Daniel Marchese |Daniel.marchese@okstate.edu| LinkedIn | Twitter |GitHub

Portfolio Project

Jupyter Notebook for this model

**Beer Rating Prediction Model**

**Introduction**

This report details creating a model for predicting beer rating, starting from the data stored in my database. This would allow brewers to predict the reception of beer ideas without having to brew and sell their idea. This will also give us some insight into how current beers are performing against the field, if a beer is rating higher than its prediction interval then that beer has a special feature that is not accounted for and worth looking into. Perhaps where this may have the most use is selling beer to a bar or store, being able to show that a new beer will be well received. Or on the buying side take Dan's Bar, which prides itself for having the hottest new beers, instead of having to rely on his intuition this predictive model will aid Dan in selecting new beers to put on draft.

**Data**

Data is scraped daily from untappd.com, a website where users self-report beer consumption, their personal rating, and can leave comments. These self-reports will be referred to as check-ins. All check-ins from the previous day are scraped each morning for each Brewery on the Columbus Ale Trail[1]. This project will use data from July and August 2022, a total of 53,355 check-ins. Individual beers each have their own page on untappd that includes relevant information on the beers. New beers are scraped daily as they are released and are stored as a dimension table.

**Data Preparation**

*Create and aggregated beer table*

- Check-in fact table merged with Beer dimension, keeping columns: Style, ABV%, Description, IBU, and Beer Link.
- Style is split into 2 features, Style_1 and Style_2
    - The Style column contains both the main style of a beer and the secondary style. For example IPA – New England
- Beers without a Description are dropped.
    - Beers without a description are usually the result of a check-in being done incorrectly and adding a new beer, untappd merges these wild beers but it's not perfect. This drops the data down to 46,211 observations. I'm not worried about removing this data if it is missing completely at random, and I recommend investigating these observations further.
- The merged table is grouped by Beer, aggregating Rating by mean and Review_ID by count.

[1] Cbusaletrail.com , also note that The Understory is on the Ale Trail but is NOT a brewery, and has no available data.

*Remove outliers from target variable, Rating*

Figure 1 is the scatter plot of check-in count to mean rating.

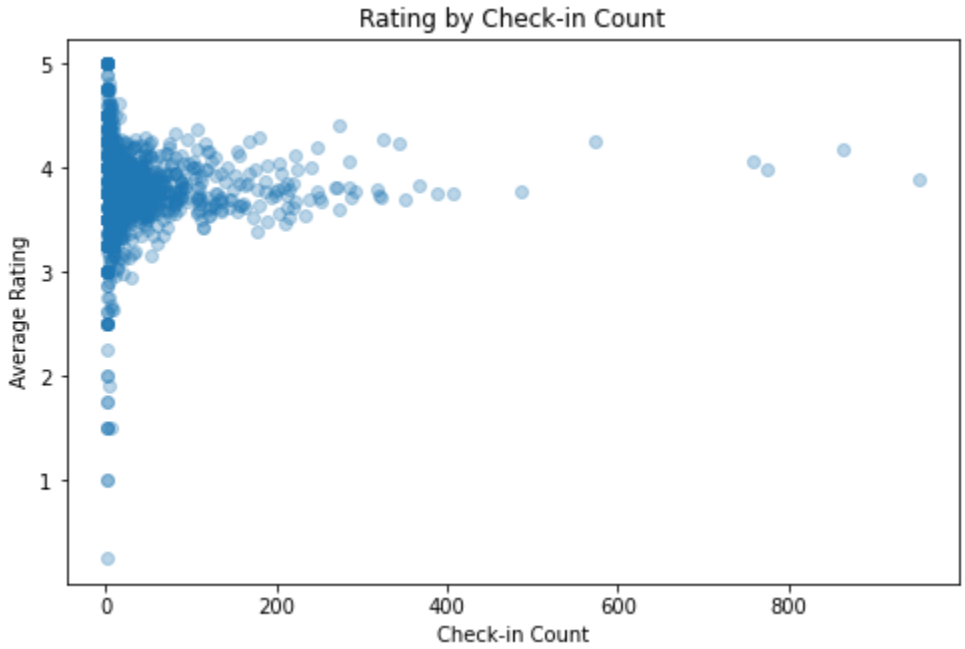


*Figure 1: Rating x Count python output*

There is a solid section of where we would expect beers made by established craft breweries to fall, but then there are points below 3 and above 4.5 that don't seem correct. They also all have very few total check-ins. Outliers were identified as any rating outside of the inner quartile boundaries ± 1.5 * IQR. In addition to identifying outliers, I set a floor of 10 check-ins to ensure beers had enough reviews to return a reliable rating. This leaves us with 638 unique beers. This will increase as more data is collected and more beers make it over the 10 check-in floor.

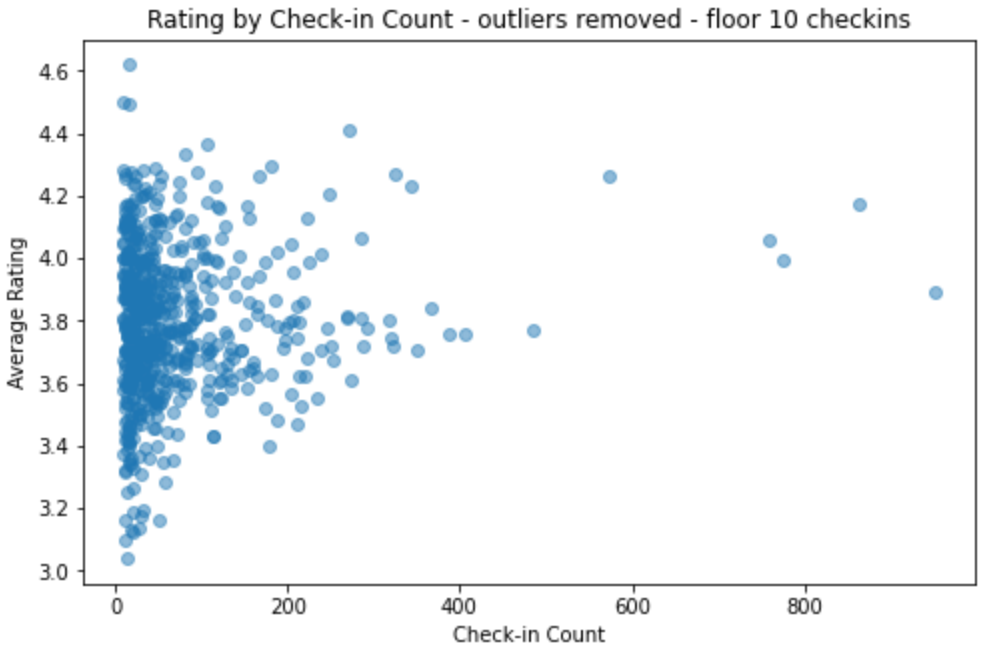


*Figure 2: Target Variable Rating with outliers removed*

*Description and Styles converted into binary features*

For each beer I broke the description into tokens with their part of speech tags. Each word that was a noun or an adjective was lemmatized and collected for each beer. If the lemma of a noun and adjective were the same, it was only included once. These lemmata were counted and the top 115 were created into features. The goal is for these features to accurately represent the beer, especially its taste. To achieve this, I removed words that were popular in beer descriptions but did not describe taste. The 115 chosen features are in the Appendix along with the removed words.

The categorical variables, Style_1 and Style_2, were dummy coded into binary flag variables.

*Numeric variable transformation*

Numeric features, ABV% and IBU, are not as normally distributed as we would hope, see Figure 3.

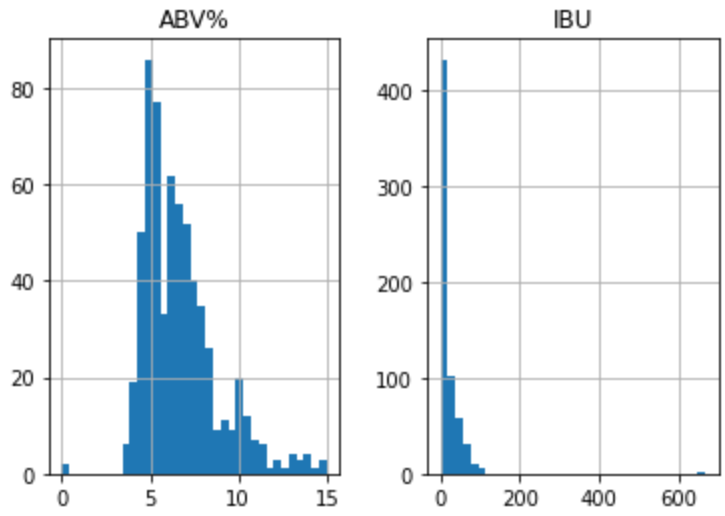


*Figure 3: Histograms of Numeric Features*

There is right skew with both distributions. 2 beers have and ABV% of 0 and 1 beer has over 600 IBUs. These values were investigated, and the correct values were obtained and updated in the database.

I transformed these features and standardized them; the resulting distributions are in Figure 4. Yeo-Johnson was used for IBU because of the 0 values and Box-Cox was used to transform ABV%.

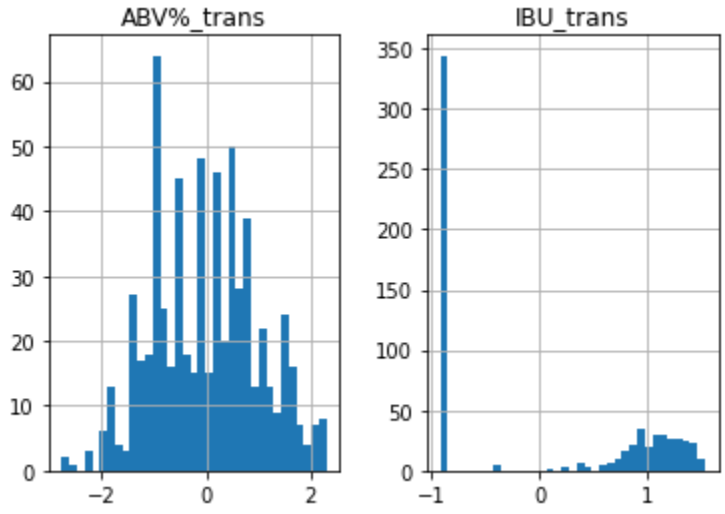


*Figure 4: Transformed ABV% and IBU*

IBU is in a tough spot, the 0 values are correct and naturally occurring and making it tough to get a good distribution. I tested a log2(x+1) transformation, but it was not as good as the Yeo-Johnson formula.

| | Skew | Kurtosis |
|---|---|---|
| ABV% | 1.35 | 2 |
| ABV% Transformed | 0.06 | -0.56 |
| IBU | 1.56 | 1.9 |
| IBU Transformed | 0.25 | -1.83 |

*Feature Selection*

I used two methods of feature selection, a recursive feature eliminator, and a random forest regression model. The recursive feature eliminator was set to select half of the available features, removing 2 at a time. A random forest regressor was trained and tested on all features and to get feature significance. I set and optimized a threshold for feature importance by finding the set of features where the R2 score RFE on a random forest and a random forest using filtered features were about equivalent. This resulted in keeping features with a feature importance greater than 0.002. From these features I further removed four features that had correlations higher than 0.5 with another feature. I removed the feature with lower feature importance of the highly correlated pair.

Note that because there are only 638 total beers in the dataset modeling is done using cross validation and R-square is the mean score of 5 splits.

**Modeling**

Four models were used, Decision Tree, Random Forest, Gradient Boosting, and Linear Regression. Models were tested on the full set of features and the filtered set of features. The best two performing models are Linear Regression and Random Forest on the filtered set of features. I combined these two models taking the average of their predictions, called Voting Regressor in Scikit-learn, which outperformed both. I recommend continuing to tune the hyperparameters for model improvement.

| Model | R2 |
|---|---|
| Random Forest (all features) | 0.179 |
| Random Forest (filtered) | 0.2135 |
| Linear Regression (filtered) | 0.2429 |
| Decision Tree (all features) | 0.0292 |
| Decision Tree (filtered) | 0.1028 |
| Gradient Boost (all features) | 0.1943 |
| Voting Regressor (Random Forest + Linear Reg) | 0.2537 |

Residuals are plotted below in Figure 5, looking very much Gaussian.

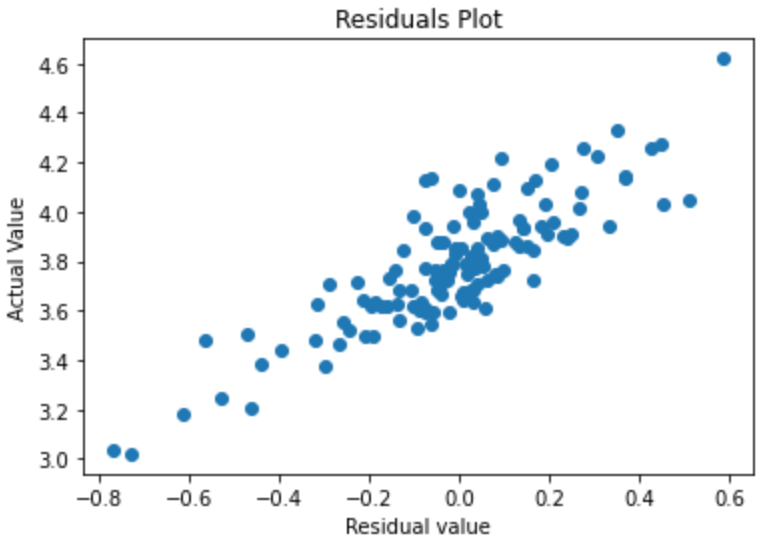


*Figure 5: Residuals Plot*

**Conclusion and next steps**

This model achieved 0.2537 R-square value, which means it's able to account for about 25% of the variance in the data. This is a great start, there are going to be features that are very tough to pick up on, I imagine each brewery has something figured out that works well for the brewery. The brewery special sauce, that maybe in time I will pick up on. My point being there will always be some variance among 33 different breweries making beers. Likewise, these are self-reported ratings by fickle humans with different taste buds.

One way I will be using this model is to help identify how individual breweries are comparing against the field. The model will be trained on all beers except ones made by the brewery up for analysis, then the model will predict those beers and compare, see Figure 6. This will be the basis for some statistical testing and analysis.

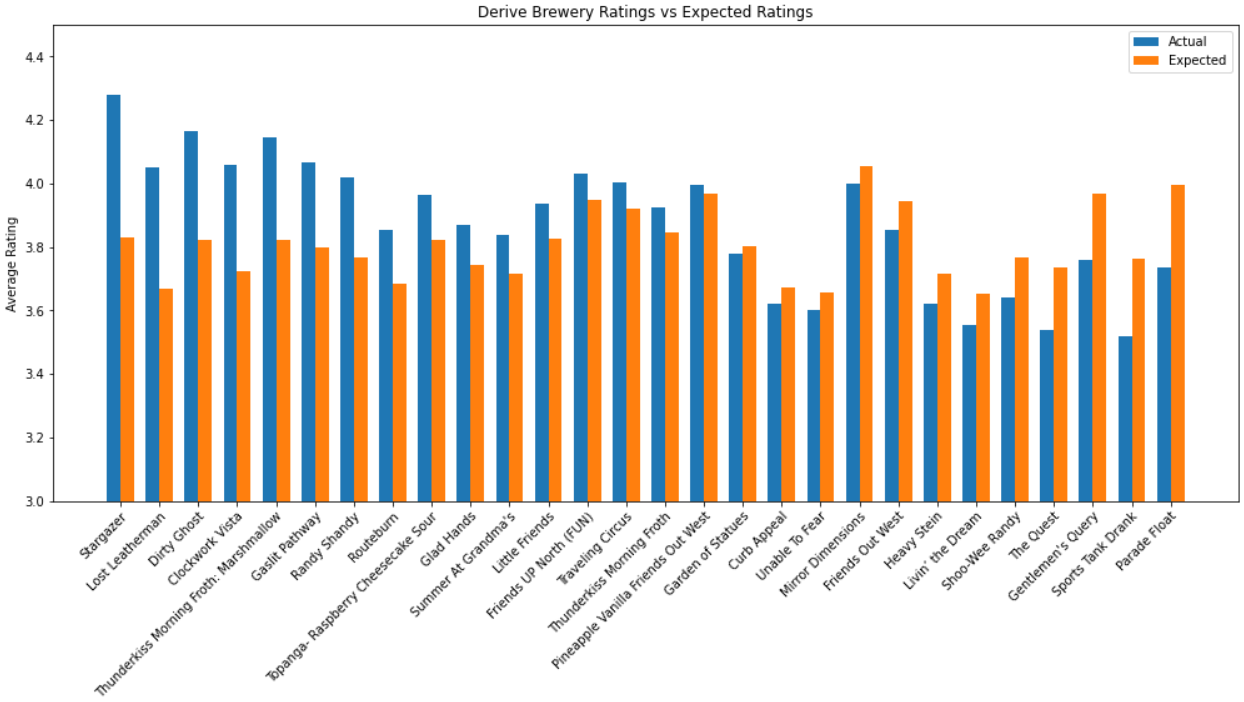


*Figure 6: Example use of Model*

## Appendix

### All Feature words

['fruit', 'vanilla', 'malt', 'hop', 'dry', 'sour', 'citra', 'coffee', 'orange', 'juice', 'tropical', 'wheat', 'american',
'hazy', 'imperial', 'pale', 'citrus', 'sweet', 'double', 'smooth', 'german', 'mosaic', 'pineapple', 'chocolate',
'golden', 'lime', 'rich', 'clean', 'tart', 'bitterness', 'classic', 'simcoe', 'cinnamon', 'lemon', 'bourbon',
'grapefruit', 'caramel', 'coconut', 'soft', 'big', 'milk', 'peach', 'subtle', 'hoppy', 'banana', 'columbus',
'amarillo', 'cream', 'dry-hopped', 'red', 'beans', 'tangerine', 'belgian', 'mango', 'raspberry', 'floral',
'bright', 'oats', 'dank', 'pine', 'strawberry', 'stone', 'full', 'bitter', 'nelson', 'cacao', 'citrusy', 'triple', 'west',
'refreshing', 'passion', 'sauvin', 'rye', 'ripe', 'traditional', 'zest', 'england', 'coriander', 'taste',
'white', 'sweetness', 'gold', 'touch', 'lactose', 'deep', 'complex', 'galaxy', 'blueberry', 'craft', 'creamy',
'super', 'purée', 'zealand', 'line', 'roast', 'kettle', 'cocoa', 'coast', 'clove', 'sabro', 'salt', 'balanced',
'india', 'oak', 'mouthfeel', 'amounts', 'take', 'brettanomyces', 'ice', 'low', 'cherry', 'lachancea', 'hints']

### Stop Words

['show','beer','ipa','stout','ipa','notes','flavor','lager','barrels','barrel','%','finish','flavors','style',\
'aroma',"'",'pilsner','yeast','blend','abv','sugar','malts','time','year','light','summer','months'\
,'brew','ale','’','dorado','friends','collaboration','s','dipa','base','drinking','body','brewing'\
,'character','blonde','peel','hint','day','nibs','ohio','aromas','batch'\
,'color','version','wolf','name','w/','crisp','addition','amount','brewer','brewery','bean','puree'\
,'-
','baseball','cap','neighbor','result','bat','tip','ash','night','reward','dark','reward','ecuador','hue','spectrum','haze'\
,'course','ghana','sandy','sunburn','land','grain','toe','week','grant','cheer','–'\
,'new','great','perfect','ale','less','favorite','little','first','more','delicious','good'\
,'el','pounds','hoof','years','fresh','real','black','farmhouse','seltzer','easy','kolsch','profile','leather','life'\
,'rugger','crystal','additions','long','medusa', 'hops','jesus', '*']

### Final Selected Features

['fruit', 'vanilla', 'malt', 'hop', 'dry', 'sour', 'citra', 'coffee', 'american', 'hazy', 'imperial', 'double', 'pineapple', 'rich',  'clean', 'bitterness', 'bourbon', 'caramel', 'cream', 'belgian', 'triple', 'traditional', 'ABV%_trans',  'IBU_trans', 'Style_1_IPA', 'Style_1_Lager', 'Style_1_Pilsner', 'Style_1_Shandy / Radler', 'Style_1_Stout', 'Style_1_Wheat Beer', 'Style_2_American', 'Style_2_American Light', 'Style_2_Imperial / Double', 'Style_2_Imperial / Double New England / Hazy', 'Style_2_Other', 'Style_2_Russian Imperial']

Note: Citra and American are the names of specific Hop varieties used in brewing beer.