Comparing the Uses and Classification Accuracy of Logistic and Random Forest Models on an Adolescent Tobacco Use Dataset

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Joseph Maginnity, B.S.

Graduate Program in Public Health

The Ohio State University

2020

Thesis Committee

Bo Lu, Advisor

Amy Ferketich

Abstract

The main purpose of research for this thesis is to compare the use of logistic and random forest classification models in machine learning to predict the outcome of adolescent tobacco use. The logistic classification model is one in which the conditional probability of the binary outcome is assumed to be equal to a linear combination of asset of independent variables, transformed by the logistic function. The random forest classification model is one in which the independent variables are used in creating many decision trees which are used to predict the binary outcome of interest. The data source is Buckeye Teen Health Study (BTHS), a survey among adolescent males to examine tobacco use behaviors. The goal of BTHS was to examine factors associated with cigarettes and smokeless tobacco product usage among urban and rural male adolescents in Ohio. Participants who answered questions at baseline and 12-month follow-up were 11- to 16- year old boys (N=1046) with 625 from the urban county and 421 from the nine rural Appalachian counties. The classification models focused on cigarette, e-cigarette, any tobacco and past 30-day tobacco usage at 12 months as the outcomes of interest. The dataset was split into two random groups with a 70:30 ratio for training and validation purposes to assess the classification accuracy of each model. The predictive capabilities of the models were assessed using ROC curves, overall classification error rates, specificity and sensitivity measurements. Overall, the logistic models performed slightly

better than the random forest counterparts, but both models had high classification accuracy in determining adolescents who did not display the outcomes of interest. The random forest models and logistic models displayed high specificity measures for all outcomes which shows that these classification models are promising techniques for determining adolescents who will not initiate tobacco use.

Dedication

This thesis is dedicated to everybody that helped get me to this point in time: my grandparents, mis abuelitos, my parents, my brother, my sister, my brother-in-law, my nieces and my friends in both the great states of California and Ohio. What a journey it has been.

Acknowledgments

I would like to acknowledge Professor Amy Ferketich and Dr. Brittney Keller-Hamilton for providing information on and allowing the use of the adolescent tobacco study dataset for this thesis. Also, I would like to acknowledge both Professor Bo Lu and Professor Amy Ferketich. Without their amazing academic and professional support, this thesis would not have been possible. Thank you.

Vita

2014 ……B.S. Biological Sciences, University of California, Davis

2014-2018…….. Research Assistant,

UC Davis Imaging Research Center

2018-present……M.S. Biostatistics, The Ohio State University

Fields of Study

Major Field: Public Health

Specialization - Biostatistics

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Section 1. Tobacco Control Background

For more than 50 years, there has been a lot of focus on the reduction of tobacco use amongst the adult and adolescent populations of the United States [1]. Scientific evidence shows that tobacco company advertising and promotion influence younger people to start using cigarette and smokeless tobacco products [2]. Even though there has been a reduction in the number of adolescents who identify as having used cigarettes within the past 30 days, the use of different smoking products is increasing. The introduction of the electronic cigarette created a whole new area in which adolescents could use tobacco products. The creation of different flavors and forms in which to use the electronic smoking devices had many appeals to the adolescent population. In a cross-sectional survey conducted in 2019 that included over 19,000 participants, the prevalence of self-reported current e-cigarette use was 27.5% among high school students and 10.5% among middle school students [3].

As the detrimental effects of tobacco use started to become apparent in medical research, efforts were started to reduce the different forms of advertising by tobacco companies. In 1971, a ban was created on cigarette advertisements on television and radio, and in 1986 advertisements for smokeless tobacco products were also banned [4]. In 1998 the Master Settlement Agreement, a civil litigation brought by 46 U.S. states,

D.C. and five territories imposed more restrictions on tobacco companies [5]. New rules brought on from the litigation included bans on transit and billboard advertisements, paid brand placement, cartoons, tobacco brand sponsorships of sporting events and concerts as well as advertising and marketing practices that targeted individuals under 18 [5].

E-cigarettes, however, have not been subject to the same restrictions as other tobacco products. E-cigarette advertising across media channels increased from $6.4 million in 2011 to $115 million in 2014 [6]. Due to this increase in advertising more than 10 million high school students and nearly 8 million middle school students were exposed to e-cigarette ads in 2014 [6].

This huge surge in advertisement spending by tobacco companies saw a sudden rise in e-cigarette use amongst the U.S. population. This raises the concern that adolescents who are surrounded by so many different forms of advertising about e-cigarettes may be more influenced to use tobacco products beyond e-cigarettes. Most e-cigarettes contain nicotine which is the addictive drug contained in regular cigarettes, cigars and other tobacco products. Nicotine can harm the developing adolescent brain and can harm parts of the brain that control attention, learning, mood and impulse control [7]. There are studies being conducted to see if there is evidence that young people who use e-cigarettes may be more likely to smoke cigarettes in the future. A 2018 National Academy of Medicine report found that there was some evidence that e-cigarette use increases the frequency and amount of cigarette smoking in the future [8].

As e-cigarette and cigarette use amongst the adolescent population began to rise, several groups began advocating for the increase of age of sale for tobacco products from

18 to 21, an initiative known as “Tobacco 21”. In March 2015, a report from the National Academy of Medicine revealed that “Tobacco 21” could prevent 223,000 deaths among people born between 2000 and 2019, including reducing lung cancer related deaths by 50,000 [9]. At the release of the report, 16 states as well as the District of Columbia had raised their minimum age of sale for all tobacco products to 21. In October 17, 2019, Ohio signed a Tobacco 21 law which banned the sale of tobacco products, including alternative nicotine products like e-cigarettes, to anyone under the age of 21 [9]. In December 2019, legislation within the year-end federal legislative package was passed by both the U.S. Senate and House of Representatives and signed into law by the president on December 20, 2019 [10]. The 2015 report from the National Academy of Medicine predicted that with the passage of Tobacco 21, tobacco use can decrease by 12 percent by the time today’s teenagers were adults and smoking initiation can be reduced by 25 percent for 15-17 year olds and 15 percent for 18-20 year olds [9]. Therefore, there is evidence that several measured factors and variables may be useful in building a predictive, statistical model to help address the effects from increased cigarette, e-cigarette and tobacco use amongst the adolescent population of the United States.

As the numbers of tobacco use begin to rise within the adolescent population, it is important to look at certain factors that may lead to the initiation of tobacco use. Ways to possibly predict this behavior would be to create a classification model with the dependent outcome variable being a whether an adolescent used some form of tobacco. There are several statistical classification models that can be used to build these models using a given dataset. As previously stated in this introduction, there have been several

studies that show that tobacco products contain nicotine, which has detrimental effects on adolescent's behavior and health. The recent surge in popularity of the e-cigarette and attitudes toward certain advertisements might cause a rise in adolescent use of tobacco products. Certain social factors and behaviors of the adolescent population of the United States may have keys to this increase in e-cigarette and nicotine containing tobacco products. Therefore, building classification models based on certain variables of interest from a current tobacco study may help give public health professionals an idea of what factors may predict behavior of adolescent tobacco use.

One of the most common classification models for a binary outcome classification is a logistic regression model. However, several machine learning methods, such as a random forest, also produce accurate predictive classification models. The question arises of which model should be utilized for looking at these factors related to possible adolescent initiation of tobacco use. The research hypothesis of this thesis is that due to their handling a large feature space, given a dataset on tobacco use related factors, a random forest model will have better predictive accuracy than their logistic regression counterparts. Before the data can be utilized, the methods of building and using these predictive models must first be defined in the next section.

Section 2. Statistical Predictive Models

There are several regression methods that can be used to build statistical predictive models using available data sets. The goal of a regression model is to find the best fitting and most parsimonious model to describe the relationship between a dependent outcome variable and a set of independent variables. Given data on predictor variables (inputs, X) and the response variable (output, Y) one can build a regression model for predicting the value of the response from these predictors or to understand the relationship between the predictors and the response. With continuous outcome variables, some regression models that can be built using the data are multiple linear regression, nonlinear regression (parametric) or nonparametric regression (smoothing). The simplest linear regression model is one with a single continuous outcome variable and a single independent, predictor variable defined as:

$$E[Y] = \widehat{\beta_0} + \widehat{\beta_1} * X$$

where $\widehat{\beta_0}$ represents the expected value of Y when the predictor variable is equal to zero and $\widehat{\beta_1}$ represents the change of the expected value of Y with a one-unit increase in X [11]. This can be extended to a multiple linear regression model with k predictor variables defined as:

$$E[Y] = \widehat{\beta_0} + \widehat{\beta_1} * X_1 + \widehat{\beta_2} * X_2 + \cdots \widehat{\beta_k} * X_k$$

Where $\widehat{\beta_0}$ represents the expected value of Y when all predictor variables are equal to zero and $\widehat{\beta_k}$ represents the change of the expected value of Y with a one-unit increase in X_k when all other predictor variables are held constant [11].

With classification outcome variables, one regression model that can be used is a logistic regression model. What distinguishes a logistic regression model from a linear regression model is that the outcome variable in logistic regression is binary or dichotomous. Techniques used in building linear regression models are similar to the methods used in building logistic regression models. In linear regression it is assumed that the expected outcome Y may be expressed as an equation linear in X taking on infinite values. However, with logistic regression, the outcome can only take on the values of 0 or 1. Thus, the goal of building classification models, like logistic regression, is understanding which variables within a given data set are most useful in helping to predict the binary outcome with a lower classification error rate. Predictive accuracy is assessed using the expected mean square error in regression analysis and expected error rate in classification analysis. Another statistical classification model that can be used for predicting a binary outcome is a random forest model, which is created from a large amount of randomly generated decision tress. In order to understand the foundation of random forest model building, one must understand the basis behind the creation of a decision tree.

For a given data set with n-covariates, a decision tree can be created with n-branches and n-leaves. The goal of the decision tree is to determine branches that reduce the residual sums of squares and provide the best predictive leaves. With continuous variables, a cut point is created that best fits the outcomes of the data. For example, if the model uses an age variable to predict a binary outcome at one branch, those aged below 15 years of age may be put in the branch leading to the outcome of “no past 30-day

cigarette usage" and those above 15 years of age go onto the next branch where another predictive variable decides the outcome of these remaining subjects in the next branch. As seen in the example decision tree in Figure 1, those under the age of 15 at baseline were put into the "yes" category and those above 15 were put in the "no" category. The cut point can be extended to classification variables as well. In a classification setting, we are creating branches that reduce the classification error and increase the predictive accuracy of the outcome. The decision trees "grow" in the way that best reduces error in the model.

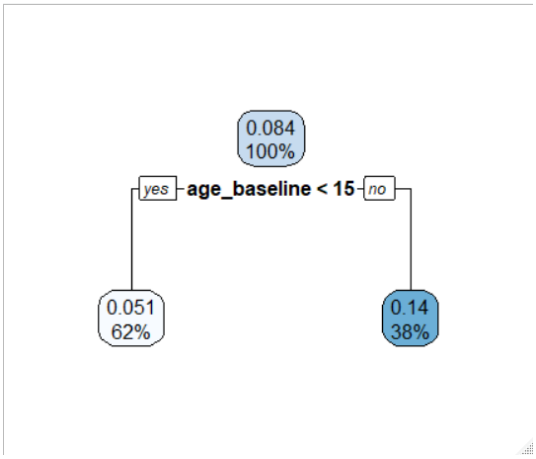

Figure 1. Example decision tree created from baseline data

The algorithms behind the creation of decision trees can lead to model over-fitting and model over generalization. This problem is resolved by the creation of a "random forest" from the many different randomly generated decision trees created from the chosen predictor variables. An additional algorithm takes a random sample of *m* predictors at each split of the branch. All decision trees created from this algorithm are aggregated back together and the model is built by assigning data to the "best" majority

vote of all the decision trees that were created in the random forest. For a given observation, the model can predict the outcome by observing the class each decision tree outputs for that observation. Then the model looks across all decision trees to see how many times that observation was predicted. An outcome is then assigned to that observation if it is the outcome predicted from the majority of decision trees.

Section 3. Adolescent Tobacco Study Dataset

This research presented in this thesis used data from a large prospective cohort study focused on examining the impact of advertising and marketing on tobacco use initiation among adolescents in urban and rural Ohio counties. The goal of the parent study was to examine factors associated with cigarettes and smokeless tobacco products among urban and rural male adolescents in Ohio. Participants who answered questions at baseline and 12-month follow-up were 11- to 16- year old boys (N=1046) with 625 from the urban county and 421 from the nine rural Appalachian counties. Participants were recruited using both probability sampling and non-probability recruitment from the community. The probability sampling method was address-based sampling that utilizes the U.S. Postal Services address list to select households to contact about the study. The selected addresses were sent a packet about the research study and a screening letter in order to determine if there were any eligible boys in the household. The non-probability sample was recruited at community events, through advertisements, and by word-of-mouth.

For each eligible household an interviewer contacted the parent or legal guardian to set up a meeting time, obtain informed permission and assent and complete the baseline session. The baseline questionnaire started with a tobacco screen to see if an adolescent had ever used cigarettes, e-cigarettes, cigars, pipe tobacco, hookah or smokeless tobacco products. There were a series of questions that were asked after this screen depending on whether the participant indicated “yes” or “no” to ever having used that tobacco product. A series of social questions were then asked surrounding deviant behavior, sensation seeking, perceptions about tobacco products, peer use of tobacco products, their school performance and media use. After these introductory questions were given, the 5 ads of interest were randomly shown to the participant for the viewing activity.

For the advertisement viewing activity, participants viewed a series of five randomly ordered advertisements: one advertisement from each of five categories (alcohol, cigarette, smokeless tobacco, e-cigarette and non-alcoholic beverage). After every advertisement was viewed, participants were asked to numerically rank it (0-10, from low to high) on three attitude measures: whether it was enjoyable, likeable and appealing. The three attitude items (enjoyable, likeable, appealing) were measured on a Likert scale (from 0 to 10) and averaged together and used as the overall attitude scale. After the adolescent was show the five advertisements, they were asked again questions regarding their perception of harm and benefits of tobacco.

Chapter 2: Methods

Section 1. Classification Models Overview

After seeing the differences between the logistic and random forest classification models in the introduction, the question arises of how each of the models predictive performances can be measured and compared. The way to look at which model is better at predicting the outcome of interest would be to look at the predictive error rate from each model using the data set of interest. The outcomes of interest utilized by the two different classification models will center on whether the individual has used a tobacco product in the past 30 days, which is considered an indicator that the individual may be a current and regular user of that tobacco product. The two different classification models will also center on the outcomes of whether the individual has ever used a certain tobacco product in their lifetime. There are many different variables that can be included in building the most efficient and parsimonious logistic or random forest classification model. The data sets contain a wide range of continuous and categorical variables that will be analyzed to use in building each of the respective base classification models. With machine learning techniques there are several methods involved with logistic regression model and random forest model building.

Section 2. Logistic Regression Models

With logistic regression, Y denotes the binary outcome variable of interest and X_1, …,X_k denote the independent random variables considered to predict the outcome. The logistic regression model uses the conditional probability [11]:

$$\pi(x) = P(Y = 1 \mid X_1, \ldots X_k) = \frac{\exp(\beta_0+\beta_1 X_1 \ldots.+\beta_k X_k)}{1+\exp(\beta_0+\beta_1 X_1 \ldots.+\beta_k X_k)}$$

$$1\text{-}\pi(x) = P(Y = 0 \mid X_1, \ldots X_k) = \frac{1}{1+\exp(\beta_0+\beta_1 X_1 \ldots.+\beta_k X_k)}$$

where β_0, β_1,…,β_k are the regression coefficients which are estimated by the log-likelihood estimates from the given dataset. A transformation of π(x) is the logit transformation, defined as [11]:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$$

$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k$$

where x is the odds of the presence of the outcome variable of interest given all the variables included in the model. The logistic formula is in terms of the predictive probability *p* that Y=1 and 1-*p* that Y=0 for a new given subject is estimated by replacing the β's by their estimated counterparts and the X's by their given values of the new subject:

$$\ln\left(\frac{p}{1-p}\right) = \hat{\beta} * X_i$$

The new subject is assigned to the outcome of Y=1 if P(Y=1)>c, where c is a fixed threshold (standard cutoff point is c=0.5) [11].

There are several advantages to logistic classification models. Logistic classification models have convenient probability outcome scores for observations. They have efficient implementations available across many different statistical tools and programs. Also, there is widespread understanding in statistics for logistic regression solutions and interpretability of the estimated coefficients. On the other hand, logistic classification models also have several disadvantages. These models do not perform well when feature space is too large. They do not handle a large number of categorical features/variables well. Also, they sometimes rely on transformations of variables for non-linear features. The predictive performance of the logistic regression model depends on whether the data follows the singular classification model. In contrast, a random forest predictive method does not rely on any singular model as it is built from the collection of randomly built decision trees.

Section 3. Random Forest Models

The random forest algorithm for regression and classification was first created by Tin Kam Ho [12] and an extension of the algorithm was developed by Leo Breiman and Adele Cutler [13] and has gained popularity since its introduction. It has grown useful in the classification approach as an alternative to logistic model classification. Clinical datasets can be limited in size, thus restricting the applications of some machine learning techniques for predictive modeling. Among various machine learning classifiers, decision trees and random forests are well suited for classification tasks. Decision trees are easy to

interpret by non-statisticians and can be more easily followed. Some studies have found that random forest models can outperform the counterpart logistic regression predictive models from real datasets [14].

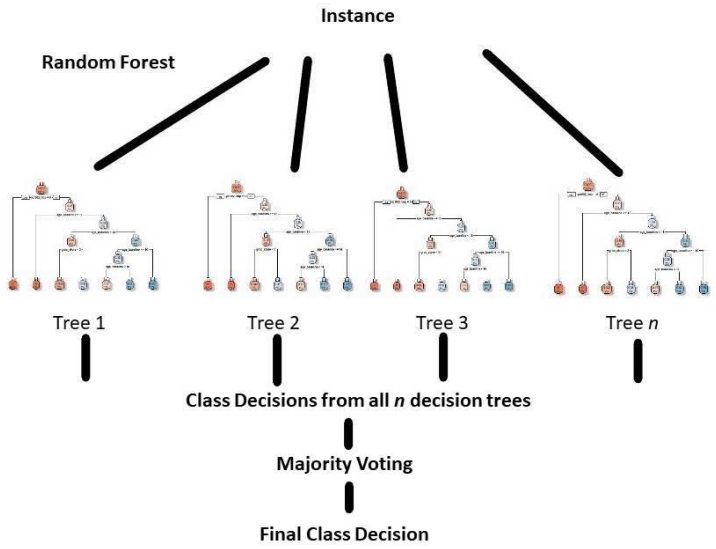

Figure 2. Simplified Random Forest Model Building Diagram

Random forest uses a class ensemble tree-based method with bagging methods to generate subsets of the entire training set to build multiple individual decision trees. Ensemble methods combine the predictions from multiple models to find the most suitable prediction to make more accurate predictions than any singular model. Bagging (Bootstrap Aggregation) is an ensemble method that is used to reduce variance from the multiple models, and in the case of a random forest, decision trees. It accomplishes this by training the model on each of the decision trees and averaging the predictions from each model. The aggregated information from all these individual decision trees are combined into a final prediction, choosing the most popular voted class as shown in

Figure 2. This classification technique requires parameters to be set, such as number of trees (*ntree*) and number of variables tried at each branch used to grow the tree (*mtry*). An advantage of using the random forest modeling over one single decision tree classifier is that it reduces over-fitting of the training data in order to increase the predictive accuracy of the random forest classification model.

Random forest model building also provides two pieces of information on top of its predictive abilities: a measurement of the importance of each predictor variable included in the model and a measurement of the proximity of different data points to one another. The variable importance measurement can be difficult to define in general because the importance of the predictor variable may be due to interactions with other variables. The random forest algorithm estimates the importance of a predictor variable by looking at how much prediction error increases when data for that variable is omitted while others are left unchanged. The calculations for the importance measurement are carried out tree by tree as the random forest model is being built. The proximity measurement is created via a proximity matrix in which the (i, j) element of this matrix is the fraction of trees in which elements i and j fall in the same terminal node. More similar observations should be in the same nodes more often than dissimilar ones. This proximity matrix can be used to identify the random forest model building process from the structure in the data.

There are several advantages with using random forest classification models. With random forest models, the predictive performance can be enhanced with the best supervised learning algorithms from the machine learning methods. They provide a

reliable feature with the variable importance estimates that help reduce the predictive error rate in variable selection. They also offer efficient estimates of the test error without the cost of repeated model training associated with cross-validation. On the other hand, as with any classification models, random forest models also have disadvantages. Ensemble models can be less interpretable than individual decision trees. Also, training a large number of complex trees can have high computational costs. The predictions can be slower and thus may create challenges for applications of the model.

Section 4. Variable Selection

In machine learning, the algorithms in creating the predictive models work by making data-driven predictions through building a mathematical model. In order to build these final classification models, two subsets of the given data will be used in the different stages of the creation of the logistic and random forest models. The predictive models will initially be fit on a training dataset, which is a randomly selected subset of the given data sets, used to fit the parameters of the model. The models are trained on the training datasets using a supervised learning method. The current models are run using the training datasets producing a result where the parameters of the model are adjusted. Then the fitted model is used to predict the responses for the observations in a second subset of the data called the test dataset. The test dataset provides an unbiased evaluation of the models created from the training dataset. The test set can help identify if the model increases the error rate which is an indication of overfitting of the data. In order to avoid overfitting, it is necessary to have the test datasets in addition to the training data sets.

When splitting the dataset into training data, less training data causes parameter estimates to have greater variance. However, putting too many subjects in the training data can cause the performance statistic to have greater variance. Therefore, the concern with dividing the data set is that neither variance is too high. With the given data set of over 1,000 subjects and about 100 subjects that display the outcomes for each variable of interest, a 70:30 split can be used for the training:test data sets.

Variable selection is a very important process in building these predictive logistic and random forest classification models. The aim of variable selection is to construct a parsimonious model that achieves a balance between predictive error and interpretability of the estimated test statistics. Automatic variable selections are not guaranteed to be accurate with these goals. Stepwise methods use a restricted search through the space of potential models and use hypothesis testing- based methods for choosing between models. Selected variables from each of the given datasets will be used in building the base model. For random forest classification model variable selection, the *Boruta* algorithm in R is useful in selecting variables based off the calculated variable importance within the predictive classification model [15]. The *Boruta* package algorithm runs models with all possible combinations of the variables to determine which of these variables has the highest cumulative variable importance measurement, and thus the lowest prediction error rate. There are a set of certain variables deemed relevant to the study from the given data sets that will be included in the statistical predictive models regardless of variable importance. The *Boruta* package labels variables as either having important, tentative or not important attributes to the random forest classification model

[15]. If the *Boruta* algorithm deems the variables of interest as either having important or tentative attributes for the models, it will be included for use in constructing the base models. If the *Boruta* algorithm deems the variables of interest as not having important attributes, it will be removed from building the base model unless it is an advertisement attitude score variable pertaining to tobacco products.

Once these base variables have been determined from the *Boruta* algorithm for the construction of the random forest classification models, the logistic regression models will be built independently using stepwise backward elimination. All variables will be included in the base logistic regression models and then using the R function *stepAIC* in the package *MASS*, the package will start with all predictors in the full model while iteratively removing the least contributive predictors and stop at a model where all variables are significant in a model with the lowest BIC (Bayesian information criterion) [16]. When fitting models, it is possible to increase the likelihood functions of the logistic regression model by adding more predictive variables, but this may result in overfitting of the model. When calculating the BIC value, the algorithm attempts to resolve this problem by introducing a penalty term for the number of variables in the model, which helps create a more parsimonious and accurate model. In other words, the BIC measurement of selection deals with both overfitting and underfitting of the model. As with the random forest models, the set of attitude advertisement score variables pertaining to tobacco products deemed relevant to the study from the given data set will be included in the statistical predictive models regardless of the BIC relevant selection using backwards stepwise selection process. If the *stepAIC* function does not include

these variables in the final selected model, these variables will be added back into the model. After the final models have been fit with all variables deemed significant or important to each respective classification model, measurements and tests will be run to assess the fit and accuracy of each predictive model.

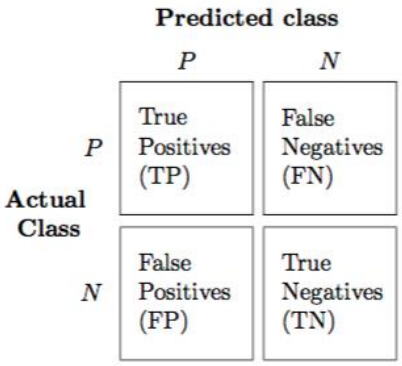

Figure 3. Confusion Matrix Diagram [11]

To compare the predictive accuracy of each model, we must discuss what errors will arise when the models make predictions on the training set. The classification measures are determined with true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Correct or incorrect classifications predicted by each model are counted and input into a confusion matrix, as seen in Figure 3. Accuracy is defined as the overall success rate of the classifier and is equal to the sum of the TP and TN divided by the total number of subjects in the training set. Sensitivity measures the fraction of correctly classified positive subjects given that they are actually in the positive class. Specificity measures the fraction of correctly classified negative subjects given they are actually in the negative class. Sensitivity and specificity rely on a single cut point to classify a test result as (Y=1).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

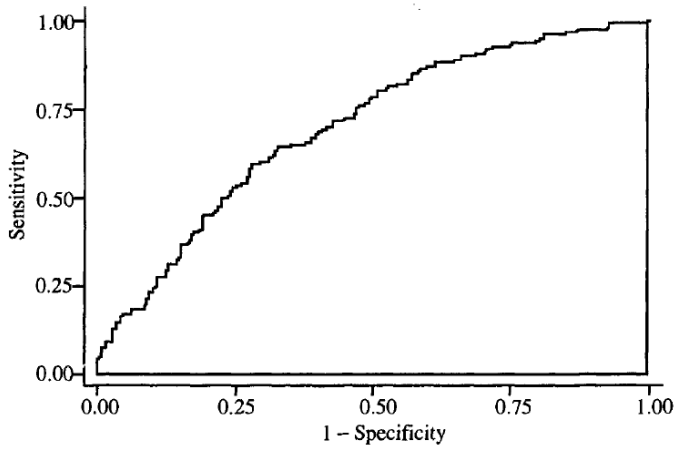

Figure 4. Example ROC Curve [11]

A plot of *sensitivity* versus 1-*specificity* over all possible probability cut points, as seen in Figure 4, is known as the *ROC* (Receiver Operating Characteristic) Curve and the area under this curve provides a measure of discrimination for the predictive model. The discrimination is the likelihood that a subject who is actually in the positive class has a higher P(Y=1) than a subject who is in the actual negative class. As a general rule [11]:

If ROC = 0.5: this suggests no discrimination (might as well flip a coin)

If 0.7≤ROC<0.8: This is considered acceptable discrimination

If 0.8≤ROC<0.9: This is considered excellent discrimination

If ROC≥0.9: This is considered outstanding discrimination

Section 5. Data Preparation and Variable Creation

Before the logistic regression and random forest classification models can be built, the data set was preprocessed which involved recoding for correct interpretation and checking for missing values. Looking at the baseline and 12-month datasets, several of the variables of interest were recoded to 0 and 1 from 1 and 2 outcomes in order to make statistical interpretations correctly from the estimated model coefficients. In order to compare the baseline values of variables to the outcomes of 12-month follow-up for each individual, the two data sets were merged via the individual adolescents unique personal identifying number (*ppid*) given to them when they entered the study (Appendix A). After merging the datasets, there were a total of 1,046 subjects that matched study criteria, answered questionnaires at baseline and at 12-month follow-up and had no missing data for the variables of interest.

Table 1		
Race of Adolescent	0 = White	n = 805
	1= Non-white	n = 241
Highest Education of Parents	0 = Bachelor's Degree or Higher	n = 609
	1 = Less than Bachelor's Degree	n = 437
Highest Income of Parents	0 = Greater than or equal to $50,000	n = 707
	1= Less than $50,000	n = 339
Adolescent Region	0 = Franklin	n = 625
	1= Appalachia	n = 421
Any Tobacco Use at 12 Months	0 = No tobacco use	n = 815
	1 = Use of any kind of tobacco	n = 231
Peer Use of Tobacco	1 = None	n = 783
	2 = A few	n = 180
	3 = Some	n = 66
	4 = Most	n= 17

Table 1. Collapsed Categorical Variables of Interest

Certain categorical variables were reduced for simplicity and due to some levels not containing enough subjects (Table 1). If the adolescent had indicated ever using either a cigarette, e-cigarette, cigar, hookah, smokeless tobacco or pipe, they were indicated as a "yes" for the any tobacco use variable and "no" if they had never used any of the tobacco products indicated. If the adolescent had indicated ever using either a cigarette, e-cigarette, cigar, hookah, smokeless tobacco or pipe in the past 30 days, they were indicated as a "yes" for the any tobacco use in the past 30 days variable and "no" if they had never used any of the tobacco products indicated in the past 30 days. Individuals that indicated that they "didn't know" an answer about tobacco use were very small (between 2-3 individuals for each variable) and recoded as "no" for all tobacco use variables where this outcome appeared. The variable for peer use of tobacco collapsed the top two categories of "all" and "most" friends use tobacco, since there were so few subjects in the top two categories (Appendix A).

As previously stated in the methods section, the study deemed certain variables of interest to be included in the classification models regardless of variable importance or statistical significance. These variables were related to the impact of advertisements and/or electronic cigarette use and whether they can be predictive of cigarette and tobacco usage amongst adolescent males in rural and urban areas of Ohio. Adolescents in the study viewed a series of five randomly ordered advertisements: one advertisement from each of five categories (alcohol, cigarette, smokeless tobacco, e-cigarette and soda). After every advertisement was viewed, participants were asked to numerically rank it (0-10, from low to high) on three attitude measures: whether it was enjoyable, likeable and

appealing. The variable for each advertisement type was created using the three averaged attitude measurements on a 0-10 Likert scale, 0 meaning not at all and 10 meaning very enjoyable, likeable or appealing. The averaged variables for e-cigarette, cigarette and smokeless tobacco advertisements are the variables that will be included in all classification models created, regardless of statistical significance or variable importance. As previously stated in the Methods section, the dataset was randomly split into training and test data sets via a 70:30 split. The training data set included 732 subjects and the test data set included the other 314 subjects from the merged datasets.

Chapter 3. Results

Section 1: Data Analysis

The variables for any cigarette usage, any e-cigarette usage and any tobacco usage at 12 months were used as the outcome variables of interest for both classification models (0 = never used at 12-months, 1=have used at 12-months). When looking at the variables for past 30-day usage of cigarettes, e-cigarettes and any tobacco product at 12 months, the number of outcomes for cigarettes and e-cigarettes were small compared to the total number of subjects. Therefore, the only classification model created for past 30-day use was a model with the outcome for any tobacco product usage in the past 30 days (0=have not used tobacco in the past 30 days, 1=some use of tobacco in past 30 days).

All variables of interest (Appendix A) were put into a full model to predict for each of these outcomes using the training data set. For the random forest models, the *Boruta* package was used for each different model to remove any non-important variables from each model until only important or tentative variables were left in the model [15]. If any of the averaged tobacco related ad scores were excluded, they were added back into the final model. For the logistic regression models, the *stepAIC* function was used with backward selection and BIC measurement as the criterion to select the reduced model (the R function is called *stepAIC* , but the default selection criterion is BIC) [16]. After the reduced model was determined, if any averaged tobacco related ad scores were excluded, they were added back into the final model.

After each of the classification models are determined for each outcome, a confusion matrix was created for each model to determine accuracy of the predictions.

Using the models and optimal cutoff points for predictive probability of each outcome, ROC curves were graphed and area under the curve was calculated. Finally, to determine that the model did not overfit the data, each model was used to predict the outcomes of the test data set and confusion matrixes were created to determine if classification accuracy was similar to the models built on the training data set.

For the random forest models, variable importance plots were created for each model. The variable importance function creates two graphs: mean decrease in accuracy and mean decrease in GINI. The graph for mean decrease accuracy shows that if this variable were removed from the model, how much the mean decision accuracy of the random forest would decrease overall from all the decision trees. The higher the decrease in average accuracy, the higher the variable importance for the random forest model. The graph for mean GINI decrease measures the average loss of accuracy by splits of that given variable. If the variable is useful, it tends to split mixed labeled nodes into pure single class nodes. GINI importance is closely related to the local decision function that the random forest model uses to select the best variable to make the split for those branches. Therefore, mean GINI decrease in local splits, is not necessarily what is most useful to measure overall model performance on classification accuracy. Therefore, the only graphs shown will be of the mean decrease accuracy of the top 5 variables used in the final model. The list of variables used for each of the final models appears in Appendix A.

Section 2: Results from Data Analysis

The first classification models used the outcome of having ever used a cigarette at 12 months. Looking at the random forest model predicting for whether a subject had used a cigarette by 12 months, the model had an overall classification error rate of 9.02% (specificity =0.981, sensitivity= 0.250) and the variables with the highest importance were marijuana usage, GPA and averaged ad scores for cigarette ads (Figure 5). Looking at the models ROC curve (Figure 6), the area under the curve was 0.8497 which means the model has excellent discrimination. Running the model on the test set, the overall error rate was 10.51% (specificity = 0.971, sensitivity = 0.265), which is similar to the results of the model using the training data set, therefore the model did not overfit the data.

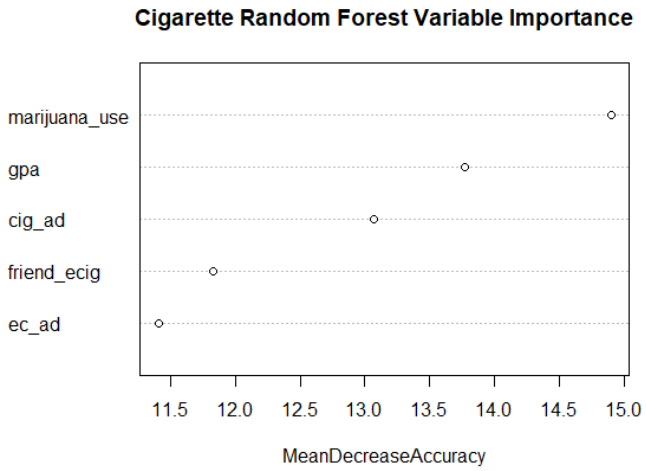

Figure 5. Random Forest Variable Importance for Cigarette Usage at 12 Months

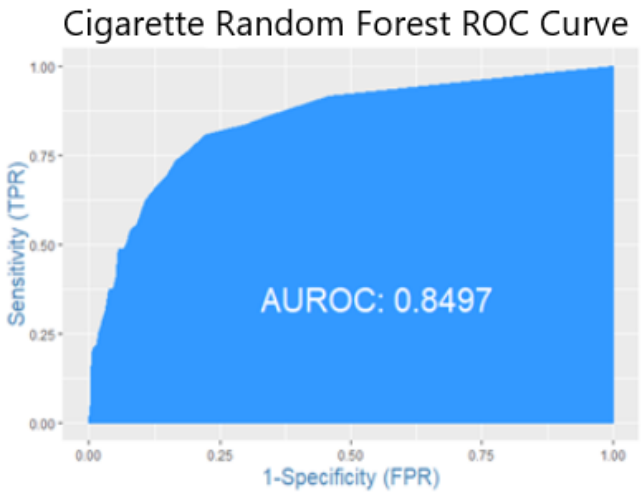

Figure 6. Random Forest ROC Curve for Cigarette Usage at 12 Months

Looking at the logistic regression model predicting whether a subject had used a cigarette by 12 months, the model had an overall classification error rate of 6.83% (specificity = 0.983, sensitivity = 0.458).

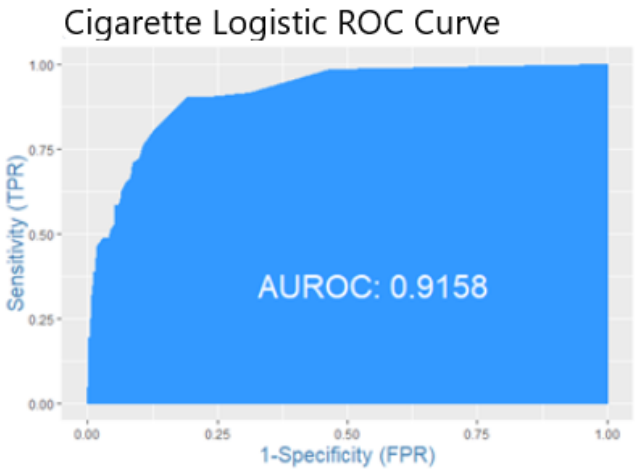

Figure 7. Logistic Regression ROC Curve for Cigarette Usage at 12 Months

Looking at the model's ROC Curve (Figure 7), the area under the curve was 0.9158 meaning that the model has outstanding discrimination. Running the model on the test

set, the overall classification error rate was 7.01% (specificity = 0.989, sensitivity = 0.441) which is similar to the results of the model using the training data set, therefore the model did not overfit the data.

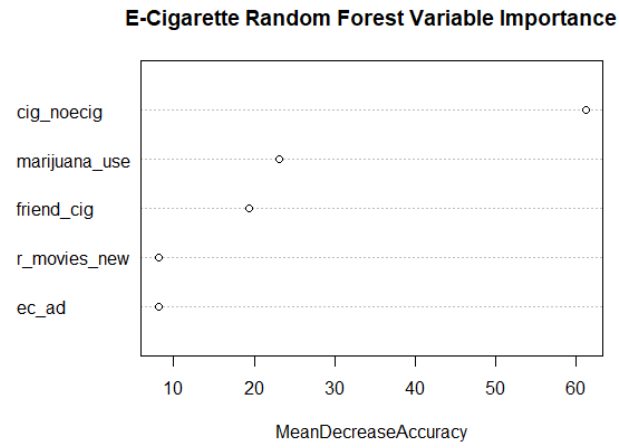

Figure 8. Random Forest Variable Importance for E-Cigarette Usage at 12 Months

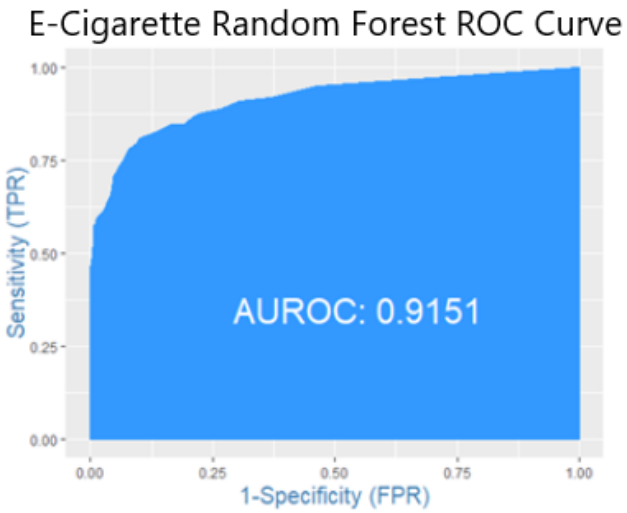

Figure 9. Random Forest ROC Curve for E-Cigarette Usage at 12 Months

The next classification models used the outcome of having ever used an e-cigarette at 12 months. Looking at the random forest model predicting for whether a subject had used an e-cigarette by 12 months, the model had an overall classification error rate of 6.56% (specificity = 0.991, sensitivity= 0.575) and the variables with the highest importance were previous cigarette usage, marijuana usage and thoughts about friend's offering them a cigarette (Figure 8). Looking at the models ROC curve (Figure 9), the area under the curve was 0.9151 which means the model has outstanding discrimination. Running the model on the test set, the overall error rate was 7.32% (specificity = 0.967, sensitivity = 0.650), which is similar to the results of the model using the training data set, therefore the model did not overfit the data.

Looking at the logistic regression model predicting whether a subject had used a cigarette by 12 months, the model had an overall classification error rate of 4.37% (specificity = 0.991, sensitivity = 0.737).

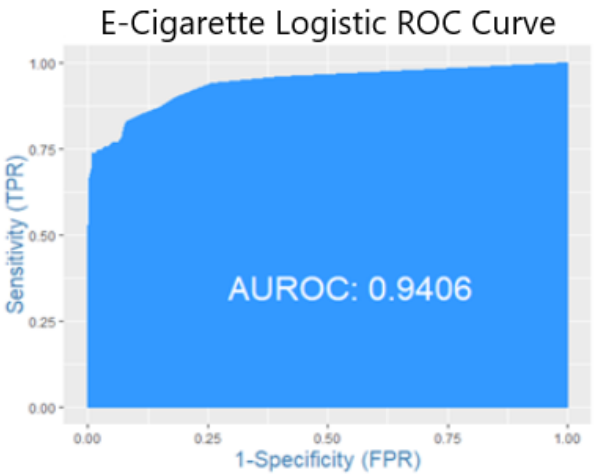

Figure 10. Logistic Regression ROC Curve for E-Cigarette Usage at 12 Months

Looking at the model's ROC Curve (Figure 10), the area under the curve was 0.9406 meaning that the model has outstanding discrimination. Running the model on the test set, the overall classification error rate was 6.69% (specificity = 0.982, sensitivity = 0.625) which is similar to the results of the model using the training data set, therefore the model did not overfit the data.

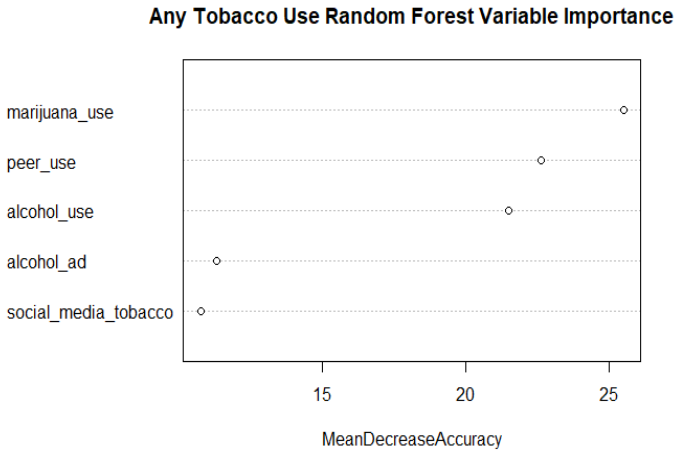

Figure 11. Random Forest Variable Importance for Any Tobacco Usage at 12 Months

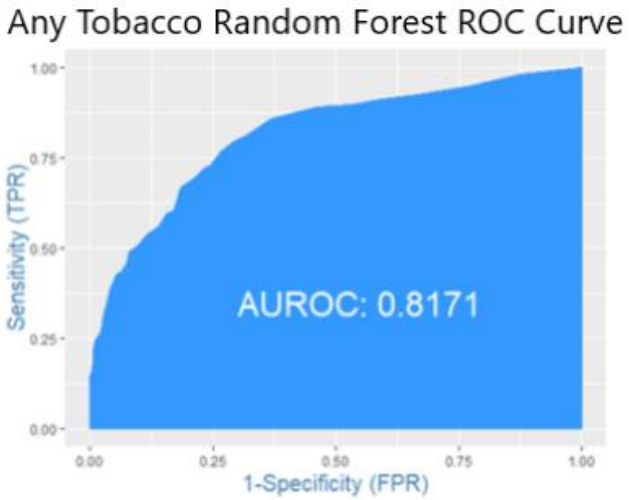

Figure 12. Random Forest ROC Curve for Any Tobacco Usage at 12 Months

The next classification models used the outcome of having ever used any tobacco product at 12 months. Looking at the random forest model predicting for whether a subject had used any tobacco product by 12 months, the model had an overall classification error rate of 16.8% (specificity =0.951, sensitivity= 0.417) and the variables with the highest importance were marijuana usage, alcohol usage and tobacco use amongst peers (Figure 11). Looking at the models ROC curve (Figure 12), the area under the curve was 0.8171 which means the model has excellent discrimination. Running the model on the test set, the overall error rate was 17.5% (specificity = 0.923, sensitivity = 0.471), which is similar to the results of the model using the training data set, therefore the model did not overfit the data.

Looking at the logistic regression model predicting whether a subject had used any tobacco product by 12 months, the model had an overall classification error rate of 14.2% (specificity = 0.951, sensitivity = 0.534).

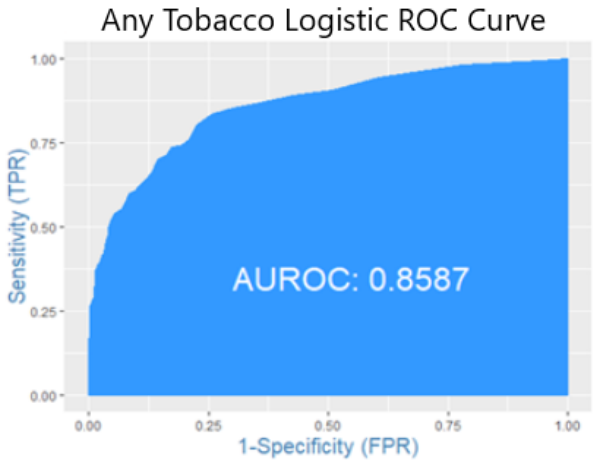

Figure 13. Logistic Regression ROC Curve for Any Tobacco Usage at 12 Months

Looking at the model's ROC Curve (Figure 13), the area under the curve was 0.8587 meaning that the model had excellent discrimination. Running the model on the test set, the overall classification error rate was 16.2% (specificity = 0.939, sensitivity = 0.471) which is similar to the results of the model using the training data set, therefore the model did not overfit the data.

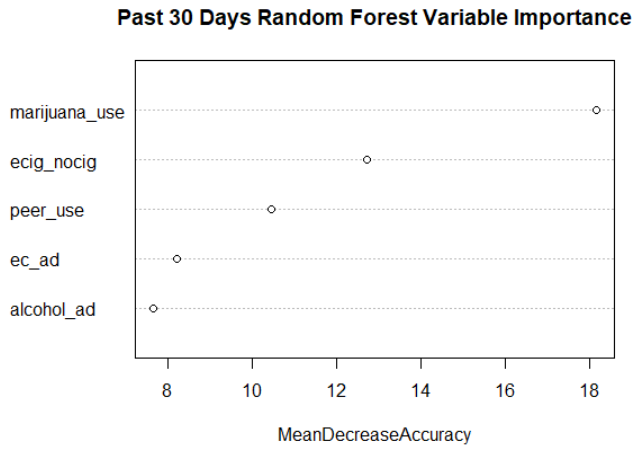

Figure 14. Random Forest Variable Importance for Past 30-Day Usage at 12 Months

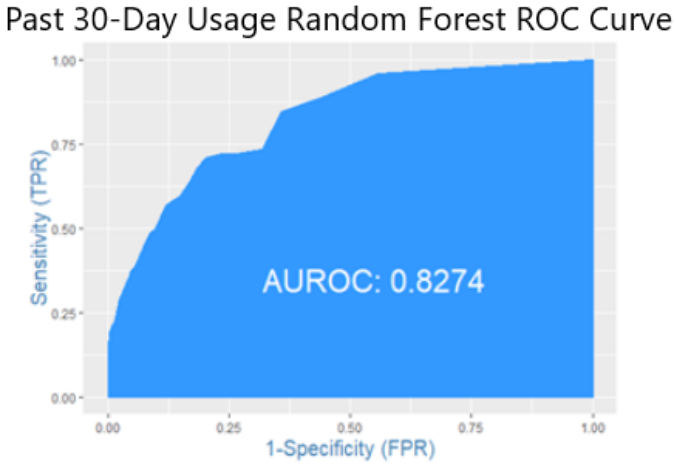

Figure 15. Random Forest ROC Curve for Past 30-Day Usage at 12 Months

The last classification models used the outcome of having ever used any tobacco product in the past 30 days at 12 months. Looking at the random forest model predicting

for whether a subject had used any tobacco product by 12 months, the model had an overall classification error rate of 8.33% (specificity = 0.995, sensitivity = 0.194) and the variables with the highest importance were marijuana usage, tobacco use amongst peers and previous e-cigarette usage (Figure 14). Looking at the models ROC curve (Figure 15), the area under the curve was 0.8274, which means the model has excellent discrimination. Running the model on the test set, the overall error rate was 10.5% (specificity = 0.982, sensitivity = 0.222), which is similar to the results of the model using the training data set, therefore the model did not overfit the data.

Looking at the logistic regression model predicting whether a subject had used any tobacco product in the past 30 days at 12 months, the model had an overall classification error rate of 7.38% (specificity = 0.995, sensitivity = 0.319).

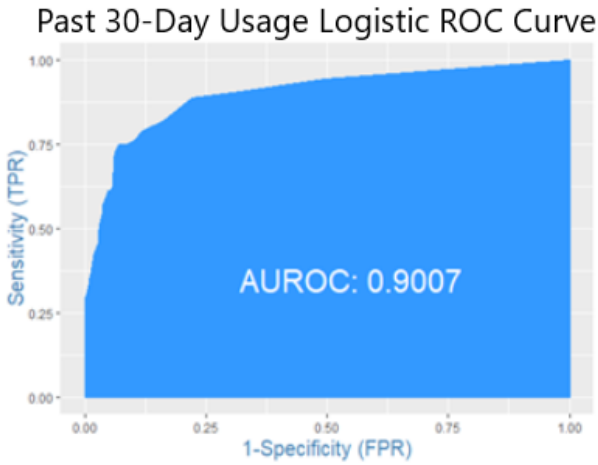

Figure 16. Logistic Regression ROC Curve for Past 30-Day Usage at 12 Months

Looking at the model's ROC Curve (Figure 16), the area under the curve was 0.9007 meaning that the model had outstanding discrimination. Running the model on the test

set, the overall classification error rate was 9.24% (specificity = 0.985, sensitivity = 0.324) which is similar to the results of the model using the training data set, therefore the model did not overfit the data.

	Random Forest	Logistic Regression
Cigarette		
Classification Error	9.02%	6.83%
Sensitivity	0.25	0.451
Specificity	0.981	0.983
AUROC	0.8497	0.9158
E-Cigarette		
Classification Error	6.56%	4.37%
Sensitivity	0.575	0.737
Specificity	0.991	0.991
AUROC	0.9151	0.9406
Any Tobacco		
Classification Error	16.80%	14.20%
Sensitivity	0.417	0.534
Specificity	0.951	0.951
AUROC	0.8171	0.8587
Past 30-Day Use		
Classification Error	8.33%	7.38%
Sensitivity	0.194	0.319
Specificity	0.995	0.995
AUROC	0.8274	0.9007

Table 2. Summary of Predictive Performance Measures of the Classification Models

Looking at all the different models, it appears that for all outcomes of interest, the random forest and logistic regression models had either excellent or outstanding discriminatory power. When it comes to the classification error rates of each of the models, it appears that for all outcomes the random forest models performed less efficiently than the logistic regression model counterparts. However, both models had excellent specificity measurements in that most of the subjects that did not have the outcome of interest were correctly labeled as not having the outcome. On the other hand, the sensitivity measurements were extremely low for both models and ranged from as low as 0.194 to as high as 0.737, which is generally not very good sensitivity. This could

be due to the fact that there were 174 subjects lost to follow-up between baseline questionnaire and 12-month follow-up. Of these loss to follow-up, 32 had the outcome for having ever smoked a cigarette and 28 had the outcome of ever having smoked an e-cigarette at baseline. Losing these many subjects with outcomes of interest could explain the high error rates among the models predicting for subjects that had the outcome of interest at 12 months.

On another point, all the logistic regression models excluded some of the tobacco related advertisement score variables when the *stepAIC* function and BIC criterion were used. Therefore, by including them in the final model, this affected the BIC of the model and may have caused error rates to slightly change for the logistic regression models. In future construction of model building, it could be considered to drop some of the insignificant advertisement score variables to reduce possible error rates in the logistic regression models. However, the random forest models always marked all of the tobacco related advertisement score variables as contributing some importance or tentative importance to the model. However, for all random forest models, they performed less efficiently than the logistic regression models predicting for the same outcome. Therefore, these high error rates amongst predictions in the random forest models cannot be attributed to including the variables that were deemed important to the study.

Chapter 4: Discussion & Future Works

This thesis presented an idea for comparing the performance of logistic and random forest classification. For predictive classification model building, where the classes can be linearly separated, and data set size may affect training and testing of the machine learning methods, the random forest and logistic classification models proved to have high discriminatory power and specificity. On the contrary, the random forest and logistic classification models showed low sensitivity, recommending against their use in classification of adolescents that display the outcomes of cigarette, e-cigarette or any tobacco usage. The overall results on the outcomes at 12-month follow-up subjects showed better discriminatory power and accuracy for all logistic models when compared to their random forest counterparts. However, overall the results support the possible increase in use of predictive random forest classification models as they also had high discriminatory power and utilized more of the variables related to the interests of the parent study relating to advertisements.

These models provided useful information on variables that are capable of predicting adolescents that do not demonstrate smoking behaviors, e-cigarette smoking behaviors or any tobacco usage behavior at 12 months. From this information, these models show that it is possible to construct models to predict tobacco related behaviors of adolescents at 24-month follow-up datasets building on variable values at baseline and 12-month follow-up. However, the information from these models would most likely be only useful in predicting non-smoking and non-tobacco usage behavior at 24-month follow-up. In the future, models could be built looking at the urban and rural adolescents

separately. Some of the collapsed variables may prove to be more useful or important for rural adolescents, but not for urban adolescents (e.g. highest income of parent, highest education of parent, etc.)

This thesis mainly focused on random forest model building using default parameters as implemented in the *randomForest* package in R [17]. This choice was made to simplify the model building. However, in future model building using random forests in R, it could be investigated using reliable and practical parameter tuning strategies, such as the number of decision trees generated, and the number of variables tested at each split of the nodes in the decision trees. Using better tuning parameters outside of the default settings may provide better sensitivity measures or overall classification accuracy for each of the individual random forest classification models.

For future works, it should be noted that there are other popular machine learning tools for classification. For example, another machine learning statistical model is support-vector machines (SVM). SVMs are supervised learning models with associated learning algorithms that analyze data used for classification analysis [18]. Given a data set, an SVM training algorithm builds a model that assigns new examples to one category or the other. In addition to performing linear classification, SVMs can also efficiently perform a non-linear classification using what are called kernel methods, which look at pattern analysis in a high-dimensional feature space [18]. So future works could look at implementing other machine learning methods, like SVMs, and comparing their performance amongst the random forest and logistic classification models.

In conclusion, this study performed using the 1046 subjects that answered questions at baseline and 12-month follow-up, showed good average predictive performance and classification accuracy for both the logistic and random forest classification models. This thesis should be seen both as an illustration of the application of random forest models in classification and a motivation to pursue the use of random forests not only on possibly more variables related to advertisements and tobacco exposures, but also on strategies to improve their use on longitudinal data.

Bibliography

1. U.S. Department of Health and Human Services. (2012). *Preventing tobacco use among youth and young adults: A report of the Surgeon General.* Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. http://www.surgeongeneral.gov/library/reports/preventing-youth-tobacco-use/.

2. Perks SN, Armour B, Agaku IT. "Cigarette Brand Preference and Pro-Tobacco Advertising Among Middle and High School Students—United States, 2012–2016." Morbidity and Mortality Weekly Report 2018;67(4):119–24. Accessed January 15, 2020

3. Cullen KA, Gentzke AS, Sawdey MD, et al. "E-Cigarette Use Among Youth in the United States", 2019. *JAMA.* 2019; 322(21):2095–2103. doi:10.1001/jama.2019.18387

4. "CDC - Information by Topic - Regulation - Smoking & Tobacco Use." Edited by Office on Smoking and Health National Center for Chronic Disease Prevention and Health Promotion, *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 13 Dec. 2017, www.cdc.gov/tobacco/data_statistics/by_topic/policy/regulation/.

5. Truth Initiative. "What Do Tobacco Advertising Restrictions Look like Today?" *Truth Initiative*, 2019, truthinitiative.org/research-resources/tobacco-industry-marketing/what-do-tobacco-advertising-restrictions-look-today.

6. "E-Cigarette Ads and Youth." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 23 Mar. 2017, www.cdc.gov/vitalsigns/ecigarette-ads/index.html#anchor_1490283089.

7. US Department of Health and Human Services. *E-cigarette Use Among Youth and Young Adults: A Report of the Surgeon General* Atlanta, GA: US Department of Health and Human Services, CDC; 2016. Accessed January 15, 2020.

8. National Academies of Sciences, Engineering, and Medicine. 2018. *Public health consequences of e-cigarettes* Washington, DC: The National Academies Press.

9. United States, Congress, Health and Medicine, et al. "Public Health Implications of Raising the Minimum Age of Legal Access to Tobacco Products." *Public Health Implications of Raising the Minimum Age of Legal Access to Tobacco Products*, The National Academies Press, 2015.

10. U.S. Food and Drug Administration. “Retail Sales of Tobacco Products.” *U.S. Food and Drug Administration*, FDA, 2019, www.fda.gov/tobacco-products/compliance-enforcement-training/retail-sales-tobacco-products.

11. Hosmer, David W., and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley, 2010.

12. Ho, Tin Kam (1995). *Random Decision Forests* Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

13. Breiman, Leo. “Random Forests.” *Machine Learning*, vol. 45, no. 1, Oct. 2001, pp. 5–32., doi:10.1023/a:1010933404324.

14. Couronné, R., Probst, P. & Boulesteix, A. “Random forest versus logistic regression: a large-scale benchmark experiment.” *BMC Bioinformatics* **19,** 270 (2018) doi:10.1186/s12859-018-2264-5

15. Miron B. Kursa, Witold R. Rudnicki (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11), 1-13. URL http://www.jstatsoft.org/v36/i11/.

16. Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0, http://www.stats.ox.ac.uk/pub/MASS4.

17. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

18. Cortes, C., Vapnik, V. Support-vector networks. *Mach Learn* **20,** 273–297 (1995). https://doi.org/10.1007/BF00994018

Appendix A: R Code

```
# Reading in the data set and merging baseline and 12-month follow-up

basedata <- read_excel("C:/Users/jdmag/Downloads/p12_baseline_imputed_8_8_19.xlsx")

twelvemdata <- read_excel("C:/Users/jdmag/Downloads/p12_12month_imputed_11_13_17.xlsx")

data <- merge(basedata, twelvemdata, by="ppid")


# Variable re-labeling and collapsing

data$int_region[data$int_region == "1"] <- "0"

data$int_region[data$int_region == "2"] <- "1"

data$race_eth_imp <- ifelse(data$race_eth_imp=="1","0","1")

data$highest_edu_par_imp <- ifelse(data$highest_edu_par_imp<8,"1","0")

data$highest_income_par_imp <-
ifelse(data$highest_income_par_imp<5,"1","0")

data$peer_use[data$peer_use==5] <- 4

data$st_ad <- (data$st_like+data$st_enjoy+data$st_appeal)/3

data$cig_ad <- (data$cg_like+data$cg_enjoy+data$cg_appeal)/3

data$ec_ad <- (data$ec_like+data$ec_enjoy+data$ec_appeal)/3

# Creation of training and test data sets

set.seed(130)
sample_size = floor(0.70*nrow(data))
train_indicator <- sample(seq_len(nrow(data)),size = sample_size)
training_set <- data[train_indicator,]
test_set <- data[-train_indicator,]


# Example Random Forest Model Creation

Boruta(CIG_EVER_12m ~ . ,data=training_set)
```

```
# Example Accuracy Measures for Random Forest Model

varImpPlot(cig_ever_model_rf, type= 1, n.var =5,main = "Cigarette
Random Forest Variable Importance")

plotROC(training_set$CIG_EVER_12m,cig_ever_model_rf$votes[,2])

pred_test <- predict(cig_ever_model_rf,test_set, type = "class")

table(pred_test, test_set$CIG_EVER_12m)

# Example Logistic Model Building

cig_ever_logit <- glm(CIG_EVER_12m ~ .,data = training_set,
family=binomial)

cig_ever_logit2 <- stepAIC(cig_ever_logit, direction="backward")

#Example Accuracy Measures for Logistic Model

predicted_tr <- predict(cig_ever_logit2, training_set, type="response")

optCutOff_tr <- optimalCutoff(training_set$CIG_EVER_12m,
predicted_tr)[1]

confusionMatrix(training_set$CIG_EVER_12m, predicted_tr, threshold =
optCutOff_tr)

misClassError(training_set$CIG_EVER_12m, predicted_tr, threshold =
optCutOff_tr)

plotROC(training_set$CIG_EVER_12m, predicted_tr)

predicted <- predict(cig_ever_logit2, test_set, type="response")

optCutOff <- optimalCutoff(test_set$CIG_EVER_12m, predicted)[1]

confusionMatrix(test_set$CIG_EVER_12m, predicted, threshold =
optCutOff)

misClassError(test_set$CIG_EVER_12m, predicted, threshold = optCutOff)
```

```
#Cigarette Model - Random Forest

cig_ever_model_rf <- randomForest(CIG_EVER_12m ~ soda_ad+alcohol_ad+ec_
ad+    cig_ad+st_ad+post_ad_relaxing+post_ad_energize+ecig_nocig+alcoho
l_use+       marijuana_use+gpa, data=training_set, ntree= 500, importan
ce = TRUE)
```

```
# Cigarette Model - Logistic Regression

cig_ever_logit <- glm(CIG_EVER_12m~cig_ad + convenience_store_ad + int_
region + st_ad + ec_ad + marijuana_use + gpa + highest_income_par_imp +
r_movies_new+ alcohol_ad + soda_ad, data = training_set, family = binom
ial)


#E-Cigarette Model - Random Forest

ecig_ever_model_rf <- randomForest(ECIG_EVER_12m ~ soda_ad + ec_ad + ci
g_ad +st_ad + post_ad_stress + post_ad_relaxing + post_ad_energize + ci
g_noecig +     social_media_tobacco + alcohol_use + marijuana_use + con
v_store + gpa +       r_movies_new,  data=training_set, ntree= 500, imp
ortance = TRUE)


#E-Cigarette Model - Logistic Regression

ecig_ever_logit2 <- glm(ECIG_EVER_12m~age_baseline + soda_ad + ec_ad +
cue_ec_q2 + cue_sd_q2 + int_region + cig_noecig + alcohol_use + marijua
na_use + conv_store + gpa + r_movies_new+st_ad+alcohol_ad+cig_ad,
data = training_set, family = binomial)

#Any tobacco use model - Random Forest

any_ever_model_rf <- randomForest(any_tobacco_ever_12m ~ age_baseline +
soda_ad + alcohol_ad + ec_ad + cig_ad + st_ad + convenience_store_ad +
int_region + post_ad_stress + post_ad_relaxing + post_ad_energize +
social_media_tobacco + smart_phone + alcohol_use + marijuana_use +
conv_store + gpa + highest_income_par_imp + r_movies_new,
data=training_set, ntree= 500, importance = TRUE)

#Any tobacco use Model - Logistic Regression

any_ever_logit2 <- glm(any_tobacco_ever_12m~age_baseline + soda_ad + st
_ad+  cig_ad + cue_sd_q2 + int_region + watch_tv + alcohol_ad + ec_ad +
alcohol_use + marijuana_use + gpa, data = training_set, family = binomi
al)


# Tobacco 30 day use model - Random Forest

any_30_model_rf <- randomForest(any_tobacco_30_12m ~ soda_ad + alcohol_
ad +  ec_ad + cig_ad + st_ad + convenience_store_ad + magazine_ad + pos
t_ad_stress + post_ad_relaxing + social_media_tobacco + alcohol_use + m
arijuana_use + gpa+ r_movies_new, data=training_set, ntree= 500, import
ance = TRUE)

#Tobacco 30 Day use Model - Logistic Regression

any_30_logit2 <- glm(any_tobacco_30_12m~age_baseline + soda_ad + cue_sd
_q2 + alcohol_use + marijuana_use + conv_store + gpa  +cig_ad + ec_ad +
st_ad +         alcohol_ad, data = training_set, family = binomial)
```

```
# List of All variables of interest used build the models

(“age_baseline,"cue_al_q2","cue_cg_q2","cue_ec_q2","cue_sd_q2",
"cue_st_q2,"int_region,“watch_tv”,"CIG_EVER_IMPR","gpa",
"highest_edu_par_imp","magazine_ad","convenience_store_ad","conv_store"
,"hh_adult_user_imp","race_eth_imp","highest_income_par_imp","r_movies_
new","you_tube”,”ecig_nocig”,”cig_noecig”,”post_ad_stress”,
“post_ad_relaxing”,”post_ad_energize”,“social_media_tobacco”,“you_tube”
,“alcohol_use”,“marijuana_use”,“alcohol_ad”,“soda_ad”,“st_ad”,“cig_ad”,
“ec_ad”)
```