Gender Disparities in Written Expression: Predicting Gender in Online Teacher Applications

Divya Krishnan
Ridge High School (NJ)
reachdkrishnan@gmail.com

Research conducted under supervision of Dr. Suma Bhat, University of Illinois at Urbana-Champaign

ABSTRACT

Although the disparity of gender in the workplace has long been analyzed, linguists continue to research the semantic norms that differentiate texts between female and male writers. In this paper, we analyze gender predictions on a teacher application forum known as DonorsChoose. DonorsChoose is a nonprofit that connects teachers to donors for in-class projects. Using content heavy fields such as essays, we analyze how part of speech can differentiate text written by male versus female authors, as well as the broader impact that such understandings can have.

Our approach identified features using natural language processing via the spaCy toolkit, a tokenizer to break large chunks of text into digestible portions, and a lemmatizer to extract the base form of each token and the part of speech tag. The machine learning model used was a decision tree/random forest classifier. Ground truth was ascertained using self-identified prefixes (like Mr., Ms.) in the data. Initial results show an accuracy of 0.91 with an F1-score of 0.56. Furthermore, we found that a greater number of first person pronouns may be indicative of a female writer. Future work includes use of extraneous gender data and variations of content-related features, such as sentiment and style, to improve accuracy.

INTRODUCTION

What is DonorsChoose?

DonorsChoose is a non-profit organization based in the US that provides necessary educational resources to high-need communities. By building a platform that connects public school teachers with donors that can fund classroom projects, the nonprofit organization works to enhance educational opportunities in lowly funded areas. When they need extraneous funding for a project, teachers are able to submit an application with relevant information such as the type of resources they need, as well as essay questions detailing information about their students and project. After each request is made, DonorsChoose accepts certain applications that are then funded by a community of generous donors. The application information, applicant's details, and the acceptance of the donation are all made public.

In order to better understand the demographics of accepted donations, Kaggle compiled the application information from DonorsChoose. The information is anonymous, and each request name is substituted for a unique donor ID.

What is the problem?

The problem set falls into two main categories, each which can be described through an indication of their target, or prediction, variables. Both the nature of these problem sets and the relevance of a target variable will be further analyzed.

1. Project_is_approved
2. Teacher_prefix

1. After the relevant information and essays are provided by each teacher looking to fund a classroom project, the application is screened by DonorsChoose and certain applications are accepted. On the open dataframe for application screening, the boolean column, “Project_is_approved,” details this decision. When exploring this column as a prediction factor, the goal becomes to analyze the relevant features in the other application data that could affect this column, such as prefix, state, or essay. Given this information, predictions could be made on unlabeled data to determine if an application will be accepted or not. However, one possibly relevant feature, “teacher_prefix,” remains slightly ambiguous when it comes to gender, which could be a key factor when analyzing approval. This brings about the second subset of the problem.

2. The “teacher_prefix” field provides details about the gender of the applicant through prefixes such as “Mrs.” and “Ms.,” which indicates a female applicant, and “Mr.,” which reveals a male applicant. However, when exploring the data, it was found that there are close to four thousand genders that are unknown through prefixes such as “teacher” and “Dr.”

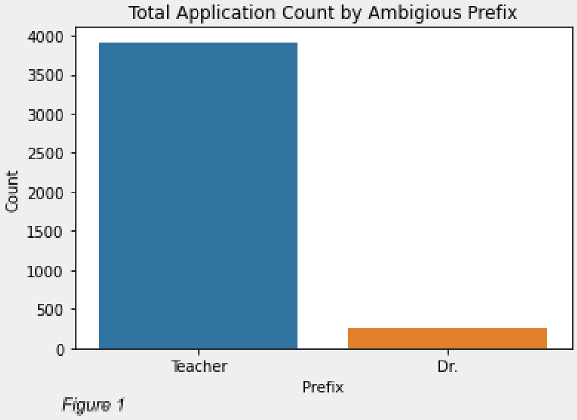


Figure 1

In Figure 1, it is evident that four thousand applicants labeled their prefix-fields as “Teacher” or “Dr.,” both gender ambiguous. In order to predict the gender of these applicants, the field that we explored was the “Essay” prompt. Each essay was lined with the individual writers’ unique semantics and choices - choices that could yield valuable information as to the gender of the applicant.

MOTIVATION

Why is it important?

The question then arises as to why predicting this field is important. It’s relevance lies both in the DonorsChoose context and in the general field of gender study. This analysis can aid the prediction of whether a project is approved. With an added feature that could make an effect on the target variable “project_is_approved,” the chance increases of correctly predicting that field. In the overall view of gender studies, recognizing gender biases in texts is a developing research field that remains widely unknown. Developing an identification to differentiate male and female writers could aid in gender bias research.

What are some challenges?

As an ever evolving field, gender analysis is ridden with ambiguity. Fully perfecting gender prediction in a subset is near impossible, as it requires labeling texts with overarching stereotypes or patterns, when all examples may not fit that same convention. This applies for most prediction algorithms.

The second challenge that arises is the lack of given feature fields. Since each feature is curated to answer the question "project_is_approved," new feature vectors must be created to identify gender.

One last factor that poses as a challenge is the test set. The test set is made up of applicant prefix data who voluntarily chose to put down "Dr." or "Teacher." Choosing to apply "Dr." indicates either a PhD or an MD. Are there observable gender biases in the degree itself?

In a male dominated education spectrum, are men more likely to apply as PhD? Research from the past eleven years actually points to the contrary, however, with a higher percentage of the doctoral degree graduates given to females than males. How could this have an impact on the data? Similar questions arise when addressing the "Teacher" column. As either a method to hide the applicant's gender, or hide their marital status, this voluntary decision may also have some skew. In this set of data, the other columns or sets of possible skewing information are removed, so only the essays can be analyzed. Since no further data on these questions is given, this challenge poses the problem of adding slight uncertainty to the test results.

What are your hypotheses?

The gender analysis features include various part-of-speech counts, including first, second, and third person pronouns, superlative, comparative, and positive adjectives, as well as nouns and verbs. (The full list of feature vectors and predictor variables is available later). Certain hypotheses were created off of just the initial features and how these counts can affect the overall connotations and the reader's impression of the text. For example, a high count of first person pronouns may give the writing a greater perspective of the writer, making the text more personal. A greater use of second person pronouns may make the writer more direct, such as a clear request from the reader to either accept the application or fund the project. Superlative adjectives may make the writer seem more assertive. Verbs may tilt the text toward an action-oriented approach; nouns may point toward a more descriptive approach.

EXPLORATORY ANALYSIS

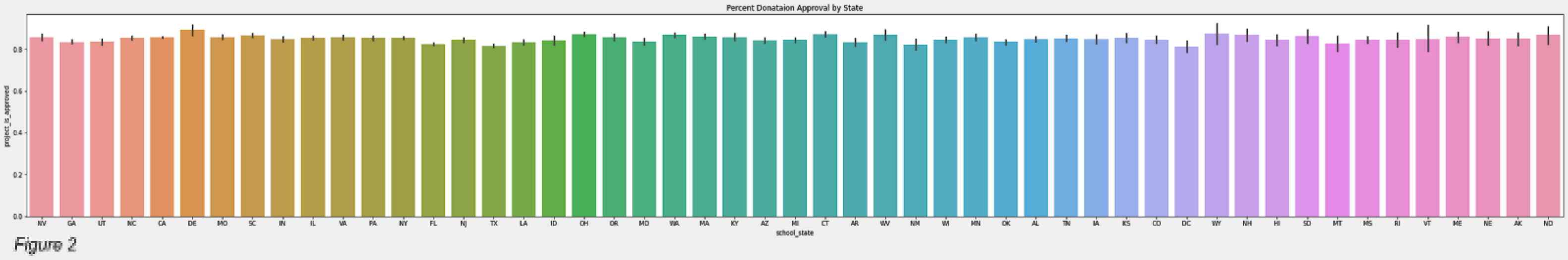


Figure 2

What trends are noticeable in the data?

Before analyzing the data for the gender prediction however, the data must first be grouped and visualized to notice any notable trends. To understand the relevance of this data visualization, the first subset of the problem mentioned above deals with multiple factors. In order to identify these and notice trends in acceptance, Seaborn, a python library built on Matplotlib, was used. This allows for easy data visualization and therefore quick identification of relevant features

The list of features in the DonorsChoose data includes, but is not limited to:

school_state
teacher_prefix
project_submitted_datetime
project_grade_category

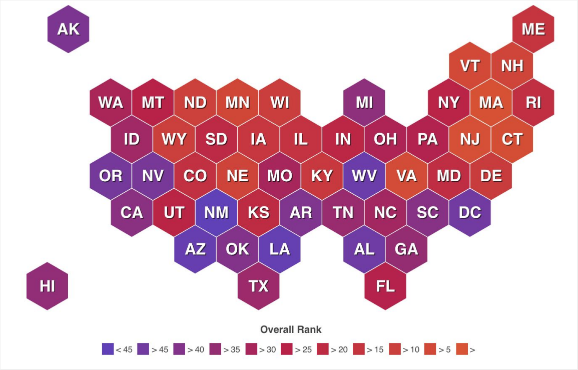


Figure 3

Figure 2 represents the approval rate and its changes based on the state of the applicant, which is labeled on the X-axis. It seems as though there are only minimal shifts in approval rate. This initially gives the impression that data on the applicant's state may not be representative of an adequate prediction feature for approval. However, it is important to note that grouping states together may lead to results, whether that is by socio-economic status, academic ranking, or average standardized test scores, or other.

Figure 3 is a heat map that ranks preliminary education on a state by state basis. Although this mixes different metrics to calculate the rankings, another scale that only factors one metric into the identification of a state ranking might help to visualize a pattern, if it exists. This might be useful for a later investigation.

Another basis to graph the data by could be a grouping between the date. For example, in Figure 4, the heat map has the state labeled on the Y-axis, while the date (month) is on the X-axis. The pigment of the box indicates the percent approval of the item at a certain date in each specific state. The lighter the column, the lower the percentage of project approval.

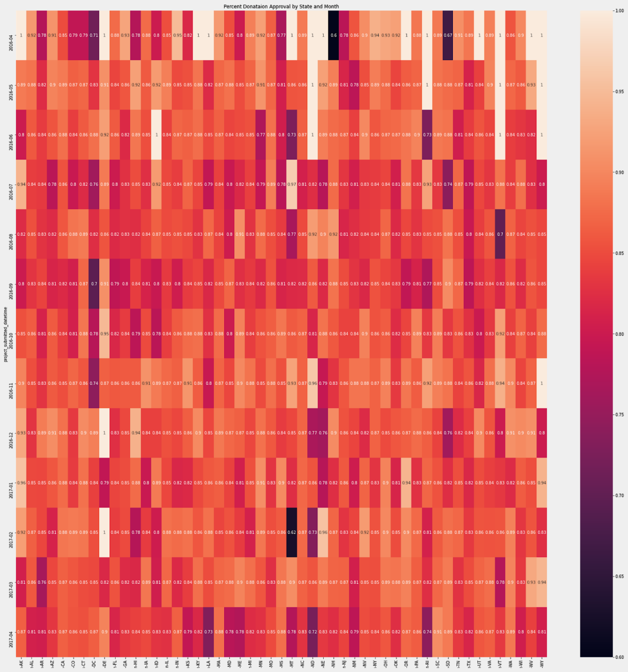


Figure 4

Although there are some trends visible, there aren't clear visual indications of a sway in one direction or another. However, there are some rows or columns that are darker or contain a larger number of lighter boxes.

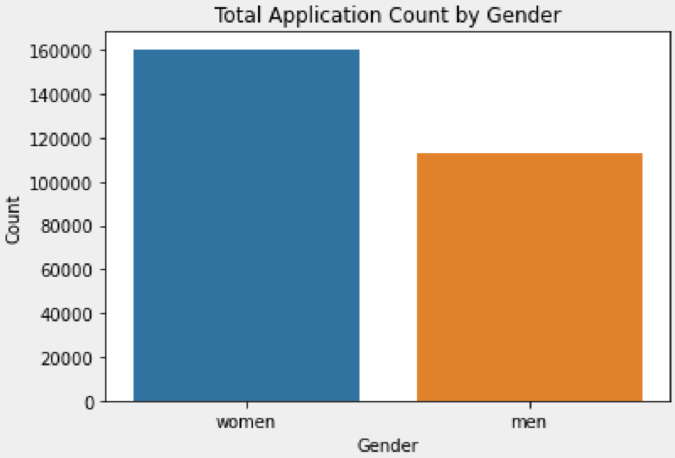


Figure 5

Another feature graphed was the gender, that was dependent on the prefix. For example, "Mrs." and "Ms." would indicate a "Female," while "Mr.," would indicate "Male."

The bar graph in Figure 5 shows a clear indication that a higher percentage of the applicants were female. The X-axis reveals the gender of the applicant and the Y-axis shows the count for each gender. However, this above graph removes the prefixes "Teacher" and "Dr." In order to categorize these into a specific gender, we require prediction with the essays.

PREVIOUS RELATED WORK

What has been done before to address the challenges?

Gender analysis to identify both bias and gender prediction in texts is a long analyzed field, but it is still researched, as information and observations are ever-evolving. Some notable gender prediction studies I observed when looking into my DonorsChoose gender analysis include the following:

"Author Profiling: Predicting Age and Gender from Blogs" K Santosh, Romil Bansal
"Gender Differences in Written Expression at the Elementary Level" Ashley D. Melloy
'Examining Gender Differences in Writing Skill with Latent Factor Modeling" Laurel Woods
(All of the above are graduate papers with similar tasks: predicting the gender of a writer.) Each of these studies touched on slightly different factors of gender analysis. K Santosh's team worked to profile different anonymous blog writers and analyze both their age and their gender. His features included content based analysis (similar to the current project), style based analysis, and topic based. As these anonymous bloggers had no prompt to style their writing, both topic and style were appropriate measures to look into.

Melloy's analysis was more centered on the academic abilities of elementary students and gender prediction at that age. Her features focused on length, spelling, and word sequence. Her final results were both a prediction of gender and an understanding of the developmental stereotypes in children.

Laurel's analysis paralleled Melloys in that the results would also be indicative of performance skill when comparing male and female writers. Given the gender at elementary levels, Melloy sought to analyze how that gender effected the skill of their writing.

Each of these previous projects has analyzed and researched the factors of expressive writing that differ a male writer from a female writer. Using this outside data, we will later formulate our quantitative results and qualify the results.

METHOD, RATIONALE & APPROACH

Creating the DataFrame

Solving this problem first began with creating the data set that would serve as the feature variables and predictor variables. These would include the following:

First_person_pron

Second_person_pron

Third_person_pron

Superlative_adj

Comp_adj

Positive_adj

Verbs

Nouns

Essay Length

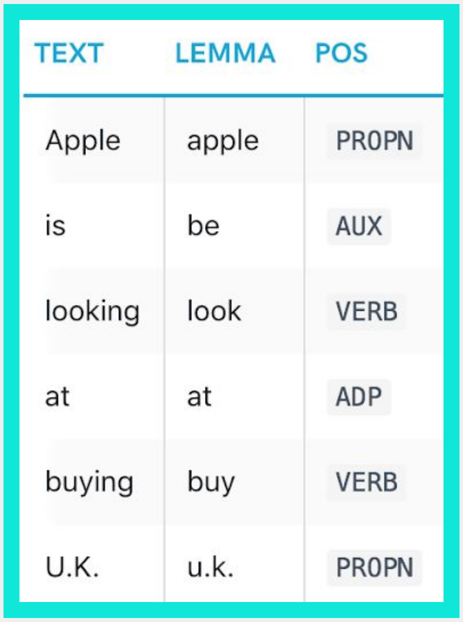
Gender

(Each of the initial eight columns listed above are a count total of the listed quantity)

To identify these variables, I utilized an advanced natural language processing toolkit known as SpaCy. The tokenizer broke the large chunk of text into digestible portions, the lemmatizer took the base form of each token, and the part of speech tag listed the part of speech of each lemma.

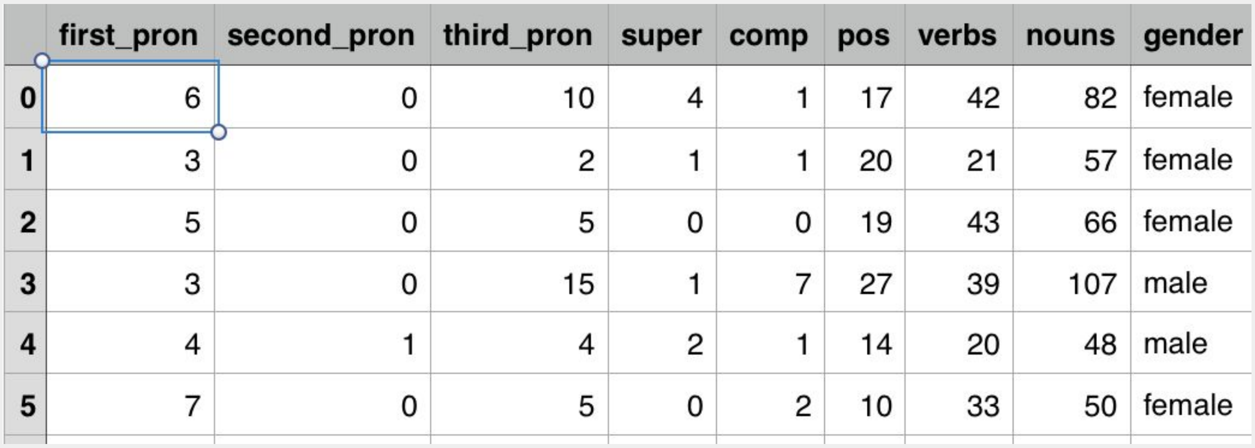
Figure 6 is a visual depiction of how SpaCy would break up a chunk of text. However, certain sets of the feature variables require more than just a simple “ADJ,” such as superlative, comparative, and positive adjectives. SpaCy has an additional tag function that gives such details about the tokens. This information contains details such as “superlative,” “comparative,” and “positive.”

Given all of this data, I created a dataframe that I used to run a machine learning model to predict the gender.



TEXT	LEMMA	POS
Apple	apple	PROPN
is	be	AUX
looking	look	VERB
at	at	ADP
buying	buy	VERB
U.K.	u.k.	PROPN

Figure 6



	first_pron	second_pron	third_pron	super	comp	pos	verbs	nouns	gender
0	6	0	10	4	1	17	42	82	female
1	3	0	2	1	1	20	21	57	female
2	5	0	5	0	0	19	43	66	female
3	3	0	15	1	7	27	39	107	male
4	4	1	4	2	1	14	20	48	male
5	7	0	5	0	2	10	33	50	female

Figure 7: an example of the first five lines representative of the 180K row dataframe.

The approach to solving the problem.

The next objective was to split the data between a training and validation set, then to run this model through a machine learning algorithm. This is a classification problem as the predicted results would be a boolean with values, “Female” or “Male.” Decision Tree or Random Forest could be used for this type of prediction. Both algorithms are available through the scikit-learn API. The algorithm splits the tree at each node, creating branches that answer “yes” or “no” questions; it then determines the final prediction at the leaf. A random forest algorithm takes an average of multiple trees to make a prediction.

Depending on how accurate the model was, we would tweak hyper-parameters such as max depth, min sample size, and training and testing split to determine the model with the highest accuracy.

How do you measure success?

Given that we will be tweaking the hyperparameters to get the most effective model, the question comes up as to how to measure success. Machine learning models have various methods to measure success, some of which were utilized, and others of which can be used later.

The following are popular metrics used to assess success.

1. Accuracy Score
2. Precision/Recall
    a. False Positive
    b. False Negative
    c. True Positive
    d. True Negative
3. F1 Score
4. Specificity and Sensitivity

1. The Accuracy Score is the most common measure used to show the effectiveness of a prediction. This calculates the total correct predictions and creates a simple ratio that compares the total correct predictions to the total predictions.

2. Precision and recall are measures that take into account the relative instances with false and true positives and negatives.

In terms of false and true positives and negatives, the following provides a description for their meanings with boolean data.

True positives:

identified as positive and correct

False positives:

identified as positive and incorrect

True negatives:

identified as negatives and correct

False negatives:

identified as negatives and incorrect

Precision measures true positives over true positives+false negatives. (This is correctly predicted positives over all positive predictions).

Recall measures the ratio of true positives over true positive+false positives. (This is correctly predicted positives over all positive predictions).

3. The F1 score also takes into account the precision and accuracy. The F1 score measures the harmonic mean of precision and accuracy.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

4. Specificity and sensitivity similarly deal with the aforementioned categories that are also pictured in Figure 8. Sensitivity measures the ability to measure the true positives, while specificity measures the ability to know the true negatives. These measures are also used when data is skewed in a certain direction. This measure was not utilized for the following experiment, but given that the data is skewed, this may be a valuable metric to later analyze. Figure 9 shows that the data that separates the number of applications that were approved and those that were not approved is monumental. Since the data is skewed toward that direction, an analysis on that field would require either a balancing of the data or an indication of skewed predictions by a low specificity or sensitivity.

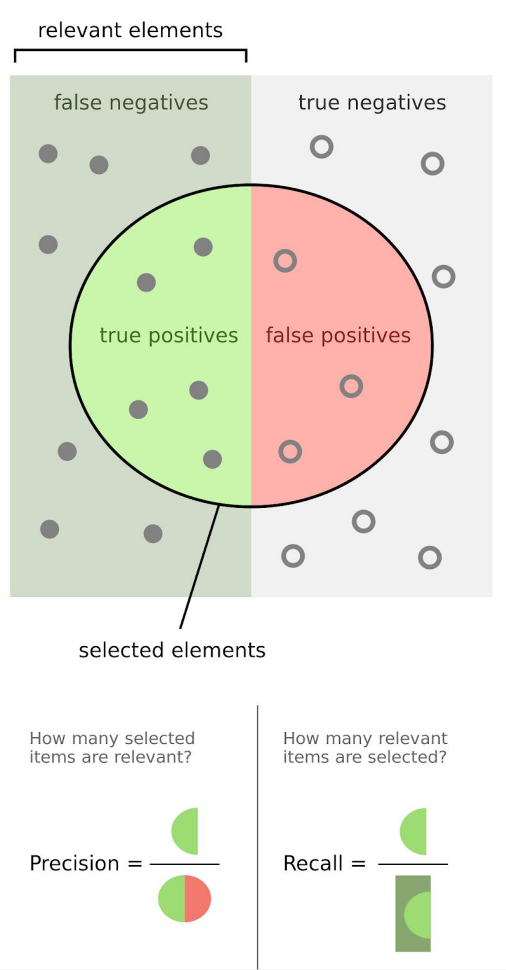


Figure 8 (credits Wikipedia)

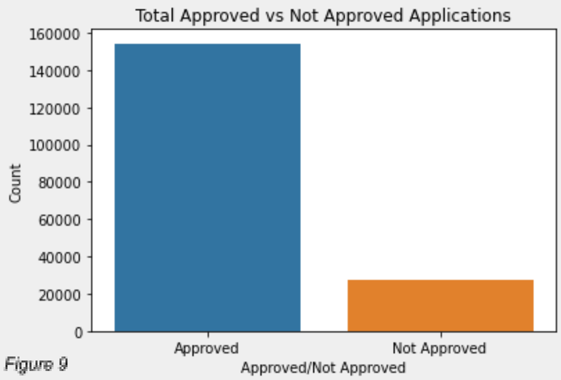


Figure 9

Similarly, the gender data also has a light skew. As seen before, the graph in Figure 10 represents the difference between male and female applicants.

However, this contrast is not as drastic. Therefore, although there is a difference and this should be taken into account, it's effects should not deter the research.

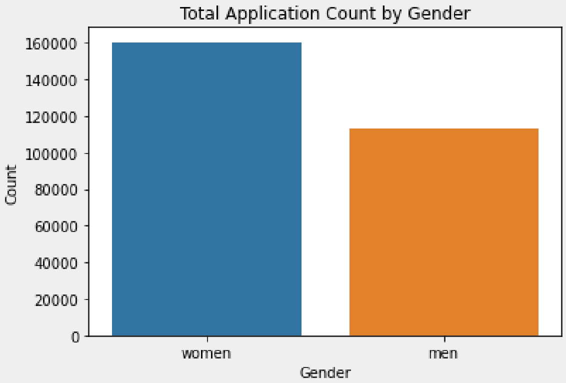


Figure 10

The data from the gender dataframe was split into a testing and training split initially with a 75% training 25% validation split. The data was then utilized to create and run a decision tree algorithm with the default hyperparameters. When the algorithm was run on the validation data, it attempted to predict the boolean value of gender.

As mentioned before, each tree can be tweaked with various hyperparameters.

The main ones that were utilized are listed below:

1. Max_depth
2. Min_leaf_samles
3. Validation and training split

1. The max_depth shows how deep the tree goes. The length of the tree and how many times it splits can be dictated by this.

2. The min_leaf_samples shows how many samples are at least required for a prediction to be made.

3. The validation and training split allow for the percent of data that is placed in validation and that placed in training to be split differently based on the user's needs.

What did you measure your approach with?

Each of the accuracy metrics would detail how well the tree was working given the hyperparameters assigned. The model usually fits under three main or broad categories.

1. Underfitting
2. Balanced
3. Overfitting

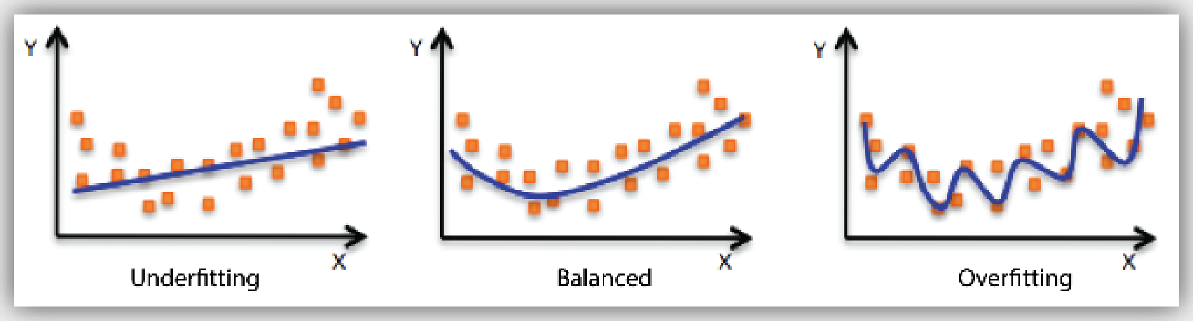


Figure 11: Example of underfitting, balanced, and overfitting curves (credits Medium)

1. Underfitting shows that the model is unable to accurately fit to the data and is therefore unable to predict the new data.
2. Balanced shows that the model reads the overall curve or pattern and predicts new data accordingly.
3. Overfitting shows that the data follows every small trend and gets confused by noise in the data. It will therefore be unable to capture trends and will be unable to predict new data well.

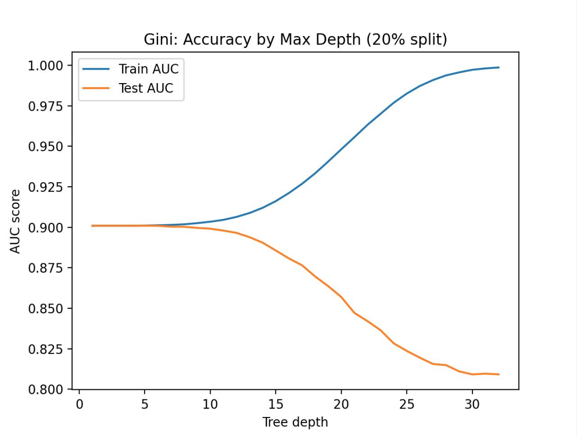


Figure 12

The first accuracy metric used was the accuracy score, and the first hyperparameter tweaked was the max_depth.

As seen in Figure 12, the data revealed a 90% accuracy with a maximum tree depth around five, but showed that it was overfitting by a tree depth of ten. It was correctly predicting the training data, but was unable to accurately predict new data in the validation split.

The next hyperparameter tweaked was min leaf samples.

Figure 13 has data that shows how the accuracy score would flatline after around five, but was inaccurately predicting the data for newer values below this minimum number of leaf samples.

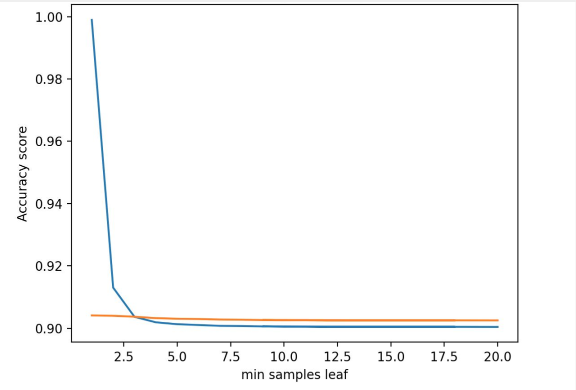
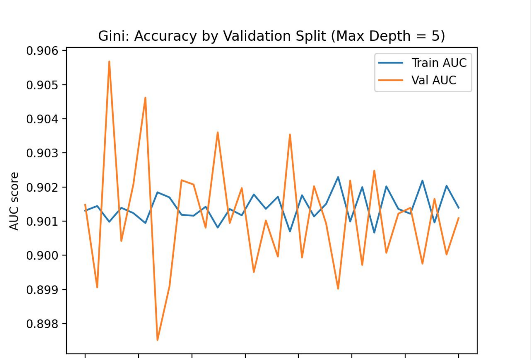


Figure 13



The final value that was tweaked was the split between the training and the validation split.

Figure 14

Although the data initially seemed to be noise (seen in Figure 14), it was later identified that the noticeable trends was due to a zoomed in flat line that was adequately noticing general patterns.

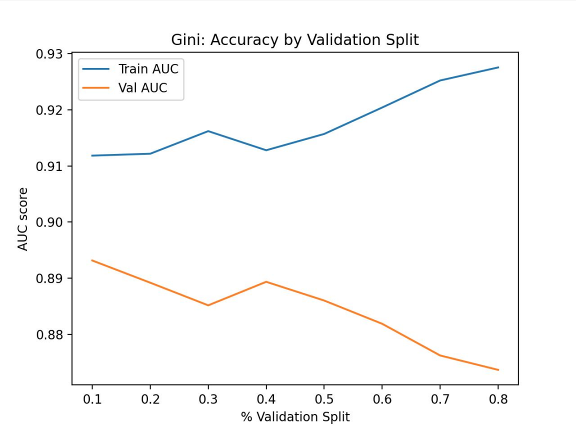


Figure 15

The graph that was later utilized was shown in Figure 15 after the maximum depth was increased. This showed clearer trends.

Another hyperparameter that was altered as well as was the method by which the Decision tree chose to split. There are two choices for this.

1. Gini
2. Entropy

This seemed to have a minimal impact on the accuracy score, except for a few noticeable shifts.

As seen in the small graphs in Figure 16, the difference between the "Entropy" (left) and "Gini" (right) is minimal, as the entropy graph only seems to overfit at around eight, while the gini graph splits around five.

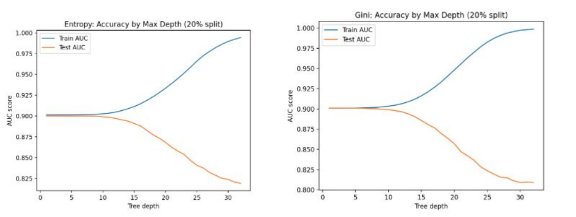


Figure 16

The data was then balanced to view which skews were caused by an unbalanced data set. This was done by deleting "female" columns to make the number of male and female columns similar. This did not have monumental impacts on the data.

RESULTS AND IMPROVEMENTS

How can the method be redone and edited given the results?

The data and the model were then reanalyzed given the accuracy graphs of the decision tree model. The same was completed with a Random Forest model. Hyperparameters were tweaked, and both the accuracy and F1 scores were created.

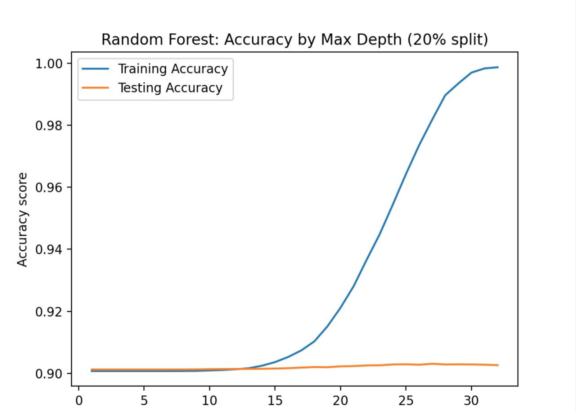


Figure 17 is the graph of Random Forest accuracy by maximum depth. By observing this data, it is evident that although the accuracy scores are similar around 90%, the overfitting of the data

Figure 17

only begins around 15. This shows the benefits of Random Forest, which aids in balancing an overfit Decision Tree.

Figure 18 shows the graph of F1 score for Random Forest by maximum depth. The average F1 score was around 0.56 a for max depth of around 10. This value is not perfect, since a value closer to 1 is more accurate, but it still reveals that the precision and recall is adequate.

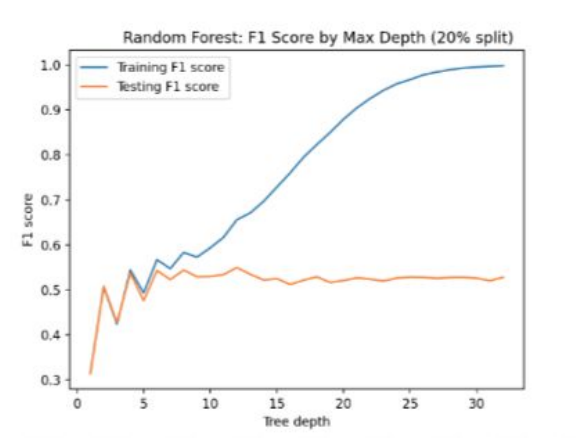


Figure 18

Similarly, graphs were created for the F1 and accuracy scores for variations of the validation and training split.

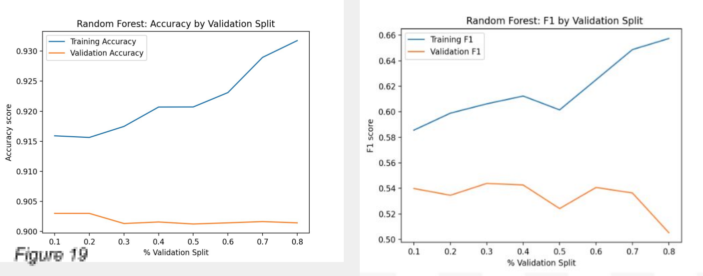


Figure 19

The graphs in figure 19 reveal an accuracy score of around 90% with an F1 score of .56 at the maximums for the validation set.

DISCUSSION

When plotting the tree, it becomes more evident as to which features have a greater impact on the prediction. Splits higher up reveal that the feature is more impactful for the final prediction.

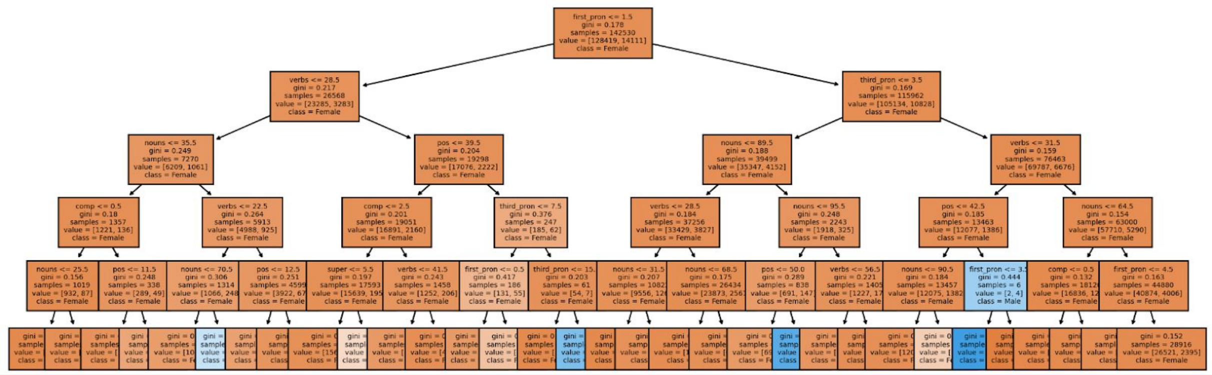


Figure 20: A plot of a decision tree for gender prediction

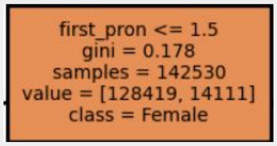


Figure 21: The first split begins with the first person pronoun.

The prediction reveals that a greater number of first person pronouns may be indicative of a female writer. Further analysis into other gender prediction studies may also reveal parallels. When continuing this machine learning algorithm, it might be beneficial to look into extraneous data. For example, as was done with the state data, looking into extraneous gender data may also be a valuable tool to better analyze which features are valuable and whether other variations of content related features, such as style or sentiment, may be applicable as well.

ACKNOWLEDGMENT

I would like to thank Dr. Suma Bhat, Professor at UIUC, and the team of mentors for their valuable advice; they have been a great help in guiding me through this process. I would also like to offer a special thanks to the TA and the other research students as well.

CITATIONS

“Author Profiling: Predicting Age and Gender from Blogs” K Santosh, Romil Bansal
http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-SantoshEt2013.pdf

“Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures” Renuka Joshi
https:/exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

DonorsChoose Application Screening Kaggle
https://www.kaggle.com/c/donorschoose-application-screening

DonorsChoose
https://www.donorschoose.org/

‘Examining Gender Differences in Writing Skill with Latent Factor Modeling” Laurel Woods
https://digital.lib.washington.edu/researchworks/bitstream/handle/1773/36586/Woods_washington_0250O_16104.pdf?isAllowed=y&sequence=1

“Gender Differences in Written Expression at the Elementary Level” Ashley D. Melloy
https://pdfs.semanticscholar.org/e50f/598846ef8eb1ea45813e8cdb8bc4a365a04e.pdf

“Overfitting vs. Underfitting” Nabil M Abbas
https://medium.com/swlh/overfitting-vs-underfitting-d742b4ffac57