TWITTER HATE SPEECH ANALYSIS

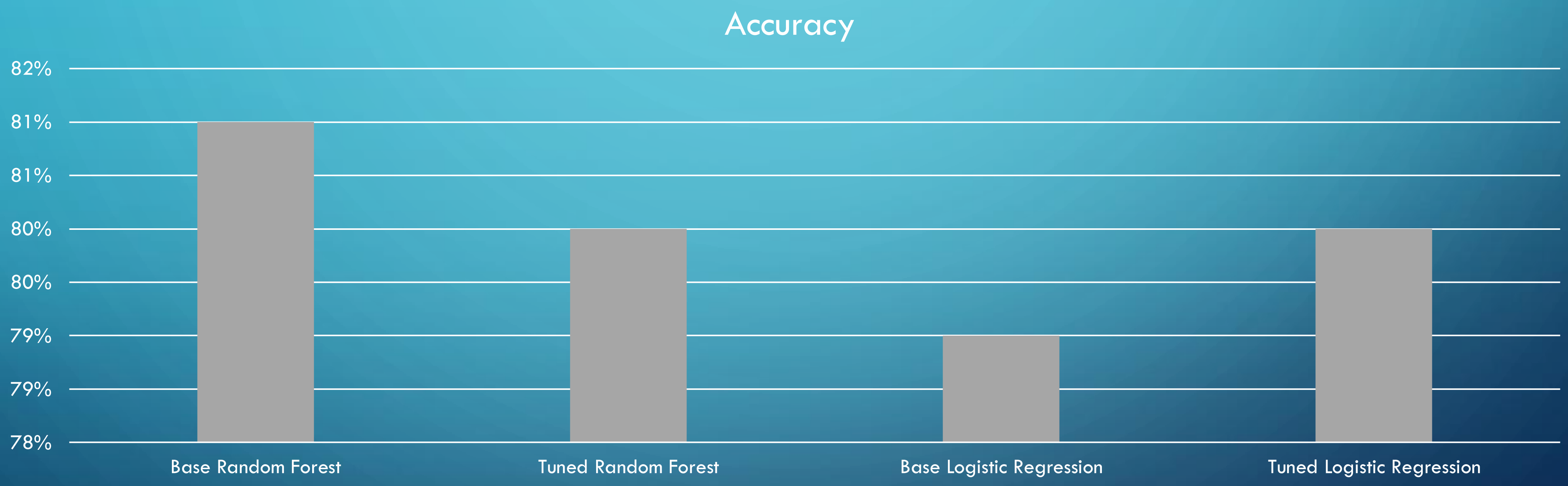
MILESTONE THREE

COLIN GREEN & SEAN ZHANG

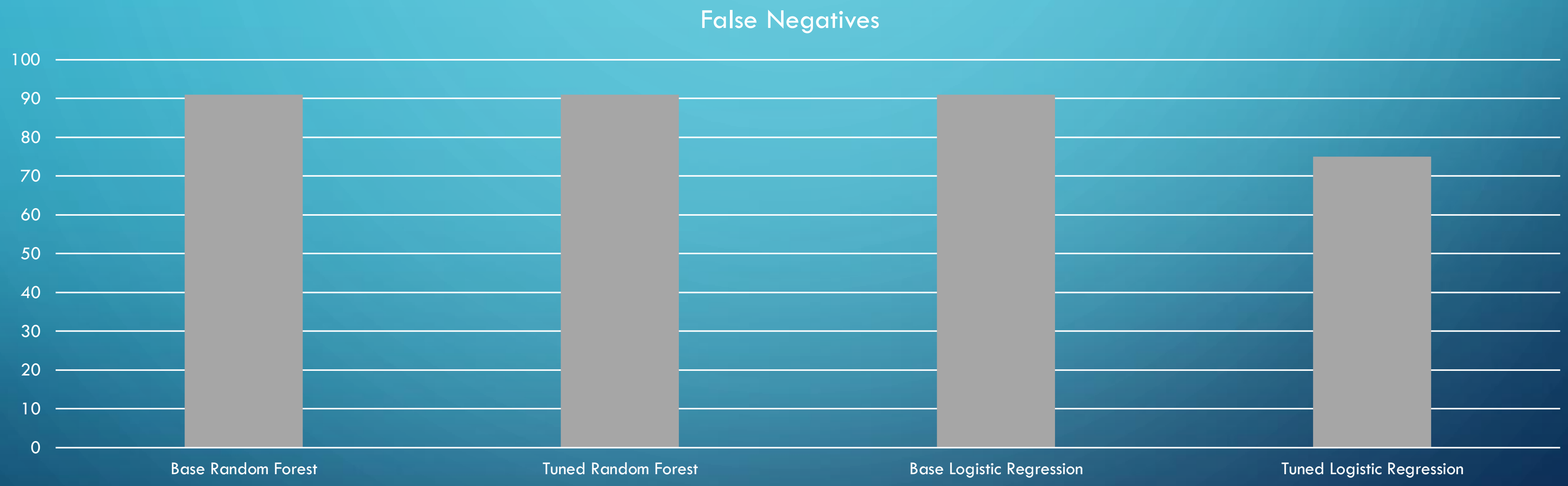
BALANCED DATASET

- Ran Random Forest and Logistic Regression as a benchmark
- Compared with tuned versions of Random Forest and Logistic Regression
- Used the bagging method for those 4 models as well as untuned:
  - Extra Trees, KNN, SVC, Ridge Classifiers
- Also used Ada Boost, Grad Boost, XG Boost and an Ensemble

LOGISTIC REGRESSION VS RANDOM FOREST

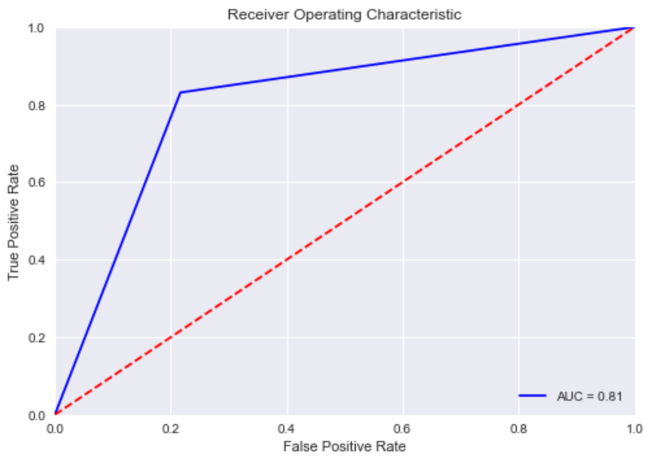


FALSE NEGATIVES

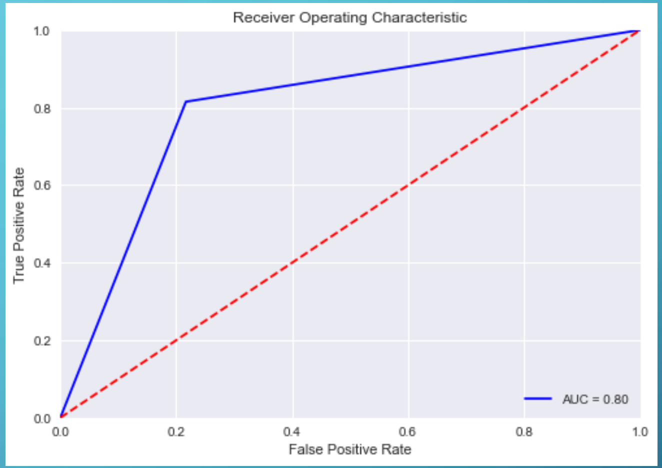


ROC CURVES

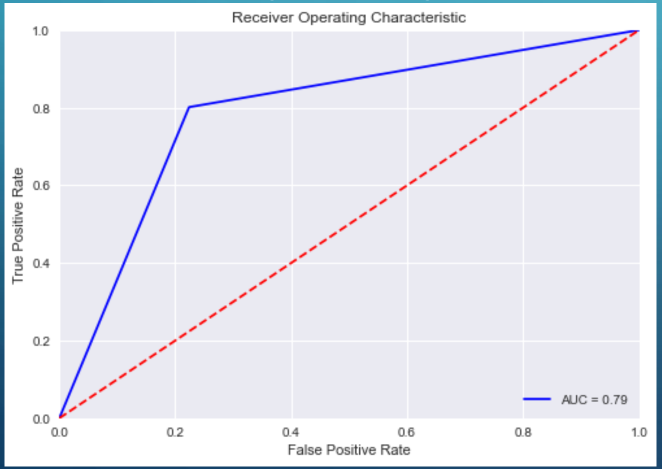
Base Random Forest



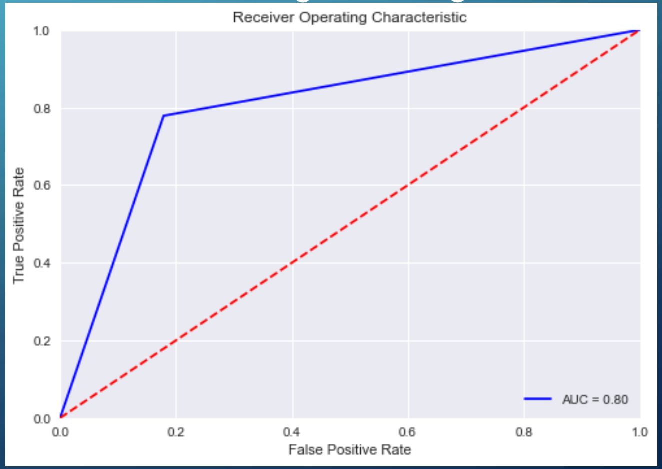
Tuned Random Forest



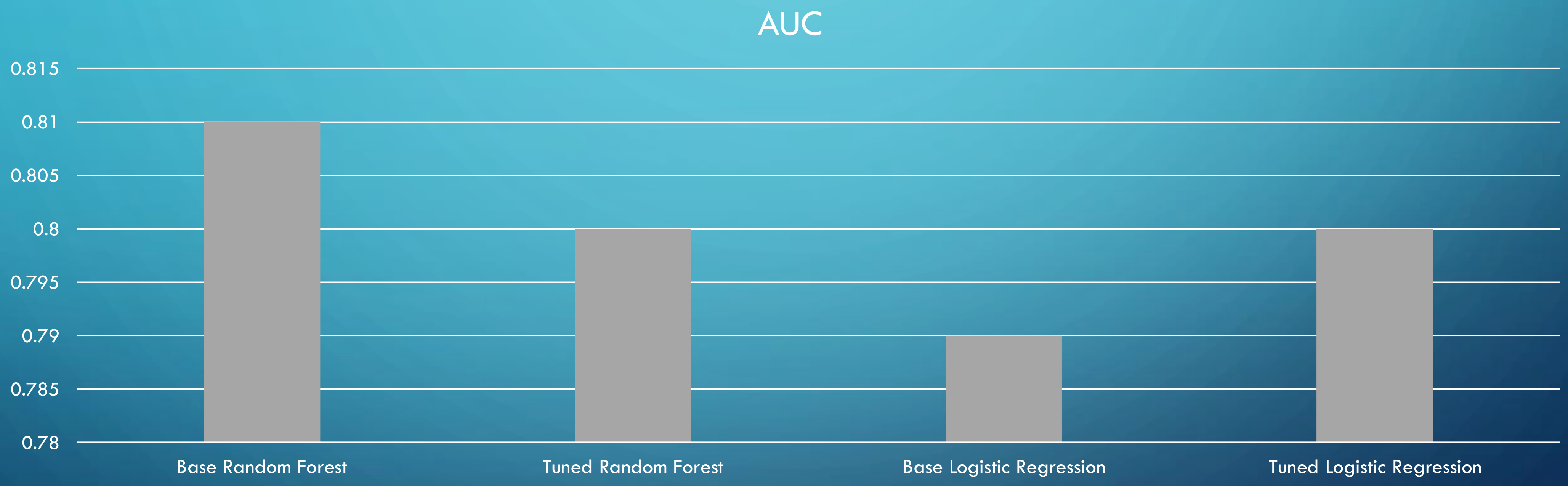
Based Logistic Regression



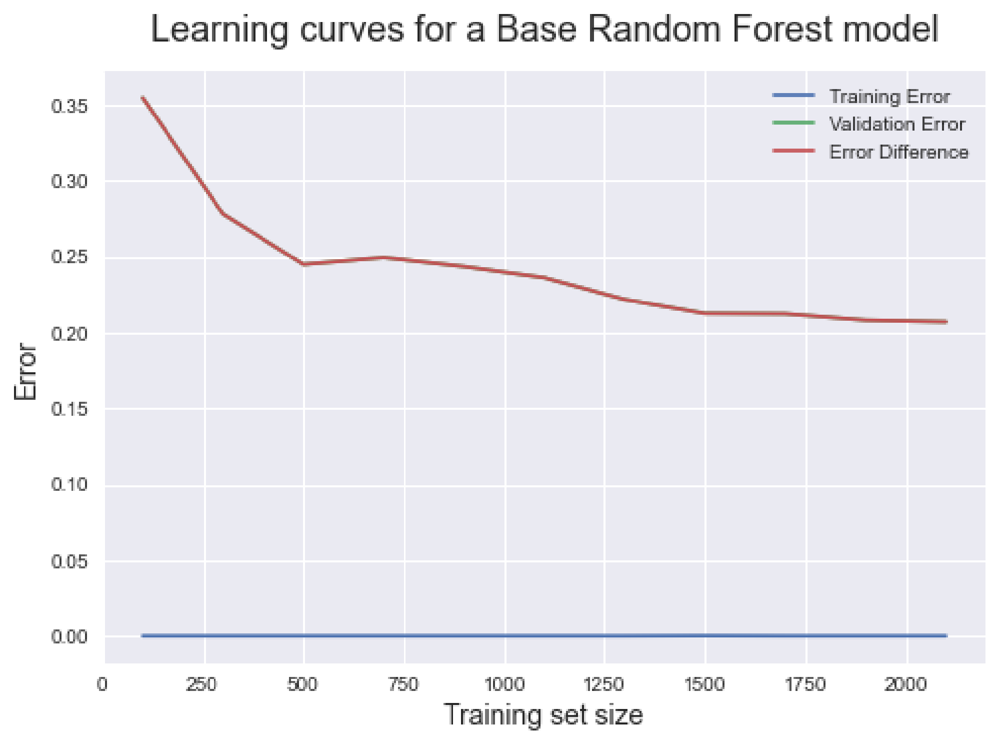
Tuned Logistic Regression



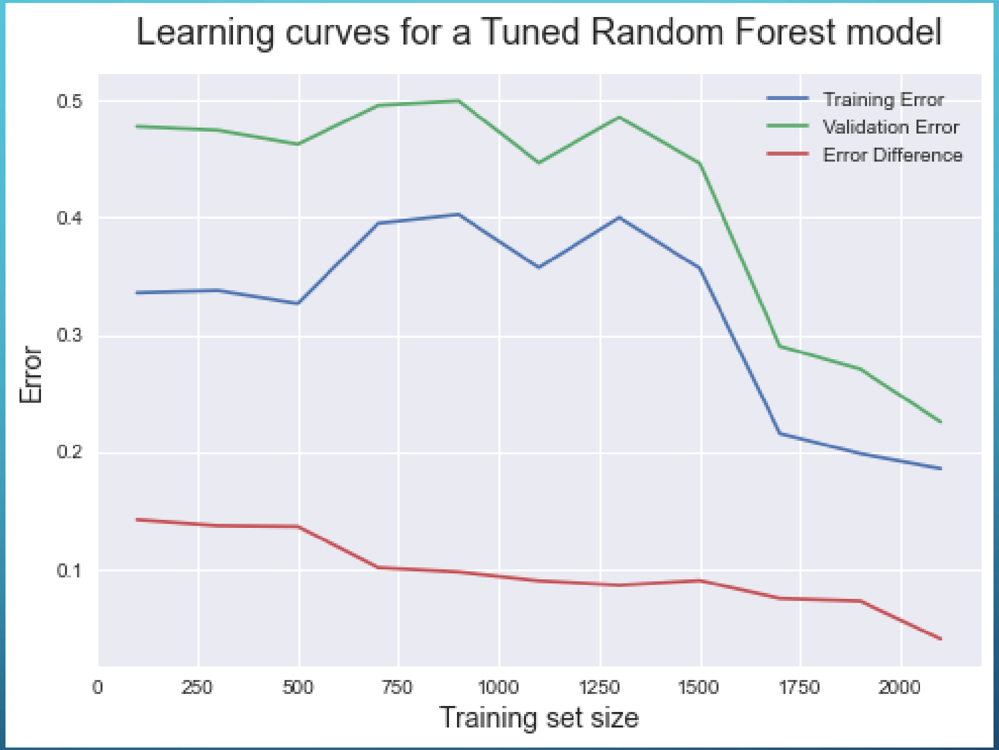
AREA UNDER THE CURVE



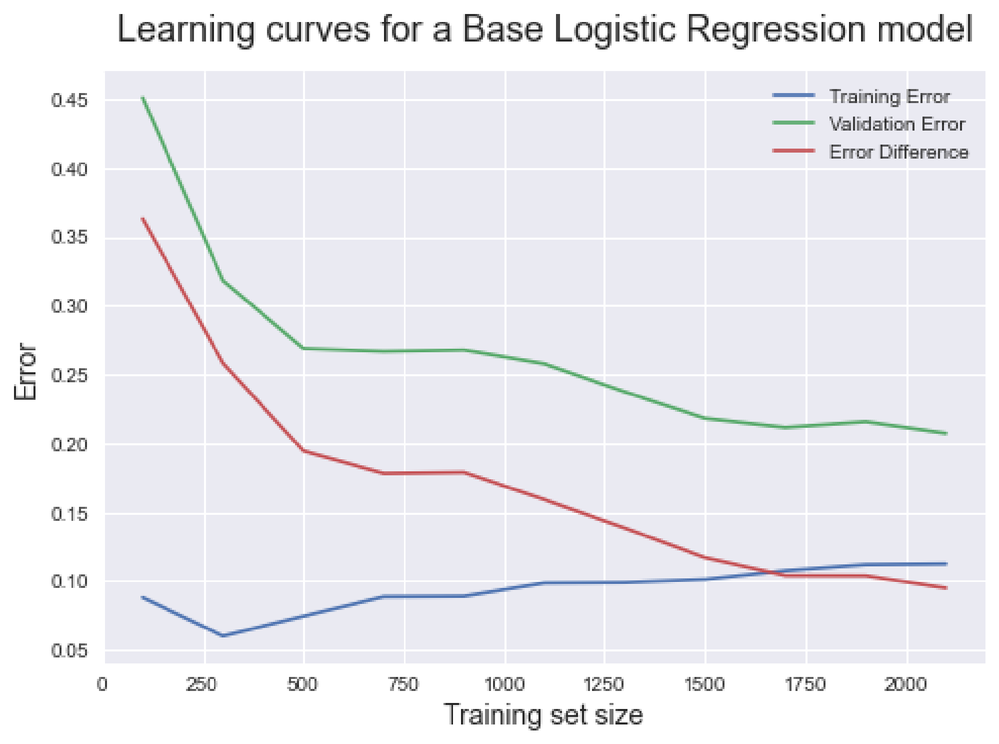
LEARNING CURVES – BASE RANDOM FOREST



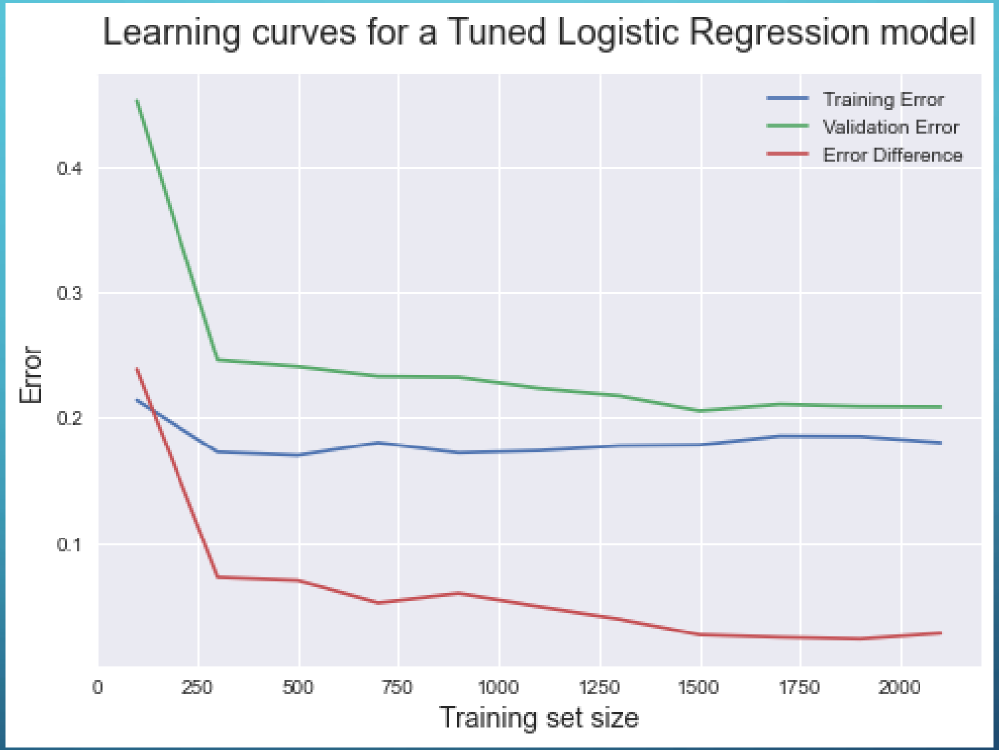
LEARNING CURVES – TUNED RANDOM FOREST



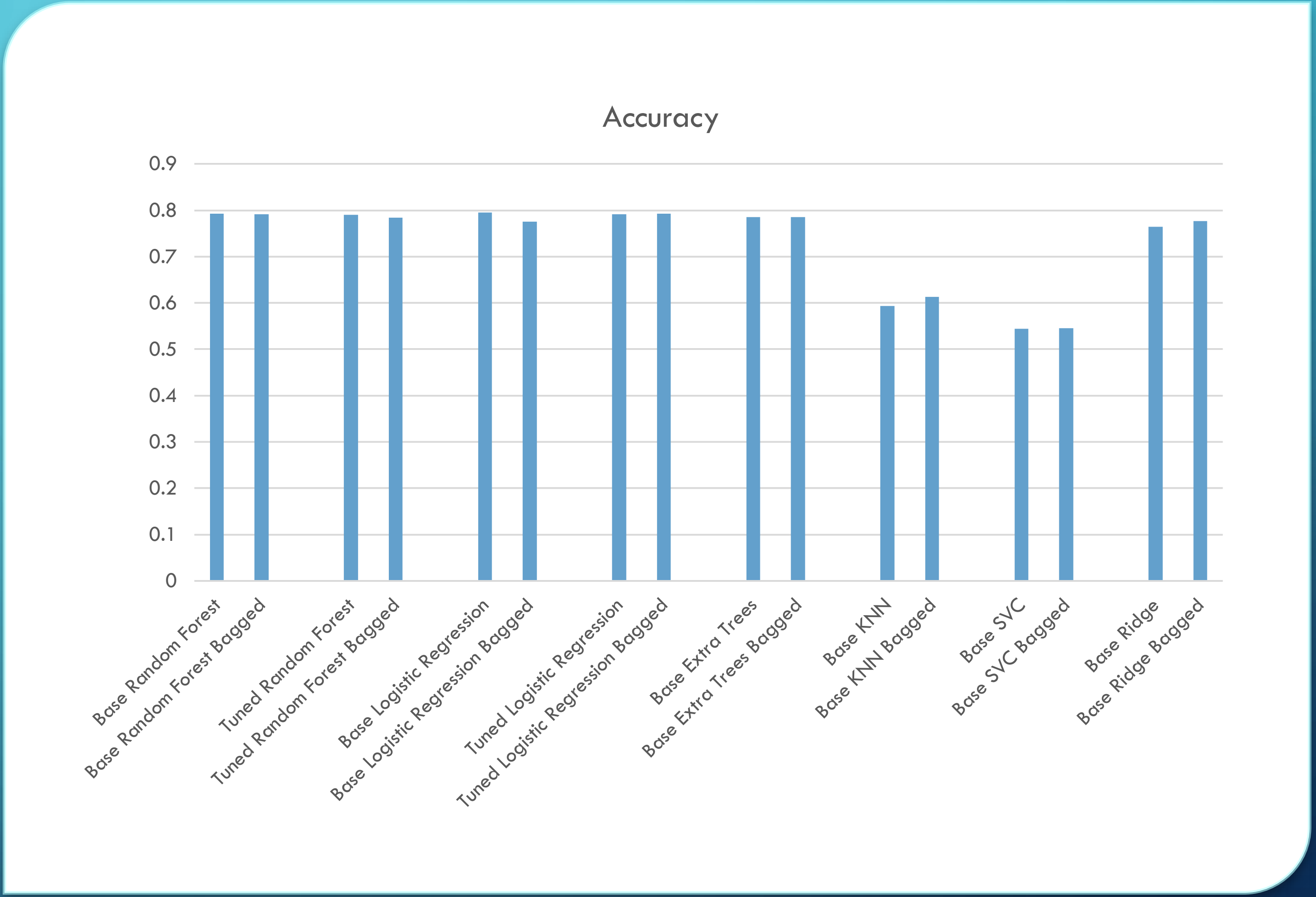
LEARNING CURVES – BASE LOGISTIC REGRESSION



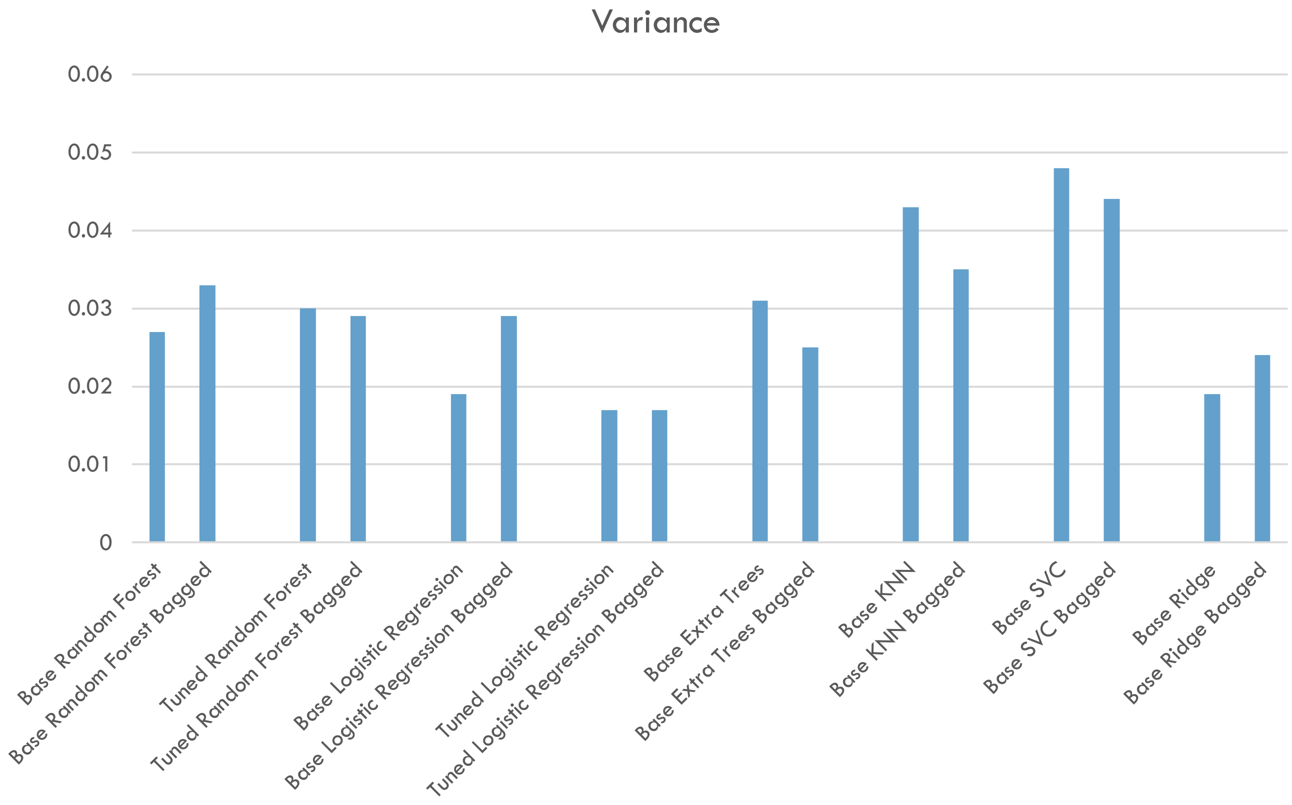
LEARNING CURVES – TUNED LOGISTIC REGRESSION



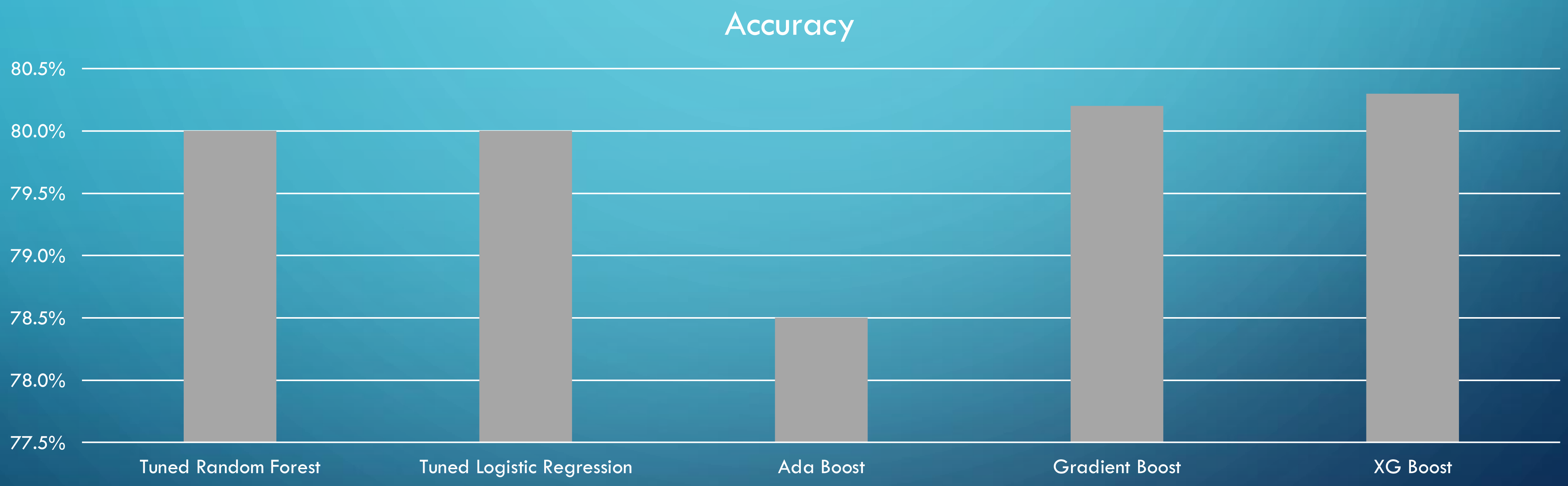
BAGGING


Accuracy
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Base Random Forest
Base Random Forest Bagged
Tuned Random Forest
Tuned Random Forest Bagged
Base Logistic Regression
Base Logistic Regression Bagged
Tuned Logistic Regression
Tuned Logistic Regression Bagged
Base Extra Trees
Base Extra Trees Bagged
Base KNN
Base KNN Bagged
Base SVC
Base SVC Bagged
Base Ridge
Base Ridge Bagged

BAGGING TO REDUCE VARIANCE



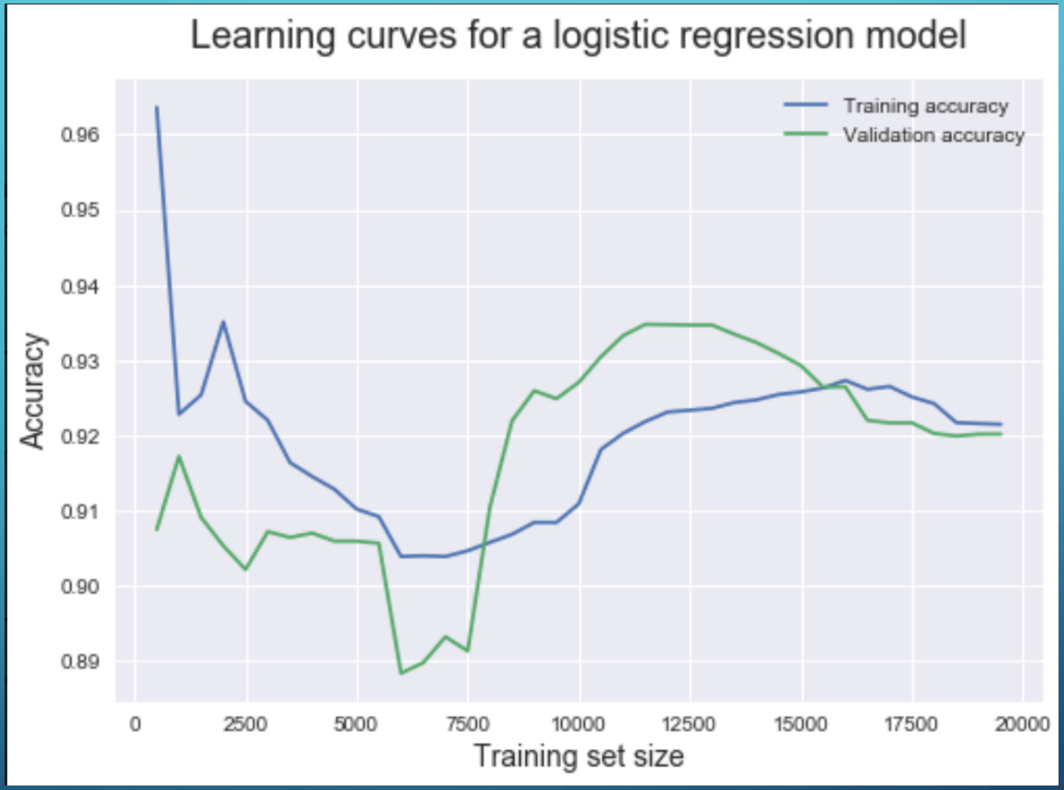
BOOSTING


Accuracy
80.5%
80.0%
79.5%
79.0%
78.5%
78.0%
77.5%
Tuned Random Forest
Tuned Logistic Regression
Ada Boost
Gradient Boost
XG Boost

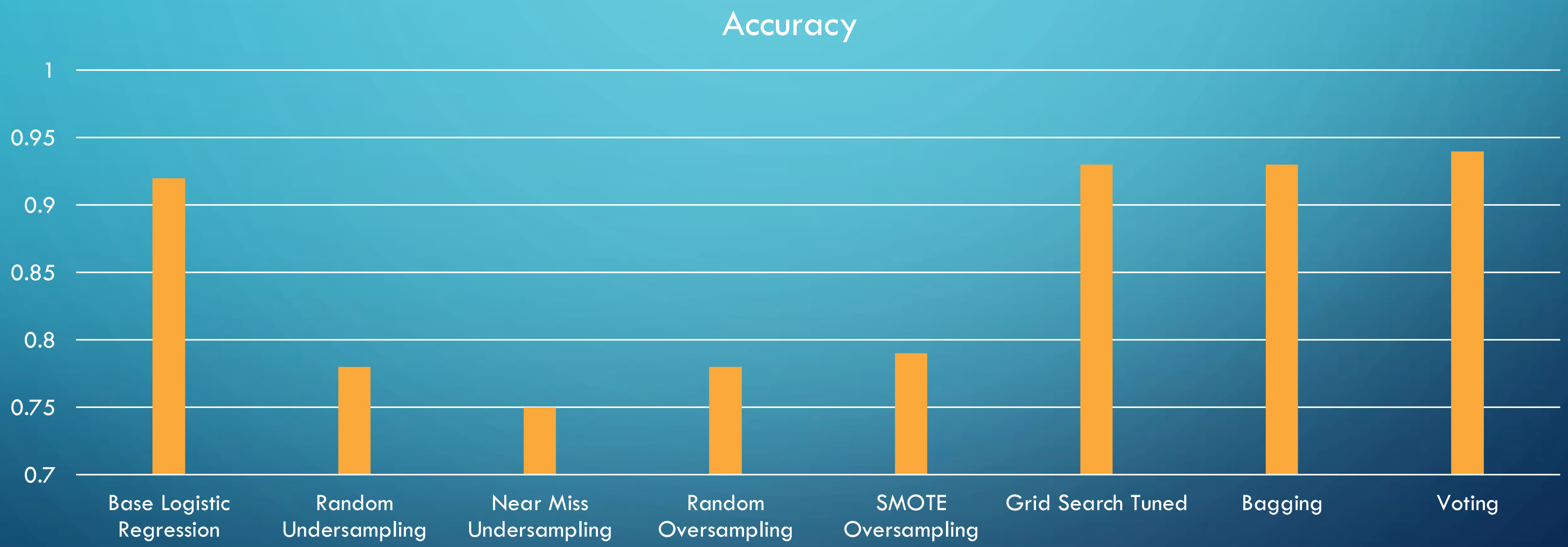
FULL DATASET

- Used Logistic Regression as Benchmark
- Ran to see whether learning curves would converge
- Comparison of: Base, under/oversampling, hyperparameter tuning, bagging, ensemble (voting)
- Decision boundary visualization

LEARNING CURVES – TUNED LOGISTIC REGRESSION (FULL DATA)



ACCURACY METRICS (FULL DATA)



Imbalanced sampling done with **imblearn** package

Voting ensemble: Random Forest, Extra Trees, KNN, Support Vector Machine, Logistic Regression

DECISION BOUNDARIES

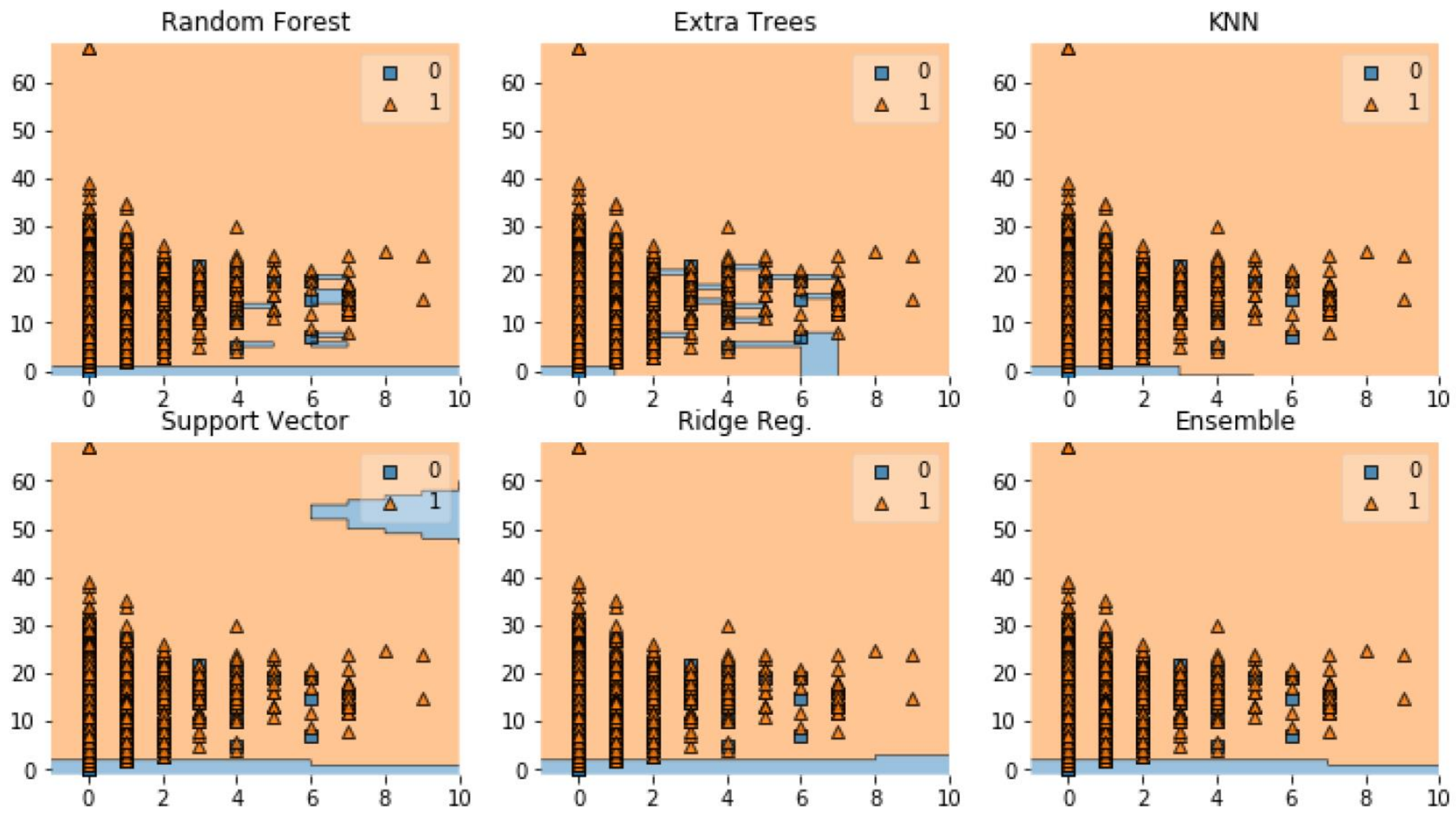
- Chose numerical variables based on importance and the fact that word features are too sparse
- Observations with lower number of tokens tends to be classified as hate speech (this was something noticed in previous EDA, feature importance, and past research)

X: hashtag count

Y: number of tokens

0: Hate speech

1: Non-hate speech



CONCLUSIONS

- Algorithm converged in full dataset compared to balanced subset
- Tuning slightly increased accuracy
- Bagging slightly increased accuracy and decreased variance in most cases
- Imbalanced sampling significantly decreased accuracy with no improvement in sensitivity or precision
- Ensemble learning slightly improved accuracy