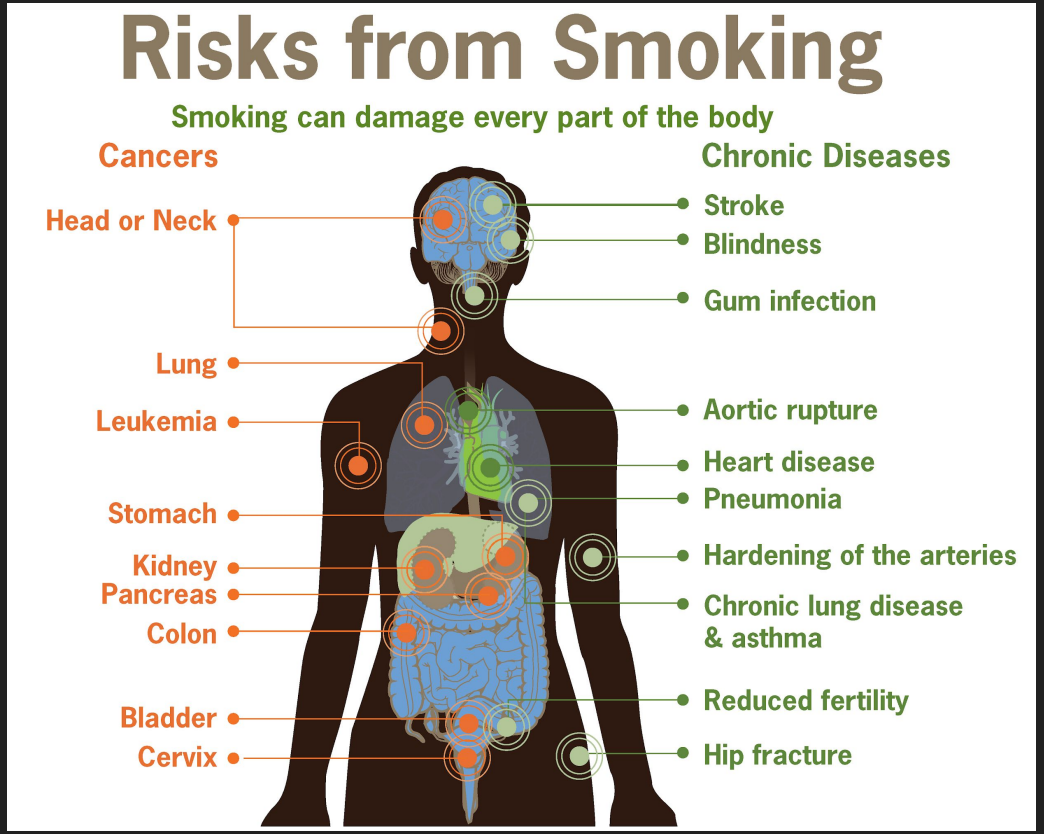
Personality Type Nicotine Consumption Risk Assessment

Predicting your risk of being a smoker given your personality traits.

Springboard Data Science Capstone Project
May 26th 2020 Cohort
Filiberto Aguilar

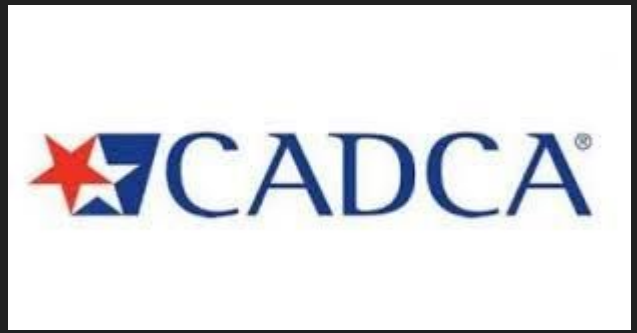
The Problem

- Smoking = poor health
- Smoking is the leading cause of preventable death in the US
- Smoking is very addictive



American
Addiction Centers

Can we predict who is at high risk of becoming a smoker based on their personality profile?

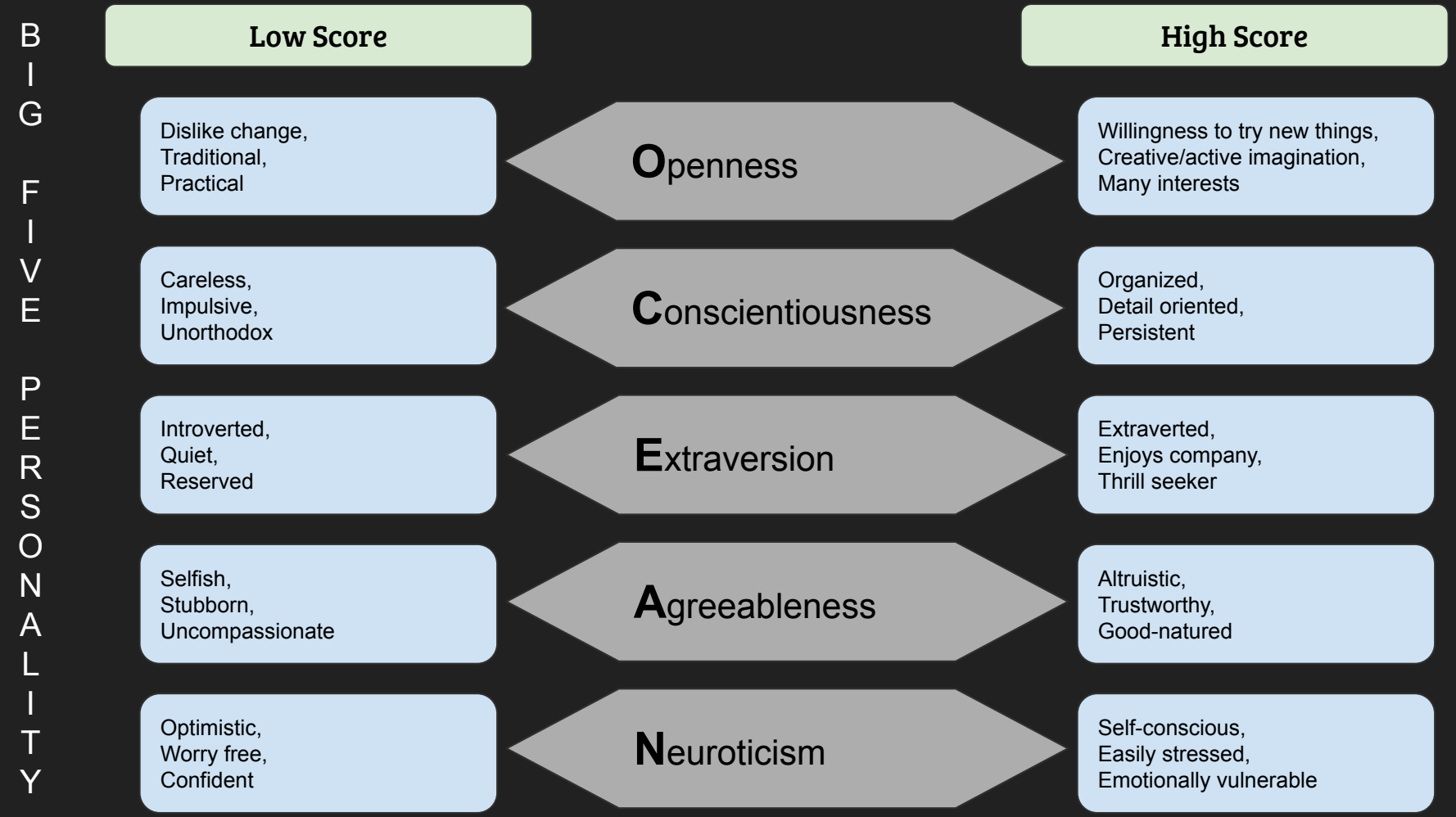


The Data

Features:

- Education
- Gender
- Country of residence
- Ethnicity
- 7 personality traits:
  - Big five OCEAN traits
  - BIS-11 or impulsiveness
  - ImpSS or sensation seeking
- 19 psychoactive drugs

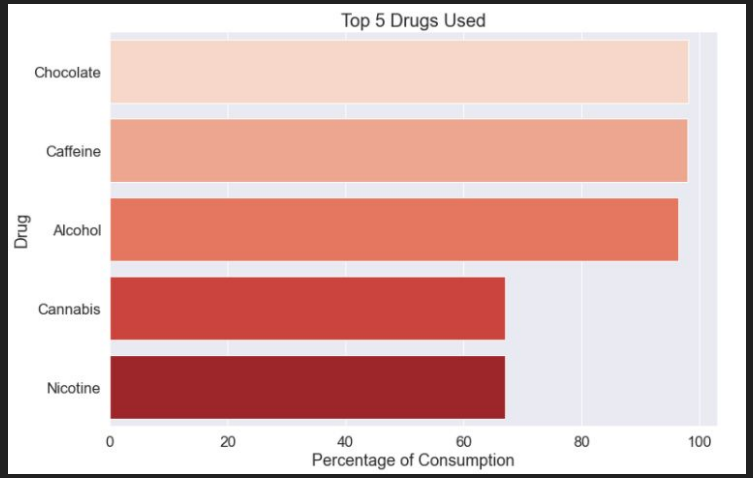
All categorical except the personality traits.

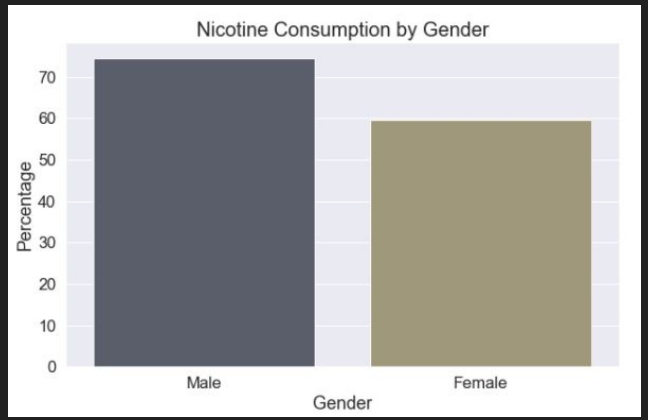

BIG FIVE PERSONALITY
Low Score
High Score
Openness
Dislike change, Traditional, Practical
Willingness to try new things, Creative/active imagination, Many interests
Conscientiousness
Careless, Impulsive, Unorthodox
Organized, Detail oriented, Persistent
Extraversion
Introverted, Quiet, Reserved
Extraverted, Enjoys company, Thrill seeker
Agreeableness
Selfish, Stubborn, Uncompassionate
Altruistic, Trustworthy, Good-natured
Neuroticism
Optimistic, Worry free, Confident
Self-conscious, Easily stressed, Emotionally vulnerable

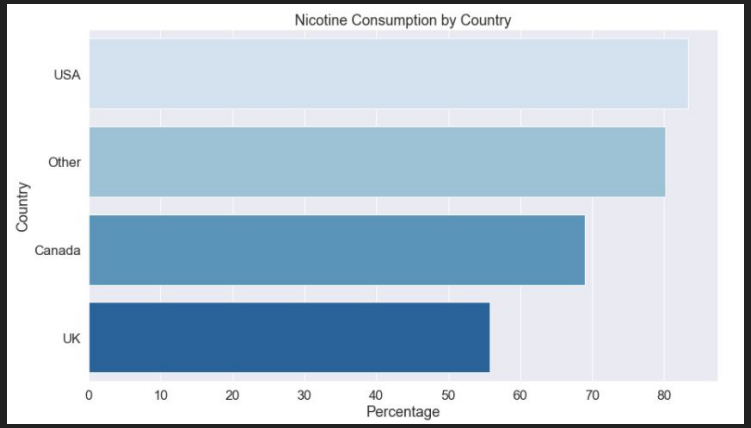
Disclaimers

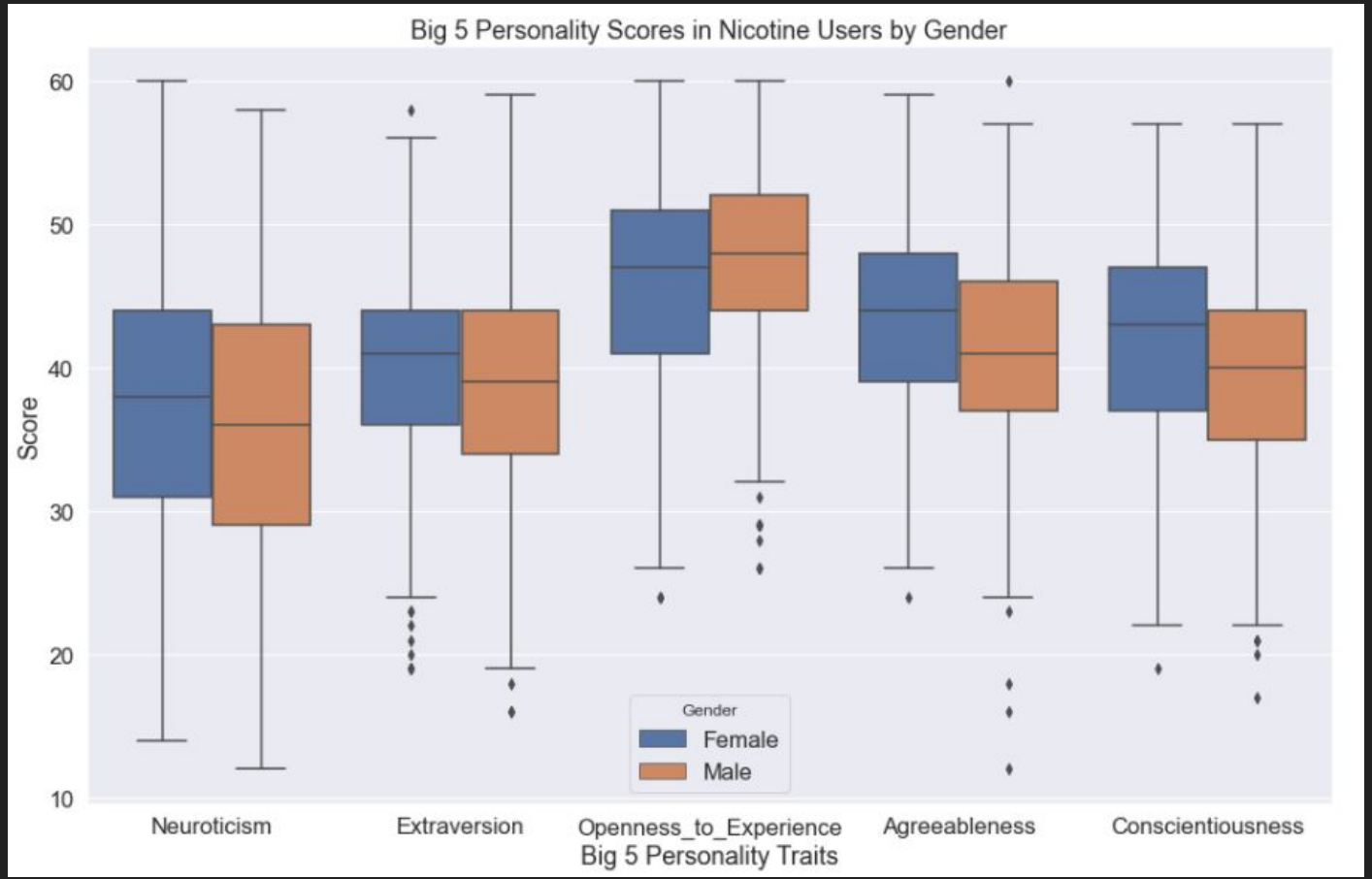
1. Given that the BIS-11 and ImpSS features are used primarily to measure substance abuse the big five or OCEAN traits were the primary focus of this project.
2. Nicotine consumption is assumed to be in the form of cigarette usage.
3. Personality profiles are subject to change as a person matures but have been proven to be valid and reliable assessments by researchers.

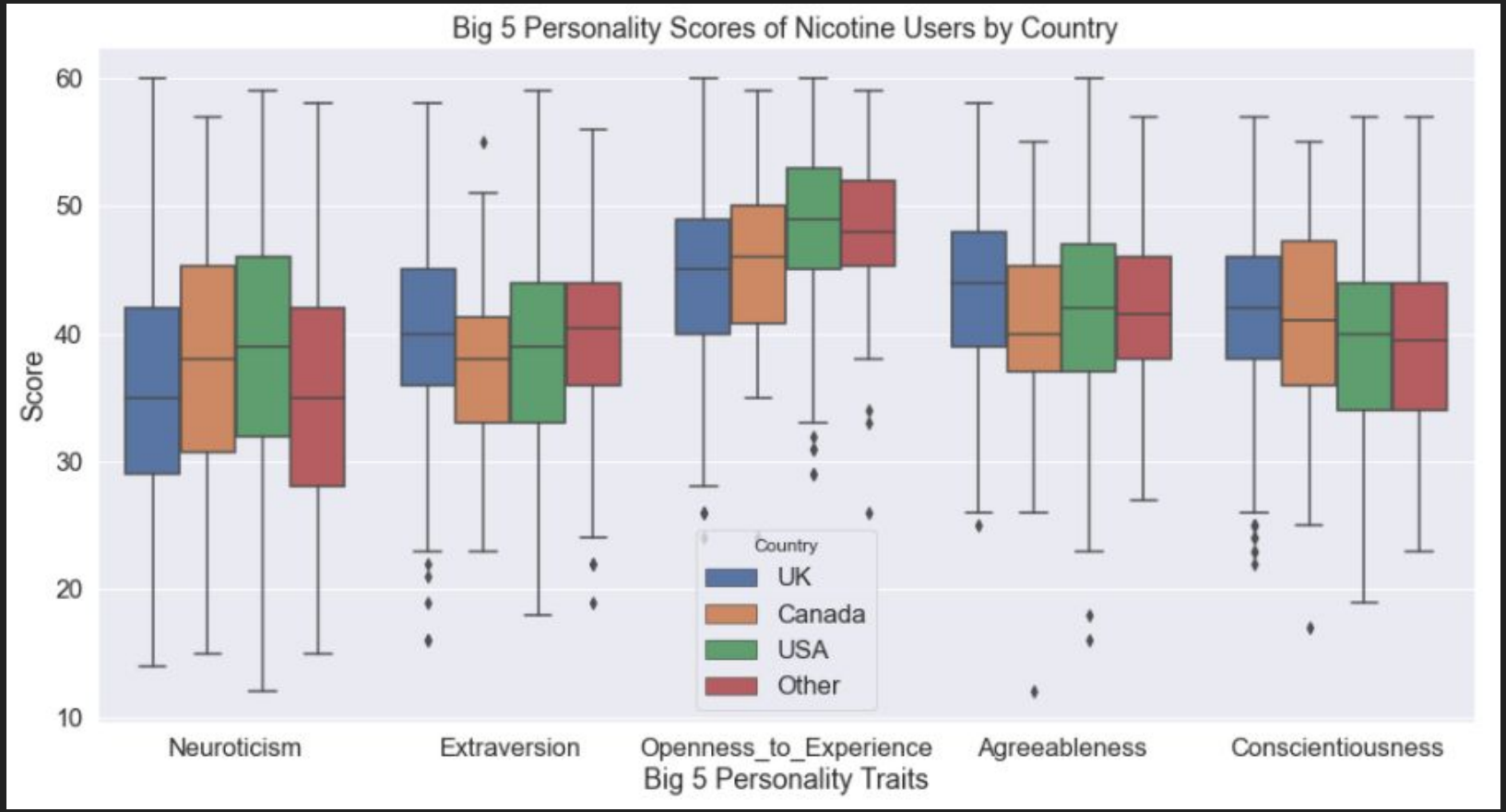
Exploratory Data Analysis

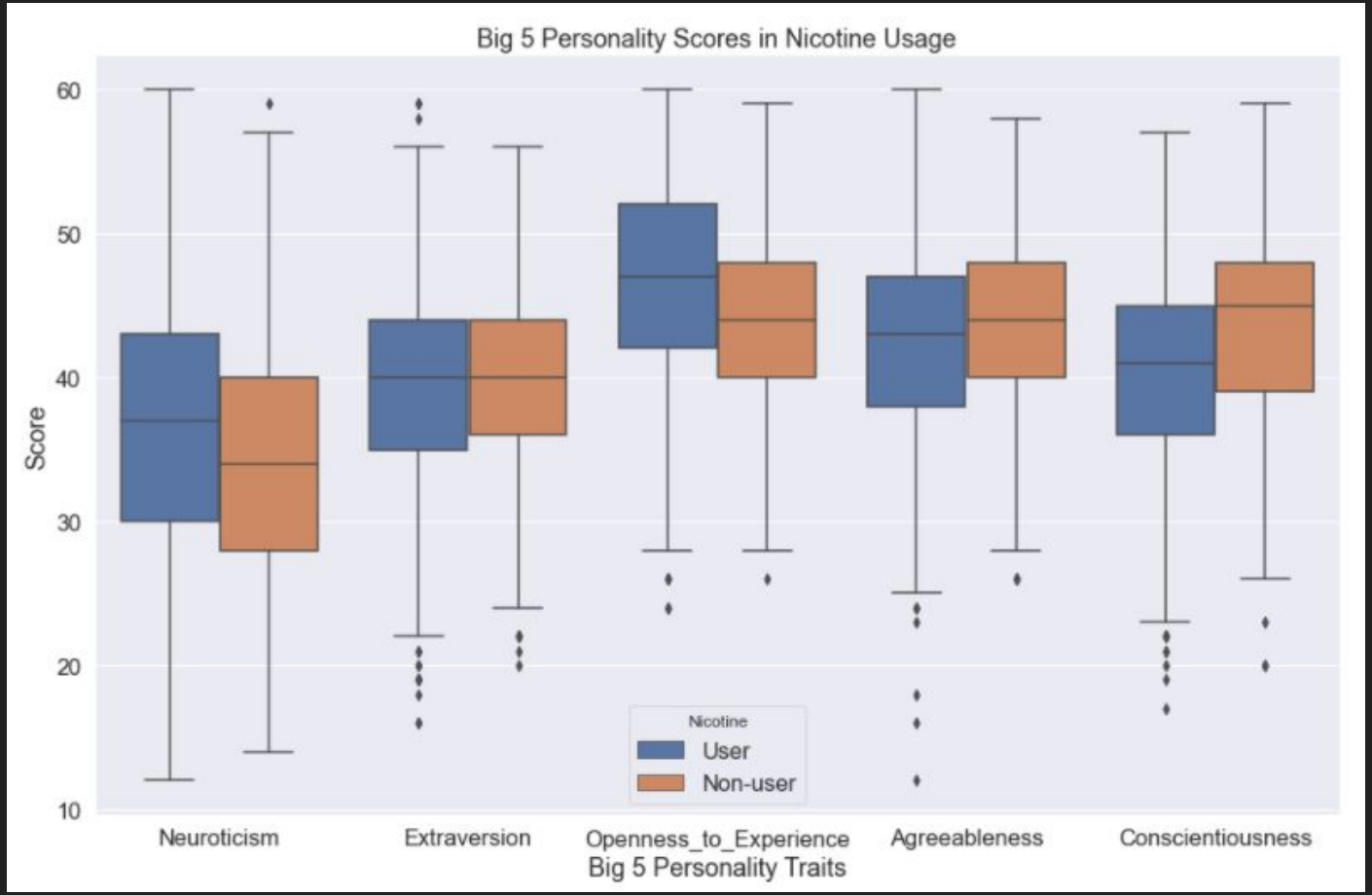

Top 5 Drugs Used
Chocolate
Caffeine
Alcohol
Cannabis
Nicotine
Drug
0
20
40
60
80
100
Percentage of Consumption


Nicotine Consumption by Gender
70
60
50
40
30
20
10
0
Percentage
Male
Female
Gender


Nicotine Consumption by Country
USA
Other
Canada
UK
Country
0
10
20
30
40
50
60
70
80
Percentage


Big 5 Personality Scores in Nicotine Users by Gender
Score
60
50
40
30
20
10
Gender
Female
Male
Neuroticism
Extraversion
Openness_to_Experience
Agreeableness
Conscientiousness
Big 5 Personality Traits


Big 5 Personality Scores of Nicotine Users by Country
Score
60
50
40
30
20
10
Country
UK
Canada
USA
Other
Neuroticism
Extraversion
Openness_to_Experience
Agreeableness
Conscientiousness
Big 5 Personality Traits


Big 5 Personality Scores in Nicotine Usage
Score
60
50
40
30
20
10
Nicotine
User
Non-user
Neuroticism
Extraversion
Openness_to_Experience
Agreeableness
Conscientiousness
Big 5 Personality Traits

Modeling

Six models were considered:

- K-nearest Neighbors (KNN)
- Logistic Regression
- Support Vector Machines (SVM)
- Naive Bayes
- Random Forest
- Gradient Boosting

Modeling steps

Pre-Processed Data:

1. Encoded categorical variables
2. Scaled features
3. Split into training and test sets by 75% - 25%

Feature Importance:

1. Checked for multicollinearity through VIF scores
2. Explored the effects of a one unit increase through odds ratios

Trained and tuned parameters via grid search cross validation:

- 5 fold cv
- Each model performance was evaluated by the ‘ROC-AUC’ score

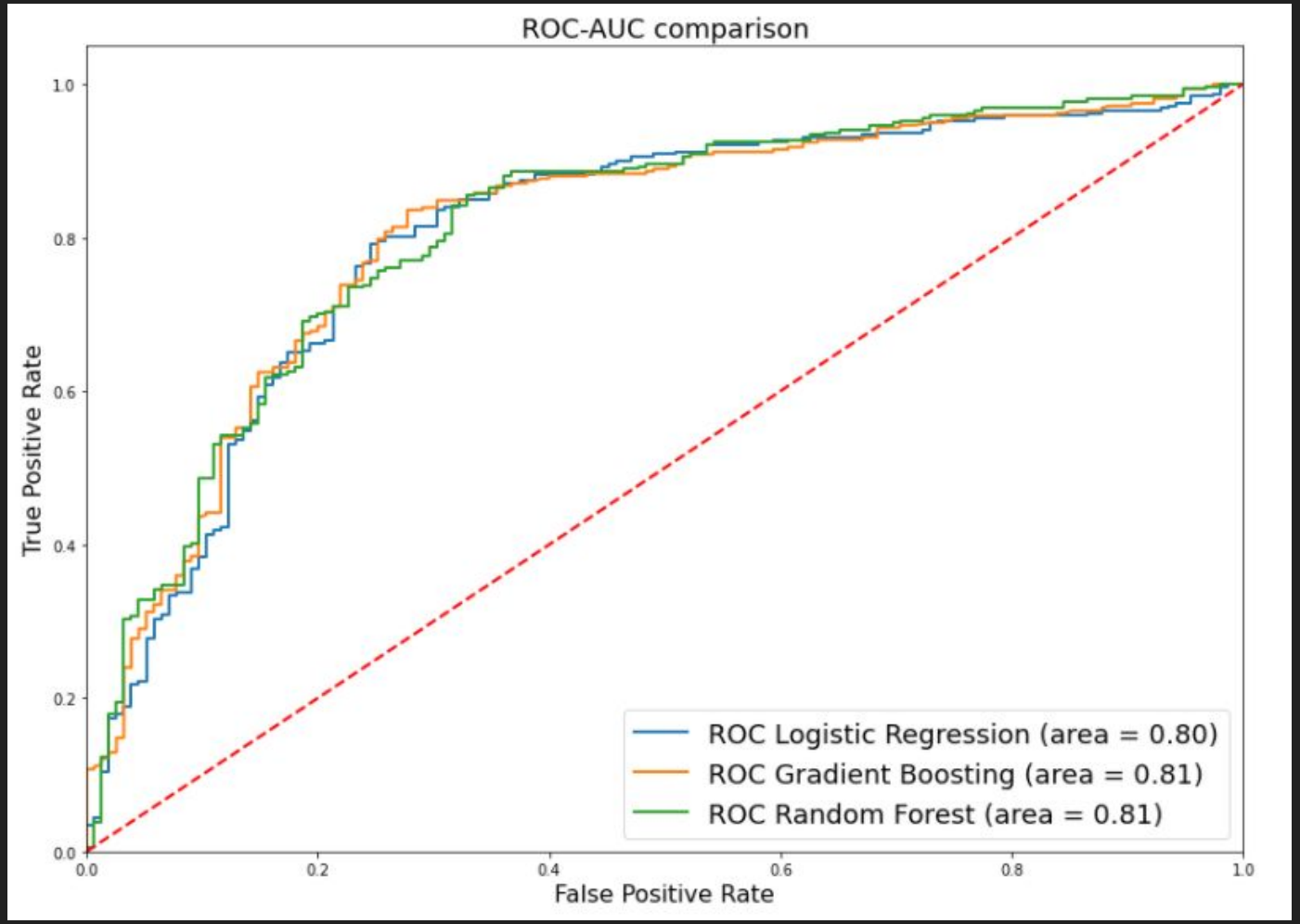
Feature Importance

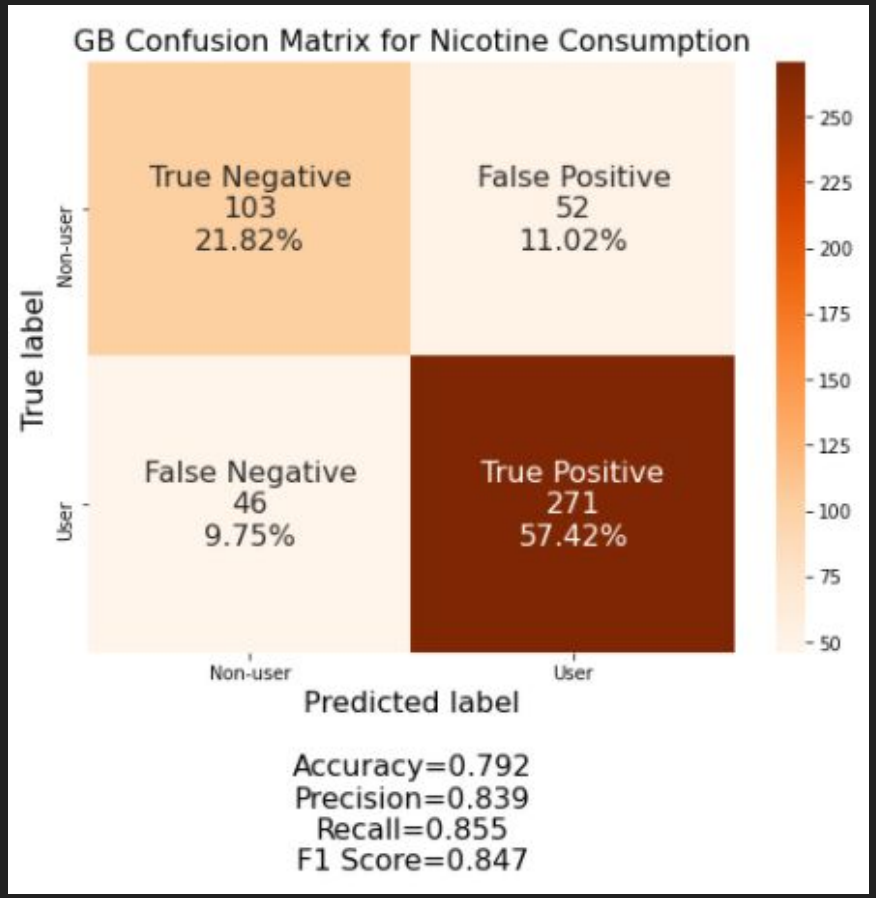
OCEAN Features	VIF Score	Standardized Regression Coefficients	Odds Ratios
Openness to experience	1.1	0.3904	1.0695
Neuroticism	1.4	0.1212	1.0171
Extraversion	1.4	0.0623	1.0123
Agreeableness	1.1	-0.1564	0.9734
Conscientiousness	1.3	-0.3200	0.9476

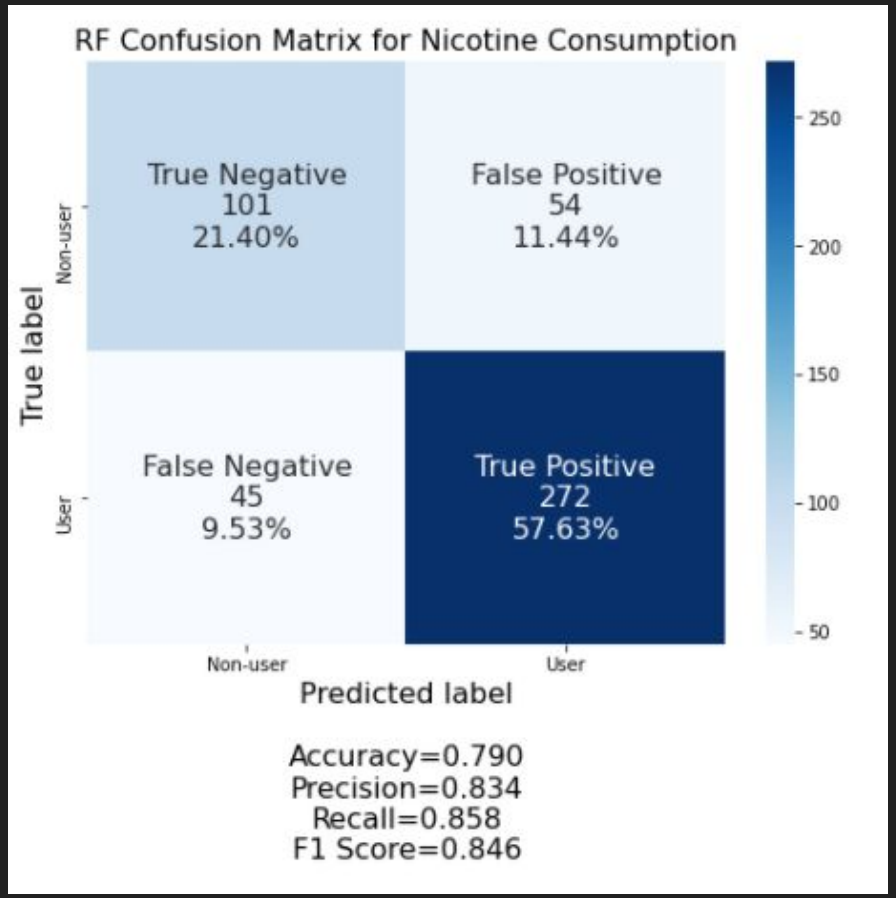
Model Performance

Model	ROC-AUC	Brier Score
KNN	0.8225	0.1683
Logistic Regression	0.8337	0.1598
SVM	0.8305	0.1570
Naive Bayes	0.8116	0.2720
Random Forest	0.8362	0.1578
Gradient Boosting	0.8341	0.1548

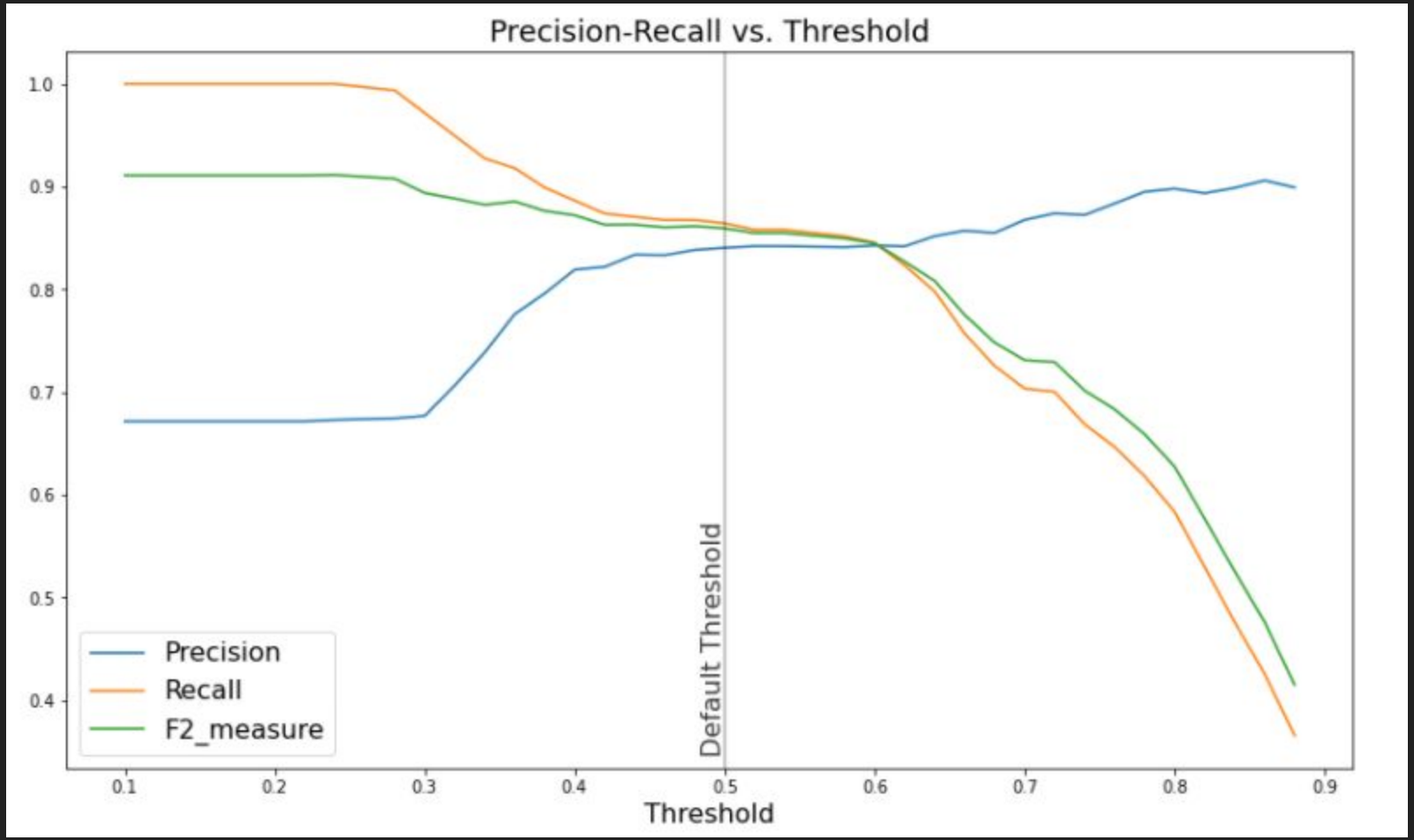
Best Performers

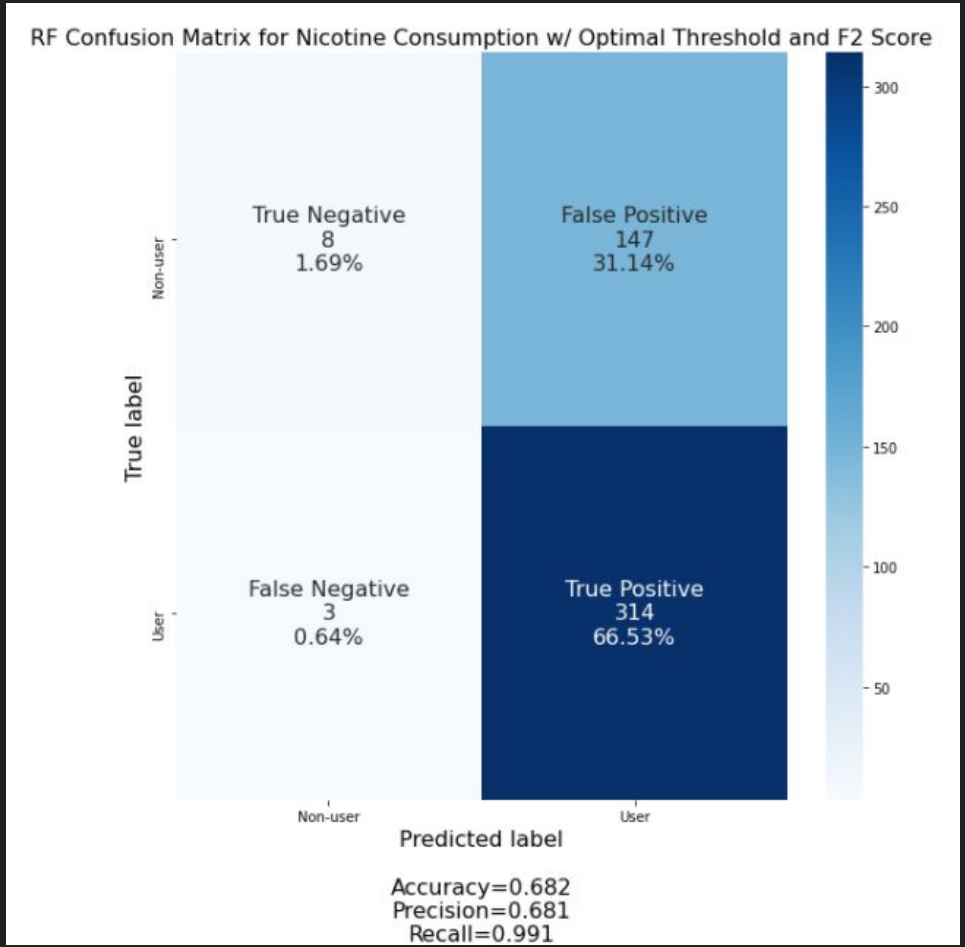

ROC-AUC comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Logistic Regression (area = 0.80)
ROC Gradient Boosting (area = 0.81)
ROC Random Forest (area = 0.81)


GB Confusion Matrix for Nicotine Consumption
True Negative
103
21.82%
False Positive
52
11.02%
False Negative
46
9.75%
True Positive
271
57.42%
True label
Non-user
User
Non-user
User
Predicted label
250
225
200
175
150
125
100
75
50
Accuracy=0.792
Precision=0.839
Recall=0.855
F1 Score=0.847


RF Confusion Matrix for Nicotine Consumption
True Negative
101
21.40%
False Positive
54
11.44%
False Negative
45
9.53%
True Positive
272
57.63%
True label
Non-user
User
Non-user
User
Predicted label
250
200
150
100
50
Accuracy=0.790
Precision=0.834
Recall=0.858
F1 Score=0.846

Model Evaluation


Precision-Recall vs. Threshold
1.0
0.9
0.8
0.7
0.6
0.5
0.4
Precision
Recall
F2_measure
Default Threshold
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Threshold


RF Confusion Matrix for Nicotine Consumption w/ Optimal Threshold and F2 Score
True Negative
8
1.69%
False Positive
147
31.14%
False Negative
3
0.64%
True Positive
314
66.53%
True label
Non-user
User
Non-user
User
Predicted label
Accuracy=0.682
Precision=0.681
Recall=0.991
300
250
200
150
100
50

Conclusion

- Openness and neuroticism were the big five personality traits common in most smokers.
- Agreeableness and conscientiousness were indicative of non-smokers.
- Random forest model was able to virtually reduce the number of false negatives at the expense of its precision.
- With collection of more features and observations, perhaps through surveying, better model performance can be achieved.

Special thanks to:
- Benjamin Bell, Springboard mentor
- Springboard community

Sources

Centers for Disease Control and Prevention (2020, May 21). Tobacco Fast Facts. Retrieved September 20, 2020, from https://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/

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. *Data Science Studies in Classification, Data Analysis, and Knowledge Organization,* 231-242. doi:10.1007/978-3-319-55723-6_18

Lim, A. G. (n.d.). The Big Five Personality Traits. Retrieved September 02, 2020, from https://www.simplypsychology.org/big-five-personality.html

National Institute on Drug Abuse. (2020, July 24). Cigarettes and Other Tobacco Products DrugFacts. Retrieved September 20, 2020, from https://www.drugabuse.gov/publications/drugfacts/cigarettes-other-tobacco-products

Raypole, C. (2019, January 26). Big Five Personality Traits: How They're Measured, What They Mean. Retrieved September 20, 2020, from https://www.healthline.com/health/big-five-personality-traits

Tavares, E. (n.d.). Variance Inflation Factor (VIF) Explained. Retrieved September 20, 2020, from https://etav.github.io/python/vif_factor_python.html

UCLA. (n.d.). FAQ: How Do I Interpret Odds Ratios in Logistic Regression? Retrieved September 20, 2020, from https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/