Given description: the test set, containing the user ids, in the expected submission format as of 11/06/2017
  - Zipped: 29.25 MB, Unzipped: 42.7 MB
- train_v2.csv.7z
  - Given description: the train set, containing the user ids and whether they have churned as of 11/06/2017
  - Zipped: 31.03 MB, Unzipped: 45.6 MB
  - 970960 entries, 2 columns
    - msno: KKBox user IDs
      - 970960 non-null objects
      - All unique/no duplicates
      - Made of combo of 44 characters/nums/symbols
      - All end with '='
    - is_churn: IDs whether user has or has not churned
      - 970960 non-null integers
      - 0 for non churners; 1 for churners
        - 883630 non churners; 91.0%
        - 87330 churners; 8.99%
- transactions_v2.csv.7v
  - Given description: transactions of users as of 11/06/2017
  - Zipped: 46.59 MB, Unzipped: 115.4 MB
  - 1431009 entries, 9 columns
    - msno: KKBox user IDs
      - 1431009 non-null objects; none missing
      - 1197050 unique values; 233959 duplicates
      - Made of combo of 44 characters/nums/symbols
      - All end with '='
    - payment_method_id: payment method
      - 1431009 non-null integers; none missing
      - 37 values 2:41
      - 48% used method 41
      - Some mbrs used multiple payment methods: 2, 3, 4, 5, and as high as 8 methods
    - payment_plan_days: length of membership plan in days
      - 1431009 non-null integers; none missing
      - 31 values 0:450

      - Some mbrs tried different payment plan day options: up to 5 different plans
    - plan_list_price: in New Taiwan Dollar (NTD)
      - 1431009 non-null integers; none missing
      - 48 values 0:2000
    - actual_amount_paid: in New Taiwan Dollar (NTD)
      - 1431009 non-null integers; none missing
      - 53 values 0:2000
    - is_auto_renew
      - 1431009 non-null integers; none missing
      - 2 values 0 and 1
    - transaction_date: format %Y%m%d
      - 1431009 non-null integers; none missing
      - Range: 2015-01-01 to 2017-03-31
    - membership_expire_date: format %Y%m%d
      - 1431009 non-null integers; none missing
      - Range: 2016-04-19 to 2036-10-15
      - Majority of subscriptions will expire in 2017, in the month of April, or on the 30th day
    - is_cancel: whether or not the user canceled the membership in this transaction.
      - 1431009 non-null integers; none missing
      - 2 values 0 and 1
  - New Variables Created
    - methods_used: whether mbr used multiple payment methods
      - create Boolean – used multiple methods
    - plans_used: whether mbr tried different payment plans
      - create Boolean – used multiple plans
    - discount: plan_list_price – actual_amount_paid
      - create Boolean – received a discount
    - vip: mbr paid 0 dollars for subscription (Boolean)
    - trans_count: number of transactions per user
    - mbr_time: registration_init_time – membership_expire_date
- user_logs_v2.csv.7z
  - Given description: daily user logs describing listening behaviors of a user as of 11/06/2017
  - Zipped: 654.17 MB, Unzipped: 1.43 GB
  - 18396362 entries, 9 columns
    - msno: KKBox user IDs
      - 18396362 non-null objects; none missing
      - 1103894 unique values; 17292468 duplicates
      - Made of combo of 44 characters/nums/symbols
      - All end with '='
    - date: format %Y%m%d

    - 18396362 non-null objects; none missing
    - Range: 2017-03-01 to 2017-03-31
  - num_25: # of songs played less than 25% of the song length
    - 18396362 non-null objects; none missing
  - num_50: # of songs played between 25% to 50% of the song length
    - 18396362 non-null objects; none missing
  - num_75: # of songs played between 50% to 75% of of the song length
    - 18396362 non-null objects; none missing
  - num_985: # of songs played between 75% to 98.5% of the song length
    - 18396362 non-null objects; none missing
  - num_100: # of songs played over 98.5% of the song length
    - 18396362 non-null objects; none missing
  - num_unq: # of unique songs played
    - 18396362 non-null objects; none missing
  - total_secs: total seconds played
    - 18396362 non-null objects; none missing
- New Variables Created
  - days_used: number of total logs for that member
    - create boolen – over 20 days
  - total_day: total songs listened to in that one day
  - fav_songs: # songs listened to multiple times that day (total songs – unique songs)
  - sum_songs: total number of songs the mbr listened to throughout membership
  - p25: percentage of total # of songs played less than 25% of the song length throughout entire membership
  - p50: percentage of total # of songs played less than 50% of the song length throughout entire membership
  - p75: percentage of total # of songs played less than 75% of the song length throughout entire membership
  - p985: percentage of total # of songs played less than 985% of the song length throughout entire membership
  - p100: percentage of total # of songs played less than 100% of the song length throughout entire membership