

University of Arkansas, Fayetteville

ScholarWorks@UARK

---

Industrial Engineering Undergraduate Honors  
Theses

Industrial Engineering

---

5-2018

## Developing an HPV Infection Risk Prediction Model for Adult Females

Rachel Holmer

*University of Arkansas, Fayetteville*

Follow this and additional works at: <https://scholarworks.uark.edu/ineguht>



Part of the [Operations Research, Systems Engineering and Industrial Engineering Commons](#)

---

### Citation

Holmer, R. (2018). Developing an HPV Infection Risk Prediction Model for Adult Females. *Industrial Engineering Undergraduate Honors Theses* Retrieved from <https://scholarworks.uark.edu/ineguht/55>

This Thesis is brought to you for free and open access by the Industrial Engineering at ScholarWorks@UARK. It has been accepted for inclusion in Industrial Engineering Undergraduate Honors Theses by an authorized administrator of ScholarWorks@UARK. For more information, please contact [scholar@uark.edu](mailto:scholar@uark.edu), [uarepos@uark.edu](mailto:uarepos@uark.edu).

# **Developing an HPV Infection Risk Prediction Model for Adult Females**

An Undergraduate Honors College Thesis

Department of Industrial Engineering

College of Engineering

University of Arkansas

By

Rachel Holmer

# Signature Page

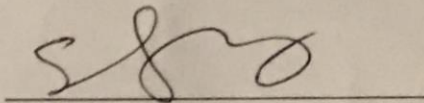
Developing an HPV  
Infection Risk Prediction Model  
for Adult Females

A thesis submitted in partial fulfillment  
of the requirements for the degree of  
Bachelor of Science in Industrial Engineering with Honors

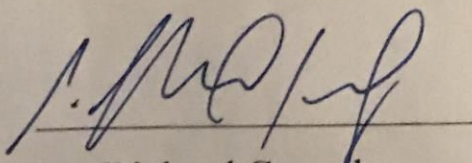
by

Rachel Holmer

Spring 2018  
University of Arkansas



Dr. Shengfan Zhang  
Thesis Director



Dr. Richard Cassady  
Committee Member

## **Acknowledgements**

I would like to take this opportunity to thank some people who have had a tremendous impact on my achievements at the University of Arkansas. I would like to thank the Department of Industrial Engineering. It has grown to feel like home with all the support and encouragement I receive from faculty and staff every day. I would like to thank the Arkansas Department of Higher Education for the State Undergraduate Research Fellowship I was awarded to fund this research. Additionally, I would like to thank my parents for always reminding me there are things in life besides school and grades.

I would like to thank Fan Wang for providing me an introduction to this research and the basics of RStudio.

I would like to thank Dr. Cassady for his investment in me over the last four years. It's not often in college you have a professor every semester, and through all the classes I took with Dr. Cassady, I learned a lot of valuable life lessons.

I would also like to thank Dr. Zhang. She is an incredible professor who really cares about the success and personal growth of her students. Over the past four years, she has opened many doors for me that I never imagined would be possible for myself before coming to college.

## **Abstract**

According to the Centers for Disease Control and Prevention (CDC), nearly one in four people are currently infected with human papillomavirus (HPV) in the United States. Although most people with HPV never experience symptoms, there is a risk of developing different types of HPV-related cancers after infection. These cancers and other related diseases result in almost \$8 billion spent annually for treatment. Currently, all boys and girls ages 11 or 12 years are recommended to receive HPV vaccination. Catch-up vaccines are recommended for males and females through the age of 21 and 26, respectively, if they did not get vaccinated previously. However, the uptake rates among young adult females remain low in the United States.

This research seeks to create a risk prediction model with a focus on adult females that will assist these individuals to estimate the risk of HPV infection based on demographic, sexual behavior, and lifestyle factors. The focus of this thesis is on the impact diet and exercise have on risk of infection. A variety of predictive models were applied to the data collected to determine the best fit. These models include logistic regression, lasso regression, ridge regression, elastic net regression, and the random forest algorithm.

Our results corroborate findings in other studies. Similar factors are recognized as significant such as sexual partners, age at first sexual activity, alcohol use, smoking habits, poverty level, and marital status. This study also found daily nutrition and sedentary activity has a significant role in HPV infection but was not able to show significance of daily exercise due to data constraints.

## Table of Contents

1. Introduction .....	3
2. Methodology.....	7
2.1 Introduction to Data .....	7
2.2 Data Processing.....	8
2.3 Dealing with Imbalanced Data .....	10
2.4 Predictive Models .....	11
2.5 Performance Measure .....	13
3. Results .....	14
3.1 Results for Logistic Regression with ROSE Sampling Data .....	14
3.1.1 Demographic Factors: .....	17
3.1.2 Sexual Behavior Factors: .....	18
3.1.3 Lifestyle Factors: .....	18
3.2 Rose Elastic Net Regression: .....	19
3.3 Results for Random Forest Algorithm.....	22
4. Conclusions .....	25
References .....	27

## 1. Introduction

More than half of all people in the United States will have a sexually-transmitted disease or infection during their lifetime (American Sexual Health Association, 2016). The Centers for Disease Control and Prevention (CDC) claims more than \$16 billion is spent annually on direct medical costs associated with sexually-transmitted diseases and infections. Among all sexually transmitted infections (STI), human papillomavirus (HPV) is the most common infection in the United States. Approximately 79 million people, or one in four people, are currently infected with HPV. Additionally, 14 million people become infected each year. This is a common disease, which means up to 80 percent of sexually-active people will contract some type of HPV in their lifetime (CDC 2016 A). It is estimated that \$8 billion is spent annually in direct medical costs for preventing and treating the diseases associated with HPV and there are approximately 6,000 deaths each year (Chesson et al. 2012).

There are over 200 types of HPV, but the majority are asymptomatic. Approximately 40 subtypes can be transmitted through sexual contact, with high-risk subtypes 6, 11, 16, 18, 31, 33, 45, 52, and 58 being commonly related to health problems. It can be seen a large portion of HPV infections are harmless, going away within a couple years and never causing cancer or other negative health effects; however, there are several diseases associated with the high-risk subtypes. These HPV-associated diseases include genital warts and cancers, such as cervical, vaginal, vulvar, anal, and even throat cancer (oropharyngeal cancer) (CDC 2016 A). Types 6 and 11 cause 90% of genital warts in addition to respiratory papillomatosis, which is the formation of benign tumors in air passages from the nose to the lungs. Type 16 is responsible for 95% of anal cancers, 70% of oropharyngeal cancers, and 65% of vaginal cancers. Types 16 and 18 are

responsible for up to 80% of all cervical cancers (National Cancer Institute 2015). The figure below shows this data for males and females in the U.S.

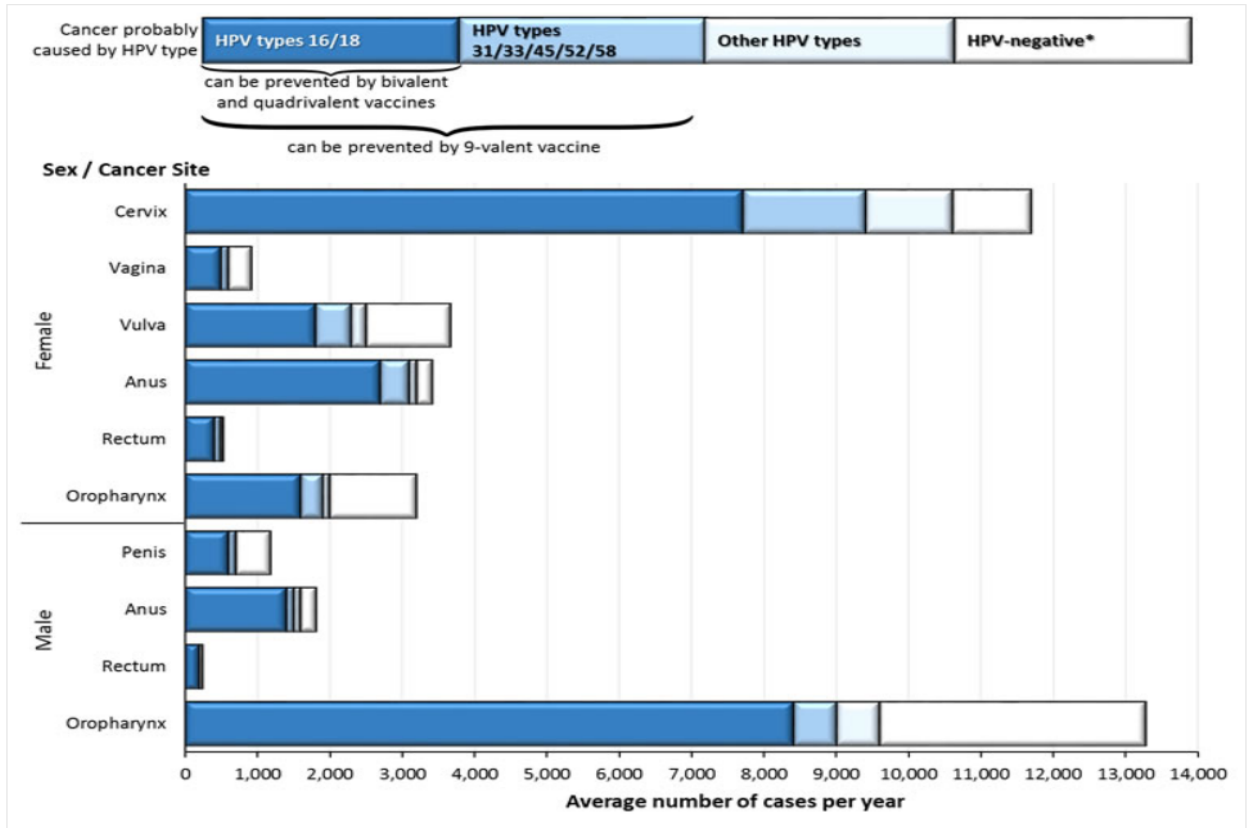


Figure 1. HPV Statistics for Cancer Types (CDC 2017)

Although a majority of diagnosed cases are caused by the high-risk types of HPV (16, 18), they can be prevented by vaccines. The CDC recommends all boys and girls receive the Gardasil or Cervarix vaccines around 11-12 years of age. Both of these vaccines are a three-dose series proven to reduce the spread of HPV along with reducing the medical costs associated with the disease (CDC 2016 B). Research has shown the HPV vaccine significantly reduced the proportion of women who are diagnosed with HPV, but there are still women who choose not to have the vaccine or do not have the vaccine by the recommended vaccination age



(Markowitz et al. 2005). Those who have not received the vaccine at childhood, or have not completed the three-dose vaccination, can receive a catch-up vaccine offered for men and women until the age of 21 and 26, respectively, within the United States.

Not all women or men have the same risk of HPV infection, and their catch-up vaccine plans should vary to achieve the greatest outcome. Should the cut off age for catch-up vaccines be determined based on a woman's risk characteristics, rather than a one-size-fits-all age of 26? Exploring risk factors associated with HPV infection and predicting each individual's HPV risk can help determine whether it is cost-effective to receive the catch-up vaccine, and whether some population may still benefit from the vaccine beyond the age of 26. To make the personal decision whether to receive the vaccine, it is necessary to predict infection based on identified risk behaviors.

This research seeks to develop a risk model to estimate the risk of HPV infection for adult females. Because approximately 43% of women ages 14-59 have HPV (Satterwhite et al., 2008), this research will focus on predicting adult female's risk for HPV infection. As a result of this research, the probability of being infected with HPV at every age can be estimated for every female based on her personal risk characteristics, which provides insights toward a more personalized recommendation on HPV catch-up vaccination.

Much research has been performed on various populations to identify behaviors associated with a higher risk of infection. Sexual history is clearly an important factor in HPV infection. Identified risk behaviors include age at first intercourse, lifetime number of sexual partners, and number of recent sexual partners (Vail-Smith et al., 1992, Moscicki et al., 1990). Having a younger age at first intercourse, or a large number of lifetime and recent sexual

partners, significantly increases risk for HPV infection because this increases the time exposed to someone with the infection. Rosario et al. (2014) researched the effect of sexual orientation on HPV infection and concluded minorities, such as homosexual and bisexual, experience a heightened risk for HPV infection when compared with heterosexual counterparts. Winer et al. (2006) performed a study to understand the impact condom use has on HPV infection. They found that women whose partners use condoms less than 5% of the time are over two times more likely to develop cervical and vulvovaginal HPV infection. There is currently no study that has combined all these risk factors together to estimate the associated risk for subjects. This research will look into the effects of all these factors and look for the best way to combine all to provide the best HPV risk prediction.

Methodologically, there are three different approaches to HPV risk prediction analyzed in the literature: multivariate analysis, Markov model, and regression models. Shi, et al. (2014) performed multivariate analysis to analyze the prevalence of HPV infection in the female population using the National Health and Nutrition Examination Survey (NHANES) data. It combined a decade of data to analyze the prevalence and quantify the extent of HPV infection in the female population. This study found individuals with more sexual partners, a lower education level, non-Hispanic black race, and no insurance were the populations at greatest risk. Meier et al. (2008) developed a Markov model to predict the age-specific incidence rate of HPV for the entire population. They also analyzed the cost-effectiveness of prevention strategies using the Markov model. Wang et al. (2018) developed a penalized regression model to predict personal HPV infection risk at the population level using a combination of data

sources, such as NHANES and a variety of other CDC datasets. There are a limited number of predictors included in their study.

This research seeks to ascertain whether nutritional and exercise lifestyle behaviors have a significant impact on HPV infection risk. Nutritional factors, such as vitamin intake, have been studied before (Breda, A. 2005); however, this research will combine a variety of nutritional aspects with the physical exercise to discover what role they play on HPV infection. The goal is to provide adult females with the risk prediction necessary to make an educated decision on the best option for risk prevention.

## 2. Methodology

We introduce the methodology for this research in this section. This research started by building a logistic regression model, then expanding to different penalized regression models, such as lasso, ridge, and elastic net. After creating these regression models, the random forest machine learning algorithm was used. All analysis is conducted in R version 3.4.2.

### 2.1 Introduction to Data

The CDC implements a program of studies, National Health and Nutrition Examination Surveys (NHANES), annually to gain insight on the health and nutritional status of adults in the United States. Questionnaire topics range from physical activity to tobacco use. Data is also collected in a laboratory setting to test for different diseases, such as hepatitis, herpes, chlamydia, and, most importantly regarding this research, human papillomavirus (NHANES 2013-2014).

Datasets from NHANES 2013-2014 will be used for this research project. The specific datasets used to formulate the predictor variables include 'Alcohol Use', 'Diet Behavior and Nutrition',

'Income', 'Physical Activity', 'Reproductive Health', 'Sexual Behavior', and 'Smoking Behavior – Cigarette Use.'

## 2.2 Data Processing

The response variable for this study is whether the subject has HPV or not. The HPV dataset from NHANES was used for the response variable. It contains 2,164 observations of 42 variables (i.e., 42 different types of HPV). This dataset was filtered down to include only those identified as high-risk subtypes: 6, 11, 16, 18, 31, 33, 45, 52, and 58. A subject was considered HPV infected (thus "1" for the response variable) if the subject tested positive for one or more of the high-risk subtypes, otherwise the subject was assigned a '0.' Rows that entirely consist of 'NA' values were discarded. There remain 1,995 observations for the dataset to use as the response variable.

After performing literature reviews to identify acknowledged risk behaviors, several variables are chosen to incorporate in this analysis as predictors. These predictors include information on demographics, sexual behavior, and lifestyles. We will discuss each category of these predictors in detail in the following.

Demographic variables include age, ethnicity, marital status, highest education achieved, and poverty level. Previous literature has stated the significant influence of sexual behavioral factors on HPV infection. A variety of behaviors are considered in this research, including lifetime sexual partners, sexual intercourse per year without condom, age at first menarche, age at first sexual activity, use of birth control, sexual orientation, live birth, and new sexual partners in the last year.

Four lifestyle factors are considered in the model: alcohol use, smoking habits, nutrition and exercise. One challenge with these variables was missing data. Although missing data was quite large for the following variables, the best manner in which to incorporate these risk behaviors was identified to ensure they were present in the model. Average daily alcohol use was calculated from the data first. There were two columns in the NHANES Alcohol Use dataset that were used to calculate this variable. The first column asked how frequently (per day, week, month, or year) the respondent would have a drink, and then the second column specified whether that number was per day, week, month, or year. The number of drinks per time period were then divided by that time period to get it into a daily value. A binary variable was created to indicate whether the subject has on average one or more drinks per day. Similarly, the model includes a binary variable indicating whether the subject has smoked 100 cigarettes in their lifetime or not, with no missing data. Nutrition information are reflected through two variables: times per week subject eats meals not from home and times per week subject eats fast food meals. Regarding exercise, we included five variables: whether subject performs vigorous work activity, whether subject performs moderate work activity, whether subject performs vigorous recreational activity, whether subject performs moderate recreational activity, and time spent sedentary daily (not including sleeping).

In total, the data frame includes 21 predictor variables. Each subject is assigned a sequence number, which was used to combine all variables into the data frame. Once all variables were merged to create one set of data, all variables were transformed into categorical variables to deal with the missing data still present in different variables. The entire list of predictor variables and categories can be found in Appendix A. The possibility for dependence

among these predictors was large, which is why penalized regression was ultimately used to predict risk of infection.

### 2.3 Dealing with Imbalanced Data

We randomly split the data into a training set with 75% of the original set and a testing set with the remaining 25% of the original data. The main concern with the data set was the presence of data imbalance. Within the training set, only 13% of the 1,496 subjects had HPV infection.

Sampling methods are commonly used to deal with data imbalance (Analytics Vidhya, 2016). In this research, we applied and compared four different sampling methods: oversampling, undersampling, both, and ROSE (i.e., random over sampling examples, a synthetic data creation method).

Oversampling works with the minority class, in this case, those with HPV. It replicates the underrepresented responses randomly until it attains the same number of responses as the larger one. This is an advantageous sampling method because it does not allow for information loss, unlike other sampling methods. The issue with oversampling is the increased possibility of overfitting a model due to duplicate responses (Analytics Vidhya, 2016).

Undersampling works with the majority class, in this case, those without HPV. It randomly samples from those who do not have HPV until the number of those without HPV is the same as those with HPV. Undersampling removes valuable information from numerous respondents, especially in this case since it deleted over 1,100 respondents' information.

The combination of undersampling and oversampling involves oversampling the smaller class with replacement and undersampling the larger class without replacement. The

advantages and disadvantages are similar to the previous two sampling methods, but to a lesser extent. Lastly, we applied the ROSE method, which involves the synthetic creation of data for the underrepresented response category (Analytics Vidhya, 2016). Refer to Table 1 for the summarized counts for the majority and minority class.

Table 1. Response Variable Counts for Sampling Methods

<b>Classification</b>	<b>0</b>	<b>1</b>
Original	1,299	197
Oversampling	1,299	1,299
Undersampling	197	197
Both	1,035	960
ROSE	788	708

## 2.4 Predictive Models

Multiple predictive modeling approaches are applied and compared in this research for the best prediction performance. Because the response variable is dichotomous, we constructed a logistic regression model and used the results as baseline for comparison. Multicollinearity is a concern in a model with so many predictor variables, especially with the variables being closely related in topic. For example, number of recent sexual partners and marital status are likely to

be related because married women will have a lower number of recent partners. To account for this, penalized regression including lasso, ridge, and elastic net regression were performed.

Lasso regression takes into account the L1 norm and regularization (Jain 2017). The L1 norm loss is essentially minimizing the sum of the absolute differences between the estimated values of the model and the target value. As regularization, L1 is the sum of the weights for the variables, which prevents the model from overfitting. The lasso regression is useful because it punishes high values of the coefficients for the model, even setting them to zero if they are irrelevant. Because of this, the lasso models in this research actually have fewer variables than the other regression models performed. The parameter alpha is set at 1 in the lasso regression mode, which is a common value chosen (R-bloggers 2017). Ridge regression takes into account the L2 norm and regularization (Jain 2017). The L2 norm loss is different than the L1 norm loss in that it is minimizing the sum of the squares of differences between the estimated model values and the target values. Similarly, L2 regularization is the sum of the squares of the weights applied to the variables. The visible difference between ridge regression and lasso regression is seen in the model output. Ridge regression will not set variable coefficients to zero, instead just lowering the coefficients. It essentially attempts to minimize their impact on the model without completely excluding variables. In R Studio, both the ridge regression and lasso regression were performed with 10-fold cross-validation on the lambda value and then tested using the test set of data. Lastly, elastic net regression, which takes into account both the L1 and L2, was performed. Different values for alpha and lambda were tested until the model performed best, using a function defined in R-Bloggers (2017).



In addition to regression-based models, one machine learning algorithm, random forest, was applied. Machine learning is a relatively recent development in which computer systems 'learn' from data and make predictions, without being programmed to do so. There are a variety of machine learning algorithms that have been developed (Le 2018).

Random forest is a classification algorithm; it has many advantages, such as its strength in preventing overfitting of the model and its ability to model for categorical values (Medium 2017). The random forest algorithm works in two steps: creation of the 'random forest' and prediction from the random forest. The algorithm can create hundreds of different decision tree processes, and it will then randomly select features from the trees to make a prediction (Medium 2017). With machine learning, the algorithms tend to 'learn' to predict the majority class, in this case no HPV infection, which will affect the ability of the model to accurately predict infection risk.

## 2.5 Performance Measure

For this study, the performance measure used to compare all the models will be area under curve (AUC). AUC measures the area under the Receiver Operating Characteristic (ROC) curve. The aim is to have as close to one as possible for the AUC value.

Odds ratios (OR) will also be used to evaluate significance of variable levels in the logistic regression models. An OR value less than one indicates that the specific variable level with that OR value has a lower risk of, with regards to this study, HPV infection, whereas a value greater than one indicates the opposite.

### 3. Results

In this section, we present the results for all the models and the machine learning algorithm.

The AUC values for all models will be displayed below. The results for the original logistic regression can be found in Appendix B. The results for only the original sampling method and the ROSE sampling method will be shown for all subsequent penalized regression models because these two performed best.

#### 3.1 Results for Logistic Regression with ROSE Sampling Data

From Table 3 below, it can be seen the best performing logistic regression model was the one using the ROSE sampling training set because it had the highest AUC. The ROC curve for the ROSE sampling logistic regression can be seen in Figure 2. Compared to the logistic regression using the original training set, a greater number of variables were found to have a significant impact on HPV risk of infection.

As mentioned earlier, multicollinearity was a concern, so several different regressions were performed after the logistic regression that account for the multicollinearity. The results for all the regression models can be seen in Table 3, but only the best performing model will be discussed in the following sections.

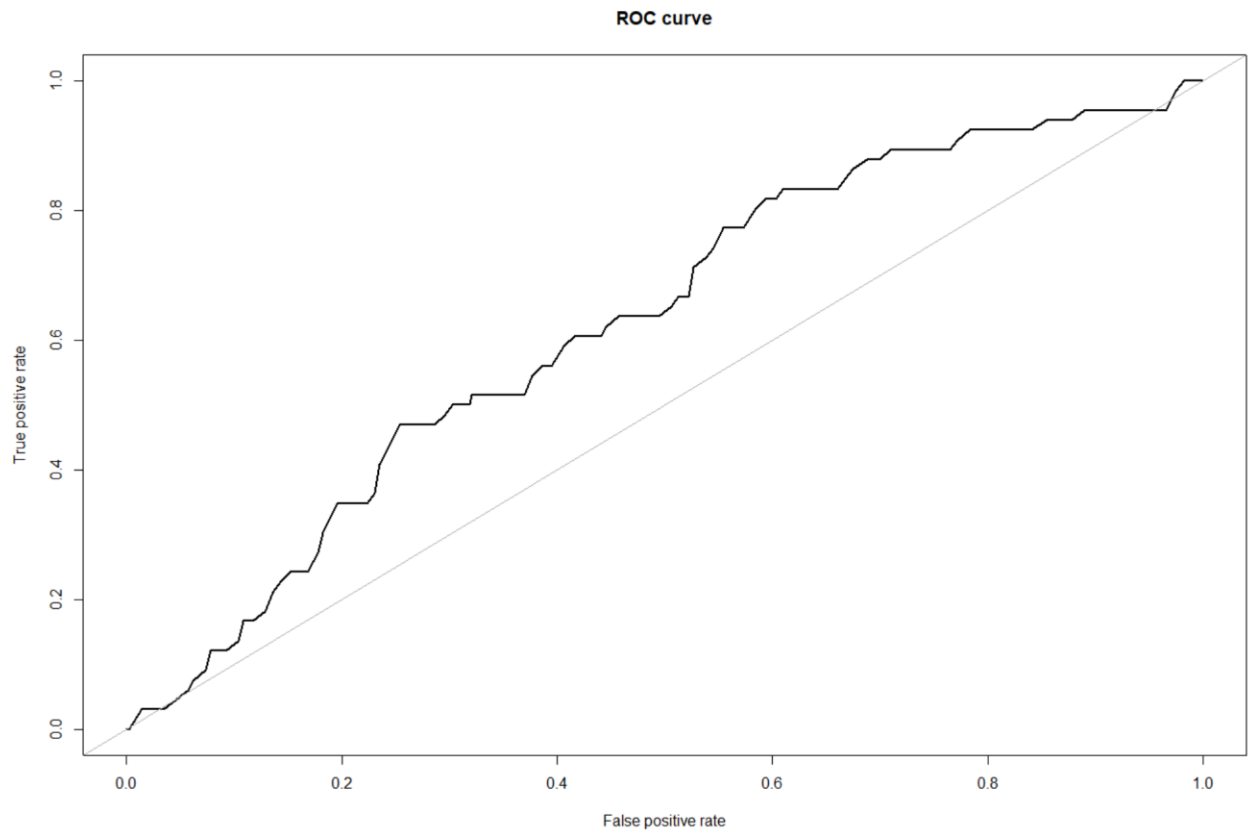
Table 2. Logistic Regression Results for Rose Sampling Method

Variables	Categories	% of Total	OR	Coef	P-Value	VIF
<b>Demographic Factors</b>						
Marital Status	Married	39.7%	--	--	--	8.65
	Widowed	1.00%	3.18	1.159	<b>0.054</b>	
	Divorced	10.5%	1.70	0.5290	<b>0.022</b>	
	Separated	5.3%	1.75	0.5613	<b>0.074</b>	
	Never Married	24.9%	1.27	0.2371	0.264	
	Living with Partner	10.2%	1.66	0.5070	<b>0.036</b>	
	Unanswered	8.4%	0.00	-16.90	0.987	
Age	18-24	21.3%	--	--	--	6.02

	25-29	12.3%	2.18	0.7825	<b>0.003</b>	
	30-34	11.9%	1.04	0.0374	0.888	
	35-39	11.5%	0.83	-0.1838	0.501	
	40-44	11.9%	0.69	-0.3725	0.203	
	45-49	9.7%	0.77	-0.2600	0.392	
	50-54	12.2%	0.59	-0.5273	<b>0.078</b>	
	55+	9.3%	0.62	-0.4815	0.158	
Highest Education Level	Less than 9 <sup>th</sup> Grade	4.9%	--	--	--	2.71
	9 <sup>th</sup> – 11 <sup>th</sup> Grade	16.0%	0.68	-0.3832	0.272	
	Graduate / GED	22.2%	0.89	-0.1138	0.747	
	Some College / AA	35.6%	0.67	-0.3977	0.247	
	College Graduate	21.2%	0.53	-0.6392	<b>0.090</b>	
Ethnicity	Mexican American	13.7%	--	--	--	2.73
	Other Hispanic	11.3%	0.88	-0.1277	0.615	
	Non-Hispanic White	38.2%	0.72	-0.3310	0.138	
	Non-Hispanic Black	25.1%	1.17	0.1531	0.509	
	Non-Hispanic Asian	8.1%	0.22	-1.535	<b>0.000</b>	
	Other Race – Including Multi-Racial	3.6%	1.19	0.1700	0.649	
Poverty Level Index	<= 1.30	52.8%	--	--	--	1.70
	>1.30	44.3%	1.29	0.2610	<b>0.078</b>	
	Unanswered	2.9%	0.44	-0.8207	<b>0.044</b>	
<b>Sexual Behavior Factors</b>						
Total Sexual Partners	<= 1	15.3%	--	--	--	9.84
	2-3	15.4%	2.56	0.9417	<b>0.001</b>	
	4-6	20.7%	6.01	1.793	<b>0.000</b>	
	7-10	16.1%	5.95	1.784	<b>0.000</b>	
	>10	18.9%	15.8	2.762	<b>0.000</b>	
	Unanswered	13.6%	1.43	0.3520	0.694	
How many times in one year do you have sex without a condom?	<= 2	29.0%	--	--	--	3.97
	3-4	10.4%	1.43	0.3581	0.110	
	>= 5	30.7%	0.67	-0.4074	<b>0.021</b>	
	Unanswered	29.8%	0.96	-0.0364	0.927	
Age at First Menarche	<= 12	45.9%	--	--	--	2.20
	13-15	36.1%	1.17	0.1584	0.277	
	> 16	7.1%	1.26	0.2303	0.399	
	Unanswered	11.0%	0.00	-17.61	0.981	
Age at First Sexual Activity	<= 12	2.9%	--	--	--	4.17
	13-15	19.1%	7.28	1.986	<b>0.000</b>	
	16-19	47.5%	7.26	1.982	<b>0.000</b>	
	>= 20	14.5%	1.00	2.305	<b>0.000</b>	
	Unanswered	15.9%	1.90	0.6403	0.483	
Birth Control Use	Yes	61.0%	--	--	--	6.17
	No	27.7%	1.20	0.1855	0.253	
	Unanswered	11.2%	0.41	-0.8887	0.999	
Live Birth	Yes	60.0%	--	--	--	7.54
	No	20.5%	0.86	-0.1540	0.439	
	Unanswered	19.5%	5.58	17.83	0.986	
Sexual Orientation	Heterosexual	77.6%	--	--	--	3.30
	Homosexual	0.67%	0.27	-1.311	<b>0.076</b>	
	Bisexual	4.6%	0.63	-0.4639	0.135	
	Something else	1.4%	0.69	-0.3776	0.517	
	Not Sure	0.94%	0.29	-1.238	0.131	
	Unanswered	14.8%	54.4	3.997	<b>0.000</b>	
<b>Lifestyle Factors</b>						

Daily Alcohol Usage	< 1 drink per day	68.1%	--	--	--	3.25
	>= 1 drink per day	2.1%	6.85	1.925	<b>0.000</b>	
	Unanswered	29.8%	1.34	0.2897	0.187	
Daily Sedentary Activity (not including sleeping)	< = 120 minutes	9.8%	--	--	--	1.63
	120-360	34.3%	1.28	0.2515	0.272	
	360-720	52.3%	0.87	-0.1389	0.540	
	>720	3.3%	0.60	-0.5168	0.221	
	Unanswered	0.27%	0.00	-14.62	0.981	
Smoking Habits	>=100 cigarettes in lifetime	36.6%	--	--	--	1.62
	< 100 cigarettes in lifetime	63.4%	0.75	-0.2893	<b>0.065</b>	
Meals Prepared Out of Home in One Week	< = 7	91.2%	--	--	--	1.95
	> 7	8.8%	0.56	-0.5816	<b>0.056</b>	
Meals Eaten at Fast Food Restaurant in One Week	< = 7	74.6%	--	--	--	2.52
	> 7	3.9%	2.21	0.7949	<b>0.082</b>	
	Unanswered	21.5%	0.63	-0.4737	<b>0.005</b>	
Does your work require at least 10 min of vigorous physical activity a week?	Yes	15.2%	--	--	--	1.33
	No	84.8%	1.12	0.1142	0.557	
Does your work require at least 10 min of moderate physical activity a week?	Yes	35.2%	--	--	--	1.35
	No	64.8%	1.16	0.1479	0.312	
Do you perform at least 10 min of vigorous recreational activity per week?	Yes	21.2%	--	--	--	1.34
	No	78.8%	1.10	0.0991	0.565	
Do you perform at least 10 min of moderate recreational activity per week?	Yes	41.5%	--	--	--	1.19
	No	58.5%	1.15	0.1428	0.290	

Figure 2. Logistic Regression ROC Plot for ROSE Sampling Method



### 3.1.1 Demographic Factors:

Out of the five demographic variables in the model, all were found to have some significance in the logistic regression model with ROSE sampling (Table 2). All marital status levels excluding 'Never Married' and 'Unanswered' were significantly different than the baseline level, 'Married.' When compared to women who are married, widowed women and separated women reported the highest ORs at 3.18 and 1.75, respectively.

Females in the age level, "25-29" are twice as likely to be infected with HPV when compared to the "18-24" level. Furthermore, when evaluating p-values, the "25-29" level and the "50-54" level were found to be significantly different than the baseline level, "18-24." Women between 30 and 50 statistically have the same level of risk of HPV infection as 18-24

year olds, with the risk decreasing after 50 years of age. Regarding education level, women who have graduated college have a much lower chance of HPV infection. Similarly, non-Hispanic Asians are significantly less likely to be diagnosed with HPV.

### 3.1.2 Sexual Behavior Factors:

Sexual behavior is acknowledged as a leading influencer for HPV infection. Much research has been done in this area, and the results of this model line up with other research available. The total number of sexual partners a woman has in her lifetime is a significant factor for HPV infection. All levels above the baseline level ( $\leq 1$  sexual partner) significantly increase the subject's chance for HPV infection. For example, females falling into the '2-3', '4-6', '7-10', '>10' levels report ORs of 2.6, 6.01, 5.95, and 16, respectively. It is clear the greater number of sexual partners, the greater the risk of HPV infection. Additionally, the age at which a female first has sex plays a significant role in HPV infection. The OR for ages 13-19 is approximately 7.3, a tremendous increase in the chance for infection.

### 3.1.3 Lifestyle Factors:

Alcohol intake and smoking were previously identified risk behaviors, and the model clearly supports this claim. Females who drink more than one alcoholic beverage a day reported OR of almost 7, and a p-value of approximately 0.00; both of these metrics show the increased risk of infection that is tied with an increase in daily alcohol use. The same pattern can be seen with smoking, where women who have smoked fewer than 100 cigarettes in their lifetime reported an OR of 0.75 and they are found to be significantly different than women who have smoked greater than 100 cigarettes in their lifetime.

The new additions of this model include the meal prepared outside the home, daily sedentary activity, and the physical activity performed at work and recreationally. Out of all these variables, the only ones found to be significant included the meals prepared outside of the home and meals eaten at fast-food restaurants. Women who eat more than seven meals from fast-food restaurants weekly have double the chance of HPV infection compared to women who eat fewer than seven meals.

Table 3. AUC Values for All Regression Models

Model	Training Set	Area Under Curve
Logistic	Original	0.615
	ROSE	<b>0.625</b>
	Over	0.603
	Under	0.559
	Both	0.602
Lasso	Original	0.522
	ROSE	<b>0.634</b>
	Over	0.610
	Under	0.586
	Both	0.606
Ridge	Original	<b>0.640</b>
	ROSE	0.633
	Over	0.612
	Under	0.612
	Both	0.604
Elastic Net	Original	0.603
	ROSE	<b>0.640</b>
	Over	0.606
	Under	0.608
	Both	0.606

### 3.2 Rose Elastic Net Regression:

An alpha value of 1 gave the best results for the elastic net regression. It can be seen that the value of alpha was close to one because it selected values for all the variable levels,

reducing a great number of them to zero. The third column represents the standardized coefficients for the variables, with only those marked as significant assigned non-zero values.

The variables selected by the model include variables identified as significant in previous models, such as smoking habits, meals eaten at fast-food restaurants, total sexual partners, age at first sexual activity, daily alcohol usage, ethnicity, and a few others. One new addition this model keeps is the daily sedentary activity, specifically the females who spend ‘120-360’ minutes per day on sedentary activities. This is interesting because it points to the possibility of exercise having a significant impact on HPV infection.

Table 4. Elastic Net Regression Results

Variables	Categories	S0
<b>Demographic Factors</b>		
Marital Status	Married	--
	Widowed	.
	Divorced	.
	Separated	.
	Never Married	0.0083
	Living with Partner	.
	Unanswered	0.1563
Age	18-24	--
	25-29	0.3702
	30-34	.
	35-39	.
	40-44	.
	45-49	.
	50-54	.
	55+	.
Highest Education Level	Less than 9 <sup>th</sup> Grade	--
	9 <sup>th</sup> – 11 <sup>th</sup> Grade	.
	Graduate / GED	.
	Some College / AA	.
	College Graduate	-0.1271
Ethnicity	Mexican American	--
	Other Hispanic	.
	Non-Hispanic White	.
	Non-Hispanic Black	0.1844
	Non-Hispanic Asian	-0.7871
	Other Race – Including Multi-Racial	.
Poverty Level Index	<= 1.30	--
	>1.30	0.3256
	Unanswered	.
<b>Sexual Behavior Factors</b>		
Total Sexual Partners	<= 1	--



	2-3	-0.0461
	4-6	0.0862
	7-10	0.0614
	>10	0.5328
	Unanswered	.
Total Recent Sexual Partners	0	--
	1	.
	>= 2	0.2454
	Unanswered	.
How many times in one year do you have sex without a condom?	<= 2	--
	3-4	0.3110
	>= 5	-0.1374
	Unanswered	.
Age at First Menarche	<= 12	--
	13-15	.
	> 16	.
	Unanswered	.
Age at First Sexual Activity	<= 12	--
	13-15	0.1411
	16-19	.
	>= 20	.
	Unanswered	.
Birth Control Use	Yes	--
	No	.
	Unanswered	.
Live Birth	Yes	--
	No	-0.1312
	Unanswered	0.0237
Sexual Orientation	Heterosexual	--
	Homosexual	.
	Bisexual	.
	Something else	.
	Not Sure	-0.0584
	Unanswered	0.5160
<b>Lifestyle Factors</b>		
Daily Alcohol Usage	< 1 drink per day	--
	>= 1 drink per day	0.7058
	Unanswered	.
Daily Sedentary Activity (not including sleeping)	<= 120 minutes	--
	120-360	0.1732
	360-720	.
	>720	.
	Unanswered	.
Smoking Habits	>=100 cigarettes in lifetime	--
	< 100 cigarettes in lifetime	-0.0796
Meals Prepared Out of Home in One Week	<= 7	--
	> 7	.
Meals Eaten at Fast Food Restaurant in One Week	<= 7	--
	> 7	.
	Unanswered	-0.0915
Does your work require at least 10 min of vigorous physical activity a week?	Yes	--

	No	.
Does your work require at least 10 min of moderate physical activity a week?	Yes	--
	No	.
Do you perform at least 10 min of vigorous recreational activity per week?	Yes	--
	No	.
Do you perform at least 10 min of moderate recreational activity per week?	Yes	--
	No	.

### 3.3 Results for Random Forest Algorithm

Two different variations were performed for the random forest machine learning. The first random forest model was built using the original sampling method, and the corresponding ROC curve and AUC value can be seen in Figure 3 and Table 5, respectively. The second random forest model was built using the ROSE sampling method, and the corresponding ROC curve and AUC value can be seen in Figure 4 and Table 5, respectively. The original sampling method produced a slightly higher AUC value than did the ROSE method.

Figure 3. ROC Plot for Original Sampling Random Forest Model

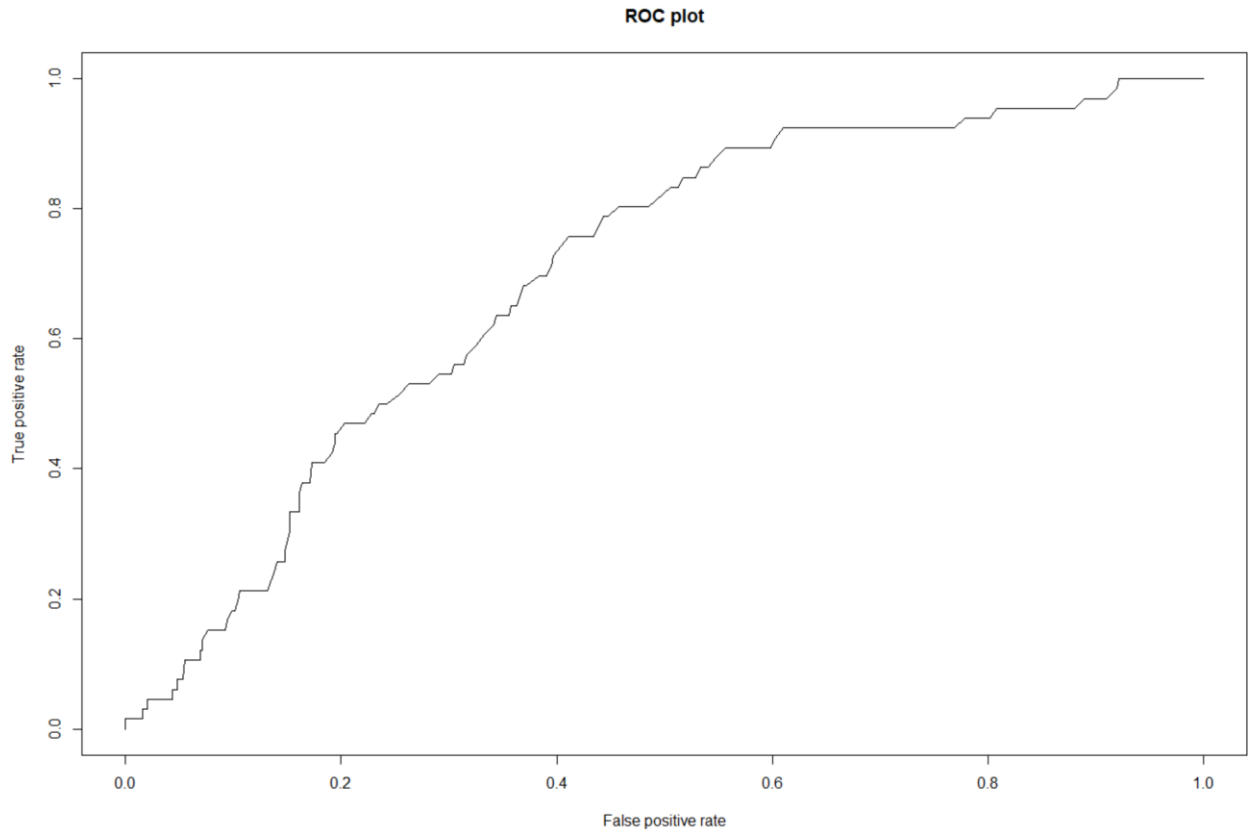


Figure 4. ROC Plot for ROSE Sampling Random Forest Model

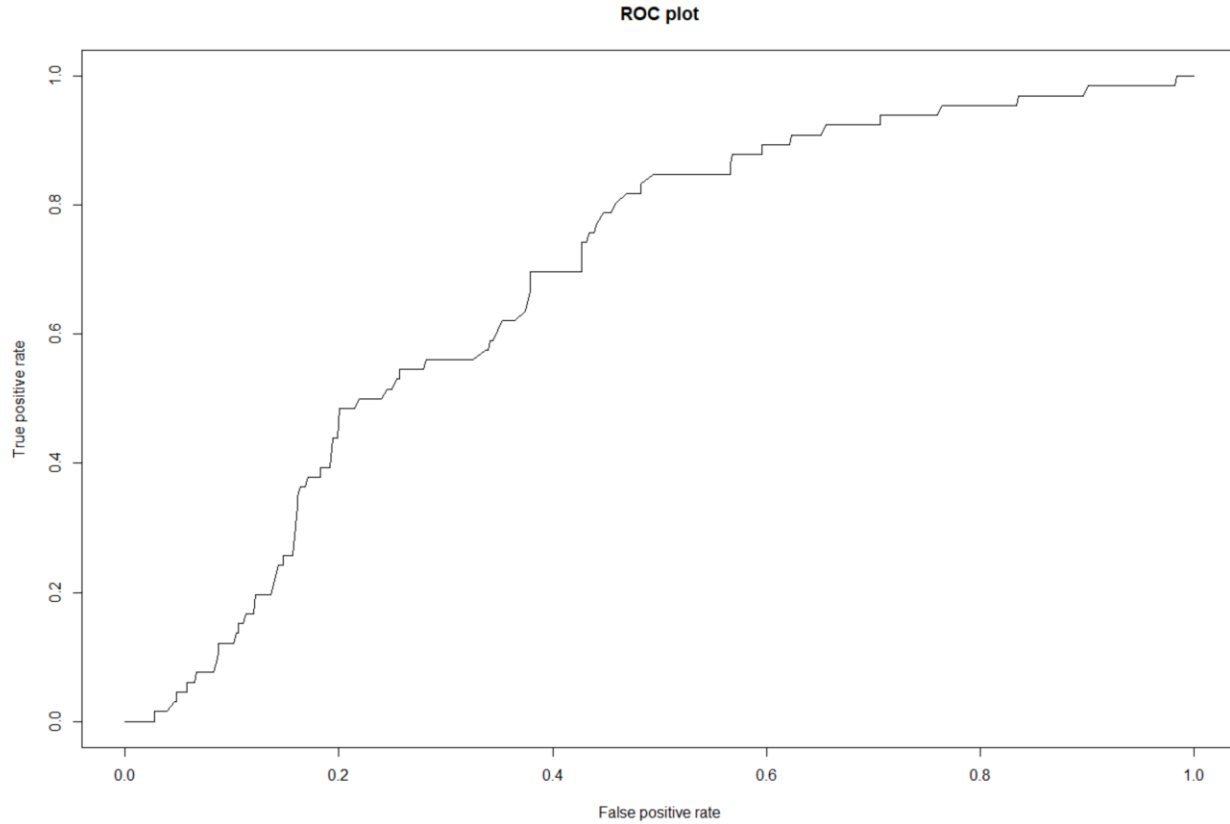


Table 5. Machine Learning Results

Machine Learning Method	Sampling Method	AUC
Random Forest	Original	0.69923
	ROSE	0.68999

In Table 6 below, the importance rankings for both random forest models are displayed. The importance ranking is worth noting because it details which variables help make better predictions. These two models placed a heavy importance on similar variables as did the previous models. Such variables include total sexual partners, sexual orientation, age, ethnicity, marital status, and daily sedentary activity.

Table 6. Random Forest Importance Ranking for Two Sampling Methods

<b>Variables</b>	<b>Mean Decrease Accuracy Original Sampling</b>	<b>Variables</b>	<b>Mean Decrease Accuracy ROSE Sampling</b>
Sexual Orientation	14.71	Age	84.73
Total Sexual Partners	14.29	Ethnicity	77.80
Total Recent Sexual Partners	11.72	Total Sexual Partners	67.91
Birth Control Use	11.03	Marital Status	67.23
Live Birth	10.21	Highest Education Level	64.73
Ethnicity	8.657	Daily Sedentary Activity (not including sleeping)	50.47
Age at First Sexual Activity	7.834	How many times in one year do you have sex without a condom?	44.65
Marital Status	5.160	Age at First Sexual Activity	42.35
Daily Alcohol Usage	3.630	Age at First Menarche	40.87
Age at First Menarche	3.605	Birth Control Use	35.99
Poverty Level Index	3.354	Meals Eaten at Fast Food Restaurant in One Week	35.91
Does your work require at least 10 min of moderate physical activity a week?	3.080	Does your work require at least 10 min of moderate physical activity a week?	34.78
Highest Education Level	2.851	Do you perform at least 10 min of vigorous recreational activity per week?	33.65
How many times in one year do you have sex without a condom?	2.471	Do you perform at least 10 min of moderate recreational activity per week?	33.24
Smoking Habits	2.236	Total Recent Sexual Partners	32.96
Meals Eaten at Fast Food Restaurant in One Week	0.8573	Poverty Level Index	32.67
Do you perform at least 10 min of moderate recreational activity per week?	0.4490	Daily Alcohol Usage	32.51
Do you perform at least 10 min of vigorous recreational activity per week?	0.1522	Sexual Orientation	30.00
Meals Prepared Out of Home in One Week	-0.4322	Live Birth	29.78
Does your work require at least 10 min of vigorous physical activity a week?	-0.4673	Does your work require at least 10 min of vigorous physical activity a week?	26.88
Daily Sedentary Activity (not including sleeping)	-0.5710	Smoking Habits	26.47
Age	-1.968	Meals Prepared Out of Home in One Week	22.30

## 4. Conclusions

The results of this research show the importance of considering a variety of factors. Several sources in the literature review have identified variables consistently found significant, such as

lifetime sexual partners, age at first sexual intercourse, ethnicity, and marital status, to name a few. However, this research shows exercise and nutrition could play a significant role.

Although the overall results for the models were not excellent when considering AUC, the results provide a decent baseline to further improve on in the future, which is discussed in the future work section following.

The main difficulty with this research was the data collection stage due to the high levels of missing data encountered. Many of the new variables added to the model (exercise and nutrition) were unable to be incorporated the way that are more directly related. This may impact the final results. For example, there are four variables in the model that try to display whether the subject exercises, and how much they exercise. When originally looking into the data, I planned to use two variables: one for recreational activity and one for physical activity in the workplace. These variables had high missing percentage, so they were transformed into binary variables with an extra level to account for the missing data.

Additionally, for nutrition, missing data was a prevalent issue. Previous research has stated lutein and other antioxidants are factors for HPV infection, but when looking into putting these types of variables into the model, over 93% of the data were missing. These issues could possibly be mitigated through the use of several years of data.

There is a great opportunity to expand this study. Since this study was completed, the newest year of NHANES data was released. Because of this, there is the possibility of improving model results by using longitudinal data to allow better inferences to be drawn about the

interactions between variables. NHANES only allows a snapshot of the respondents in that specific year. Observing over a longer period would benefit the model.

The machine learning technique performed the best when comparing the AUC value to all other models, but it could still be improved. For this I would recommend adding an ensemble method, such as bagging or boosting. These combine results from several weaker models and improve the accuracy of the results. The machine learning results here are a good baseline, but using boosting or bagging will create a more complex model that can more accurately portray the data. Additionally, it could be useful to develop more models using different machine learning algorithms. Such models may include K-Nearest Neighbors, Gradient Boosting Machine, and Support Vector Machine (Le 2018).

## References

1. "Statistics." American Sexual Health Association. 2016.  
<http://www.ashasexualhealth.org/stdsstis/statistics/>

2. "Human Papillomavirus: What is HPV?" Centers for Disease Control and Prevention (CDC). 19 May 2016. <http://www.cdc.gov/std/hpv/stdfact-hpv.htm>
3. "Human Papillomavirus: HVP Vaccines." The Center for Disease Control and Prevention. 21 July 2016. <http://www.cdc.gov/hpv/parents/vaccine.html>
4. Satterwhite, Catherine Lindsey PhD, MSPH, MPH; Torrone, Elizabeth PhD, MSPH; Meites, Elissa MD, MPH; Dunne, Eileen F. MD, MPH; Mahajan, Reena MD, MHS; Ocfemia, M. Cheryl Bañez MPH; Su, John MD, PhD, MPH; Xu, Fujie MD, PhD; Weinstock, Hillard MD, MPH. "Sexually Transmitted Infections among US Men and Women, Prevalence and Incidence Estimates, 2008" Sexually Transmitted Diseases. The American Sexually Transmitted Diseases Association.
5. Chesson HW, Ekwueme DU, Saraiya M, Watson M, Lowry DR, Markowitz LE. "Estimates of the annual direct medical costs of the prevention and treatment of disease associated with human papillomavirus in the United States." National Center for Biotechnology. 14 September 2012. <http://www.ncbi.nlm.nih.gov/pubmed/22867718?report=abstract>
6. Moscicki, Anna-Barbara, Joel Palefsky, John Gonzales, and Gary K. Schoolnik. "Human Papillomavirus Infection Sexually Active Adolescent Females; Prevalence and Risk Factors." *Pediatric Research* 28, no. 5 (1990): 507-513.
7. Winer, R. L., Lee, S. K., Hughes, J. P., Adam, D. E., Kiviat, N. B., & Koutsky, L. A. (2003). "Genital human papillomavirus infection: incidence and risk factors in a cohort of female university students." *American Journal of Epidemiology*, 157(3), 218-226.
8. Breda, A. (2005). "The Correlation Between Lifestyle, Nutrition, Vitamin Deficiency and Human Papillomavirus (HPV) Cervical Changes." Universal-Publishers.
10. Markowitz, L.E., Hariri, S., Lin, C., et al. (2013). "Reduction in human papillomavirus (HPV) prevalence among young women following HPV vaccine introduction in the United States," *National Health and Nutrition examination surveys, 2003–2010. Journal of Infectious Diseases*, 208(3), 385–393.
11. Meier, L., Van De Geer, S., & Bühlmann, P. (2008). "The group lasso for logistic regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 53-71.
12. "Fast Facts." American Sexual Health Organization. 2017. <http://www.ashasexualhealth.org/stdsstis/hpv/fast-facts/>



13. "HPV and Cancer." National Cancer Institute. 09 Feb 2015. <https://www.cancer.gov/about-cancer/causes-prevention/risk/infectious-agents/hpv-fact-sheet>
14. "HPV (Human papillomavirus)" US Food and Drug Administration. 24 Jan 2018. <https://www.fda.gov/forconsumers/byaudience/forwomen/ucm118530.htm>
15. "HPV and Cancer: How Many Cancers are Linked with HPV Each Year?" Centers for Disease Control and Prevention. 03 March 2017. <https://www.cdc.gov/cancer/hpv/statistics/cases.htm>
16. Rosario, M., Corliss, H. L., Everett, B. G., Reisner, S. L., Austin, S. B., Buchtnig, F. O., & Birkett, M. (2014). "Sexual Orientation Disparities in Cancer-Related Risk Behaviors of Tobacco, Alcohol, Sexual Behaviors, and Diet and Physical Activity: Pooled Youth Risk Behavior Surveys. *American Journal of Public Health*, 104(2), 245-254. Doi:10.2105/ajph.2013.301506
17. Vail-Smith, K., White, D. M. (1992). "Risk Level, Knowledge, and Preventive Behavior for Human Papillomaviruses among Sexually Active College Women." *Journal of American College Health*, 40(5), 227-230. doi:10.1080/07448481.1992.9936284
18. Winer, R., Hughes, J., Feng, Q., O'Reilly, S., Kiviat, N., Holmes, K., Koutsky, L. "Condom Use and the Risk of Genital Human Papillomavirus Infection in Young Women." *The New England Journal of Medicine*. [www.nejm.org/doi/full/10.1056/nejmoa053284](http://www.nejm.org/doi/full/10.1056/nejmoa053284)
19. "How Random Forest Algorithm Works in Machine Learning." Medium. AI Technology & Review. 24 Oct 2017. <https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674>
20. "Human Papillomavirus DNA – Vaginal Swab: Roche Cobas & Roche Linear Array." National Health and Nutrition Examination Survey. 2013-2014. [https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/HPVSWR\\_H.htm](https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/HPVSWR_H.htm)
21. Wang, F., Zhang, S., Jozkowski, K. (2018) "Personalized Modeling for Assessing Human Papillomavirus (HPV) Vaccination Policies for Women." Working paper.
22. Shi, R., Devarakonda, S., Liu, L., Taylor, H., Mills, G. "Factors associated with genital human papillomavirus infection among adult females in the United States, NHANES 2007-2010." 18 August 2014. doi:10.1186/1756-0500-7-544

23. Analytics Vidhya Content Team. "Practical Guide to Deal with Imbalanced Classification Problems in R." Analytics Vidhya. 28 March 2016. <https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/>
24. "Variable Selection with Elastic Net." R-bloggers. 3 September 2017. <https://www.r-bloggers.com/variable-selection-with-elastic-net/>
25. Jain, Shubham. "A comprehensive guide for Linear, Ridge, and Elastic Net Regression." Analytics Vidhya. 22 June 2017. <https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/>
26. Le, James. "A Tour of the Top 10 Algorithms for Machine Learning." Medium. 20 January 2018. <https://towardsdatascience.com/a-tour-of-the-top-10-algorithms-for-machine-learning-newbies-dde4edffae11>

#### Appendix A: Predictor Variables and Corresponding Levels

Variables	Categories
Demographic Factors	

Marital Status	Married Widowed Divorced Separated Never Married Living with Partner Unanswered
Age	18-24 25-29 30-34 35-39 40-44 45-49 50-54 55+
Highest Education Level	Less than 9 <sup>th</sup> Grade 9 <sup>th</sup> – 11 <sup>th</sup> Grade Graduate / GED Some College / AA College Graduate
Ethnicity	Mexican American Other Hispanic Non-Hispanic White Non-Hispanic Black Non-Hispanic Asian Other Race – Including Multi-Racial
Poverty Level Index	< = 1.30 >1.30 Unanswered
<b>Sexual Behavior Factors</b>	
Total Sexual Partners	< = 1 2-3 4-6 7-10 >10 Unanswered
Total Recent Sexual Partners	0 1 >= 2 Unanswered
How many times in one year do you have sex without a condom?	< = 2 3-4 >= 5 Unanswered
Age at First Menarche	< = 12 13-15 > 16 Unanswered
Age at First Sexual Activity	< = 12 13-15 16-19 >= 20 Unanswered
Birth Control Use	Yes No Unanswered

Live Birth	Yes No Unanswered
Sexual Orientation	Heterosexual Homosexual Bisexual Something else Not Sure Unanswered
<b>Lifestyle Factors</b>	
Daily Alcohol Usage	< 1 drink per day ≥ 1 drink per day Unanswered
Daily Sedentary Activity (not including sleeping)	< = 120 minutes 120-360 360-720 >720 Unanswered
Smoking Habits	≥100 cigarettes in lifetime < 100 cigarettes in lifetime
Meals Prepared Out of Home in One Week	< = 7 > 7
Meals Eaten at Fast Food Restaurant in One Week	< = 7 > 7 Unanswered
Does your work require at least 10 min of vigorous physical activity a week?	Yes No
Does your work require at least 10 min of moderate physical activity a week?	Yes No
Do you perform at least 10 min of vigorous recreational activity per week?	Yes No
Do you perform at least 10 min of moderate recreational activity per week?	Yes No

## Appendix B: Logistic Regression Results for Original Training Set

Variables	Categories	% of Total	OR	Coef	P-Value	VIF
<b>Demographic Factors</b>						
Marital Status	Married	45.4%	--	--	--	2.04

	Widowed	1.6%	1.90	0.6404	0.368	
	Divorced	11.7%	1.15	0.1413	0.644	
	Separated	3.7%	1.36	0.3100	0.439	
	Never Married	21.2%	1.22	0.1978	0.469	
	Living with Partner	8.2%	1.06	0.0587	0.849	
	Unanswered	8.2%	0.00	-0.1423	0.985	
Age	18-24	19.2%	--	--	--	4.08
	25-29	10.3%	1.01	0.0078	0.980	
	30-34	11.4%	0.81	-0.2143	0.523	
	35-39	11.9%	0.69	-0.3588	0.315	
	40-44	13.3%	0.66	-0.4158	0.262	
	45-49	11.7%	0.56	-0.5838	0.145	
	50-54	11.3%	0.45	-0.8049	<b>0.049</b>	
	55+	10.8%	0.56	-0.5795	0.185	
Highest Education Level	Less than 9 <sup>th</sup> Grade	4.7%	--	--	--	2.12
	9 <sup>th</sup> – 11 <sup>th</sup> Grade	14.5%	0.88	-0.1313	0.774	
	Graduate / GED	21.7%	1.23	0.2092	0.642	
	Some College / AA	34.6%	0.95	-0.0527	0.905	
	College Graduate	24.5%	0.71	-0.3463	0.478	
Ethnicity	Mexican American	15.1%	--	--	--	2.11
	Other Hispanic	10.2%	0.85	-0.1613	0.631	
	Non-Hispanic White	38.6%	0.87	-0.1364	0.639	
	Non-Hispanic Black	20.6%	1.04	0.0375	0.900	
	Non-Hispanic Asian	11.7%	0.31	-1.173	<b>0.010</b>	
	Other Race – Including Multi-Racial	3.6%	1.16	0.1464	0.757	
Poverty Level Index	< = 1.30	58.7%	--	--	--	1.38
	>1.30	38.0%	1.61	0.4774	<b>0.011</b>	
	Unanswered	3.2%	0.76	-0.2729	0.596	
<b>Sexual Behavior Factors</b>						
Total Sexual Partners	< = 1	20.4%	--	--	--	7.85
	2-3	17.3%	1.64	0.4969	0.199	
	4-6	20.8%	3.17	1.155	<b>0.002</b>	
	7-10	14.5%	3.44	1.238	<b>0.003</b>	
	>10	16.2%	6.39	1.855	<b>1.81E-5</b>	
	Unanswered	10.8%	4.46	1.495	0.240	
How many times in one year do you have sex without a condom?	< = 2	26.3%	--	--	--	3.17
	3-4	9.5%	1.09	0.0885	0.739	
	>= 5	33.6%	0.66	-0.4087	<b>0.072</b>	
	Unanswered	30.5%	0.99	<b>-0.0131</b>	0.979	
Age at First Menarche	< = 12	46.7%	--	--	--	4.04
	13-15	37.6%	1.18	0.1647	0.384	
	> 16	5.9%	1.04	0.0369	0.919	
	Unanswered	9.6%	0.00	-0.1958	0.989	
Age at First Sexual Activity	< = 12	2.8%	--	--	--	4.55
	13-15	18.8%	3.29	1.193	<b>0.071</b>	
	16-19	45.2%	3.35	1.212	<b>0.067</b>	
	>= 20	17.9%	5.25	1.658	<b>0.020</b>	
	Unanswered	15.3%	0.67	-0.3886	0.766	
Birth Control Use	Yes	62.2%	--	--	--	5.49
	No	28.1%	1.23	0.2055	0.327	
	Unanswered	9.8%	0.52	3.948	0.998	
Live Birth	Yes	65.3%	--	--	--	2.23
	No	17.1%	0.75	-0.2805	0.279	
	Unanswered	17.6%	0.00	0.1458	0.985	
Sexual Orientation	Heterosexual	79.8%	--	--	--	1.85

	Homosexual	1.05%	0.35	-1.052	0.208	
	Bisexual	4.9%	0.86	-0.1547	0.674	
	Something else	0.80%	0.87	-0.1379	0.870	
	Not Sure	1.8%	0.29	-0.1234	0.259	
	Unanswered	11.6%	0.17	2.849	<b>0.000</b>	
<b>Lifestyle Factors</b>						
Daily Alcohol Usage	< 1 drink per day	70.2%	--	--	--	2.38
	>= 1 drink per day	1.2%	2.43	0.8886	0.165	
	Unanswered	28.6%	1.22	0.2004	0.451	
Daily Sedentary Activity (not including sleeping)	< = 120 minutes	9.4%	--	--	--	1.34
	120-360	35.7%	1.19	0.1795	0.568	
	360-720	50.7%	1.04	0.0385	0.902	
	>720	4.1%	0.46	-0.7745	0.220	
	Unanswered	0.15%	0.00	-0.1313	0.987	
Smoking Habits	>=100 cigarettes in lifetime	32.4%	--	--	--	1.47
	< 100 cigarettes in lifetime	67.6%	0.77	-0.2591	0.195	
Meals Prepared Out of Home in One Week	< = 7	91.7%	--	--	--	1.89
	> 7	8.3%	0.94	-0.0658	0.869	
Meals Eaten at Fast Food Restaurant in One Week	< = 7	77.3%	--	--	--	2.19
	> 7	2.9%	0.99	-0.0029	0.995	
	Unanswered	19.7%	0.81	-0.2094	0.353	
Does your work require at least 10 min of vigorous physical activity a week?	Yes	13.2%	--	--	--	1.28
	No	86.7%	0.87	-0.1450	0.558	
Does your work require at least 10 min of moderate physical activity a week?	Yes	33.5%	--	--	--	1.26
	No	66.5%	1.12	0.1129	0.552	
Do you perform at least 10 min of vigorous recreational activity per week?	Yes	23.4%	--	--	--	1.23
	No	76.6%	1.17	0.1606	0.465	
Do you perform at least 10 min of moderate recreational activity per week?	Yes	43.5%	--	--	--	1.15
	No	56.5%	0.99	-0.0055	0.975	

## Appendix C: Logistic Regression ROC Curve for Original Sampling Method

