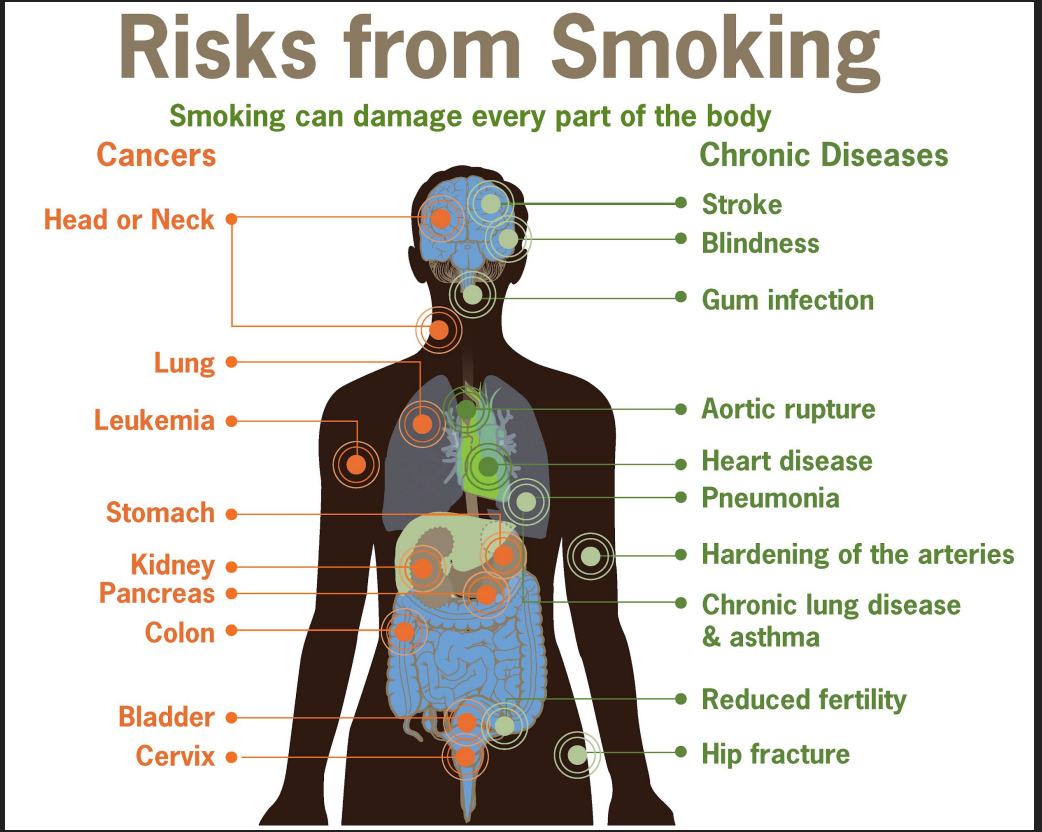
# Personality Type Nicotine Consumption Risk Assessment

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

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

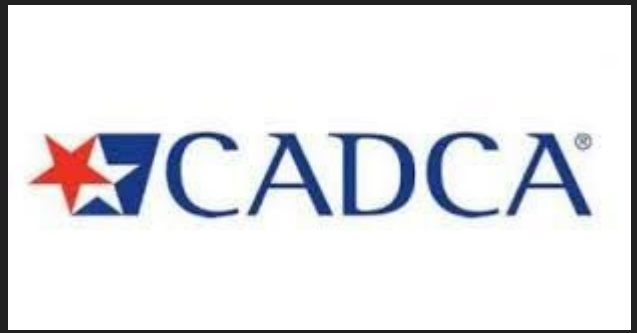
# The Problem

- Smoking is attributed to several forms of cancers and chronic diseases
- Smoking is the leading cause of preventable death in the US, killing approximately 480,000 Americans each year
- Smoking is very addictive




American
Addiction Centers

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

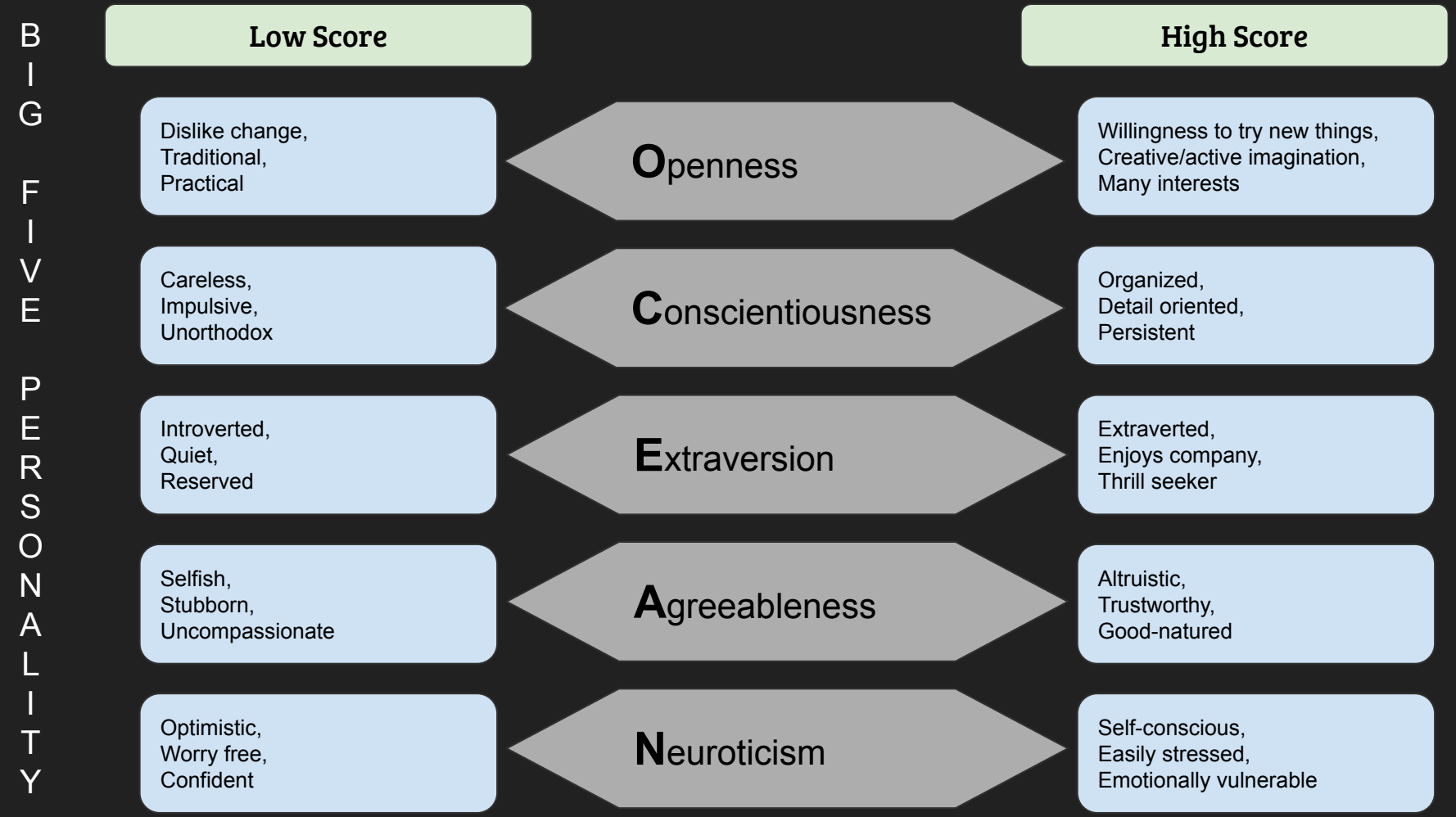



# The Data

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.

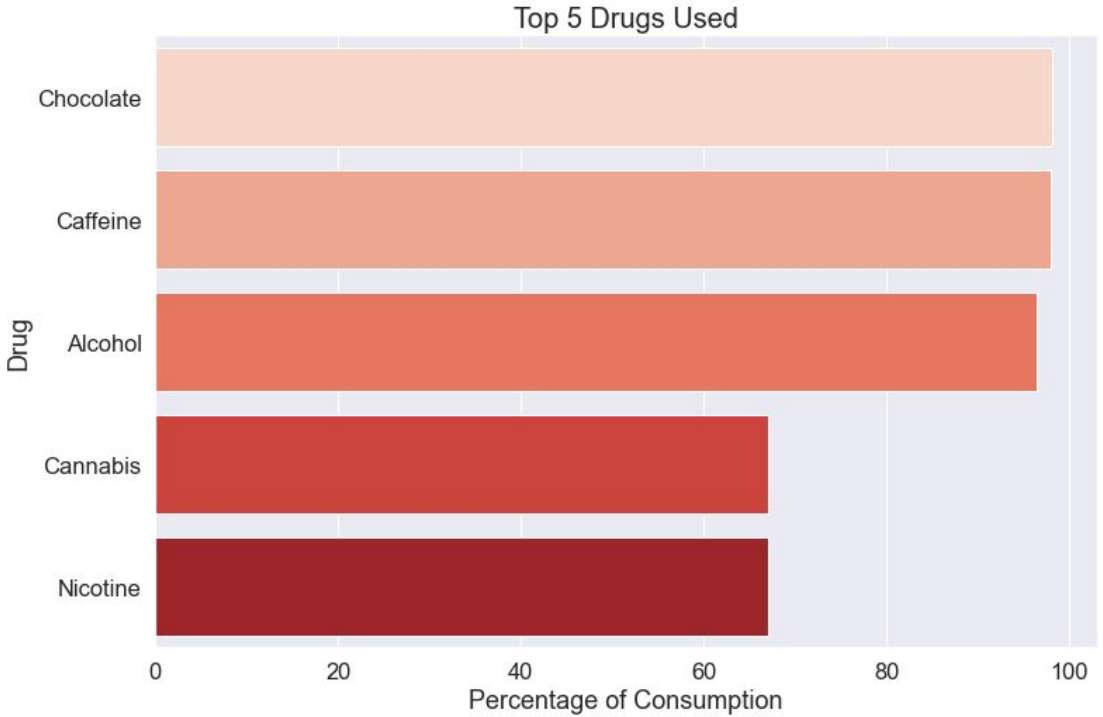

BIG FIVE PERSONALITY
Low Score
High Score
Openness
Dislike change, Traditional, Practical
Willingness to try new things, Creative/active imagination, Many interests
Conscientiousness
Careless, Impulsive, Unorthodox
Organized, Detail oriented, Persistent
Extraversion
Introverted, Quiet, Reserved
Extraverted,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

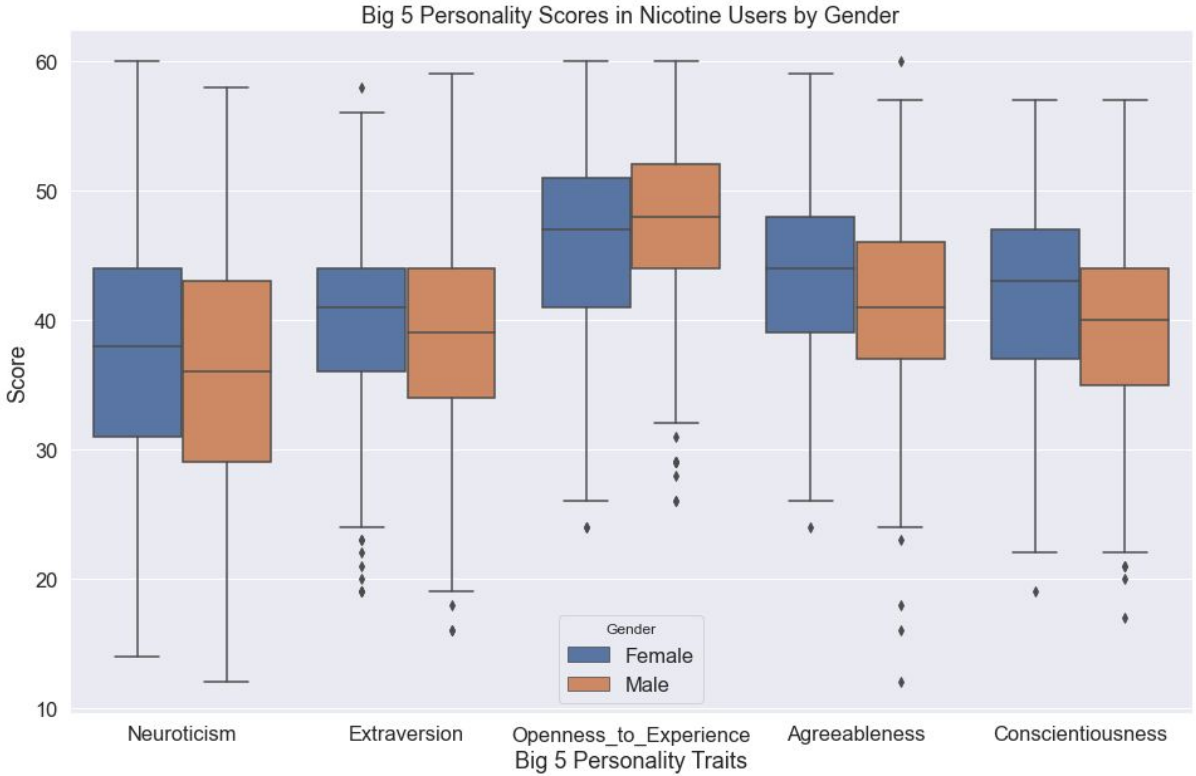
# Disclaimers

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

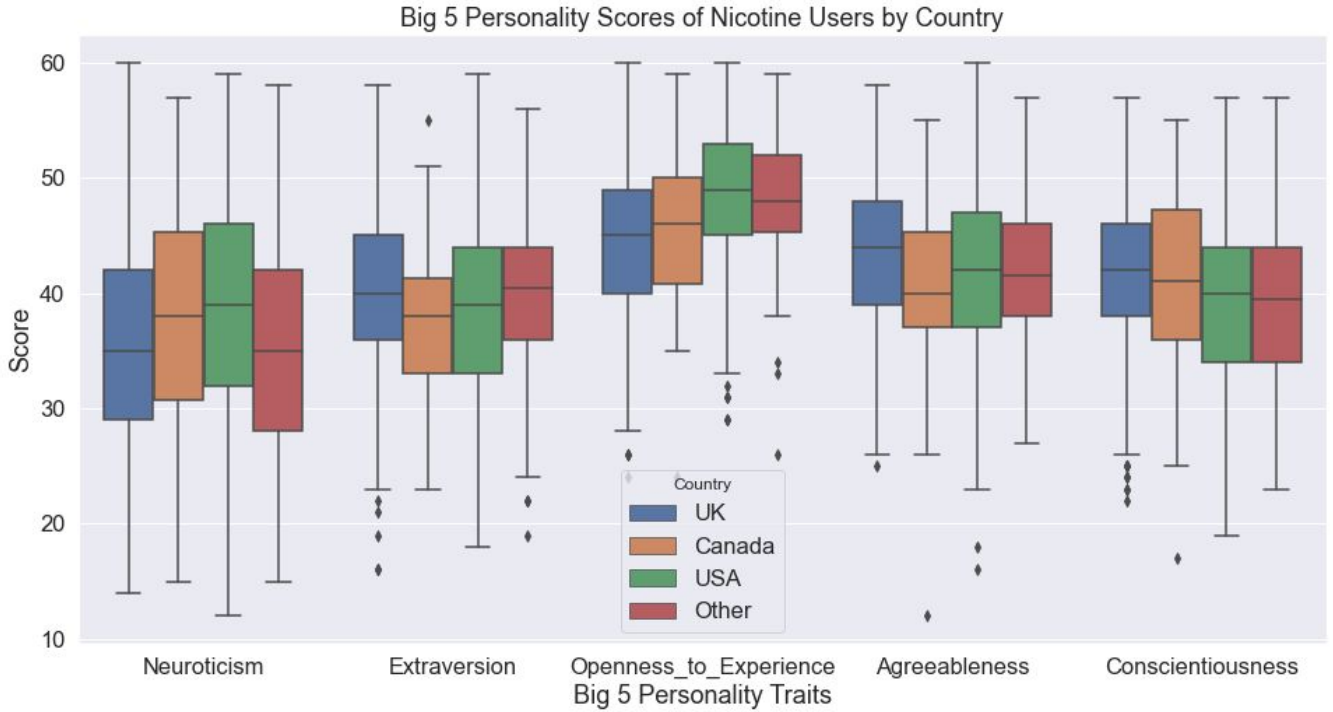
- Nicotine was the 5th most consumed drug ~ 65% of the population
- 1225 out of 1885 participants potentially at risk of health ailments

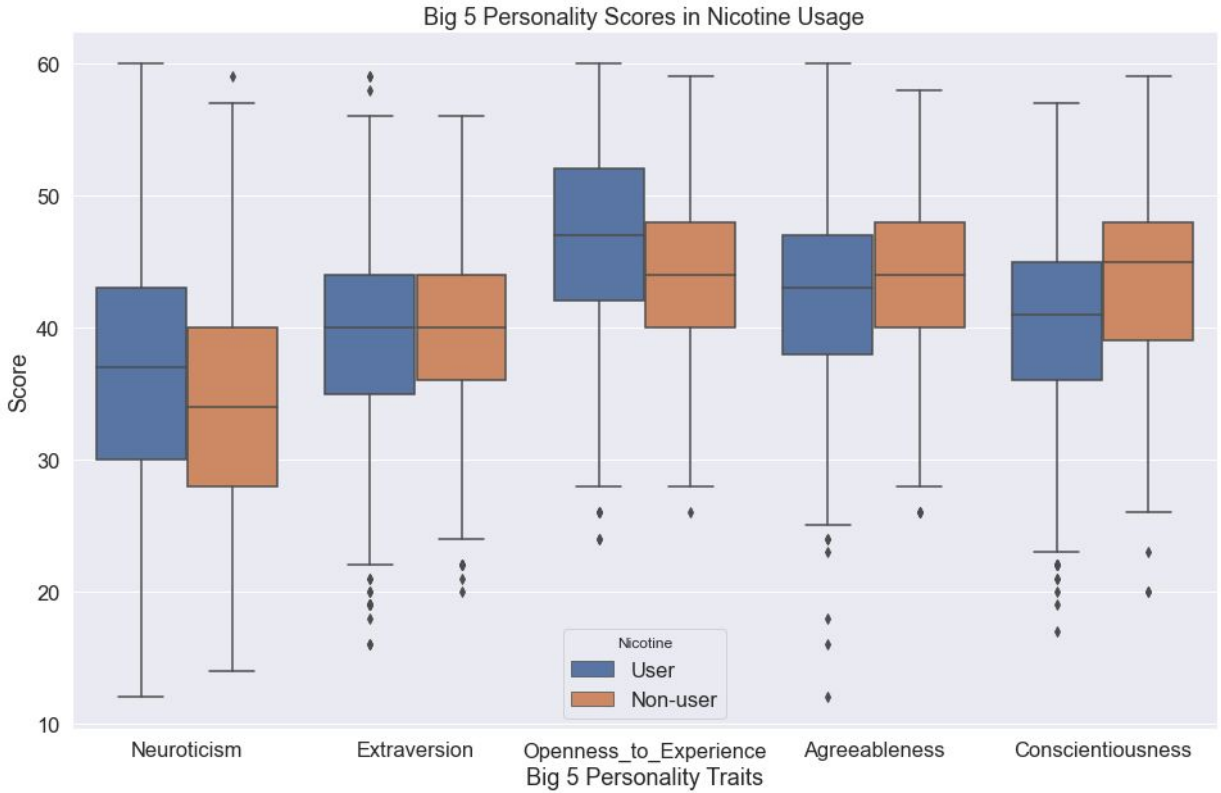





- Openness to experience scores were high for both male and female smokers

- Similar pattern continued; users score high for openness to experience
- Americans in particular held the highest scores






- Users indeed have high openness to experience scores but also score high for neuroticism
- Non-users have relatively higher agreeableness and conscientiousness scores than users

# 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

- VIF scores indicate OCEAN traits did not exhibit multicollinearity
- The features with a positive correlation to nicotine usage were openness to experience, neuroticism and extraversion

| 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

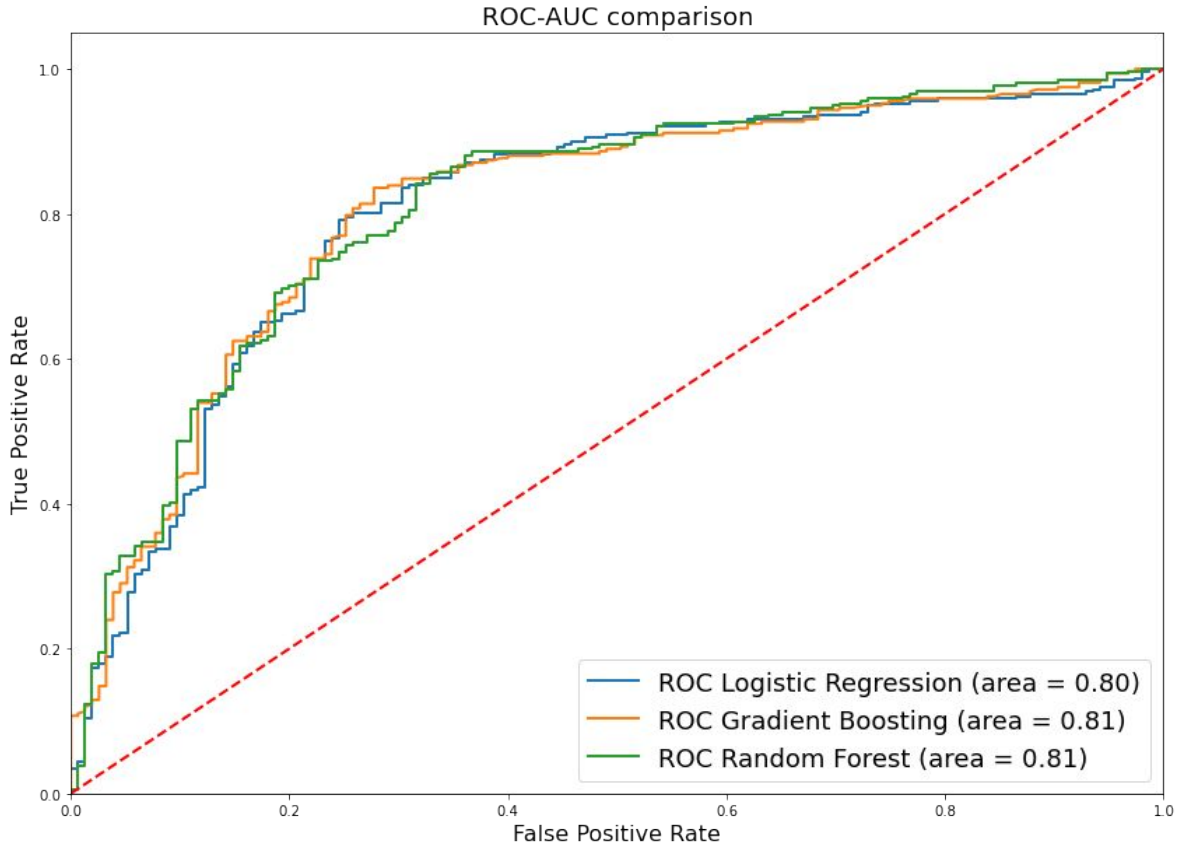
The models with the best performance in grid search cross validation were:
***Logistic Regression, Random Forest and Gradient Boosting***

| 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

- From the ROC-AUC curves it is apparent that Random Forest and Gradient Boosting slightly outperformed Logistic Regression

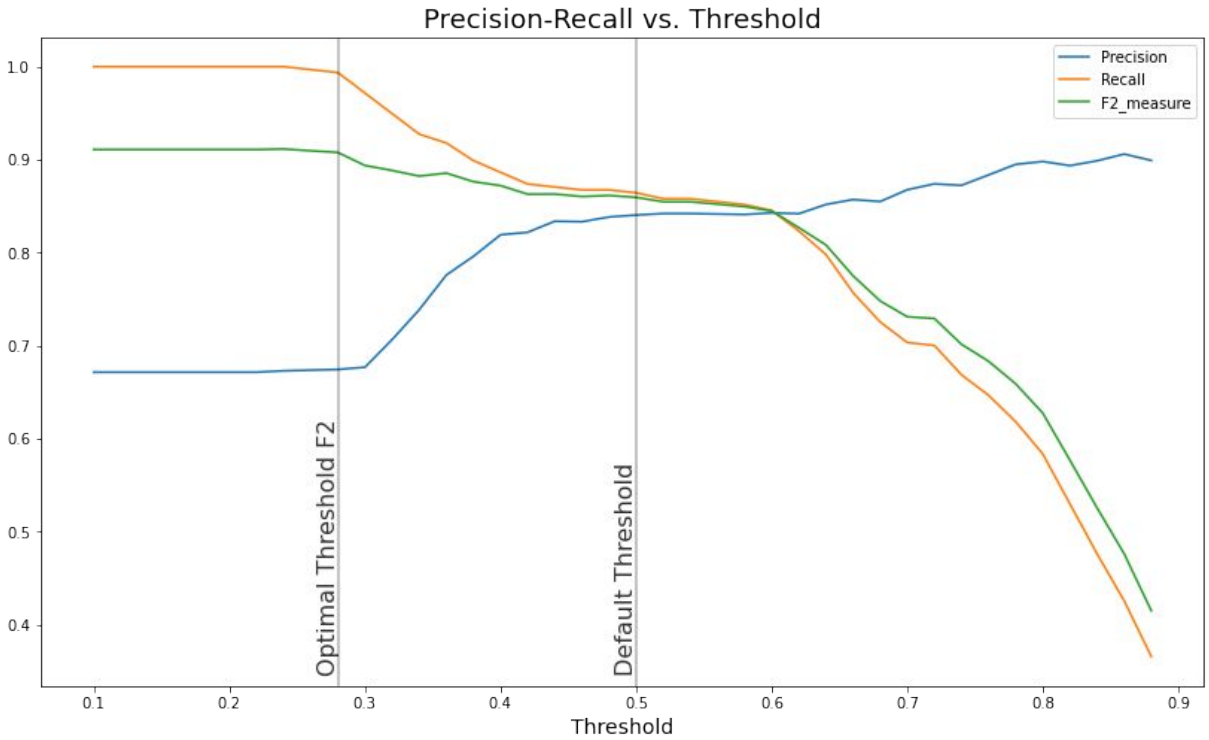
- The decision was made to move forward with the Random Forest Classifier

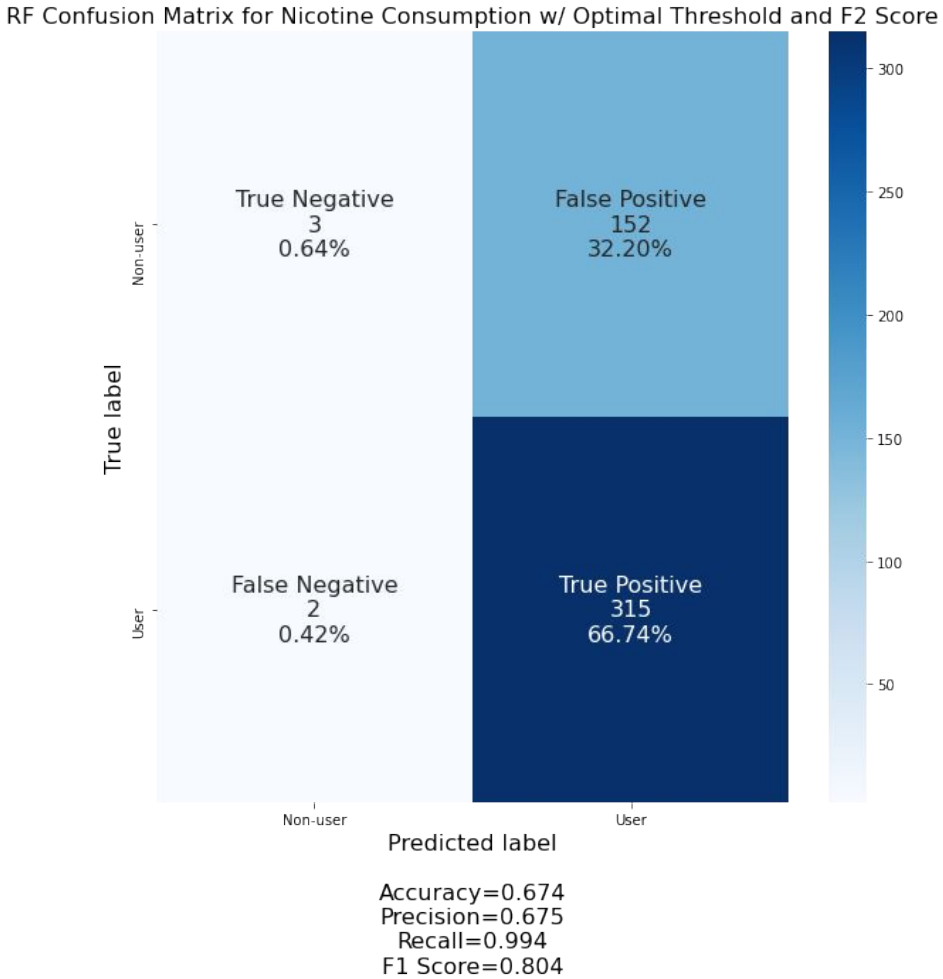


# Model Evaluation

# Choosing a Metric

- To help at risk individuals it was of greater benefit to reduce the number of predictions of actual smokers as not smokers
- Optimized for recall
- F2 score doubled weight for recall in thresholding as compared to precision
- Optimal threshold value was approximately 0.28






- Model resulted in almost no false negatives meaning that the model was successful in correctly identifying people with high nicotine consumption risk
- Greatly misclassified people who were not at high consumption risk

# 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/