Automated Classification of GES DISC User Support Tickets (ACOUSTICS)

ESIP Summer 2020
July 19-23, 2021
scs5@clemson.edu

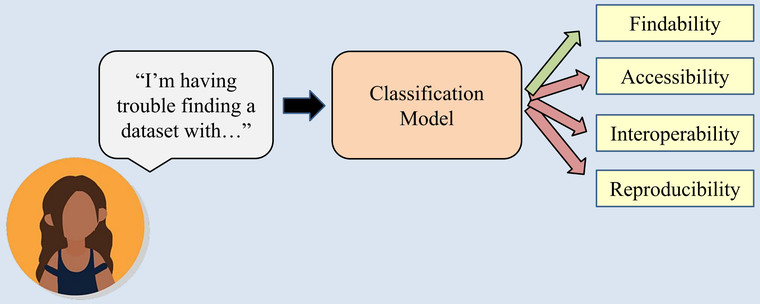
https://disc.gsfc.nasa.gov/ NASA/Goddard Earth Sciences Data and Information Services Center (GES DISC)

Sam Smith[1], Jennifer Wei[1], Armin Mehrabian[1,2], Binita KC[1,2], Alexis Hunzinger[1,2]
[1]Code 610.2, NASA Goddard Space Flight Center, Greenbelt, MD, USA [2]ADNET Systems Inc., Lanham, MD, USA

Overview

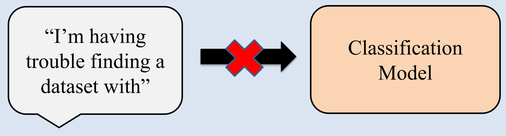
- The NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) is a leading data center that provides earth science data, information, and services to users globally.
- Users send in help tickets when they have a specific problem using the GES DISC, at which point the GES DISC help desk will help solve the problem

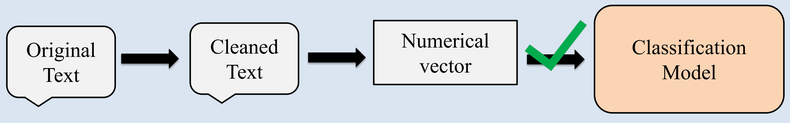
Goal
- Develop a machine learning model to classify GES DISC help desk user tickets according to the F.A.I.R. model (findability, accessibility, interoperability, reproducibility)



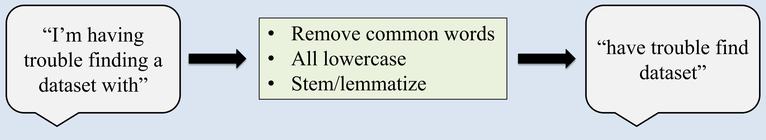
Training Data
- GES DISC has been recording user tickets since 2013
- Training data is ~700 manually labeled tickets from 2020

Data Pre-processing

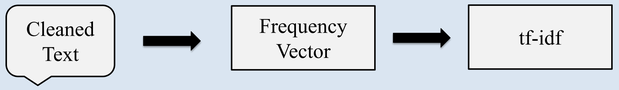




Text Cleaning
- Get rid of noise while keeping important textual themes



Feature Engineering
- Pulling numerical features from non-numerical data such as text

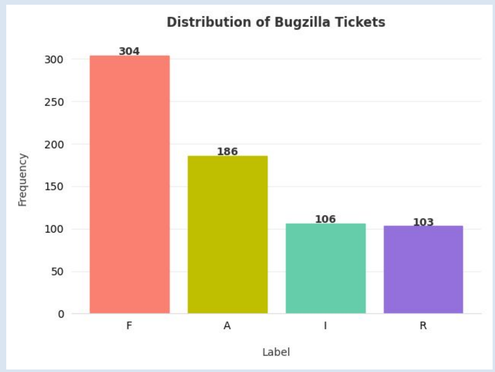


Frequency Vectorizer
'That is a cat.' □ {'that', 'is', 'a', 'cat'} □ {1, 1, 1, 1, 0, 0}
'This is a dog. □ {'this', 'is', 'a', 'dog'} □ {0, 1, 1, 0, 1, 1}

Tf-idf
- Statistical measure of word relevancy
- TF: term frequency
- IDF: inverse document frequency

Imbalanced Data

Our data does not have a uniform distribution, as shown below.



Possible Problem
- Most classification algorithms work better with balanced data.

Solutions:

Image Source: Resampling to Properly Handle Imbalanced Data

When to oversample?
- Oversampling the entire dataset will lead to overoptimistic model performance due to similar patterns emerging in both training and testing set
- We only want to oversample the training set

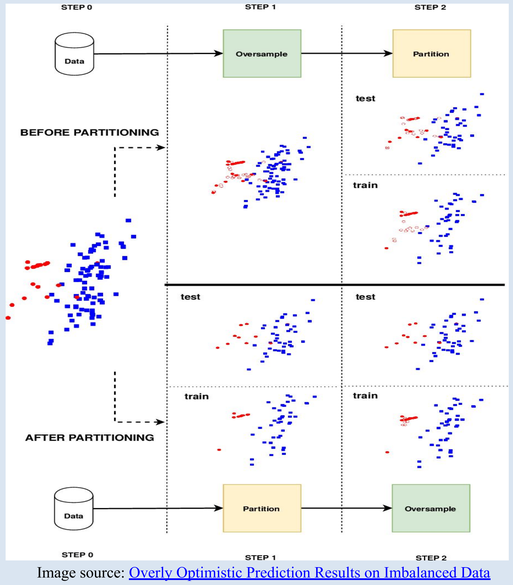


Image source: Overly Optimistic Prediction Results on Imbalanced Data

Results

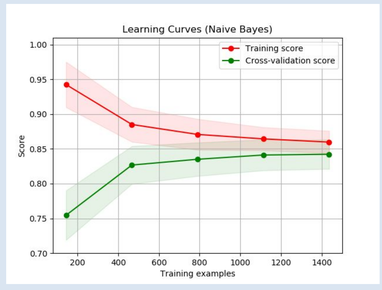
Accuracies of Classification Models (%)			
Model	**Original**	**Undersampled**	**Oversampled**
Multinomial NB	52.2	55.6	59.5
Linear SVC	63.1	59.9	62.5
Logistic Regression	60.3	54.5	60.8
Random Forest	58.9	54.3	60.1
Polynomial SVC	47.5	46.4	53.5
RBF SVC	56.9	59.5	57.5
Sigmoid SVC	62.8	61.4	62.5
K-NN	56.5	48.0	33.6

Accuracies are average over 3 iterations of 10-fold cross validation.

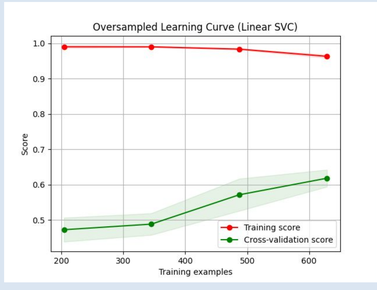
Resampling improved performance on certain models, while other models weren't affected. This is likely due to how well each model handles imbalanced data.

Learning Curves
- Show the evolution of the model performance as it is trained on more data
- Can be a useful tool for checking how well the model will generalize outside of the training data

Ideal Learning Curve
- Training accuracy decreases
- Testing accuracy increases
- Converge to some level of performance



a) Ideal Learning Curve



b) Our Learning Curve

Image Souce: Wikipedia: Learning Curve

As seen above, our learning curve is starting to show the expected trend of training accuracy decreasing while testing accuracy increases. However, it has not reached the level of convergence which we are looking for.

The takeaway from this learning curve is that more data will increase our model's performance. In addition to the 2020 ticket data we are already using, adding the 2019 data will be extremely helpful.

Future Work

- Manually label more data to increase our model's performance
- Use ML techniques to analyze help desk responses, hopefully automatically mapping a user ticket to a response
- Explore deep learning techniques as a classification method

Special Thanks

A special thanks to Dr. Jennifer Wei for mentoring this project and manually labeled the user tickets, Alexis Hunzinger for extracting ticket data for easy access, and Armin Mehrabian, Binita KC, and Rohan Dayal for offering their support through the project.