Capstone Project Report - Personality Type Nicotine Consumption Risk Assessment

Problem

How likely are you to smoke given your personality profile? According to the Centers for Disease Control and Prevention (CDC) smoking is the leading cause of preventable deaths in the United States. Smoking is attributed to cancer, chronic diseases, and increases risks for other problems. Typically most of the adverse health defects are linked to tobacco in cigarettes, but nicotine is the drug that makes smoking addictive. Nicotine is a psychoactive drug, it promotes release of adrenaline and increases the levels of the neurotransmitter dopamine which makes a person feel good. Irritability, lack of focus and sleep, and spike in appetite are just some of the withdrawal symptoms that make quitting smoking difficult. Ultimately, the aim of this project is to develop a model that can assess a population of people who are currently smokers or are at high risk of picking up the habit of smoking based on their personality profiles.

Context

Studies have shown an individual’s personality profile plays a role in becoming a drug user. In this project different personality features, demographic information, and different drug usage of an individual will be explored to identify how likely an individual is to smoke by predicting a person’s nicotine consumption risk.

The five-factor model of personality states that there are five basic facets of human personality (known as the big five). These traits are neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness (OCEAN). Below are common traits demonstrated by individuals with high and low scores of a given feature:

- **Openness to experience**
    - High score
        - Willingness to try new things
        - Creative or active imagination
        - Broad interests
    - Low score
        - Dislike change
        - Traditional
        - Practical
- **Conscientiousness**
    - High score
        - Organized
        - Detail oriented
        - Persistent
    - Low score
        - Careless
        - Impulsive
        - Unorthodox

- **Extraversion**
  - <u>High score</u>
    - Extraverted
    - Enjoys company
    - Thrill seeker
  - <u>Low score</u>
    - Introverted
    - Quiet
    - Reserved

- **Agreeableness**
  - <u>High score</u>
    - Altruistic
    - Trustworthy
    - Good-natured
  - <u>Low score</u>
    - Selfish
    - Stubborn
    - Uncompassionate

- **Neuroticism**
  - <u>High score</u>
    - Self-conscious
    - Easily stressed
    - Emotionally vulnerable
  - <u>Low score</u>
    - Optimistic
    - Worry free
    - Confident

The Barratt Impulsiveness (BIS-11) and Impulsive Sensation Seeking (ImpSS) are also considered in the dataset. BIS-11 quantifies an individual's tendency to demonstrate impulsive behavior with a higher score signifying constant/frequent impulsive behavior and the opposite for a low score. Very similarly, the ImpSS scale is used to quantify an individual's likelihood to use drugs or other psychoactive substances.

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

Clientele

The model developed may be of use to anti-substance abuse organizations such as the Substance Abuse and Mental Health Services Administration (SAMHSA), American College Health Association (ACHA), American Addiction Centers, etc. As an at risk consumer or consumer identifier, the model can help these orgs cater to these individuals by customizing outreach plans, improving rehab programs and hopefully reducing the number of consumers of an ailment provoking drug. Allowing them to save time and money all while offering more effective programs.

The Data

The features in the dataset include an ID column, education, age, gender, country of residence, ethnicity, 7 personality measurements, and 19 different drugs for 1885 subjects. All values except the personality features are multilevel categorical values.

The age column is not numerical instead its values are binned into different age groups. The classes for all drugs are: "Never used", "Used over a decade ago", "Used in last decade", "Used in last year", "Used in last month", "Used in last week", and "Used in last day". They will be joined to make the drug features binomial. “Never used” and “Used over a decade ago” will be considered as a “User” and the rest as “Non-user”.

18 central nervous system psychoactive drugs and 1 fake drug are considered in the dataset. All drugs are considered abusable psychoactive drugs, whether legal or not. These drugs are used to provoke different mental effects such as pleasure, increased focus, or behavioral changes aside from just treating a disease or ailment. The drugs consist of:

- Alcohol
- Amphetamines
- Amyl nitrite
- Benzodiazepine
- Caffeine
- Cannabis
- Chocolate
- Cocaine
- Crack
- Ecstasy
- Heroin
- Ketamine
- Legal highs
- Lysergic acid diethylamide (LSD)
- Methadone
- Mushrooms
- Nicotine
- Semeron
- Volatile substance abuse (VSA)

Exploratory Data Analysis

In the exploratory data analysis portion of the project the features of the data were visually analyzed to see if any trends or correlations could be seen before conducting feature selection. A few questions about the data were explored and delved into further if the initial visual plot revealed any outstanding observations. The following diagrams show the most prominent findings.

Which drugs were the most consumed?

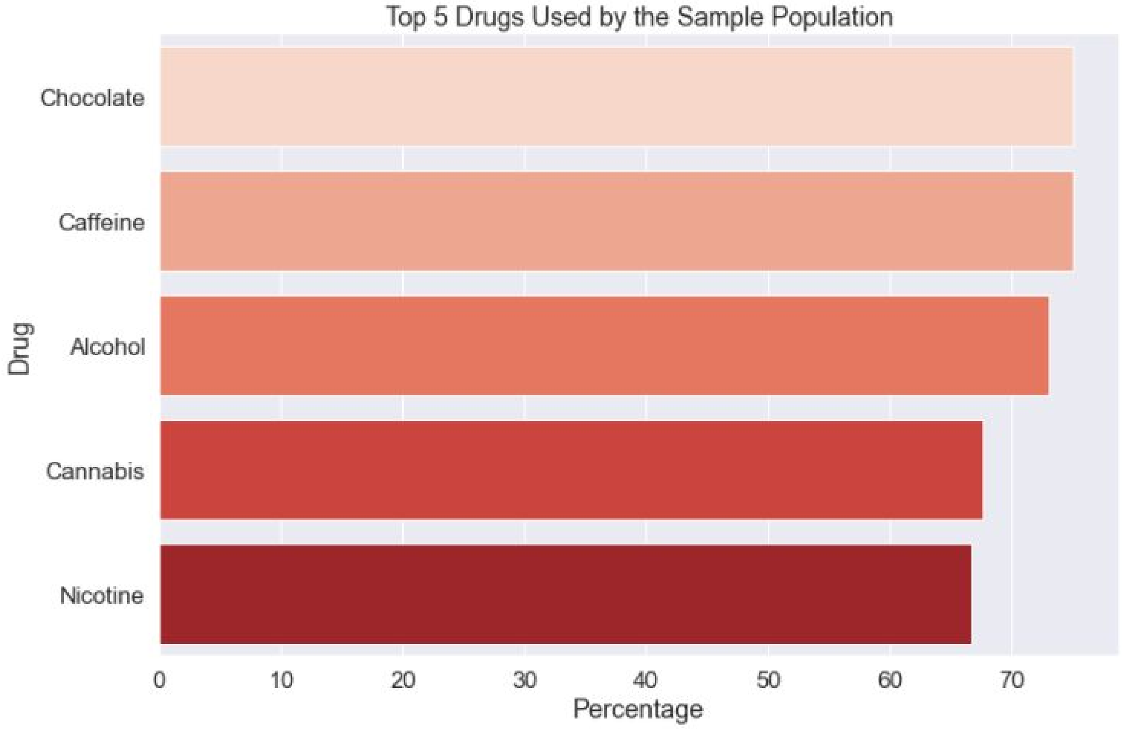

Figure 1: Top five drugs consumed by the sample population

Among the 1885 participants, the top five drugs consumed were chocolate, caffeine, alcohol, cannabis and nicotine in descending order shown in Fig. 1. About 65% of the sample population consumed nicotine. Any of these individuals could be at high risk of suffering from any of the aforementioned health complications.

Who were more likely to smoke, men or women?

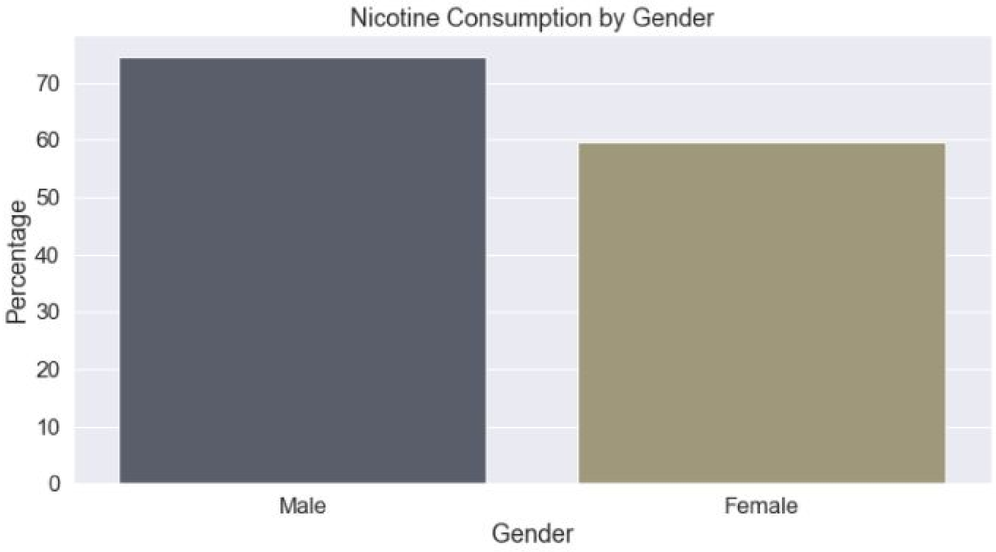

Figure 2: Percentage of smokers by gender

Roughly half of the participants of the study were women, however, men were more likely to be smokers with over 70% of nicotine consumers being male to only 60% being female (refer to Fig. 2). Given the significant difference derived from a chi-square test, the variability of the OCEAN traits between genders was explored.

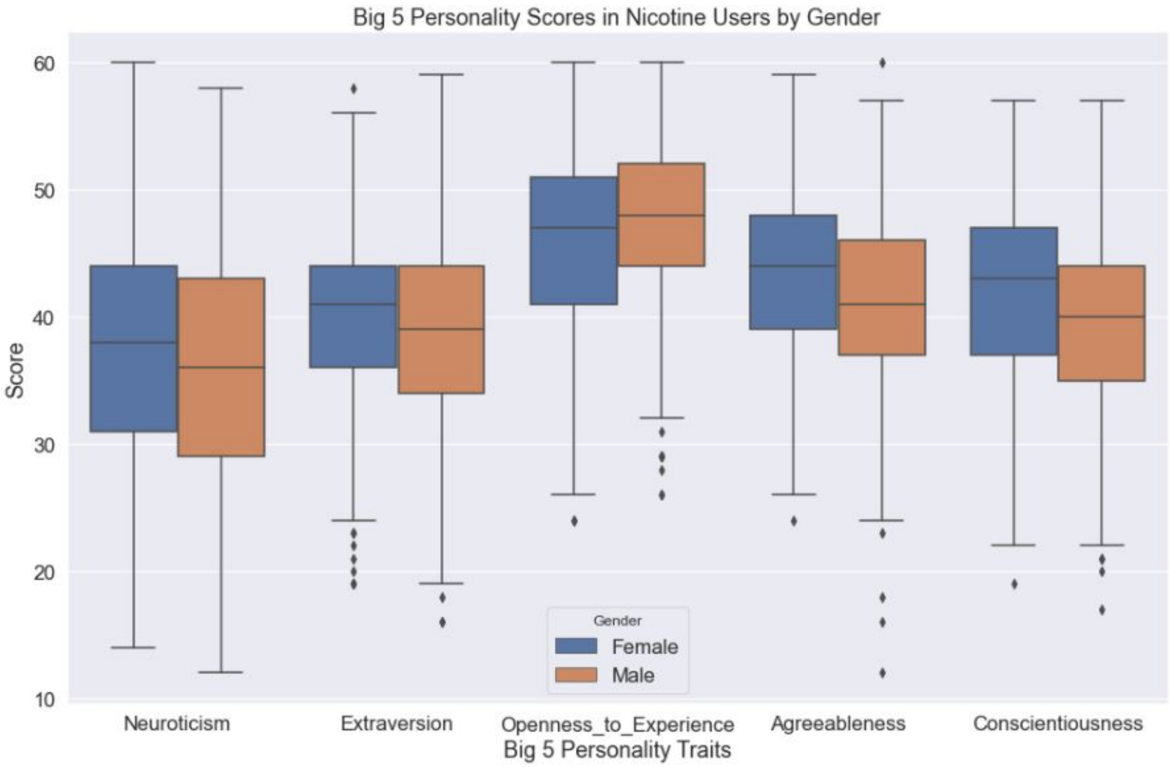

Figure 3: Variability of OCEAN traits by gender

The OCEAN personality traits were explored between men and women in an attempt to discover if men exhibited certain traits more than women. After plotting a box plot of the big 5 personality traits depicted in Fig. 3, it was apparent that nicotine users that were men scored significantly higher for the openness to experience trait and women had higher scores of the agreeableness and conscientiousness traits. T-testing determined that there were significant differences between genders and each OCEAN trait.

Were you more likely to be a smoker based on the country you live in?

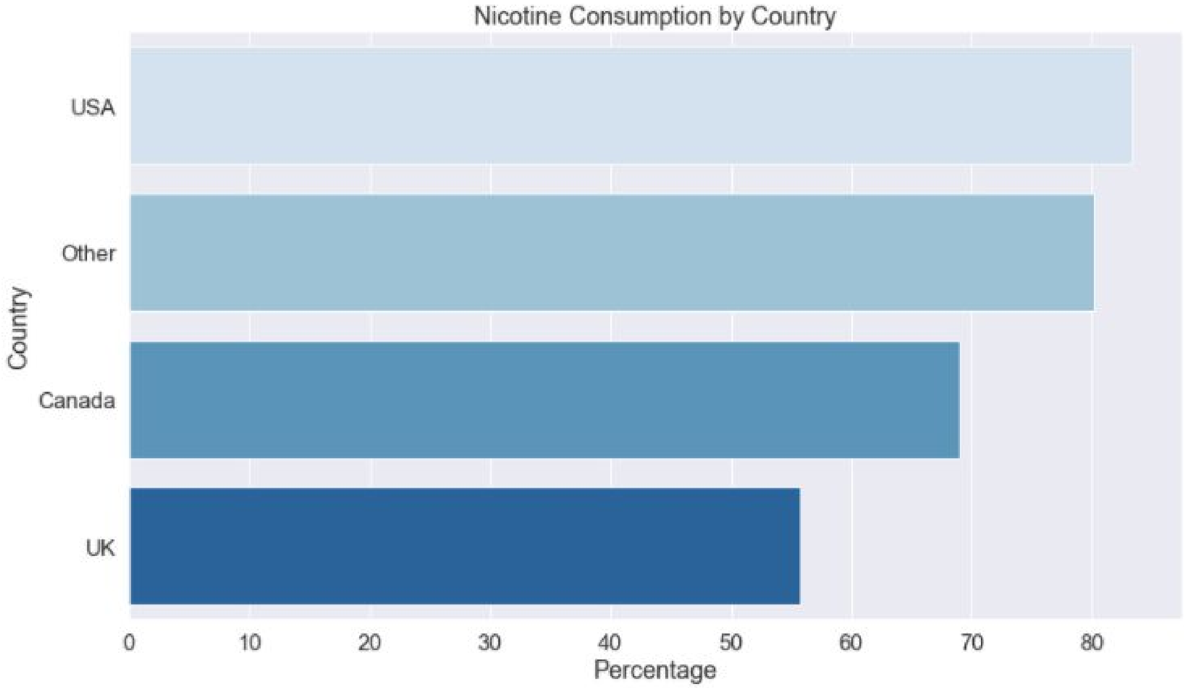

Figure 4: Percentage of smokers based on country of residence

From Figure 4, it is evident that of all Americans that participated in the study an outstanding 85% of them were smokers. The United Kingdom had the least amount of smokers with a little over half of them being smokers. Chi-square testing suggested that this was not due to chance (P-value virtually equal to 0), therefore the big five traits between countries was further explored.

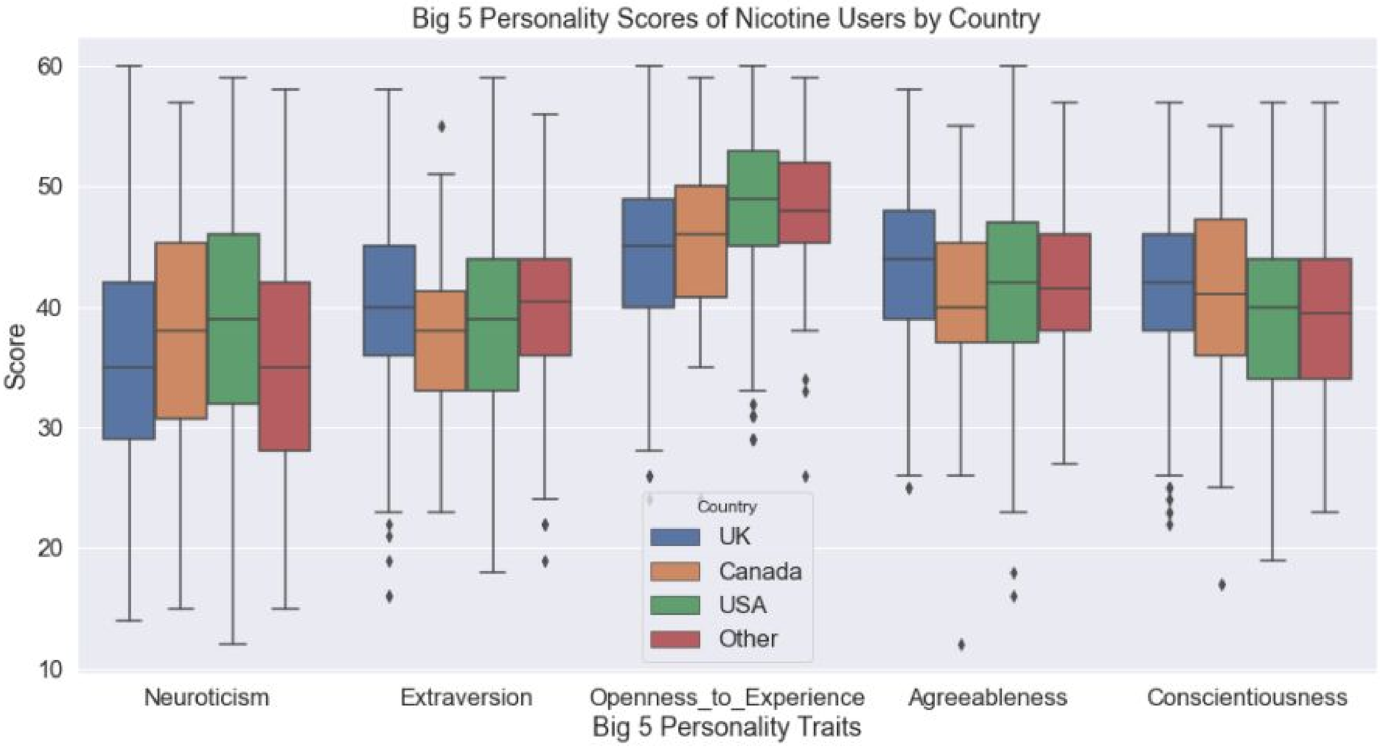

Figure 5: Variability of OCEAN traits by country

Americans scored higher for the openness to experience and neuroticism traits. Openness to experience seems to be the most common trait found in smokers based on these findings and the gender findings. No other obvious observations were extracted from the following box plots shown in Fig. 5. One way ANOVA testing resulted in P-values well below 0.05 meaning that there is a significant difference between smokers who have certain OCEAN traits and their country of origin.

Which big five OCEAN personality traits are the most common in people who smoke?

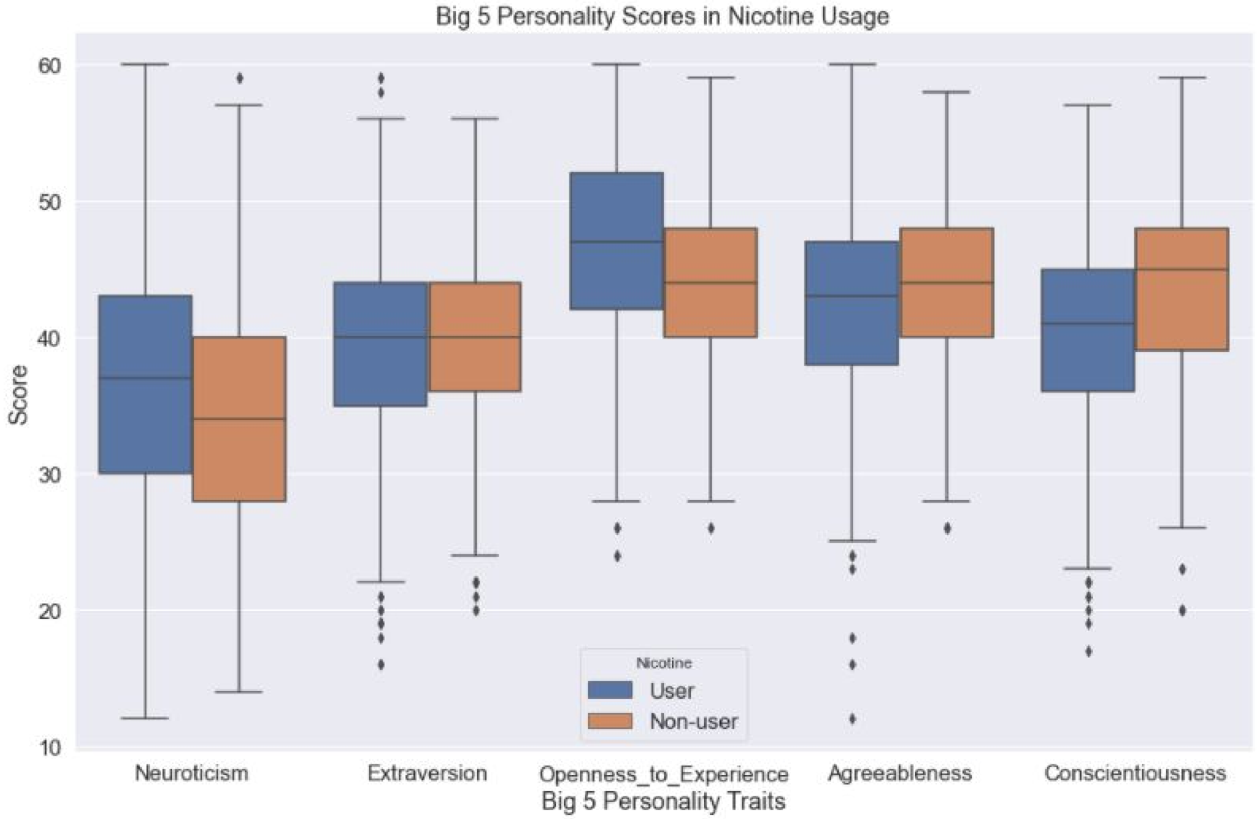

Figure 6: Variability of OCEAN traits among smokers and non-smokers

According to Figure 6, smokers had higher scores for the neuroticism and openness to experience traits and non-smokers had higher scores for the agreeableness and conscientiousness traits. Extraversion demonstrated no statistical difference between smokers and non-smokers, all other traits resulted in a P-value well below 0.05. These findings support the previous findings that people with the openness to experience trait are likely to be smokers and therefore likely to be at greater risk for nicotine consumption.

Which features have the strongest correlation to nicotine usage?

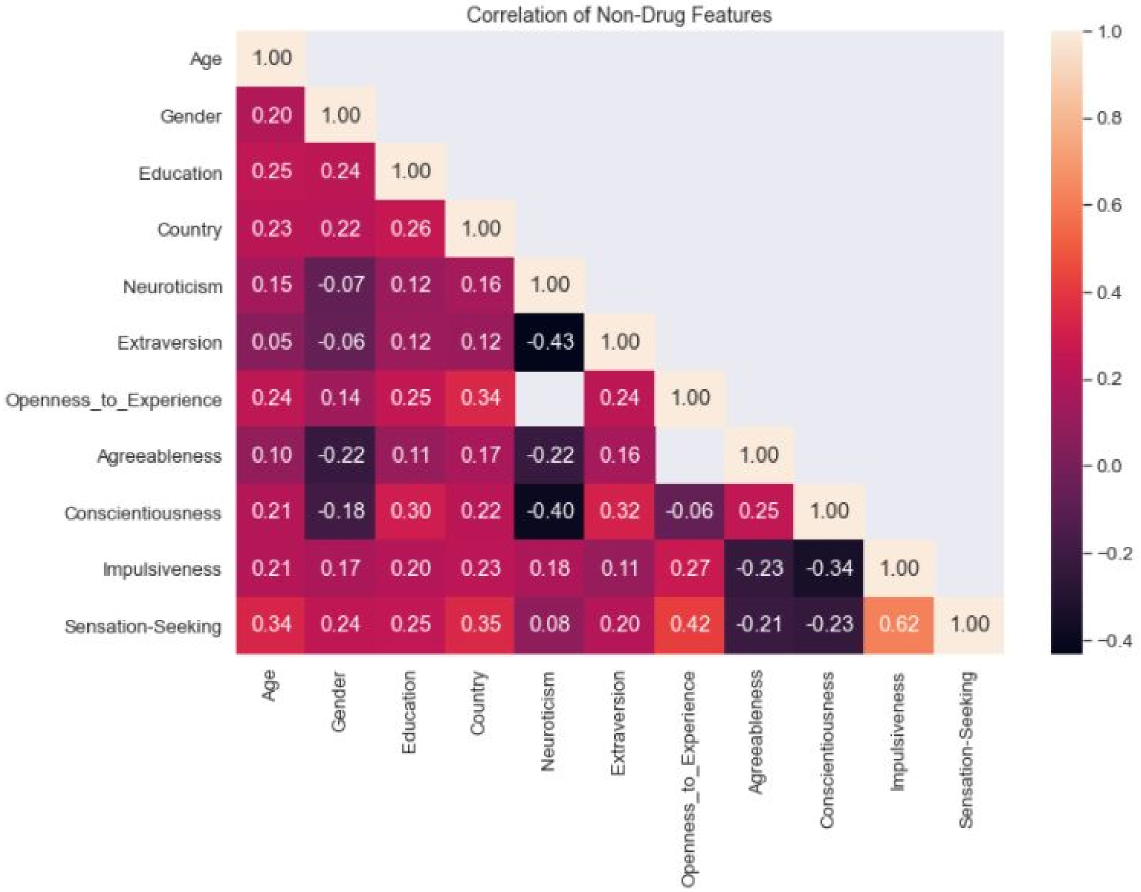

Figure 7: Statistical significant correlation features among non-drug features

All drug features were more correlated with other drug features than personality features, therefore the statistical significant correlations among the other features were explored (Fig. 7). The most prominent correlation from these features is the correlation between impulsivity and sensation-seeking traits. Openness to experience also had a strong correlation with the sensation-seeking trait. Among the OCEAN traits all were statistically correlated with each other except for openness to experience with neuroticism and agreeableness.

Feature Importance

All categorical values were encoded to prepare the data for modeling using pandas ‘get_dummies’ method with the ‘drop_first’ parameter set to true to reduce collinearity among the features. Then, the data was standardized to ensure all features were on the same scale using scikit- learn's ‘StandardScaler’. Finally, the data was split into x and y test and train variables for modeling, but before models were created the continuous features, in particular the OCEAN traits, were checked for multicollinearity and how each trait would affect the prediction of nicotine consumption through their odds ratios.

Table 1: VIF scores, regression coefficients, and odds ratios of OCEAN traits

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

Checking for multicollinearity among the personality traits was accomplished by using the ‘variance_inflation_score’ function from the statsmodels package. None of the VIF scores, shown in Table 1, were above the conventional value of 5 indicating little to no multicollinearity among these traits.

After running a logistic regression on the standardized features using statsmodels. The traits that attribute to nicotine consumption are openness to experience, neuroticism and extraversion. The finding supports our previous discovery of the openness to experience trait being present among different groups of smokers.

The personality features with the greatest percent increases on average per unit increase were neuroticism, extraversion and the openness to experience traits. To put simply, if a person scores high for the neuroticism and extraversion trait they have an increased odds on average of about 1% of being a nicotine consumer and a 7% increase for having the openness to experience trait. As suggested from the EDA portion of the project there is a decrease in odds if a person has a high agreeableness and conscientiousness score.

Note: *Sensation-seeking and impulsiveness traits were not considered as these two traits pertain to tests that are commonly used to assess drug consumption in clinical studies and were only available in their standardized forms making it difficult to assess their odds ratios per unit increase.*

Modeling

The aim of this project is to predict whether someone is a smoker based on different personality traits. Other drug usage features were also included as part of the training features to increase the training data's robustness. As this a classification problem six different models were considered:

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

Table 2: ROC-AUC and Brier Score of six different models

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

The models were evaluated through grid search cross validation to determine the best performing models through the grading criteria of 'ROC-AUC' as well as their brier scores. Table 2 depicts the performances of each model. Given the results, the logistic regression, random forest and gradient boosting models were further explored by using each to plot their ROC-AUC and derive a confusion matrix.

The best parameter for the logistic regression model was an inverse of regularization strength value of 0.01. For random forest, the parameters of max depth = 6, max features as 'auto', and number of estimators = 50 yielded the best results. For gradient boosting, the best parameters were learning rate = 0.1, max depth = 3 and number of estimators = 45.

Model Evaluation

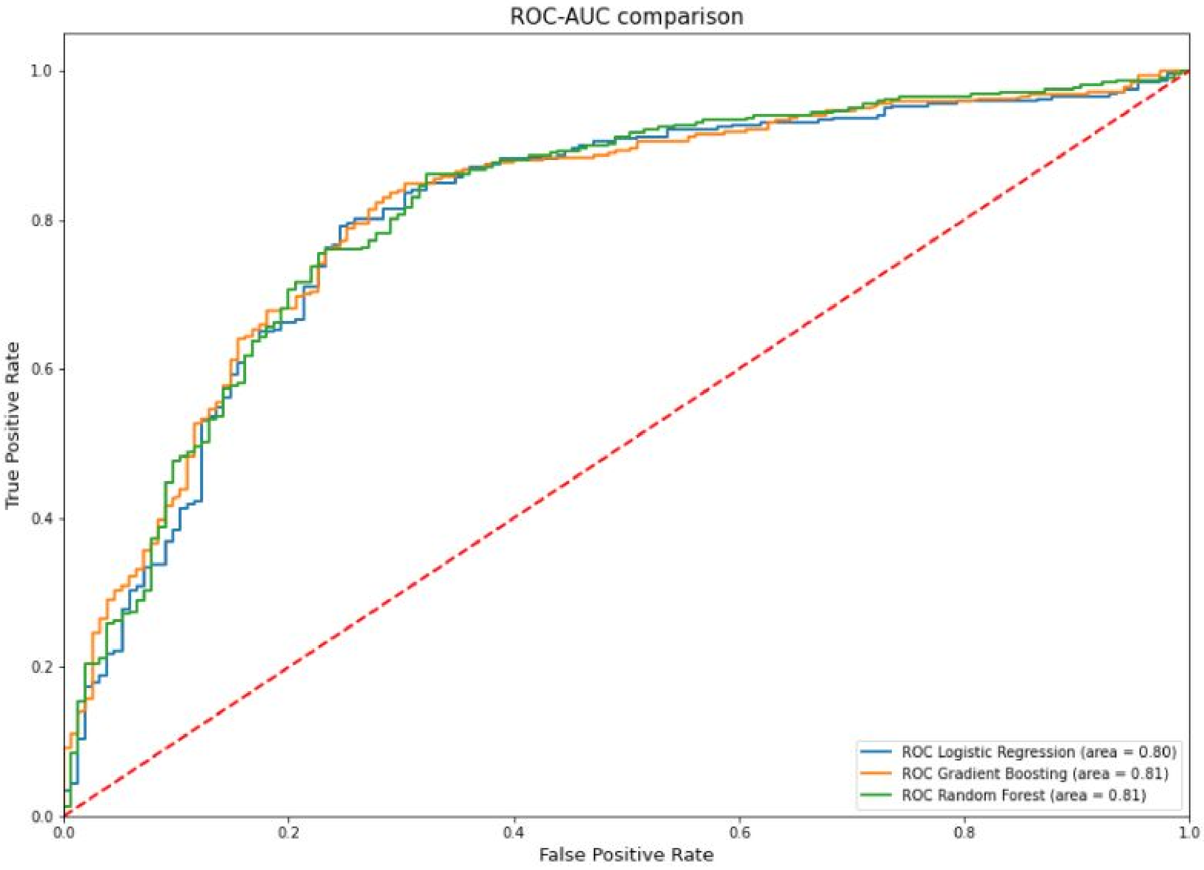

Figure 8: ROC-AUC comparisons between best performing models

From the ROC-AUC curves (Fig. 8) it is difficult to assess which model stands out from the rest. Considering the business problem in order to help at risk individuals or consumers of nicotine, it is of greater benefit to reduce the number of predictions of actual smokers as not smokers (false negatives). If the model were to predict non-smokers as smokers (false positives) it would not be that big of a deal as those individuals probably wouldn't pick up the habit of smoking based on the data. Hence, the confusion matrices were created for the random forest classifier and gradient boosting classifier to determine the model with the highest recall score.

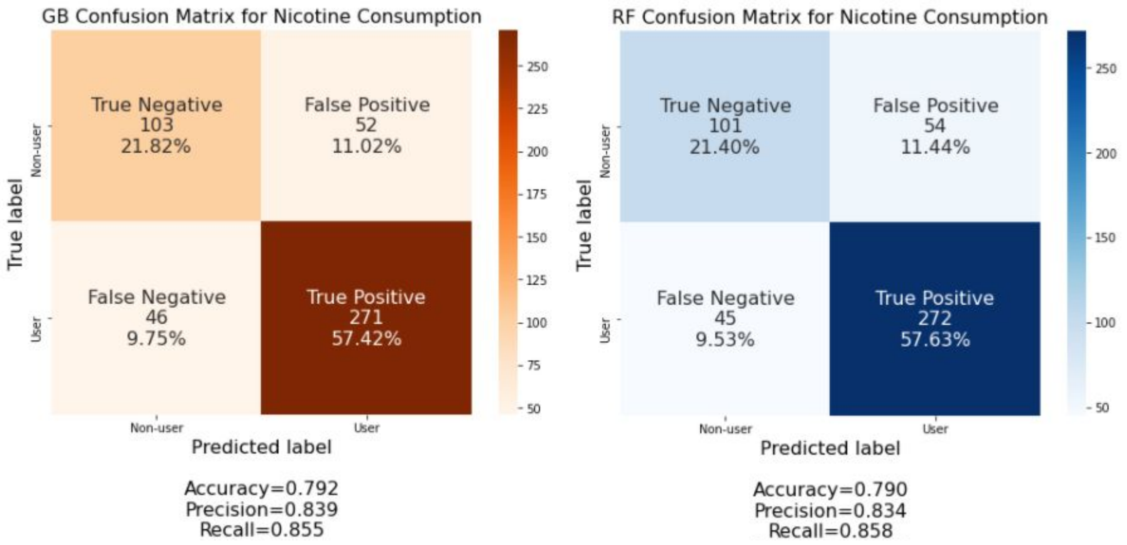

Figure 9: Confusion matrices for gradient boosting and random forest classifiers

Both classifiers resulted in nearly similar results (Fig. 9). Yet, the random forest classifier had one less false negative prediction than the gradient boosting one. Therefore, the random forest classifier was selected to be the model for production.

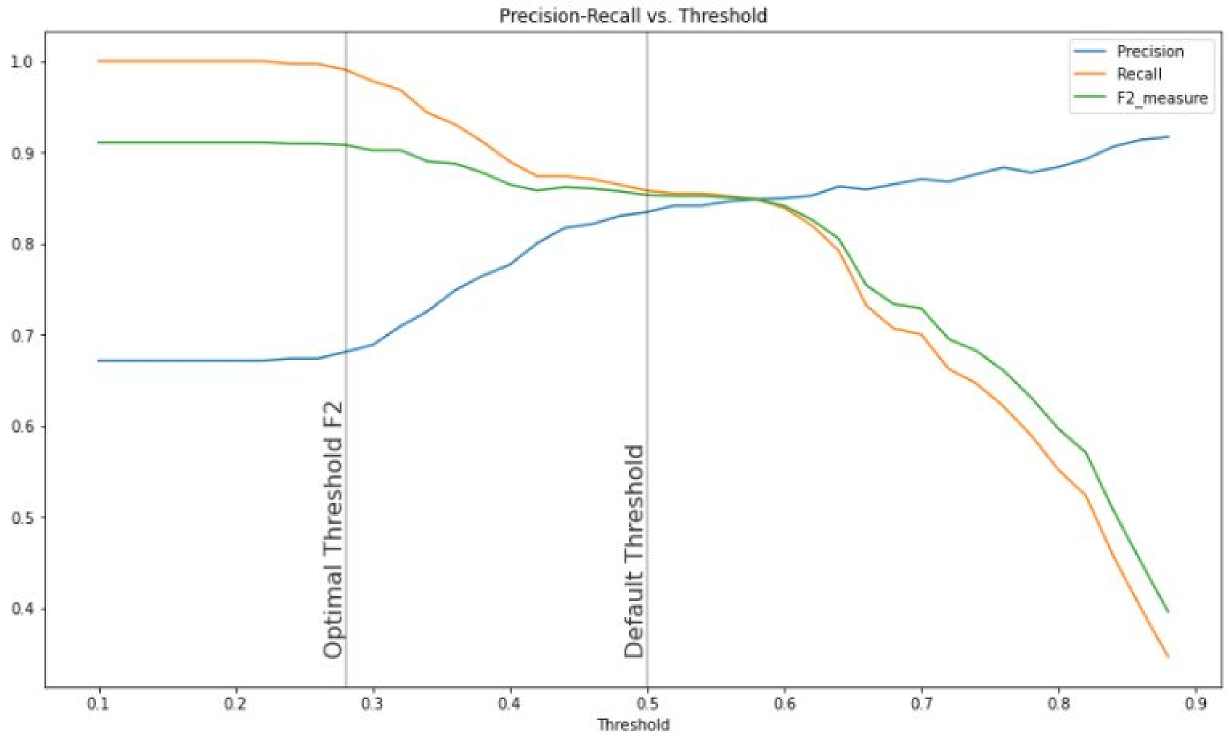

Figure 10: F-beta score, precision and recall of random forest model at different thresholds

To enhance its performance, a threshold value for classifications was determined by looping over different thresholds and plotting the model's precision, recall and f-beta score (with a beta value of 2 to favor recall) which are depicted by Fig. 10.

Given that the goal was to reduce the amount of false negatives our model produces, the graph suggested using a threshold value of about 0.28. This threshold produced a model that results in almost no false negatives meaning that the model was successful in correctly identifying people with high nicotine consumption risk, at the cost of including people who are not at high consumption risk. A final confusion matrix was developed (Fig. 11) to see the classifications of the model with this new threshold.

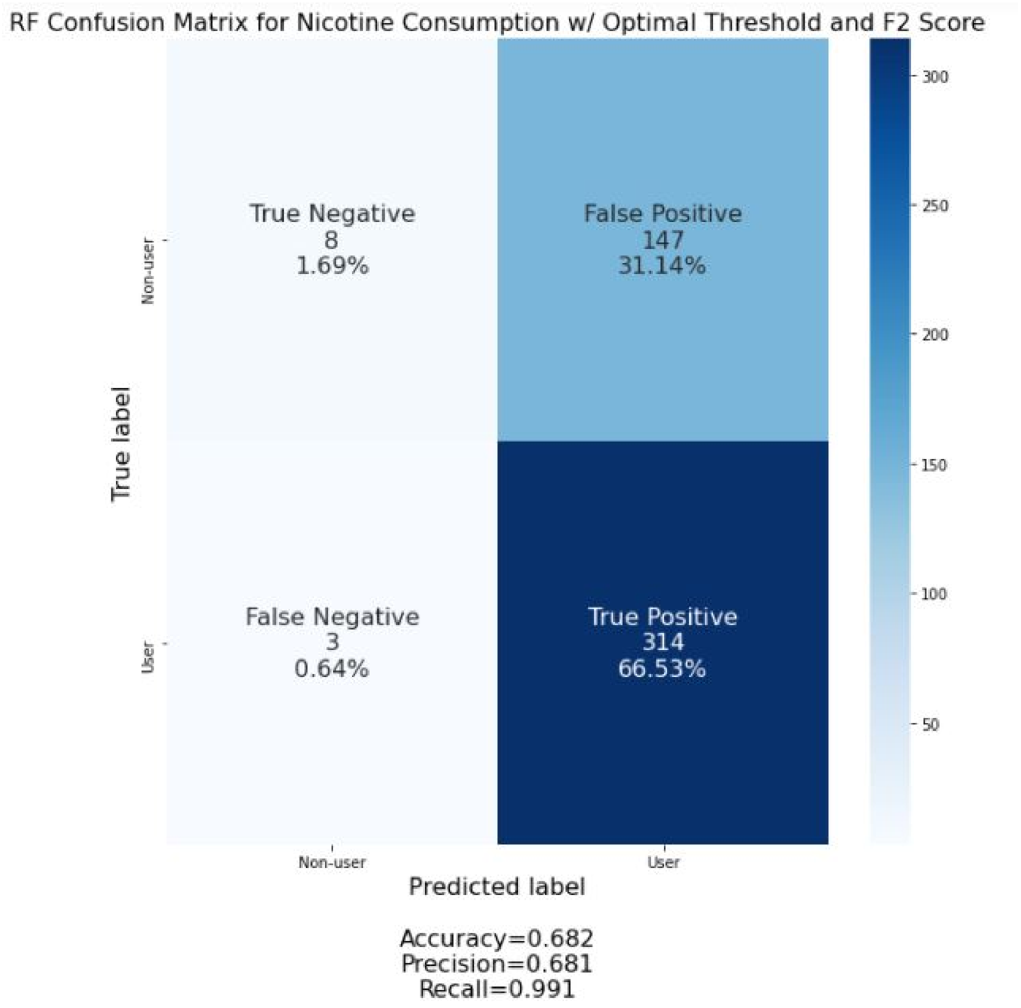

Figure 11: Confusion matrix for optimal random forest model

Conclusion

A personality profile assessment can give insight to an individual's drug consumption, particularly the big five personality, BIS-11 and ImpSS assessments. All personality traits evaluated demonstrated significant correlation with nicotine consumption except the extraversion trait. Anti-substance organizations can create a survey to gather the data needed to use the model. This way the orgs can save resources focusing on helping the population at highest risk.

An interesting finding was the variability of personality traits between gender and people from different countries. Women exhibited higher scores for agreeableness and conscientiousness, traits indicative of non-smokers. Whereas men showed to be generally more open to experiences, the trait most prominent within smokers. What's more the openness to experience trait along with neuroticism (the other big five personality trait associated with smokers) trait, were generally scored higher on by people in the U.S. which happened to be the country with the highest percent of smokers in the sample population.

To further improve model performance and validate the findings, more data should be gathered such as:

- demographic, GDP and life expectancy of a country
- other general characteristics of the individuals who participated in the study
- a substantial amount of more participants

Overall, a quick individual profile can be assessed from the data exploration. American men who are open to new experiences, are easily stressed and emotionally vulnerable are the population group that are at highest risk for nicotine consumption. On the other end of the spectrum are driven, altruistic women from the UK.

References

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/