Cancer Screening Behaviors among men & women in the US from 2015-2016

Manoj Gootam, Deepti Joshi

December 18, 2016

Abstract

Cancer has proven to be a major burden on the health care industry as well as on the economy. Even though there are no cures for cancer, cancer screening has been shown to be effective in increasing survival rates among individuals with the disease. Hence, several efforts have been made to encourage cancer screening behaviors among healthy adults. In the effort to understand cancer screening behaviors, literature suggests that there are certain factors that make it more likely for an individual to get screened for cancer. Literature has identified individual level factors such as age, gender, race and ethnicity among other factors that predict engagement in cancer screening. However, there is comparatively little evidence on the importance of social context factors in guiding cancer screening. Hence, the goal of the current project was to examine if among men and women in the US between 2015 and 2016, cancer screening would be based on familial and community level factors and also assess their likelihood of getting screened for cancer in the past year based on these social context factors. Results suggest among adults in the US, family-related factors such as family structure (have family members who were limited in any way, or aged 65 years and older) family income as well as community-related factors such living in a close-knit and helpful neighborhood, exposure to second-hand smoke at work were predictive of cancer screening.

Motivation

Cancer has a major impact on society in the United States and across the world (NCIT, 2015). In 2016, an estimated 1,685,210 new cases of cancer will be diagnosed in the United States and 595,690 people will die from the disease (NCIT, 2015). Given the high incidence of cancer in the US, there have been several efforts in the past, encouraging cancer screening among adults. Cancer screening can include physical exams, laboratory tests, imaging procedures as well as genetic tests. Unfortunately, even though cancer screening technologies exist widely, they are often underused (Selvin et al., 2003). While there are no known cures for cancer, an early diagnosis is more likely to be treated successfully. Cancer screening can help find cancer at an early stage, before symptoms appear (Selvin et al., 2003). For instance, more than 90% of women diagnosed with breast cancer at the earliest stage survive their disease for at least 5 years compared to around 15% for women diagnosed with the most advanced stage of disease. Similarly, around 70% of lung cancer patients will survive for at least a year if diagnosed at the earliest stage compared to around 14% for people diagnosed with the most advanced stage of disease. Given the high incidence of cancer and the importance of getting screened in increasing survival rates, it is important to understand the factors that predict cancer screening behaviors among individuals. This would allow us the encourage cancer screening among those who are least likely to get screened for cancer. Awareness of such behavioral patterns could help health care and public health professionals develop intervention materials to encourage cancer screening among those least likely to get screened for cancer.

Literature Review

Broadly, two levels of cancer screening predictors have been identified in the current literature, individual-level and community-level factors. For instance, individual level factors such as age, race, ethnicity, education, employment, income and insurance have been shown to be predictive of cancer screening (Wong et al., 2013, Calle et al., 1983; Gandhi et al., 2015) such that being under 50 years of age, Hispanic, having lower education, earning a lower income and having no insurance have been shown to be significant predictors of not getting screened for cancer. Indeed, racial differences in cancer incidence and mortality rates are well-established. For instance, although the rate of breast cancer incidence among White women is higher compared to Blacks and Hispanic women, Black and Hispanic women are usually diagnosed at later stages of cancer and thus has lower survival rates compared to White women. Hence, while there is considerable research on individual level factors predicting cancer screening, what is less well understood is how one's social context can impact engagement in screening behaviors. Although majority of the literature discusses the impact of individual level factors in predicting cancer screening, there is some evidence that familial and community level factors also play a major role. For instance, family-level factors such as marital status and having children have all been shown to predict cancer screening such that being married having children make cancer screening more likely (Straughon et al., 1998). On the other hand, community-level factors such as type of urban area has been shown to be predictive of cancer screening such that living in makes cancer screening more likely (Barry et al., 2005). Given the promising impact of community and family level factors the goal of the current study is to extend this literature and examine if other family level factors (e.g., family structure, family income) and community-level factors such (e.g., individual's involvement in community, work-related factors) can predict engagement in cancer screening behaviors.

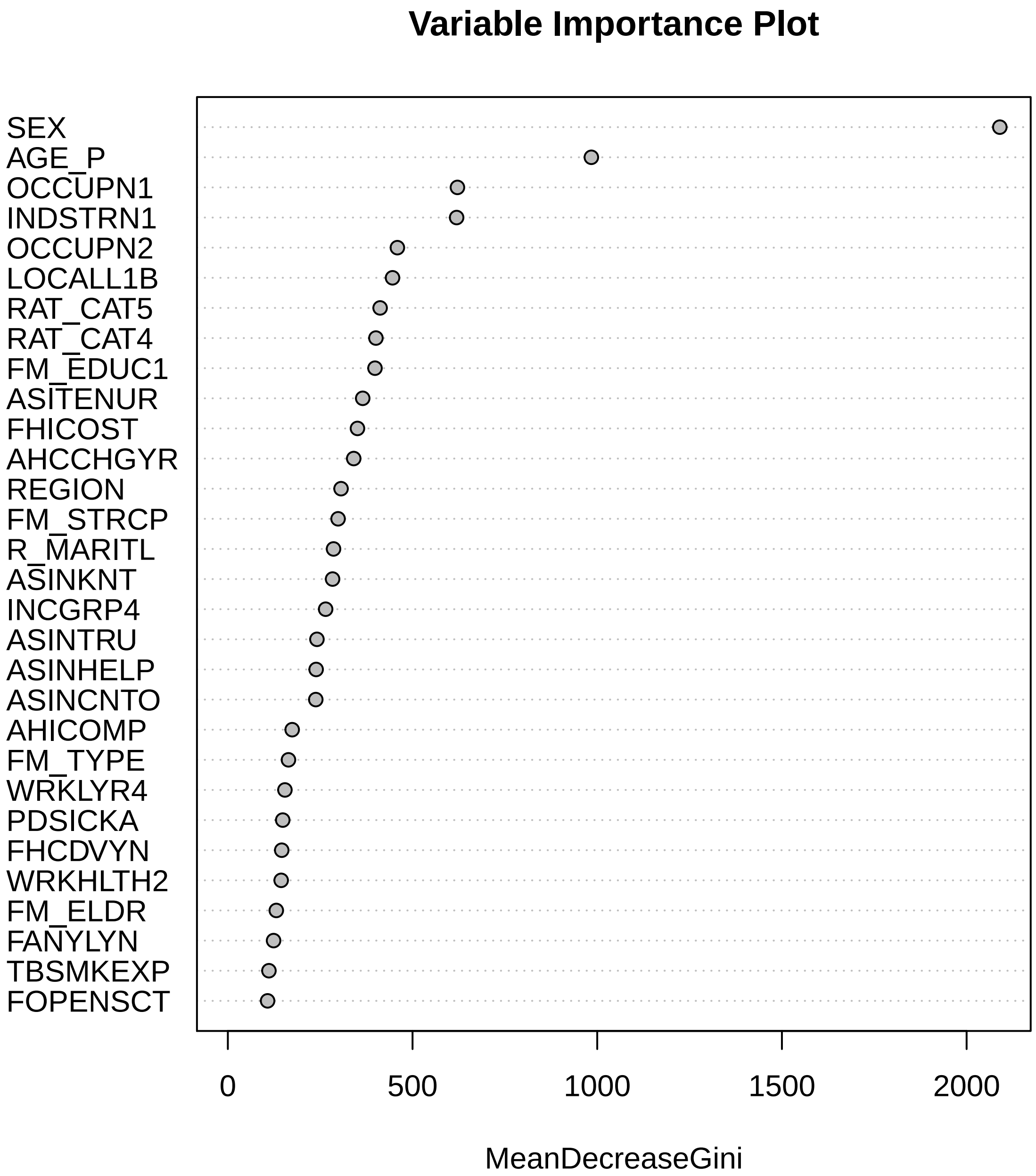
Exploratory Data Analysis

1. **Data Preprocessing**

The data was retrieved from the National Health Interview Survey (NHIS) 2015-2016. NHIS, a major data collection program of the National Center for Health Statistics (NCHS) which is part of the Centers for Disease Control and Prevention (CDC; NHIS, 2016). NHIS conducts annual in-person interviews to survey the population of US. Data is collected from 1 adult per family to obtain nationally representative estimates. Additionally, NHIS surveys oversample Black and Hispanic populations. NHIS collects household, family, injury and child related data. In order to prepare the data for analysis, we merged two separate data sets containing data on family-level factors and community-related factors on the basis of common identifiers. Additionally, we recoded the variables that were relevant. For instance, given that our output variables was a categorical one (i.e., if you got screened for cancer or not), we recoded not getting screened, not knowing or refusing to answer = 0 and getting screened in the past year = 1. Besides, all missing data were encoded as zeros. Finally, variables that had variance of 0 or approximately 0 were removed from analysis. Relatedly, in order to cross-validate out data, we split the data into train data and test data with 70% and 30% of our total sample respectively.

2. **Feature Selection**

To empirically assess the relative importance of individual predictors in the model, first we looked at the absolute value of the t-statistic for each model parameter and the corresponding p.value. All variables with p.values less than 0.001 were considered signigicant. In order to further validate these findings, we also conducted the random forest algorithm, using mean decrease in gini coefficient as a metric. As expected we found a lot of variables that were common to both the analyses. We have selected a total of 32 predictors common to both the feature selection results.



3. **Description of the final dataset after feature selection**

The following are the details of the final dataset

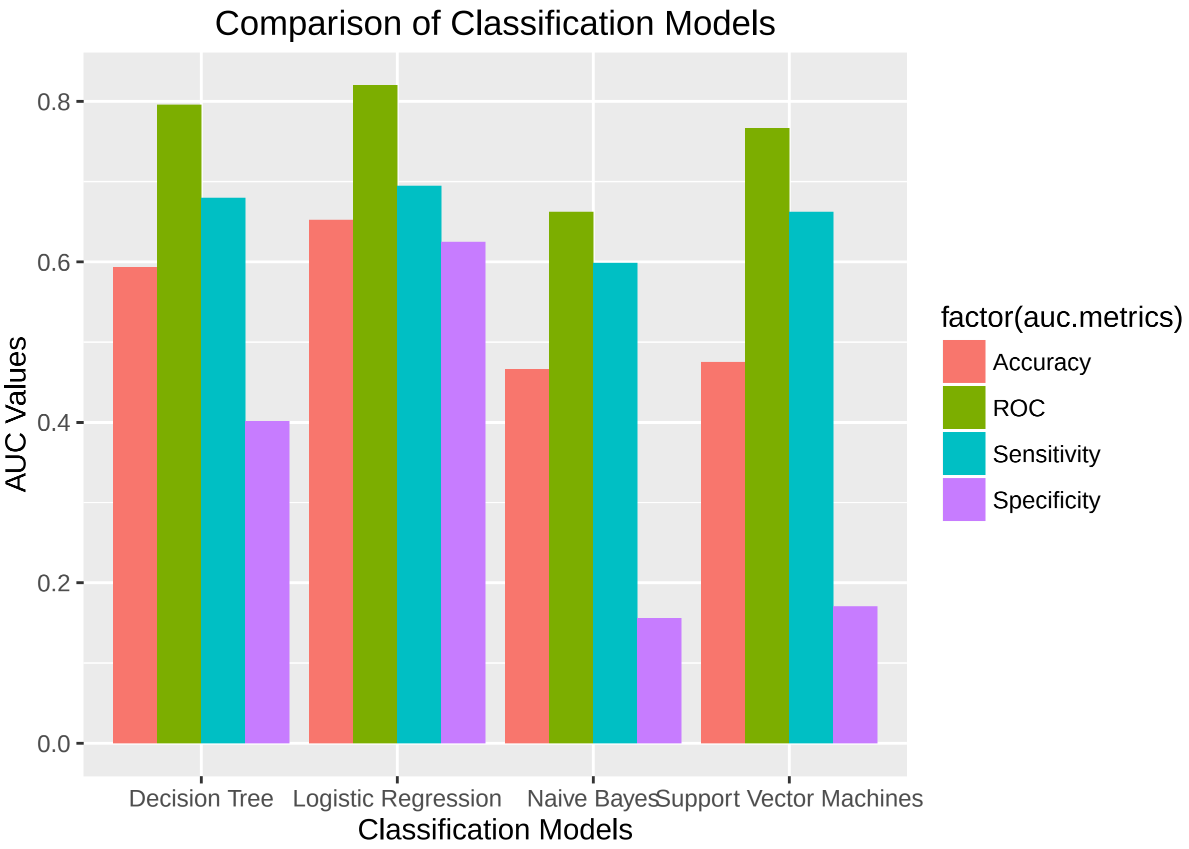
- Number of observations: 33672
- Number of independent variables of interest: 32
- Dependent Variable: Cancer Screening Test (Dichotomous)
  - 0 if Cancer Screening is not done (a total of 20,401 out of 33,672 people)
  - 1 if Cancer Screening is done (a total of 13,271 out of 33,672 people)
- Cross Validation: Data Partitioning(70% - 30% split)
  - Train Data set size: 23,570 x 32
  - Test Data set size: 10,102 x 32

4. **Data Modelling - Model Selection**

Since the dependent variable is binary, and the independent variables are a mixture of categorical variables like FM_TYPE(Family Type) and continuous variables like FHICOST(cost of family health insurance), we have looked at the following binary calssification techniques.

- Logistic Regression: Logistic regression is one of the most frequently used technique for binary classification but this method assumes exponential relationship between the odds of success and predictors. The error distributions are assumed to be normal with observations being independent and identically distributed.
- Decision Trees: This is a simple modelling technique and is difficult to model complex relationships using this technique.
- Naïve Bayes: This modelling technique uses bayesian approach to find the classes assuming conditional Independence of the predictors given the data. This assumption is almost always not true for real life datsets since there is always some kind of dependence between the variables.
- Support Vector Machines: This is also a commonly used binary classification technique which is robust to noise as well. This modelling technique assumes gaussian error distributions with observations being independent and identically distributed.

- Random Forests: Although this technique is a very good predictive modelling technique, it is difficult to infer the direct relationship(as a functional form) between the output and the inputs. So, we have not used this technique because we were keen on identifying the functional relationship between the output and the inputs.



All the classification metrics are tabulated below for all the classifiers.

Classifier	Sensitivity	Specificity	Accuracy	ROC

Classifier	Sensitivity	Specificity	Accuracy	ROC
Logistic Regression	0.6916204	0.6268045	0.6526359	0.8184249
Decision Tree	0.6730167	0.4023921	0.5948274	0.7875568
Naive Bayes	0.5987141	0.1583360	0.4711874	0.6640253
Support Vector Machines	0.6555370	0.1704261	0.4796934	0.7584951

Looking at the above plot and the above classification metrics table for all the modelling techniques it is very clear that logistic regression classifier outperforms the rest of the classifiers. Also, specificity in particular is very important to our problem because we are interested in identifying people who do not opt for screening behavior and their characteristic profiles. And, logistic regression classifier does a very good job in terms of specificity.

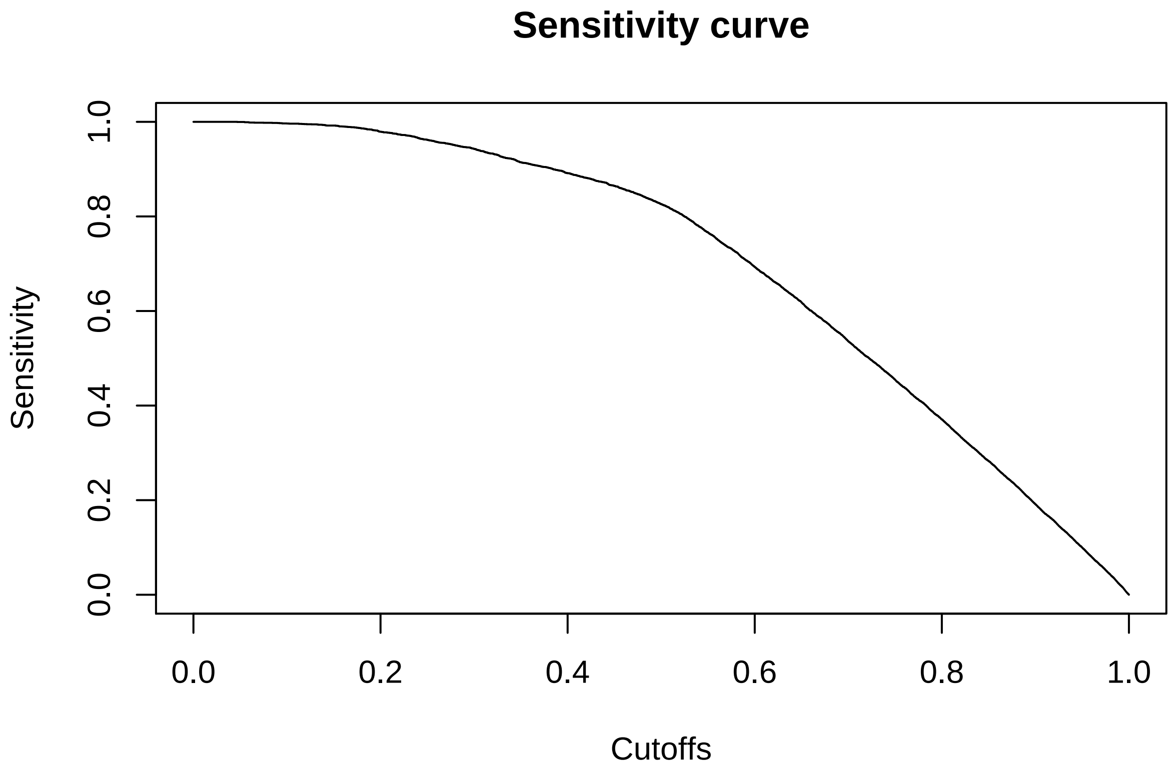
5. **Model Diagnostics - Goodness of Fit Tests**

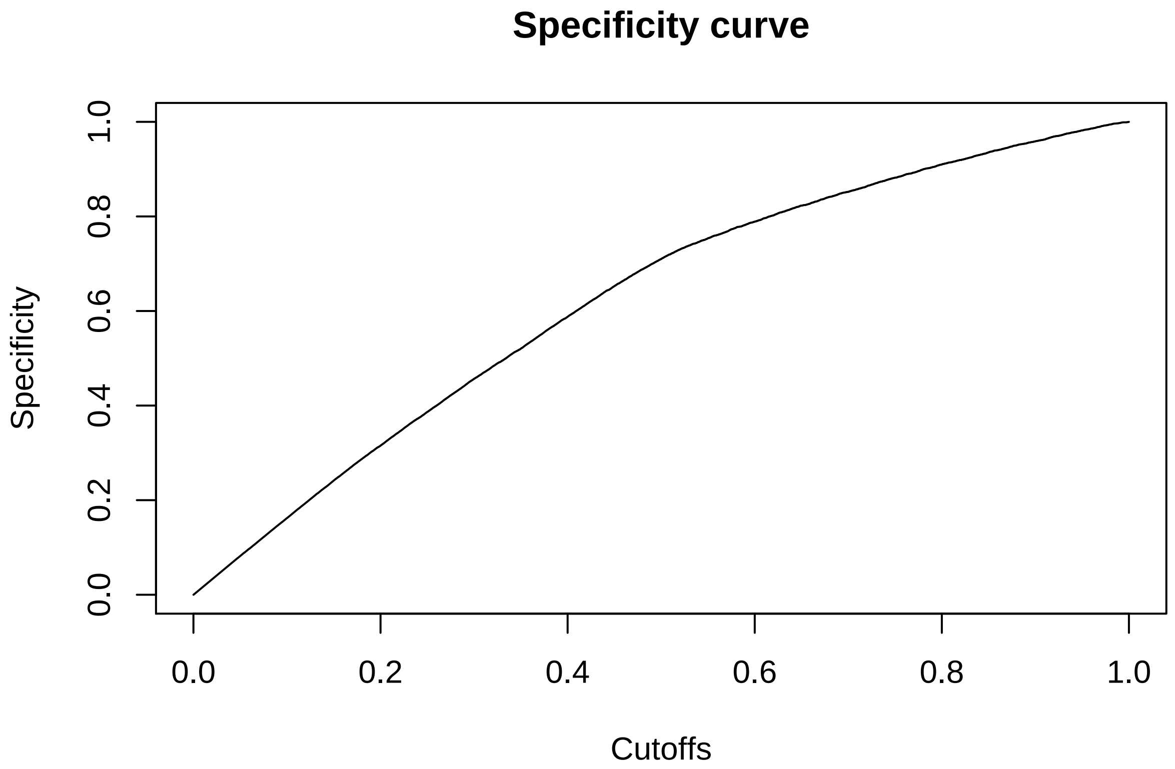
In the following section, we will discuss the goodness of fit test and other model diagnostic metrics to assess the performance of the logistic regression classifier. From the summary of the logistic classifier, printed below, 20 of the chosen 32 predictors are statistically significant for predicting the screening behaviors. In order to further improve the results, we have tried to build the classifier on these 20 predictors but the results didn't improve significantly. So, we have used all the 32 predictors as inputs for the classifier. Now, the following section discusses the goodness of fit test of the classifier with these 32 predictors.

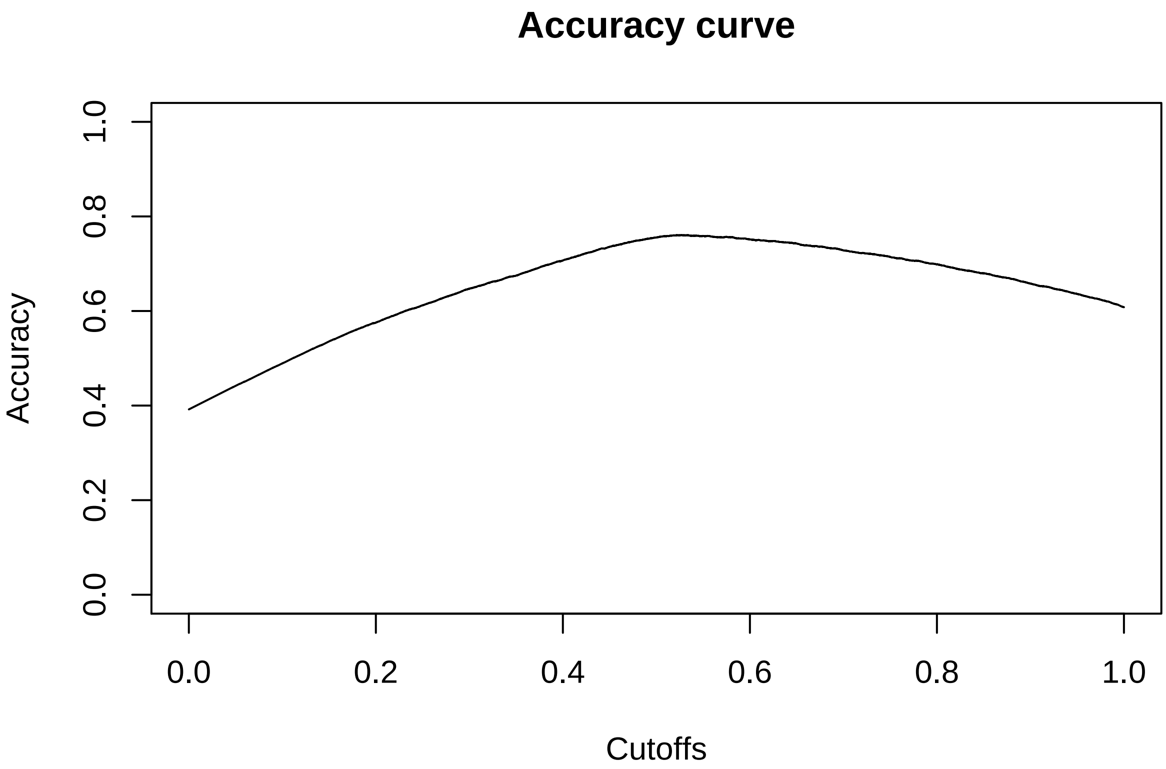
```
##
## Call:
## glm(formula = train.outputs ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1534  -0.6882  -0.3839   0.8750   2.7364
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.6064124  0.2098590 -17.185  < 2e-16 ***
## SEX          2.3206260  0.0365787  63.442  < 2e-16 ***
## AGE_P        0.0151115  0.0014876  10.158  < 2e-16 ***
## OCCUPN1      0.0013368  0.0009074   1.473 0.140684
## OCCUPN2     -0.0045054  0.0021325  -2.113 0.034626 *
## INDSTRN1     0.0042553  0.0007544   5.641 1.69e-08 ***
## AHCCHGYR     0.5382047  0.0278568  19.320  < 2e-16 ***
## LOCALL1B    -0.0016226  0.0009005  -1.802 0.071550 .
## RAT_CAT5    -0.0030863  0.0019193  -1.608 0.107822
## RAT_CAT4     0.0019670  0.0018250   1.078 0.281106
## FM_EDUC1     0.0005908  0.0034630   0.171 0.864537
## ASITENUR     0.0082207  0.0134173   0.613 0.540077
## FHICOST      0.0400104  0.0104968   3.812 0.000138 ***
## REGION      -0.0465281  0.0153849  -3.024 0.002492 **
## ASINKNT      0.0185760  0.0161053   1.153 0.248743
## R_MARITL    -0.0501879  0.0077441  -6.481 9.12e-11 ***
## ASINHELP    -0.0135275  0.0170068  -0.795 0.426373    
## FM_STRCP    -0.0556725  0.0167328  -3.327 0.000877 ***
## ASINCNTO    -0.0417508  0.0178553  -2.338 0.019372 *  
## ASINTRU      0.0400160  0.0152694   2.621 0.008776 ** 
## INCGRP4      0.0006897  0.0010365   0.665 0.505767    
## FM_ELDR     -0.2101820  0.0355299  -5.916 3.31e-09 ***
## FM_TYPE      0.5649076  0.1762453   3.205 0.001350 ** 
## FANYLYN      0.1824148  0.0384376   4.746 2.08e-06 ***
## FHCDVYN     -0.3733536  0.0333855 -11.183  < 2e-16 ***
## FPENSYN     -0.1464590  0.0471840  -3.104 0.001909 ** 
## FOPENSCT     0.1932419  0.0419083   4.611 4.01e-06 ***
## PDSICKA     -0.0822540  0.0172720  -4.762 1.91e-06 ***
## WRKLYR4     -0.1091839  0.0195009  -5.599 2.16e-08 ***
## TBSMKEXP    -0.1211046  0.0233027  -5.197 2.03e-07 ***
## AHCCHGHI     0.3172639  0.0348510   9.103  < 2e-16 ***
## AHICOMP     -0.0873400  0.0205162  -4.257 2.07e-05 ***
## WRKHLTH2    -0.5714225  0.0379072 -15.074  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 31628  on 23569  degrees of freedom
## Residual deviance: 23779  on 23537  degrees of freedom
## AIC: 23845
## 
## Number of Fisher Scoring iterations: 6
```

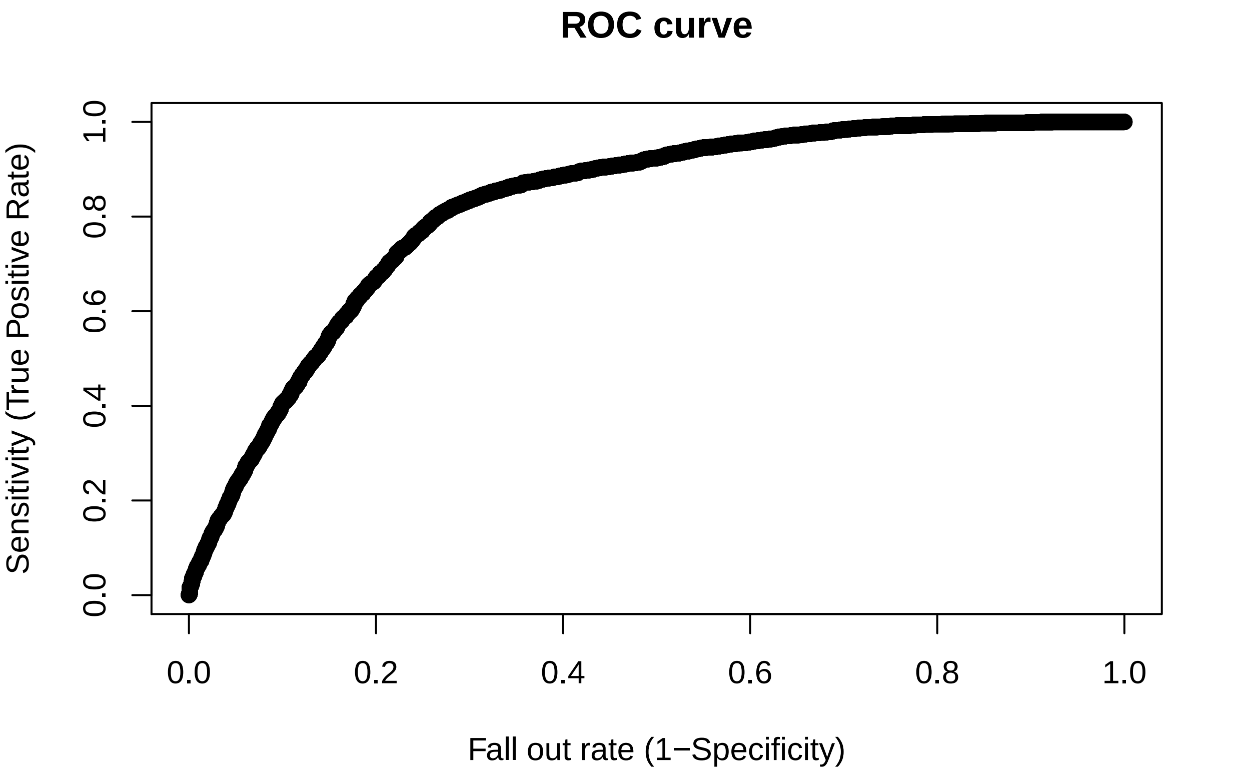
```
## 
##  Hosmer and Lemeshow goodness of fit (GOF) test
## 
## data:  logit.model$fitted.values, as.numeric(train$train.outputs)
## X-squared = -42770, df = 8, p-value = 1
```

```
##           llh       llhNull            G2      McFadden          r2ML 
## -1.188975e+04 -1.581423e+04  7.848970e+03  2.481615e-01  2.832347e-01 
##          r2CU 
##  3.834490e-01
```


Sensitivity curve
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Cutoffs
0.0
0.2
0.4
0.6
0.8
1.0


Specificity curve
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Cutoffs
0.0
0.2
0.4
0.6
0.8
1.0


Accuracy curve
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Cutoffs



Looking at the summary of the logit model, it is obvious that there are multiple significant predictors and the residual deviance is smaller compared to the null deviance implying that the fit of the model is good. Also, Hosmer and Lemeshow goodness of fit (GOF) test confirms that the model is good with a p-value of 1. From the above plots, AUC values(the area under the curve) for all the classification metrics are high and close to 82% which implies that the logistic regression classifier models the data well and that the conclusions drawn from the results have a strong statistical significance.

Conclusion

From the above results, it is clear that social context, like family-related factors such as family structure (have family members who were limited in any way, or aged 65 years and older), family income, as well as community-related factors such as living in a close-knit and helpful neighborhood, exposure to second-hand smoke at work were predictive of cancer screening behavior among men and women in the US. For a more accurate model, individual level factors should be controlled for (e.g., age, income, education level) because they are strong confounding variables that affect the accuracy of the model. Another limitation is the presence of multicollinearity between predictors, which obscures obvious relationships between the predictors and the output variable. Using techniques like principal component analysis that account for multicollinearity between variables would be better. One take away for policy makers is that they should target the following sections of the society, like medical facilities, insurance organizations, neighborhoods with poor living conditions, work place surroundings etc,in order to improve screening behaviors.

References

1. Barry, J., & Breen, N. (2005). The importance of place of residence in predicting late-stage diagnosis of breast or cervical cancer. Health & place, 11(1), 15-29.
2. Calle, E. E., Flanders, W. D., Thun, M. J., & Martin, L. M. (1993). Demographic predictors of mammography and Pap smear screening in US women. American Journal of Public Health, 83(1), 53-60.
3. Dillard, A. J., & Main, J. L. (2013). Using a health message with a testimonial to motivate colon cancer screening associations with perceived identification and vividness. Health Education & Behavior, 40(6), 673-682.
4. Gandhi, P. K., Gentry, W. M., Kibert II, J. L., Lee, E. Y., Jordan, W., Bottorff, M. B., & Huang, I. C. (2015). The relationship between four health-related quality-of-life indicators and use of mammography and Pap test screening in US women. Quality of Life Research, 24(9), 2113-2128.
5. NHIS (2016). Retrieved from https://www.cdc.gov/nchs/nhis/nhis_2015_data_release.htm
6. Selvin, E., & Brett, K. M. (2003). Breast and cervical cancer screening: sociodemographic predictors among White, Black, and Hispanic women. American journal of public health, 93(4), 618-623.
7. Straughan, P. T., & Seow, A. (1998). Fatalism reconceptualized: a concept to predict health screening behavior. Journal of Gender, Culture and Health, 3(2), 85-100.
8. Weaver, K. E., Ellis, S. D., Denizard-Thompson, N., Kronner, D., & Miller, D. P. (2015). Crafting Appealing Text Messages to Encourage Colorectal Cancer Screening Test Completion: A Qualitative Study. JMIR mHealth and uHealth, 3(4).
9. Wong, R. K., Wong, M. L., Chan, Y. H., Feng, Z., Wai, C. T., & Yeoh, K. G. (2013). Gender differences in predictors of colorectal cancer screening uptake: a national cross sectional study based on the health belief model. BMC Public Health, 13(1), 1.