Gaining Insight into the Craft Beer Industry:
What Makes A Successful Craft Beer?

Abstract: This paper investigates to what extent beer-related data on the ratings platform Untappd can inform us about which factors have the largest influence on consumer ratings and how tastes have changed over time. A random forest model was trained to predict the chance of a highly rated beer with 88.1% accuracy confirming the two most important predictors of rating were a high ABV and specific beer styles (i.e., Lambic). A Latent Dirichlet Allocation algorithm was also trained to identify topics of tasting notes, revealing that popularity for fruity and tropical tasting notes has grown rapidly since 2010 at the expense of more bitter beers.

I. Introduction

Over the past decade the face of the beer industry has been changing. Craft beer has grown rapidly at the expense of "Big Beer" – named after the handful of multinational companies that largely controlled production.

In America the two largest brewers, Anheuser-Busch Inbev and MillerCoors, have seen their combined market share decline from 78.3% in 2009 to 62.5% by 2019, whereas the share of craft beer has doubled[1].

While it is important for Big Beer to arrest their share declines, it is also important for small breweries to understand how they can capitalize on these developments to grow and differentiate themselves. One of the most successful craft beer companies – Brewdog – got their first initial break by winning two brewing competitions, shortly thereafter being stocked in Tesco[2].

The rationale often quoted for this trend is "changing consumer tastes"[3] as well as increasing support for local industry. To investigate how these tastes are changing I will source data from the "Untappd"[4] beer-rating website and mobile app, used by over 7 million consumers.

I believe my project is important for the two following stakeholder groups:

- Large brewers looking to understand how they can adapt their existing beer ranges to stay relevant with consumers.

- Smaller breweries needing to know which flavour profiles and beer styles have the greatest chance of gaining a successful rating from customers.

II. Analytical Questions and Data

A. Analytical Questions

My project aims to answer the following questions:

i. What are the components of a highly rated beer on Untappd – looking at ABV, beer style and tasting notes.

ii. How have consumer tastes for beer changed over time?

iii. To what extent can the information provided on Untappd be used to predict consumer ratings, and what does this tell us about the beer industry?

B. Key Characteristics of Dataset

The dataset has been web-scraped from the Untappd website containing 22,000 rows of beer data, across 9 columns. 6 columns are categorical - beer name, brewery, beer style, date added, availability and beer description. 3 are numerical - ABV, number of ratings received over time and consumer rating out of 5.

The categorical columns mostly exhibit high cardinality, with 3,892 different breweries and over 200 beer styles, excluding availability, which is a boolean column.

The ABV and number of ratings numerical data is highly positively skewed and impacted by extreme outliers, e.g., Guinness with 680,000 ratings (Fig.1):

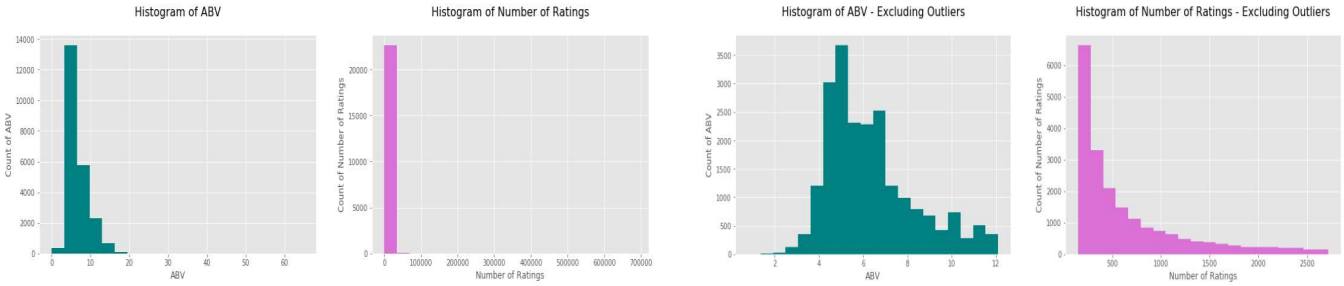

Fig. 1 ABV and Number of Ratings Distributions

The dataset is suitable for answering my research questions because Untappd is the most popular and comprehensive listing of beers, and their associated consumer ratings, in the world[5].

Because of the number of users of the platform, each brewery has an incentive to upload their own beer and ensure the data is both complete and accurate. Untappd is an independent entity and so should not have any incentive to alter or mis-present the data.

Demographics of users on Untappd are unlikely to reflect the real world, but are still largely representative of the beer industry's target market. The user's likely skew male, white[6] and with a preference towards craft beer over traditional beers.

I have made an assumption that a high rating on Untappd correlates with monetary gain, however this ignores other aspects behind commercial success such as advertising budgets.

III. Analysis

A. Data Preparation

The dataset has been constructed by web-scraping information about the top 50 rated beers in each beer-style category from the Untappd website. This prevented the database from becoming too large, as over 500,000 beers are added to Untappd each year, and is pertinent to the research question as it focuses only on successful beers and avoids obscure beers with few ratings.

I believe the beer descriptions may contain rich data that can inform my understanding of the individual predominant flavours of each beer and act as a relevant input when clustering and classifying the beers. I included beers from the UK, US and Australia to ensure the beer descriptions were all in English. I then dropped all 3,409 rows where the beer description was blank, leaving 19,373 rows.

To clean the beer descriptions, I first tokenized the words in each row and then removed any punctuation and numbers before changing all words to lower case, lemmatizing to remove plurals, and utilizing a modified list of stop words to retain only relevant words describing beer tasting notes.

To determine "success" I chose the top 20% of the beers rated within my dataset (which is already made up of just the top 50 beers in each beer style category). This cut-off resulted in beers rated >=3.96 out 5.00 being defined as successful, which correlates with my own domain knowledge from being a user on Untappd.

Since the ABV and number of ratings columns were impacted by severe outliers and neither was normally distributed, I trimmed the values that were more than 1.5 standard deviations from the mean.

B. Data Derivation & Construction of Models

With the cleaned beer descriptions, I created a corpus of words to be used by the Latent Dirichlet Allocation (LDA) algorithm[10], implemented utilizing Python's Gensim library. The purpose of this was to create a model to identify recurring "topics" within the corpus, with each topic representing a collection of dominant keywords. Standard LDA using Variational Bayes sampling was compared with Mallet LDA, using Gibb's sampling[11], to find the model with the best coherence score.

Before the best LDA model was applied, I split the data 70:30 into a training and test set. I did this as I plan to use the topics created as an input into a random forest model for predicting whether a beer will be rated as successful or not on Untappd. Therefore, the LDA model needs to be trained using only the training data, and then applied to the test set in order to generate test set topic values. A grid search was performed on the training data to find the optimal number of topics to maximise the coherence score of the model. See Fig. 2 below for the graph of the grid search:

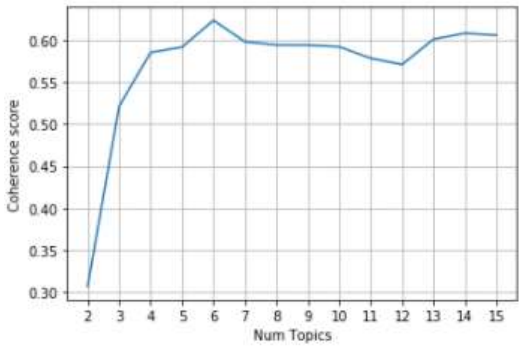

Fig. 2 LDA Grid Search for Optimal Topic Number (6)

I believe LDA is a better approach than using 1-of-C encoding for each tasting note, as it reduces dimensionality of the dataset, and thus training time for the model. It also produces a small list of topics that can be more easily interpreted and understood by users of the data.

I further derived additional data inputs by separating the date-added column into month and year columns in order to analyse any seasonal or temporal patterns in the data. I also created a condensed version of the beer style column, creating an amalgamated list of 19 styles to be more easily analysed.

Due to the number of categorical variables with high cardinality in the dataset, I chose to use the H2O implementation of the distributed random forest (DRF) classifier. The DRF maps categorical columns in lexicographic order with integer indices[7]. Random forest was chosen because it is a low bias, moderate variance model, which doesn't overfit easily[8]. It uses random sampling to be able to reduce model variance, building a forest out of an ensemble of decision trees where each tree then votes for the most popular target class given the feature inputs.

C. Validation of Results

The LDA topic output was validated in the following ways. Firstly, the coherence score measures the extent to which the topics made sense being grouped together. Secondly, they were visualized using multidimensional scaling to evaluate "distance", i.e. dissimilarity between topics (Fig. 3). A greater spread of topics indicates greater difference between each topic.

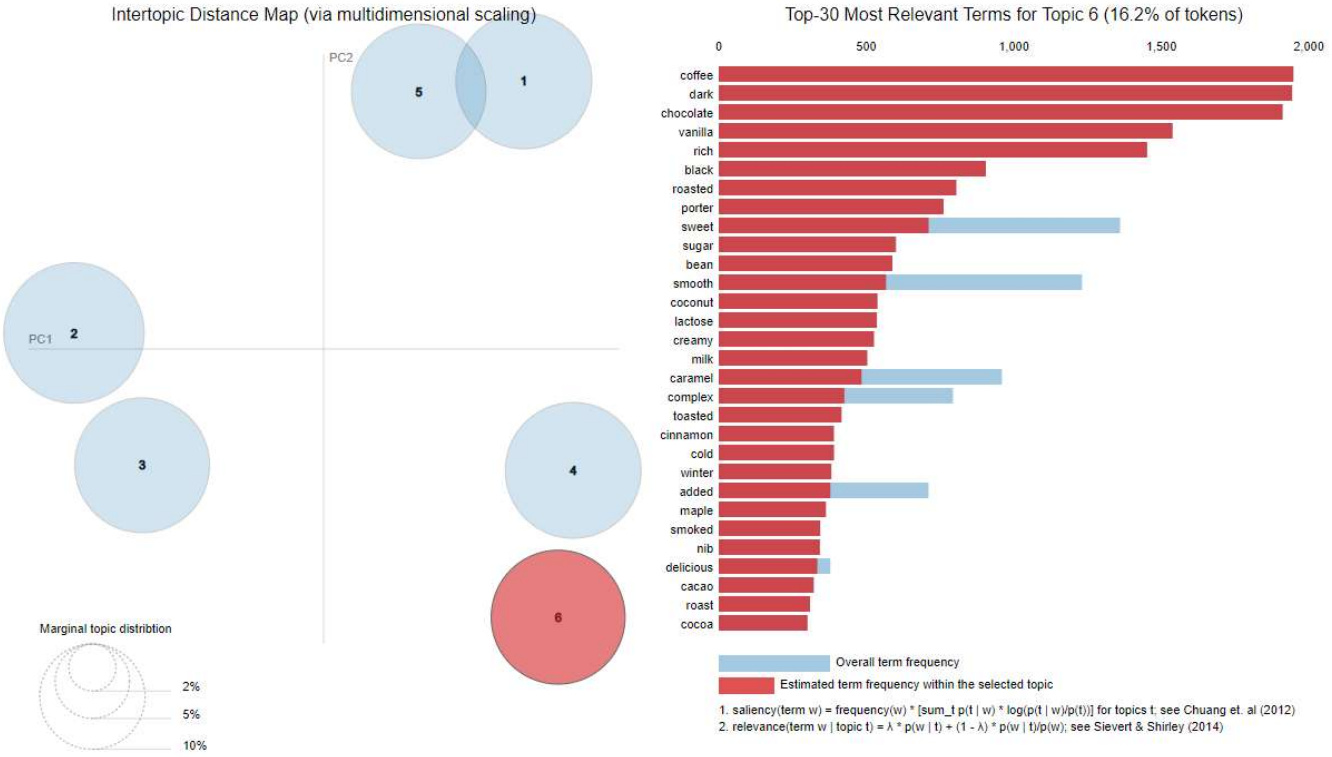

Fig. 3 Inter-topic Distance Visualization (Topic 6 highlighted)

I also compared flavour trends, represented by the topics, to hops acres harvested data, essentially comparing beer inputs (i.e. hops) to outputs (i.e. flavours). This provided some form of external validation to my results.

The random forest model was evaluated based on both accuracy and F1 scoring metrics. F1 score is important given the imbalanced nature of the data – 20% are considered successful. A confusion matrix was constructed to visualize model predictions and the result was compared to a baseline model that predicted every beer as unsuccessful.

Lastly, the impact of the addition of the topics column to the data was evaluated using a Student's T-Test measurement. This determined whether topics were a statistically significant model input.

IV. Findings, Reflections and Further Work

A. Findings

I found that ABV was an important component in making a beer rate highly on Untappd. After excluding outliers there is a clear positive correlation between a stronger ABV and a higher rating (Fig. 4), and this can be seen more dramatically when bucketing ABV into 5 equal quantiles (Fig. 5), where the chance of a rating above 3.96 increases from 3.8% for beers with an ABV of less than 4.7%, to 54.8% for high strength beers over 8.5% ABV.

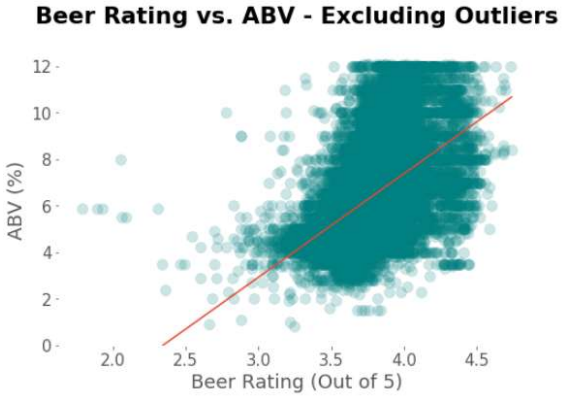

Fig. 4 Scatterplot of Beer Rating vs. ABV Excluding Outliers

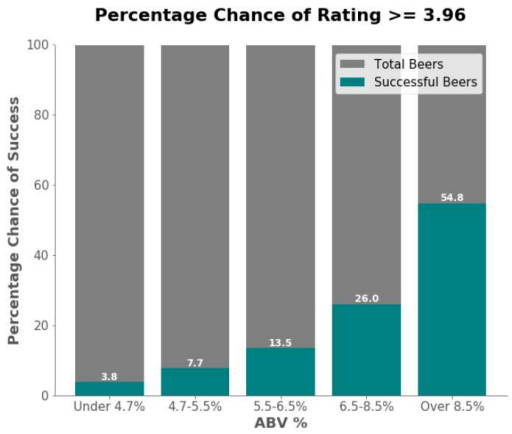

Fig. 5 ABV Quantile Percentage Chance of Success

Beer style was also an important predictor, with certain styles having a dramatically higher chance of success than others – i.e., Lambic, a wild fermented Belgian beer, has a 63.6% chance of success vs. lager, which has the lowest chance of rating highly at 1.5% (Fig. 6).

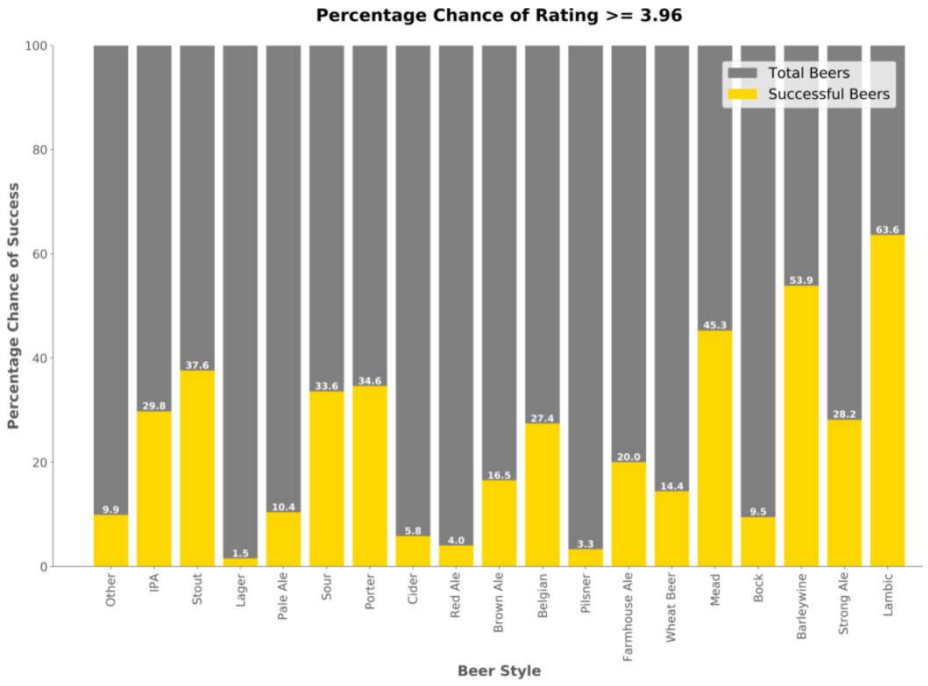

Fig. 6 Beer Style Percentage Chance of Success

There is a large disparity between the number of beers being brewed compared to chance of success, as some of the most successful types were some of the least frequently made (Fig.7). This likely reflects the fact that some beers are easier or quicker to brew than others, however some breweries may have the time and expertise to invest in perfecting these styles.

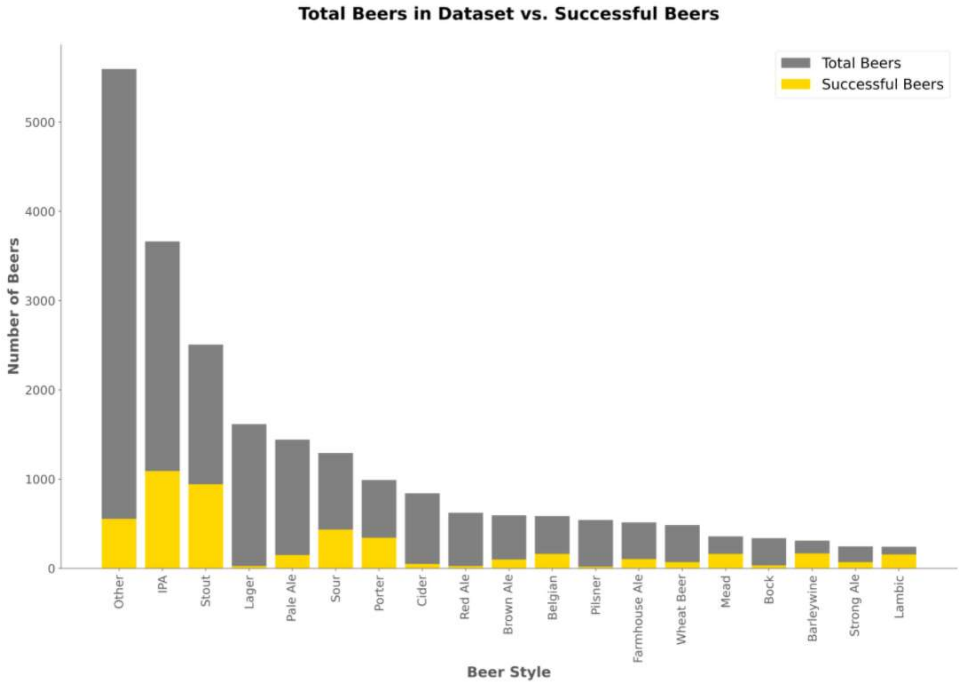

Fig. 7 Success Showing Count by Beer Style

The LDA mallet model outperformed the standard model, increasing coherence of the topics from 36.6 to 59.2. The optimal number of topics found to maximize coherence was six (Fig. 2). Here are the dominant keywords for each topic (Fig. 8):

Fig. 8 Topic Wordclouds displaying the top 100 words from each topic according to the LDA Model

Plotting the frequency of these topics over time gives us insight into how consumer tastes have changed:

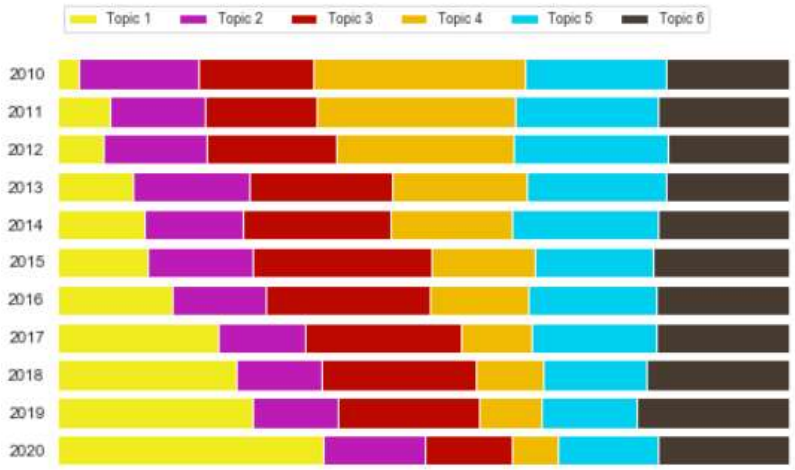

Fig. 9 Annual Count of Topics in Beer Descriptions

We can see that Topic 1, representing fruity, tropical, dry hopped beers has increased rapidly since 2010, largely to the detriment of Topic 4 - bitter, orange, spiced beers (Fig. 9). Knowing that hops are used to develop these characteristic flavours I validated these findings using hops acres harvested data from the USDA[9] (Fig.10).

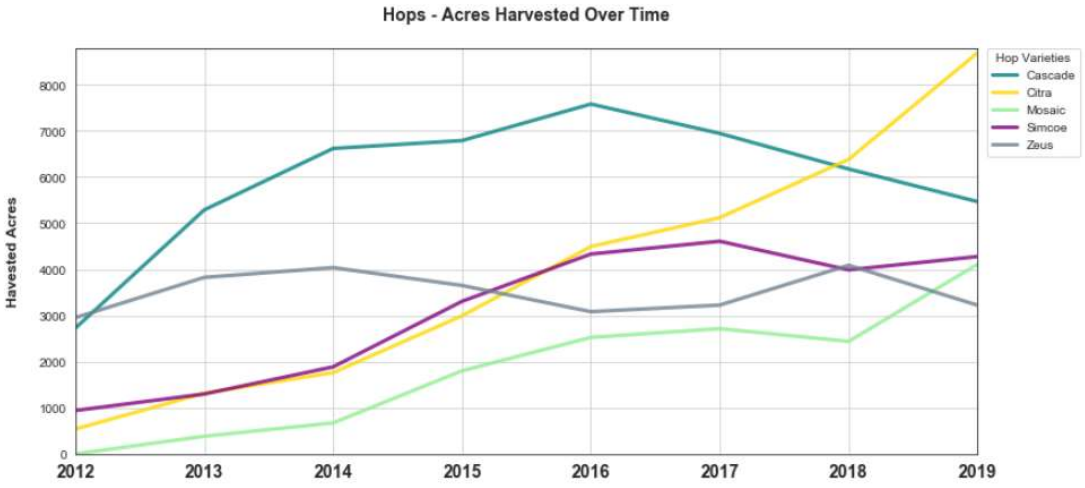

Fig. 10 Line Chart of Hops Acres Harvested Over Time

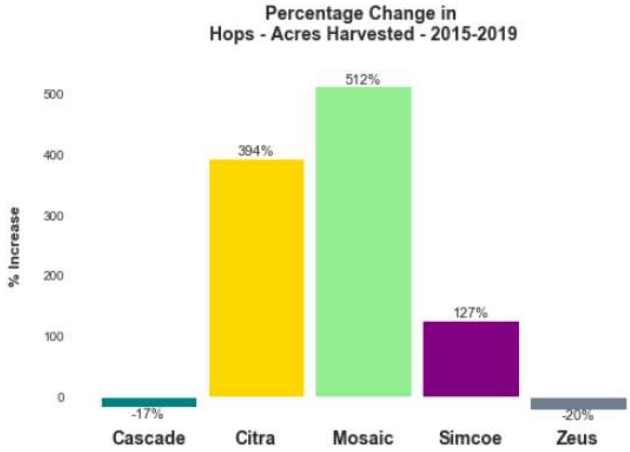

Fig. 11 Five Year Percentage Change in Hops Acres Harvested

These charts (Fig. 11) demonstrate that farmers have responded to consumer demands for more fruity tastes by planting more of the Mosaic and Citra hops, known for these flavour profiles, and less Cascade and Zeus hops, known

to produce more bitter tasting beer. This information can therefore aid brewers as well as farmers by signposting which popular hops to use/plant.

The random forest model was able to predict success on the test set with 88.1% accuracy, outperforming the baseline model by 13.8%. The F1 score of 75.6% also showed the model was successful at identifying both positive and negative target classes:

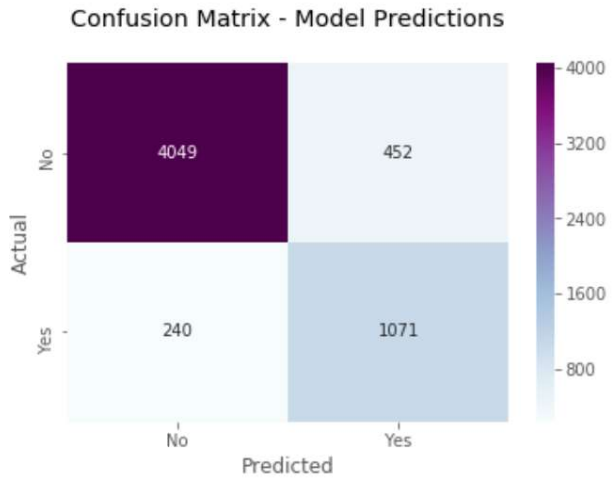

Fig. 12 Random Forest Confusion Matrix

The most important predictors were shown to be:

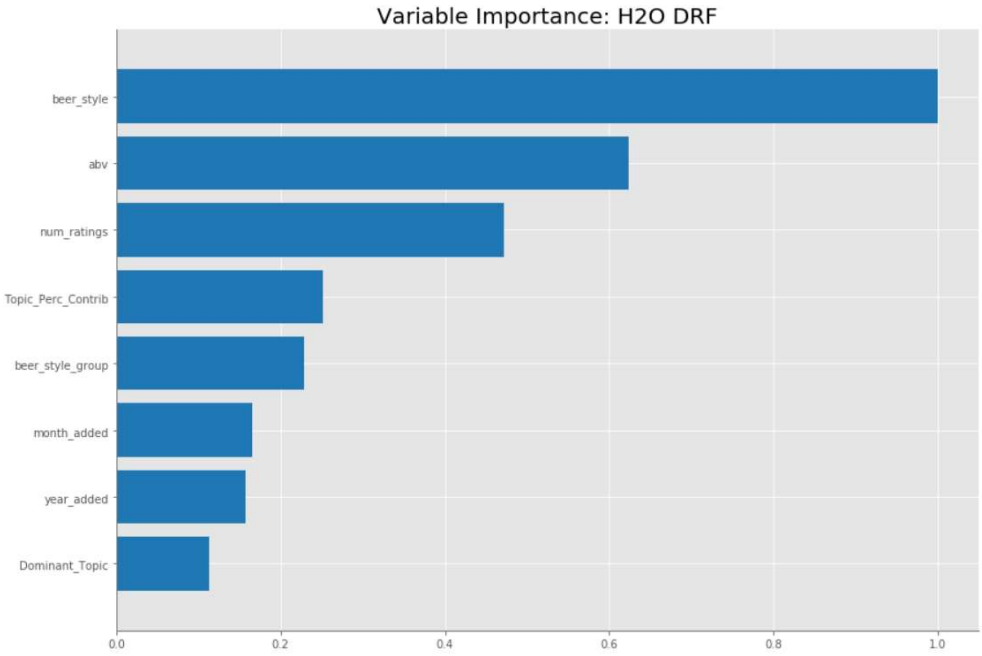

Fig. 13 Variable Importance as Measured by the H20 Random Forest Model

The accuracy of the model and importance of ABV and beer style tells us that there is a high level of predictability to the beer industry and to the types generally preferred by consumers on Untappd.

Despite "Topic" not being an important input, "Topic Percentage Contribution" was. This signifies how well each topic is represented by the words used, showing that a clear description of flavours on a beer label can positively impact a consumers overall rating (see Figure 14 below).

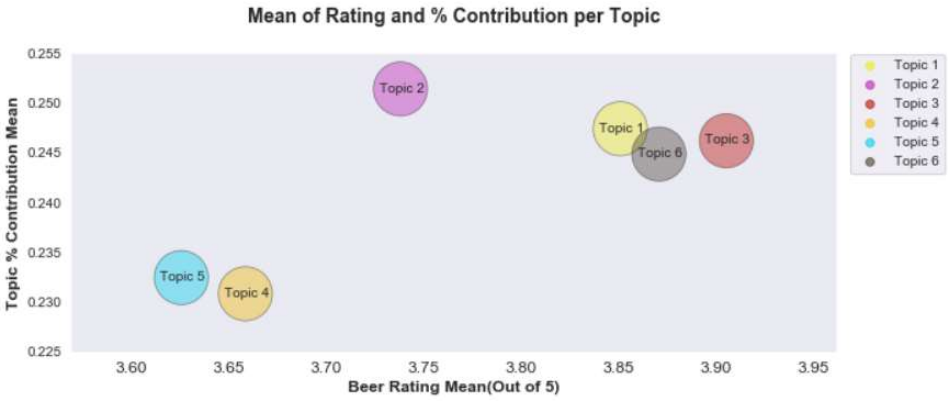

Fig. 14 Scatterplot showing the correlation between average topic percentage contribution and overall rating

B. Suitability and Further Work

The data was suitable for the level of analysis required for this project, however, to investigate deeper it would be important to understand the relative costs and times of producing beers of various styles and ABV's to evaluate commercial viability of the most successful combinations.

For a more comprehensive analysis the data could be sourced from non-English speaking countries, and expanded beyond the top 50 highest rated beers in each category.

To build on my topic work, certain keywords associated with success, such as 'barrel' and 'aged' could be 1-of-C encoded and fed into the existing or a more complex neural network model to potentially improve the overall accuracy of the model.

Scraping the data at a user level would provide more granularity on when and where beers are rated. This would allow for a deeper temporal and spatial understanding of the data.

REFERENCES

[1] NBWA: America's Beer Distributors. "Industry Fast Facts," January 6, 2015. https://www.nbwa.org/resources/industry-fast-facts.

[2] Headspace. "Entrepreneurs: How BrewDog Started from Nothing," March 21, 2019. https://www.headspacegroup.co.uk/entrepreneurs-how-brewdog-started-from-nothing/.

[3] Thompson, Derek. "Craft Beer Is the Strangest, Happiest Economic Story in America." The Atlantic, January 19, 2018. https://www.theatlantic.com/business/archive/2018/01/craft-beer-industry/550850/.

[4] Team, The Untappd. "Untappd." Untappd. Accessed December 6, 2020. https://untappd.com/home.

[5] "Untappd." Untappd. Accessed December 6, 2020. https://untappd.com/blog/post/181615466178/year-in-beer-2018.

[6] Brewers Association. "Shifting Demographics Among Craft Drinkers," June 12, 2018. https://www.brewersassociation.org/insights/shifting-demographics-among-craft-drinkers/.

[7] "Distributed Random Forest (DRF) — H2O 3.32.0.2 Documentation." Accessed December 6, 2020. http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html.

[8] Brieman, Leo (2001), Random Forests, Machine Learning 45, 5-32.

[9] "Publication | National Hop Report | ID: S7526c41m | USDA Economics, Statistics and Market Information System." Accessed December 6, 2020. https://usda.library.cornell.edu/concern/publications/s7526c41m.

[10] Blei, David, Ng, Andrew, Jordan, Michael. *Latent Dirichlet Allocation,* Journal of Machine Learning Research 3 (2003) 993-1022, Submitted 2/02; Published 1/03.

[11] Steyvers, Mark, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. "Probabilistic Author-Topic Models for Information Discovery." In *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04*, 306. Seattle, WA, USA: ACM Press, 2004. https://doi.org/10.1145/1014052.1014087.