Applied Data Science: Final Project Report

**Smoking Behavior Analysis**

**Team Members:**

Marvin Mananghaya (msm796)

Rufei Sheng (rs6431)

Tanya Nabila (tn1050)

Urwa Muaz (um357)

Xiao Jing (xj655)

**Individual roles and contributions:**

Marvin Mananghaya: Data acquisition and cleansing, exploratory data analysis , paper report & presentation, Elohim, Documentarist

Rufei Sheng: Acquire and process the tax data; build classification model(Random Forest (RF) Classifier, K-nearest-neighbour Classifier, Naive Bayes (NB) Classifier, Random guessing); Performed counterfactual causality test, paper report & presentation

Tanya Nabila: Data wrangling & preprocessing, exploratory data analysis, feature engineering, code integration & management, catering services (cookies and snacks), paper report & presentation,

Urwa Muaz: Feature engineering, logistic regression, random forest, class imbalance treatment, hyper parameter tuning, model evaluation.

Xiao Jing: Chi-square test, code integration & management, model explanations, Paper Report & Presentation

**Abstract**

Smoking has been the significant contributor to many health problems. This paper investigates the specific features that have correlations with the electronic smoking(e-smoking) behaviors and then trains models for prediction. From the logistic regression model we find correlated features including health, demographic features, life habits and behaviors, socioeconomic features, tax and price of tobacco. There is also a causal relationship between cigarette tax policy and smoking behavior. Next, classification models(RFC) are trained to predict a potential e-smoker. In conclusion, the drinking socioeconomic, health and lifestyle conditions are the most important factors that influence people in choosing smoking behavior.

**Introduction**

Smoking has been the significant contributor to many health hazards. Understanding the smoking behaviors and its relation with socioeconomic and behavioral variables is vital to formulate relative public policy. This report focus on two questions.

1) What behavioral, health and socioeconomic factors influences an individual's habit to smoke and more specifically which of these factors is more likely to make you an e-smoker. For completeness we also including external factors that could be important like state laws, tax rates and cigarette prices.

2) Secondly we investigate to what extent can we predict a person's smoking behavior based on the socio economic, lifestyle and health data. In this phase we explore the non linear hypothesis space and aim for accuracy rather than interpretability.

There have been numerous studies about the effect of these characteristics on smoking behavior, reviewing all the literature was beyond the scope of our project. But we observed that all these features are really modelled in unison, so that is what we perceive to be our novel contribution to the research problem. The insights generated by this study can aid the government in designing effective tobacco policy and evaluating the existing programs.

**Methodology**

**Data Gathering**

**BRFSS -** Our primary source of data comes from the Behavioral Risk Factor Surveillance System (BRFSS) from the Centers for Disease Control and Prevention (CDC). The database is a system of ongoing health-related telephone surveys designed to collect data on health-related risk behaviors, chronic health conditions, and use of preventive services from the non-institutionalized adult population (≥ 18 years) residing in the United States.

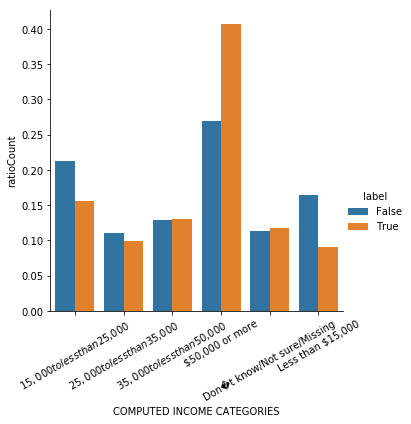
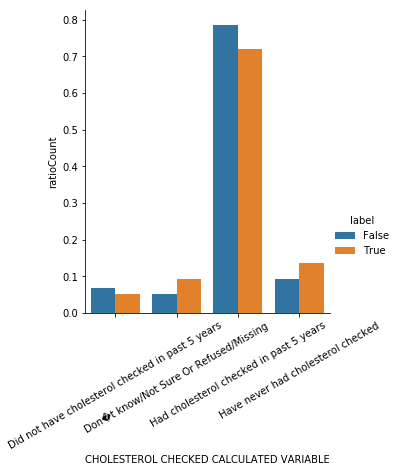
**State Level Features:**

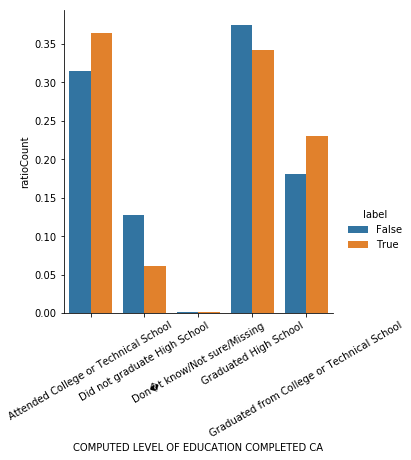
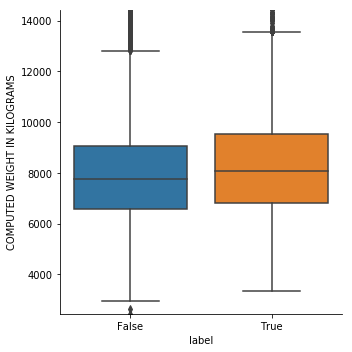
* **Cigarette Retail Tax** - The amount of cigarette tax in each state was gathered from the State Tobacco Activities Tracking and Evaluation (STATE) System records.
* **US State Shapefile -** We performed spatial residual analysis of our linear model by visualizing it over the State boundaries shapefile.
* **Tobacco Retail Price -** This data was captured from U.S. Food & Drug Administration
* **Prevalence of key search terms on Google** : We used google trends to extract ‘Mormon’ key term and used it as proxy for state-wise prevalence on anti-smoking religious doctrine.

**Exploratory Data Analysis-Data Distribution Plot**

We plotted the labels against all our observations as our initial exploratory analysis to decide which features seem correlated or of significance according to our observation. Our labels for the exploratory analysis are of smoker type (traditional smoker / e-cigarette smoker). We plotted the categorical features with bar chart type and the numeric features with box chart.

The group of 40 plots shows that the following list of columns,which have significant difference distribution between e-smoker group, labeled as True, and traditional smoker group, labeled as False.





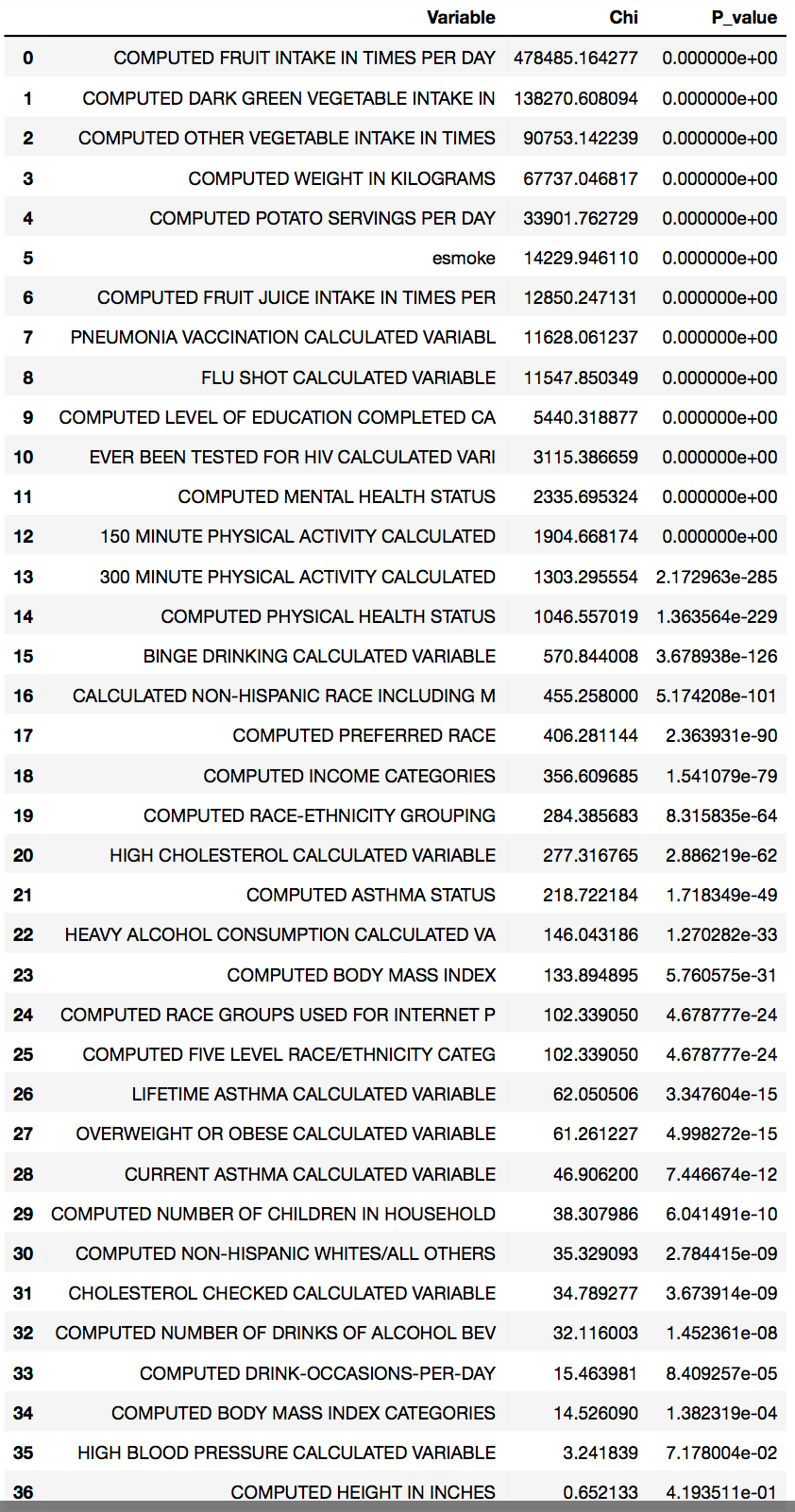
**Feature Engineering**

We transformed certain continuous features columns into ordinal or binary ones. These are a few of the features we included.

|  |  |
| --- | --- |
| Significant Features from Plotting | Feature Type |
| High cholesterol | Binary |
| Asthma | Binary |
| Computed Body Mass Index(Obese/Not Obese) | Ordinal |
| Level of education | Ordinal |
| No. of children(< 3/others) | Ordinal |
| Computed income(<50k/others) | Ordinal |

**Hypothesis Testing - Chi-square test**

We formulated Chi-square hypothesis test to evaluate the significance of their linear relationship with the smoking habits, and used to inform feature selection. According to the results, 34 out of 36 features were significantly related to the label [smokers/non smokers]. In subsequent analysis for [esmokers/smokers], we will put more emphasis on the 28 out of 36 features with p-value less than 0.05 in the following part. The 6 features not passed the second test are mostly disease related features, which means e-smoker probably will not improve the health status.



**Class Imbalance**

We observe the samples sizes of two groups have large differences, which was giving us a false impression of high accuracy. However, precision recall analysis informed our evaluation and decision to treat the imbalanced classes. We tried both upsampling the minority class and downsampling the majority class and empirically deduced the better strategy for our model based on Area under the curve(AUC). We chose upsampling technique.

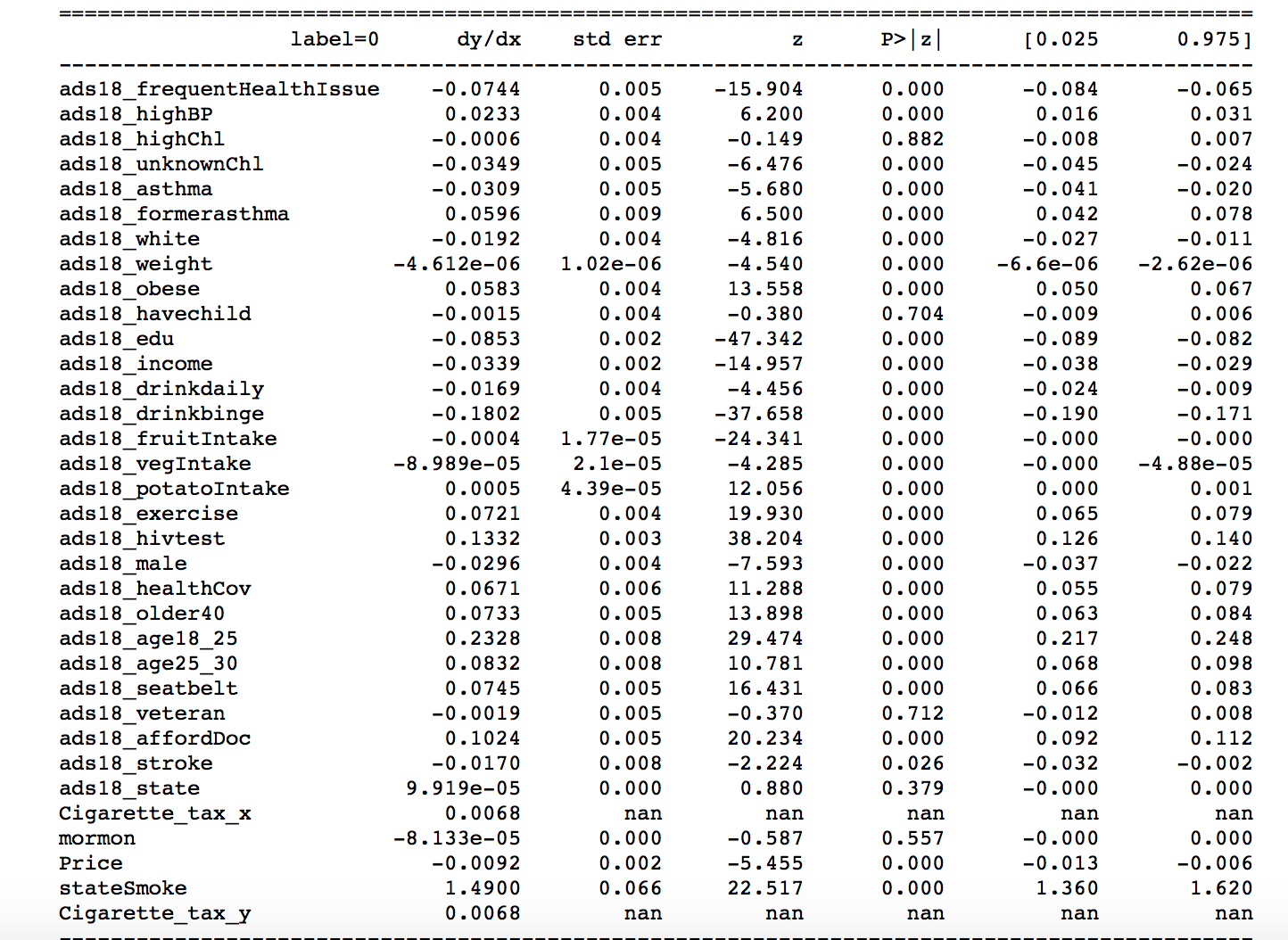
**Model building**

In this part, we separate our model fitting process into two phrases because of two different goals.

**Logistic Linear Regression**

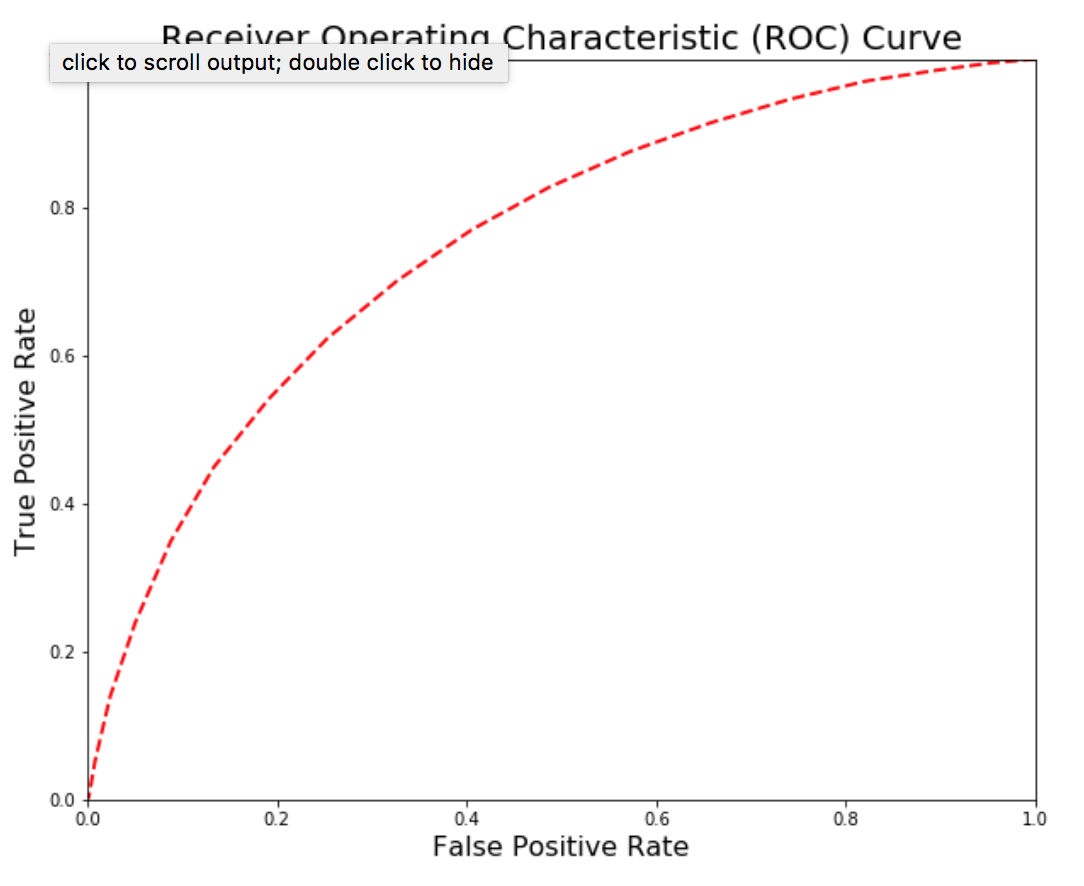
We iteratively fit a series of logistic regression models to examine the linear relationship between each independent variable with dependent variable( e-smoker/non e-smoker). Our goal is to interpret the coefficients and their significance to evaluate their relative importance in determining the label class.

* **Smoker / non-smoker: Label 0 = smoker**

****

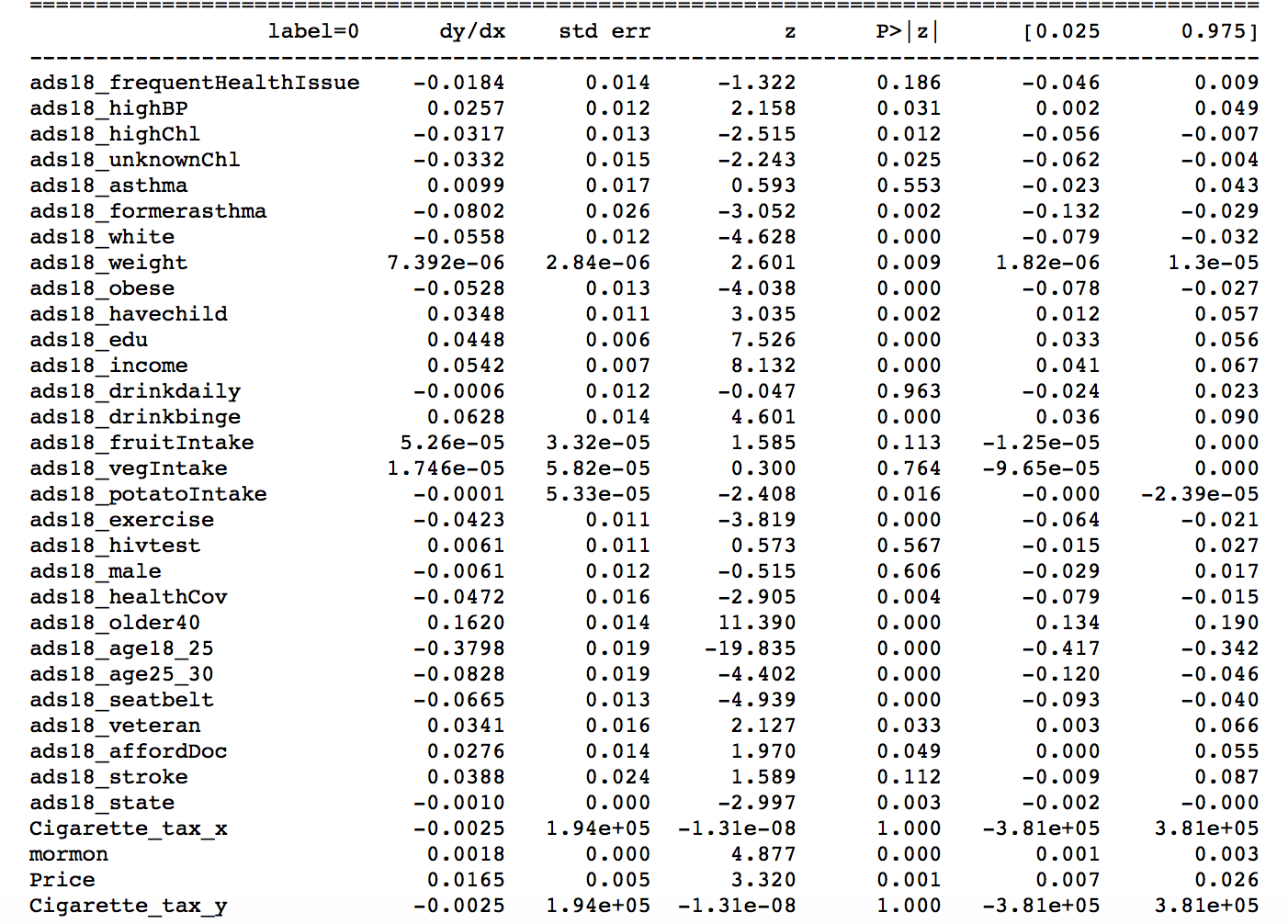
The second column of table above shows the marginal effect, which measures the impact of the input variable on the output class. The negative sign would mean that increasing this feature would make an individual more likely to be a smoker and vice versa. The magnitude captures the intensity of the impact.

|  |  |  |
| --- | --- | --- |
| **Best threshold** | **Highest Accuracy** | **AUC** |
| **0.5** | **0.688** | **0.752** |

****

We see that our logistic classifier is actually doing pretty well with an AUC of over 0.7. Which is decent performance considering that we are modelling human behavior. Drinking habits, frequent and acute medical issues, smoking price and health awareness appear to be major drivers behind smoking.

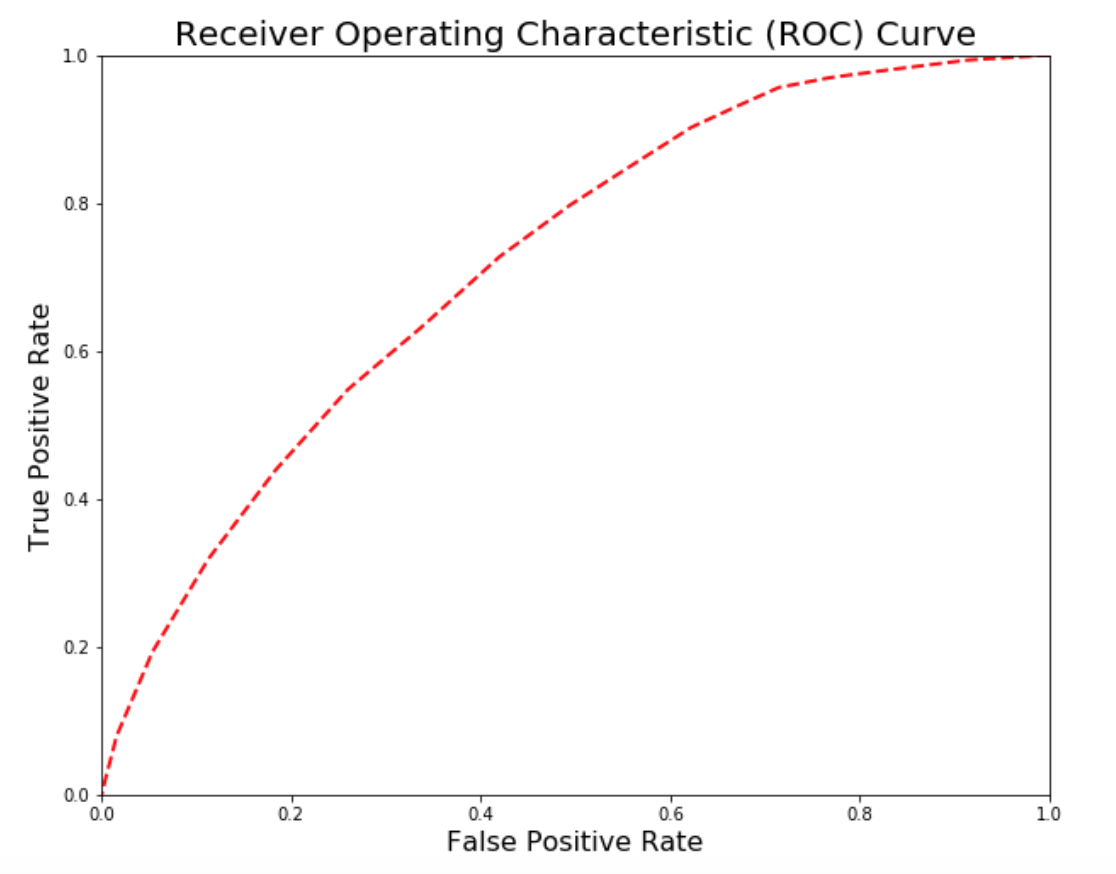
* **non-smoker/ smoker**



The table shows marginal effect of smoker e-smoker classifier.

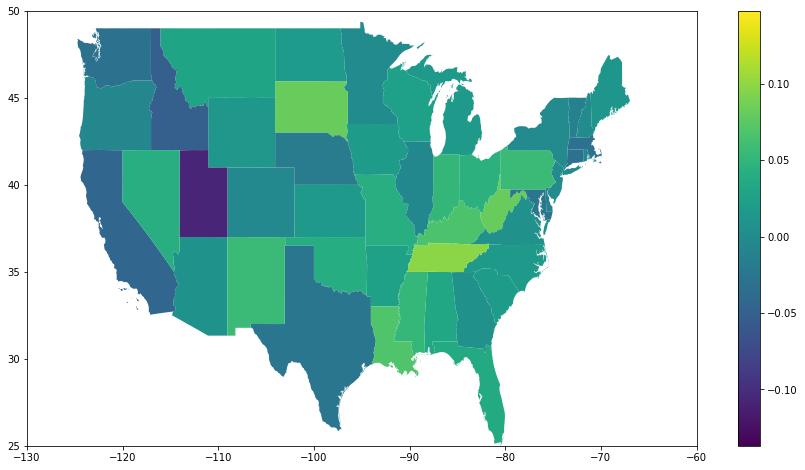
|  |  |  |
| --- | --- | --- |
| **Best threshold** | **Highest Accuracy** | **AUC** |
| **0.5** | **0.65** | **0.72** |

Age is an important predictor for this classifier, probably because e-smoking is a more prominent trend in younger people. We also find that people with disease history and chronic smoking related illness are more likely to esmsmoke.

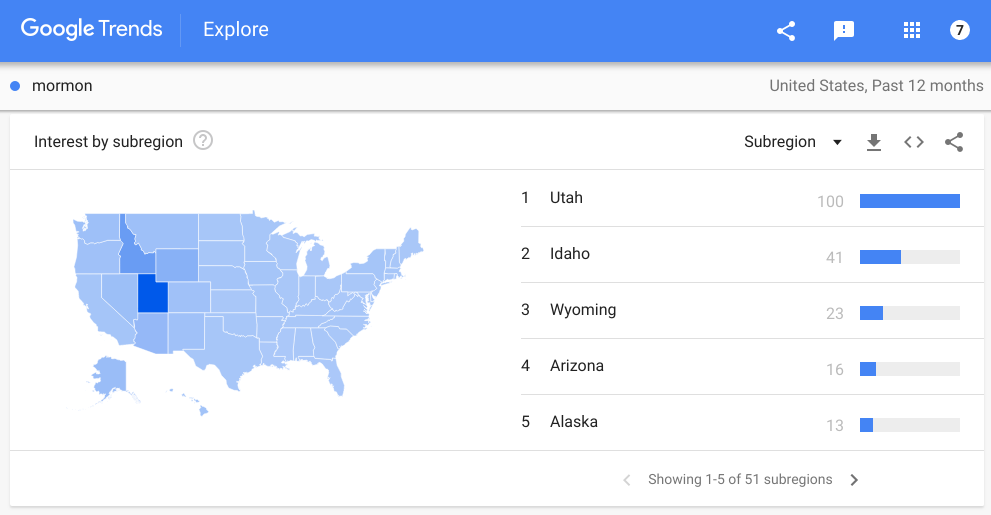


### **Residual Analysis**

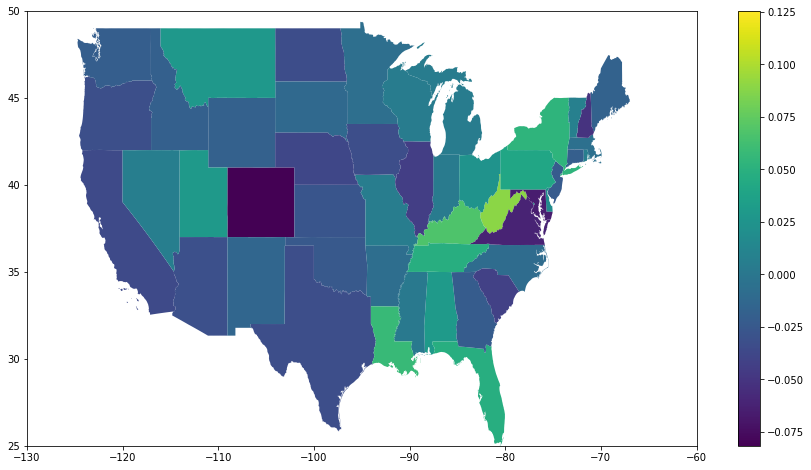
In order to consider the spatial heterogeneity of our features, we map the residuals of the logistic regression onto the basemap of US, at the state level, as a way to explore possible features to improve the predictive power of our model.



The map initially showed a high error on specifically Utah and Idaho, and it was made known to the group that maybe the Mormon population could be influential to Tobacco consumption as they abstain from smoking. As a proxy to the effect of religious reverence and practices, we incorporated google trends data using the keyword “mormon” as a proxy.

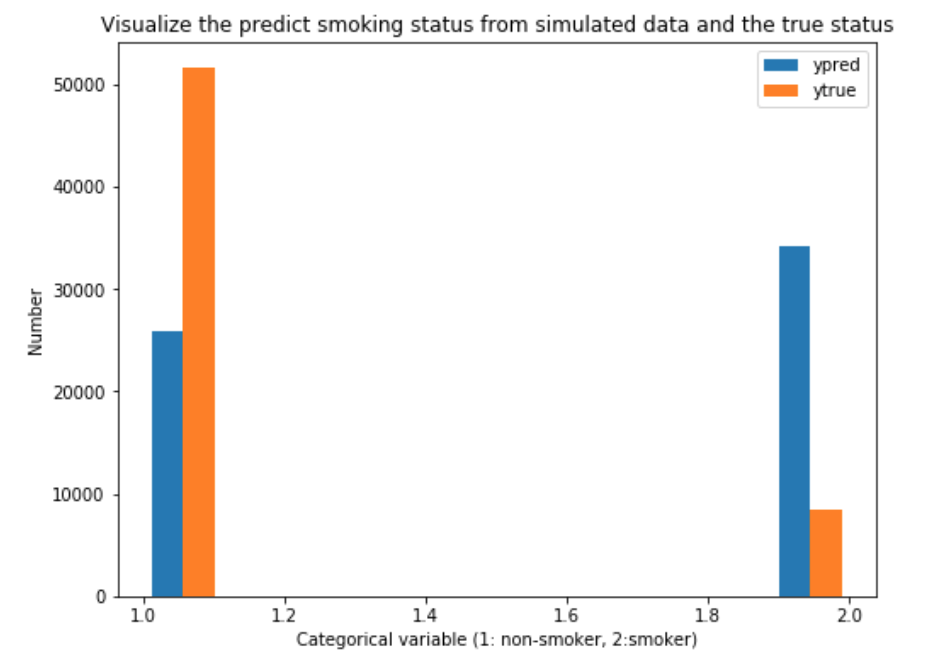
****

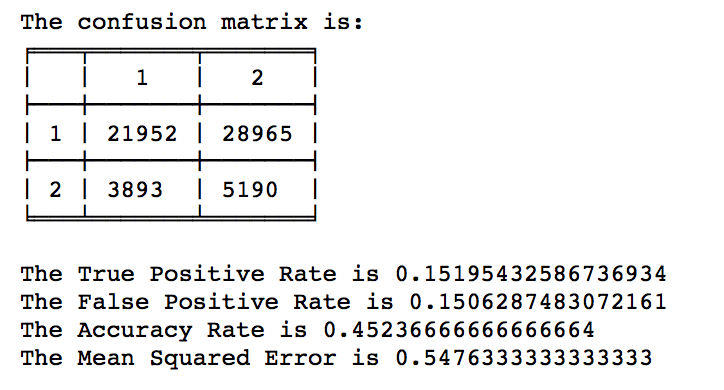
This proved effective in reducing the bias of the model as seen in the figure below that the error measure has substantially decreased. Later, added features such as cigarette tax and cigarette retail price effectively as consumption will be function of these features.



### **Counterfactual Causality Test**

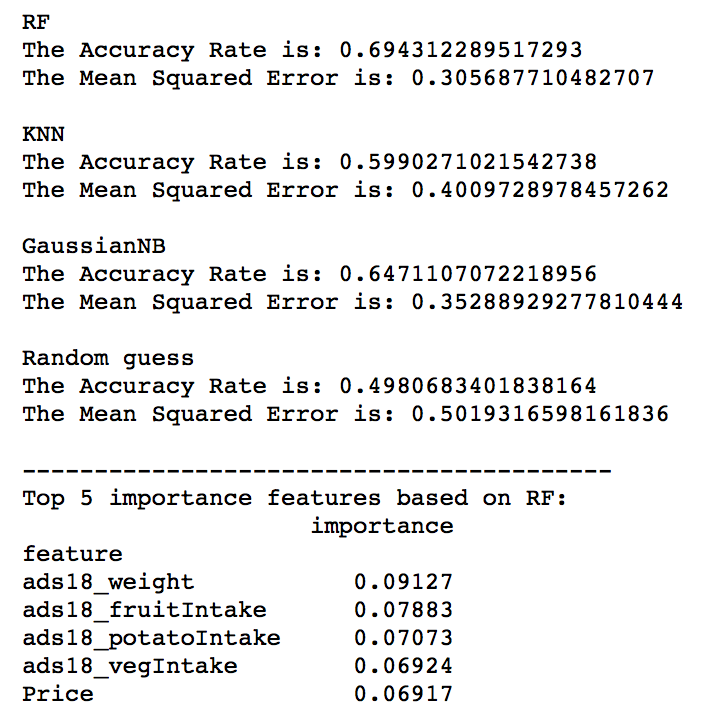
On April 1, 2009, the federal cigarette excise tax increased from 39 cents to $1.01 per pack; thus, the idea of the test was to test the causation between cigarette tax and smoking by evaluating whether there was a difference between the group who did exposure the changing tax policy (2012 data) and who didn’t exposure (2008 data). Cigarette smoking is always associated with cancer death and chronic diseases, and it is suggested that most people start to quit the smoking because of health reasons (Schoren, 2017). Therefore, the health consciousness related variable “routine check-up”, which indicated how long has it been since the respondent last visited a doctor for a routine check-up, was selected as the feature of interest. Other variables such as age, sex, and income were chosen to preserve the moment of the future set through sampling with replacement. Those four variables consisted the synthetic Dataframe that ‘look like’ the actual Dataframe. First, we feed the Random Forest (RF) model with 2008 true data, then using the synthetic Dataframe to see the different between the predicted e-smoker and true e-smoker (from 2012 data). From the histogram, it indicated that there was large difference between the fake results and the true results; and the accuracy rate for the prediction was just 45.24%. In addition, the number of predicted smokers was far greater than the number of true smokers. Therefore, it was reasonable to infer that the tax policy had huge impacts on the smoking behavior, resulting the difference between counterfactual and real-life data.



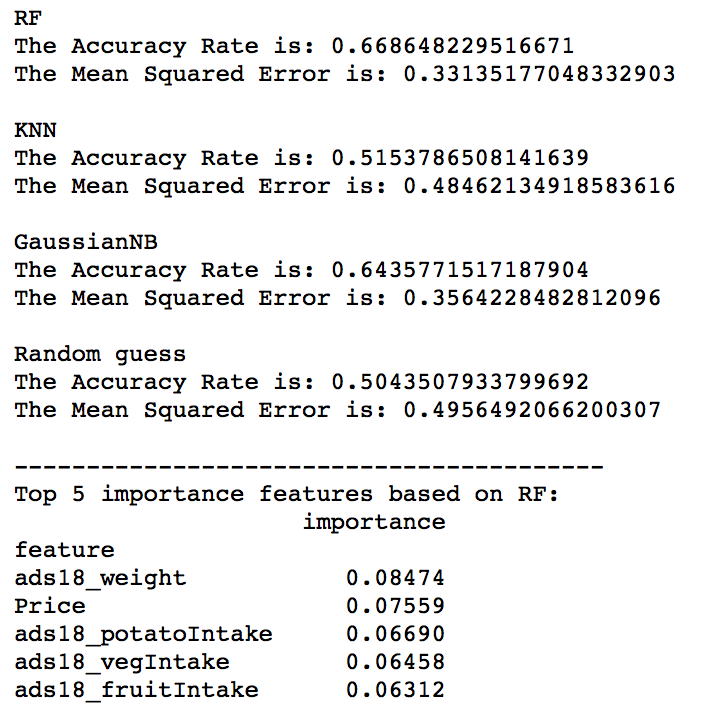


**Classification Model**

Three classification methods were selected to build the classification model: Random Forest (RF) classifier, K Nearest Neighbors (KNN) classifier, and the Naïve Bayes (NB) classifier. The KNN predictors will determine the nearest K points in the training set based on a Euclidian distance metric. Given the known labels in the training set, the predictor will assign a label based on a majority vote. Random Forest is a non-parametric advanced classification and regression tree (CART) analysis method that have an inherent procedure of producing variable importance and has been adopted widely in many scientific fields. And naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features. All those methods scale very nicely with massive data because the algorithms can be easily parallelized. We trained the model aiming to predict whether a person is a smoker, or whether a person is a e-smoker. The prediction results are shown below, compared with the result of random guessing.



The objective for above model was to predict whether a person is smoker or non-smoker. The three classification methods all done better than the random guess, and Random Forest method indicated the best prediction ability, around 0.694 accuracy rate and 0.306 mean squared error. Based on RF, the top 5 important factors that influence a person to be a smoker are weight, fruit intake, potato intake, vegetable intake, and cigarette price.



The objective for above model was to predict whether a person is e-smoker or smoker. The three classification methods all done better than the random guess, and Random Forest method indicated the best prediction ability, around 0.67 accuracy rate and 0.33 mean squared error. KNN predicted worst since its accuracy rate only 0.01 higher than the random guess. Based on RF, the top 5 important factors that influence a person to be a smoker are weight, cigarette price, potato intake, vegetable intake, and fruit intake.

**Conclusion**

The increased tax rate on cigarette in 2009 has the casual relationship with the smoking behavior in a large degree. People’s smoking behaviour will be changed as the consequence of changing tax policy. Therefore, in order to effectively reduce the smoking, tax indeed a useful manner.

We see that the drinking habit is highly correlated with the habit to smoke and so is the frequency of health issues. It is interesting to note that people who do not know their cholesterol level are more likely to smoke, because they are probably not that concerned about their health. Furthermore, habit of exercising and more education are features that make you less likely to smoke.

Socioeconomic, health and lifestyle conditions can determine the smoking habit of the individual with fair accuracy. We get AUC greater than 0.7 which is very good when modelling human behavior. We also see that there is a certain spatial trend, because after accounting for all the known variables the residuals still show some for some states model over predicts the smoking occurence and for some it under predicts it.

**Limitation & Improvement**

The Behavioral Risk Factor Surveillance System(BRFSS) dataset being used is answered by phone interviewees and the answer to some questions, as features, are descriptive sentences according to individuals’ own standards. So we have perceived health and socioeconomic condition than the actual one. We based some of our feature engineering on this fact. We created indicator column for missing value in cholesterol level, and the individuals who didn’t know there cholesterol level turned out to be significantly more inclined towards smoking. The ones with former acute breathing/heart disease were less likely to smoke than the ones with current disease. So a lot revolves around one’s own perception of his health. There might be an inherent bias in the sampling technique because the people who are willing to talk might not be representative of the whole population. Lastly, we observed some multicollinearity in our feature space and logistic regression model operates under the assumption of independence of feature space.

**Reference**

Schoren, C., Hummel, K., Vries, H. D. (2017). Electronic cigarette use: comparing smokers, vapers, and dual users on characteristics and motivational factors. Tobacco Prevention & Cessation, 3(April), 8. https://doi.org/10.18332/tpc/69392  
  
Warner KE. Smoking and Health Implications of a Change in the Federal Cigarette Excise Tax. JAMA.1986;255(8):1028–1032. doi:10.1001/jama.1986.03370080050024