Categorization of Clinical Trials using Machine Learning and Natural Language Processing

Michael Welford (Z-1679714)

and

A S M Shahadat Hossain (Z-1907296)

Department of Computer Science

Northern Illinois University

DeKalb, IL, USA

Overview

- ❑ Introduction
- ❑ Objective
- ❑ Related Work
- ❑ Dataset Information
- ❑ Implementation
- ❑ Experimental Results
- ❑ Conclusion and Future Work

Introduction

“**Clinical trials** are research studies performed in people that are aimed at evaluating a medical, surgical, or behavioral intervention. They are the primary way that researchers find out if a new treatment, like a new drug or diet or medical device (for example, a pacemaker) is safe and effective in people.”

[https://www.nia.nih.gov/health/what-are-clinical-trials-and-studies]

Most recent example: **Clinical trials of COVID-19 vaccines**

Objective

- To categorize clinical trials using Machine Learning and Natural Language Processing techniques
- To observe the performances of different types of classifiers for this experiment: general purpose classifiers (Decision Tree, Support Vector Machine, Logistic Regression, and Naive Bayes) and ensemble methods (Random Forest, Bagging, and Boosting)

Related Work

- Factors on Clinical Trial Accrual
  - Iruku et al. - investigating organization type and location of primary investigator to participants
  - Lara Jr. et al. – socioeconomic status
- Trial Citations
  - Thewall and Kousha – studies cited on ClinicalTrials.gov have higher citation counts
- Working with Clinical Trial Data
  - Thiele et al. used the Random Forest algorithm -> linkages between German clinical trials on ClinicalTrials.gov.
  - Suggested Clinical trial titles could be useful for categorization.
- Altmetrics in scientific fields
  - Finch – ornithology
  - Wang – neurosurgery
  - Muñoz-Velandia – endocrinology
  - Suzan and Unal – malnutrition
  - Studies have also been done in the fields such as urology, critical care medicine, and sports medicine.
  - Erskine – types of mentions on Twitter

Dataset Information

- Altmetric Clinical Trial Dataset
  - Altmetric website
  - 50,330 clinical trials
  - 45 features
  - Contains information about media mentions

- Dimensions Clinical Trial Dataset
  - Dimensions.ai
  - 32,728 clinical trials
  - 57 features
  - Contains general information about trials like abstracts, title, and categories.

Implementation

Steps in the implementation:

- Data Preprocessing
- Feature Extraction
- Applying Classifiers

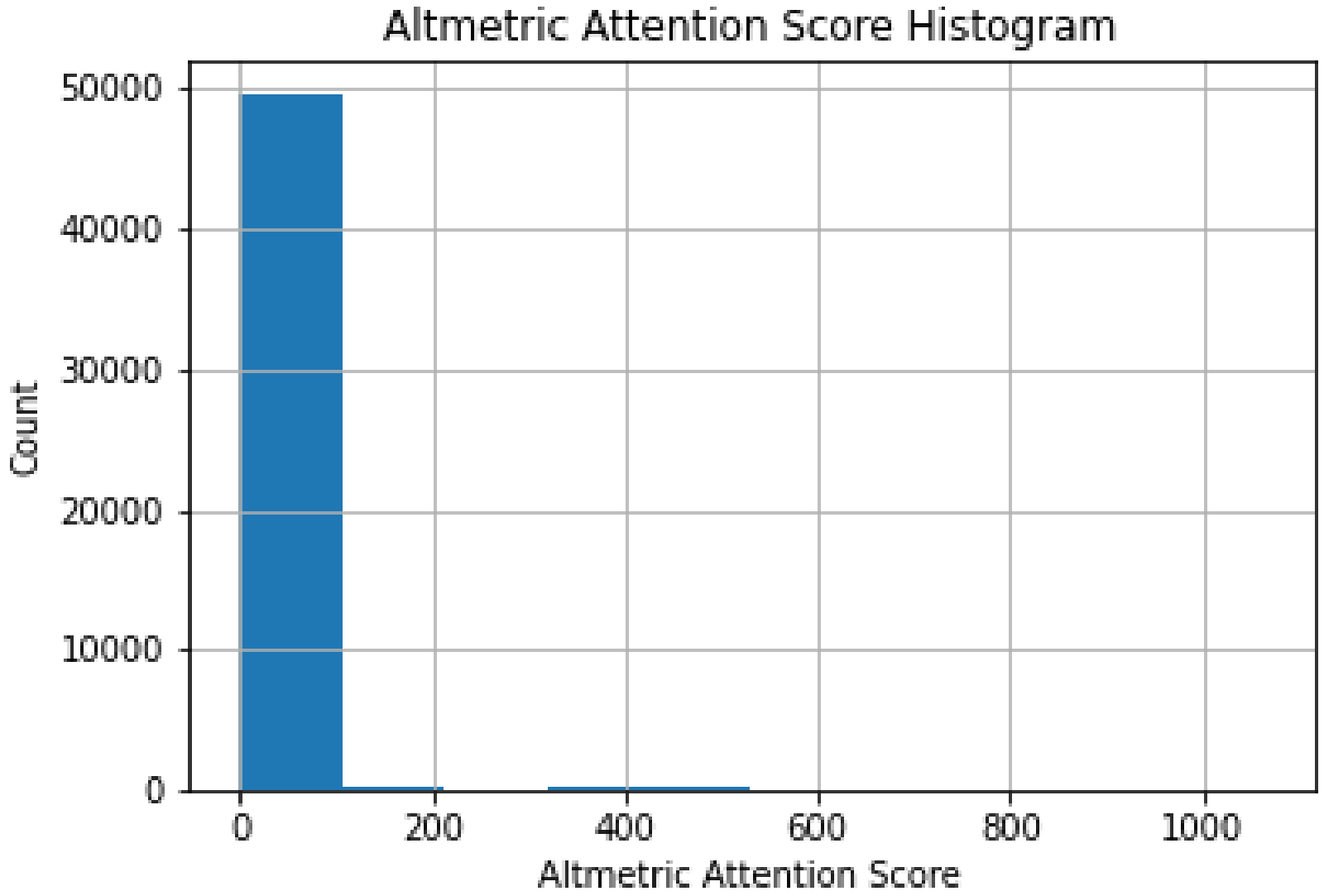
Implementation |Numeric Data Preprocessing

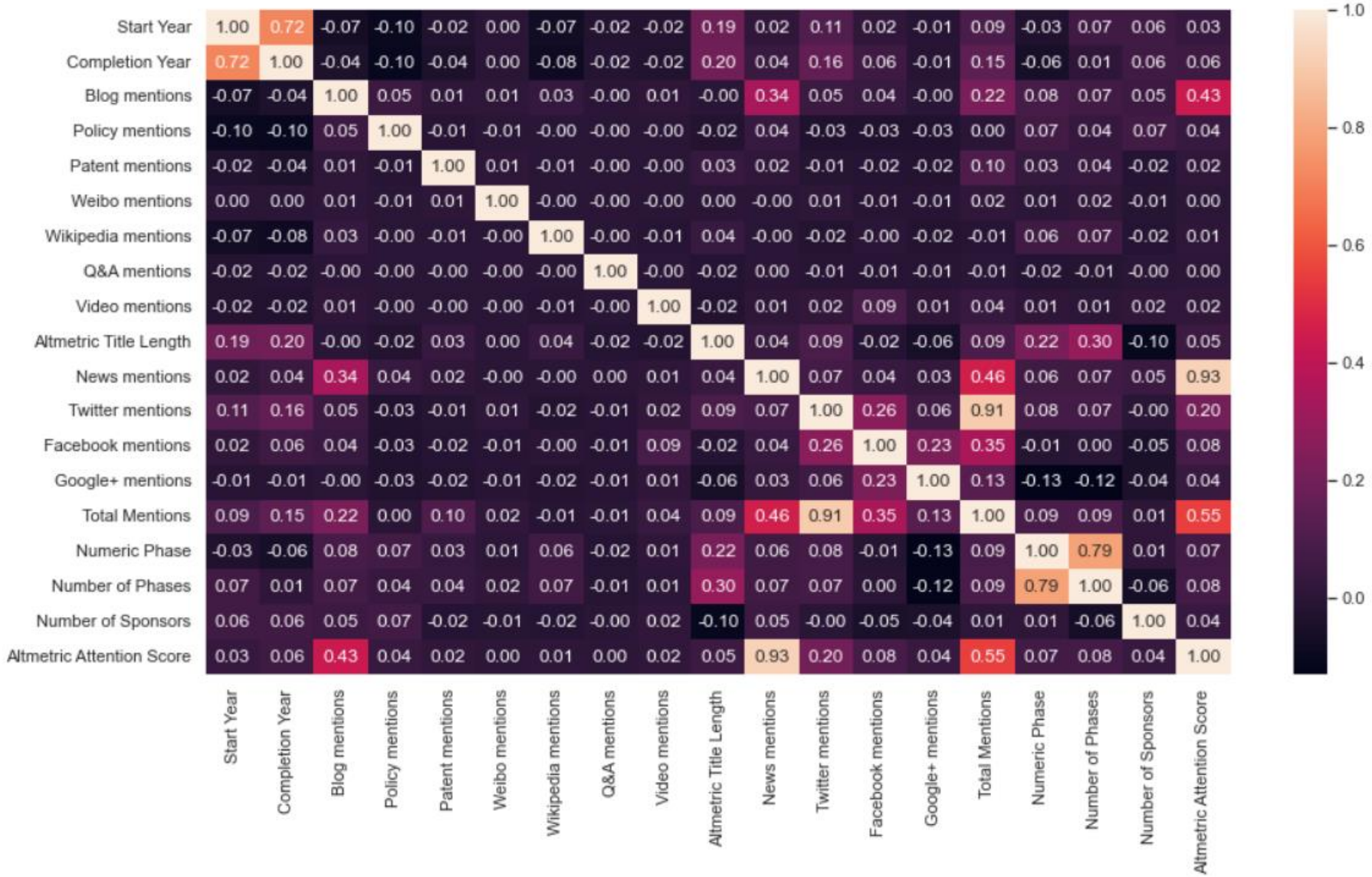
- Clean the data by removing blank columns
- Combine the Altmetric and Dimensions datasets into a joined dataset using an inner join (6202). (left='Trial ID', right='National Clinical Trial ID')
- Select the features needed for the task at hand
- Perform feature engineering using current features
  - Altmetric Title Length
  - Total Mentions
  - Numeric Phase
  - Number of Phases
  - Number of Sponsors
  - AAS Two Class

Implementation | Numerical Feature Engineering

- Altmetric Title Length – character count of the clinical trial title
- Total Mentions – the sum of the values for all mention types
- Numeric Phase – an integer value for the highest simultaneous phase of the trial
- Number of Phases – the number of simultaneous phases of the trial
- Number of Sponsors – the number of sponsors of the trial
- AAS Two Class – the class label
  - 0 – AAS score > 75^{th} percentile (7)
  - 1 – AAS score ≤ 75^{th} percentile

	count	mean	std	min	25%	50%	75%	max
Altmetric Title Length	6202.0	90.650597	40.081617	18.0	61.0	83.0	113.0	279.0
News mentions	6202.0	0.862786	4.799986	0.0	0.0	0.0	0.0	147.0
Twitter mentions	6202.0	3.672203	11.063528	0.0	0.0	1.0	3.0	292.0
Facebook mentions	6202.0	0.385360	1.355107	0.0	0.0	0.0	0.0	30.0
Google+ mentions	6202.0	0.168010	0.570598	0.0	0.0	0.0	0.0	7.0
Total Mentions	6202.0	5.420187	13.047338	0.0	1.0	2.0	4.0	294.0
Numeric Phase	6202.0	1.204128	1.158436	0.0	0.0	1.0	2.0	4.0
Number of Phases	6202.0	0.682199	0.620903	0.0	0.0	1.0	1.0	2.0
Number of Sponsors	6202.0	2.279587	1.958001	1.0	1.0	2.0	3.0	68.0
Altmetric Attention Score	6202.0	8.802161	34.138401	0.0	1.0	2.0	7.0	1062.0
AAS Two Class	6202.0	0.797001	0.402264	0.0	1.0	1.0	1.0	1.0


Altmetric Attention Score Histogram
Count
50000
40000
30000
20000
10000
0
0
200
400
600
800
1000
Altmetric Attention Score


Start Year 1.00 0.72 -0.07 -0.10 -0.02 0.00 -0.07 -0.02 -0.02 0.19 0.02 0.11 0.02 -0.01 0.09 -0.03 0.07 0.06 0.03
Completion Year 0.72 1.00 -0.04 -0.10 -0.04 0.00 -0.08 -0.02 -0.02 0.20 0.04 0.16 0.06 -0.01 0.15 -0.06 0.01 0.06 0.06
Blog mentions -0.07 -0.04 1.00 0.05 0.01 0.01 0.03 -0.00 0.01 -0.00 0.34 0.05 0.04 -0.00 0.22 0.08 0.07 0.05 0.43
Policy mentions -0.10 -0.10 0.05 1.00 -0.01 -0.01 -0.00 -0.00 -0.00 -0.02 0.04 -0.03 -0.03 -0.03 0.00 0.07 0.04 0.07 0.04
Patent mentions -0.02 -0.04 0.01 -0.01 1.00 0.01 -0.01 -0.00 -0.00 0.03 0.02 -0.01 -0.02 -0.02 0.10 0.03 0.04 -0.02 0.02
Weibo mentions 0.00 0.00 0.01 -0.01 0.01 1.00 -0.00 -0.00 -0.00 0.00 -0.00 0.01 -0.01 -0.01 0.02 0.01 0.02 -0.01 0.00
Wikipedia mentions -0.07 -0.08 0.03 -0.00 -0.01 -0.00 1.00 -0.00 -0.01 0.04 -0.00 -0.02 -0.00 -0.02 -0.01 0.06 0.07 -0.02 0.01
Q&A mentions -0.02 -0.02 -0.00 -0.00 -0.00 -0.00 -0.00 1.00 -0.00 -0.02 0.00 -0.01 -0.01 -0.01 -0.01 -0.02 -0.01 -0.00 0.00
Video mentions -0.02 -0.02 0.01 -0.00 -0.00 -0.00 -0.01 -0.00 1.00 -0.02 0.01 0.02 0.09 0.01 0.04 0.01 0.01 0.02 0.02
Altmetric Title Length 0.19 0.20 -0.00 -0.02 0.03 0.00 0.04 -0.02 -0.02 1.00 0.04 0.09 -0.02 -0.06 0.09 0.22 0.30 -0.10 0.05
News mentions 0.02 0.04 0.34 0.04 0.02 -0.00 -0.00 0.00 0.01 0.04 1.00 0.07 0.04 0.03 0.46 0.06 0.07 0.05 0.93
Twitter mentions 0.11 0.16 0.05 -0.03 -0.01 0.01 -0.02 -0.01 0.02 0.09 0.07 1.00 0.26 0.06 0.91 0.08 0.07 -0.00 0.20
Facebook mentions 0.02 0.06 0.04 -0.03 -0.02 -0.01 -0.00 -0.01 0.09 -0.02 0.04 0.26 1.00 0.23 0.35 -0.01 0.00 -0.05 0.08
Google+ mentions -0.01 -0.01 -0.00 -0.03 -0.02 -0.01 -0.02 -0.01 0.01 -0.06 0.03 0.06 0.23 1.00 0.13 -0.13 -0.12 -0.04 0.04
Total Mentions 0.09 0.15 0.22 0.00 0.10 0.02 -0.01 -0.01 0.04 0.09 0.46 0.91 0.35 0.13 1.00 0.09 0.09 0.01 0.55
Numeric Phase -0.03 -0.06 0.08 0.07 0.03 0.01 0.06 -0.02 0.01 0.22 0.06 0.08 -0.01 -0.13 0.09 1.00 0.79 0.01 0.07
Number of Phases 0.07 0.01 0.07 0.04 0.04 0.02 0.07 -0.01 0.01 0.30 0.07 0.07 0.00 -0.12 0.09 0.79 1.00 -0.06 0.08
Number of Sponsors 0.06 0.06 0.05 0.07 -0.02 -0.01 -0.02 -0.00 0.02 -0.10 0.05 -0.00 -0.05 -0.04 0.01 0.01 -0.06 1.00 0.04
Altmetric Attention Score 0.03 0.06 0.43 0.04 0.02 0.00 0.01 0.00 0.02 0.05 0.93 0.20 0.08 0.04 0.55 0.07 0.08 0.04 1.00
Start Year
Completion Year
Blog mentions
Policy mentions
Patent mentions
Weibo mentions
Wikipedia mentions
Q&A mentions
Video mentions
Altmetric Title Length
News mentions
Twitter mentions
Facebook mentions
Google+ mentions
Total Mentions
Numeric Phase
Number of Phases
Number of Sponsors
Altmetric Attention Score
1.0
0.8
0.6
0.4
0.2
0.0

Implementation | Classification Models Used

- **General Classifiers**
  - **Decision Tree (DT)**
  - **Support Vector Machine (SVM)**
  - **Logistic Regression (LR)**
  - **Naive Bayes Classifier (NB)**
- **Ensemble Methods**
  - **Bagging**
    - with DT
    - with SVM
    - with LR
  - **Random Forest with DT**
    - No tuning
    - Grid Search
    - Randomized Search
  - **Boosting**
    - Adaboost with DT
    - XGBoost using DMatrix
    - XGBoost using Scikit-learn API
      - No tuning
      - Grid Search
      - Randomized Search

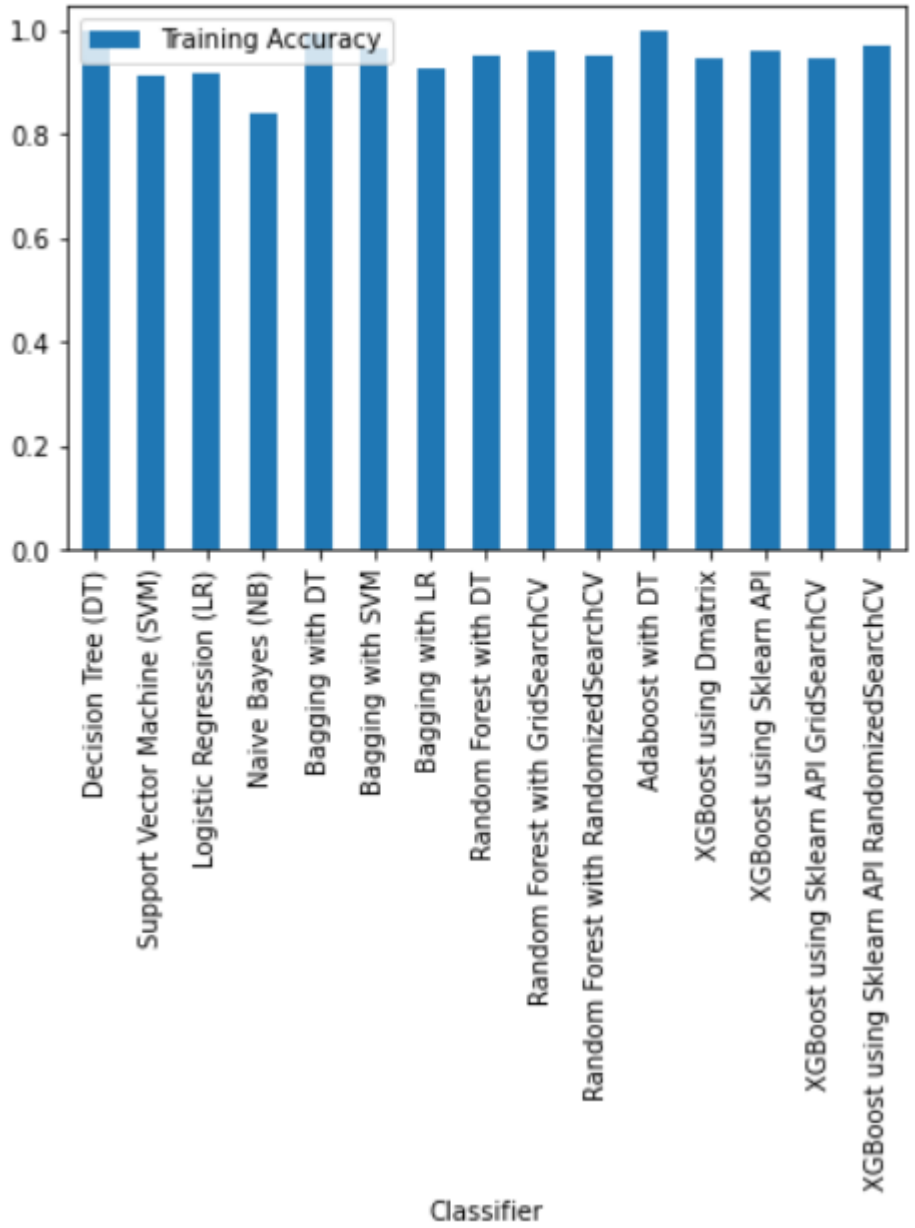
Implementation | Hyperparameters Tuning

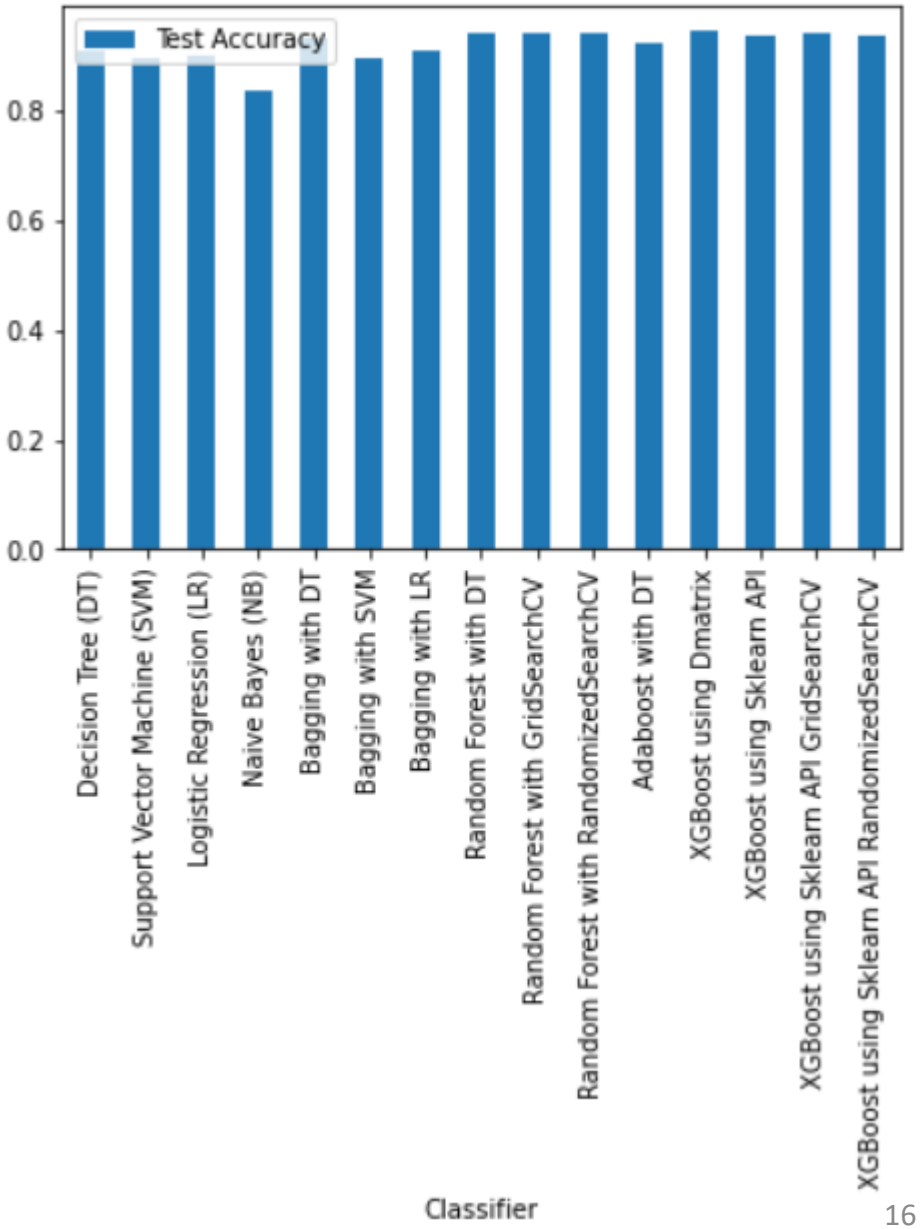
- Random Forest
  - Grid Search
    - max_depth: 8
    - n_estimators: 300
  - Randomized Search
    - max_depth: 6
    - n_estimators: 300
- XGBoost using Sklearn API (XGBClassifier)
  - Grid Search
    - learning_rate: 0.05
    - n_estimators: 200
    - subsample:0.05
  - Randomized Search
    - learning_rate: 0.05
    - n_estimators: 200
    - subsample: 0.8

Experimental Results | Evaluation Metrics Using Numeric Features

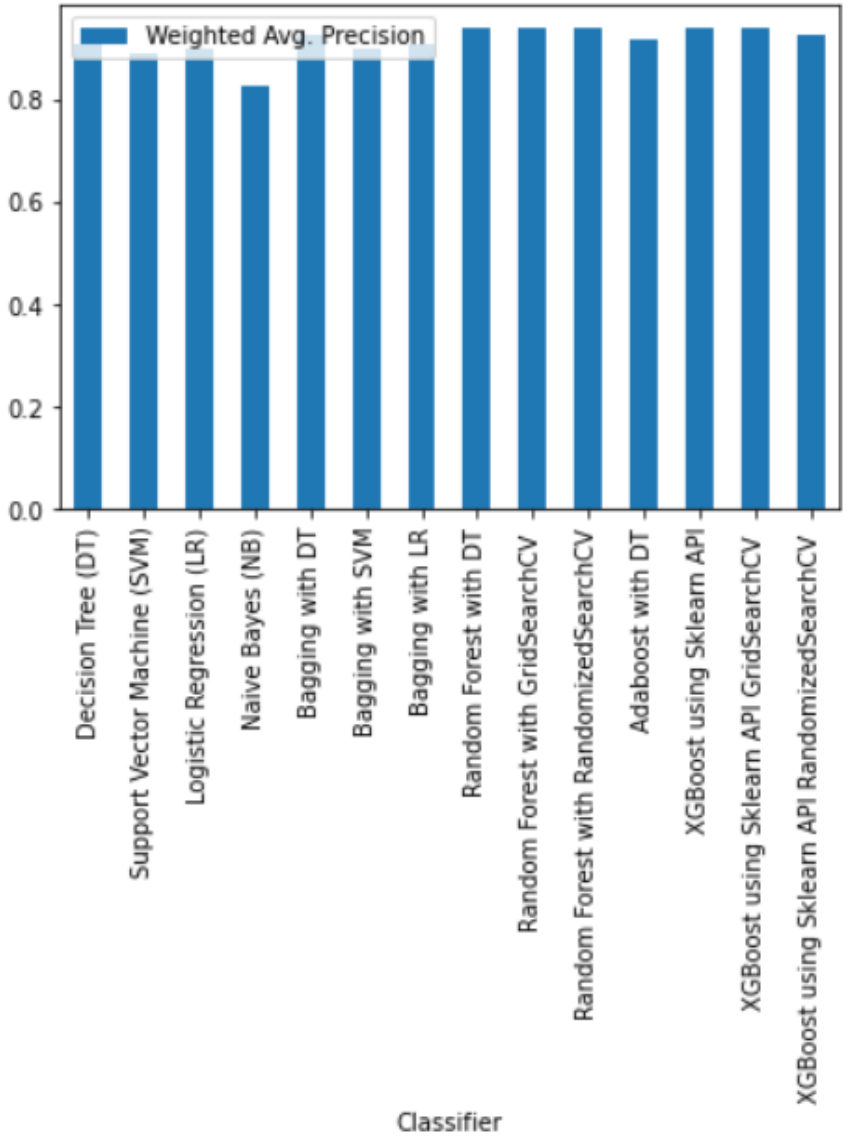
Classifier	Training Accuracy	Test Accuracy	Group 0 Precision	Group 1 Precision	Weighted Avg. Precision	Group 0 Recall	Weighted Avg. Recall	Weighted Avg. F1 Score	Specificity of Group 0	ROC/AUC	CV Accuracy
Decision Tree (DT)	0.997984	0.912168	0.77	0.95	0.91	0.8	0.91	0.91	0.801587	0.872468	0.928241
Support Vector Machine (SVM)	0.914735	0.895246	0.87	0.9	0.89	0.57	0.9	0.89	0.571429	0.947416	0.90849
Logistic Regression (LR)	0.916751	0.902498	0.88	0.91	0.9	0.6	0.9	0.9	0.599206	0.947213	0.916553
Naive Bayes (NB)	0.842572	0.836422	0.61	0.88	0.83	0.52	0.84	0.83	0.52381	0.77135	0.842977
Bagging with DT	0.993953	0.933924	0.85	0.95	0.93	0.82	0.93	0.93	0.821429	0.960622	0.937915
Bagging with SVM	0.964120	0.898469	0.77	0.93	0.9	0.71	0.9	0.9	0.706349	0.930198	0.893774
Bagging with LR	0.926628	0.911362	0.89	0.92	0.91	0.64	0.91	0.91	0.642857	0.960366	0.924211
Random Forest with DT	0.949405	0.939565	0.9	0.95	0.94	0.79	0.94	0.94	0.789683	0.958506	0.947794
Random Forest with GridSearchCV	0.959282	0.939565	0.9	0.95	0.94	0.79	0.94	0.94	0.789683	0.968639	0.947996
Random Forest with RandomizedSearchCV	0.952026	0.940371	0.9	0.95	0.94	0.79	0.94	0.94	0.789683	0.965975	0.947996
Adaboost with DT	0.997984	0.921837	0.82	0.95	0.92	0.79	0.92	0.92	0.785714	0.943459	0.935695
XGBoost using Dmatrix	0.948171	0.944855	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XGBoost using Sklearn API	0.958879	0.937953	0.88	0.95	0.94	0.8	0.94	0.94	0.801587	0.970493	0.945978
XGBoost using Sklearn API GridSearchCV	0.947390	0.939565	0.89	0.95	0.94	0.81	0.94	0.94	0.805556	0.961973	0.94598
XGBoost using Sklearn API RandomizedSearchCV	0.969764	0.936342	0.88	0.95	0.93	0.79	0.94	0.93	0.789683	0.973498	0.94477

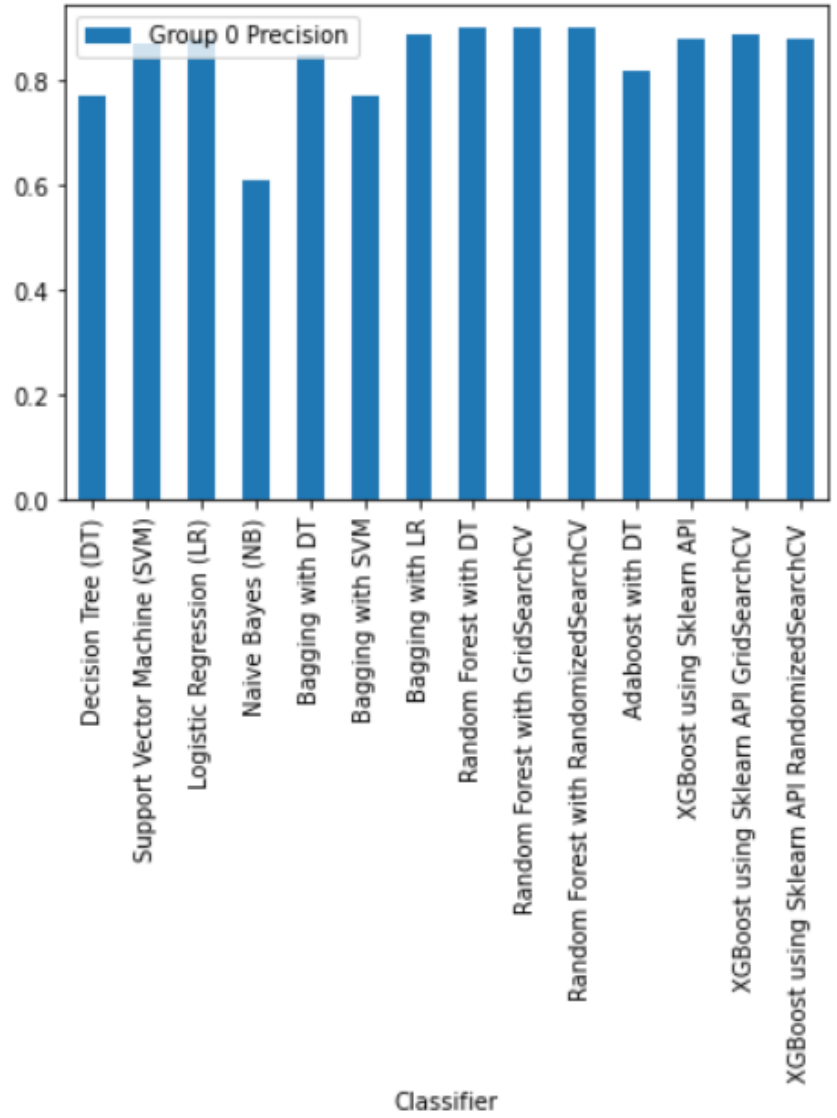
Experimental Results | Numerical Features



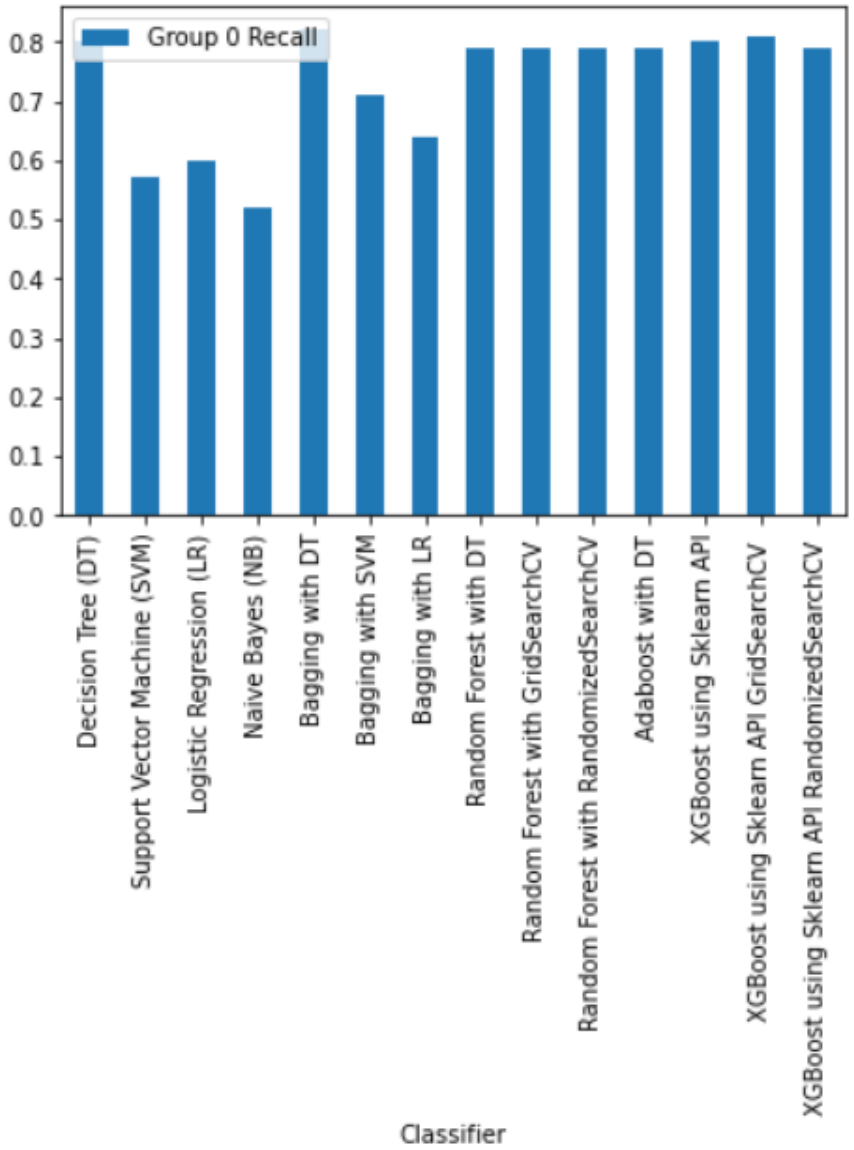


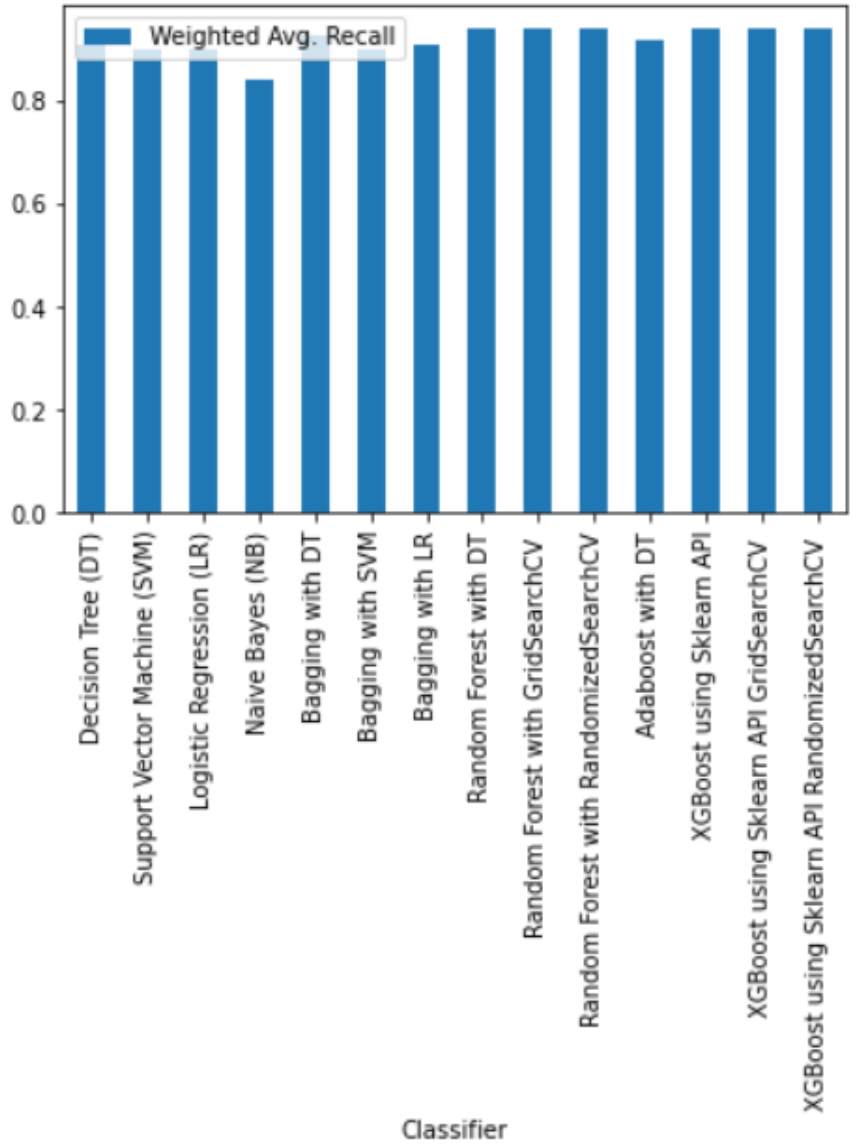
Experimental Results | Numerical Features



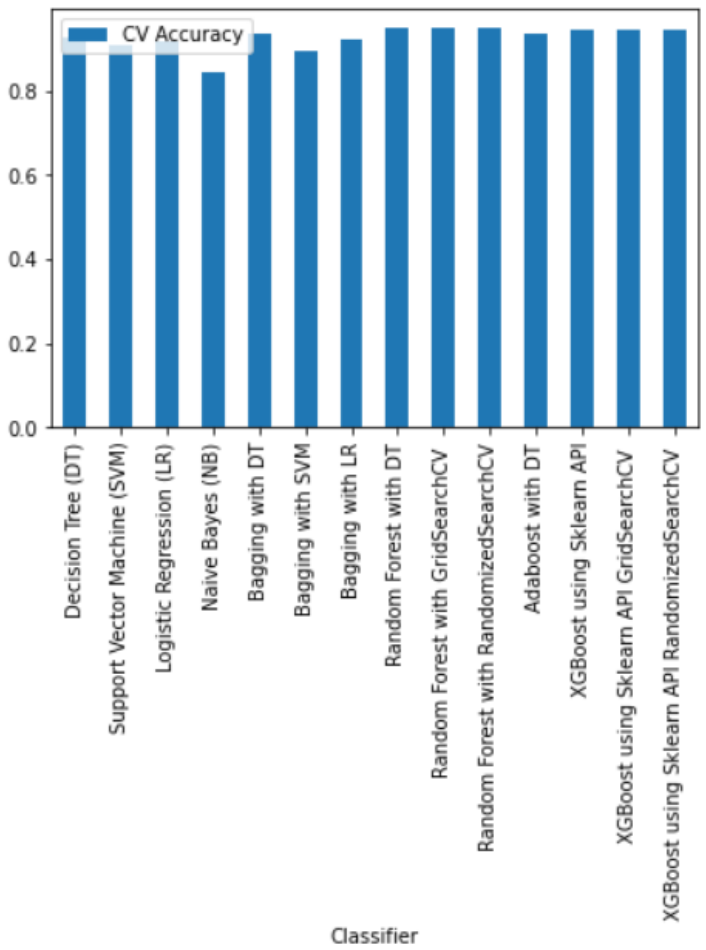


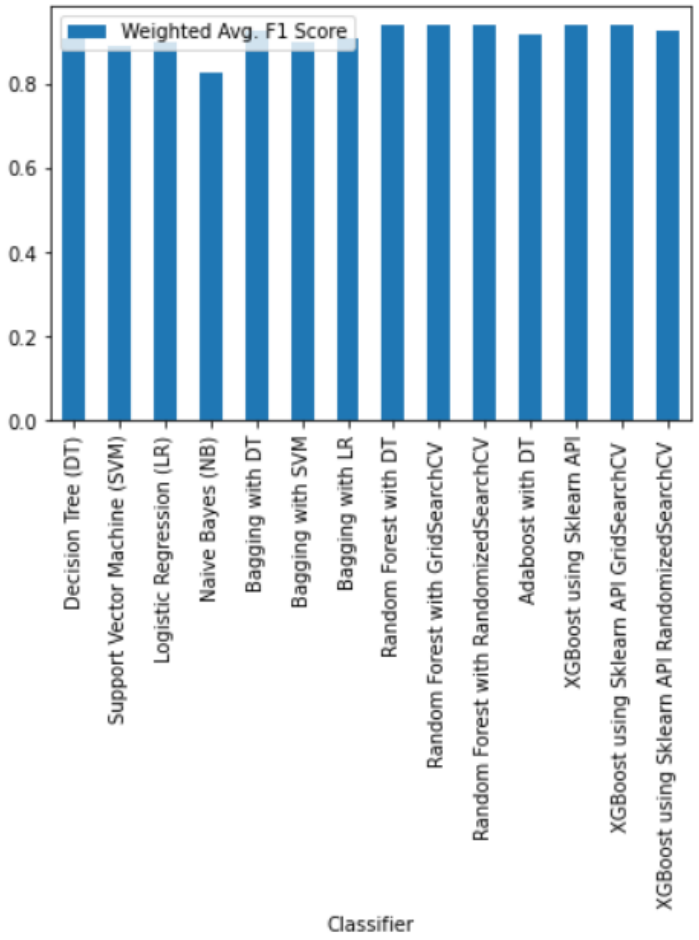
Experimental Results | Numerical Features

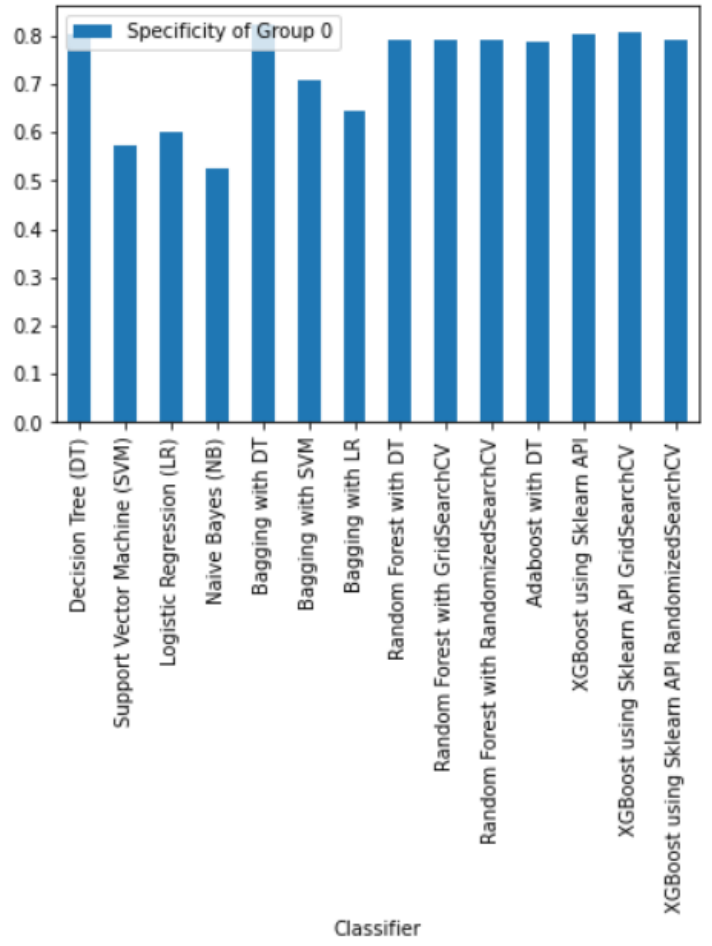




Experimental Results | Numerical Features

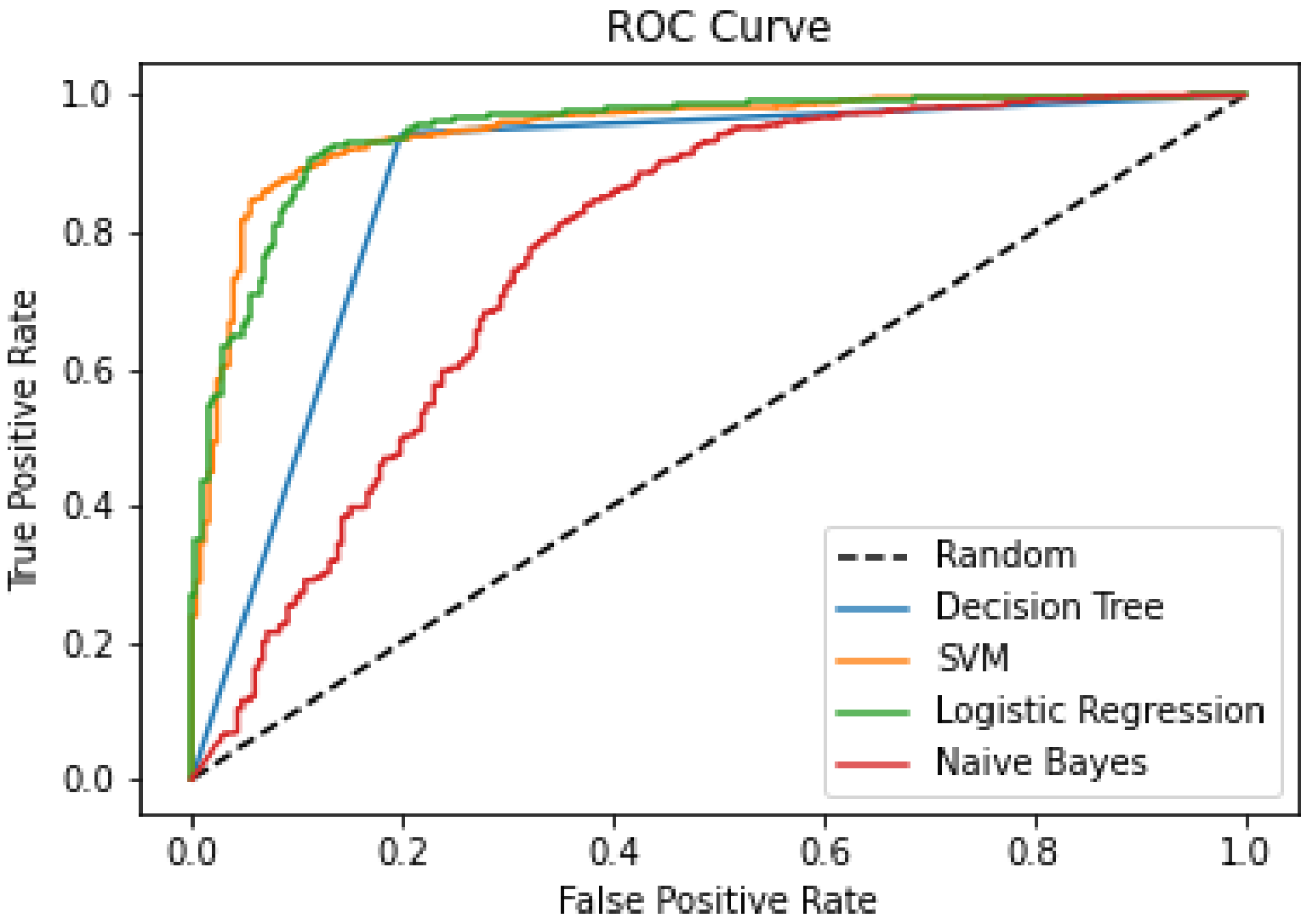




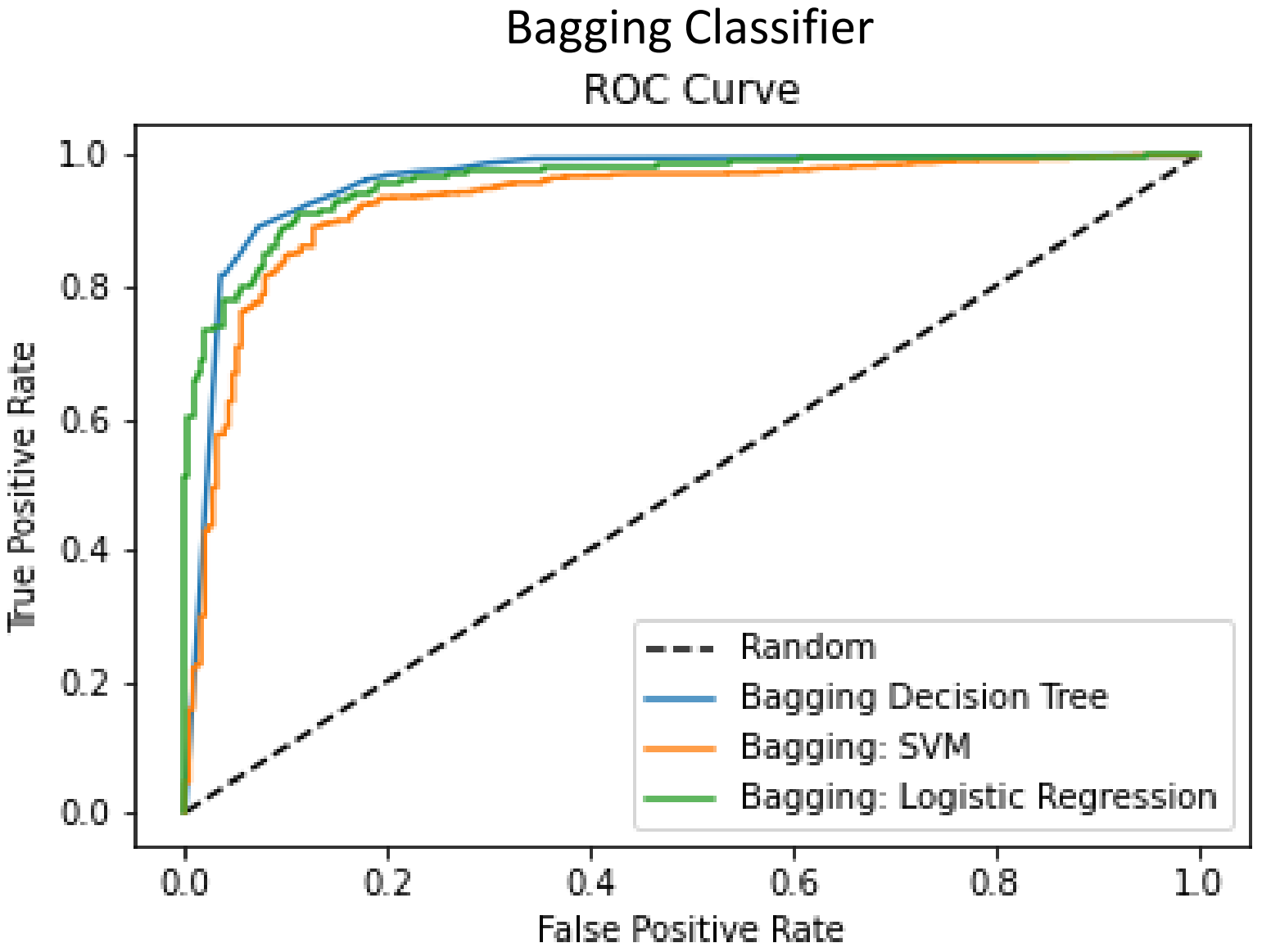


Experimental Results | Numerical Features

General Classifier

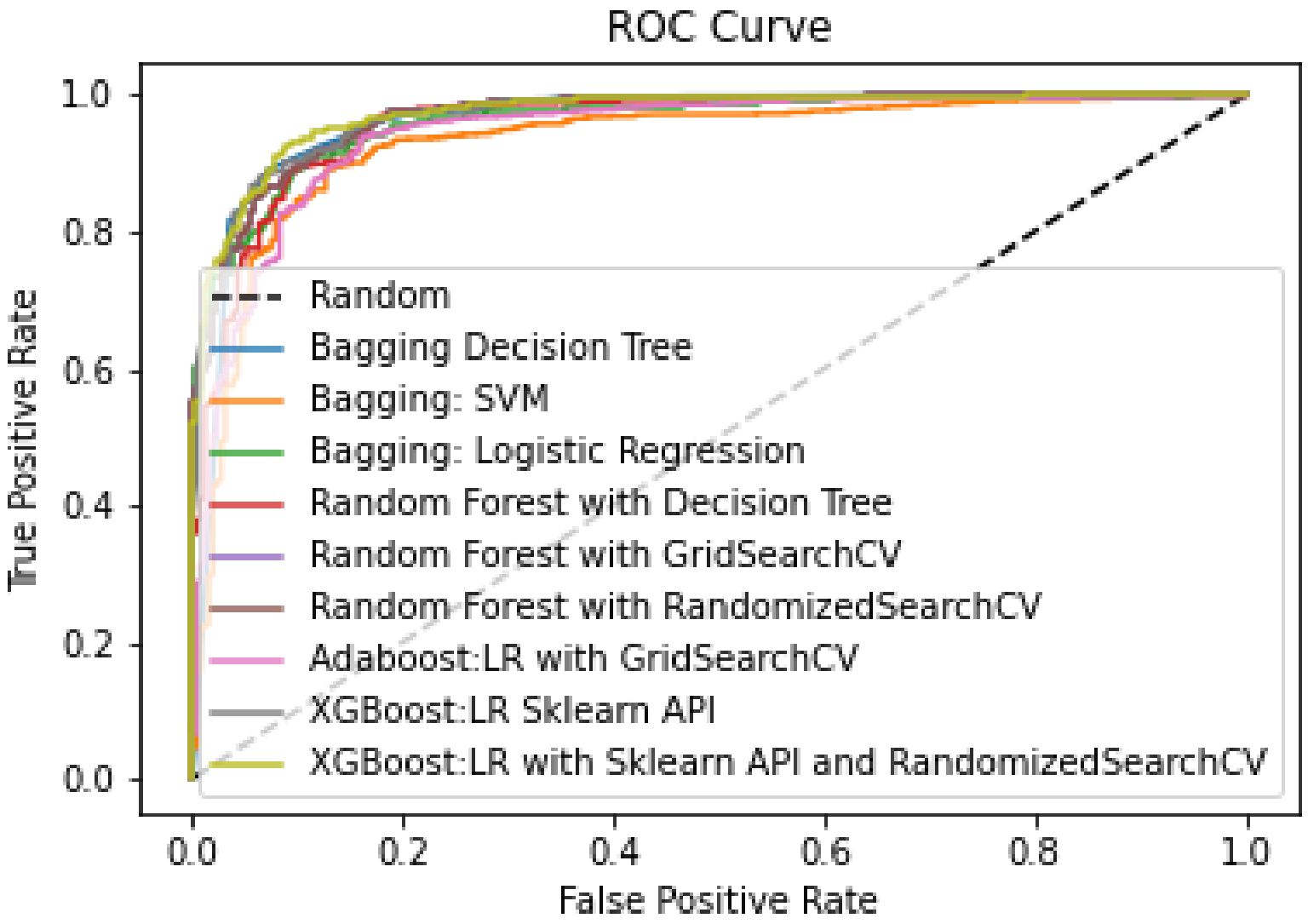


Experimental Results | Numerical Features

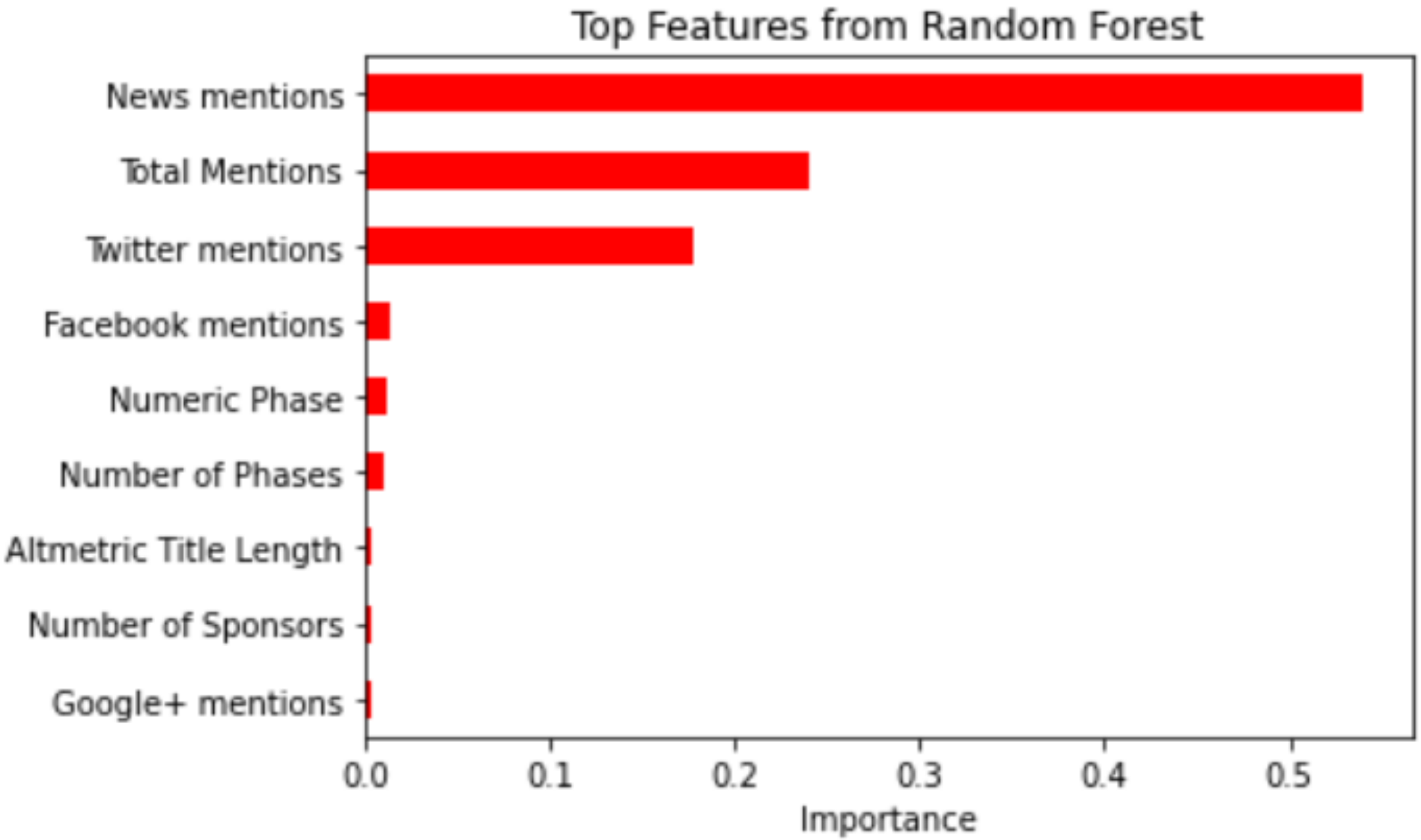


Experimental Results | Numerical Features

Ensemble Methods



Experimental Results | Numerical Features



Implementation | Text Data Preprocessing

- Steps used for Text Data Preprocessing:
  - HTML Parsing (using BeautifulSoup Python library)
  - Removing Special Characters (using Python Regular Expression function)
  - Tokenization (using word_tokenize function)
  - Converting into lower cases (using .lower function)
  - Stopwords Removal (using Python NLTK)
  - Stemming and Lemmatization (using PorterStemmer and WordNetLemmatizer from Python NLTK)

Implementation | Text Data Preprocessing

- Text in ‘Title’ before Preprocessing

'An Exploratory, Open-Label Study of Vedolizumab (Anti-alpha4beta7 Antibody) in Subjects With HIV Infection Undergoing Analytical Treatment Interruption'

- Text in ‘Title’ after Preprocessing

'exploratory open label study vedolizumab anti alpha4beta7 antibody subject hiv infect undergo analyt treatment interrupt'

Implementation | Text Feature Extraction

- Feature extraction was done using

  Term Frequency–Inverse Document Frequency (TF-IDF) model:

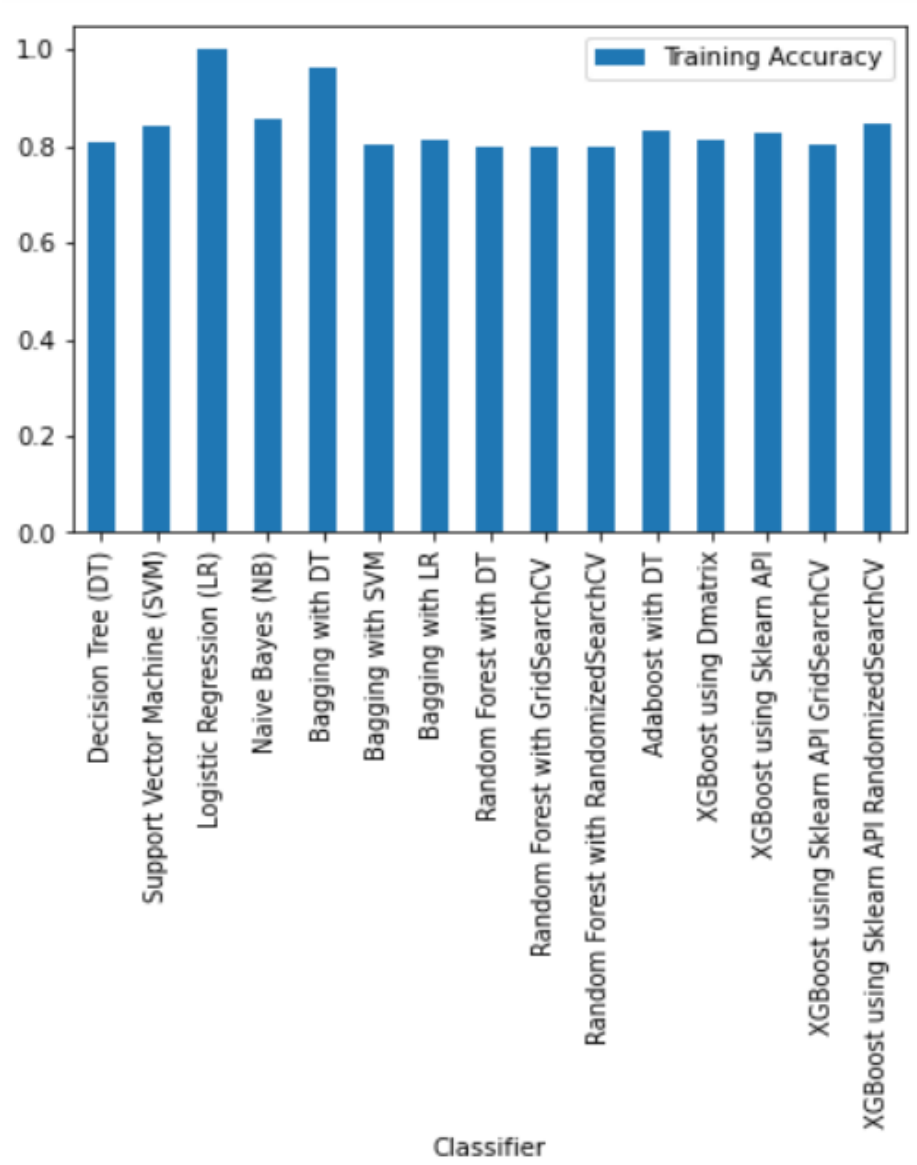
$$tfidf(t, d, D) = tf(t, d).idf(t, D)$$

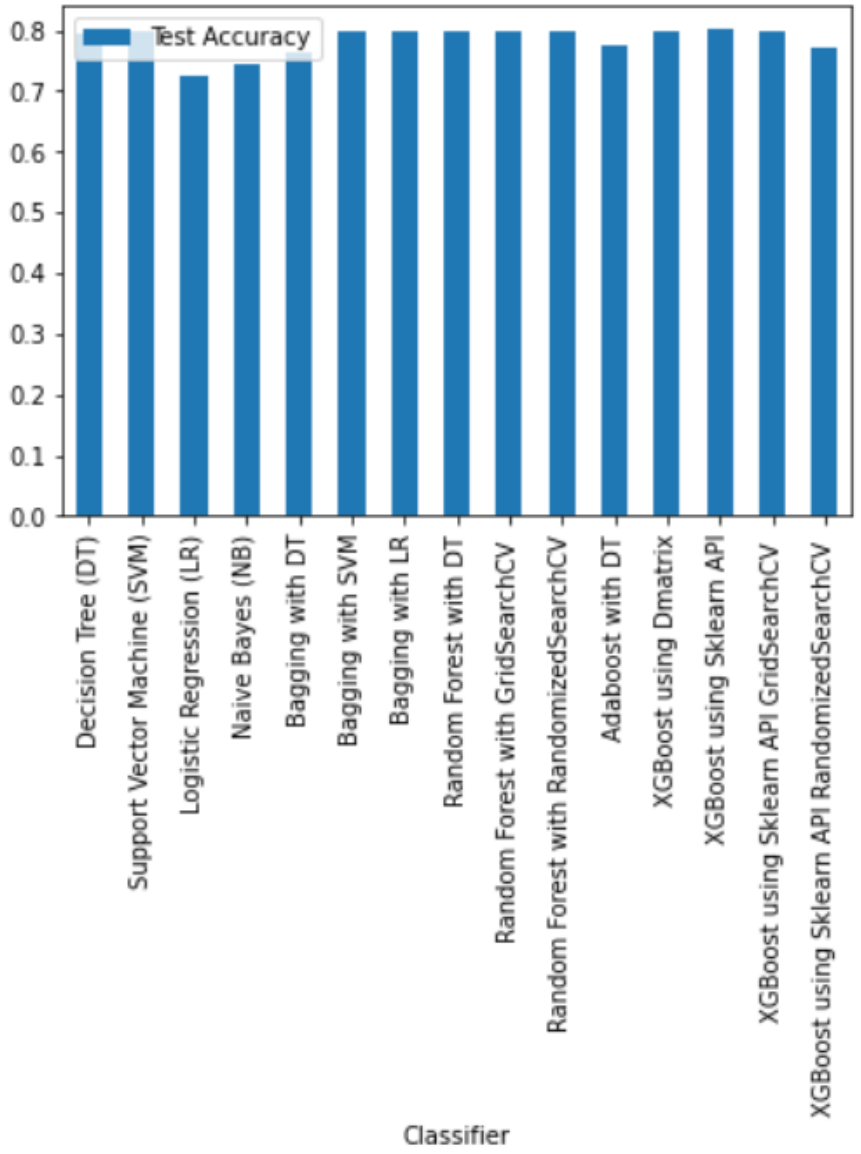
Here,
t is the term,
d is the document, and
D is the total number of documents in the corpus

Experimental Results | Evaluation Metrics Using Text Features

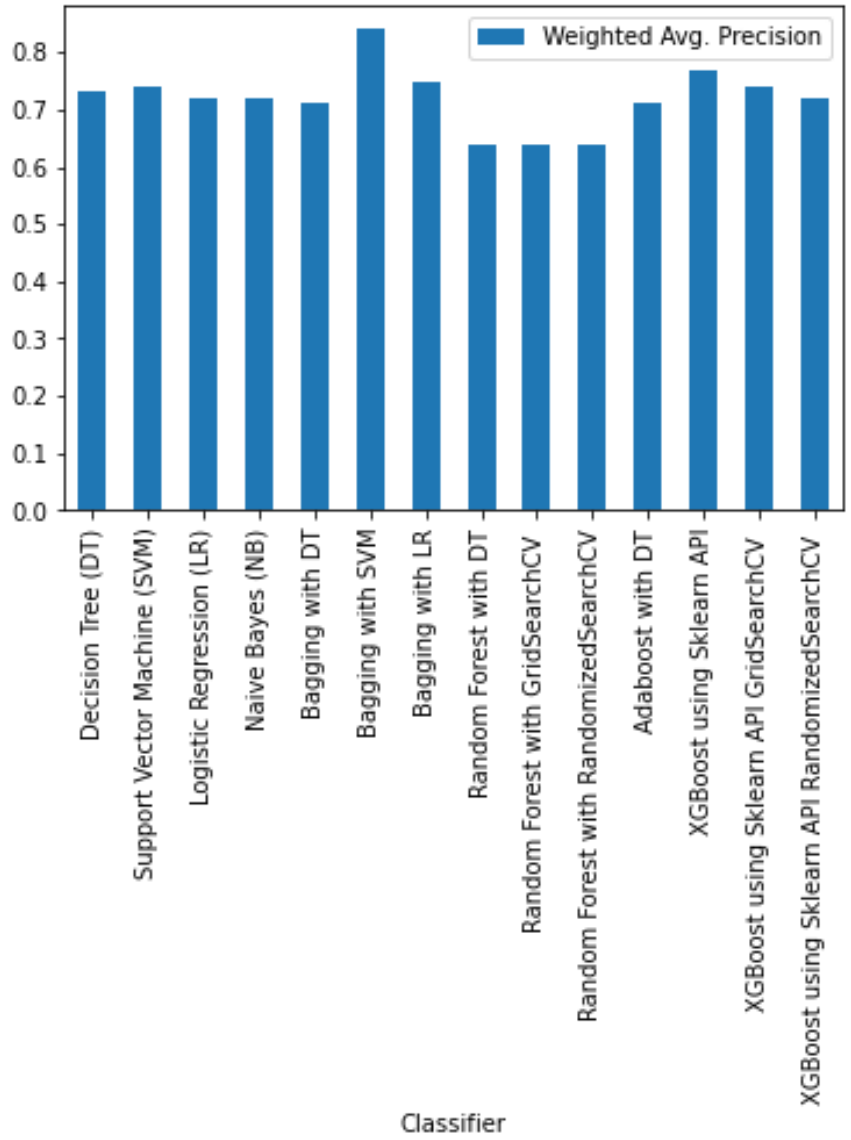
Classifier	Training Accuracy	Test Accuracy	Group 0 Precision	Group 1 Precision	Weighted Avg. Precision	Weighted Avg. Recall	Weighted Avg. F1 Score	Specificity of Group 0	ROC/AUC	CV Accuracy
Decision Tree (DT)	0.807498	0.796132	0.46	0.8	0.73	0.8	0.72	0.02381	0.569501	0.792584
Support Vector Machine (SVM)	0.842774	0.796938	0.5	0.8	0.74	0.99	0.72	0.039683	0.631297	0.795808
Logistic Regression (LR)	0.999395	0.725222	0.31	0.82	0.72	0.73	0.72	0.281746	0.592606	0.720622
Naive Bayes (NB)	0.856481	0.746172	0.34	0.82	0.72	0.75	0.73	0.253968	0.640454	0.751866
Bagging with DT	0.960089	0.763900	0.33	0.81	0.71	0.76	0.73	0.162698	0.56918	0.765774
Bagging with SVM	0.801250	0.797744	1.0	0.8	0.84	0.8	0.71	0.003968	0.643591	0.796815
Bagging with LR	0.813344	0.797744	0.55	0.8	0.75	0.8	0.72	0.02381	0.672416	0.799637
Random Forest with DT	0.797017	0.796938	0.0	0.8	0.64	0.8	0.71	0.0	0.640819	0.797017
Random Forest with GridSearchCV	0.797017	0.796938	0.0	0.8	0.64	0.8	0.71	0.0	0.619254	0.797017
Random Forest with RandomizedSearchCV	0.797017	0.796938	0.0	0.8	0.64	0.8	0.71	0.0	0.654565	0.797017
Adaboost with DT	0.831687	0.773570	0.32	0.81	0.71	0.77	0.72	0.103175	0.599399	0.762144
XGBoost using Dmatrix	0.812032	0.797482	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XGBoost using Sklearn API	0.827253	0.800967	0.62	0.8	0.77	0.8	0.73	0.051587	0.624759	0.795003
XGBoost using Sklearn API GridSearchCV	0.805080	0.796938	0.5	0.81	0.74	0.8	0.73	0.083333	0.595092	0.7948
XGBoost using Sklearn API RandomizedSearchCV	0.848821	0.770346	0.35	0.81	0.72	0.77	0.73	0.154762	0.586487	0.769805

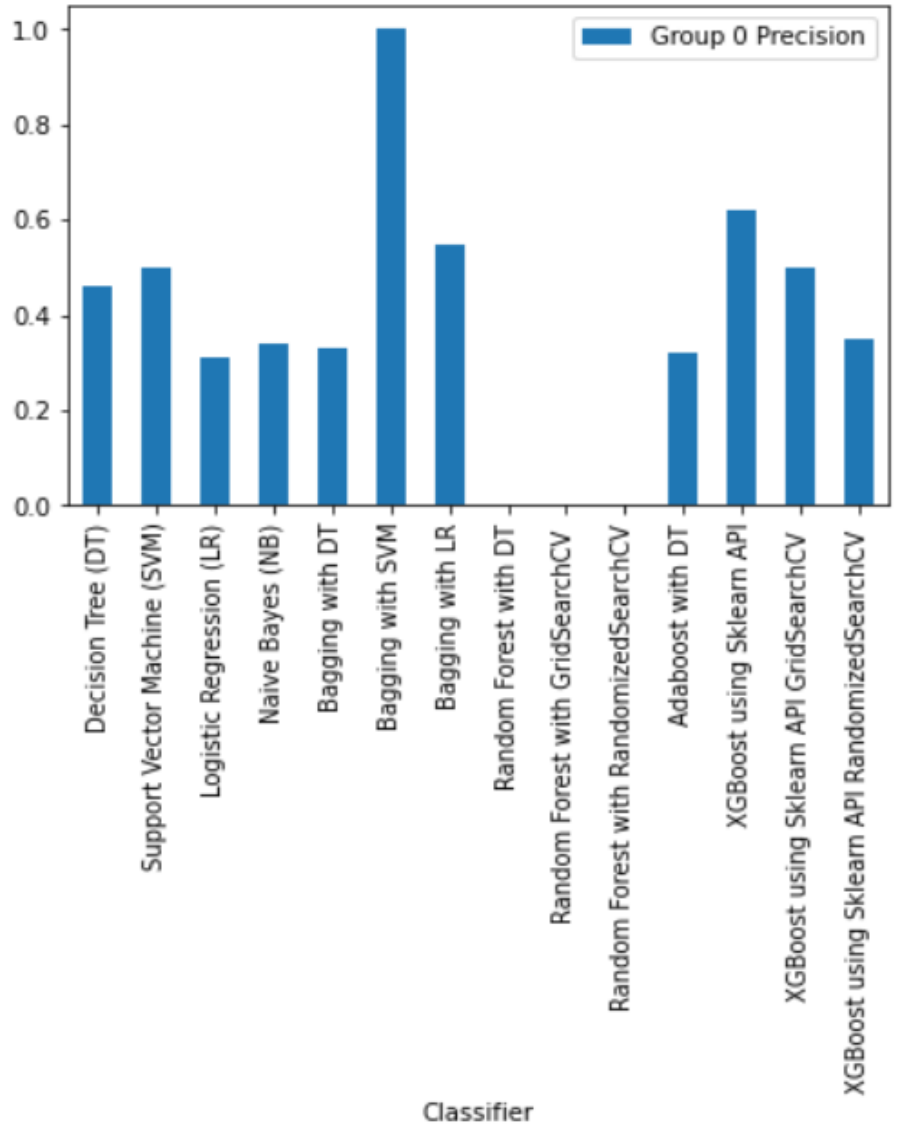
Experimental Results | Text Features



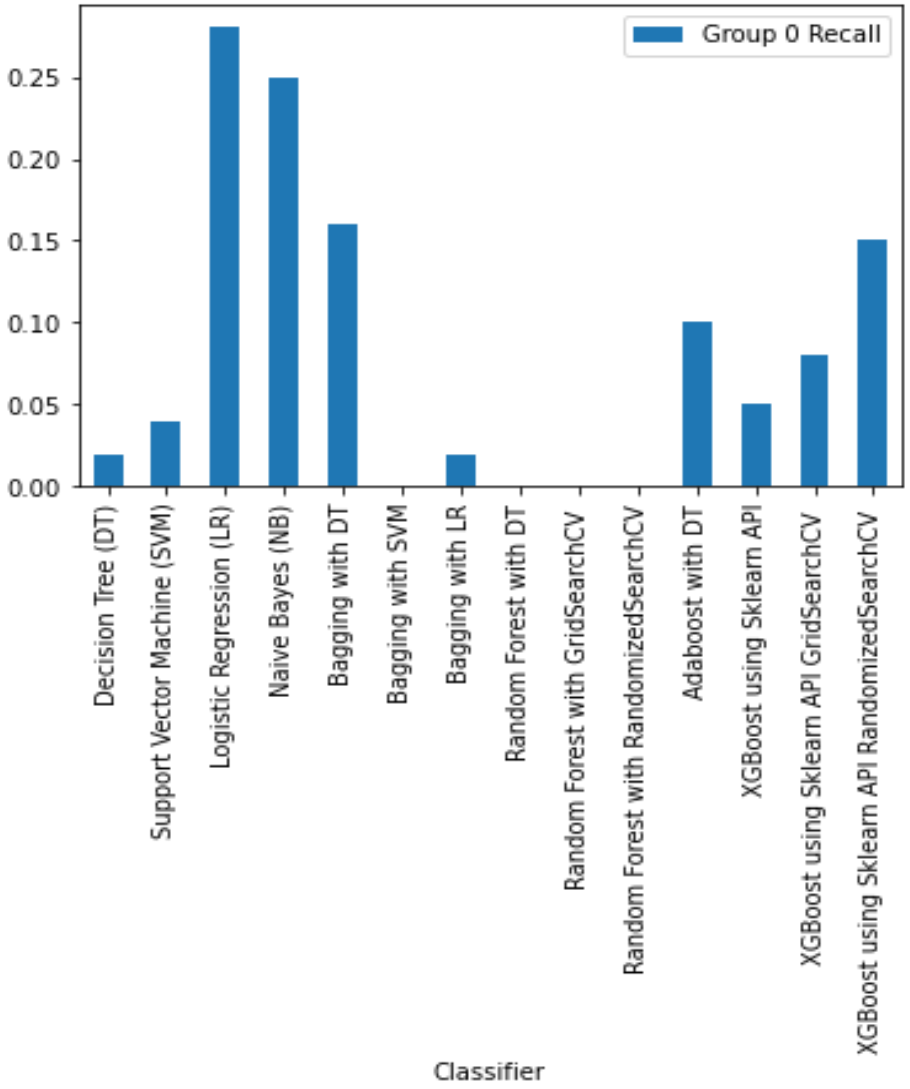


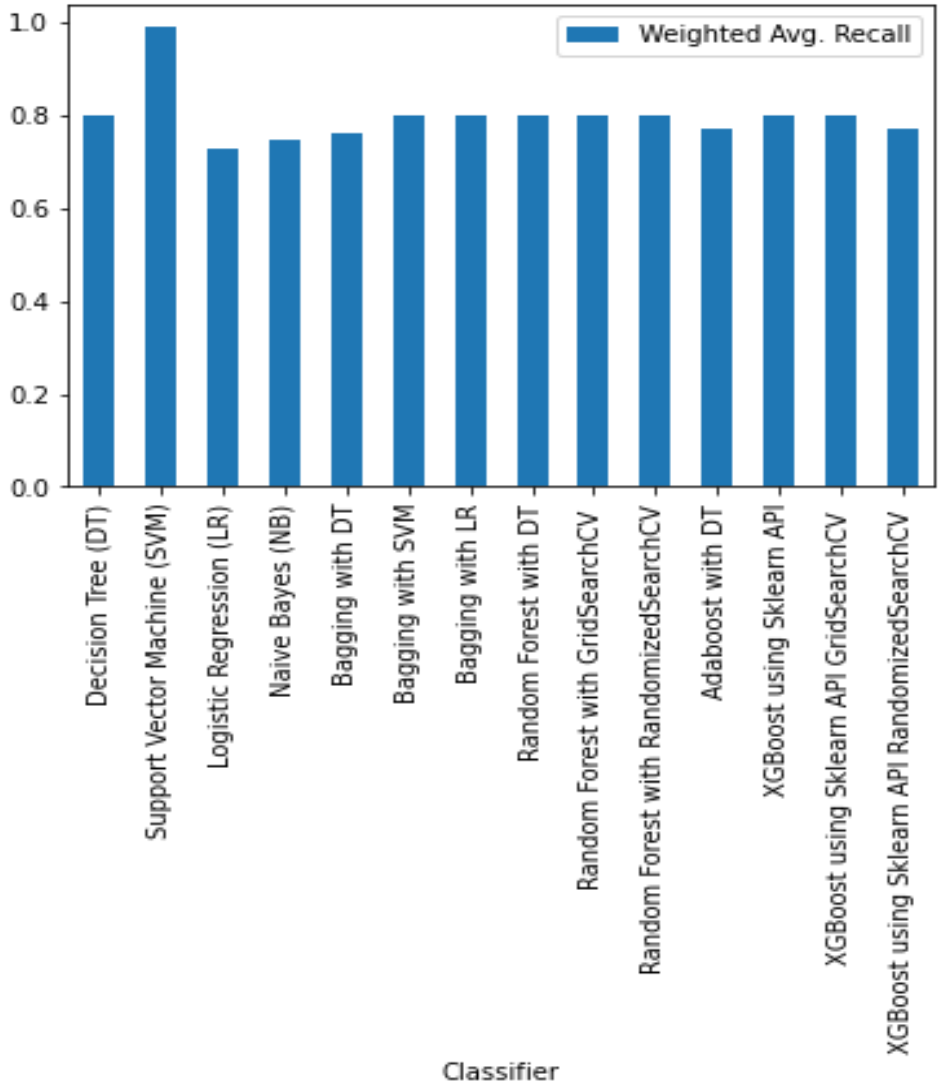
Experimental Results | Text Features



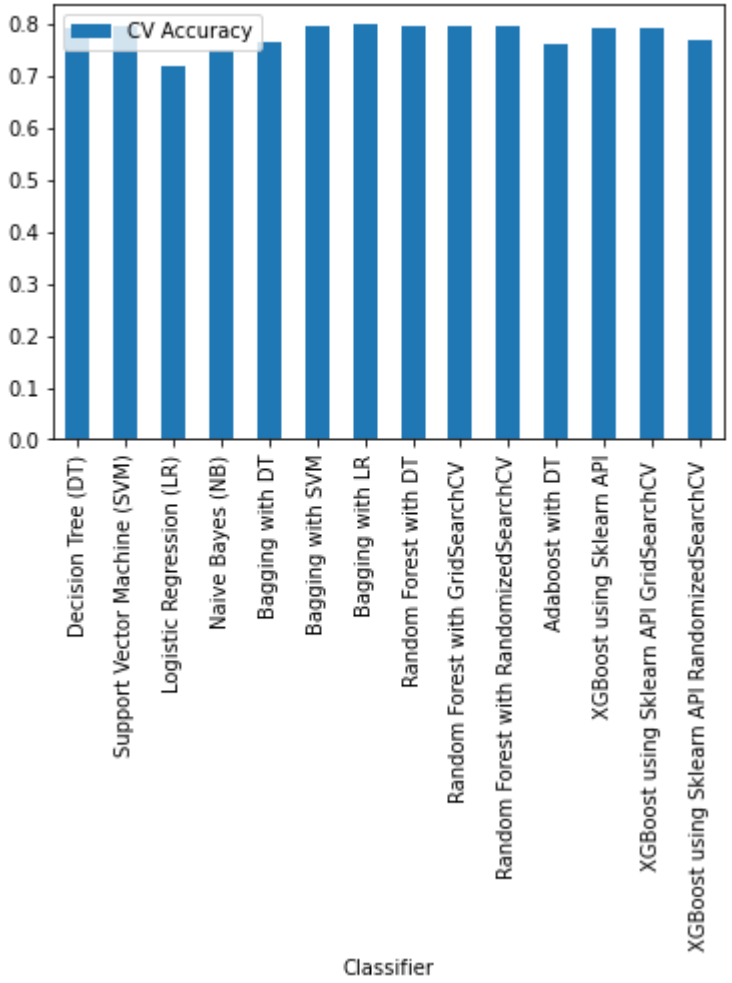


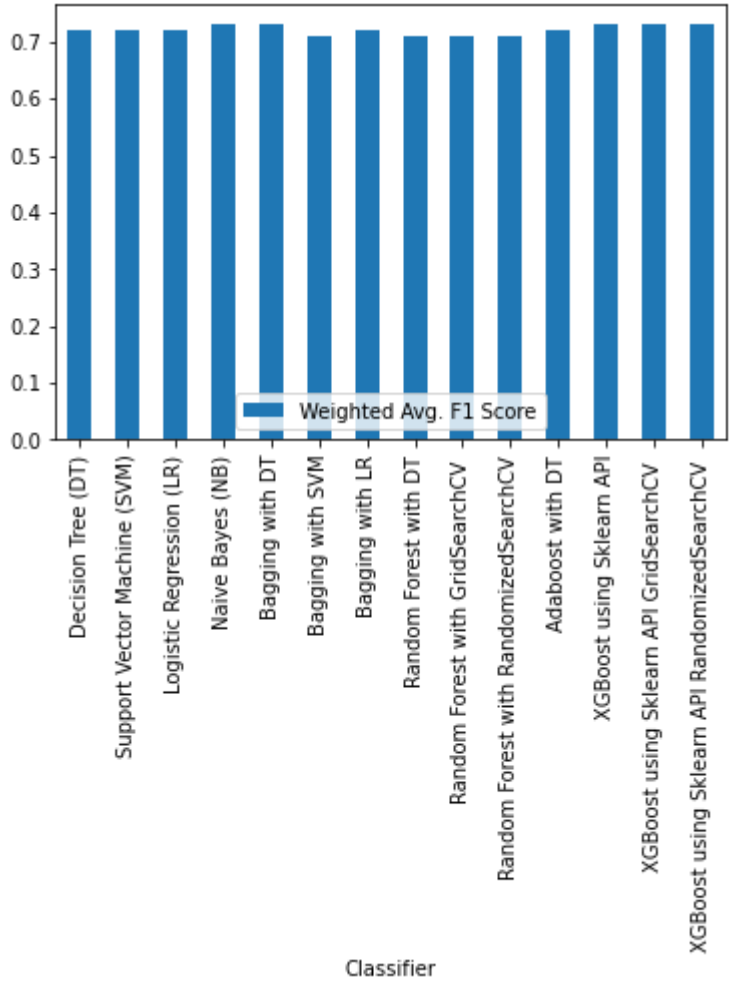
Experimental Results | Text Features

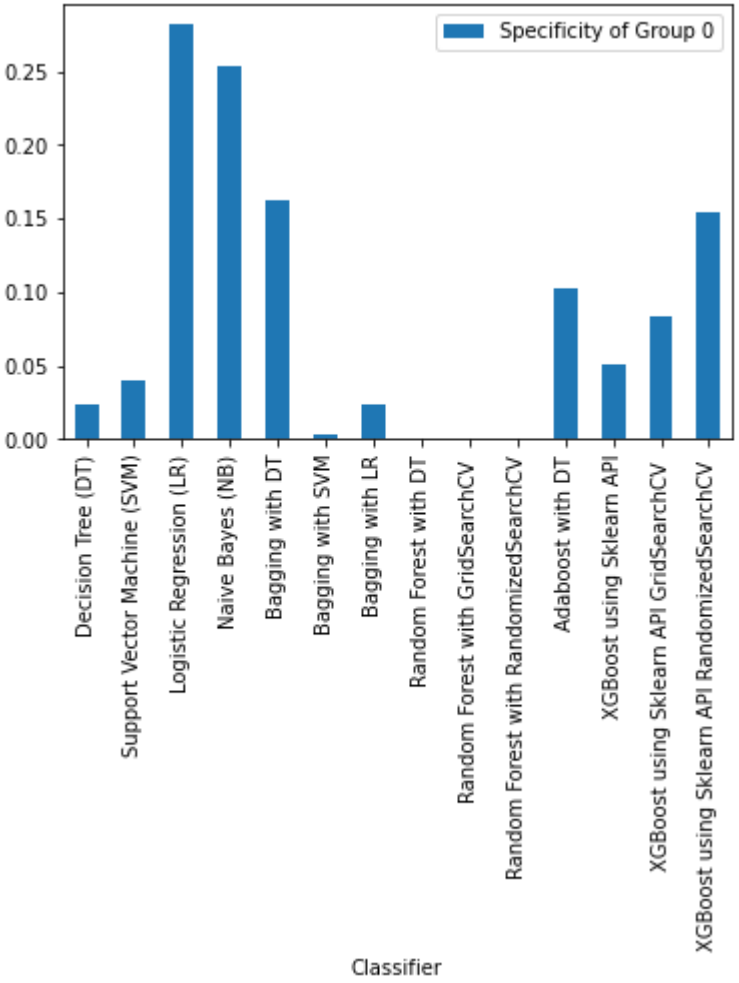




Experimental Results | Text Features

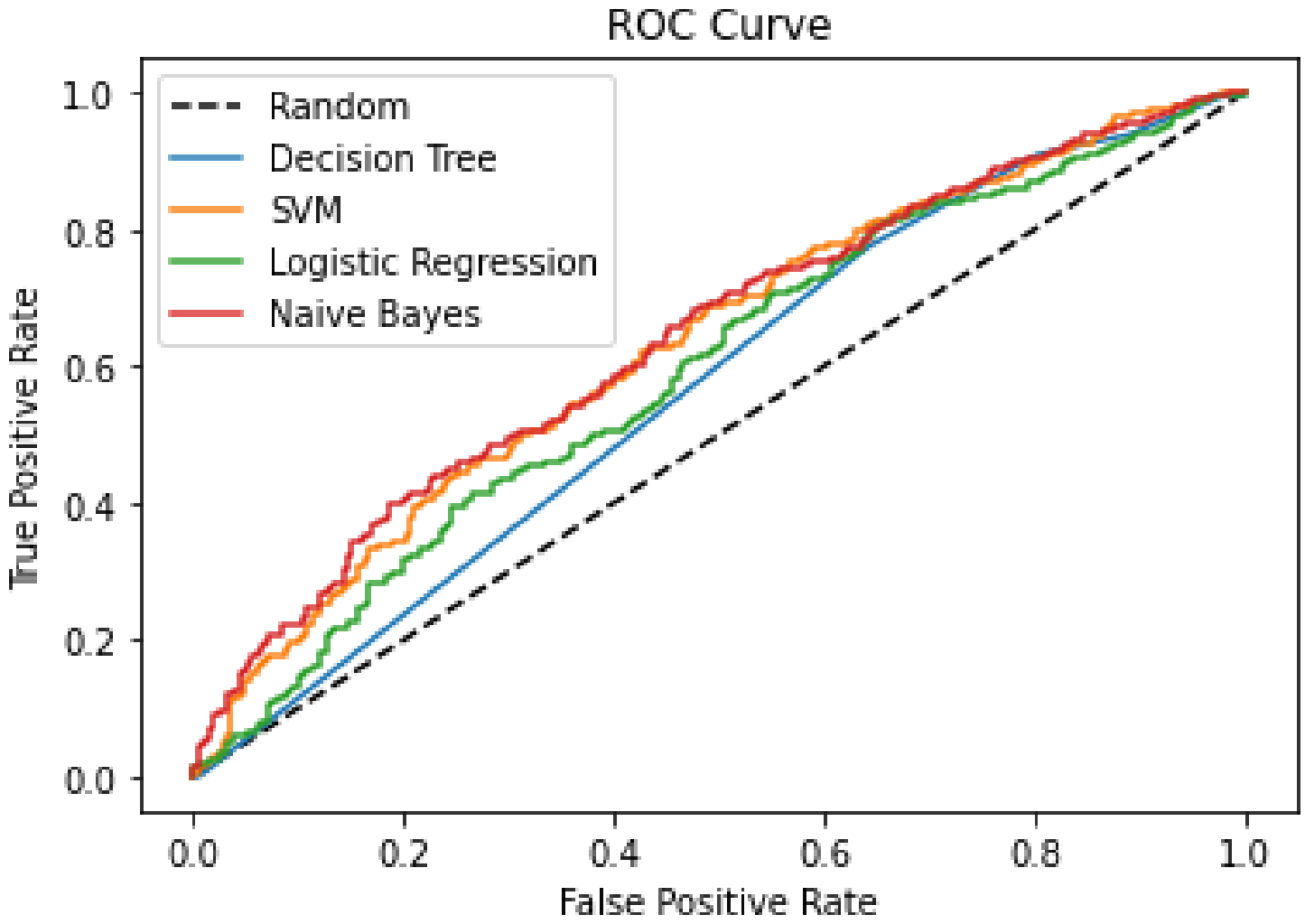






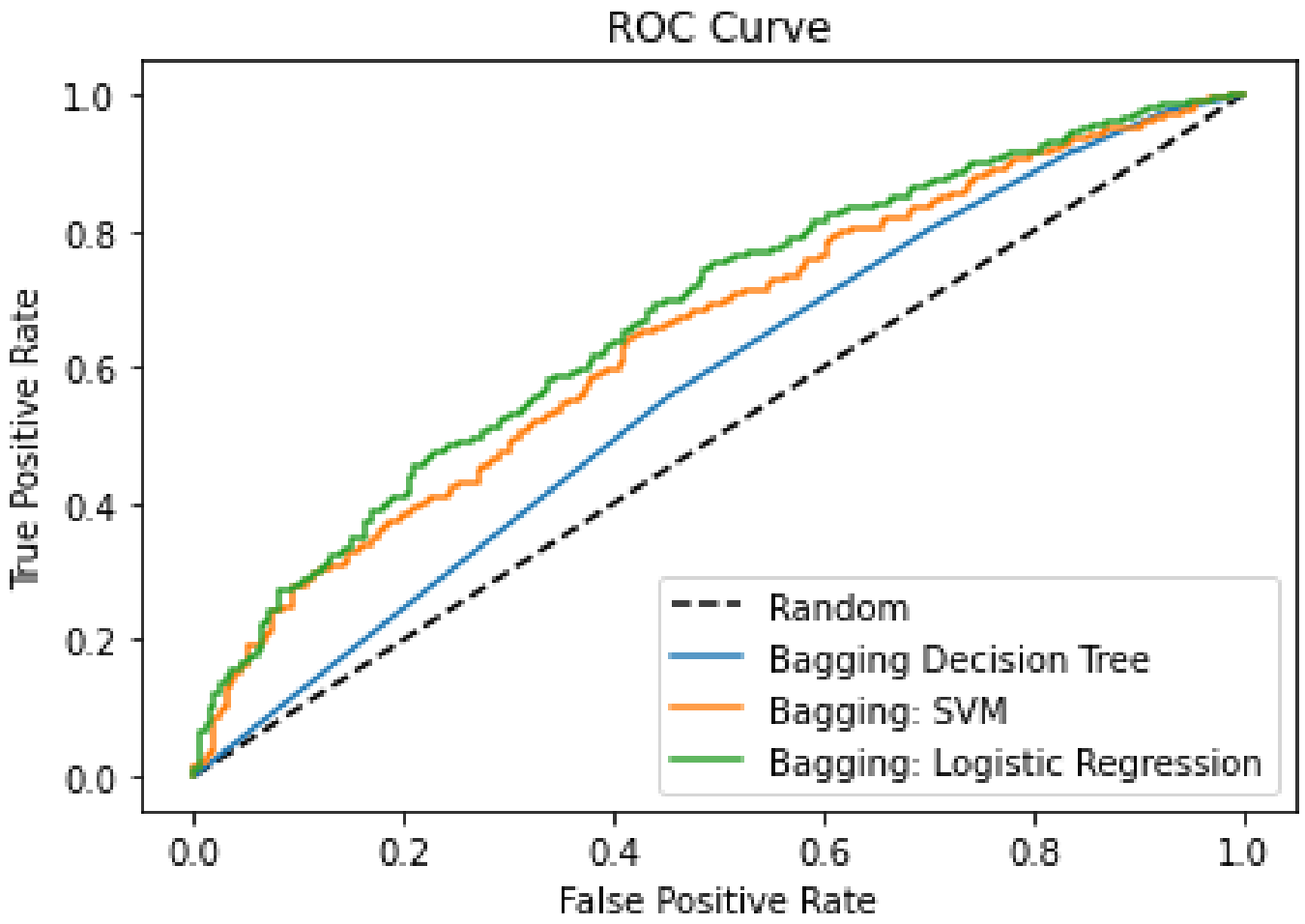
Experimental Results | Text Features

General Classifier



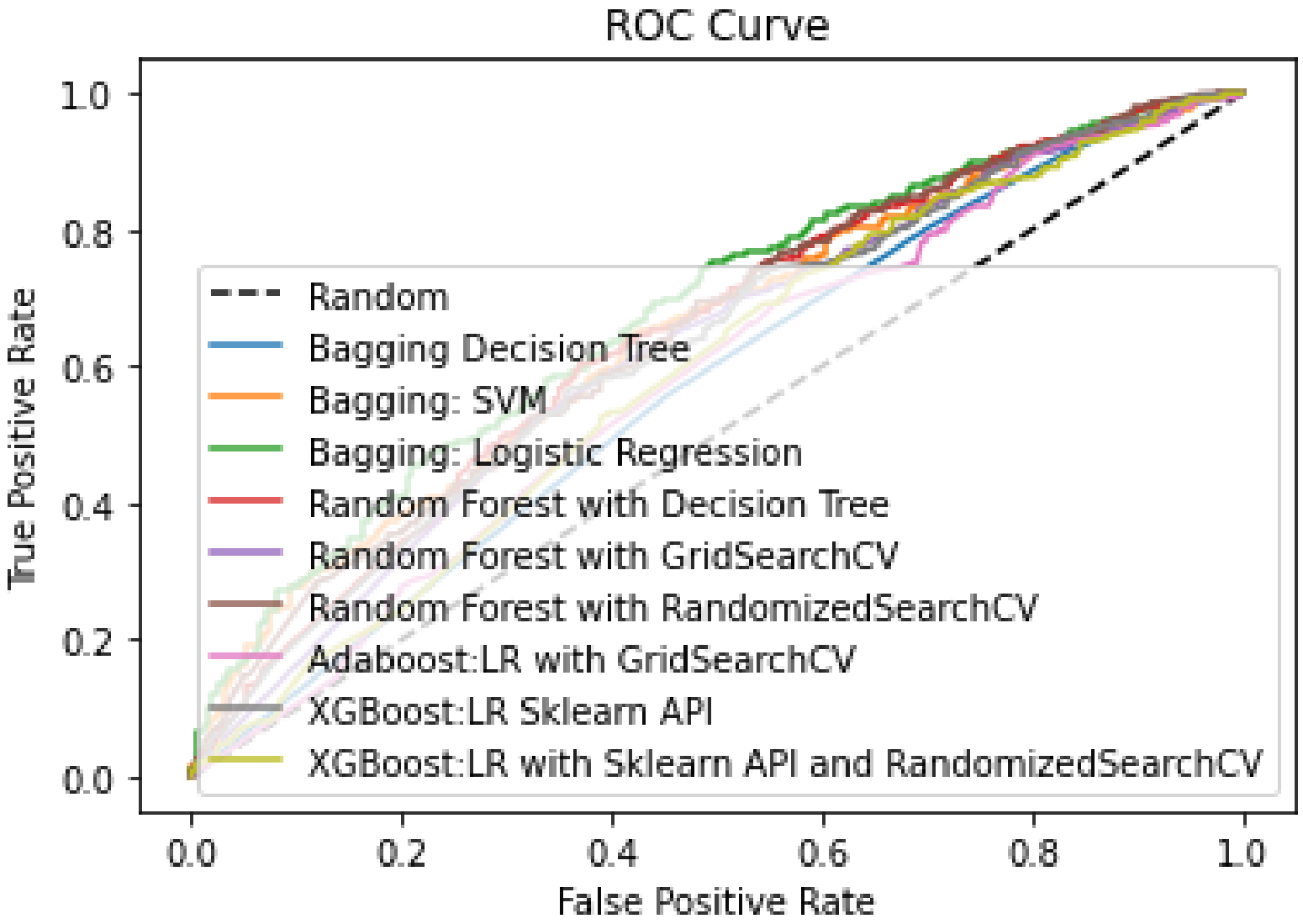
Experimental Results | Text Features

Bagging Classifier



Experimental Results | Text Features

Ensemble Methods



Conclusion and Future Work

- Numeric Models
  - Decision Tree showed the best performance of general classifiers
  - The use of ensemble models increased test accuracy by about 2%
  - Increase in precision and recall when using Random Forest.
  - Bagging with Decision Trees was found to give the largest specificity for the class with high AAS values.
  - XGBClassifier tuned with Randomized Search gave the largest ROC/AUC.
- Text Models
  - Support Vector Machine with linear kernel showed the best performance among general classifiers
  - The use of ensemble models increased test accuracy by about 1%
  - Performance of Random Forest was consistent with both the GridSearchCV and RandomSearchCV
  - Logistic Regression was found to give the largest specificity for the class with high AAS values.
  - The overall performance of XGBoost was the best among the Ensemble methods for text features

Conclusion and Future Work

- Perform clustering of clinical trials using text features
  - Titles
  - Abstracts
- Use text data to make classification models to predict trial categories
- Expand the original datasets with newer trials in hopes of improving model accuracy
- Use more complex text extraction methods in order to limit the number of features used