Craft Beer Classification

Samantha Jarrett - GTID: 903743596

ISYE7406 Final Project

20 November 2023

Table of Contents

ABSTRACT

The purpose of this study is to investigate the characteristics of craft beer to classify different beer types by style and origin. Gradient Boosting models were implemented resulting in 90% accuracy in Style Classification and 80% accuracy in Origin Classification. Using these classifications, a craft beer recommendation system could be implemented to assist beer drinkers in selecting the best options based on their desired beer tasting profile.

INTRODUCTION

In 2021, total United States beer sales volume increased by 1%, but craft beer sales volume increased by 8%. Craft beer market share increased by almost 1%. Across more than 9,500 craft breweries in the United States, 24.8 million barrels were brewed. It is now estimated that most Americans live within 10 miles of an independent craft brewer. The total beer industry supports close to 23.1 million jobs worldwide [1]. Craft beer has been on the rise for the past three decades and is a very popular libation to many Americans. There are many different characteristic styles of craft beers.

The aim of this study is to analyze these characteristics to attempt to classify different craft beer styles and origins. These classifiers could be used to implement a recommendation system to point craft beer consumers towards craft beers they would like. The input to the system would be profile characteristic ratings provided by the user and the output would be a style and/or origin of beer that matches the desired profile.

DATASET

The Beer Profile and Ratings dataset used for this study was obtained through Kaggle. The Kaggle dataset was put together using datasets mentioned in the dataset acknowledgements. The dataset includes 3197 distinct beers from 934 different breweries, their alcohol content (ABV), International Bitterness Unit (IBU), and eleven tasting profile characteristics. The tasting profile characteristics were defined using word counts in beer reviews [2]. The characteristics are listed in Table 1.

Profile	**Characteristic**
Flavor and Aroma	Fruity
	Hoppy
	Malty
	Spices
Mouthfeel	Alcohol (perceived booziness)
	Astringency
	Body
Taste	Bitter
	Salty
	Sour

	Sweet

Table 1. Beer Profiles and Characteristics

Since the dataset includes so many different styles of beer, the Brewers Association of Beer Style Guide [3] was referenced to classify each beer by style (ale vs. lager) and origin. Note that unique, hybrid beers such as pumpkin ales, Japanese rice lagers, and rye beers were excluded from the dataset as they do not typically belong to the main style and origin categories.

EXPLORATORY DATA ANALYSIS

The distributions of the features were examined to understand their ranges and concentrations. It was noted that some variables have very wide distributions with high variation, such as sweet, sour, and malty. Others have more narrow distributions with less variation, such as abv and salty (Figure 1). It is likely that the variables with wider distributions and higher variation will be more important in predicting the class of the beers.

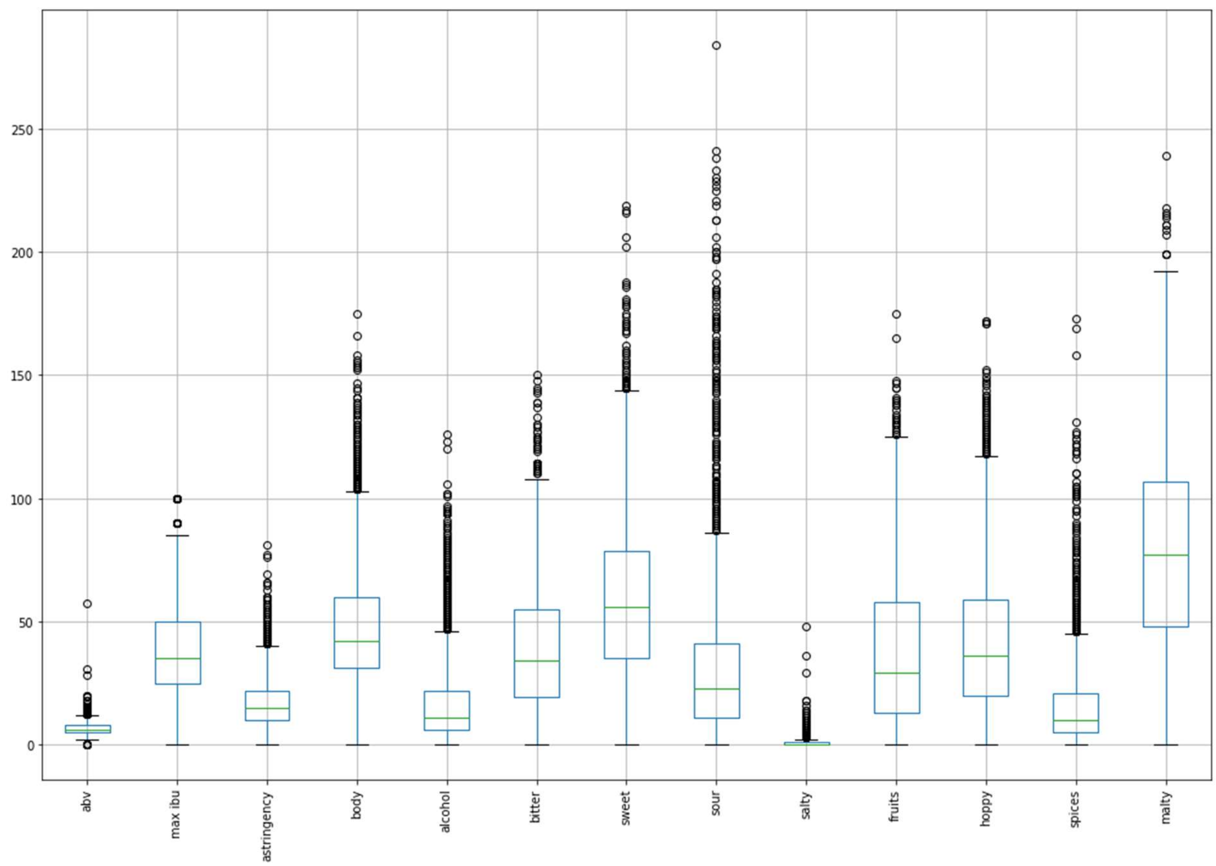


Figure 1. Boxplot of Features

Style Classification

The relationships between the features and the two classes were also examined. It was noted that the features fruits, sour, body, and sweet showed more distinctly different distributions between the two style classes, indicating that they are likely important in predicting if the beer is an ale or a lager. The ales seem to tend to have higher fruits, sour, body, and sweet values than the lagers (Table 2).

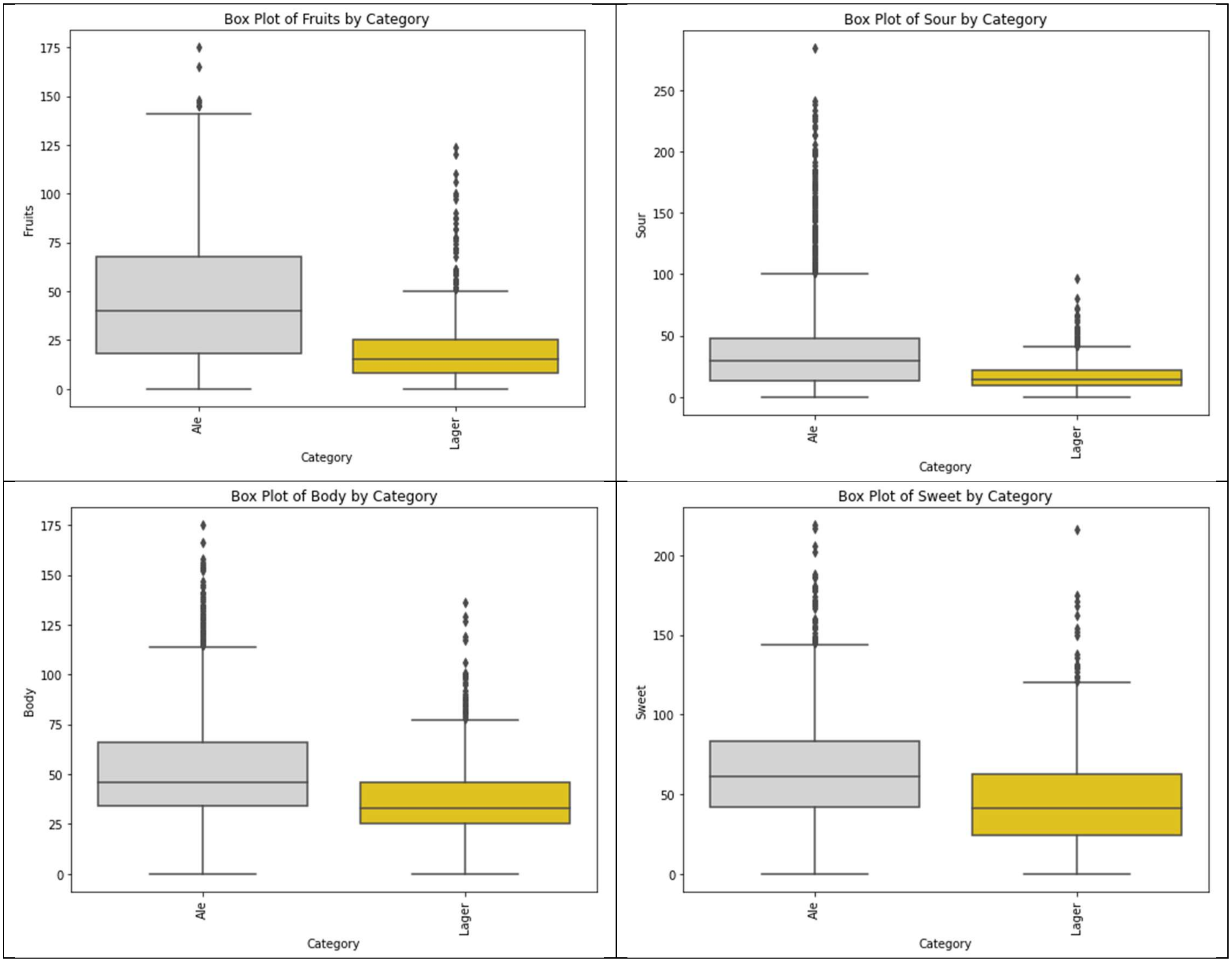


Table 2. Style Classification Boxplots

Origin Classification

The distribution of samples for each craft beer origin is presented in Figure 2. Since the Other Origin Lager and Ale classifications are small and do not provide much added value, they were excluded from the dataset. It is also noted that the German Origin Ale, North American Origin Lager, and Irish Origin Ale have smaller sample size than the remaining origins. This may make them more difficult to classify.

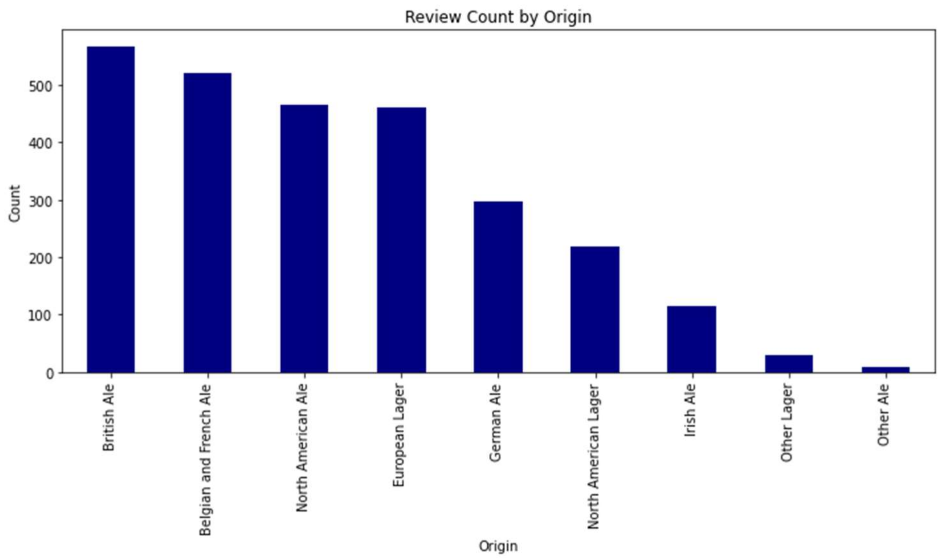
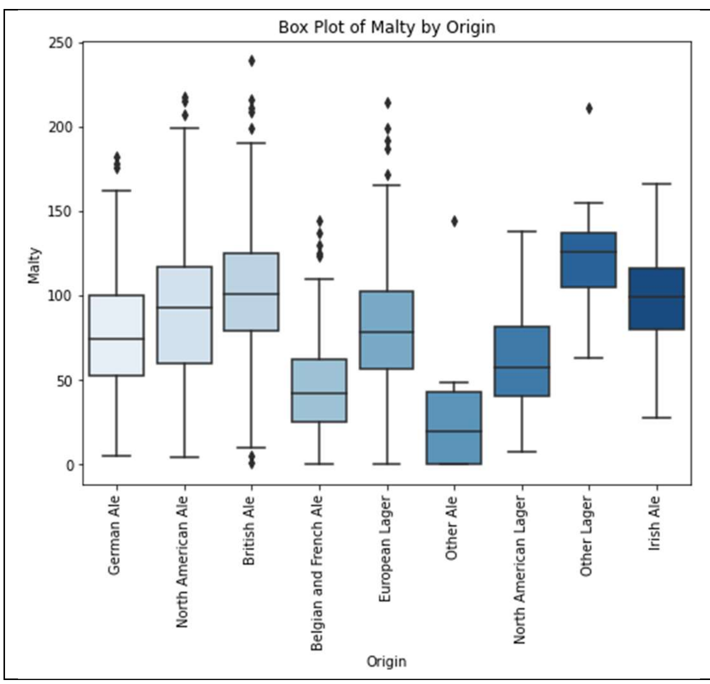
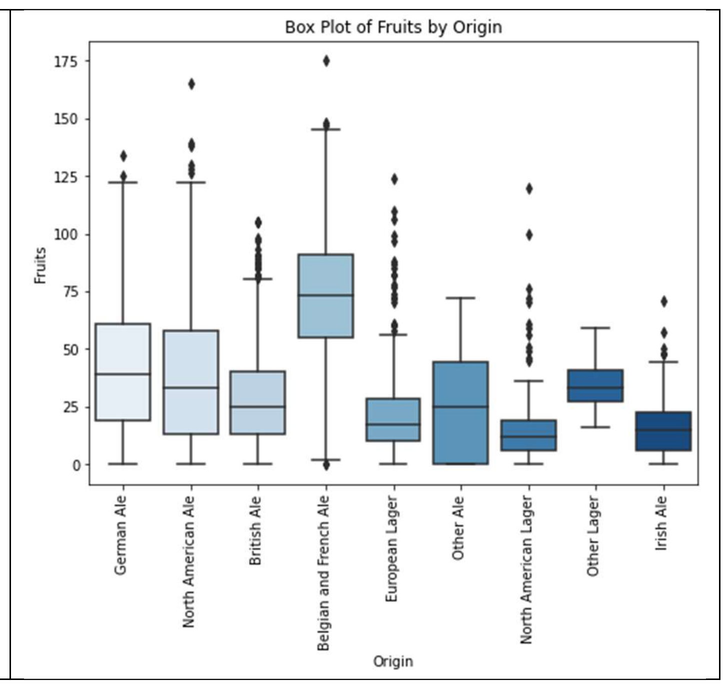


Figure 2. Distribution of Samples by Origin

Box plots of the feature characteristics by origin show that the features of malty, fruits, sour, and bitter have the most differences between origins. This indicates that these characteristics may be important in classifying beers by origin. Particularly, it is noted that Belgian and French Origin Ales tend to have higher fruits and sour scores, while North American and British and Irish Origin Ales tend to have higher bitter scores (Table 3).





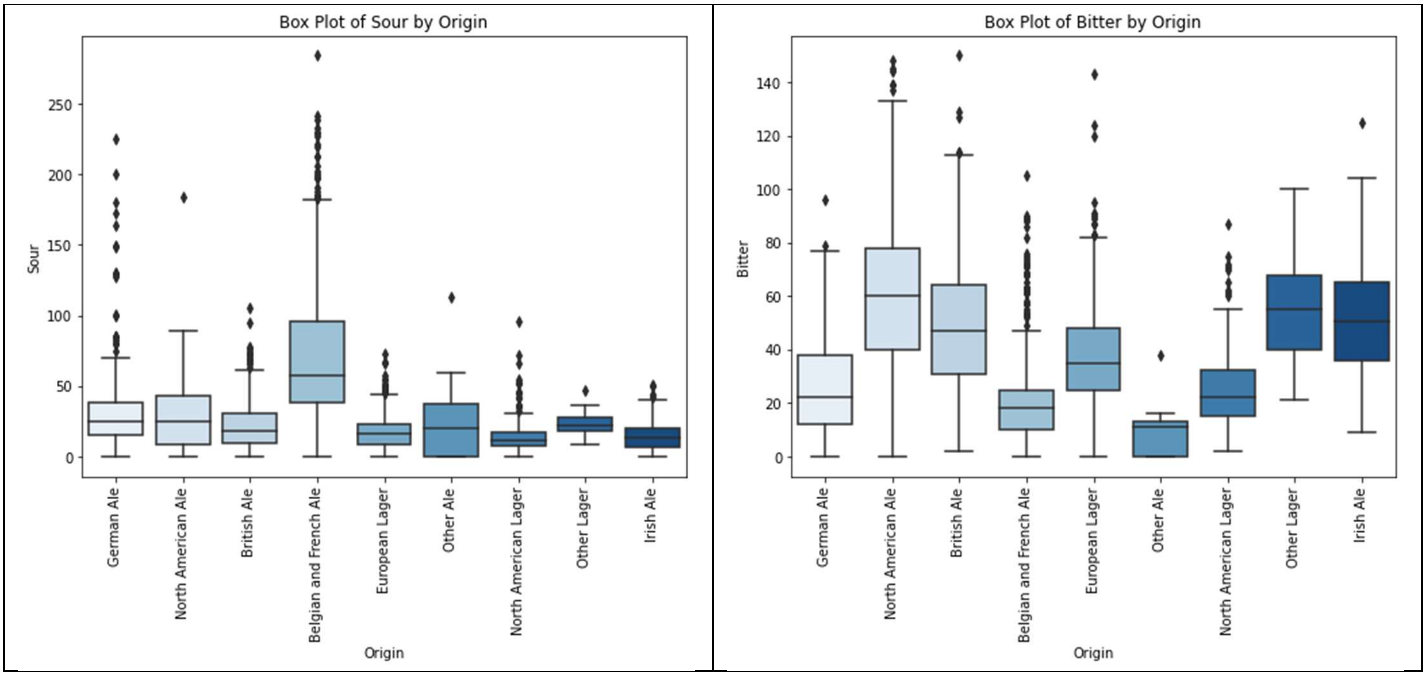


Table 3. Origin Classification Boxplots

METHODOLOGY

Feature Selection

Feature selection was performed through determining the chi-squared scores of the features and ranking them by their score. Based on these scores, the fruits, sour, body, and sweet features have the highest rankings and scores for Style Classification. This aligns with predictions made in the EDA section. Based on their low rankings and scores, the abv and hoppy features were excluded from Style Classification (Table 4).

Feature	Score
fruits	7299.18
sour	6758.46
body	2198.26
sweet	2192.71
spices	1927.87
max ibu	1632.75
alcohol	735.55
bitter	679.00
salty	304.25
astringency	88.41
malty	85.77
abv	69.87
hoppy	2.16

Table 4. Feature scoring for Style Classification

The same process was used to perform feature selection for Origin Classification. Based on these scores, it is evident that the sour, fruits, bitter, and malty characteristics have the highest scores (all about 10,000). It is shown that the sour characteristic is highly important in Origin Classification as it has a score of about 27,000, much higher than the other feature characteristics. These rankings align with EDA performed in the previous selection. Due to their very low scores, the features salty and abv were excluded from the Origin Classification (Table 5).

Feature	Score
sour	27796.50
fruits	19804.63
bitter	11513.28
malty	10980.68
max ibu	9288.65
body	6549.96
hoppy	5404.30
spices	4596.85
sweet	3849.29
alcohol	2064.76
astringency	1290.70
salty	399.98
abv	218.89

Table 5. Feature scoring for Origin Classification

Modeling

The following classification models were fitted using the training data and evaluated using the testing data. Parameter tuning was performed using a grid-search technique on the training data with a 5-fold cross validation splitting strategy and scored using prediction accuracy.

Logistic Regression utilizes the logit/sigmoid function to create a boundary between two classes. It is a linear model that performs well when there is a linear relationship between the features and the target. Regularization adds a penalty to the cost function to help prevent overfitting. The regularization parameter, C, is used to control the regularization penalty. Smaller values of C increase the regularization penalty effect.

The K-Nearest Neighbors (KNN) model classifies datapoints based on its similarity to the datapoints nearby. The parameter n-neighbors is used to specify how many neighboring datapoints will influence the classification of the datapoint being predicted. Larger numbers of neighbors may not capture the data well enough, while smaller numbers of neighbors may lead to overfitting.

The Support Vector Machine (SVM) model identifies a hyperplane that classifies the data by maximizing the margin, or distance between the hyperplane and support vectors. The support vectors are datapoints closest to the hyperplane that influence the hyperplane's placement. A regularization parameter, C, is chosen to control the prioritization of the margin. Smaller values of C prioritize a wider margin, and larger values of C prioritize correctly classified datapoints and a narrower margin. The kernel function of the SVM module transforms the data when a non-linear decision boundary is needed.

The Decision Tree model performs classification through splitting data into leaves based on node criteria. The criteria parameter indicates how the quality of the split will be measured, the max depth parameter indicates the maximum number of expansions allowed, and the minimum samples per split indicates the minimum number of samples allowed per leaf.

The Random Forest model is an ensemble method consisting of many decision trees. Each of the decision trees are constructed individually on a random subset of the data. The results from the individual trees are compiled as votes to determine the ultimate classification of the data points. Random forests are known to be robust due to their cross-validation-like technique and excel at handling high-dimensional data.

Adaptive Boosting (AdaBoost) is another ensemble method that boosts the performance of weak classifiers to create a stronger one. Classifiers are built iteratively and assigned weights based on their results. This method focuses on datapoints that are difficult to classify by assigning them higher weights in subsequent iterations. In this project, the AdaBoost model was applied to the Random Forest model used. In addition to the number of estimators, the model is assigned a learning rate. The learning rate scales the contribution of each classifier. A smaller learning rate requires more iterations to converge than a larger learning rate.

Gradient Boosting is also an ensemble learning method that works by training tree models sequentially. With each iteration, the model works to correct deficiencies in the previous model. Through this method, the weak tree learners are combined to create a stronger learner. Like AdaBoost, this model includes the number of estimators and learning rate parameters. Additionally, it includes a maximum depth parameter to be applied to each individual tree in the ensemble.

The hyperparameters used for each model in both Style and Origin Classification are presented in Table 6 and Table 7.

Model	**Tuned Parameters**
Logistic Regression	regularization parameter, C = 0.001
K-Nearest Neighbors (KNN)	n-neighbors = 5
Support Vector Machine (SVM)	regularization parameter, C = 10 kernel = radial basis function (rbf)
Decision Tree	criteria = Gini impurity maximum depth = 10 minimum samples per split = 10
Random Forest	n-estimators = 100 maximum depth = 20

	minimum samples per split = 10
Random Forest with ADA Boosting	learning rate = 0.5 n-estimators = 50
Gradient Boosting	learning rate = 0.1 maximum depth = 4 n-estimators = 200

Table 6. Modeling parameters used for Style Classification

Model	Tuned Parameters
Logistic Regression	regularization parameter, C = 1
K-Nearest Neighbors (KNN)	n-neighbors = 5
Support Vector Machine (SVM)	regularization parameter, C = 10 kernel = radial basis function (rbf)
Decision Tree	criteria = entropy maximum depth = None minimum samples per split = 2
Random Forest	n-estimators = 200 maximum depth = 20 minimum samples per split = 2
Random Forest with ADA Boosting	learning rate = 0.1 n-estimators = 50
Gradient Boosting	learning rate = 0.2 maximum depth = 5 n-estimators = 200

Table 7. Modeling parameters used for Origin Classification

After tuning and fitting the model to the training data, the target classes of the test data were predicted and compared to the actual test data targets. Metrics reported include accuracy, precision, recall, and F1-score. Note that precision, recall, and F1-scores reported are macro averages of both variables to ensure that the model performs well for all classes, regardless of sample size.

RESULTS

Style Classification

The coefficients determined by the Logistic Regression model are presented in Table 8. It is as expected that the coefficients for sour, and body are among the top highest absolute value, as they were ranked highly in feature selection. However, it is surprising that the astringency, salty, and max ibu features also had a large impact, since they were ranked lower during feature selection.

Feature	Coefficient
astringency	0.095
salty	0.069
alcohol	0.025
sweet	0.023
bitter	0.006
malty	-0.0005
spices	-0.012
fruits	-0.027
max ibu	-0.044
body	-0.047
sour	-0.055

Table 8. Style Classification Logistic Regression Coefficients

The Logistic Regression model performed the worst among the models, tied with the decision tree model in terms of accuracy and precision while the decision tree performed slightly better in terms of recall and F1-score. The lower performance of this model suggests that the data may not have a linear relationship with the target classes. The precision of the Logistic Regression model is higher than the recall, indicating that the model is better at correctly identifying positive instances, but may miss actual positive instances.

As described above, the Decision Tree model did not perform well compared to the other models but had a higher recall and F1-score than that of the Logistic Regression model. Decision trees tend to be less sensitive to outliers. Since the two classes encompass a large variety of beer styles, the Decision Tree may not be able to pick up on the differences between the classes.

The KNN model performed slightly better than both the decision tree model and Logistic Regression model, in all metrics. KNN tends to work well with decision boundaries that are not complex. The lower performance of this model may indicate that the decision boundary is more complex.

The RBF, or Gaussian, kernel that was determined for this model is used to capture complex, nonlinear relationships. The SVM model performed average compared to the other models but was the best performing non-ensemble method.

The Random Forest model performed better than all the non-ensemble methods, but slightly worse than the boosting methods. The precision of the Random Forest model was higher than the recall indicating that the model is better at correctly identifying positive instances but may miss actual positive instances.

AdaBoosting the Random Forest model resulted in a slightly higher performance. Both the accuracy and recall were slightly higher in the AdaBoost version of the Random Forest than in the simple Random Forest ensemble. This was expected since this model combines multiple Random Forest models to create a stronger result.

Overall, the Gradient Boosting model performed the best of all models, with an accuracy of 90%, a precision of 89%, recall of 86%, and F1-Score of 87%. Each of these metrics outperforms metrics in the other models. This was expected since the Gradient Boosting model has a stronger ability to capture complex patterns and relationships.

The testing metrics for each Style Classification model are presented in Table 9.

Model	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0.90	0.89	0.86	0.87
Random Forest with AdaBoost	0.89	0.87	0.84	0.85
Random Forest	0.88	0.87	0.83	0.85
Support Vector Machine (SVM)	0.86	0.83	0.83	0.83
K-Nearest Neighbors (KNN)	0.84	0.80	0.80	0.80
Decision Tree	0.83	0.79	0.79	0.79
Logistic Regression	0.83	0.79	0.76	0.77

Table 9. Style Classification Model Results

Generally, all the models had a more difficult time predicting lagers than ales. In the Gradient Boosting model, the precision of lager vs ale prediction was 0.86 and 0.92, respectively. Likewise, the recall of lager vs. ale prediction was 0.77 and 0.95, respectively. The lower recall of the prediction of the lager class indicates that the model has trouble identifying the lager style beers. Of the 149 lager style samples, only 115 were correctly predicted as lagers, whereas of the 389 ale style samples, 370 were correctly predicted (Figure 3). This could be due to the imbalance of sample size between the two styles. It is also possible that ales have more defining features than lagers.

Five of the 34 misclassified lagers were Schwarzbier and four were European Dark Lagers. Ales are typically full-bodied and bitter with a strong flavor. Schwarzbier are characterized by having a medium body and moderate to high bitterness, and European Dark Lagers are characterized by having a full body. Similar characteristics of these beers may also account for some of the misclassifications.

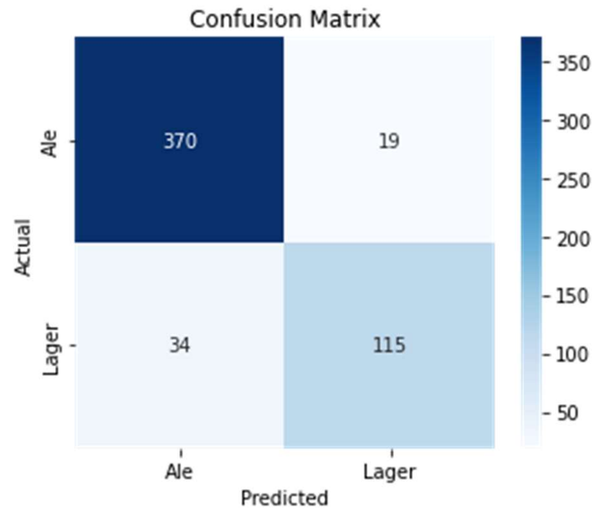


Figure 3. Confusion Matrix of Gradient Boosting Model for Style Classification

Origin Classification

The coefficients determined by the Logistic Regression model are not presented. Since the prediction is multi-class, the coefficients are not as easy to interpret. The Logistic Regression model performed the worst among the models, in all metrics. The low performance of this model suggests that the data may not have a linear relationship with the target classes.

The KNN model performed better than the Logistic Regression model in all metrics, but its performance was still quite a bit worse than the other models. KNN tends to work well with decision boundaries that are not complex. The low performance of this model may indicate that the decision boundary is more complex.

The SVM model performed quite better than the KNN and Logistic Regression models in all metrics. The RBF, or Gaussian, kernel that was determined for this model is used to capture complex, nonlinear relationships. This is the same kernel that was chosen for Style Classification. It was as expected that the SVM model performed better than KNN and Logistic Regression, due to its ability to capture more complex patterns.

The Decision Tree model performed the best of the non-ensemble methods. Decision trees tend to be less sensitive to outliers. It is interesting that this model performed well in Origin Classification, but poorly in Style Classification. The wide variety of beers in the dataset may have had a negative impact on trying to classify into only two styles instead of multiple origins, which allows for more difference in the classes.

The Random Forest model performed better than all the non-ensemble methods. The accuracy of the Random Forest model and precision were the same as the Random Forest with AdaBoost. The recall, however, was slightly lower than the AdaBoost model. It is interesting that the two performed so closely, since the AdaBoost model combines multiple Random Forest models to create a stronger prediction.

Overall, like with Style Classification, the Gradient Boosting model performed the best of all models, with an accuracy of 80%, a precision of 77%, recall of 76%, and F1-Score of 76%. Each of these metrics outperforms metrics in the other models. Like Style Classification, this was expected since the Gradient Boosting model has a stronger ability to capture complex patterns and relationships. Given that this is a multi-class prediction problem, the relationship between the classes is likely very complex.

The testing metrics for each Origin Classification model are presented in Table 10.

Model	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0.80	0.77	0.76	0.76
Random Forest with AdaBoost	0.78	0.76	0.72	0.74
Random Forest	0.78	0.76	0.71	0.73
Decision Tree	0.74	0.70	0.72	0.70
Support Vector Machine (SVM)	0.72	0.71	0.67	0.68
K-Nearest Neighbors (KNN)	0.65	0.62	0.60	0.60
Logistic Regression	0.63	0.61	0.57	0.58

Table 10. Origin Classification Model Results

Generally, all the models had a difficult time with classifying Irish Origin Ales and North American Lagers. In the Gradient Boosting model, half (9 out of 18) of the Irish Origin Ales were misclassified. Four were misclassified as British Origin Ales and 3 were misclassified as European Lagers. It is not necessarily surprising that four were misclassified as British Origin Ales, since both are ales, and the two countries are very close by and likely have similar influences. The F1-Score for Irish Origin Ales was only 53%. The low prediction power of this class could likely be due to the very small sample size.

Of the misclassified Irish Origin Ales, three were Irish Dry Stouts classified as British Origin Ales. British Ales tend to have high malty characteristics with nutty, toffee, or caramel flavors and light chocolate notes. Likewise, Irish Dry Stouts are often characterized with coffee-like and malty flavors. The similarity of these styles likely contributed to the misclassifications.

Likewise, 21 of the 58 North American Origin Lagers were misclassified. Of the misclassified samples, 12 were classified as European Origin Lagers. The F1-Score for North American Lagers was only 69%. Of the North American Origin Lagers misclassified as European Origin Lagers, six were American Amber/Red Lagers. American Amber/Red Lagers are characterized by high malt and medium caramel flavors. European Origin Lagers typically have a light malt aroma, but also some caramel flavoring. Though they don't share similar malt aromas, the caramel flavor may have contributed to the misclassifications of these origins.

See Figure 4 for the confusion matrix for Origin Classification using Gradient Boosting.

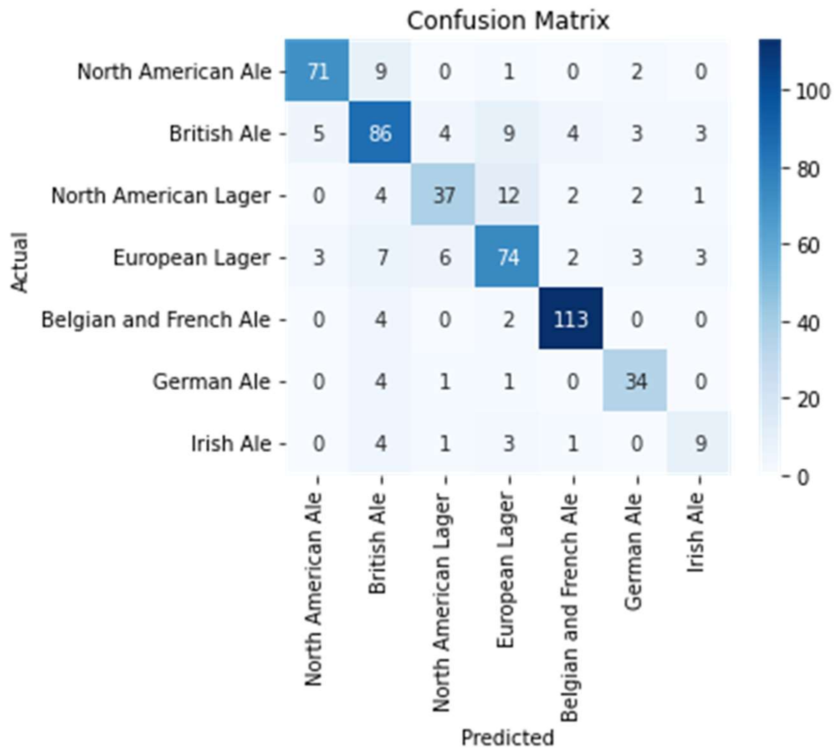


Figure 4. Confusion Matrix of Gradient Boosting Model for Origin Classification

CONCLUSION

The Gradient Boosting model performed the best in both style and Origin Classifications. This model was able to predict craft beer style (ale vs. lager) with 90% accuracy and craft beer style origin with 80% accuracy. The best performing non-ensemble method was SVM for Style Classification and Decision Tree for Origin Classification.

The fruits, sour, body, and sweet features were particularly important in Style Classification, while the abv and hoppy features were excluded. In Origin Classification, the sour, fruits, bitter, and malty features were important, while the salty and abv features were excluded.

These results indicate that using desired user characteristics, a style and origin could be recommended to craft beer drinkers. Potential improvements include first identifying ale or lager style, then narrowing down the origin by style to allow for more accurate classification. Additionally, further modeling could allow for the specific beer type to be recommended (stout, IPA, sour, etc.).

Lessons Learned

This course provided a thorough and comprehensive explanation of statistical models, with enough detail to understand their uses and nuances. The homework assignments in this course were very useful in becoming more familiar with different models and their uses as well as with writing comprehensive reports. This project allowed the building of a framework for future projects and brought attention to the importance of sample size in classification problems.

REFERENCES

[1] Biscotti, L. (2023, January 20). *Craft beer boom slows, but still grows*. Forbes. https://www.forbes.com/sites/louisbiscotti/2023/01/19/craft-beer-boom-slows-but-still-grows/?sh=6a10fc615e3e.

[2] *Beer profile and ratings data set*. Kaggle. (2021, November 18). https://www.kaggle.com/datasets/ruthgn/beer-profile-and-ratings-data-set.

[3] *Brewers Association Beer Style Guidelines*. Brewers Association. (2023, July 10). https://www.brewersassociation.org/edu/brewers-association-beer-style-guidelines/#37.

APPENDIX

Original Variables – beer_profile_and_ratings.csv

Variable Name	Data Type	Description
Name	object	Beer Name
Style	object	Beer Style
Brewery	object	Brewery Name
Beer Name (Full)	object	Brewery + Beer Name
Description	object	Notes (if available)
ABV	float64	Alcohol Content (% by Volume)
Min IBU	int64	Minimum IBU value beer can possess
Max IBU	int64	Maximum IBU value beer can possess
Astringency	int64	Astringency word count
Body	int64	Body in terms of mouthfeel word count
Alcohol	int64	Perceived booziness word count
Bitter	int64	Bitterness word count
Sweet	int64	Sweetness word count
Sour	int64	Sourness word count
Salty	int64	Saltiness word count
Fruits	int64	Fruity aroma and flavor word count
Hoppy	int64	Hoppy aroma and flavor word count
Spices	int64	Spices aroma and flavor word count
Malty	int64	Malty aroma and flavor word count
review_aroma	float64	Aroma average rating score
review_appearance	float64	Appearance average rating score
review_palate	float64	Palate average rating score
review_taste	float64	Taste average rating score
review_overall	float64	Overall average rating score
number_of_reviews	float64	Count of reviews

NOTE: Astringency through Malty – represent the tasting profile – defined by word counts found in up to 25 reviews of each beer, likely describe what is experienced rather than what is not