

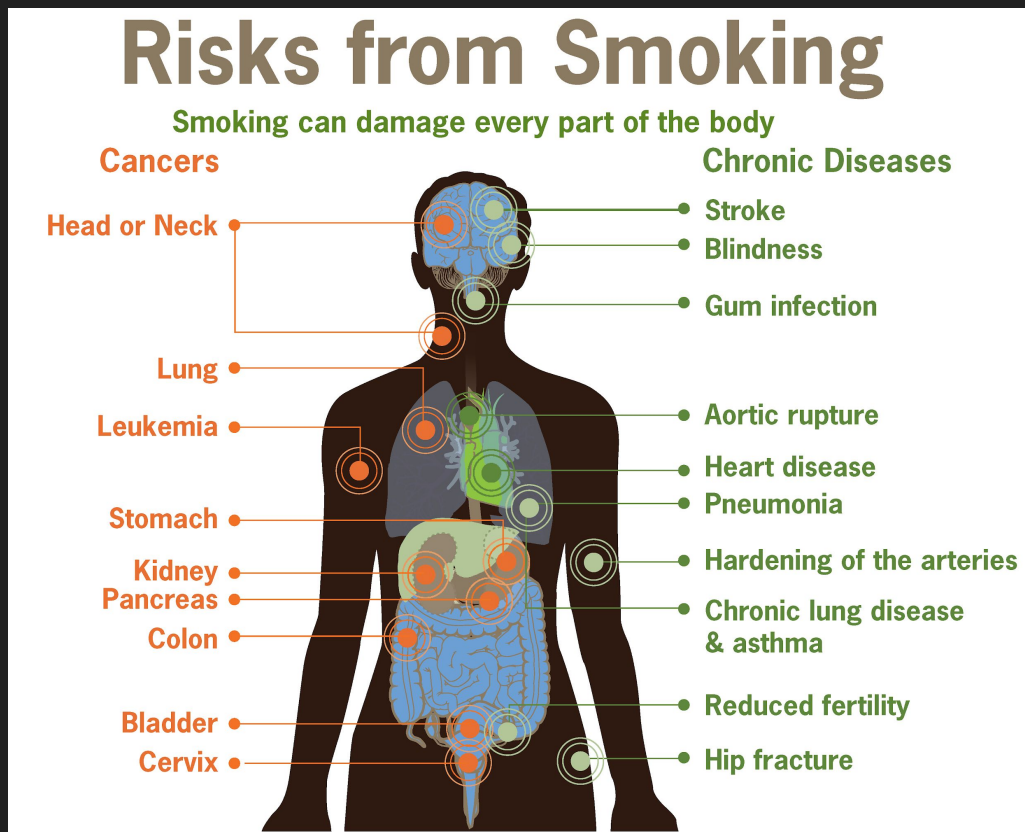
Personality Type Nicotine Consumption Risk Assessment

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



The logo for the Substance Abuse and Mental Health Services Administration (SAMHSA). It features the acronym "SAMHSA" in a large, bold, blue, italicized sans-serif font.

Substance Abuse and Mental Health
Services Administration



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.



B
I
G

F
I
V
E

P
E
R
S
O
N
A
L
I
T
Y

Low Score

Dislike change,
Traditional,
Practical

Careless,
Impulsive,
Unorthodox

Introverted,
Quiet,
Reserved

Selfish,
Stubborn,
Uncompassionate

Optimistic,
Worry free,
Confident

Openness

Conscientiousness

Extraversion

Agreeableness

Neuroticism

High Score

Willingness to try new things,
Creative/active imagination,
Many interests

Organized,
Detail oriented,
Persistent

Extraverted,
Enjoys company,
Thrill seeker

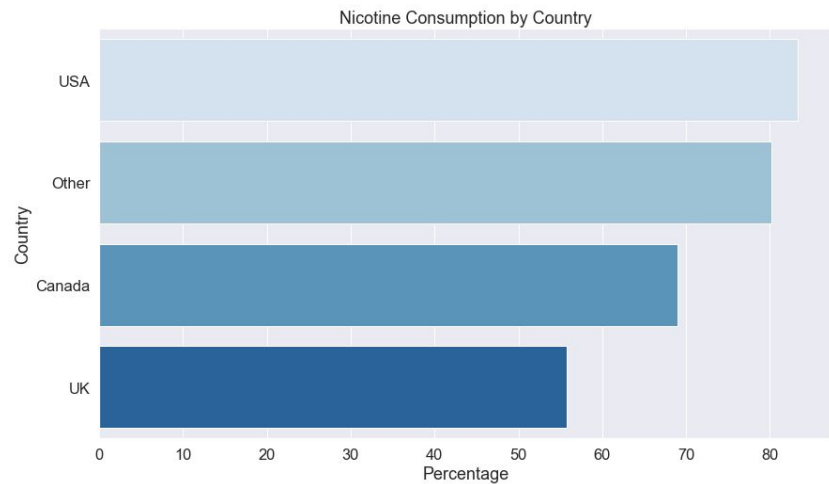
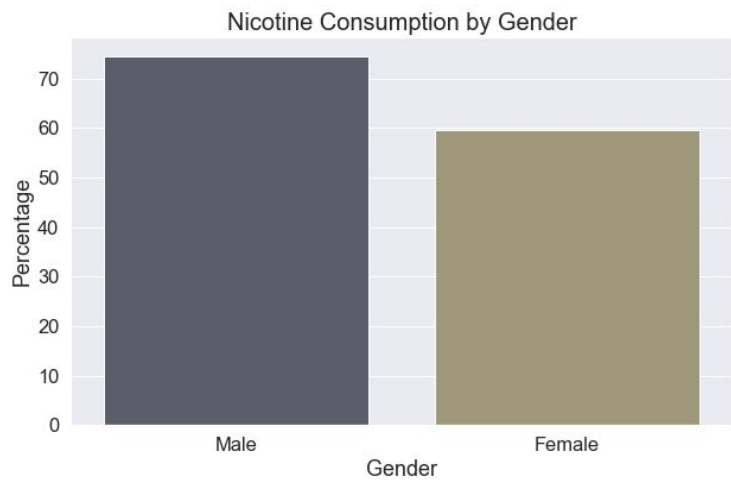
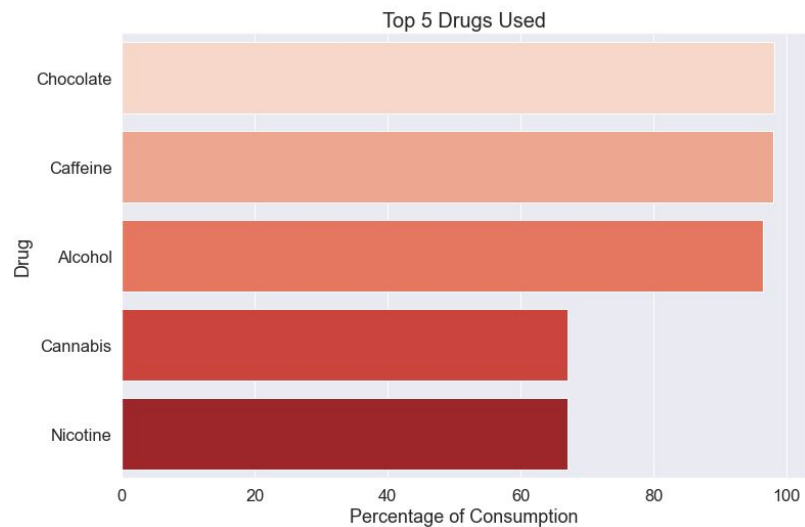
Altruistic,
Trustworthy,
Good-natured

Self-conscious,
Easily stressed,
Emotionally vulnerable

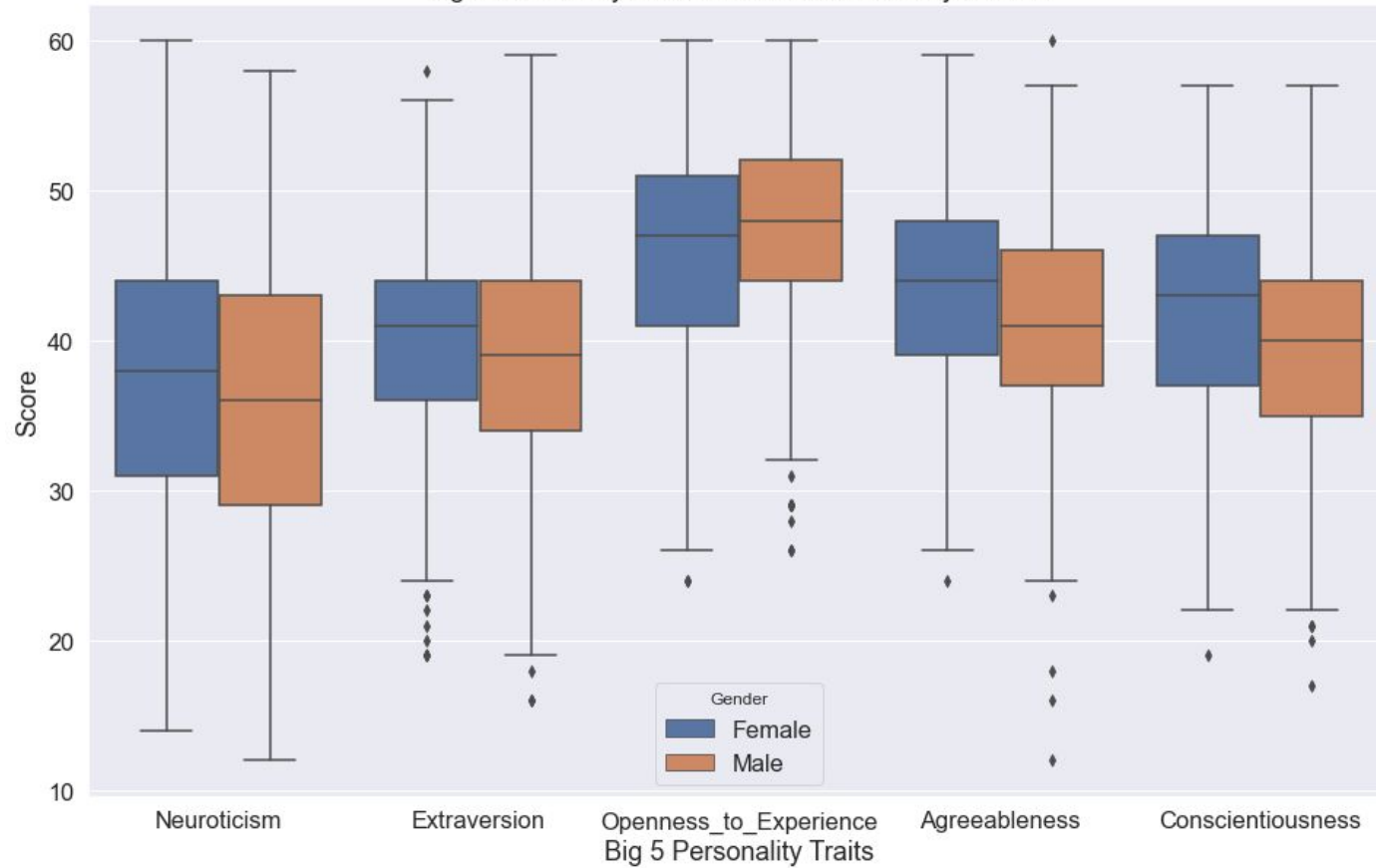
Disclaimers

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.

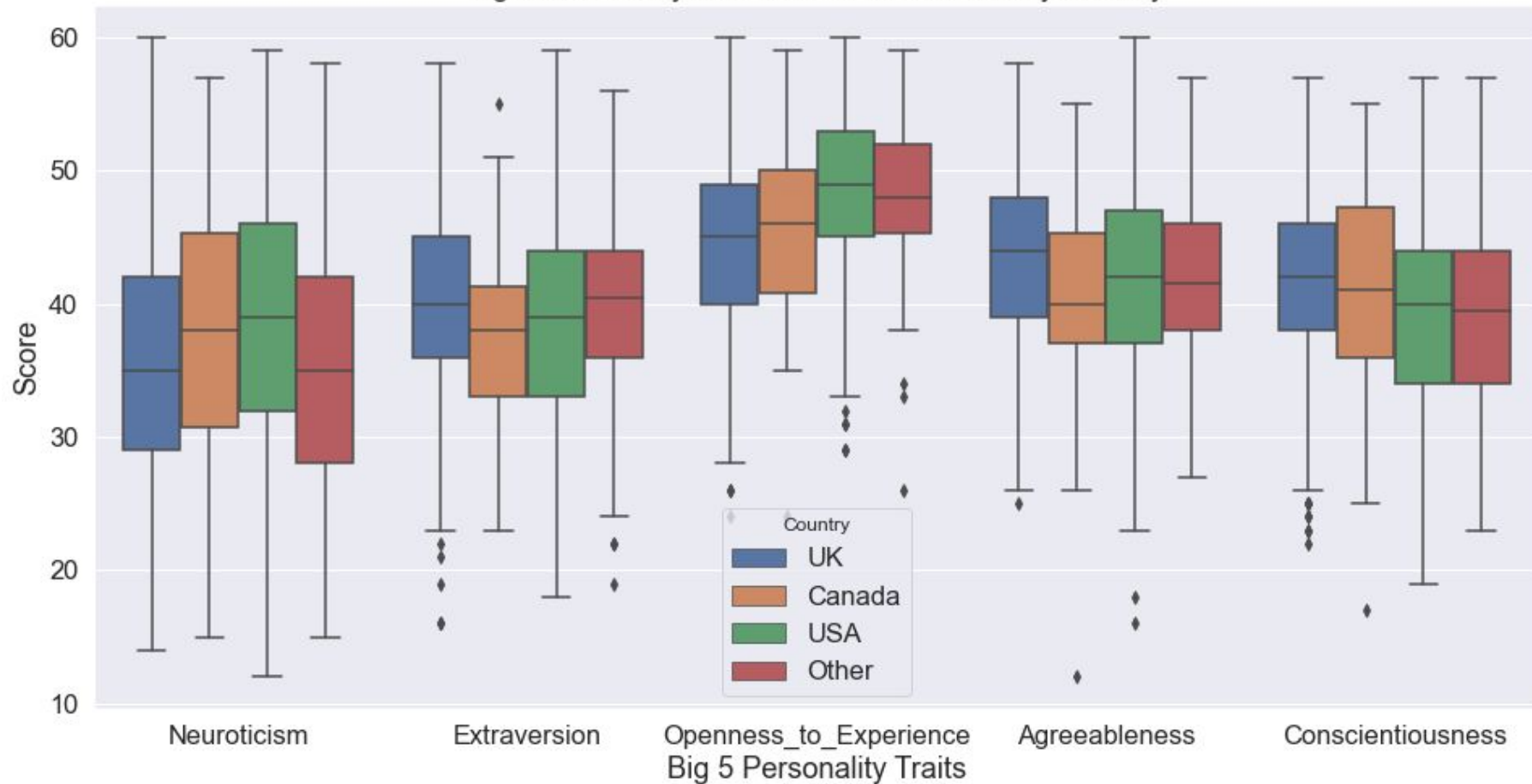
Exploratory Data Analysis



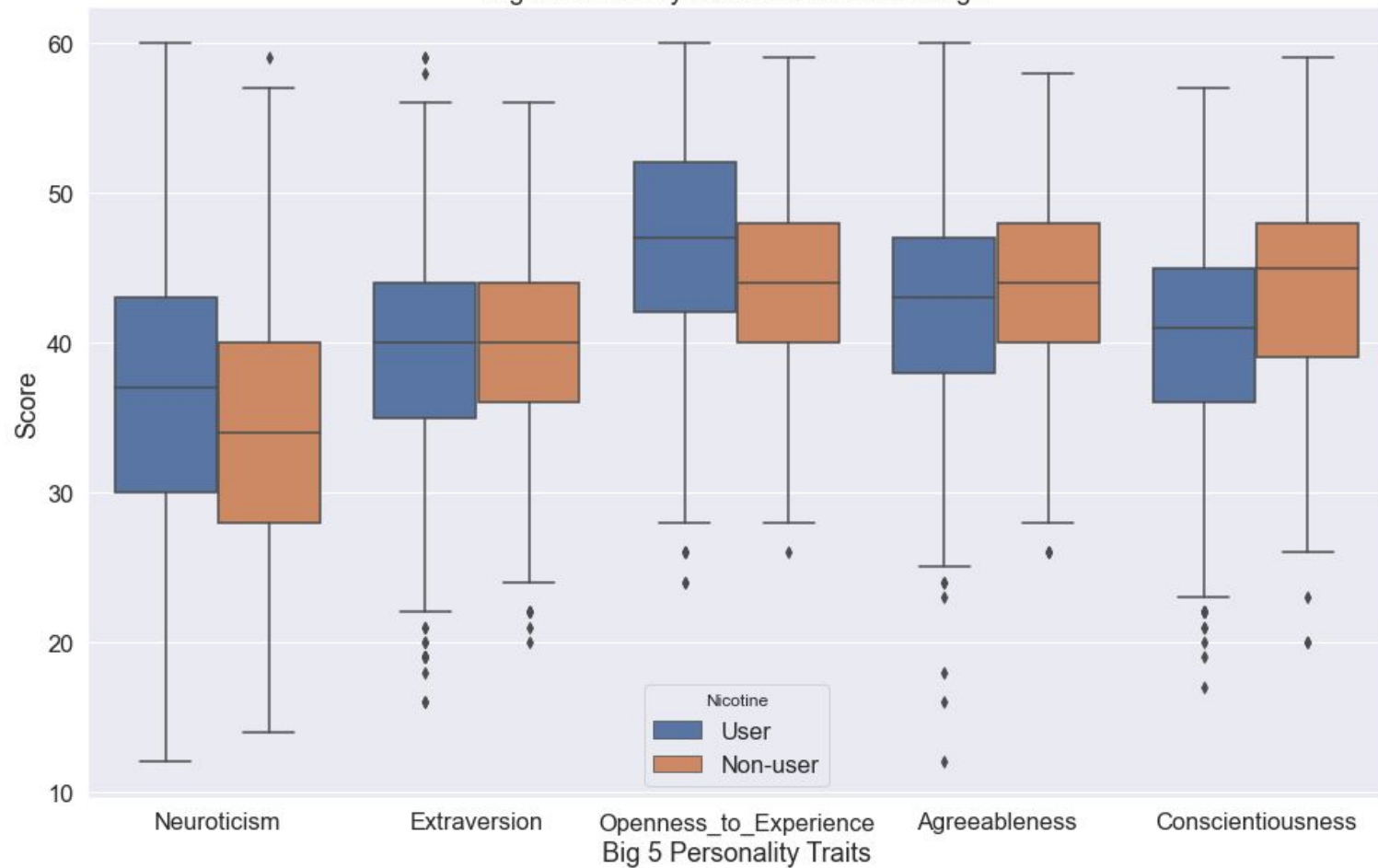
Big 5 Personality Scores in Nicotine Users by Gender



Big 5 Personality Scores of Nicotine Users by Country



Big 5 Personality Scores in Nicotine Usage



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

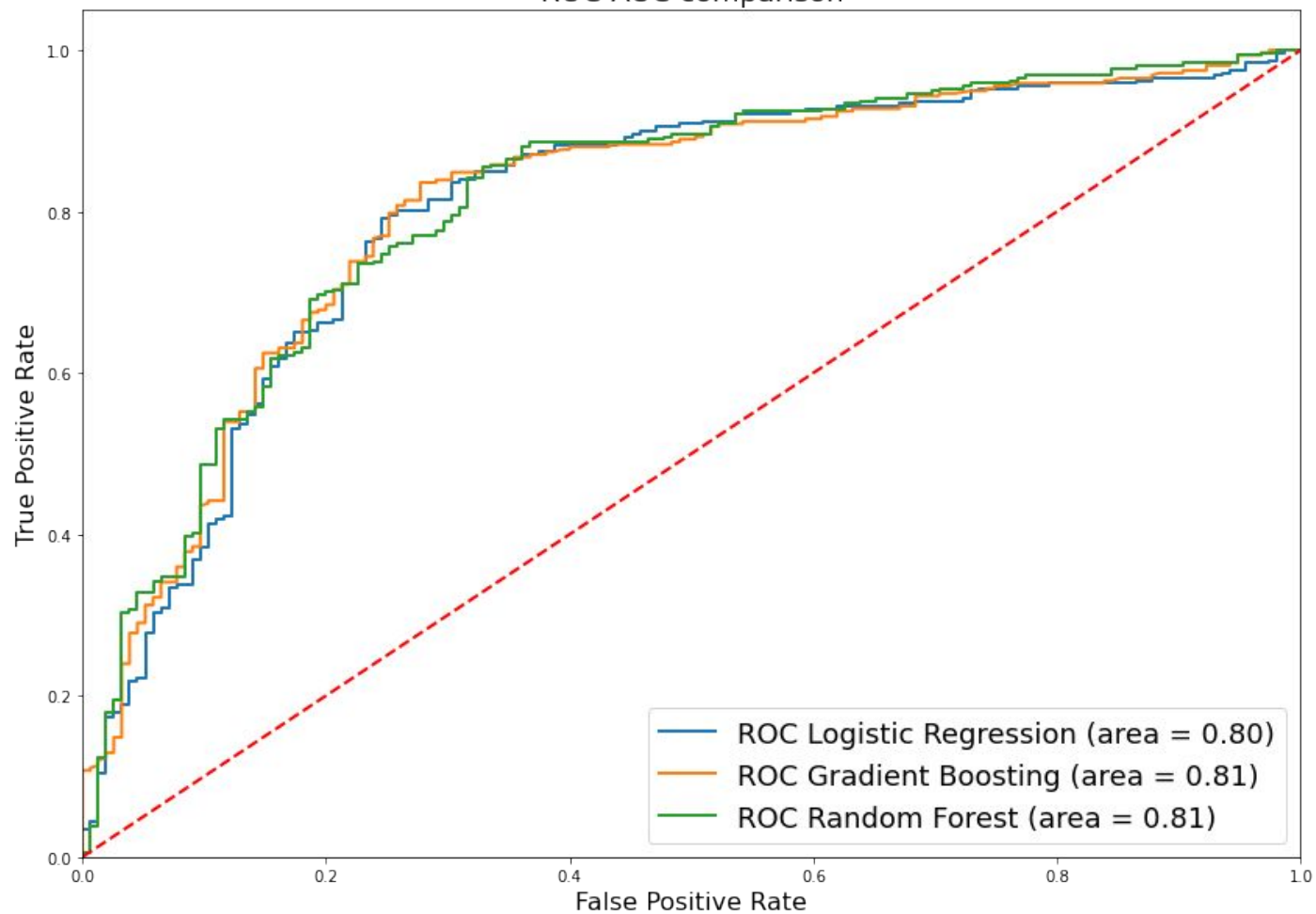
OCEAN Features	VIF Score	Standardized Regression Coefficients	Odds Ratios
Openness to experience	1.1	0.3904	1.0695
Neuroticism	1.4	0.1212	1.0171
Extraversion	1.4	0.0623	1.0123
Agreeableness	1.1	-0.1564	0.9734
Conscientiousness	1.3	-0.3200	0.9476

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

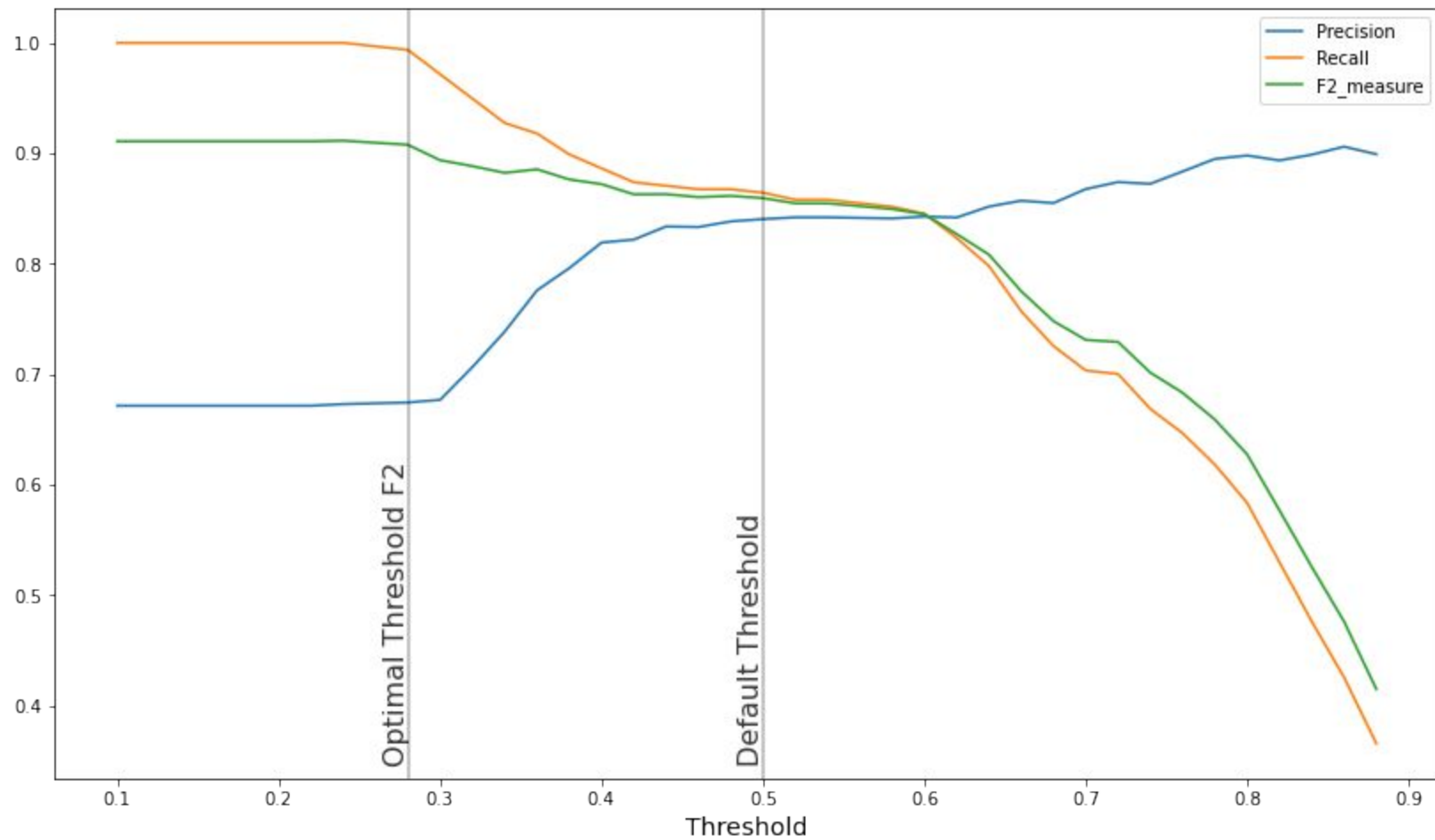
Best Performers

ROC-AUC comparison

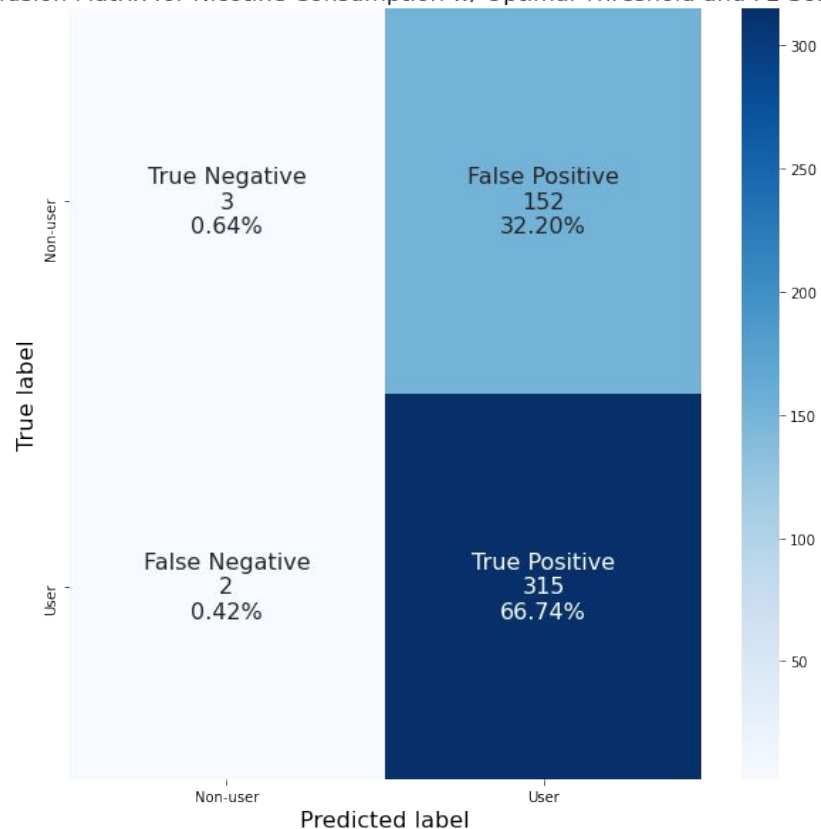


Model Evaluation

Precision-Recall vs. Threshold



RF Confusion Matrix for Nicotine Consumption w/ Optimal Threshold and F2 Score



Accuracy=0.674
Precision=0.675
Recall=0.994
F1 Score=0.804

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/>