Springboard

UNTAPPD
DRINK SOCIALLY

SPRINGBOARD CAPSTONE PROJECT ONE
PREDICTING MY NEXT CRAFT BEER

Author: Josh Mayer
Date: September 14, 2017

Springboard

UNTAPPD
DRINK SOCIALLY

OUTLINE

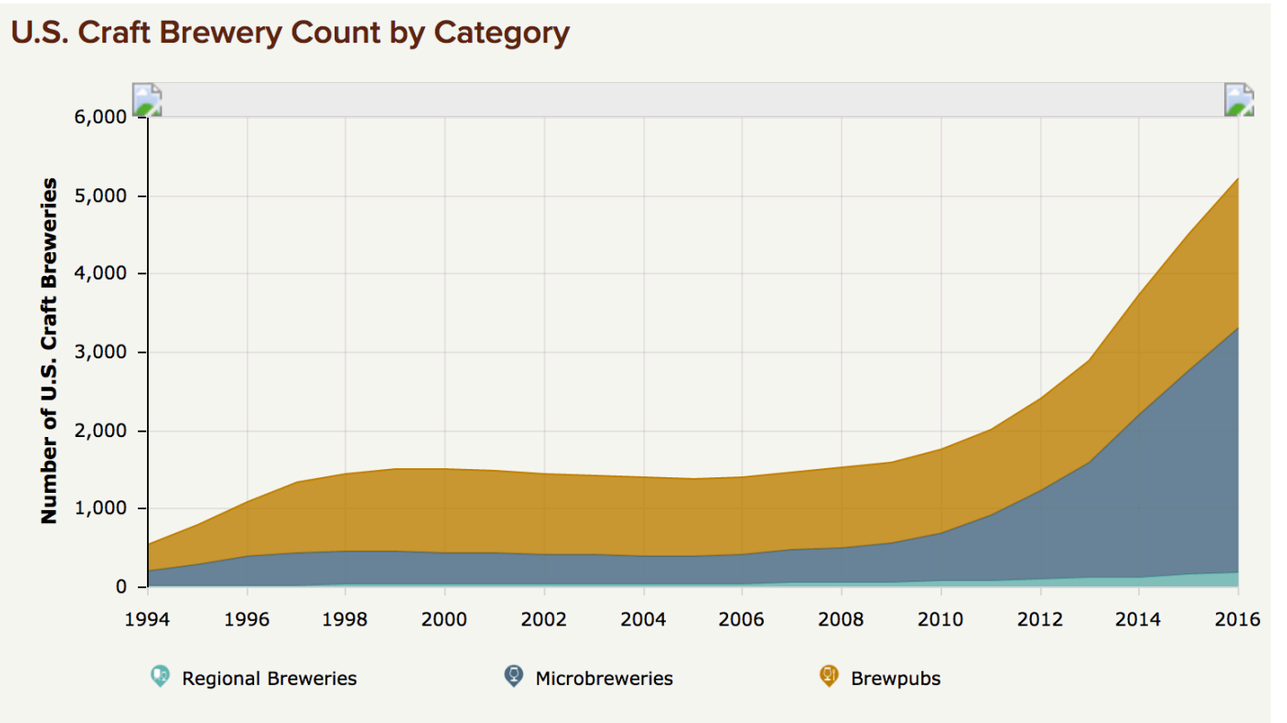
- Introduction & Problem Statement
- The Untappd Application
- Dataset
- Analysis & Findings
- Statistical Inference
- Machine Learning
- Conclusions

INTRODUCTION & PROBLEM STATEMENT

With almost 200% new craft breweries since 2010 and over 5,000 unique craft breweries in the U.S. today, beer is one of the fastest growing products in America.

THE PROBLEM

Given the substantial rise in the number of craft breweries deciding which beer to try next can be a confusing decision.



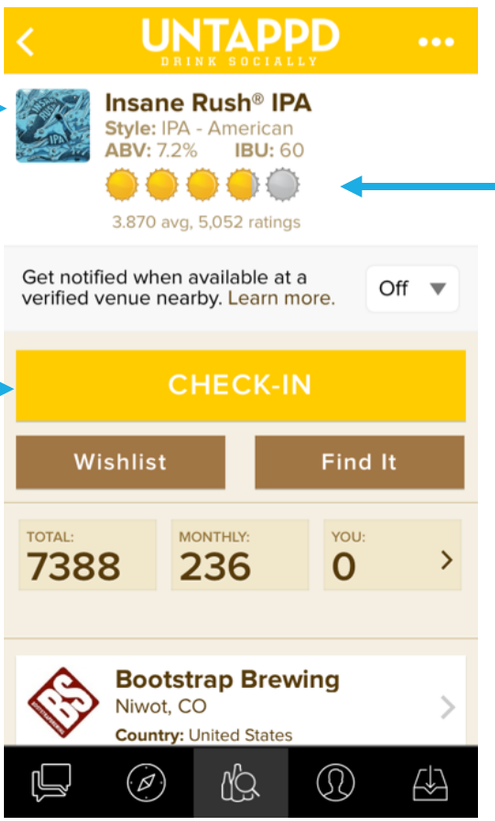
Source: https://www.brewersassociation.org/statistics/number-of-breweries/

THE UNTAPPD APPLICATION

Untappd is an existing application that allows its users to rate and track each beer sampled. This data should be leveraged to help its users determine which beer to try next.

1) In this example I tried the beer named "Insane Rush IPA".

2) Upon check-in users may elect to rate the beer based on a 1 (worst) to 5 (best) scale.



Screenshot of the Untappd iPhone Application

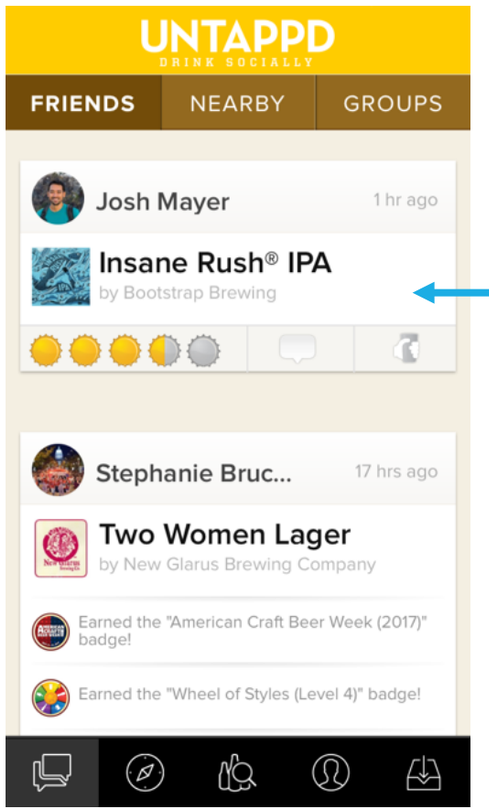
3) Each beer shows you the average user rating as well as the number of unique ratings.

Springboard

UNTAPPD
DRINK SOCIALLY

THE GOAL

The Untappd application should leverage its user ratings and provide tailored recommendations based on an individual users preferences.



Screenshot of the Untappd iPhone Application

PRIMARY OBJECTIVE

Based on a user's historical ratings create predictions for untasted beers and create a personalized recommendation on which beer to sample next.

Springboard

UNTAPPD
DRINK SOCIALLY

DATA ACQUISITION & WRANGLING

Data was acquired via calling the Untappd API. In total, almost 30,000 API calls were required to compile all of the necessary data for the project.

	Primary Dataset (My personal ratings) *Test/Train Set*	**Secondary Dataset** (Untasted beer) *Prediction Set*
Results Returned per API Call	50	1
Total Unique Records (Beers) Acquired	181	19,838
API Calls Required	4	29,146
Features Returned per Beer	35	53

DATA WRANGLING

- Duplicate Records - 9,308 records (32% of the acquired records) were dropped since they were duplicates. Driven by users creating duplicate records of the same beer (identified via unique beer ID).
- Null Values – some of the records had features with null values (e.g. beer description, brewery city).
- Outliers – records (beers) with a rating of 0.0 (meaning that I forgot to rate the beer upon check-in).

CHALLENGES WITH DATA ACQUISITION

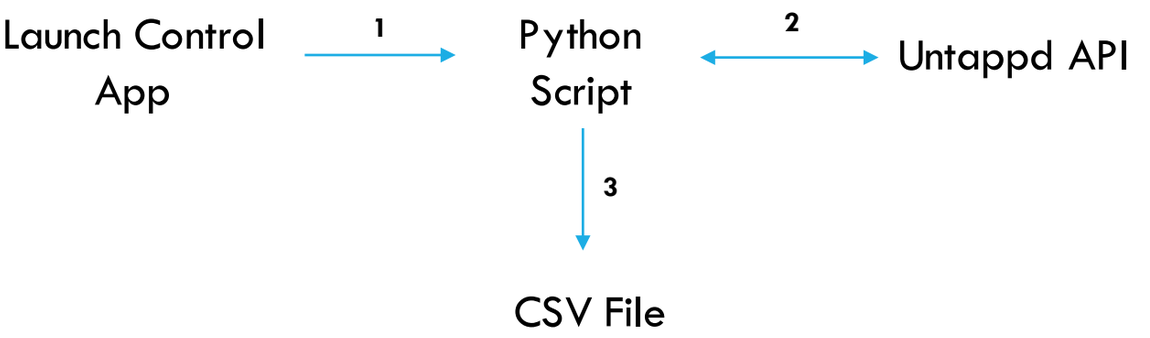
Many challenges were encountered and overcome during the data acquisition phase.

PRIMARY CHALLENGES

- The Untappd applications limits the number of API calls allowed to 100 per hour (required 300 hours to gather all of the data).
- Built scripts to acquire and load the data into CSV files hourly.
- A significant number of duplicate records returned from API calls had to be removed.
  - Believe the cause to be the fact that any user may create a new beer entry (record) even though it may already exist.

THE HOURLY PROCESS

The launch control app (1) hourly triggered a Python script to request 100 API calls, return the JSON data (2), and export the returned results to a CSV file (3).



INITIAL HYPOTHESIS

The initial hypothesis was to determine if there was a statistical difference between my personal ratings and the ratings of the average Untappd user.

TWO-SIDED T-TEST RESULTS

Null hypothesis: mean of my beer ratings is the same as the mean of average user beer ratings.

P-value result of 0.40, meaning I cannot reject the null hypothesis.

The result meant that my ratings did not differ significantly from the average user, which was a surprising result.

COMPARING MY RATINGS TO THE AVERAGE USER RATINGS*

	My Rating	Average User Rating
Unique Beer Count	181.00	181.00
Average Rating	3.39	3.43
Standard Deviation	0.64	0.39
Lowest Rating	1.50	2.31
25th Percentile	3.00	3.16
50th Percentile	3.50	3.53
75th Percentile	3.75	3.73
Highest Rating	5.00	4.56

**Rating scale is 1-5 with 5 being the highest score*

FURTHER ANALYSIS

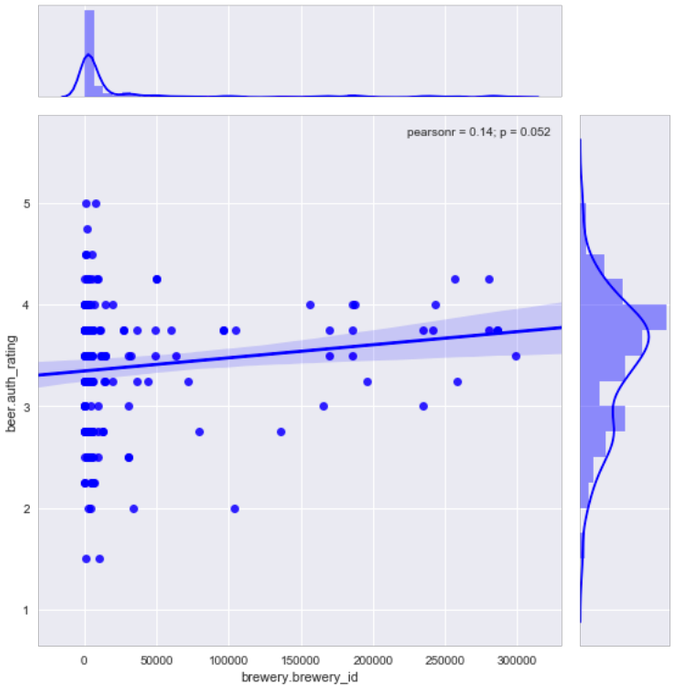
In addition to researching the initial hypothesis it was important to understand the potential impacts of other features in the primary dataset.

FEATURE ANALYSIS – R-SQUARED RESULTS

- Style of Beer: 0.69
- Alcohol by volume (ABV): 0.40
- Individual bitterness unit (IBU): 0.24
- Unique Brewery: 0.14

"Style of beer" was the feature with the highest correlation to beer rating with an R-Squared result of 0.69. This makes sense given that I prefer certain types of beer over others.

EXAMPLE: MY RATINGS VS BREWERY



Springboard

UNTAPPD
DRINK SOCIALLY

MACHINE LEARNING

After analyzing the various features the goal was to determine the best algorithm to use to predict a user's rating for a particular beer.

MACHINE LEARNING APPROACH

1) Test supervised learning algorithms with the key features: Linear Regression, Random Forest, and Gradient Boosting.

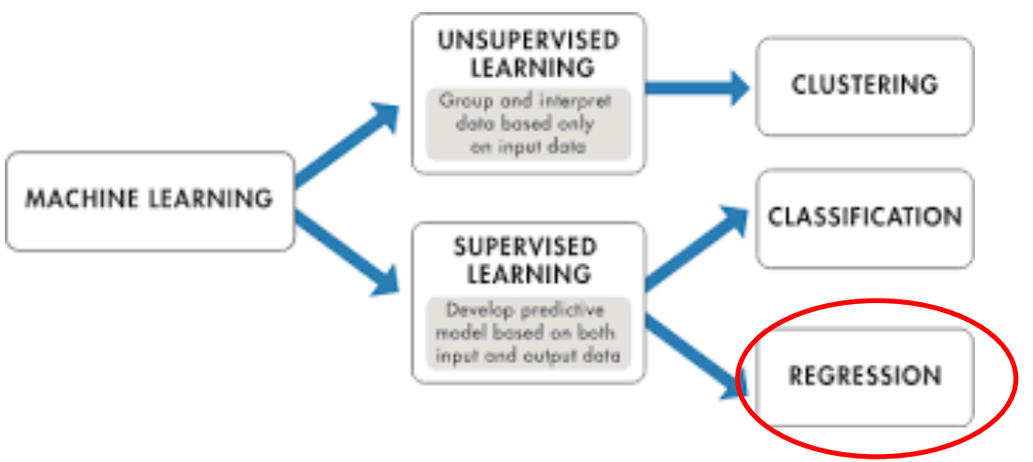
2) Leverage some of the feature analysis capabilities in the Random Forest algorithm to better understand the key features.

3) Understand model results and identify the best model.

4) Leverage the best model to create prediction ratings for the dataset of untasted beer.

FOCUS ON SUPERVISED LEARNING

Given the goal was to predict a specific numerical rating the focus was on supervised learning and regression algorithms.



Springboard

UNTAPPD
DRINK SOCIALLY

MACHINE LEARNING – MODEL RESULTS

Evaluating different regression algorithms proved that Random Forest Regression produced the best results.

	Train/Test Split	Cross-Validation (CV)	Model Score (1.0 is best)
Random Forest Regression	Yes	No	0.61
Gradient Boosting Regression	Yes	No	0.60
Linear Regression	Yes	Yes	0.48 *CV Score: 0.52*

Overall the Random Forest Regression produced the best results although Gradient Boosting Regression was a close second. Linear Regression was a distant third even with leveraging cross-validation.

Springboard

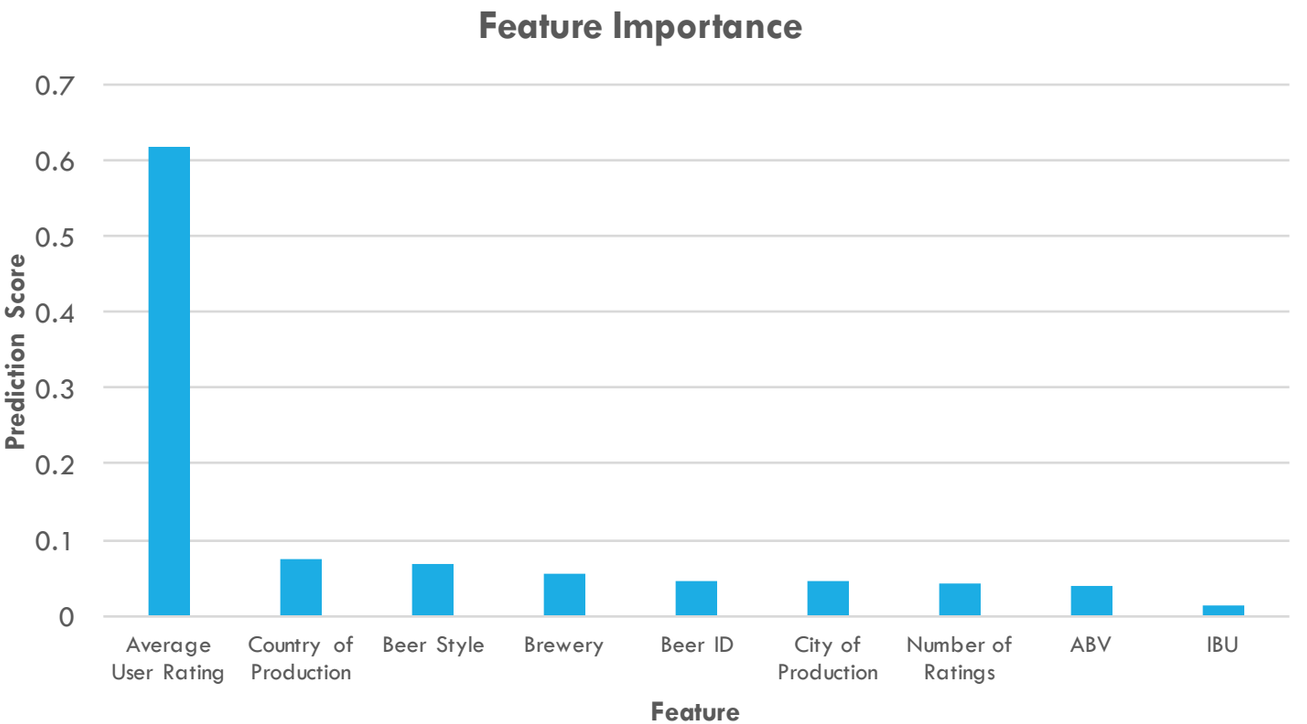
UNTAPPD
DRINK SOCIALLY

MACHINE LEARNING – FEATURE ANALYSIS

Leveraging some of the feature analysis capabilities of the Random Forest algorithm we see that average user rating is by far the best predictor of my individual rating.

FEATURES BY ORDER OF IMPORTANCE

1. Average User Rating
2. Country of Production
3. Beer Style
4. Brewery
5. Beer ID
6. City of Production
7. Number of Ratings
8. Alcohol by Volume (ABV)
9. Bitterness (IBU)



MACHINE LEARNING - PREDICTION

Implementing the newly built rating prediction model on the secondary dataset of almost 20,000 untasted beers shows that my top rated beer would be the Citra Single Hop Pale Ale from Hill Farmstead Brewery.

#1
Citra Single Hop Pale Ale

Brewery: Hill Farmstead
Location: Greensboro, NC
My Predicted Rating: 4.60
Average User Rating: 4.30

#2
Sucaba Barley Wine

Brewery: Firestone Walker
Location: Paso Robles, CA
My Predicted Rating: 4.50
Average User Rating: 4.36

#3
Angel's Share Bourbon Stout

Brewery: Hill Farmstead
Location: San Marcos, TX
My Predicted Rating: 4.48
Average User Rating: 4.29

#4
Rouge Sour Red Ale

Brewery: Brouwerij Omer Vander Ghinste
Location: Bellegem, Belgium
My Predicted Rating: 4.43
Average User Rating: 4.02

#5
Parabola Russian Stout

Brewery: Firestone Walker
Location: Paso Robles, CA
My Predicted Rating: 4.43
Average User Rating: 4.53

Springboard

UNTAPPD
DRINK SOCIALLY

RESULTS SUMMARY

KEY FINDINGS

- As users taste and rate more beer the impact of regression towards the mean is real. In this project this allowed the use of "average user rating" as a feature for model creation and prediction.
- Along with producing the best overall model results, Random Forest Regression was the simplest algorithm to practically use and understand.
- Given that the best model had a coefficient of determination (r-squared) of 0.61 there is still considerable room for improvement.
- Leveraging the best model to predict future beer ratings provided some surprising results: two of the top five beers suggested are from the same brewery and I am excited to seek out and try their beer!

Springboard

UNTAPPD
DRINK SOCIALLY

RECOMMENDATIONS & FUTURE IMPROVEMENTS

Key results show that the Untappd team could implement a prediction model to help users select their next beer.

PRIMARY RECOMMENDATIONS

- Implement a prediction model to help users select their next craft beer.
  - OK to use "average user rating" as a feature.
- Make the beer creation (record entry) process a bit more cumbersome for the user (e.g. require certain fields to have values).
- Scrub the production environment on a regular basis to remove duplicate records (beers).

FUTURE IMPROVEMENTS

- Leverage real-time, location-based data to help identify beers that may be immediately available to the user.
- Provide real-time access to API data.
- Run test/train model on a larger dataset to make a more robust prediction model.