Predict Active Nicotine Users

ALY6020 Winter 2019 Final Project

Instructor: Andrew Long

Minyi Chen

2/16/2019

Introduction

Project definition: predict if a person is an active Nicotine User

Drug consumption dataset from UCI Machine Learning Repository

Predictive model

P

Active Users = consume Nicotine in the last year, month, week, or day

56.2%

or

N

Non-active users = users consume in the last decade, over a decade ago, or never used

Data and methods

Data

Sample size: 1885 participants, feature number: 32, survey data collected 03/2011 to 03/2012

Demographics	Personality traits	Drug consumption
• Age • Gender • Education • Country • Race	• neuroticism (Nscore) • extraversion (Escore) • openness to experience (Oscore) • agreeableness (Ascore) • conscientiousness (Cscore) • impulsiveness • Sensation-seeking(SS)	• Alcohol • Heroin • Nicotine • …

Input feature engineering

No missing value

Numerical features (7)

- 7 personality traits: N, E, O, A, C, I, SS

Categorical features (13)

- Race (6)
- Gender
- Country (6)

One-hot encoding

Ordinal features (4)

- Age: 6 unique values
- Education: 9 unique values
- Heroin: 7 unique values
- Alcohol: 7 unique values

Build training/validation/test set

We will split into 70% train, 15% validation, 15% test.

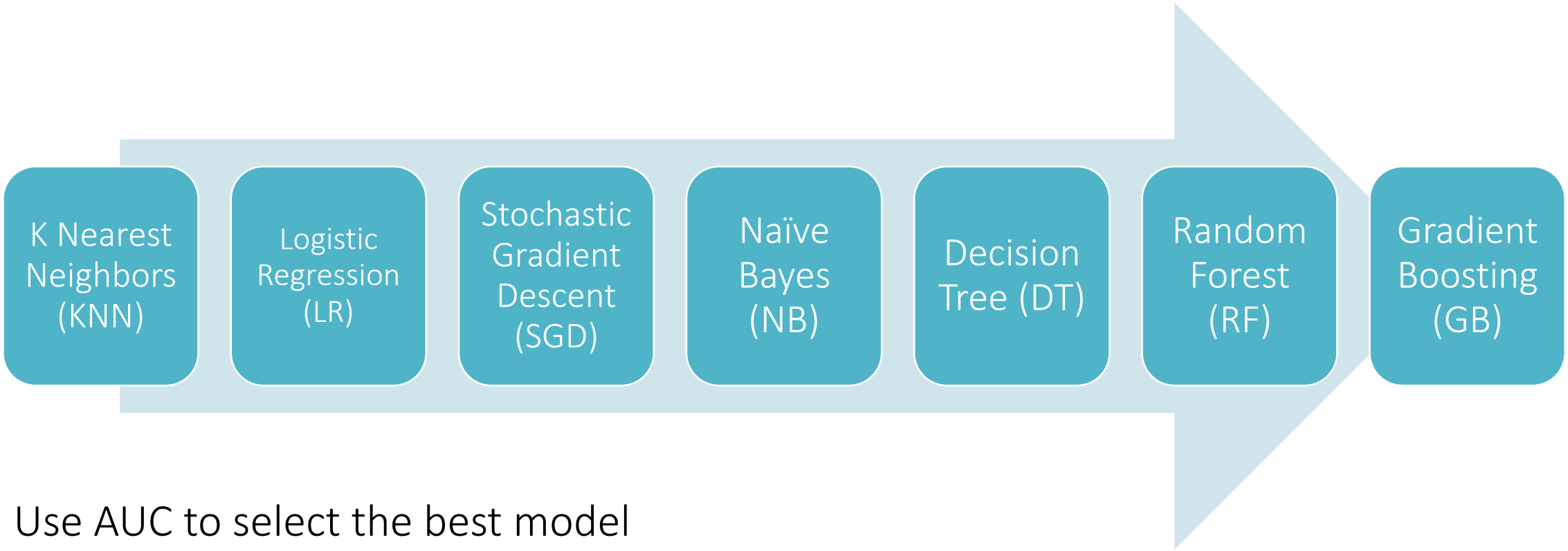
Use training data set to train our model, validation data set to see how we can improve our model, test data set to see how well the model performs.

Train all prevalence(n = 1319):0.577

Valid prevalence(n = 283):0.537

Test prevalence(n = 283):0.519

Classification methods

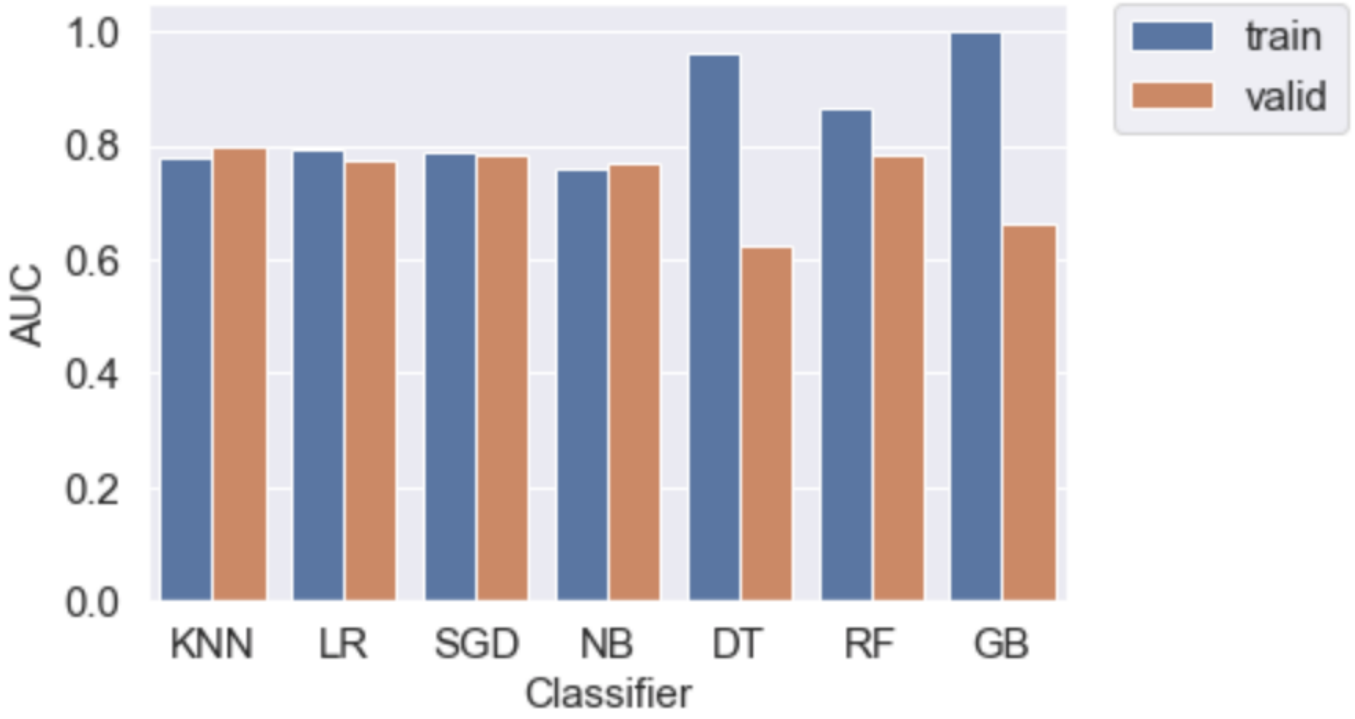


Use AUC to select the best model

- AUC does not require us selecting a threshold and helps balance true positive rate and false positive rate.

Results

Baseline models



Random forest model has an AUC of 0.782 that catches 78% of the active Nicotine users with a threshold of 0.5.

Overfitting (high variance): decision tree, gradient boosting models, random forest.

Underfitting (Bias): KNN, Logistic, SGD, NB.

Best baseline model: Random Forest

Problems: Overfitting (high variance)

Techniques:

Add more samples

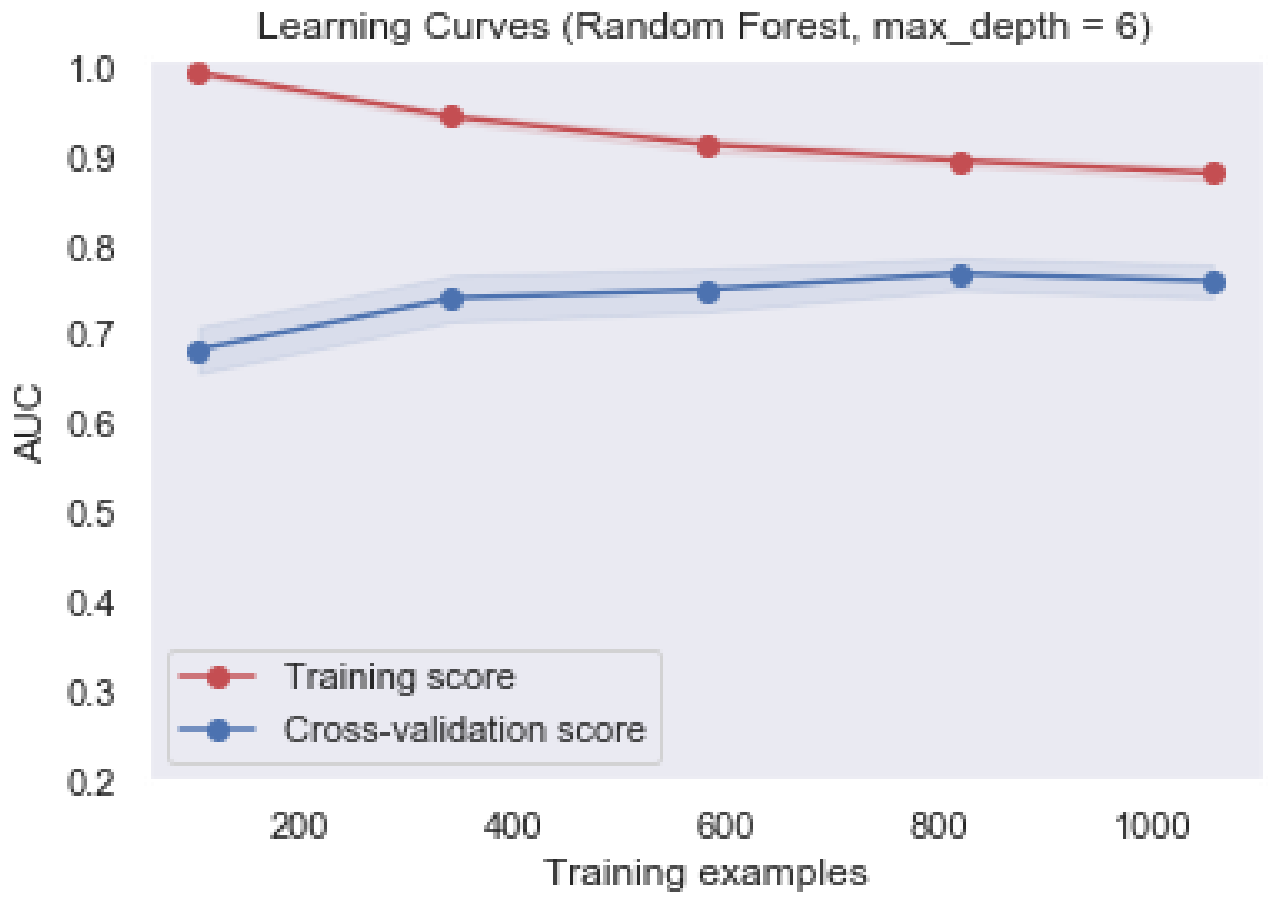
Add regularization

Reduce number of features

Decrease model complexity

Add better features

Change model architecture



Source: Andrew Ng's Coursera class paraphrased on https://www.learnopencv.com/bias-variance-tradeoff-in-machine-learning/ and Andrew Ng's Machine Learning Yearning textbook.

Input feature importance of Logistic Regression

Positive Feature Importance:

SS, C_RepofIreland, Oscore, C_USA, heroin, alcohol, C_Other, Nscore, ~~Ascore~~, C_Canada

Negative Feature Importance:

Age, Education, C_NewZealand, Cscore, R_Black/Asian, R_White, R_Black, R_White/Asian, Female, R_Other, R_Asian, ~~Escore~~, C_UK

Notes:

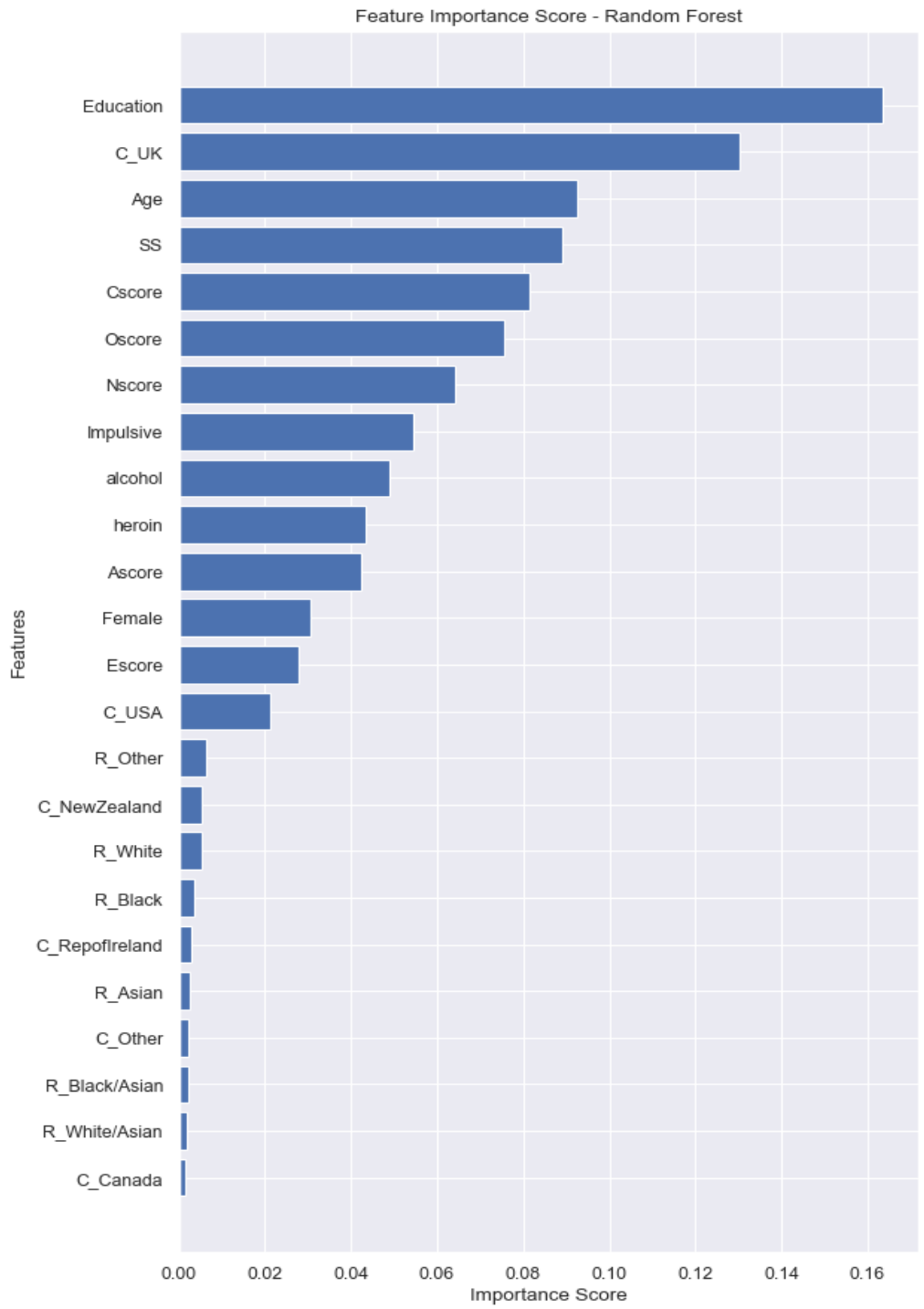
SS: high risk behavior, Impulsive: act without thinking, N: negative emotion such as anxiety, E: outgoing, talkative, etc. O: wide interests, etc. A: kindness, etc. C: organized, reliable, etc.

Input feature importance of Random Forest

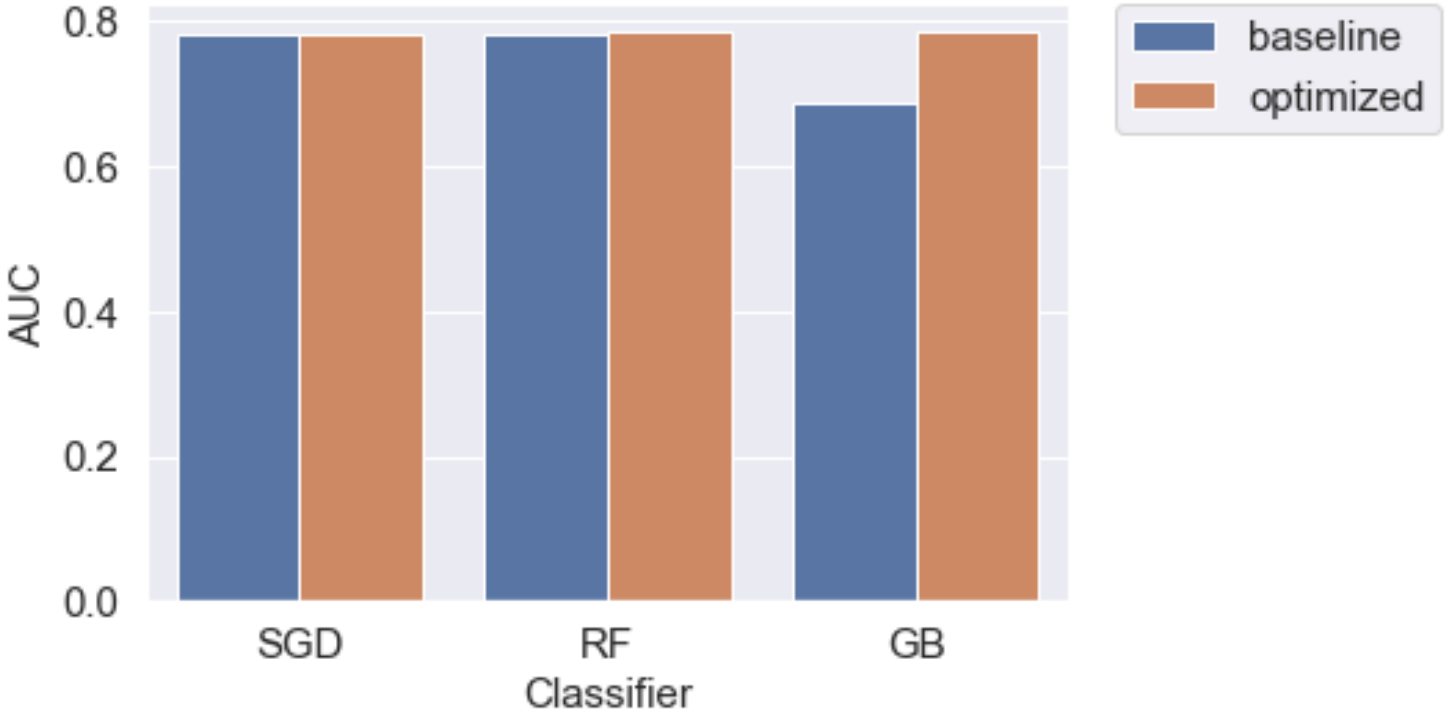
The top variables: education, UK dummy variable, age, and personality traits variables.

This makes sense since you can split continuous variables more times than categorical variables.

To reduce the overfitting problem of random forest, I remove categorical features: country and race, and also less importance features: Ascore and Escore.



Best classifier selection

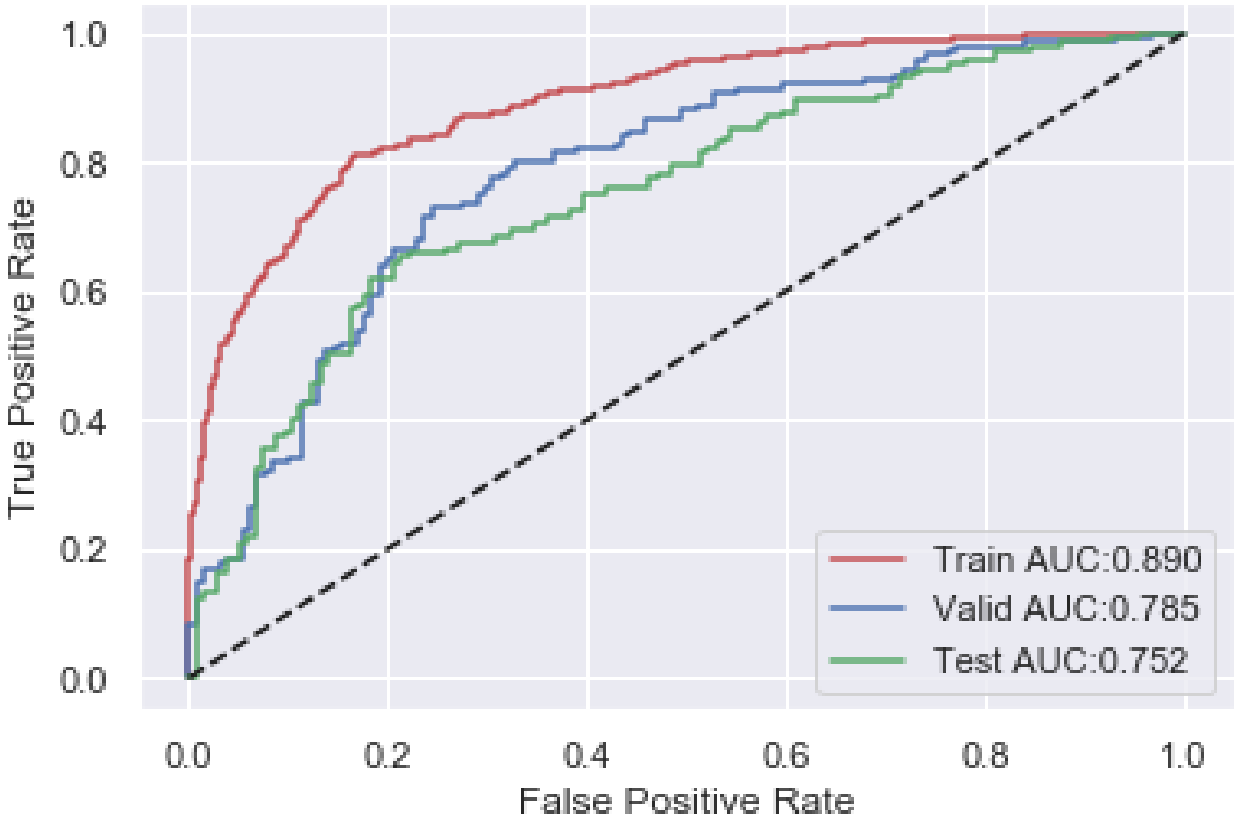


I use hyperparameter tuning for three models: SGD, RF, and GB.

From the graph, AUCs of validation set increase in the optimized models except for SGD.

RF is my best model since its train and validation AUC scores are close and its validation AUC scores is the highest too.

Model evaluation



One common method to evaluate a binary classifier is the ROC curve graph.

If a binary classifier predict every observations correctly, then it would be a straight line on the left border and then the top border.

If a binary classifier doesn't predict well it will be closer to the diagonal line.

The roc curve graph shows that my model has pretty high accuracy.

The model on the test dataset has an AUC of 0.752 that catches 71.4% of the active nicotine users with a threshold of 0.5.

Conclusion

Active nicotine users tend to have
- higher sensation seeking score (SS)
- higher openness to experience score (O)
- larger consumption of heroin and alcohol
- higher neuroticism (N) score

Non-active nicotine users:
- Older
- better educated
- more conscientiousness
- female

This suggests that younger people who have less education, and higher scores on SS, O, and N would be our focus group. We should provide more education on this group of people and help them reduce nicotine consumption and maybe other drug consumptions as well.

References

Drug consumption (quantified) Data Set. Retrieved from: https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29

Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., & Gorban, A. N. (2017). The Five Factor Model of personality and evaluation of drug consumption risk. In *Data Science* (pp. 231-242). Springer, Cham.

Thank you!