The Oxford Comma: A User Classification Model

Hadley Kruse
DSCI 303
Rice University
hmk2@rice.edu

Jae Hyun Hong
DSCI 303
Rice University
jh103@rice.edu

I. ABSTRACT

The Oxford comma has been a contentious topic in the English language for decades. People in the writing field are oftentimes adamant about whether it should be used, while the general public may go back and forth. Many surveys have pointed to various results about Oxford comma usage in America, but machine learning has yet to be used as a method to solve this potential classification problem. The classification model in this paper is designed to classify a person as somebody who uses or does not use the Oxford comma based on survey responses to grammatical questions and demographic information. In order to better understand the optimal model and the consequences of using ordinal and nominal data, the raw dataset was transformed into two separate datasets, one containing label encoded values and the other using One Hot Encoding to create binary columns. The analysis includes feature selection methods of kNN, Random Forest, and Principal Component Analysis to reduce dimensionality. The following models were used to compare testing accuracy scores: kNN, Random Forest, SVM, Logistic Regression, and Naive Bayes. The findings are consistent among both transformed datasets with Random Forest outperforming the other models' testing accuracy scores. Additionally, the dataset containing ordinal values did not seem to need feature selection as most of the models performed best using all features. On the other hand, the "Binary Dataset" saw mixed results on the optimal feature selection method. The testing accuracy scores were overall quite similar and lower than expected, likely due to limited features and some biased data.

II. INTRODUCTION

Mutual agreement in grammar is necessary to create a degree of standardization that ensures understanding across different regions, cultures, and/or dialects amongst people that speak the same language. However, the question of what constitutes grammatical correctness across different aspects of language is subject to much debate. While the Oxford English Dictionary is commonly regarded as the blueprint for the English language, there are many variations that exist within English-speaking communities.[1] The placement of adverbs or pronouns, for example, are commonly agreed upon, but the usage of the Oxford comma is not. The Oxford comma (also known as the serial comma) is used to distinguish two items at the end of a list. For example, one would use an Oxford comma when saying, "I love my parents, Lady Gaga, and Humpty Dumpty." If one were to disregard the Oxford comma, the sentence would read, "I love my parents, Lady Gaga and Humpty Dumpty." The debate over the usage of the Oxford comma is one of the more controversial topics of discussion among linguists because its usage varies vastly across English-speaking societies. While this may seem inconsequential, it has caused serious problems in the real world. For instance, in 2018, a $5 million lawsuit against the dairy industry over a missing Oxford comma in their drivers' overtime pay took place.[2] Clearly, omitting an Oxford comma can cause some harmful ambiguities for a legal team. That said, the ability to determine whether a person uses the Oxford comma or not could also prove useful to gaining insight into a person's writing style and their perspective on grammar. In fields such as journalism or academia, knowing whether a candidate will use the Oxford comma could prove beneficial, too.

A. Literature Review

There are hundreds of articles debating over the Oxford comma, with most of them arguing for its usage. FiveThirtyEight surveyed a subset of the American population to identify trends and patterns across people who use the Oxford comma and those who do not. They concluded that people who believed their own grammar was "excellent" were more likely to use the Oxford comma in an example sentence.[3] While these results are interesting, their dataset can be further examined as a classification problem. Machine learning algorithms have yet to be leveraged to solve this classification problem, so this paper presents a first look into how different models can be implemented to determine whether a person is an Oxford comma user or not.

III. METHODS

A. Dataset Description

The 1130 by 13 dataset was taken from the aforementioned FiveThirthyEight article titled, "*Elitist, Superfluous, Or Popular? We Polled Americans on the Oxford Comma*," which consists of 1,130 respondents' answers to a series of survey questions and demographic information. The 13 columns and a breakdown of the possible values that each column takes on are presented below:

1. RespondentID: unique 10-digit identifier for each respondent;

2. In your opinion, which sentence is more grammatically correct?: "It's important for a person to be honest, kind and loyal" OR "It's important for a person to be honest, kind, and loyal";

3. Prior to reading about it above, have you heard of the serial (or Oxford) comma?: "No" OR "Yes";

4. How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?: "Not at all," "Not much," "Some," OR "A lot";

5. How would you write the following sentence?: “Some experts say it’s important to drink milk, but the data is inconclusive” OR “Some experts say it’s important to drink milk, but the data are inconclusive”;

6. When faced with using the word “data,” have you ever spent time considering if the word was a singular or plural noun?: “No” OR “Yes”;

7. How much, if at all, do you care about the debate over the use of the word “data” as a singular or plural noun?: “Not at all,” “Not much,” “Some,” OR “A lot”;

8. In your opinion, how important or unimportant is proper use of grammar?: “Very unimportant,” “Somewhat unimportant,” “Neither important nor unimportant,” “Somewhat important,” OR “Very important”;

9. Gender: “Male” OR “Female”;

10. Age: “18-29,” “30-44,” “45-60,” OR “> 60”;

11. Household income: “$0 - $24,999,” “$25,000 - $49,999,” “$50,000 - $99,999,” “$100,000 - $149,999,” OR “$150,000+”;

12. Education: “Less than high school degree,” “High school degree,” “Some college or Associate degree,” “Bachelor degree,” OR “Graduate degree”;

13. Location (Census Region): “East North Central,” “East South Central,” “Middle Atlantic,” “Mountain,” “New England,” “Pacific,” “South Atlantic,” “West North Central,” OR “West South Central.”

B. Preprocessing[4]

As an initial step, observations with more than 75% of its column values missing were eliminated, leaving 1,099 total observations to work with. The dataset consists solely of categorical data (apart from the RespondentID), of which there is a combination of binary, ordinal, and nominal features. As such, binary features were mapped to 0s and 1s and ordinal features were mapped to increasing integer values. For example, answers to “Education” were mapped as follows:

- “Less than high school degree” was assigned a value of 1;
- “High school degree” was assigned a value of 2;
- “Some college or Associate degree” was assigned a value of 3;
- “Bachelor degree” was assigned a value of 4;
- “Graduate degree” was assigned a value of 5.

As for the nominal features, namely “Location,” a binary indicator was created for each of its answers, such that a value of 1 under the column of a given Census Region indicates that a respondent belongs to said Census Region, and 0 otherwise. Ultimately, after removing the RespodentID column, this dataset becomes a 1,099 by 20 dataset that will be deemed as the “Ordinal Dataset” for the purposes of this paper. A separate dataset that will be deemed as the “Binary Dataset” has been created, as well, where the values for all of the ordinal features were transformed into binary indicators similar to the nominal feature, “Location.” The “Binary Dataset,” then, expands to a 1,099 x 45 dataset. Justification for the creation of the “Binary Dataset” is discussed in Section VI.

In order to fill in for the missing data on the remaining observations, a mode imputation method was utilized for all features because the percentage of missing values for any given feature was very small. One exception concerns the Household Income feature because its column had 263 missing values (263 / 1,099 = 23.9%). As such, a logistic regression for classification based on the respondent’s answers under the Gender, Age, and Education columns was utilized to impute the missing values. Finally, the answers to the survey question, “In your opinion, which sentence is more grammatically correct?” was encoded as class labels, with the usage of the Oxford comma, i.e. “It’s important for a person to be honest, kind, and loyal,” encoded as 1, and the non-usage of the Oxford comma, i.e. “It’s important for a person to be honest, kind and loyal,” encoded as 0. However, given that the class labels were initially imbalanced, with 623, or 623 / 1,099 = 56.68%, of the observations taking on a value of 1, and 476, or 476 / 1,099 = 43.31%, of the observations taking on a value of 0, an oversampling method, which involves duplicating random records from the minority class, was implemented to uphold the assumptions of a random binary classifier. Note that results may vary according to the duplicate observations that are drawn from the minority class every instance that the code is run. For this reason, a random_state = 45 was set to ensure reproducible results. The results discussed below are purely concerned with the final datasets that were created from this paper’s particular instance of drawing random duplicate observations. With that, the final “Ordinal Dataset” is a 1,246 by 20 dataset and the final “Binary Dataset” is a 1,246 by 45 dataset.

Additionally, the datasets were not scaled because binary columns do not need to be scaled. Ordinal columns also do not benefit from scaling because they are essentially already scaled based on response severity.

IV. COMPARISON METRICS

To ensure consistent measures of success and standardized points of comparison, testing accuracy scores were determined to be the ideal metric. This is because the final datasets contained equally balanced class labels. While this research project compiled confusion matrices and examined other metrics, such as recall, precision, F-1 scores, and cross validation scores, testing accuracy scores were deemed sufficient for comparison purposes.

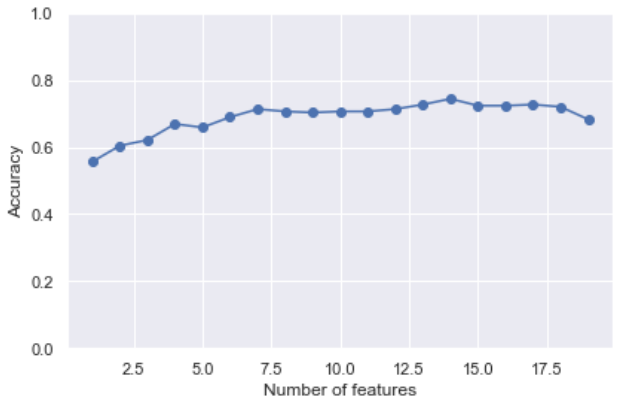
V. FEATURE SELECTION

For feature selection, this project leveraged the k-nearest neighbors (kNN), random forest (RF), and principal component analysis (PCA) algorithms to select a subset of features that were to be used as inputs in the classification models to predict a respondent's usage or non-usage of the Oxford comma.

A. kNN

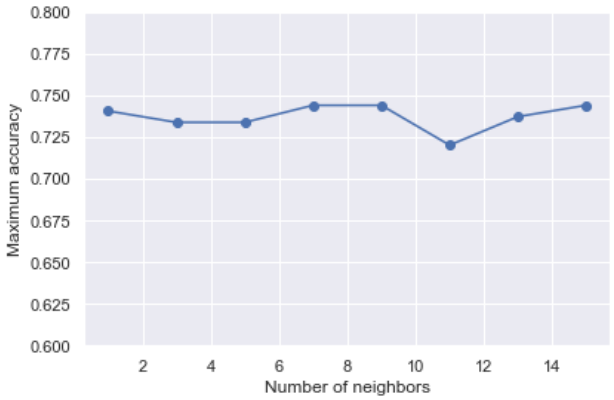
When implementing the kNN algorithm to select a subset of features on the "Ordinal Dataset," all possible combinations for *n* number of total features from 1 through 19 were tested with the training data, which consisted of 70% of the observations within the dataset, and the combination of features for each total number of features that produced the highest accuracy score was stored for examination. That said, the following graph illustrates the accuracy score obtained from the best combination for each total number of features:

Figure 1: Accuracy Scores of the Best Combination of *N* Total Number of Features



From the graph above, the following combination of 14 features was selected to be utilized as the "kNN-selected features" in the "Ordinal Dataset:" 1). Prior to reading about it above, have you heard of the serial (or Oxford) comma?, 2). How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?, 3). How would you write the following sentence?, 4). When faced with using the word "data," have you ever spent time considering if the word was a singular or plural noun?, 5). How much, if at all, do you care about the debate over the use of the word "data" as a singular or plural noun?, 6). In your opinion, how important or unimportant is proper use of grammar?, 7). Age, 8). Education, and 9). Binary indicators for "East North Central," "East South Central," "Middle Atlantic," "Mountain," "Pacific," and "West North Central." When selecting the number of neighbors for the kNN algorithm, the number of neighbors that produced the highest training accuracy was selected. The following graph illustrates these results:

Figure 2: Selecting the Number of Neighbors

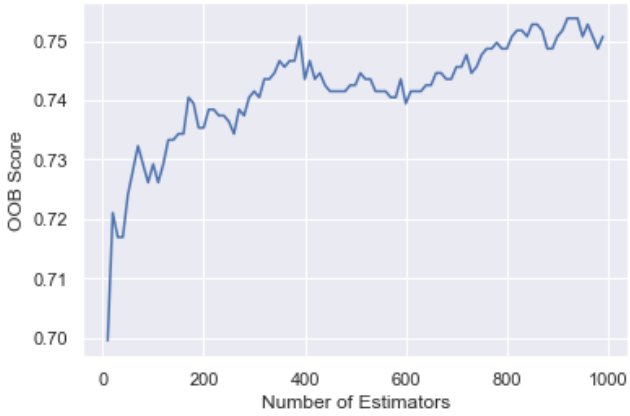


The process for selecting a subset of features for the "Binary Dataset" using kNN was the same as the one described above, except all possible combinations for n number of total features from 1 through 44 were tested, such that 14 features (different from those derived from the "Ordinal Dataset") were selected.

B. Random Forest

A second feature selection method involved the use of the RF algorithm. The number of estimators selected was based on the following graph, which plots the number of estimators from 10 to 1,000 in intervals of 10 against its corresponding out-of-basket score using the training data:

Figure 3: Selecting the Number of Estimators



Evidently, there is no major increase in accuracy or, conversely, a significant decrease in error beyond 800 estimators. Hence, this was selected as the number of estimators for the RF algorithm. With that, the RF produced the feature importance values for each of the 19 features in the "Ordinal Dataset." Setting a value of 0.05 as a threshold value, the following features were selected as the "RF-selected features:" 1). How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?, 2). Age, 3). Household Income, 4). Education, 5). How

much, if at all, do you care about the debate over the use of the word "data" as a singular or plural noun?, 6). In your opinion, how important or unimportant is proper use of grammar?, 7). Prior to reading about it above, have you heard of the serial (or Oxford) comma?, and 8). Gender.

The process for selecting a subset of features for the "Binary Dataset" using RF was the same as the one described above, except that a value of 0.02 was set as a threshold value, such that 32 out of 44 features were selected.

C. Principal Component Analysis (PCA)

The third feature selection method was through PCA. PCA is a strong feature selection method because the dimensionality of the "Binary Dataset" is quite large. As such, it was important to reduce its dimensionality, especially because the transformation of all the ordinal features into binary indicators introduced redundancy. Additionally, PCA is cited as a strong feature selection method for smaller datasets and is not significantly impacted by noise in the data.

In order to determine the optimal number of components, the variation explained by the different components was analyzed. Findings suggested that 25 components explained approximately 95% of the variance in class labels. As a result, these 25 components comprise the "PCA-selected features," reducing the "Binary Dataset" into a 1,246 by 25 dataset.

VI. CLASSIFICATION ALGORITHMS

The justification for creating the "Binary Dataset" is described here.[5] Logistic regression, kNN, and support vector machines (SVM), all three of which are implemented to predict the usage or non-usage of the Oxford comma, are classification algorithms that rely on distance metrics. However, the ordinal features, although mapped to incrementing integer values, do not represent values that have a consistent or an inherent concept of distance amongst them. For example, as is the case with the survey question, "How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?," the different answers do not hold an inherent value from which a distance can be established amongst them. As another example, as is the case with the Household Income ordinal values, there is no consistent distance that can be established between "$25,000 - $49,999" and "$50,000 - $99,999," or between "$100,000 - $149,999" and "$150,000+". With all of this in mind, the "Binary Dataset" was created to transform ordinal features into binary indicators for each of its possible values to eliminate the distance-related assumptions that the aforementioned algorithms make when solving this classification problem.

The algorithms that were implemented to predict either the usage or the non-usage of the Oxford comma for a given respondent with 1). all features, 2). "kNN-selected features, 3). "RF-selected features," and 4). "PCA-selected features" (only for "Binary Dataset") can be distinguished into two categories: a) "distance-related models" and b). "non-distance-related models." Logistic regression, kNN, and support vector machines (SVM) algorithms fall under the "distance-related models" category, whereas RF and Naive Bayes fall under the "non-distance-related models." Although the "Binary Dataset" was primarily created to implement the "distance-related models" specified above to solve this classification problem, for comparison reasons, the "distance-related models" have also been analyzed for the "Ordinal Dataset."

A. Distance Related Models

Logistic Regression

Logistic regression for classification was implemented on both the "Ordinal Dataset" and the "Binary Dataset" and tested with 1). all features, 2). "kNN-selected features," 3). "RF-selected features," and 4). "PCA-selected features." Training and test accuracy scores for both datasets are reported below:

Table 1: Logistic Regression Implementation on "Ordinal Dataset"

"Ordinal Dataset"		
Feature Selection Method	**Training Accuracy Score**	**Testing Accuracy Score**
All features	0.6953	0.6890
kNN-selected features	0.6823	0.6315
RF-selected features	0.6800	0.6531

Table 2: Logistic Regression Implementation on "Binary Dataset"

"Binary Dataset"		
Feature Selection Method	**Training Accuracy Score**	**Testing Accuracy Score**
All features	0.6835	0.6685
kNN-selected features	0.6537	0.6791
RF-selected features	0.6835	0.6684
PCA-selected features	0.6651	0.6364

Evidently, the logistic regression algorithm for this classification problem yielded higher testing accuracy scores for most feature selection methods when implemented on the "Binary Dataset," with the "kNN-selected features" having yielded the highest accuracy score for the "Binary Dataset" and all features yielding the highest accuracy score for the "Ordinal Dataset". However, it is important to note that the differences in training accuracy scores and testing accuracy scores across both datasets are fairly small. Coefficient values for each feature in each feature selection method can be examined by running the code provided. However, to highlight some interesting findings, the table below reports the

coefficients for the "RF-selected features" obtained from the implementation of logistic regression for classification on the "Ordinal Dataset:"

Table 3: Logistic Regression with Random Forest: Coefficients[6]

Feature	Coefficient Value
Prior to reading about it above, have you heard of the serial (or Oxford) comma?	0.2206
How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?	0.7405
How much, if at all, do you care about the debate over the use of the word "data" as a singular or plural noun?	-0.1556
In your opinion, how important or unimportant is proper use of grammar?	-0.0779
Gender	-0.0106
Age	-0.4684
Household income	-0.1588
Education	0.3183

To interpret a few coefficient values, the sign for the first feature is positive, for example, which shows that, all else equal, respondents who have heard of the Oxford comma prior to reading about it from the survey are more likely to use the Oxford comma than respondents who have not heard of the Oxford comma prior to reading about it from the survey. The sign for the Gender feature is negative, which shows that, all else equal, female respondents are less likely to use the Oxford comma than male respondents. As for the ordinal features, the positive coefficient value of the Education feature, for example, means that the effect of a respondent having the highest level of education (Graduate degree) is greater in magnitude than the effect from the first feature being "Yes" or the Gender feature being "Male."

kNN

Similar to the logistic regression algorithm, kNN was implemented on both the "Ordinal Dataset" and the "Binary Dataset" and tested with 1). all features, 2). "kNN-selected features," 3). "RF-selected features," and 4). "PCA-selected features." Training and test accuracy scores for both datasets are reported below:

Table 4: kNN Implementation on "Ordinal Dataset"

"Ordinal Dataset"		
Feature Selection Method	**Training Accuracy Score**	**Testing Accuracy Score**
All features	0.7487	0.7010
kNN-selected features	0.7590	0.6555
RF-selected features	0.7600	0.6627

Table 5: kNN Implementation on "Binary Dataset"

"Binary Dataset"		
Feature Selection Method	**Training Accuracy Score**	**Testing Accuracy Score**
All features	0.7362	0.6578
kNN-selected features	0.7362	0.6390
RF-selected features	0.7362	0.6417
PCA-selected features	0.6823	0.6150

In order to avoid using the same training data that yielded the "kNN-selected features" when training the kNN algorithm to solve this classification problem, different training and testing datasets were sampled from the "Ordinal Dataset" and "Binary Dataset" and utilized with the previously discussed subset of features to obtain the results shown above (random_state=0 to obtain the subset of features and random_state=123 to train and test the kNN algorithm).

Surprisingly, the training and testing accuracy scores obtained through the implementation of kNN on the "Binary Dataset" were generally not better than those obtained through its implementation on the "Ordinal Dataset." This contradicts the initial belief that transforming ordinal features' values into binary indicators would improve the performance of all "distance-related models." However, it is important to note, again, that the results presented above are solely based on the datasets that were created in this paper's particular instance of the oversampling method, which, due to its randomness, could potentially change the results when run again if different observations are duplicated from the minority class. Overall, although training accuracy scores from the kNN implementation were generally higher than those obtained from the implementation of logistic regression, testing accuracy scores were generally lower than those obtained from the implementation of logistic regression.

Support Vector Machine (SVM)

The SVM model was utilized because it is one of the strongest models that allows for flexibility in data distribution. Since both datasets have many dimensions, it made sense to sample potential kernels to ensure the boundary is separating the data as much as possible. Recognizing that the data is likely not perfectly separable, C and γ were chosen based on a grid search as a method to tune the hyperparameters. After testing these hyperparameters and various kernel options on both datasets, along with the inclusion of all feature selection models, results were mixed. RBF and Linear were the most common kernels. This is likely because the data is very difficult to map into a cleaner space and, therefore, difficult to separate using SVM. With that being said, the highest testing accuracy score was obtained from an RBF kernel with all features for the "Ordinal Dataset" and a Linear kernel with "kNN-selected features" for the "Binary Dataset." Training and test accuracy scores for both datasets are reported below:

Table 6: SVM Implementation on "Ordinal Dataset"

"Ordinal Dataset"		
Feature Selection Method	**Optimal Kernel**	**Testing Accuracy Score**
All features	RBF	0.6819
kNN-selected features	RBF	0.6819
RF-selected features	Linear	0.6651

Table 7: SVM Implementation on "Binary Dataset"

"Binary Dataset"		
Feature Selection Method	**Optimal Kernel**	**Test Accuracy Score**
All features	Linear	0.6818
kNN-selected features	Linear	0.6898
RF-selected features	Linear	0.6738
PCA-selected features	RBF	0.6524

B. Non-Distance Related Models

Random Forest

As the first "non-distance-related" model, expectations for the RF algorithm were high in regard to its performance on the "Ordinal Dataset" and the "Binary Dataset." Similar to the cases of the "distance-related" models, RF was implemented on both the "Ordinal Dataset" and the "Binary Dataset" and tested with 1). all features, 2). "kNN-selected features," 3). "RF-selected features," and 4). "PCA-selected features." Training and test accuracy scores for both datasets are reported below:

Table 8: Random Forest Implementation on "Ordinal Dataset"

"Ordinal Dataset"		
Feature Selection Method	**Training Accuracy Score**	**Testing Accuracy Score**
All features	0.9918	0.7512
kNN-selected features	0.9569	0.7153
RF-selected features	0.9467	0.7033

Table 9: Random Forest Implementation on "Binary Dataset"

"Binary Dataset"		
Feature Selection Method	**Training Accuracy Score**	**Testing Accuracy Score**
All features	0.9920	0.7353
kNN-selected features	0.8704	0.6791
RF-selected features	0.9920	0.6471
PCA-selected features	0.9794	0.6765

As was the case with kNN, in order to avoid using the same training data that yielded the "RF-selected features" when training the RF algorithm to solve this classification problem, different training and testing datasets were sampled from the "Ordinal Dataset" and "Binary Dataset" and utilized with the previously discussed subset of features to obtain the results shown above (random_state=0 to obtain the subset of features and random_state=123 to train and test the kNN algorithm).

As suspected, the training accuracy scores obtained through the implementation of the RF algorithm on both datasets were the highest out of all the other models examined thus far. This is likely because it does not rely on a distance metric to classify respondents as Oxford comma users or non-Oxford comma users. As for the testing accuracy scores, they tend to be on the higher side compared to those obtained from the implementation of "distance-related models" and, in general, are subject to less variation across the different feature selection methods. The RF algorithm yielded the highest values and outperformed the "distance-related models" on every occasion.

Naïve Bayes

The last model analyzed was naive bayes, primarily because it performs well with categorical data. It is also simple and the algorithm does not rely on distance metrics. Since Naive Bayes assumes all features are independent and works well with categorical data, this algorithm was only run on the "Ordinal Dataset." The assumption of feature independence is usually a decent assumption, however, many of the features had high multicollinearity values. As a result, features with redundancy were eliminated, including 1). How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?, 2). How much, if at all, do you care about the debate over the use of the word "data" as a singular or plural noun?, 3). In your opinion, how important or unimportant is proper use of grammar?, and 4). Education. This yielded VIF values that were less than 5, which is the typical threshold for multicollinearity. Variables with extremely unbalanced categories were also removed to reduce bias and to maximize information gain.

Naive Bayes can take many approaches, such as Categorical, Bernoulli, and Gaussian distributions. All three were tested to ensure the results made sense, and it was found that Categorical Naive Bayes provided the highest testing accuracy score, as expected. The testing accuracy score, F-1 score, precision score, and recall score were all 67%. Naive Bayes was predicted to be one of the better algorithms due to it also being a "non-distance-related model," but it presented the worst testing accuracy score out of all the models examined thus far. The testing accuracy and number of misclassified samples are presented below:

Table 10: Naïve Bayes Implementation on "Ordinal Dataset"

"Ordinal Dataset"		
	Testing Accuracy Score	**Misclassified Samples**
Categorical	.6700	139
Bernoulli	.5700	179
Gaussian	.5900	173

VII. MODEL COMPARISONS

Table 11: Model Comparisons on "Ordinal Dataset"

"Ordinal Dataset"		
Model	**Feature Selection Method**	**Testing Accuracy Score**
Random Forest	All Features	0.7512
kNN	All Features	0.7010
Logistic Regression	All Features	0.6890
SVM - Linear	All Features	0.6819
Naive Bayes - Categorical	No Multicollinearity	0.6700

Table 12: Model Comparisons on "Binary Dataset"

"Binary Dataset"		
Model	**Feature Selection Method**	**Testing Accuracy Score**
Random Forest	All Features	0.7353
SVM - Linear	kNN	0.6898
Logistic Regression	kNN	0.6791
kNN	All Features	0.6578

Another comparison metric that can be helpful is cross validation accuracy. Cross validation average testing accuracy and standard deviations were obtained using 10-Fold Cross Validation for the best version of each model (i.e. optimal feature selection method and parameters). While the cross-validation scores tended to be slightly lower than the testing accuracy scores, the trends were quite similar and Random Forest remained the strongest performer. A boxplot of these comparisons is shown below:

Figure 3: Algorithm Comparisons on "Ordinal Dataset"

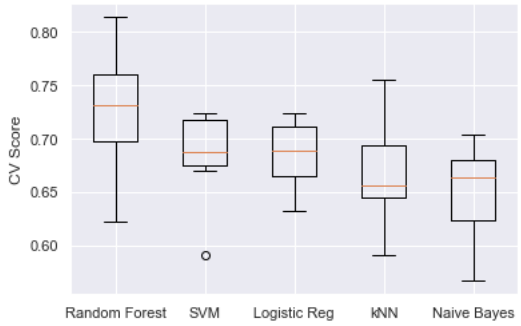
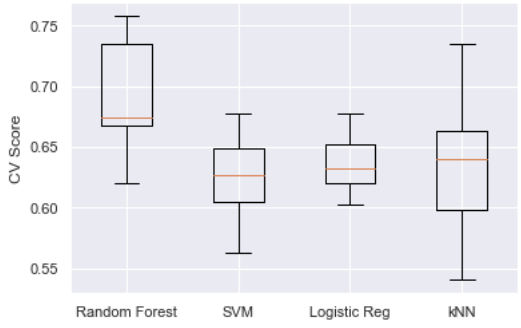


Figure 4: Algorithm Comparisons on "Binary Dataset"



VIII. DISCUSSION

Some drawbacks to the methods used above include the inherent bias in surveys and limited data. Since the dataset was originally distributed as a survey, there are likely some limitations to its efficacy. The first being the missing values. Part of the preprocessing included taking out rows with excessive missing data, which may not have been completely random. As a result, there may be fewer educated people or lower income people who finished the survey, which induces some bias. While the income distribution was normal, there could have been some bias in other features. Additionally, survey data comes with inherent issues such as people answering dishonestly. Some people may not know their exact income, or, for example, they do not know the difference between "plural" and "singular" and end up taking a guess as to how the word "data" should be interpreted. These are just a few examples of how surveys may not give accurate responses. Nonetheless, survey data can still be analyzed as long as this bias is minimal, which can be assumed due to the nature of the survey.

Not only could the data be inherently biased, but it is also quite limited. When it comes to the decision of using or not using an Oxford comma, many factors may play into that decision that were not included in the dataset, such as a respondent's school curriculum, respondents' parent's writing styles, or respondents' consumption of different media types. Not only could important factors be omitted, but the included factors may be limited as well. For example, the age feature was in ranges such as "45-60" or "60>". These intervals are not only uneven, but also spanning multiple generations within them. An 80 year old may have a very different perspective on grammar than a 61 year old. While the data could likely be improved through more continuous features, it still provided important information that could eventually lead towards a classification model. The model found some success with a maximum testing accuracy score of around 75%, however, this accuracy score is still low relative to typical machine learning success metrics. The low obtained scores can be interpreted as the inability for the models to correctly classify a person as somebody who uses the Oxford comma at least 25% of the time. Overall, these low scores can likely be attributed to the limitations discussed above, meaning the model could be approved through more extensive data collection.

IX. CONCLUSION

Once again, experimentation with different random_state values during the oversampling and the data partitioning processes found that results may vary from those presented above. However, and perhaps most importantly, the RF algorithm produced the most consistent and accurate results over the other algorithms that were implemented in this research project. In other words, the RF algorithm is ultimately deemed the best machine learning algorithm to solve this specific classification problem. Even when different number of neighbors and number of features had to be selected to produce the "kNN-selected features," or when different number of estimators and number of important features had to be selected to produce the "RF-selected features," the RF algorithm maintained a degree of consistency and high performance that other algorithms, especially kNN and SVM, were not able to attain. An examination of Figure 3 and Figure 4 further supports this conclusion, as the median testing CV accuracy score and the interquartile range of CV accuracy scores obtained through the implementation of the RF algorithm on the "Ordinal Dataset" and the "Binary Dataset" are higher than those obtained through the implementation of other algorithms.

X. CONTRIBUTIONS

Work was evenly split.

Jae Hyun Hong:

1. Creating ordinal dataset
2. Logistic Regression
3. Random Forest (model and FS)
4. kNN (model and FS)
5. Writing: Above models, data, preprocessing, conclusion

Hadley Kruse:

1. Creating binary dataset
2. PCA
3. Naïve Bayes
4. SVM
5. Writing: Abstract, intro, related work, above models, and discussion

Code is available here.

XI. REFERENCES

1 Lamberg, Jasso. "Origin of the Oxford Comma." Comdesres, 23 Feb. 2015, comdesres.com/origin-of-the-oxford-comma/.

2 Benson, Lindsay A. "A Lack of an Oxford Comma Cost Dairy $5 Million." CNN, Cable News Network, 10 Feb. 2018, www.cnn.com/2018/02/09/us/dairy-drivers-oxford-comma-case-settlement-trnd/index.html.

3 Hickey, Walt. "Elitist, Superfluous, or Popular? We Polled Americans on the Oxford Comma." FiveThirtyEight, FiveThirtyEight, 17 June 2014, fivethirtyeight.com/features/elitist-superfluous-or-popular-we-polled-americans-on-the-oxford-comma/.

4 Raschka, Sebastian. "Python Machine Learning - Code Examples." Github, 2017, github.com/rasbt/python-machine-learning-book/blob/master/code/ch04/ch04.ipynb.

5 Cheng, Xi. "Preprocessing of Categorical Predictors in SVM, KNN and KDC (Contributed by Xi Cheng)." Statistics LibreTexts, Libretexts, 17 Aug. 2020, stats.libretexts.org/Bookshelves/Computing_and_Modeling/RTG%3A_Classification_Methods/4%3A_Numerical_Experiments_and_Real_Data_Analysis/Preprocessing_of_categorical_predictors_in_SVM%2C_KNN_and_KDC_(contributed_by_Xi_Cheng).

6 Bock, Tim. "How to Interpret Logistic Regression Coefficients." Displayr, N.d., www.displayr.com/how-to-interpret-logistic-regression-coefficients/.