AG News Topic Classification

CSML 1010 - Winter 2020 - Group 20

Tony Lee, Viswesh Krishnamurthy

YORK
UNIVERSITÉ
UNIVERSITY

PROBLEM SELECTION & DEFINITION

- A text classification problem was chosen from the website https://datasets.quantumstat.com/. We chose the AG News corpus dataset

- The goal of this project is to develop a text classifier model that can accept a news 'headline' and 'content' of the news to classify the news article into one of the 4 following categories
  - World (coded as 1)
  - Sports (2)
  - Business (3)
  - Sci/Tech (4)

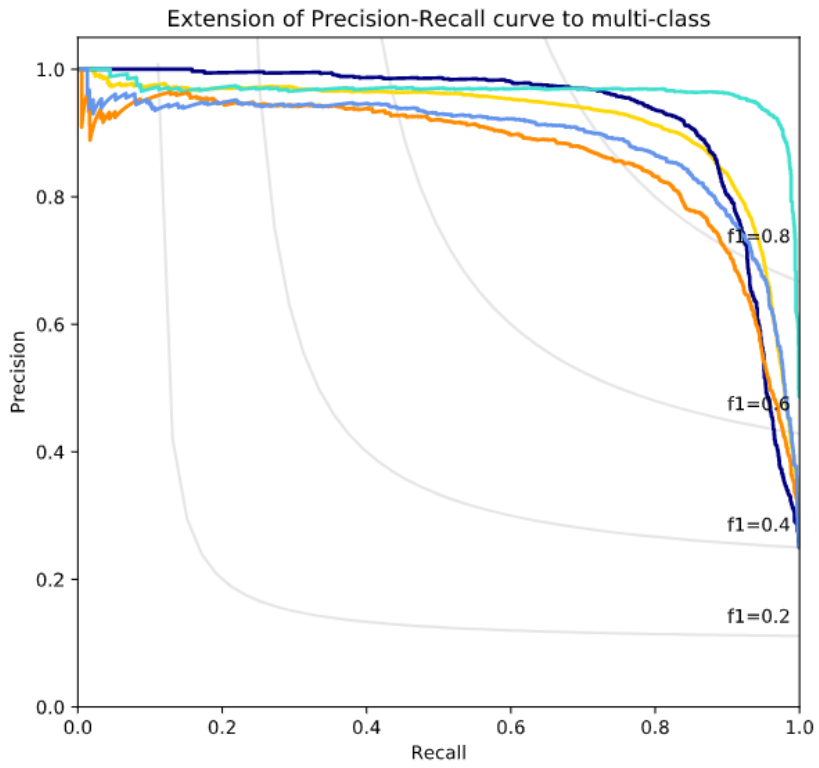
PROJECT MILESTONE 2 – 6th May 2020

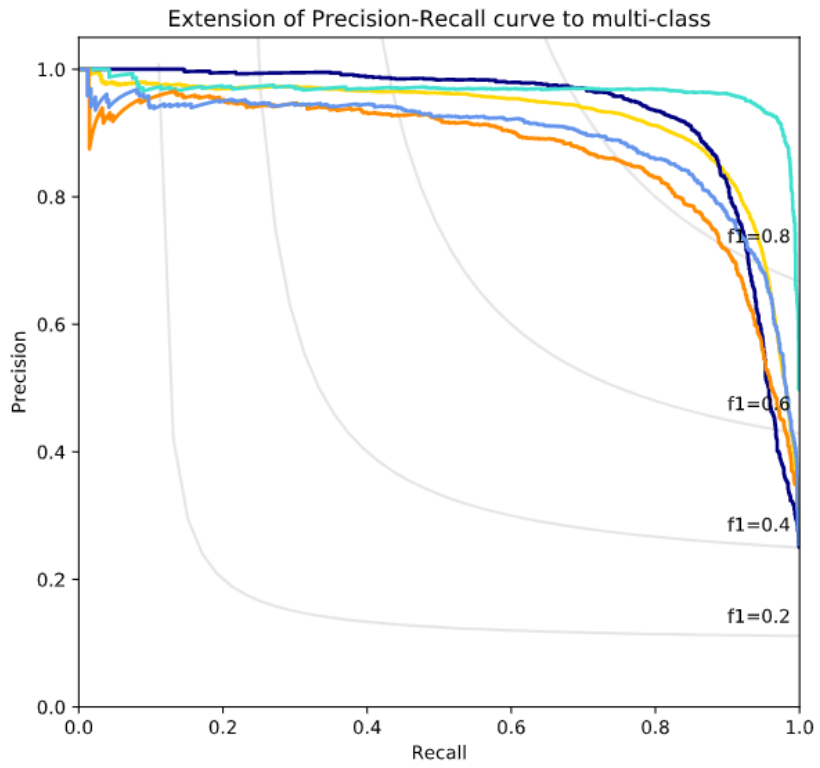
MODELS

- With the previous SVM model as the baseline, the following models were built,
  - Logistic Regression
  - Naïve Bayes
  - Decision Trees

  All the models were built using Word2Vec embedding. The embedding was trained using the 120,000 instances, the entire training set

MODELS – PERFORMANCE COMPARISON





Baseline Model – SVM

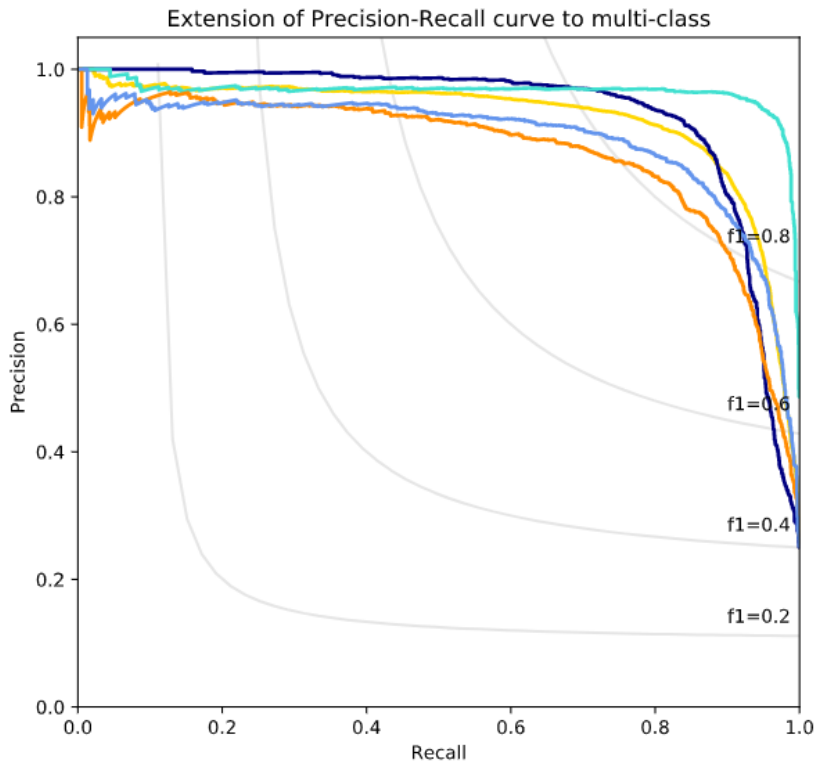
- 120,000 instances
- Min word count = 5
- No.of Dimensions = 300

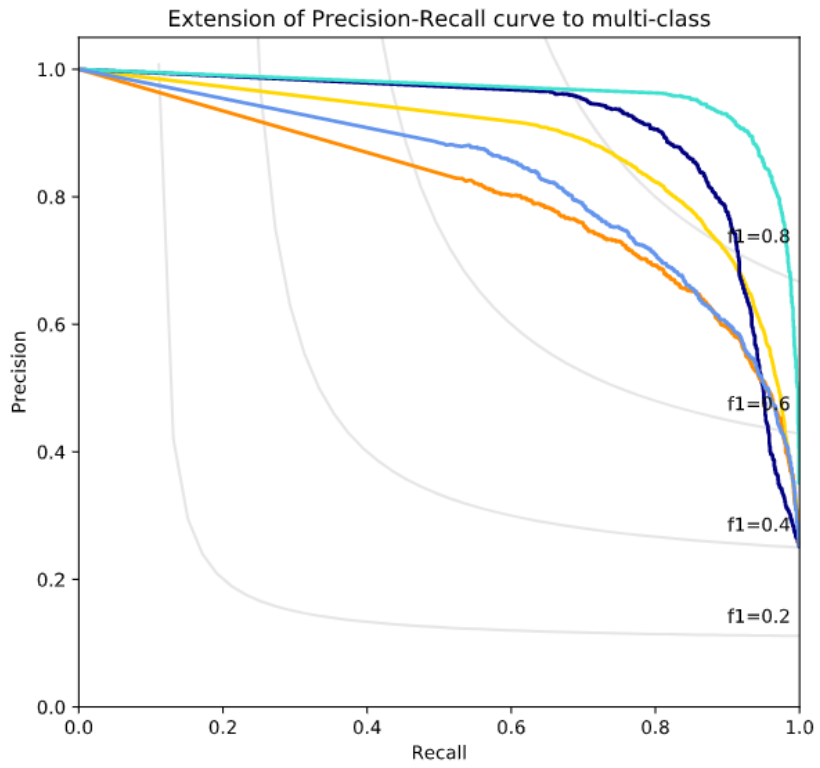
Logistic Regression

- 120,000 instances
- Min word count = 5
- No.of Dimensions = 300

YORK UNIVERSITÉ UNIVERSITY

MODELS – PERFORMANCE COMPARISON (Cont'd)





Baseline Model – SVM

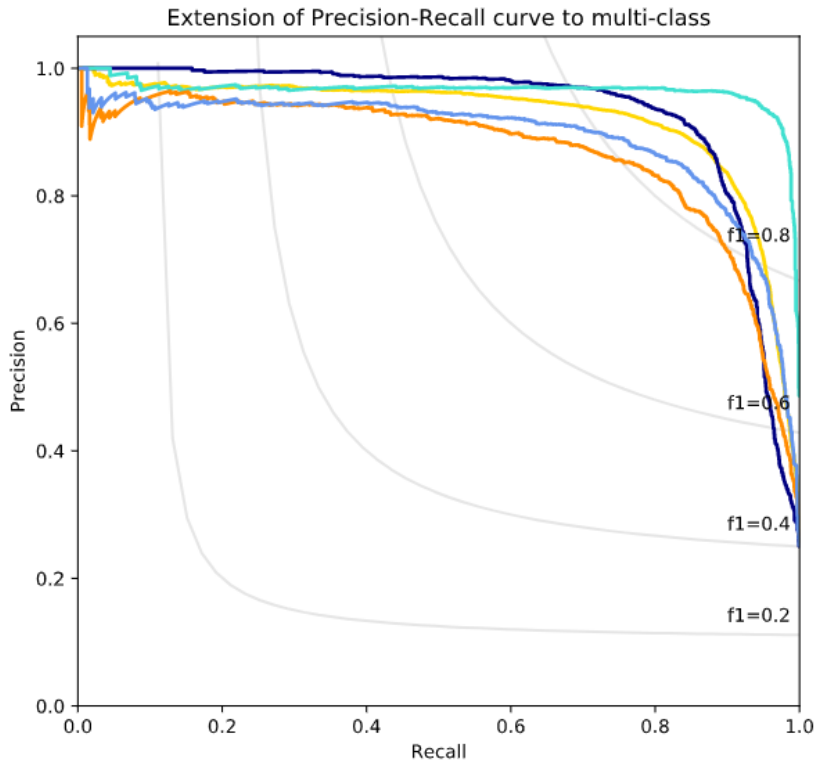
- 120,000 instances
- Min word count = 5
- No.of Dimensions = 300

Naïve Bayes

- 120,000 instances
- Min word count = 5
- No.of Dimensions = 300

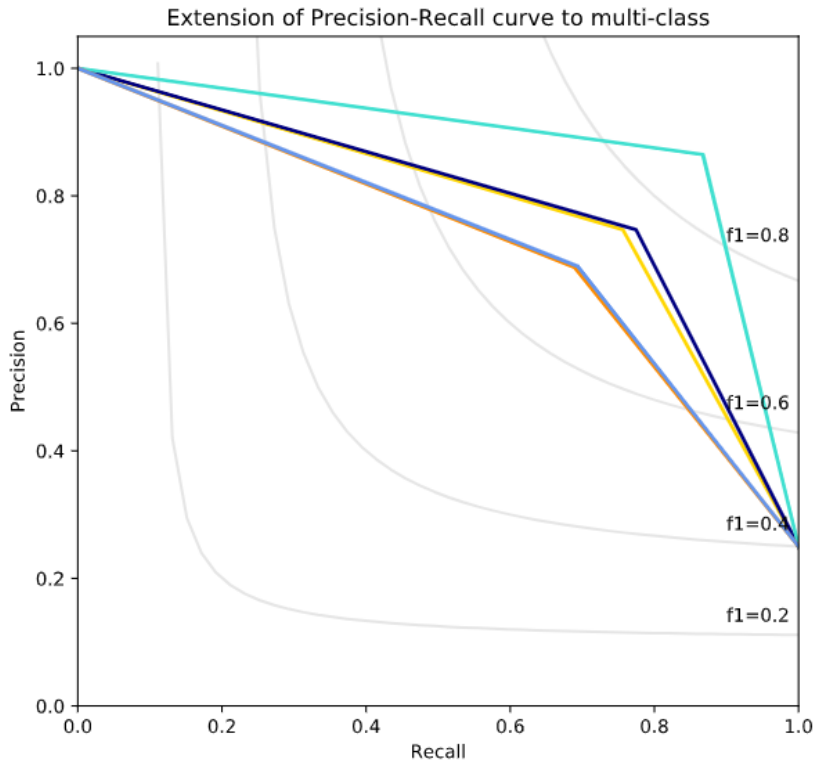
YORK UNIVERSITÉ UNIVERSITY

MODELS – PERFORMANCE COMPARISON (Cont'd)



Baseline Model – SVM

- 120,000 instances
- Min word count = 5
- No.of Dimensions = 300



Decision Trees

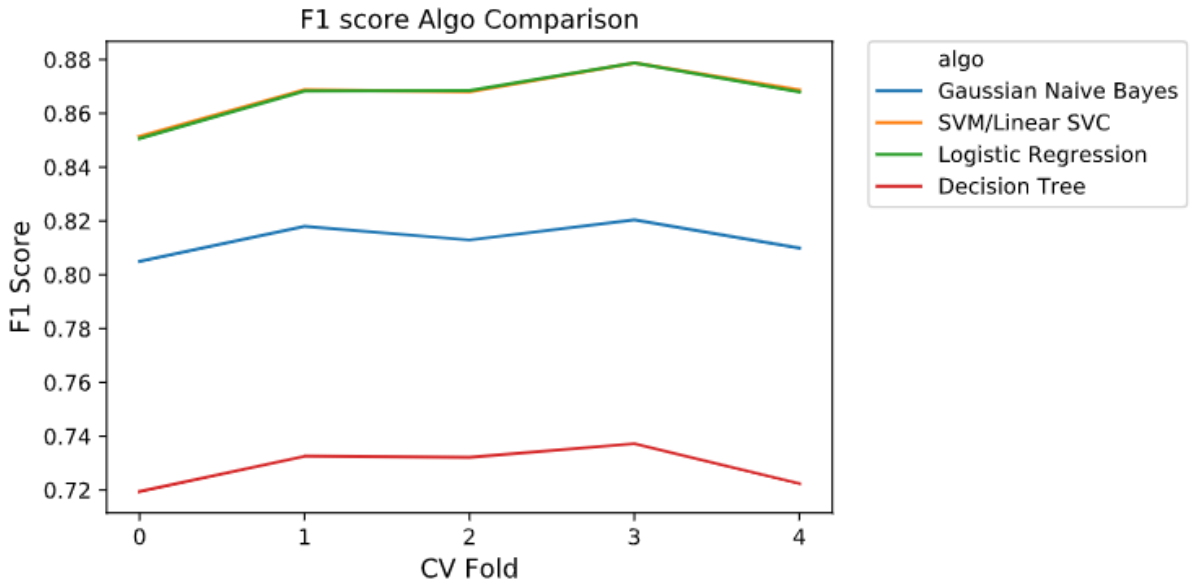
- 120,000 instances
- Min word count = 5
- No.of Dimensions = 300

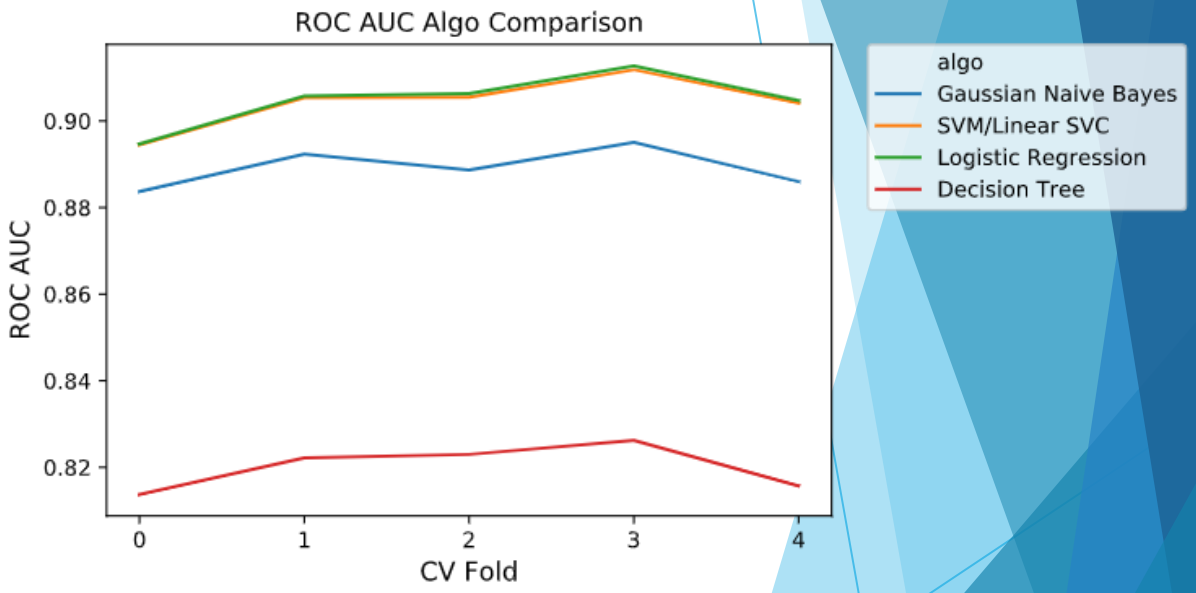
YORK UNIVERSITÉ UNIVERSITY

CROSS VALIDATION

- We ran a 5 fold cross validation on the training dataset and re-ran all the models
- The models on the "solo" runs were trained on the 120,000 instance training set and tested on a separate 7600 instance test set
- The results of cross validation closely follow the previous "solo" runs
- Logistic Regression & SVM compare very closely in terms of the metrics. However, SVM was very resource intensive

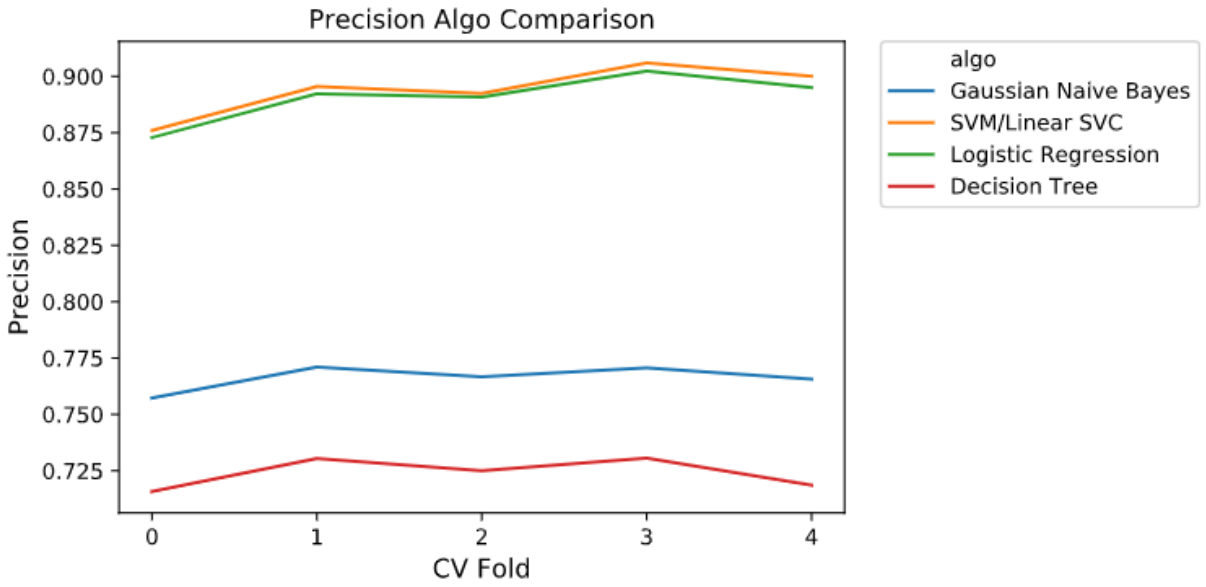
CROSS VALIDATION – COMPARISON OF RESULTS

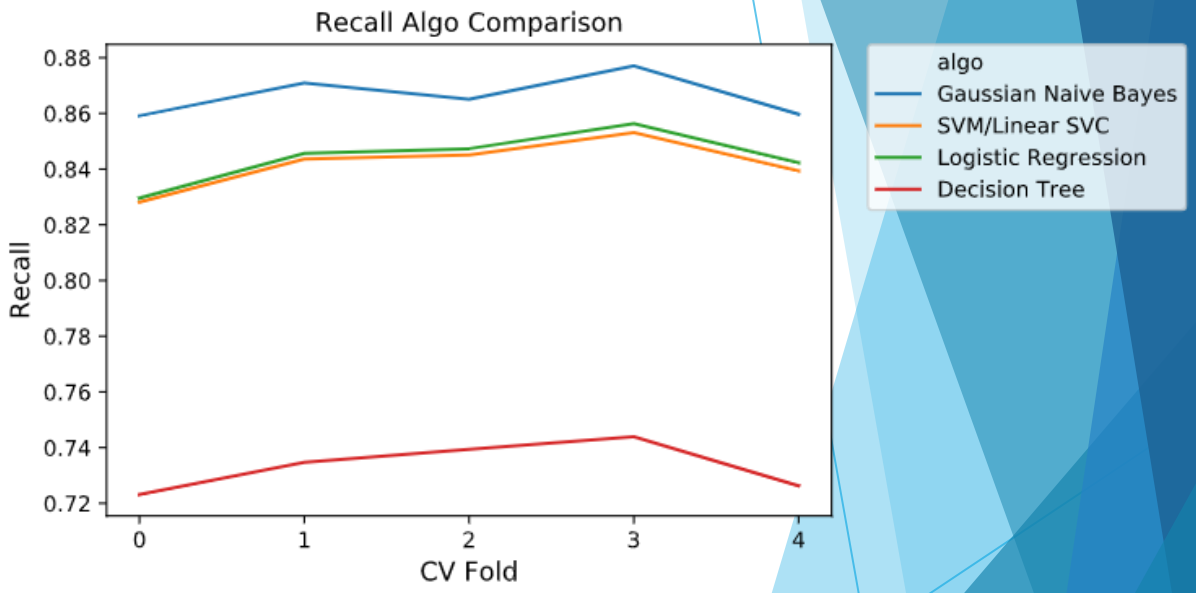




YORK UNIVERSITÉ UNIVERSITY

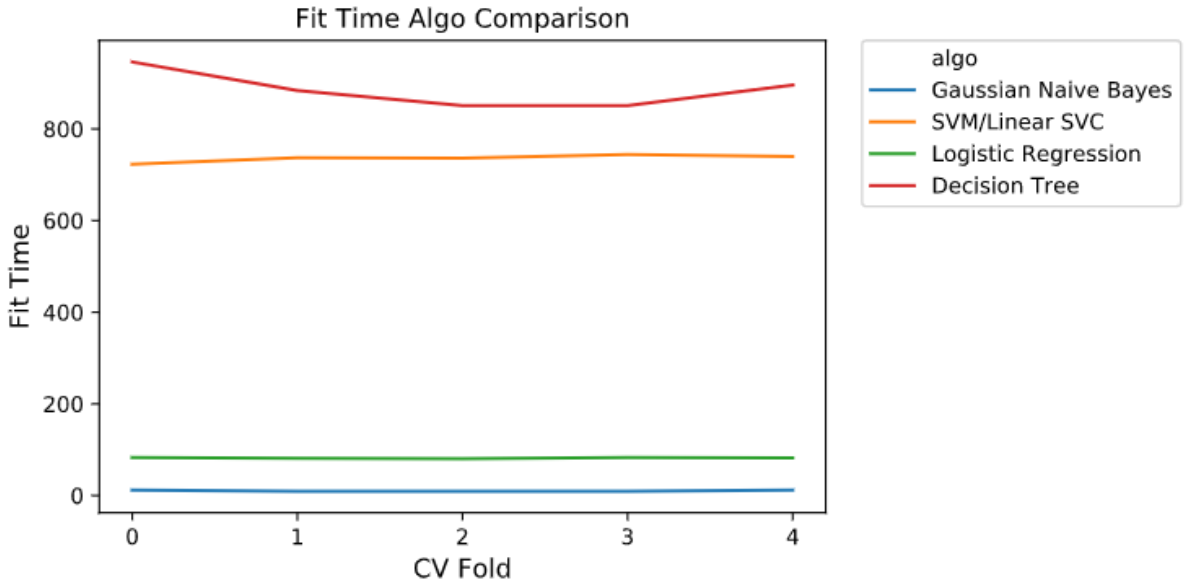
CROSS VALIDATION – COMPARISON OF RESULTS





YORK UNIVERSITÉ UNIVERSITY

CROSS VALIDATION – COMPARISON OF RESULTS

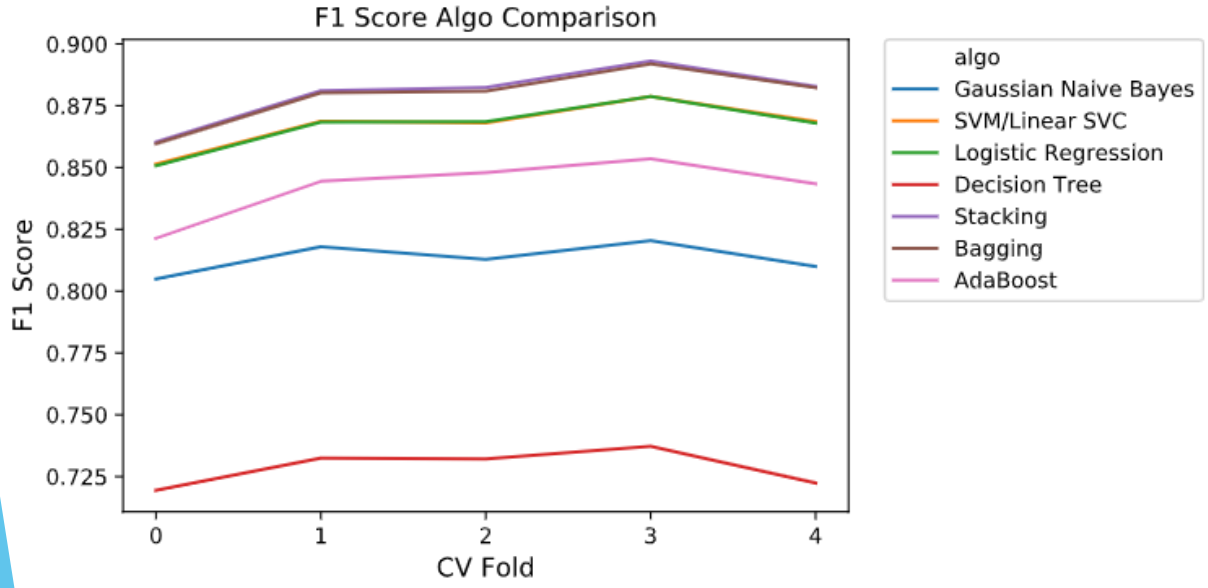


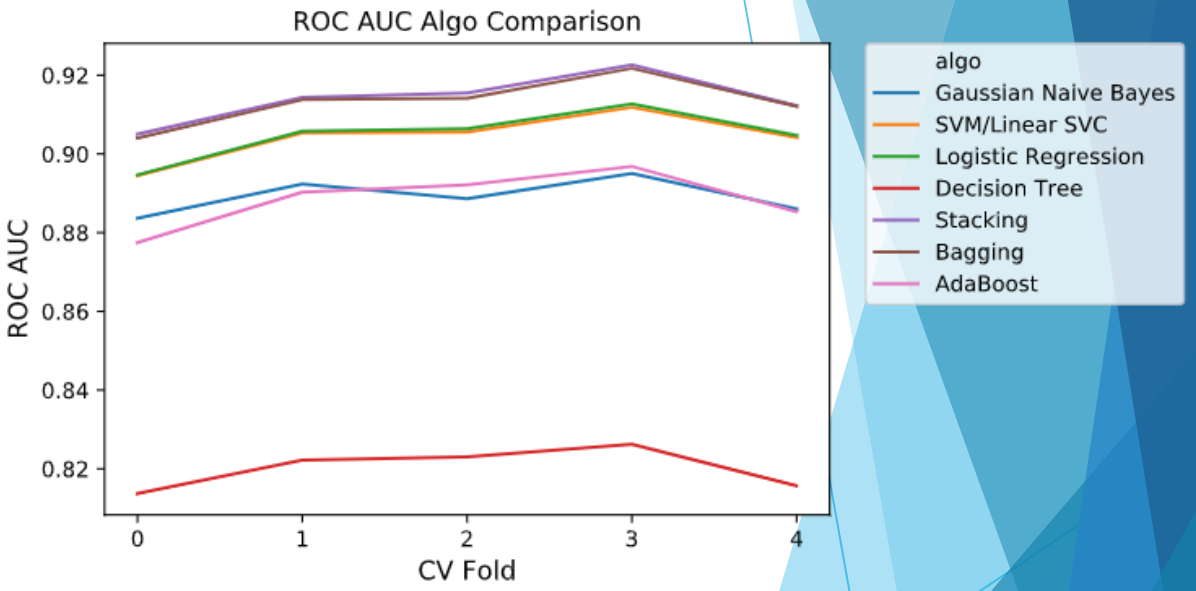
- Based on the previous metrics and comparison, Logistic Regression emerges as the model of choice

ENSEMBLE METHODS

- The following ensemble methods were used
  - Bagging
    - The base estimator was chosen to be Logistic Regression
  - Stacking
    - The initial estimators in stacking were chosen to be Naïve bayes and SVM. The meta learner was Logistic Regression
  - Boosting
    - An Adaboost classifier was used for boosting in this iteration.

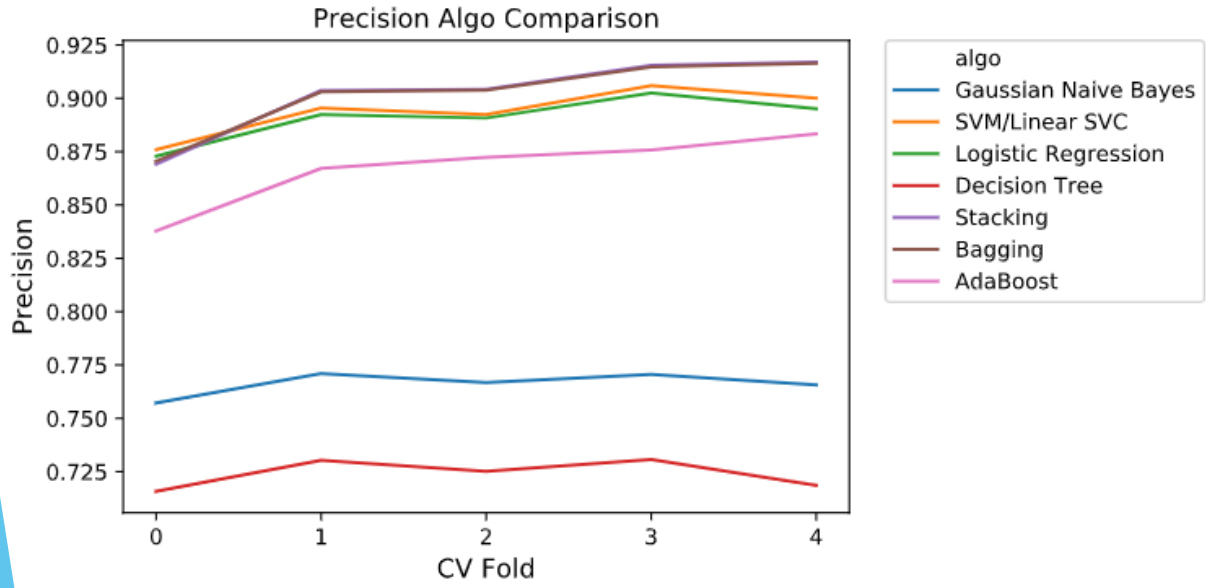
ENSEMBLE METHODS vs OTHERS

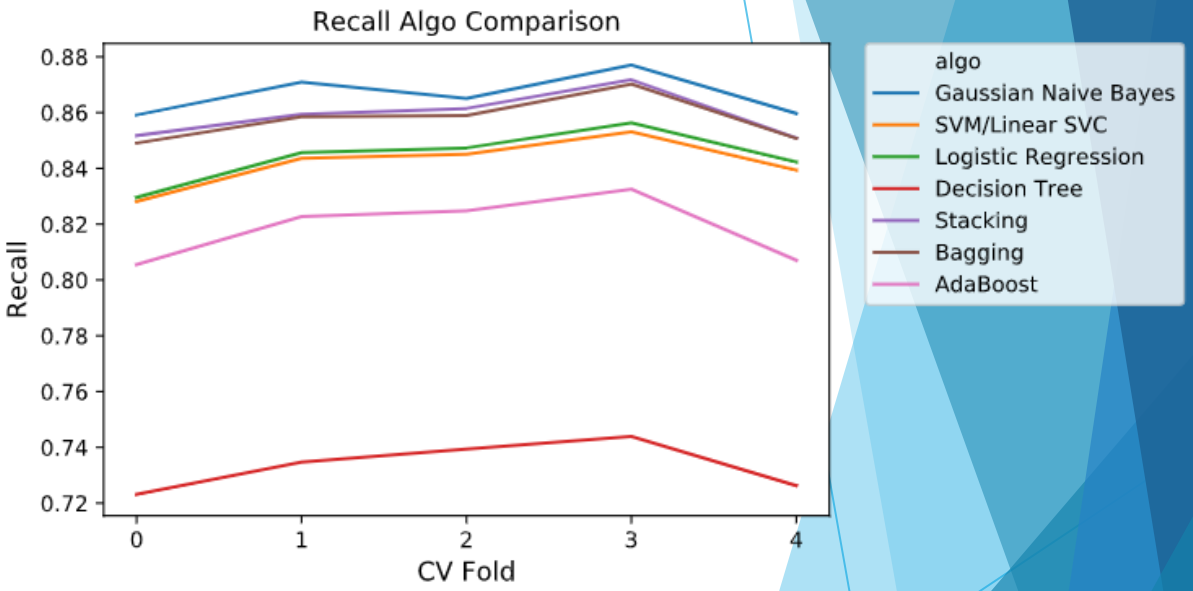




YORK UNIVERSITÉ UNIVERSITY

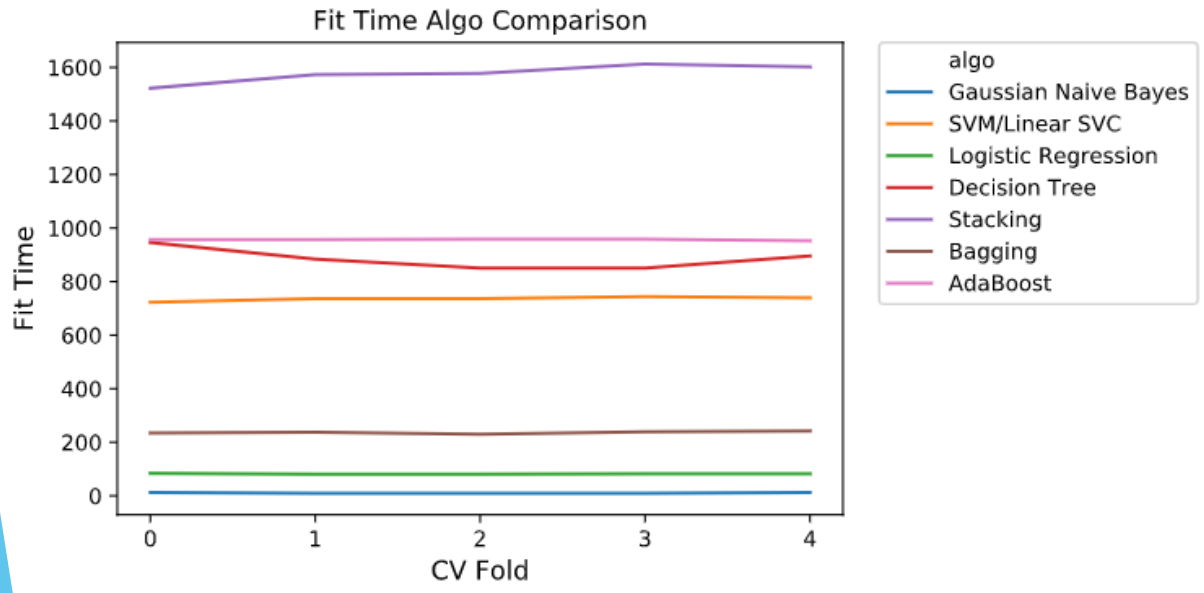
ENSEMBLE METHODS vs OTHERS





YORK UNIVERSITÉ UNIVERSITY

ENSEMBLE METHODS vs OTHERS



- Based on the metrics seen so far, it is clear that "Bagging" ensemble method is the algorithm of choice
- Stacking and Bagging perform almost similarly in terms of F1 Score, ROC AUC, Precision & Recall
- However, in terms of Fit time, stacking took almost 25 minutes to fit, while bagging took just over 3.5 mins