CS613 Final Project

Connor Secen, Richard Strouss, Tristan Marshall

PROBLEM STATEMENT

- COMPARE THE RESULTS OF MULTIPLE METHODS ON PREDICTING THE DISCRETIZED CRIME RATE OF VARIOUS LOCALITIES GIVEN A RELEVANT FEATURE SET
- TO INCREASE PRACTICAL APPLICABILITY, DIMENSIONALITY WILL BE REDUCED WHILE MAINTAINING ACCEPTABLE VALIDATION SCORES

PRIOR WORK

- Ingilevich, V. & Ivanov, S. (2018). Crime Rate Prediction in the Urban Environment Using Social Factors. *Procedia Computer Science, 136,* 472-478. https://doi.org/10.1016/j.procs.2018.08.261
  - Compared results from linear regression, logistic regression, and gradient boosted decision trees
  - Found gradient boosting to be the most appropriate technique of the three
- Alves, L. G. A., Ribiero, H. V., & Rodrigues, F. A. (2018). Crime Prediction Through Urban Metrics and Statistical Learning. *Physica A, 55,* 435-443. https://doi.org/10.1016/j.physa.2018.03.084
  - Used a random forest model
  - Out performed previous linear models
- Kshatri, S. S., Singh, D., Narain, B., Bhatia, S., Quasim, M. T., & Sinha, G. R. (2021). An Empirical Analysis of Machine Learning Algorithms for Crime Prediction using Stacked Generalization: An Ensemble Approach. *IEEE Access.* Advanced online publication. https://doi.org/10.1109/ACCESS.2021.3075140
  - Compared J48 decision tree algorithm, random forests, sequential minimal optimization (SVM), bagging, and naïve bayes classifiers
  - Best results from stacking the models
  - Random forest was nearly as good as the stacked model on all validation metrics

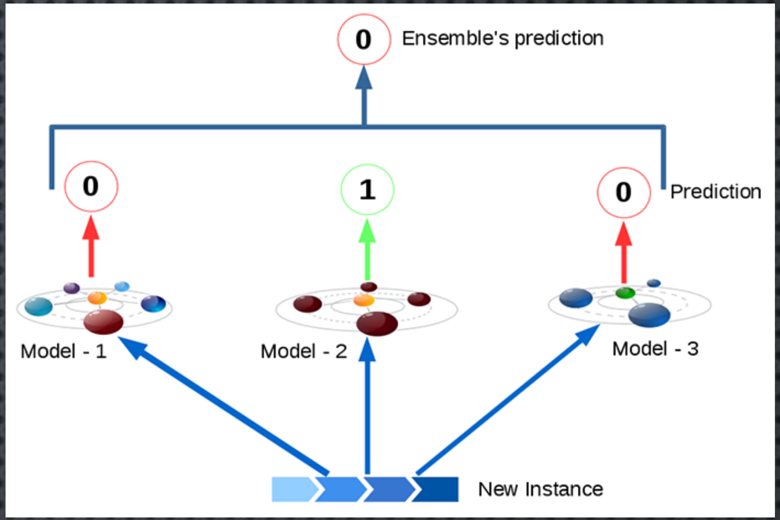
PRIOR WORK

- Zhu, J., S. Rosset, H. Zou, and T. Hastie. 2009. Multi-Class AdaBoost.Statistics and its Interface2 (3):349–360.
  - Multi-class Adaboost extension
  - Fixes problem with high multi-class error
- General lessons
  - Features are important
  - Subject similarity does not guarantee learning performance similarity

DATASET

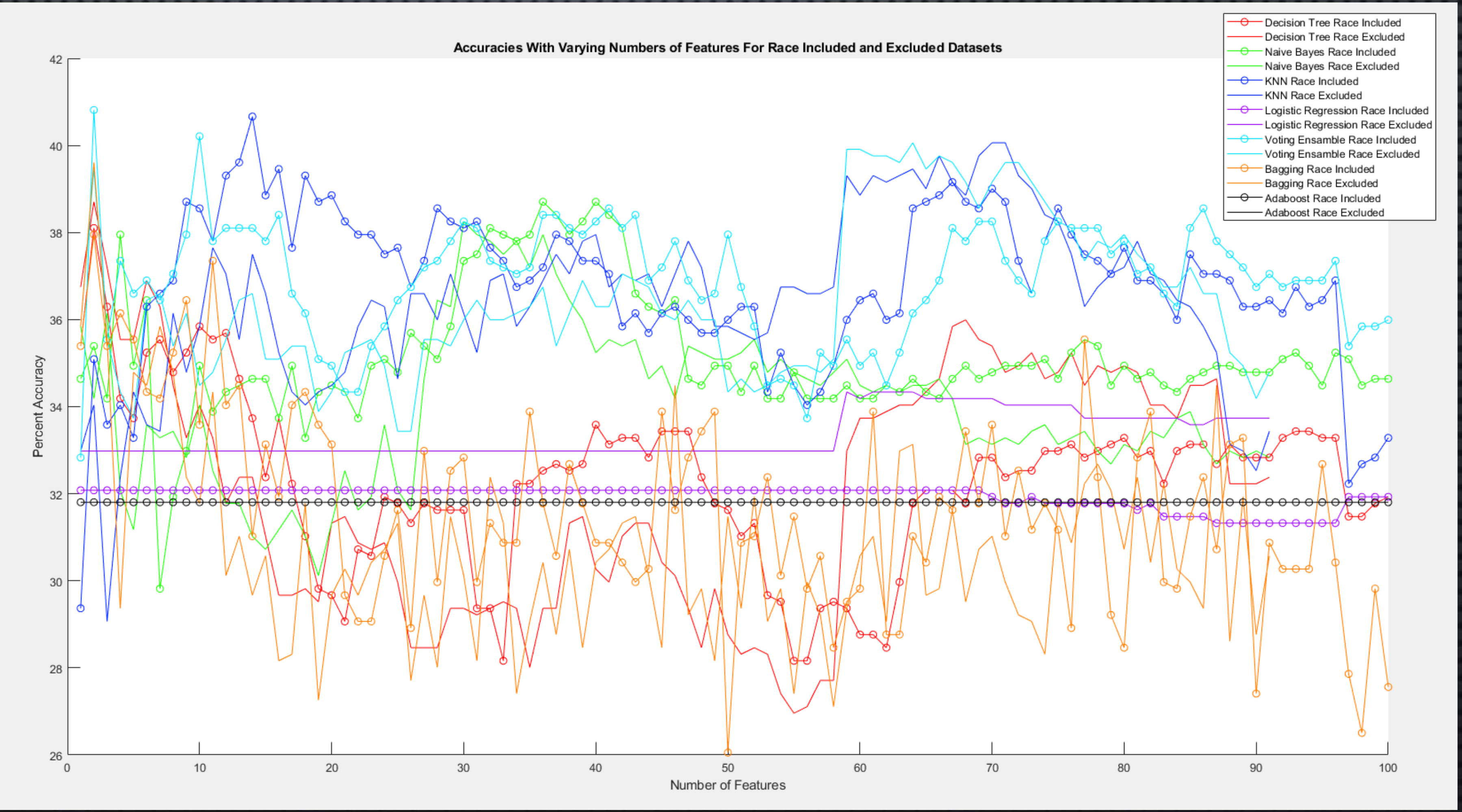
- Communities and Crimes Dataset from the UC Irvine machine learning repository
  - http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime
- Combines socio-economic data from 1990 US Census, law enforcement data from 1990 US LEMAS (*Law Enforcement Management and Administrative Statistics*) survey, and crime data from 1995 FBI UCR (*Uniform Crime Reporting*)
- Dataset contains 1994 instances and 128 Attributes
  - Mix of categorical and continuous features
- 102 attribute With removal of missing data including
- Converted the target value to categorical by binning the data into 11 different groups separated by .1

APPROACH



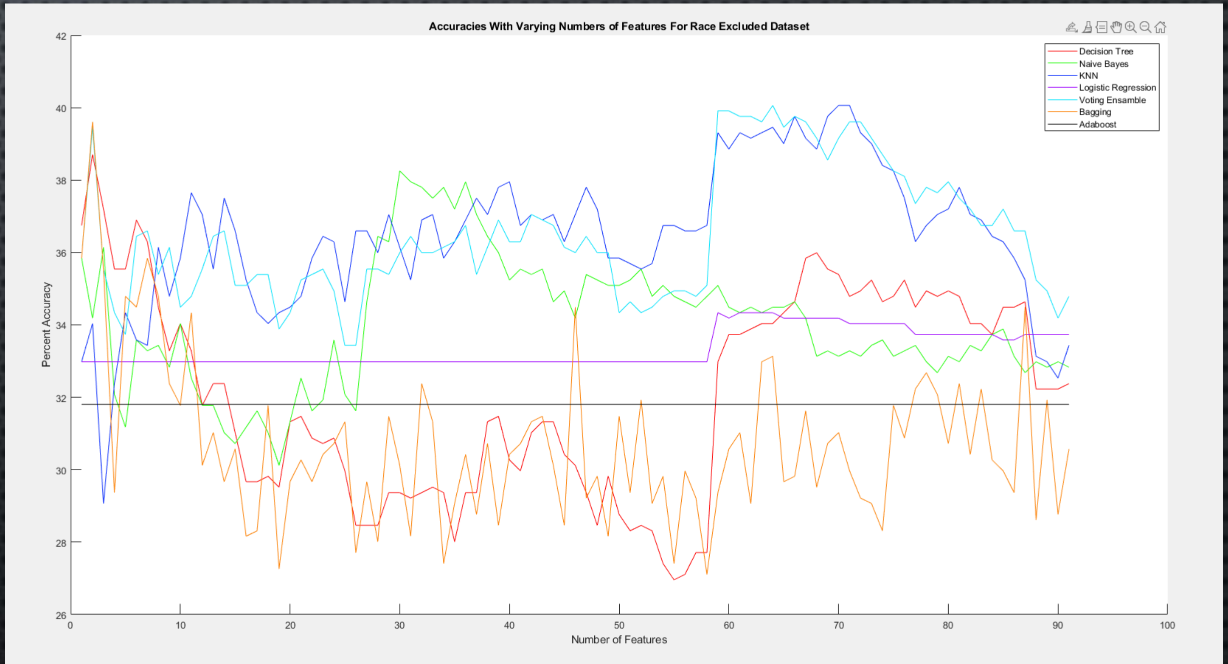
- Found the correlation between each feature and the target value
  - Split the dataset into one, one containing and one not containing race data
  - Sorted the feature from most correlated to least and progressively added the features to the dataset currently being sorted
- Compare simple classification methods
  - Logistic regression vs decision tree
- Compare more complex Ensemble methods
  - Voting
    - Logistic regression
    - ID3 Decision tree
    - KNN
    - Naïve bayes
  - Bagging
  - Boosting
    - Beta value for weight update calculate as follows: BETA = LOG(1-ERROR/ERROR)+LOG(K-1) where K is the class
- Determine which method works best for the given data by comparing the accuracy

RESULTS


Accuracies With Varying Numbers of Features For Race Included and Excluded Datasets
Percent Accuracy
Number of Features
42
40
38
36
34
32
30
28
26
0
10
20
30
40
50
60
70
80
90
100
Decision Tree Race Included
Decision Tree Race Excluded
Naive Bayes Race Included
Naive Bayes Race Excluded
KNN Race Included
KNN Race Excluded
Logistic Regression Race Included
Logistic Regression Race Excluded
Voting Ensemble Race Included
Voting Ensemble Race Excluded
Bagging Race Included
Bagging Race Excluded
Adaboost Race Included
Adaboost Race Excluded

OBSERVATIONS AND DISCUSSION

- Maximum Accuracy 40.06%
- Best Classifiers Voting Ensemble and KNN
- Race Data did not add significantly to accuracy and in many cases decreased Accuracy
- Logistic Regression and Adaboost had consistent accuracies, but adhered to specific classes



Observations and Discussion

- Low accuracy potentially a result of the data itself
- Dataset includes:
  - census data which has a 10% margin of error
  - Categories that should not hold non-zero data but do
  - Potentially subjective categories

FUTURE WORK/EXTENSIONS

- ATTEMPT TO USE MORE COMPLEX CLASSIFIERS
  - ARTIFICIAL NEURAL NETWORKS
- CHANGE THE LEARNING TASK TO BE A REGRESSION STYLE TASK
  - THE TARGET IN THE DATASET IS CONTINUOUS LENDING ITSELF TO REGRESSION
- COLLECT MORE DATA
  - MORE EXAMPLES TO REDUCE OVERFITTING
  - DIFFERENT FEATURES TO AVOID BIAS
  - REDUCE THE NUMBER OF FEATURES BY COLLECTING FEATURES THAT ARE BETTER PREDICTORS

THANK YOU!