PARTH MANISH PATEL

ISHAN HANDA

USING MACHINE LEARNING TO PREDICT SECONDARY SCHOOL STUDENT ALCOHOL CONSUMPTION

MOTIVATION

Prevalence of alcohol consumption amongst students: [1]

- *Prevalence of Drinking*: In 2013, 59.4 percent of full-time college students' ages 18–22 drank alcohol in the past month compared with 50.6 percent of other persons of the same age. [2]
- *Prevalence of Binge Drinking*: In 2013, 39 percent of college students ages 18–22 engaged in binge drinking (5 or more drinks on an occasion) in the past month compared with 33.4 percent of other persons of the same age. [3]
- *Prevalence of Heavy Drinking*: In 2013, 12.7 percent of college students ages 18–22 engaged in heavy drinking (5 or more drinks on an occasion on 5 or more occasions per month) in the past month compared with 9.3 percent of other persons of the same age.

MOTIVATION

Consequences of alcohol consumption amongst students: [5]

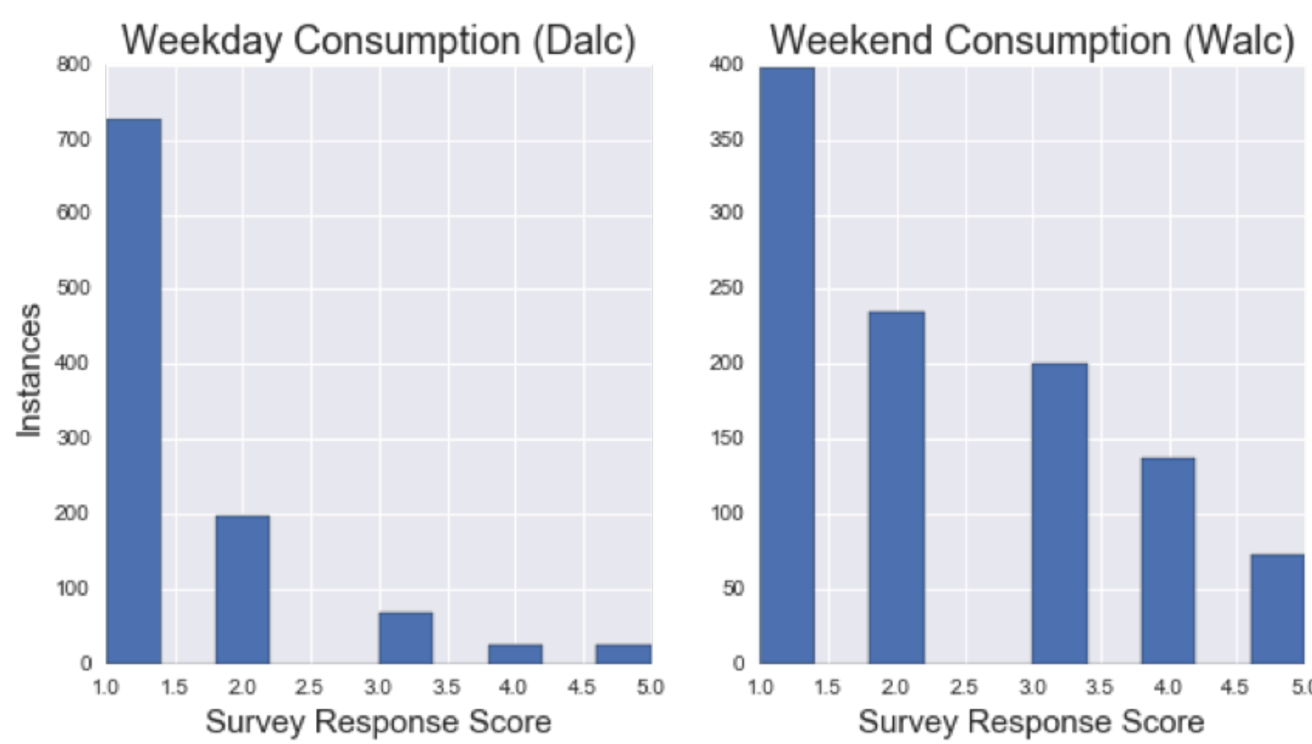
- 1,825 college students between the ages of 18 and 24 die from alcohol-related unintentional injuries, including motor vehicle crashes. [6]
- Another student who has been drinking assaults 696,000 students between the ages of 18 and 24. [7]
- Roughly 20 percent of college students meet the criteria for an AUD. [8]
- About 1 in 4 college students report academic consequences from drinking, including missing class, falling behind in class, doing poorly on exams or papers, and receiving lower grades overall. [9]

DATA-SET

- Source: Using Machine Learning To Predict Secondary School Student Alcohol Consumption(https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION)
- Collected from questionnaires responses for students spread across 2 public secondary schools in Portugal during the 2005-2006 school year
- Features:
    - Binary: gender, school attended, urban vs. rural address, internet access etc.
    - Discrete: gender, school attended, urban vs. rural address, internet access etc.
    - Weekend and Weekday alcohol consumption: Rated on a scale of 1 to 5.

PREPARATION OF DATA-SET

- Binary features(eg. 'U' vs. 'R' for urban / rural) converted to 0-1 values.
- Target variables Walc(weekend) and Dalc(weekday) converted to binary.
    - Walc > 3
    - Dalc > 1

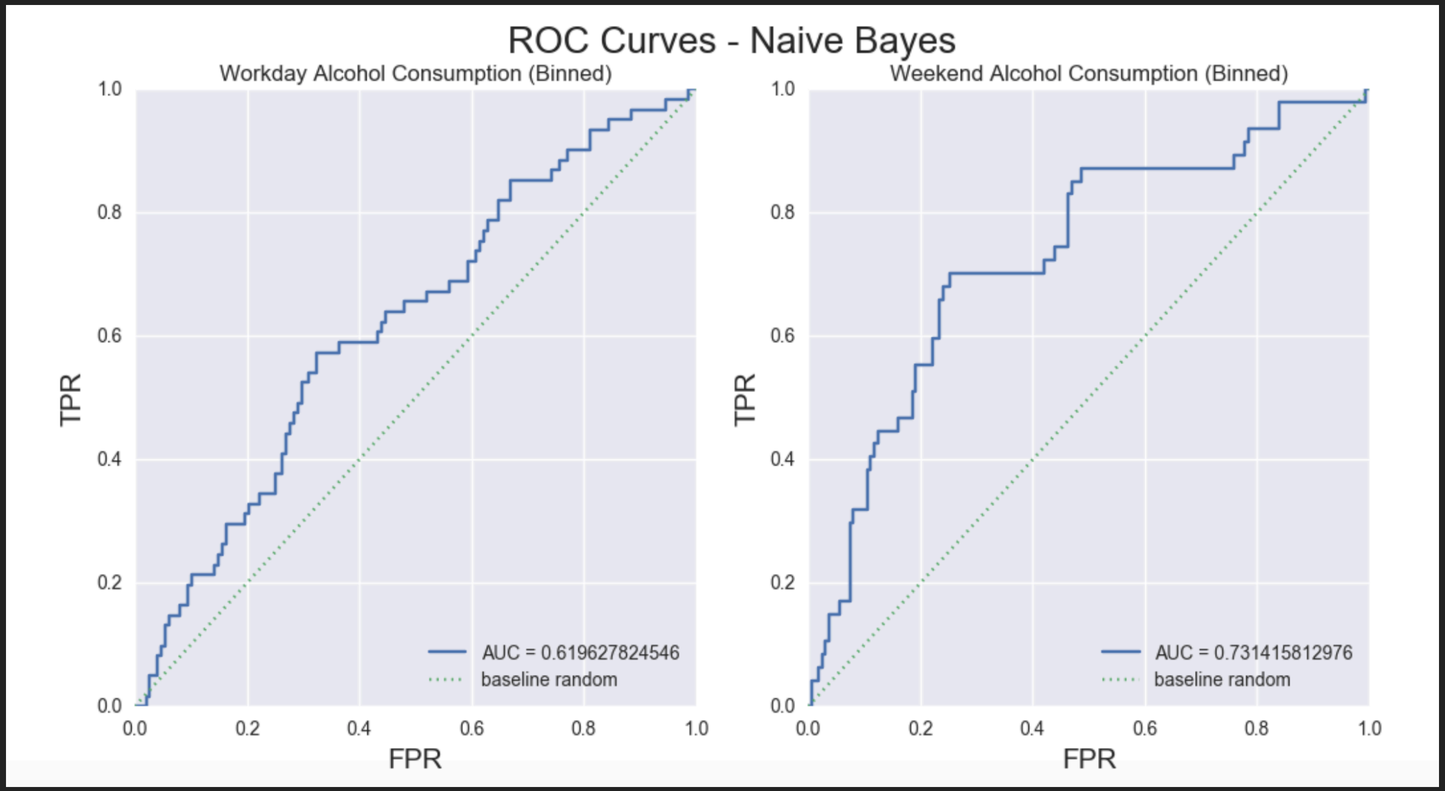


MODELS

- Naive Bayes
- Logistic Regression
- Random Forrest

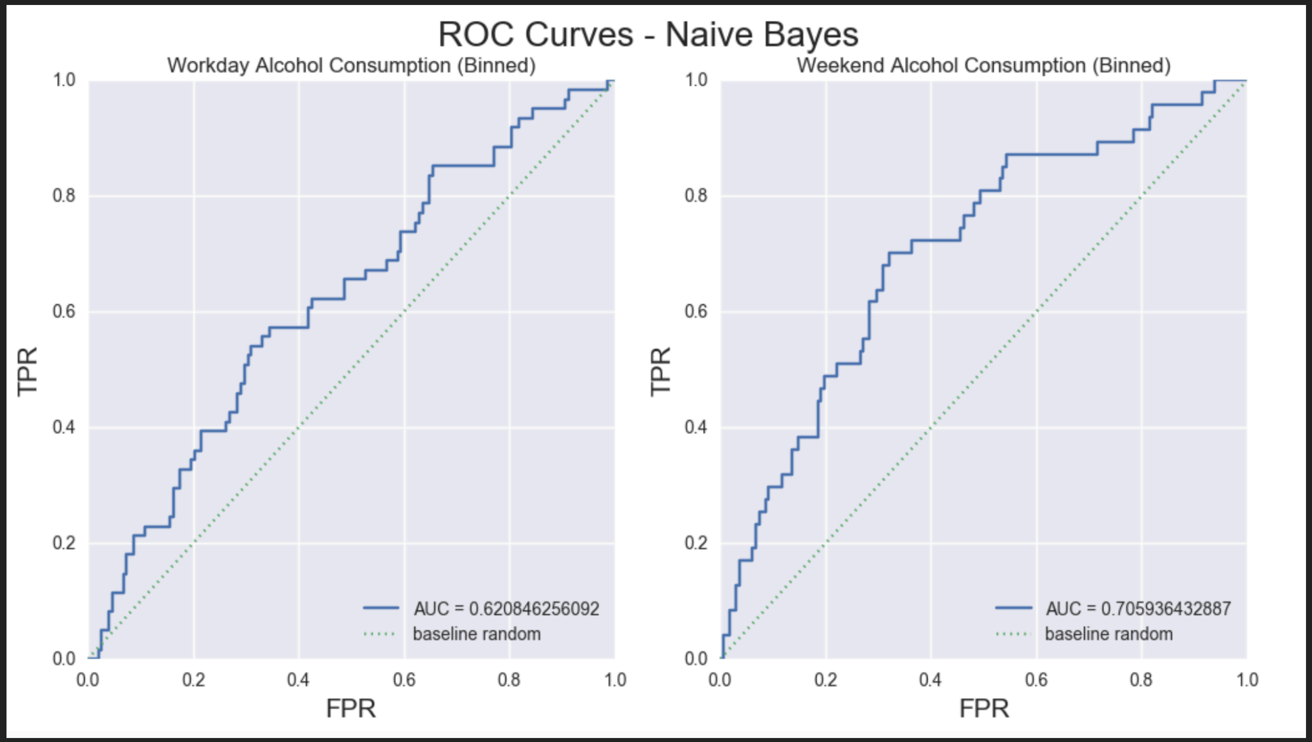
NAIVE BAYES

- Trained using 80/20 train test split for each target variable.
- Using scikit-learn BernoulliNB function
- Default parameter Scores:
    1. Dalc Accuracy: 0.650717703349
    2. Dalc Recall: 0.327868852459
    3. Dalc Precision: 0.384615384615
    4. Dalc F Score: **0.353982300885**
    5. Walc Accuracy: 0.77033492823
    6. Walc Recall: 0.148936170213
    7. Walc Precision: 0.466666666667
    8. Walc F Score : **0.225806451613**



NAIVE BAYES

- With alpha = 5 smoothening for Dalc and alpha = 7 for Walc
- Scores:
  1. Dalc Accuracy: 0.617224880383
  2. Dalc Recall: 0.573770491803
  3. Dalc Precision: 0.393258426966
  4. Dalc F Score: **0.466666666667**
  5. Walc Accuracy: 0.684210526316
  6. Walc Recall: 0.702127659574
  7. Walc Precision: 0.388235294118
  8. Walc F Score : **0.5**



CONSIDERING EVALUATION METRICS

- Precision

$$\text{Precision} = \frac{tp}{tp + fp}$$

- Recall

$$\text{Recall} = \frac{tp}{tp + fn}$$

- Accuracy

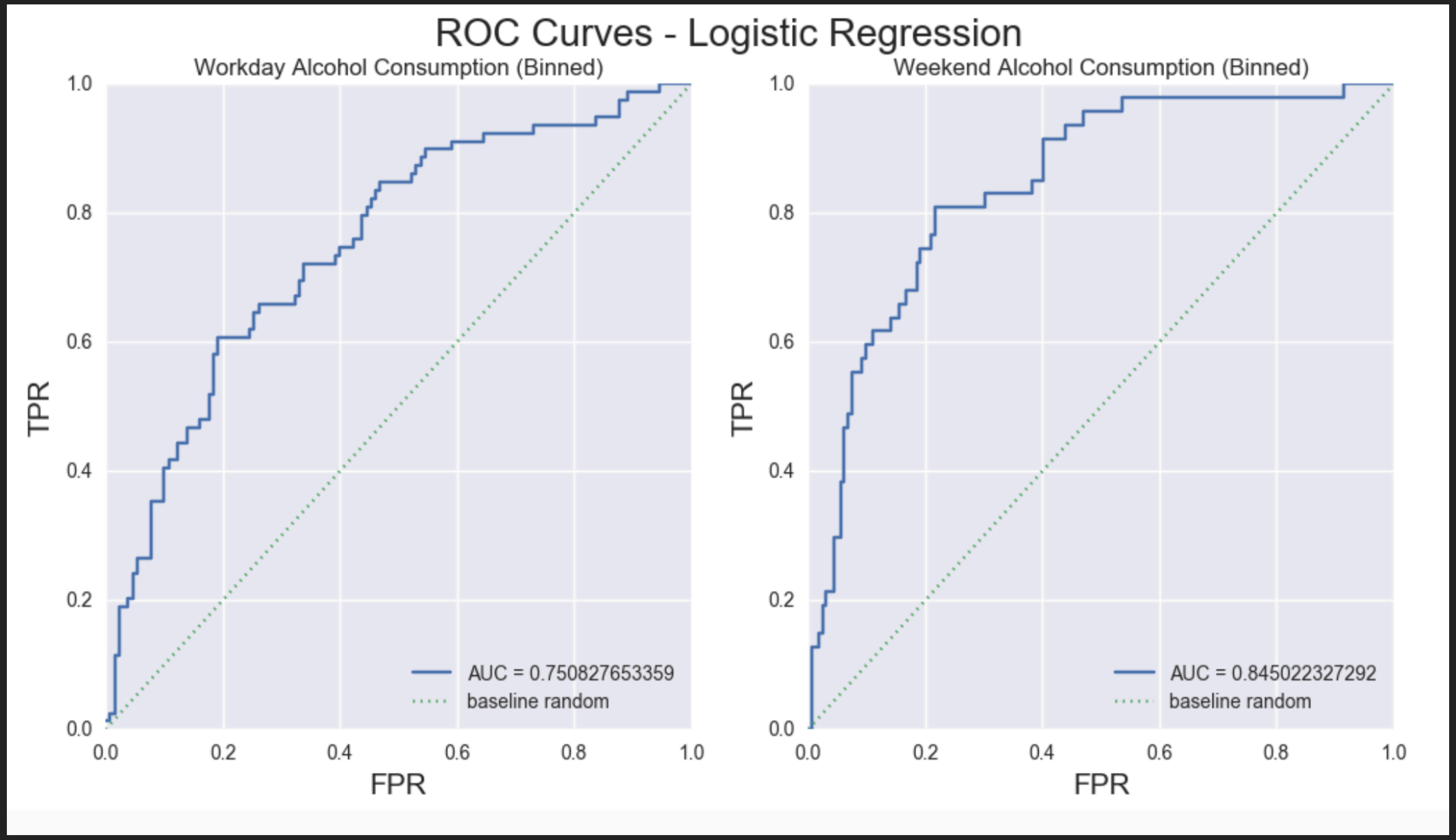
$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

- F score

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

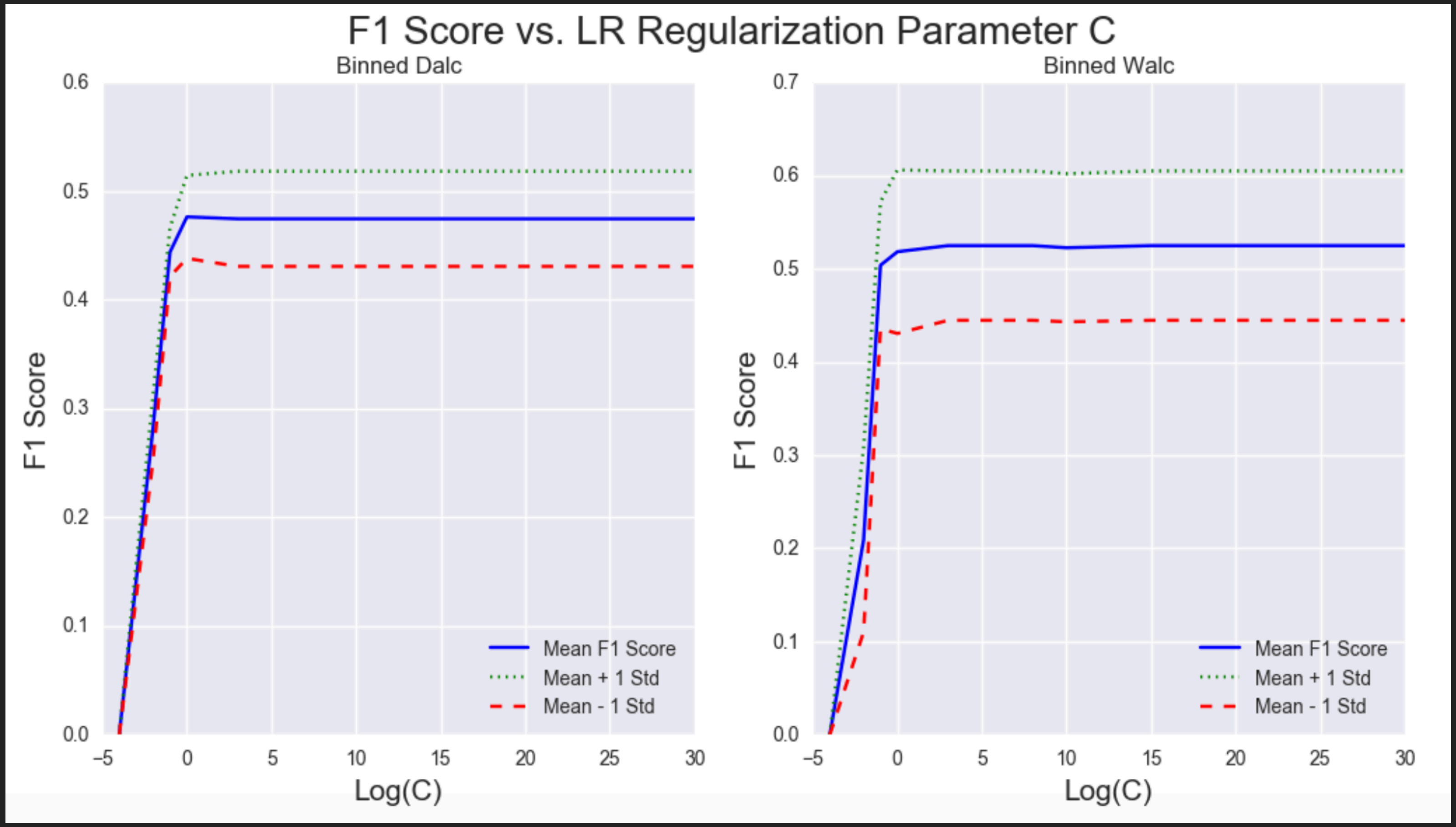
LOGISTIC REGRESSION

- Using sklearn linear_model.LogisticRegression
- Scores:
  1. Dalc F Score: 0.516129032258
  2. Walc F Score : 0.48



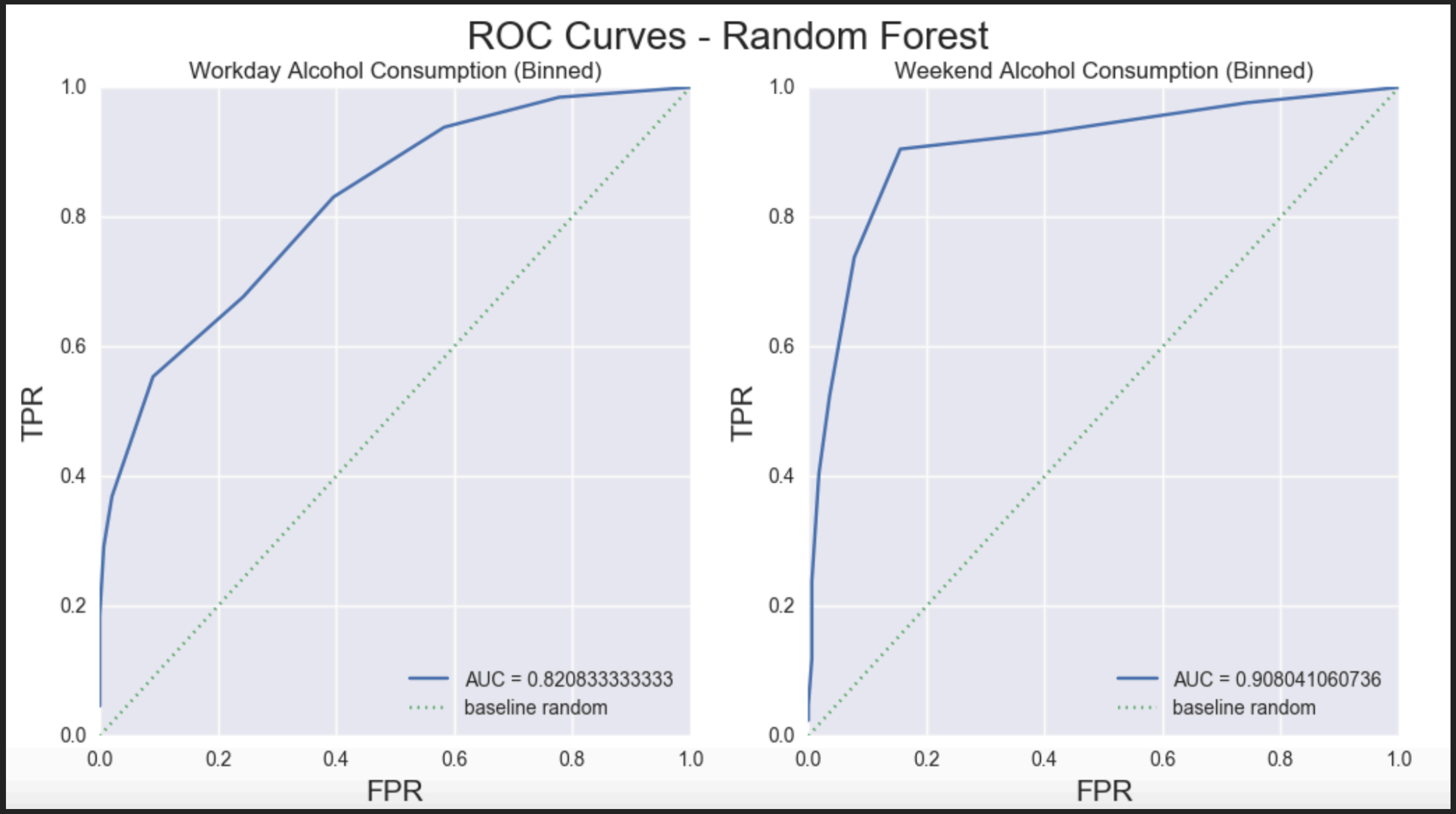
LOGISTIC REGRESSION

- Using Cross-fold validation by varying regularization parameter.
- [1e-4, 1e-2, 1e-1, 1e0, 1e3, 1e5, 1e8, 1e10, 1e15, 1e20, 1e25, 1e30]



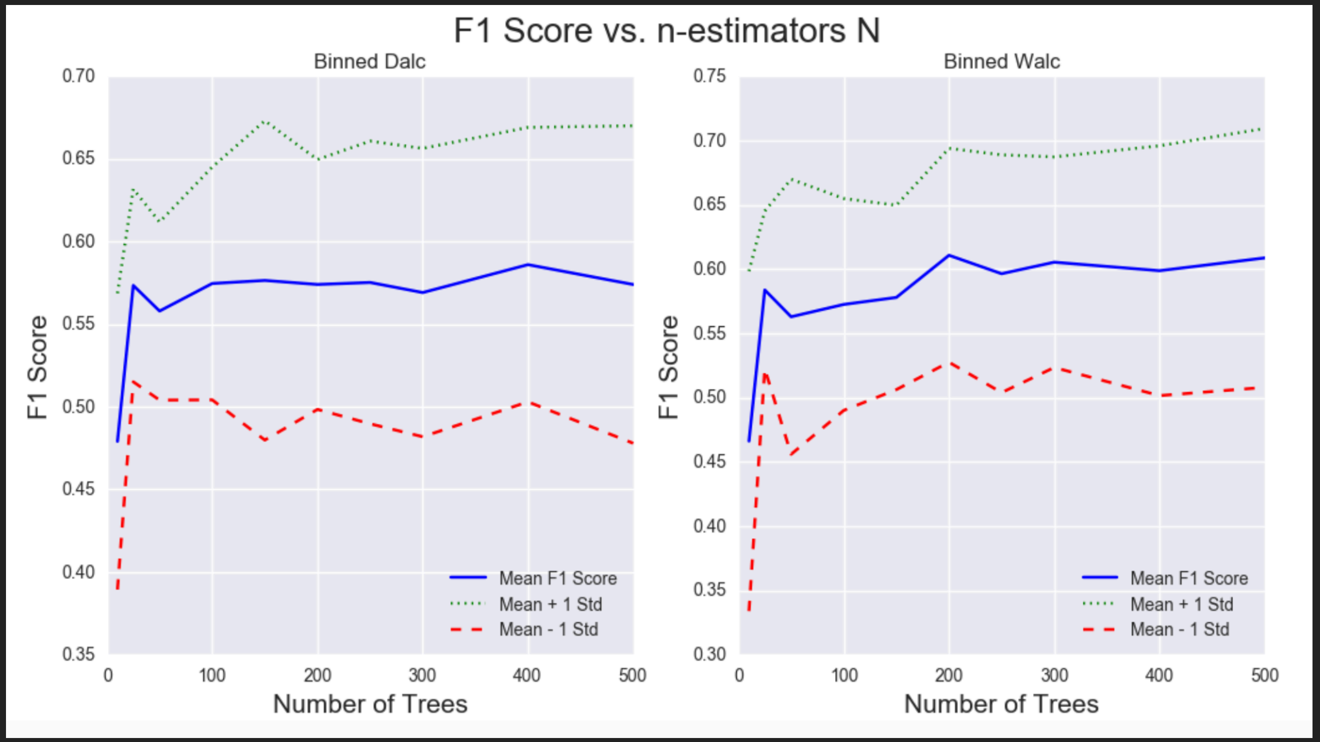
RANDOM FOREST

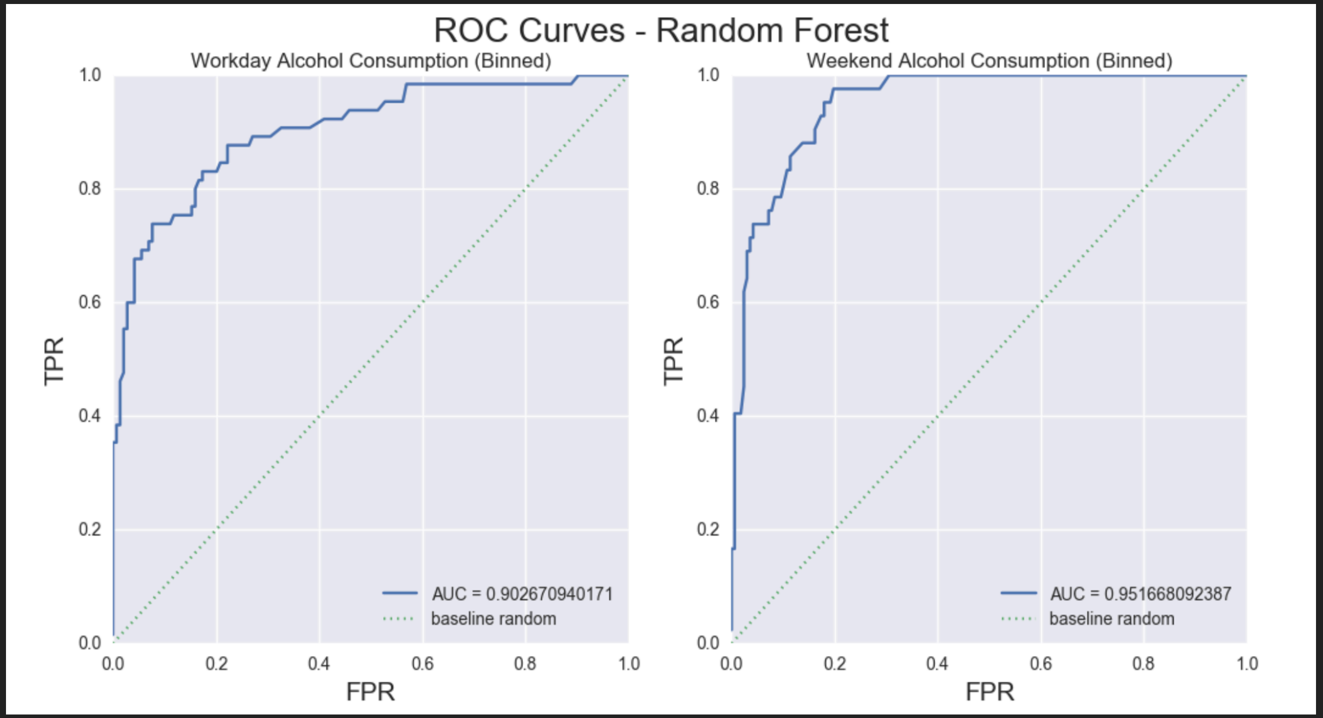
- Using RandomForestClassifier from sklearn.ensemble
- Scores with default parameters:
    1. Dalc F Score: 0.521739130435
    2. Walc F Score : 0.548387096774
- Higher Weekday score, but lower weekend score than Logistic Regression.



RANDOM FOREST

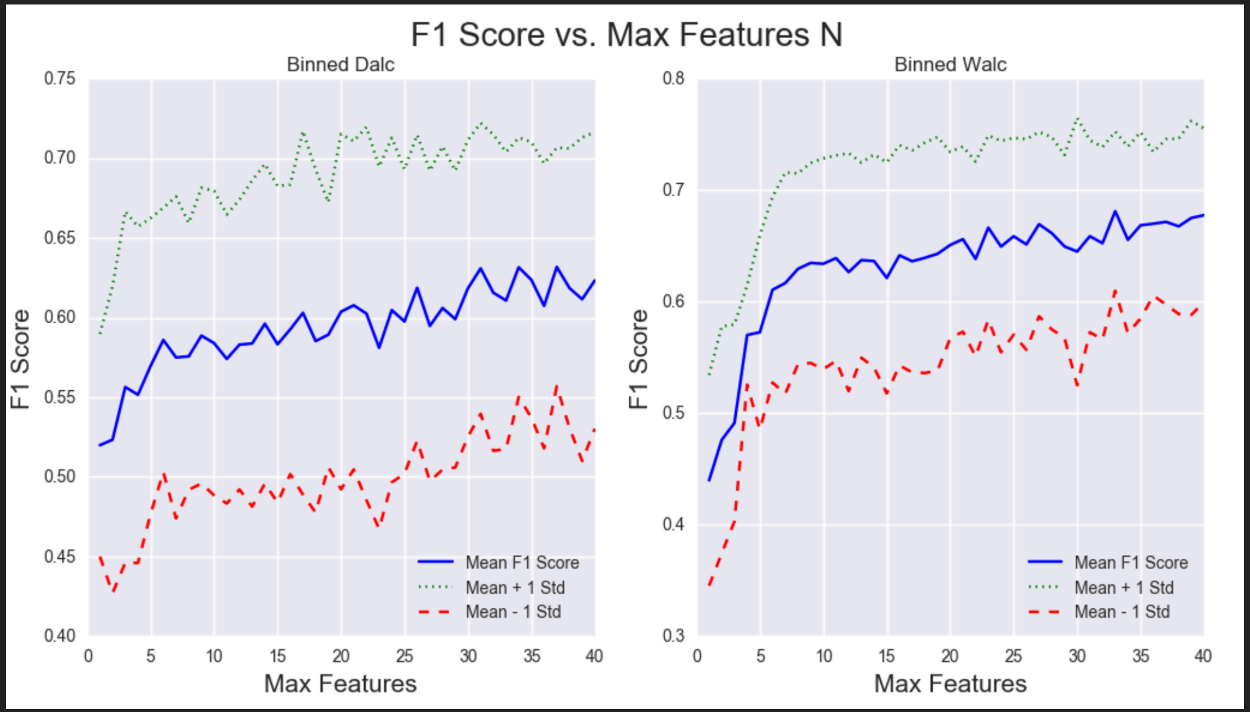
- Studying the impact of varying the n_estimators parameter using cross fold validation.
- Values: [10, 25, 50, 100, 150, 200, 250, 300, 400, 500]
- Optimum value: ~400 trees for Dalc and ~200 for Walc
- Scores:
  1. Dalc F1 Score: 0.626262626263
  2. Walc F1 Score : 0.666666666667

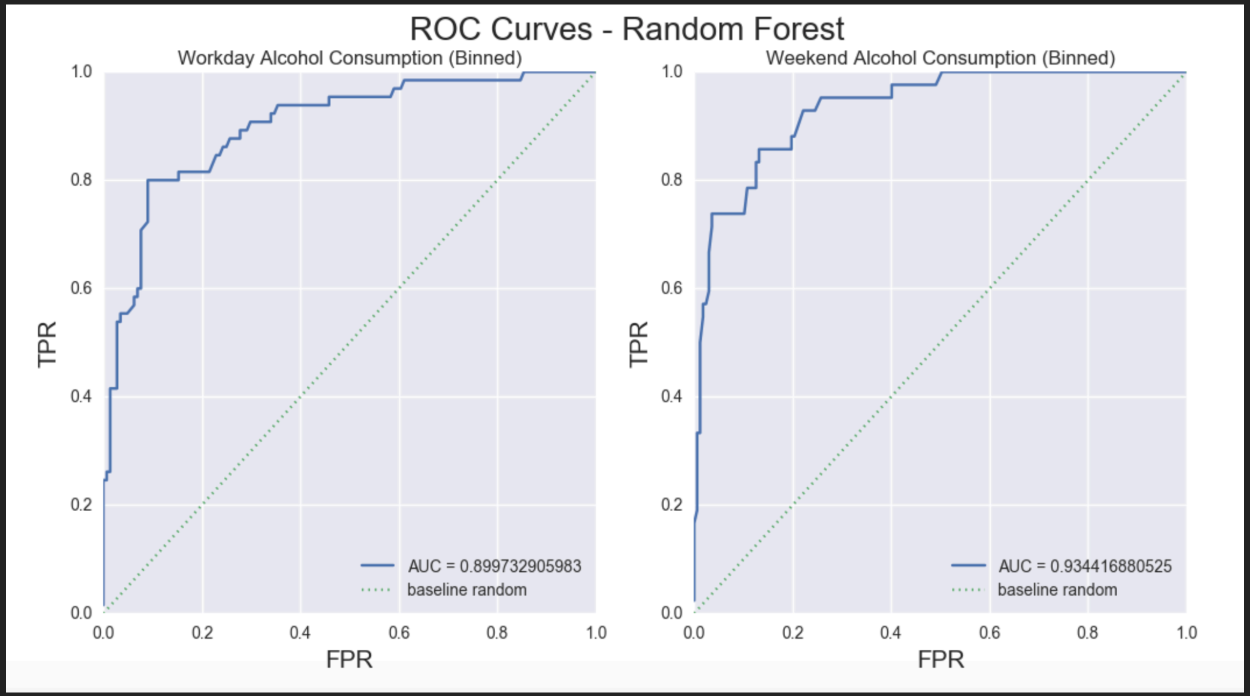




RANDOM FOREST

- Studying the impact of varying the max_features parameter using cross fold validation.
- Optimum value: ~37 trees for Dalc and ~33 for Walc
- Scores:
    1. Dalc F1 Score: 0.678571428571
    2. Walc F1 Score : 0.712328767123





COMPARISON

	F scores	
	Weekday (Walc)	Weekend (Dalc)
Naive Bayes	0.353982300885	0.225806451613
Naive Bayes tuned	0.466666666667	0.5
Logistic Regression	0.516129032258	0.48
Random Forrest	0.521739130435	0.548387096774
Random Forrest tuned	0.678571428571	0.712328767123

CONCLUSION

- Random Forrest with n_estimators=400 and max_features=37, optimal for Weekend consumption (Walc)
- Random Forrest with n_estimators=200 and max_features=33, optimal for Weekday consumption (Dalc)