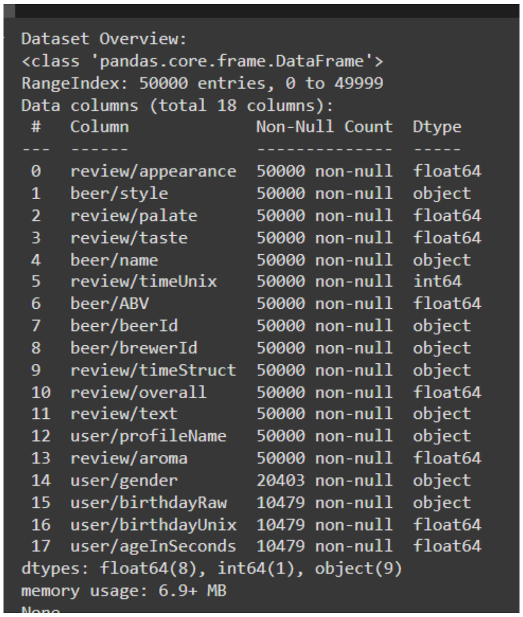
CSE158 Assignment 2: Classifying Beer Style from Review Data

Brandon Breeze, Pranu Lingineni, Ethan Cota, Jason Wong

Contents:

Part 1 - Exploratory Analysis:

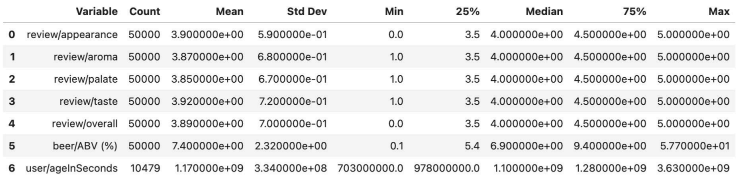
The beer_50000 dataset contains 50,000 beer reviews sourced from RateBeer, covering a temporal span from May 10, 1999, to January 11, 2012. This dataset captures over a decade of consumer feedback on various beers, encompassing detailed numeric ratings for sensory attributes such as taste, aroma, appearance, and palate, as well as overall quality. It also includes information about alcohol content (ABV), beer styles, text-based reviews, and limited demographic details about the reviewers. Figure 1 displays the full list of features the dataset captures.



```
Dataset Overview:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 18 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   review/appearance   50000 non-null  float64
 1   beer/style          50000 non-null  object
 2   review/palate       50000 non-null  float64
 3   review/taste        50000 non-null  float64
 4   beer/name           50000 non-null  object
 5   review/timeUnix     50000 non-null  int64
 6   beer/ABV            50000 non-null  float64
 7   beer/beerId         50000 non-null  object
 8   beer/brewerId       50000 non-null  object
 9   review/timeStruct   50000 non-null  object
 10  review/overall      50000 non-null  float64
 11  review/text         50000 non-null  object
 12  user/profileName    50000 non-null  object
 13  review/aroma        50000 non-null  float64
 14  user/gender         20403 non-null  object
 15  user/birthdayRaw    10479 non-null  object
 16  user/birthdayUnix   10479 non-null  float64
 17  user/ageInSeconds   10479 non-null  float64
dtypes: float64(8), int64(1), object(9)
memory usage: 6.9+ MB
None
```

Figure 1

Figure 2 summarizes the key statistics for the numeric variables in the dataset. The numeric variables show a general clustering of ratings between 3.5 and 4.5, reflecting a strong positive sentiment. ABV ranges widely, highlighting a diverse selection of beers, from low-alcohol options to strong specialty brews. Variables like user age and gender are sparsely populated (20.9% and 40.8% availability, respectively), limiting their reliability for in-depth analysis. Age anomalies, such as extremely high values, further reduce their utility. While female reviewers provide slightly higher average ratings (4.00) than males (3.90), the small sample size for females restricts the generalizability of this observation. Similarly, while the temporal variable offers insights into trends over time, its exploratory value is limited. Variations in ratings across years may be influenced by external factors like changes in beer quality, market preferences, or platform dynamics, making it challenging to attribute trends to specific aspects of the dataset. Furthermore, the wide timeline of reviews, spanning over a decade, introduces potential inconsistencies in reviewer behavior and beer availability, which could dilute the relevance of temporal patterns for deeper analysis.



	Variable	Count	Mean	Std Dev	Min	25%	Median	75%	Max
0	review/appearance	50000	3.900000e+00	5.900000e-01	0.0	3.5	4.000000e+00	4.500000e+00	5.000000e+00
1	review/aroma	50000	3.870000e+00	6.800000e-01	1.0	3.5	4.000000e+00	4.500000e+00	5.000000e+00
2	review/palate	50000	3.850000e+00	6.700000e-01	1.0	3.5	4.000000e+00	4.500000e+00	5.000000e+00
3	review/taste	50000	3.920000e+00	7.200000e-01	1.0	3.5	4.000000e+00	4.500000e+00	5.000000e+00
4	review/overall	50000	3.890000e+00	7.000000e-01	0.0	3.5	4.000000e+00	4.500000e+00	5.000000e+00
5	beer/ABV (%)	50000	7.400000e+00	2.320000e+00	0.1	5.4	6.900000e+00	9.400000e+00	5.770000e+01
6	user/ageInSeconds	10479	1.170000e+09	3.340000e+08	703000000.0	978000000.0	1.100000e+09	1.280000e+09	3.630000e+09

Figure 2

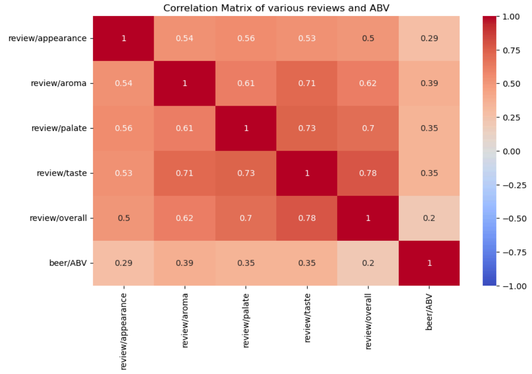


Figure 3

Sensory attributes such as taste, aroma, and palate are highly correlated with overall ratings, particularly taste (correlation = 0.78) (Figure 3). This strong predictive relationship motivates their inclusion as primary features in a modeling task.

The distribution of beer styles is heavily skewed toward a small number of popular styles (Figure 4). American Double / Imperial Stout leads with 11.93% of the reviews, followed by American IPA (8.23%) and American Double / Imperial IPA (7.77%). Together, these top three styles account for over 25% of all reviews. Conversely, niche styles like Low Alcohol Beer and Kristalweizen account for less than 0.02% each, reflecting their limited representation in the dataset. This imbalance highlights the need for careful consideration in predictive modeling. To address this imbalance, styles can be grouped into broader categories—such as ale, lager, and more—to create a more balanced classification task.

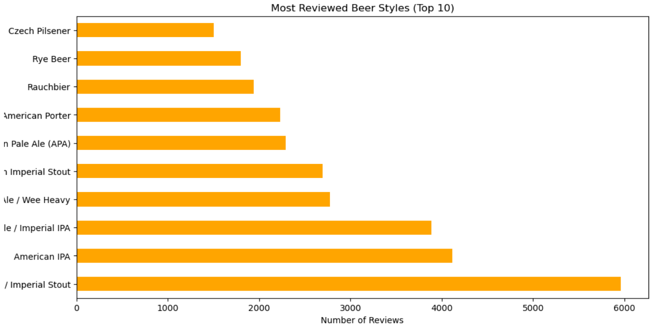


Figure 4

Average overall ratings peak for beers with ABV in the range of 10-15%, while those exceeding 20% ABV receive lower overall ratings (Figure 5). This suggests that moderate alcohol levels enhance the sensory experience, while extremely high alcohol content may detract from enjoyment for most consumers.

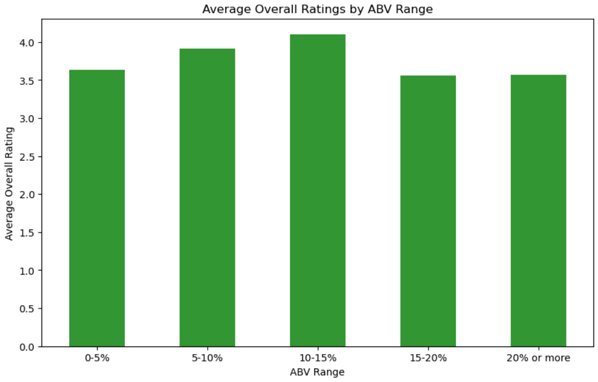


Figure 5

The exploratory analysis conducted highlights sensory ratings (taste, aroma, palate, appearance) as the strongest predictors of beer style and overall quality. The analysis reveals significant class imbalance in beer styles necessitates grouping into broader categories to enable more balanced modeling. While ABV contributes moderately, sparse demographic data, temporal data and text reviews are less relevant for initial models. Though, they may offer potential for future analyses. These findings logically inform our predictive task, ensuring that the most impactful variables are emphasized while addressing dataset challenges like class imbalance and sparsity. This approach lays a strong foundation for subsequent model development.

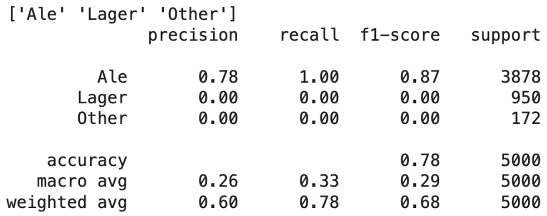
2 - Predictive Task:

The predictive task identified using the beer_50000 dataset was the classification of beer styles based on user ratings for taste, aroma, appearance, and palate. These ratings, provided as floating-point values from 0 to 5, required minimal preprocessing for the model's features. However, the beer style classification feature needed significant processing to improve model performance. Initially, using the raw beer style categories (e.g., American Double / Imperial Stout, American IPA, Scotch Ale / Wee Heavy) resulted in low accuracy (~15%) with the logistic regression classifier. To simplify the prediction task, we grouped the original beer styles into three categories: ale, lager, and 'other.' The 'other' group included beer styles that did not fit cleanly into the ale or lager categories. After this grouping, the dataset contained approximately 38,784 instances of ale, 9,390 instances of lager, and 1,826 instances of 'other.' To prepare the data for modeling, we converted the string categories ('ale,' 'lager,' and 'other') into numerical inputs using the 'LabelEncoder' from the 'sklearn.preprocessing' library. This concluded our preprocessing efforts for the dataset. To evaluate the models, we used accuracy, precision, and recall metrics across the ale, lager, and other categories. Accuracy, defined as the ratio of correct predictions to total predictions, provided an overall measure of model performance, while precision and recall offered deeper insights into how well the model performed for each specific category. The baseline model assumed all predictions to be 'ale,' reflecting the largest category, which comprised 77.568% of the dataset. This simple model provided a benchmark for comparison. Beyond the baseline and models taught in class (e.g., logistic regression, Naive Bayes, and support vector machines), we also explored additional approaches to assess their performance and suitability for this classification task.

3 - Model:

We experimented with several models and ultimately selected the Random Forest Classifier as our final model. This model operates by generating multiple decision trees, each trained on different subsets of the data to ensure diversity. These trees collectively predict the target outcome—'what beer style is this, given the ratings for taste, appearance, aroma, and palate?'—with the model's final prediction determined by the majority vote among the trees.

To validate our choice, we compared the Random Forest Classifier to models covered in class, including logistic regression, Naive Bayes, and support vector machines (SVM). Logistic regression, Naive Bayes, and SVM were chosen for their relevance and familiarity from class instruction, while the Random Forest Classifier was selected based on external research suggesting that tree-based classifiers perform well with unbalanced datasets, which aligns with our case.



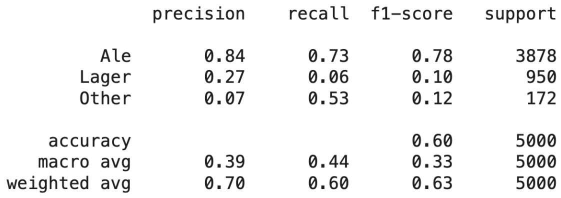
```
['Ale' 'Lager' 'Other']
              precision    recall  f1-score   support

         Ale       0.78      1.00      0.87      3878
       Lager       0.00      0.00      0.00       950
       Other       0.00      0.00      0.00       172

    accuracy                           0.78      5000
   macro avg       0.26      0.33      0.29      5000
weighted avg       0.60      0.78      0.68      5000
```

Figure 6: Baseline Model Scores

The table above shows the performance metrics for the baseline model. With an overall accuracy of 78%, the model aligns with the fact that 'ale' comprises approximately 78% of the dataset. However, the model is limited in its predictive ability, as it only predicts 'ale.' This is evident in the 0% recall for 'lager' and 'other' beer styles, indicating it fails to identify these categories entirely. Additionally, because the model always predicts 'ale,' its precision for 'ale' predictions is capped at 78%.



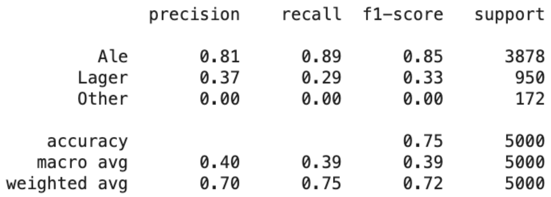
```
              precision    recall  f1-score   support

         Ale       0.84      0.73      0.78      3878
       Lager       0.27      0.06      0.10       950
       Other       0.07      0.53      0.12       172

    accuracy                           0.60      5000
   macro avg       0.39      0.44      0.33      5000
weighted avg       0.70      0.60      0.63      5000
```

Figure 7: Logistic Regression Classifier Scores

The logistic regression model achieves an overall accuracy of 60%, which is lower than the baseline model's 78%. However, unlike the baseline model, it makes predictions for non-'ale' styles. Specifically, it recalls 6% of 'lager' styles and achieves 27% precision for 'lager' predictions. The model also attempts to predict 'other' styles but with a low precision of 7%. Despite its broader range of predictions, the significantly lower overall accuracy led us to conclude that the logistic regression model underperforms compared to the baseline model.



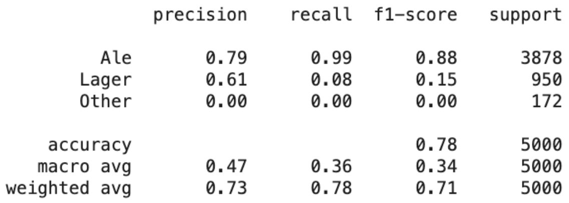
```
              precision    recall  f1-score   support

         Ale       0.81      0.89      0.85      3878
       Lager       0.37      0.29      0.33       950
       Other       0.00      0.00      0.00       172

    accuracy                           0.75      5000
   macro avg       0.40      0.39      0.39      5000
weighted avg       0.70      0.75      0.72      5000
```

Figure 8: Naive Bayes Scores

The Naive Bayes model outperformed the logistic regression model in several key areas. It made significantly more predictions for the 'lager' class, achieving 37% precision for these predictions. Additionally, the model attained an overall accuracy of 75%, slightly below the baseline model's 78%, but notably better in predicting the 'lager' category, which is a critical improvement. However, unlike the logistic regression model, the Naive Bayes model did not make any predictions for the 'other' category, which the logistic regression model attempted, albeit with very low precision.



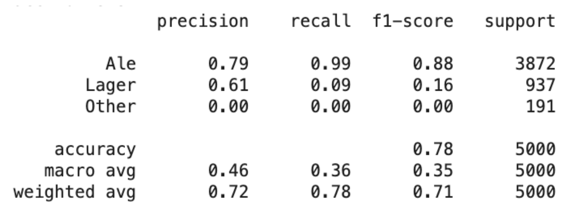
```
              precision    recall  f1-score   support

         Ale       0.79      0.99      0.88      3878
       Lager       0.61      0.08      0.15       950
       Other       0.00      0.00      0.00       172

    accuracy                           0.78      5000
   macro avg       0.47      0.36      0.34      5000
weighted avg       0.73      0.78      0.71      5000
```

Figure 9: Support Vector Machine Scores

The SVM classifier performed well on the 'ale' class, achieving 99% recall and 79% precision. However, its performance on the 'lager' class was more limited, with a recall of only 8%. Within this subset, the model achieved a precision of 61% for 'lager' predictions. The overall accuracy of the SVM model is 78%, matching the baseline model. However, it surpasses the baseline by making and correctly identifying some 'lager' predictions, which represents an improvement in predictive diversity.



```
              precision    recall  f1-score   support

         Ale       0.79      0.99      0.88      3872
       Lager       0.61      0.09      0.16       937
       Other       0.00      0.00      0.00       191

    accuracy                           0.78      5000
   macro avg       0.46      0.36      0.35      5000
weighted avg       0.72      0.78      0.71      5000
```

Figure 10: Random Forest Classifier Scores

The table above presents the performance metrics for our chosen model, the Random Forest Classifier. While its performance is comparable to the SVM classifier, it offers a significant advantage in training speed, making it more efficient for our purposes. Based on this, we selected the Random Forest Classifier and performed a grid search to optimize hyperparameters, including the number of estimators, max depth, and criterion, among others. The resulting scores show that the model performs similarly to the SVM classifier on the 'ale' class but slightly outperforms it on the 'lager' class. However, like the SVM classifier, it does not recall any 'other' beer styles.

We selected the Random Forest Classifier as our final model because it achieved one of the highest overall accuracy scores, comparable to the SVM classifier, while being significantly faster to train. This efficiency enabled us to perform an extensive grid search to optimize hyperparameters. The optimal parameters were: 'n_estimators' = 50, 'max_depth' = 5, 'min_samples_split' = 2, 'min_samples_leaf' = 1, 'criterion' = 'gini', and 'class_weight' = None. A key challenge we encountered was the dataset's heavy imbalance, with a majority of samples classified as 'ale,' causing the model to exhibit bias toward this class. To address this, we experimented with setting 'class_weight' to 'balanced.' However, while this approach reduced the bias, it significantly decreased the overall accuracy, making it an unsuitable solution.

Regarding the strengths and weaknesses of each model, the logistic regression model underperformed in terms of overall accuracy compared to the baseline but stood out as the only model that attempted to predict the 'other' class. The Naive Bayes model also had lower overall accuracy but achieved the highest F1 score for the 'lager' class, thanks to its higher recall for 'lager,' even though its precision was lower. The SVM and Random Forest Classifier performed similarly, with both excelling on the 'ale' class but struggling with 'lager' and 'other.' However, when they did recall the 'lager' class, their predictions were generally accurate. A notable downside of the SVM compared to the Random Forest Classifier was its much longer training time, making it less efficient. The primary challenge across models was addressing the dataset's imbalance. Experimenting with parameters like 'class_weight = "balanced"' showed that while it reduced bias, it significantly decreased overall accuracy, contrary to expectations. Additionally, much of the process involved iterative experimentation, refining parameters, and addressing unexpected outcomes. ChatGPT was utilized to assist in generating initial code for models and grid searches. However, these initial outputs often required significant adjustments and fine-tuning to better align with the specific goals and parameters of the project.

4 - Literature:

The dataset we are using was scraped from the beer rating website RateBeer by Julian McAuley et al. in late 2011. It consists of nearly 3 million beer reviews spanning from April 2000 to November 2011. The dataset was originally used for rating prediction tasks due to its explicit review dimensions, where each aspect of a beer (such as taste, feel, appearance, and aroma) has an individual ranking. The primary objective of the original study was to determine which parts of the written reviews corresponded to these individually ranked aspects. The authors then aimed to use these findings to generate summaries that accurately described the overall ratings. Additionally, since some individual ratings were optional in the dataset, the authors sought to predict and fill in the missing values. The authors approached this data by modeling it with an aspect sentiment model. This model enabled them to recover missing ratings, identify connections between written reviews and individual rankings, and generate summaries reflecting the overall ratings. These analyses demonstrated the dataset's suitability for rating prediction tasks [1].

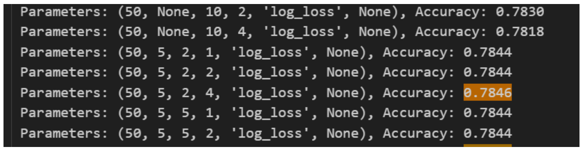
Similar datasets, such as those from Amazon and Audible described in [1], are also well-suited for studying consumer data. These datasets are ideal for various recommendation tasks, including rating prediction and product recommendations. However, when compared to the RateBeer dataset, a notable drawback is the lack of individual rankings for specific aspects of reviews in the Amazon and Audible datasets. Despite this limitation, it would be feasible to make predictions based on genres in the Amazon and Audible datasets, similar to how we predict beer styles in our task. It would also be interesting to apply sentiment analysis, as described in [1], to predict genres based on review content. For this project, however, we focused on leveraging the individual aspect rankings provided by the BeerAdvocate and RateBeer datasets, which offer a unique advantage for our predictive modeling tasks.

State-of-the-art methods for studying consumer data include techniques such as decision trees, k-nearest neighbor, regression, and neural networks. Among these, k-nearest neighbor is one of the most widely used, while decision trees are particularly popular for classification tasks. For our project, we focused on regression, decision trees, Naive Bayes, and support vector machine classifiers, selecting these methods based on our familiarity with them. Our analysis aligned with existing findings, confirming that decision trees are highly effective for classification tasks. In [3], a study with a similar objective of product classification using consumer data also concluded that decision trees perform well for datasets like ours. However, unlike our findings, their study reported greater success with support vector machines, whereas we found that decision trees outperformed support vector machines in accuracy on our dataset.

5 - Conclusion:

Using our dataset of user beer reviews—which included ratings for taste, aroma, appearance, palate, overall quality, and other review information—we aimed to predict the style of beer. In our exploratory analysis, we found no missing data or significant abnormalities, apart from some user-specific information. From this analysis, we identified that the four attribute ratings (taste, aroma, appearance, and palate) were the most strongly correlated with predicting a beer's style.

By focusing on these aspects of the review data as our model's features, we achieved the highest accuracy using a Random Forest Classifier. After implementing the model and performing a grid search to optimize its hyperparameters, we maximized its accuracy at 78.46%, minimizing the log-loss in the process.



```
Parameters: (50, None, 10, 2, 'log_loss', None), Accuracy: 0.7830
Parameters: (50, None, 10, 4, 'log_loss', None), Accuracy: 0.7818
Parameters: (50, 5, 2, 1, 'log_loss', None), Accuracy: 0.7844
Parameters: (50, 5, 2, 2, 'log_loss', None), Accuracy: 0.7844
Parameters: (50, 5, 2, 4, 'log_loss', None), Accuracy: 0.7846
Parameters: (50, 5, 5, 1, 'log_loss', None), Accuracy: 0.7844
Parameters: (50, 5, 5, 2, 'log_loss', None), Accuracy: 0.7844
```

Figure 11: Grid Search on our Model

We conducted a grid search to optimize the model’s hyperparameters, focusing on the number of estimators, maximum tree depth, and minimum samples for splitting and leaf nodes (Figure 11). The optimized parameters were 50 estimators, a maximum tree depth of 5, a minimum of 2 samples for splitting, and 4 for leaf nodes.

The number of estimators represents the total number of trees in the random forest, while the maximum tree depth defines the maximum size of each tree. The minimum samples for splitting and leaf nodes specify the smallest number of samples required to split a node and to form a leaf, respectively.

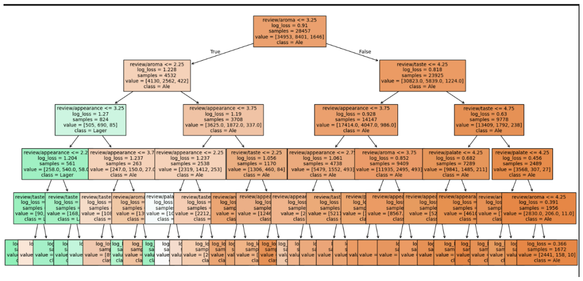


Figure 12: Model Estimator Example

Among the models we trained—logistic regression, Naive Bayes, and support vector machines (SVM)—the Random Forest Classifier outperformed all but the SVM model, with which it performed similarly. The superior performance of the Random Forest Classifier may be attributed to its ability to handle the dataset's class imbalance, skewed heavily toward the majority class of 'Ale.' Additionally, the Random Forest Classifier is less prone to overfitting compared to other classification models we tried, which likely contributed to its better accuracy.

Due to the limited size and dimensionality of our dataset, achieving exceptionally high accuracy was challenging. By comparison, papers we reviewed reported accuracies of around 78% to 80% when predicting more specific classes, such as 'light ale,' by leveraging more complex models that incorporated text modeling and sentiment analysis. While our results did not reach the levels achieved in these studies, for our task of predicting beer styles using quantitative review data as features, we successfully selected a model that delivered a solid accuracy measure for our objectives.

Citations:

[1] McAuley, J., Leskovec, J. and Jurafsky, D. (2012) Learning attitudes and attributes from multi-aspect reviews, arXiv.org. Available at: https://arxiv.org/abs/1210.3926 (Accessed: 01 December 2024).

[2] Park, D.H. et al. (2012) A literature review and classification of Recommender Systems Research, Expert Systems with Applications. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417412002825 (Accessed: 01 December 2024).

[3] Pawłowski, M. (2021) 'Machine Learning Based Product Classification for eCommerce*', Journal of Computer Information Systems, 62(4), pp. 730–739. doi: 10.1080/08874417.2021.1910880.