CSML1010 – Milestone 2

Group 11 – Alex Fung and Patrick Osborne

Instructor – Dr. En-Shiun Annie Lee

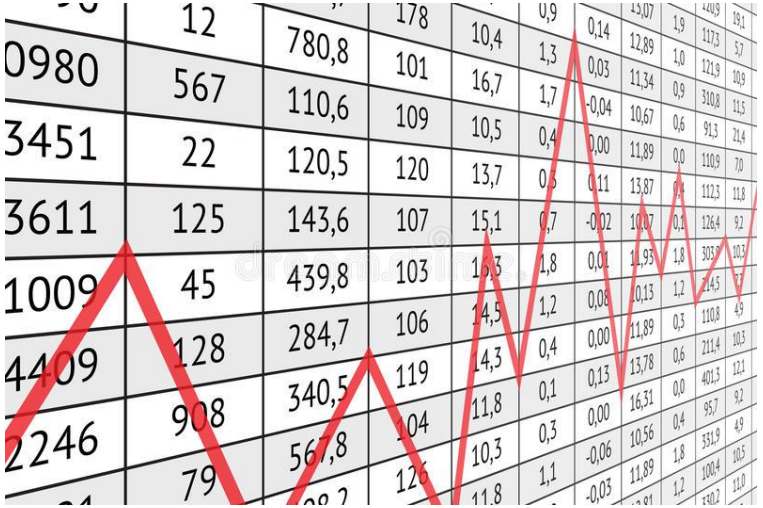
Twitter US Airline – Sentiment Analysis

- **Milestone 2**
  - Model Benchmarking
  - Hyperparameter Optimization
  - Ensemble Methods
  - Analysis and Model Selection

YORK UNIVERSITÉ UNIVERSITY

Twitter US Airline – Sentiment Analysis

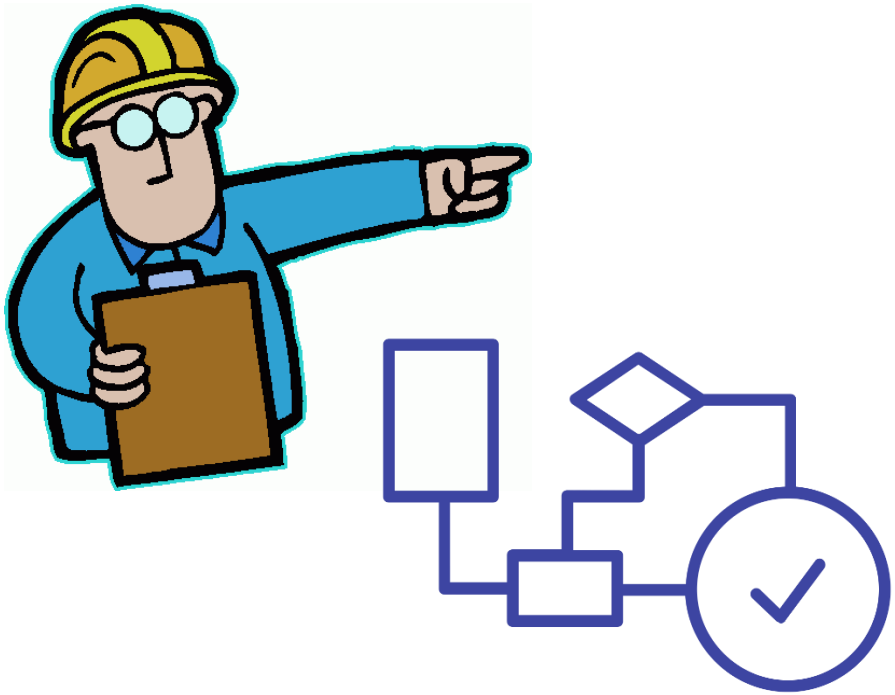
- **Combined Feature Engineered Dataset**
  - Data is fully numerical now
- **Engineered Features**
  - E.g. Emojis/Emoticons
  - Hashtags
  - Etc.
- **GloVe Word Embeddings**
  - (300 dimension)
- **Binary response variable**



YORK UNIVERSITÉ UNIVERSITY

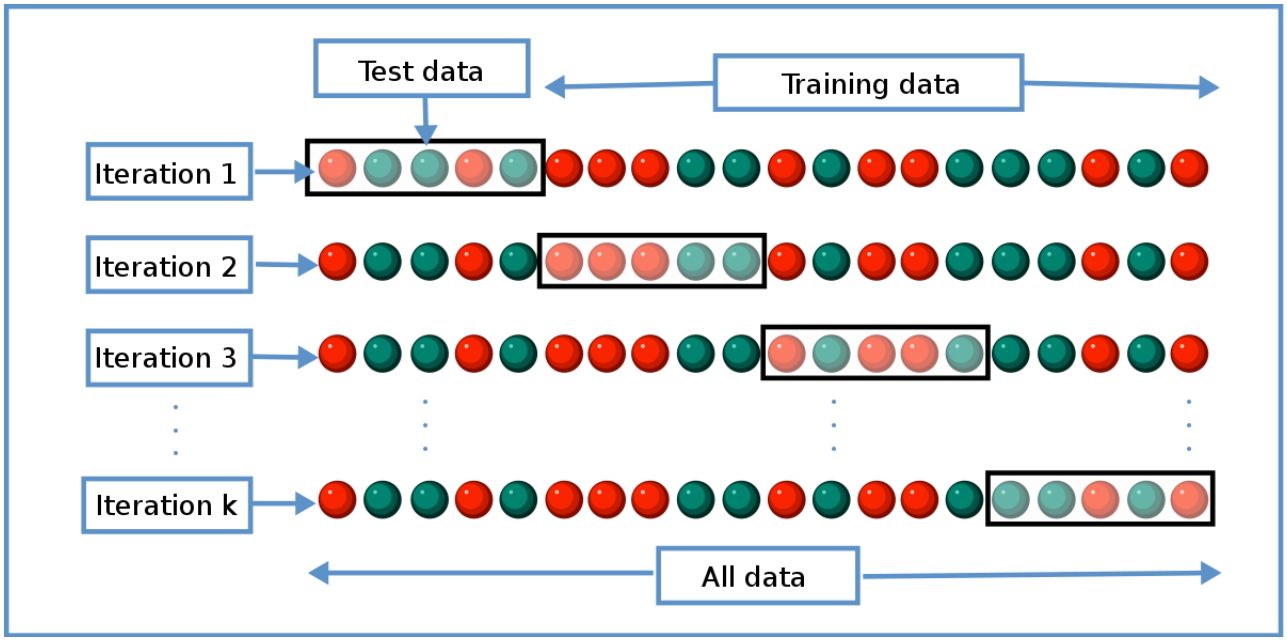
Model Benchmarking – Supervised Learning

- **Supervised Models (Classification)**
  - K Nearest Neighbors
  - Decision Tree
  - Support Vector Classifier
  - Gaussian Naïve Bayes
  - Logistic Regression



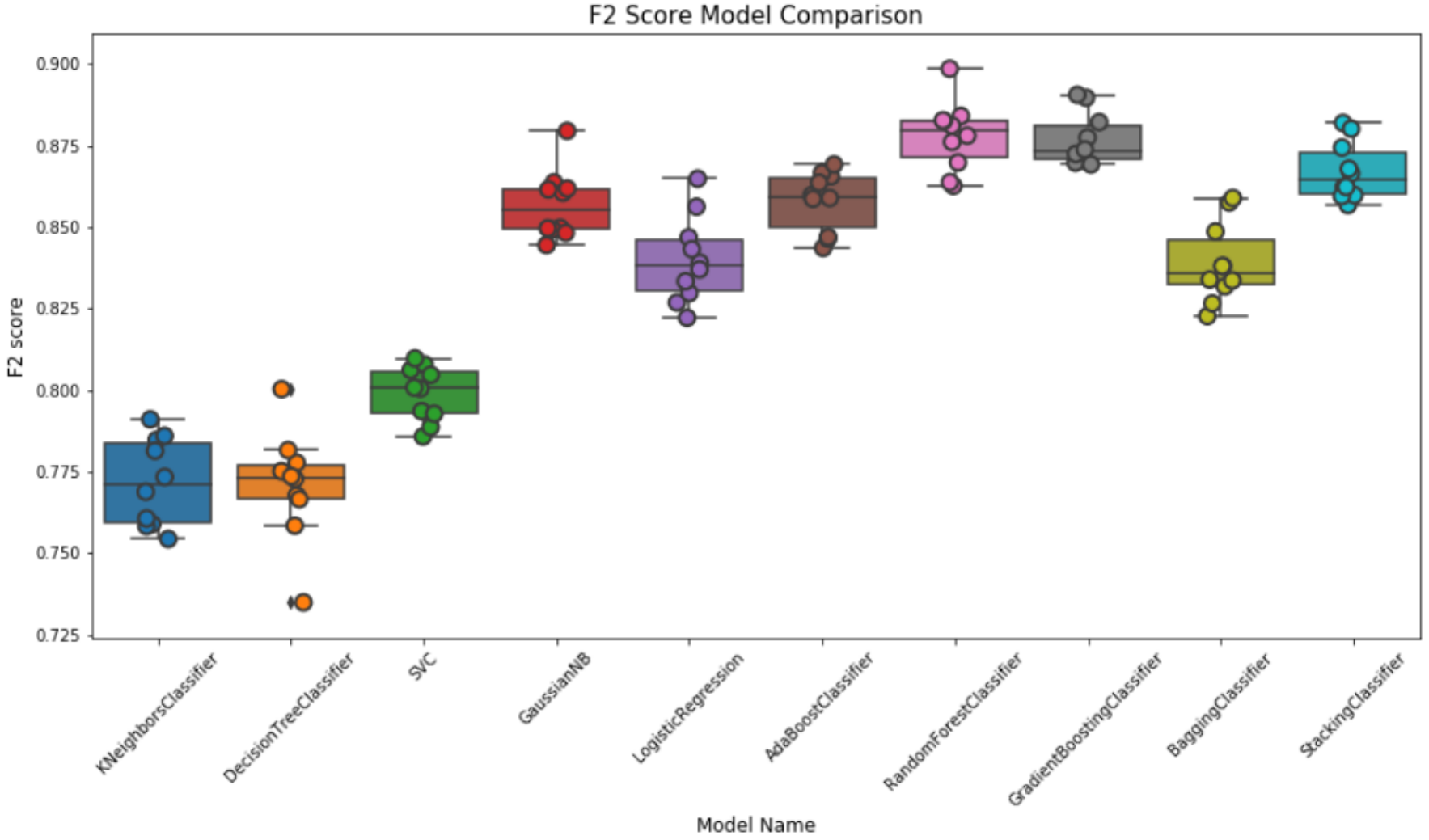
Model Benchmarking - Execution

- **Stratified K-fold Cross Validation**
  - Using sklearn
  - Various Scorers:
    - Fit time, Score time, Bal. Accuracy, F1 score, F2 score, Precision, Recall, ROC-AUC

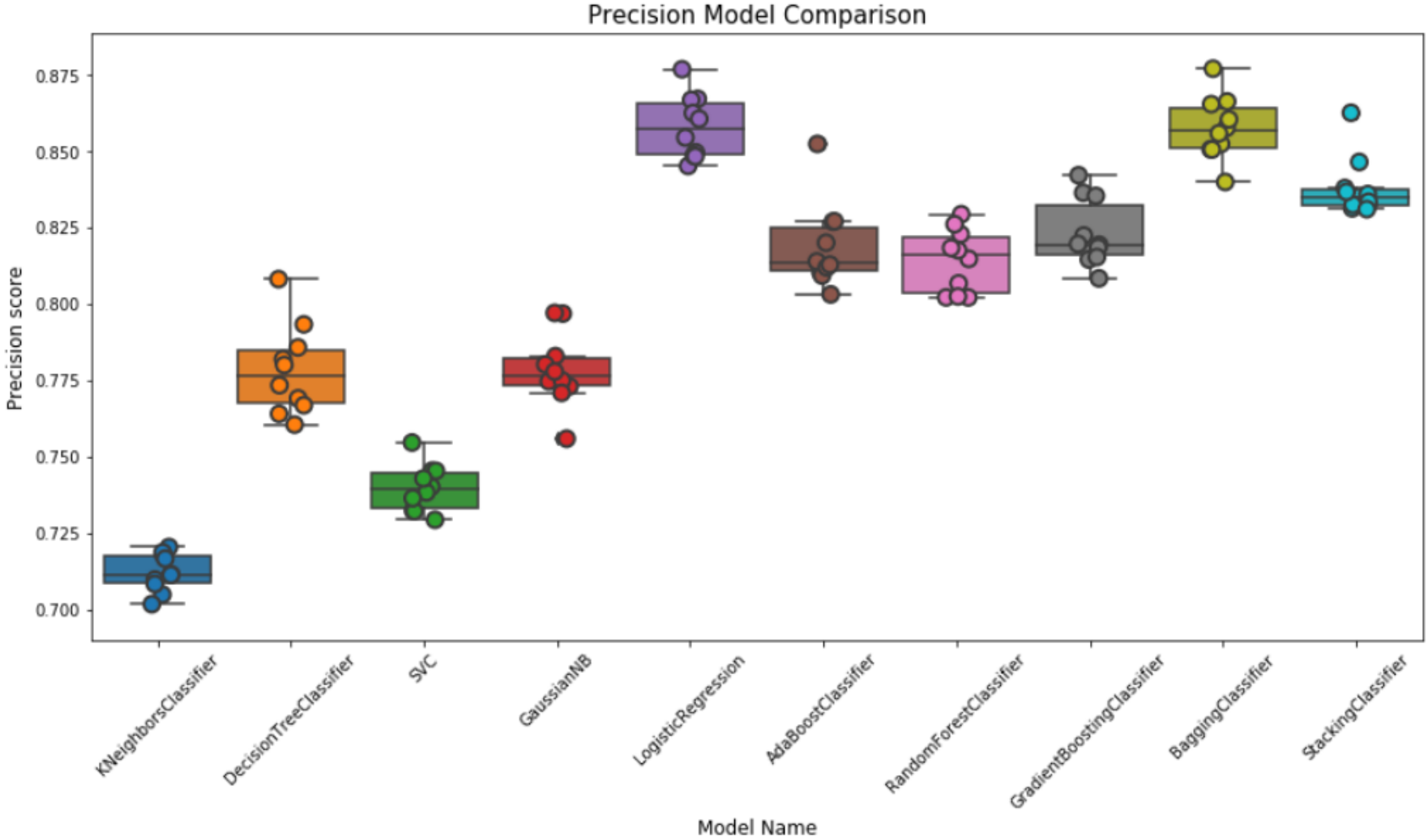


YORK UNIVERSITÉ UNIVERSITY

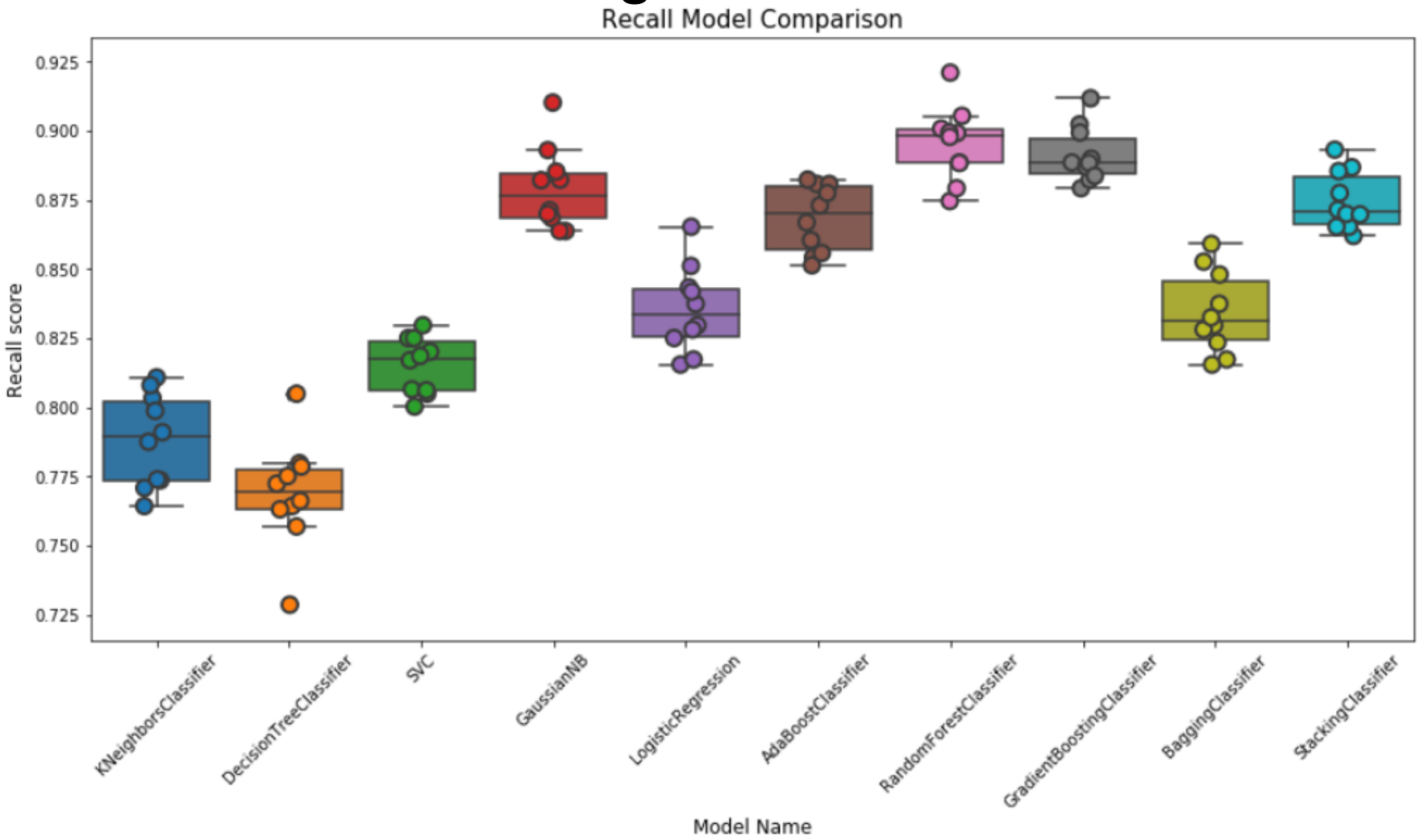
Model Benchmarking – F2 Score



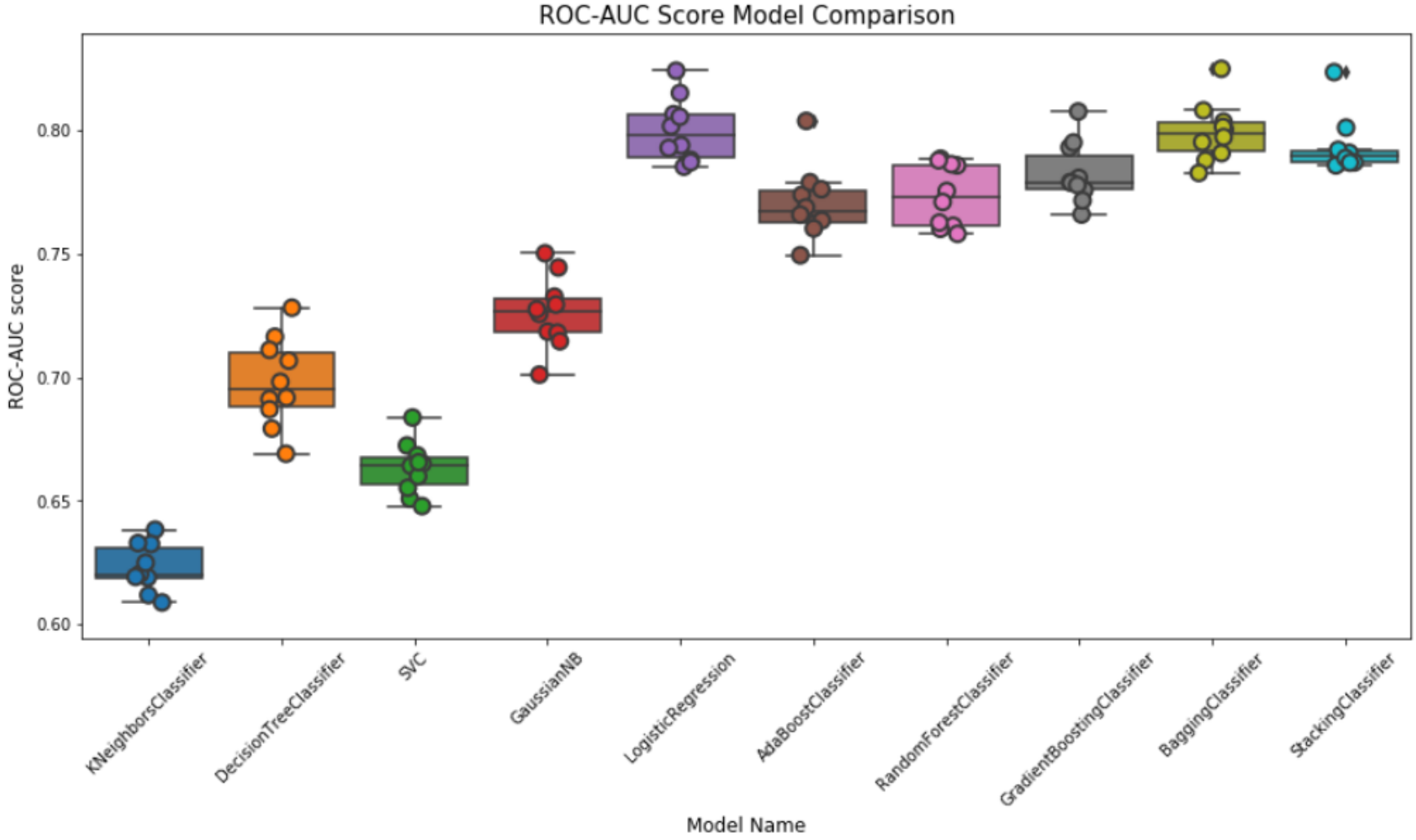
Model Benchmarking – Precision



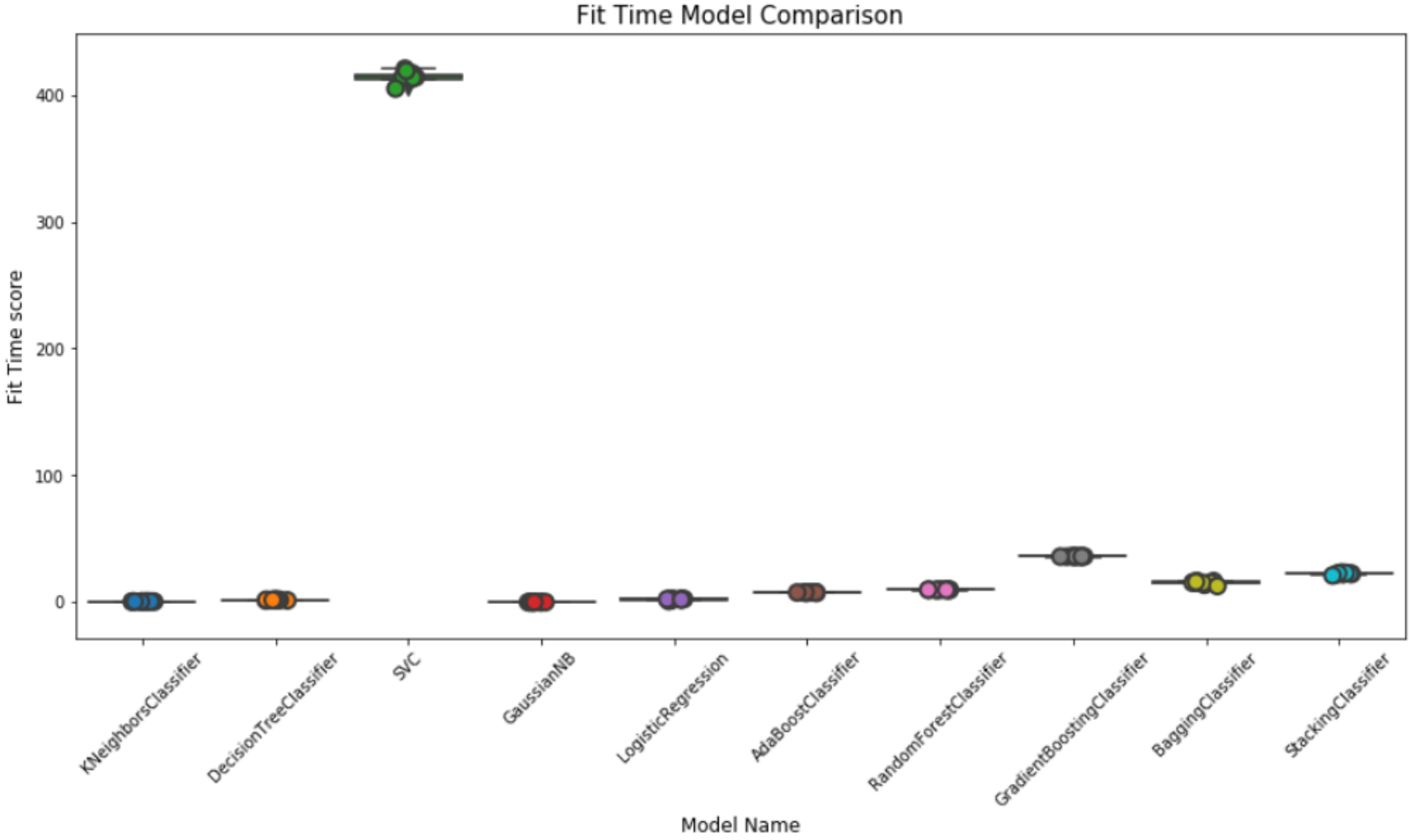
Model Benchmarking – Recall



Model Benchmarking – ROC-AUC



Model Benchmarking – Fit Time

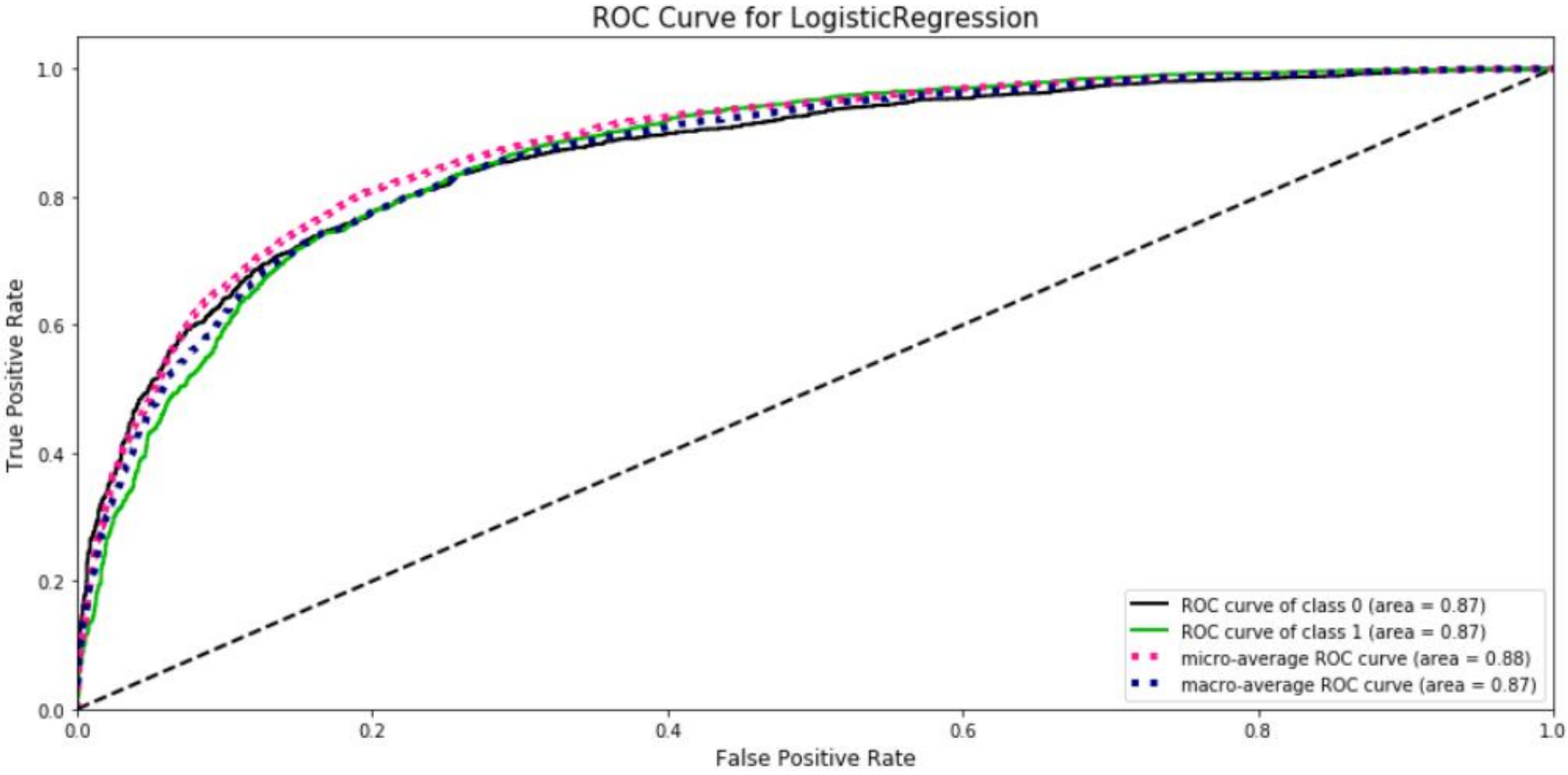


Hyperparameter Optimization

- **Grid Search**
  - Exhaustive test of parameters from a specified list
  - Time consuming, but certain to provide the best set of parameters
- **Random Search**
  - Random search of provide parameters
  - Faster but could possibly miss some combinations of parameters

Hyperparameter Optimization

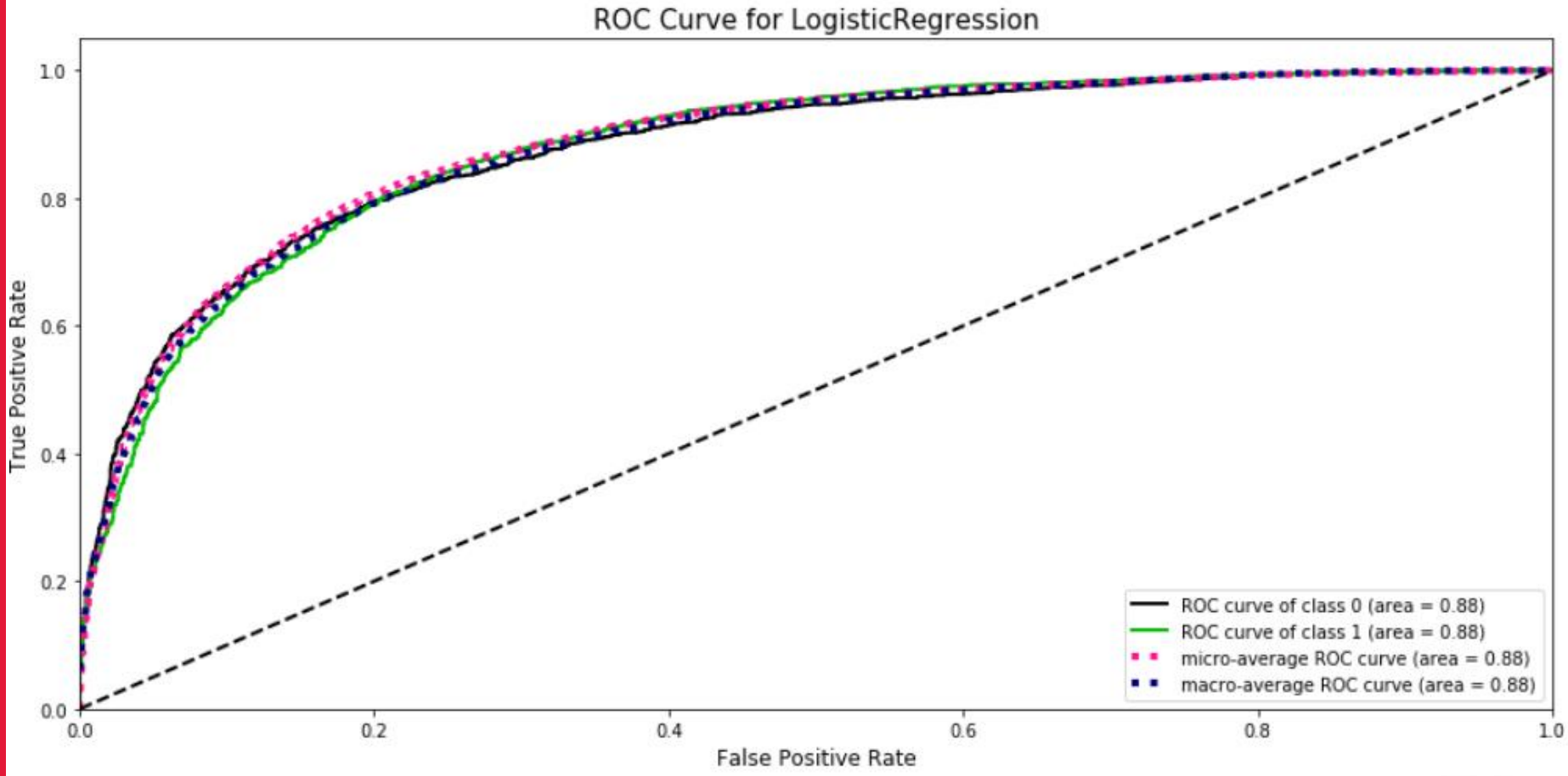
- *Logistic Regression – Pre-Optimization*



YORK UNIVERSITÉ UNIVERSITY

Hyperparameter Optimization

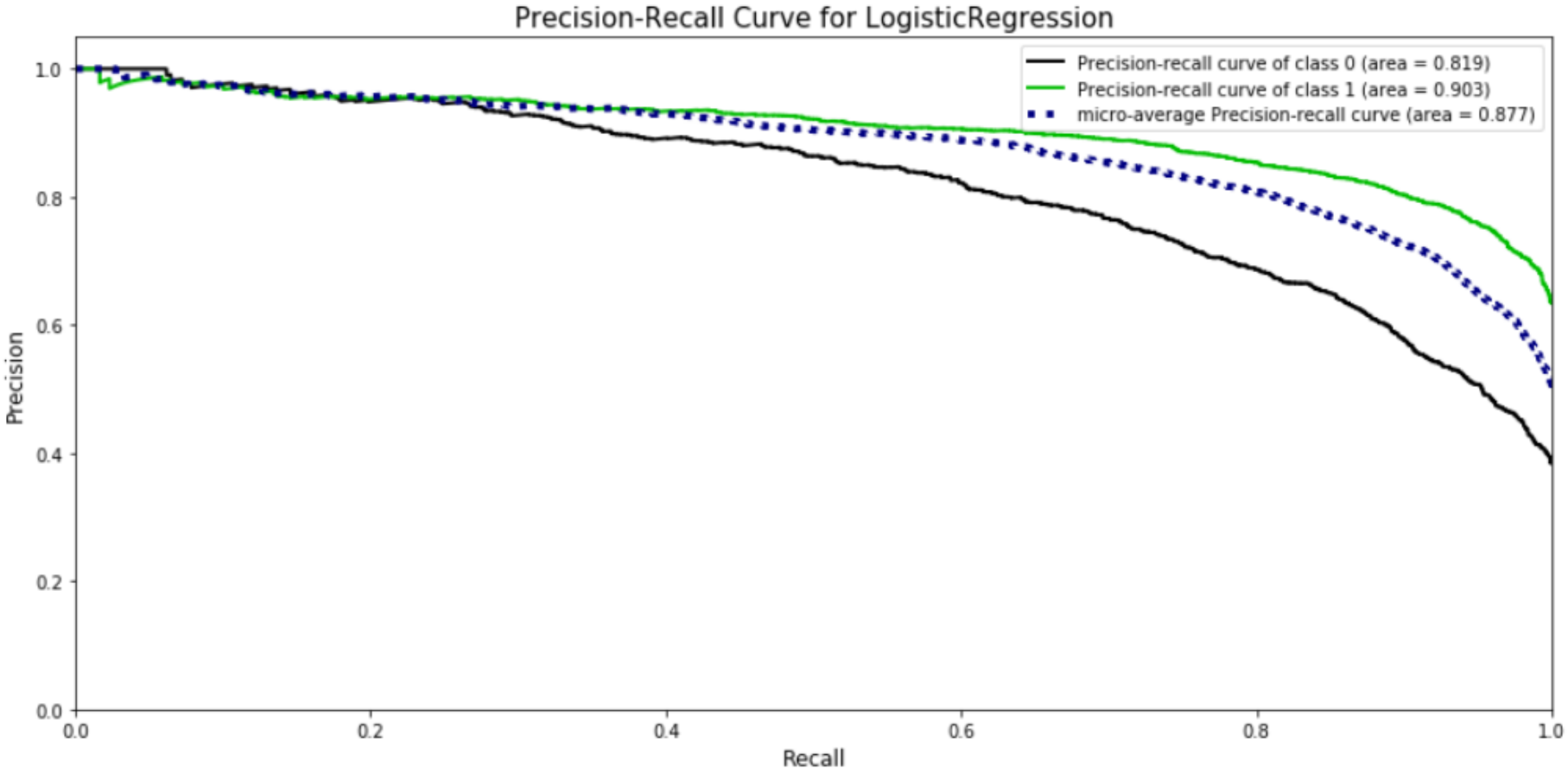
- *Logistic Regression – Post-Optimization*



YORK UNIVERSITÉ UNIVERSITY

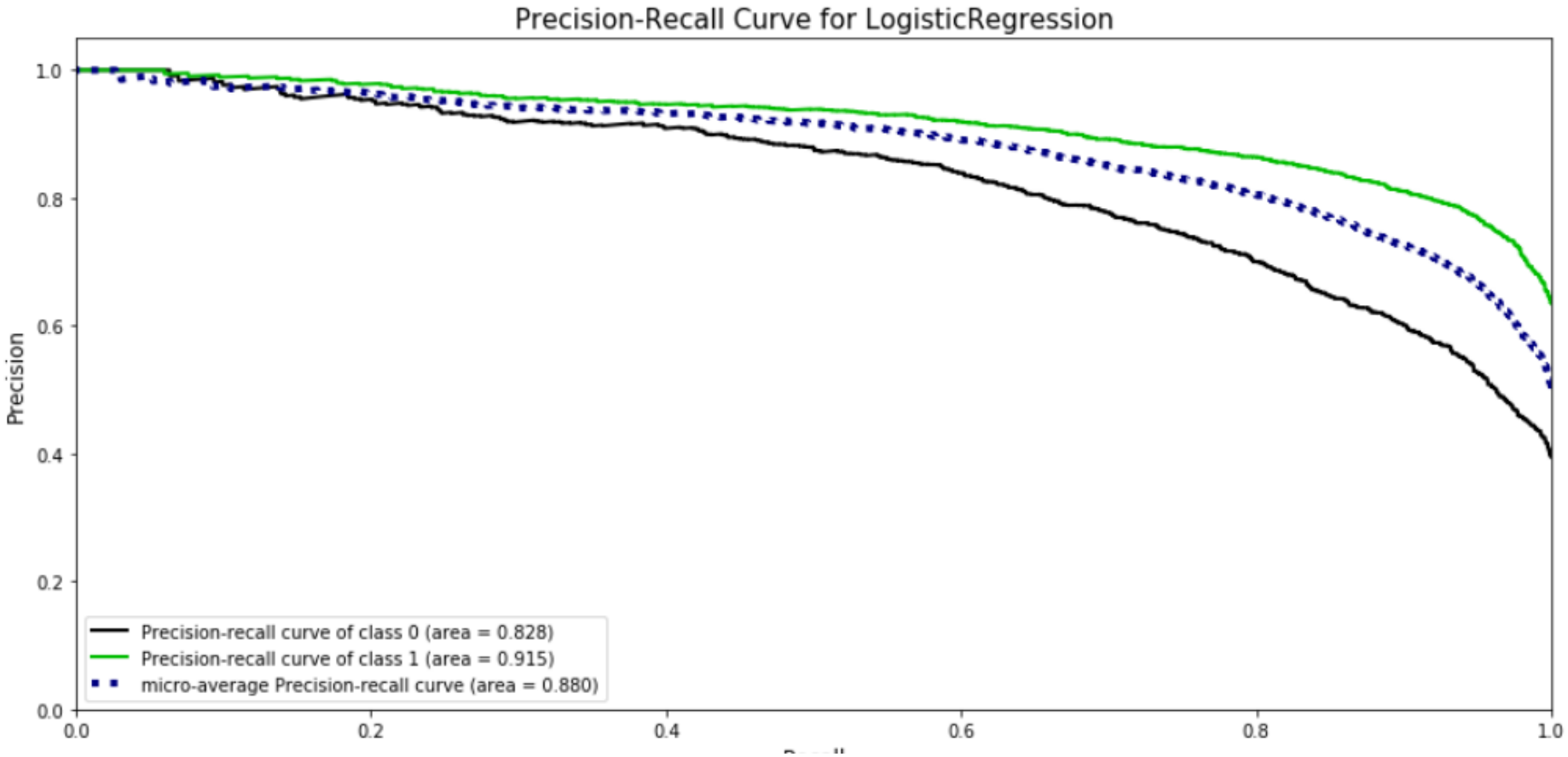
Hyperparameter Optimization

- *Logistic Regression – Pre-Optimization*



Hyperparameter Optimization

- *Logistic Regression – Post-Optimization*



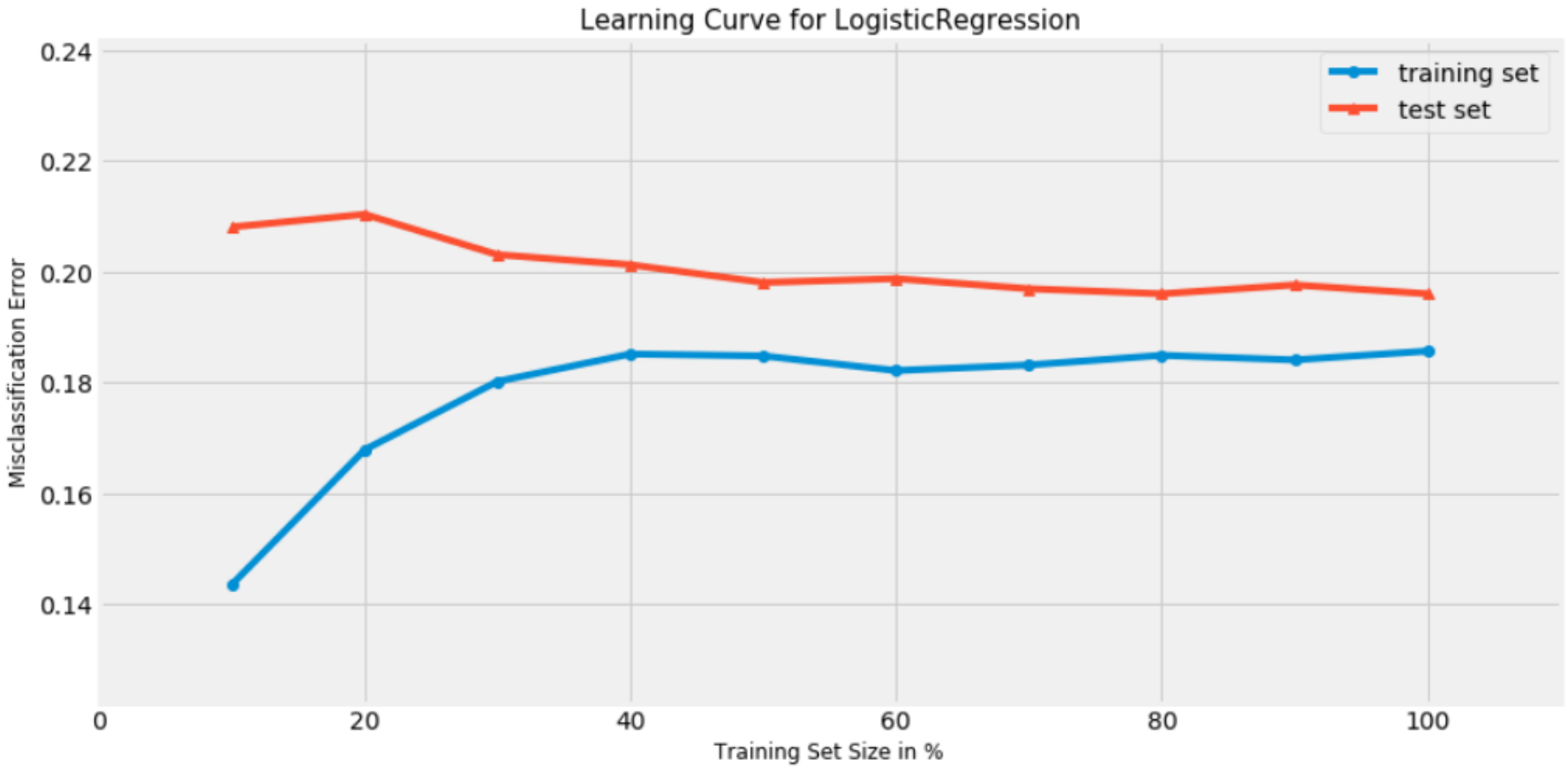
YORK UNIVERSITÉ UNIVERSITY

Hyperparameter Optimization

- *Logistic Regression – Pre-Optimization*

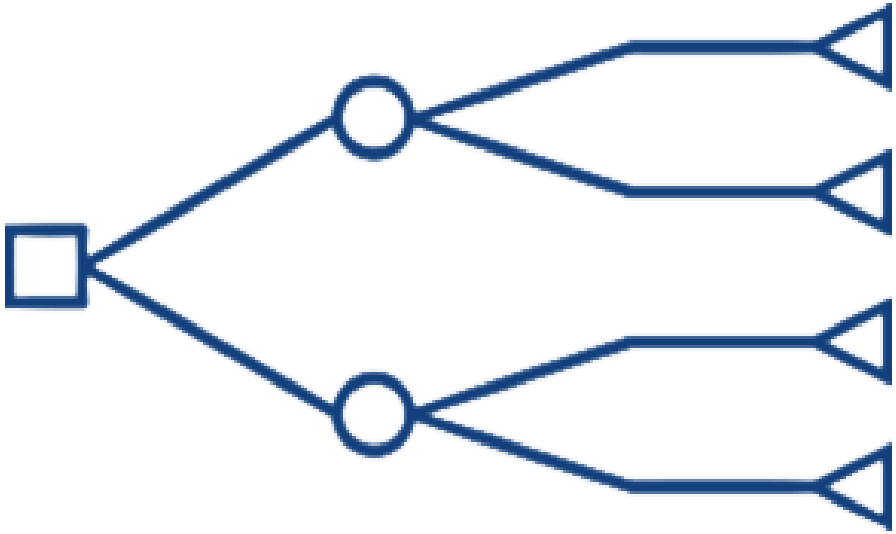
Hyperparameter Optimization

- *Logistic Regression – Post-Optimization*



Ensemble Methods

- **Boosting**
  - Gradient Boosting Machine
  - Ada Boost
- **Bagging**
  - Random Forest
  - Bagging Classifier
- **Stacking**
  - Stacking Classifier (sci-kit learn 0.22)



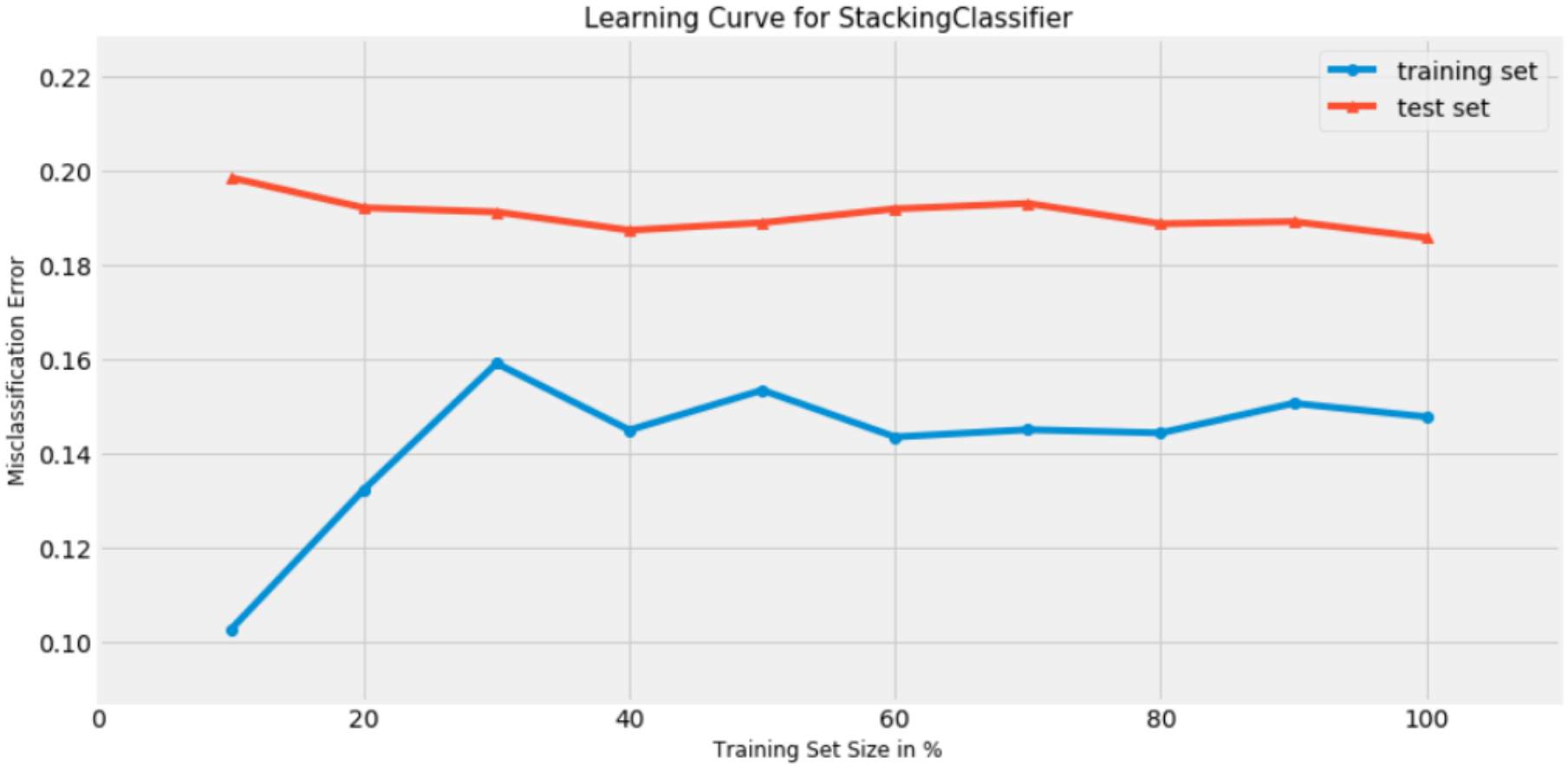
Model Selection

- **Using scores/metrics to select model**
  - F2 score
  - Fit time (as this may need to be deployed)
  - Learning Curve
  - Precision-Recall Curve

- **Narrowed down list**
  - *Logistic Regression*
    - Good fit time, good metrics - selected
  - *Gradient Boosting Machine*
    - Fit time and parameters too complex - rejected
  - *Stacking*
    - Good combination of base learners, fit time acceptable - selected

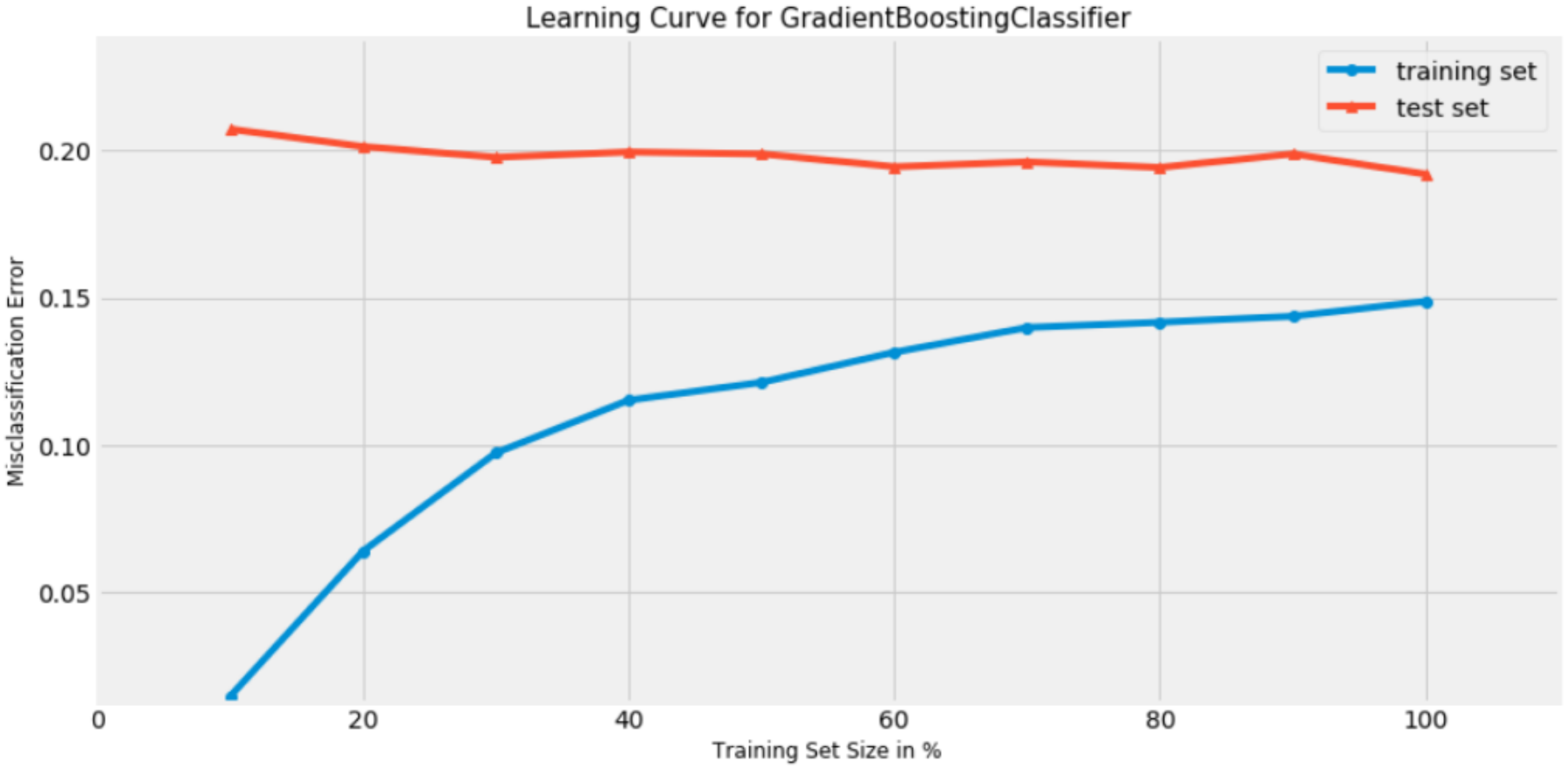
Model Selection – Learning Curve

- *Stacking Classifier*



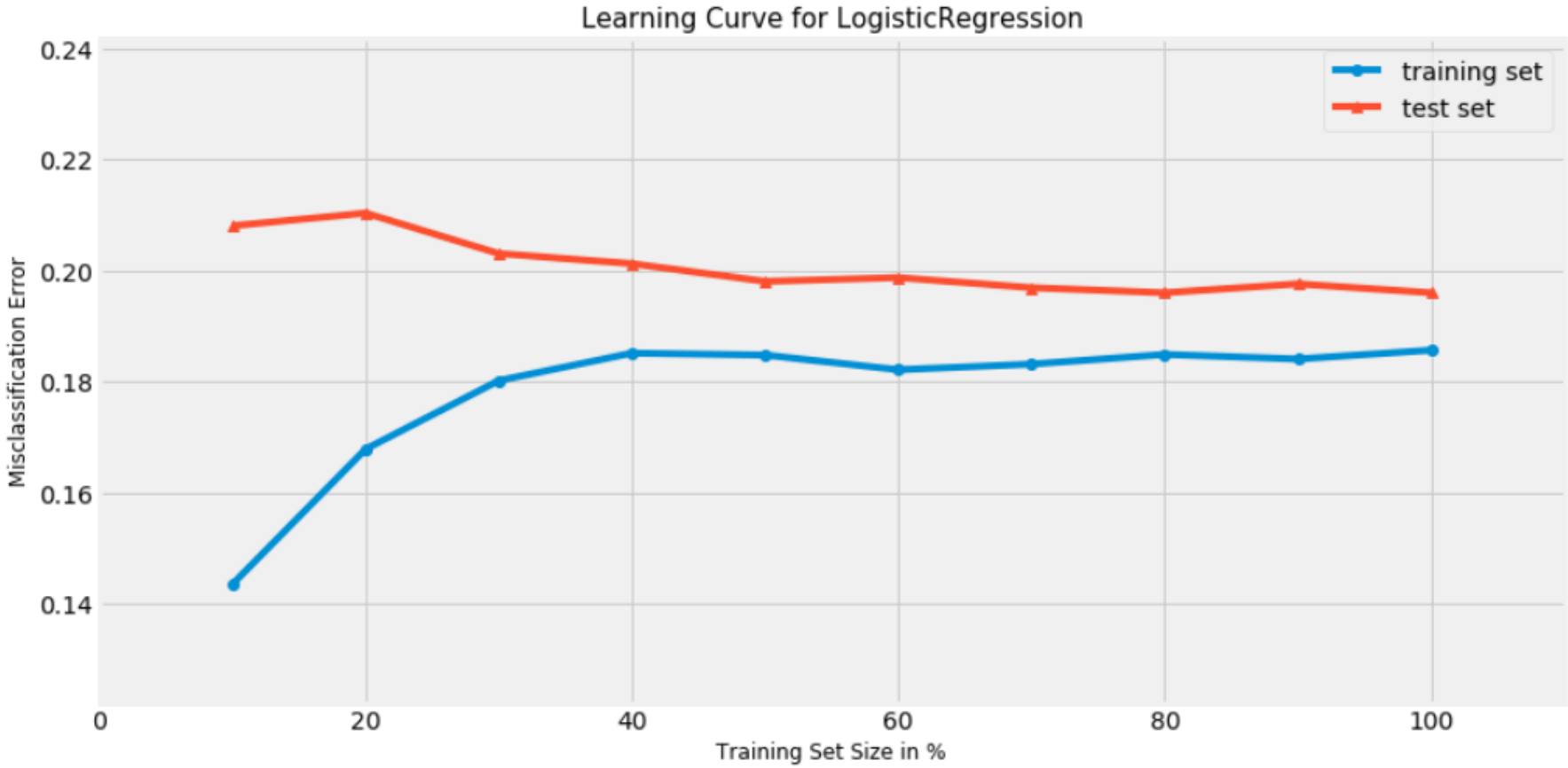
Model Selection – Learning Curve

- *Gradient Boosting Machine*



Model Selection – Learning Curve

- *Logistic Regression*




YORK
UNIVERSITÉ
UNIVERSITY
Thank You!