**Introduction:**

Physical health is the state of being free from illness or injury.[1](http://www.bridgewater.nhs.uk/bolton/the-parallel/physical-health/) It can cover a wide range of areas including healthy diet, healthy weight, dental health, personal hygiene and sleep. Physical health is vital for overall well-being. Chronic illness like cancer, diabetes, asthma, arthritis cannot be cured but can be managed. Putting into practice the old saying 'prevention is better than cure' – modelled physical health not being good for more than 14 days to identify the most significant contributors.

The data used for the analysis is “500 Cities: Local Data for Better Health” project from the Robert Wood Johnson Foundation. Three types of health-related data that are tracked in the project

1. **Health Outcomes**

Arthritis, asthma, high blood pressure, high cholesterol, cancer, diabetes, kidney disease, pulmonary disease stroke, mental health not good for >14 days, physical health not good for >14 days, all teeth lost

1. **Prevention**

Current lack of health insurance, visits to doctor within past year, visits to dentist, taking medicine for BP, cholesterol screening, mammography, pap smear, fecal occult blood test, up to date on preventative services for men (or women)

1. **Unhealthy Behaviors**

Binge drinking, smoking, no leisure time physical activity, obesity, sleeping less than 7 hours

For the analysis, all the variables from Prevention and Unhealthy Behaviors are considered as independent variables. Three new variables are feature engineered from the existing variables. One of them is the combined variable for “up to date on preventative services for men (or women)” as “up to date on

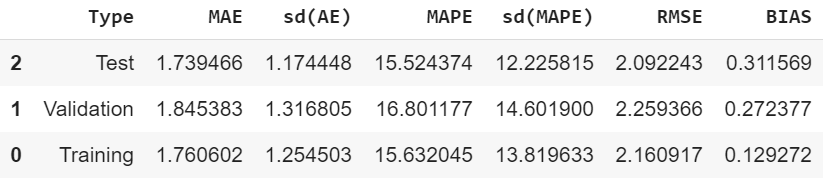
1. <http://www.bridgewater.nhs.uk/bolton/the-parallel/physical-health/>

preventative services”. The others are Unhealthy behavior score and Young Prevention Score calculated based on unhealthy behaviors and prevention categories based of age above 18.

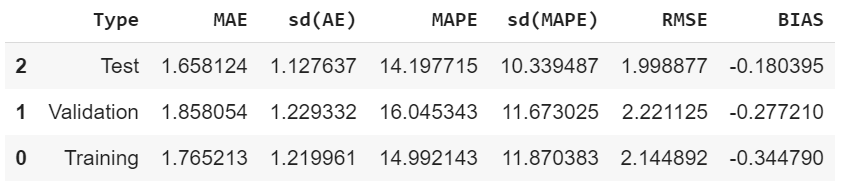
**Models, Methodology and Results:**

Two decision tree models were built and evaluated on the data in terms of error metrics and variable importance is identified on each model. We used Random Forest for Model 1 and Gradient Boosting Machine for Model 2. For easy interpretation, let’s name Model 1 as Model RF and Model 2 as Model GBM. Let’s start with Model interpretation and analyze the results. Comparing the Error metrics of Model RF and Model GBM, we can see that the RMSE of both the models is around 2 which implies a 2% Crude Prevalence error in predicted variables. Bias is positive for Model RF and Negative for Model GBM. Looking at the bias we can say that the model RF's predicted values are lesser than the actual whereas Model GBM's predicted are greater than the actual values. Overall both the models did great. Looking at the test metrics, Model GBM did a little better compared to Model RF.

Error Metrics of RF:



Error Metrics of GBM:



Now looking at the variable importance, we can see that both the models have identified the same top 4 contributors.

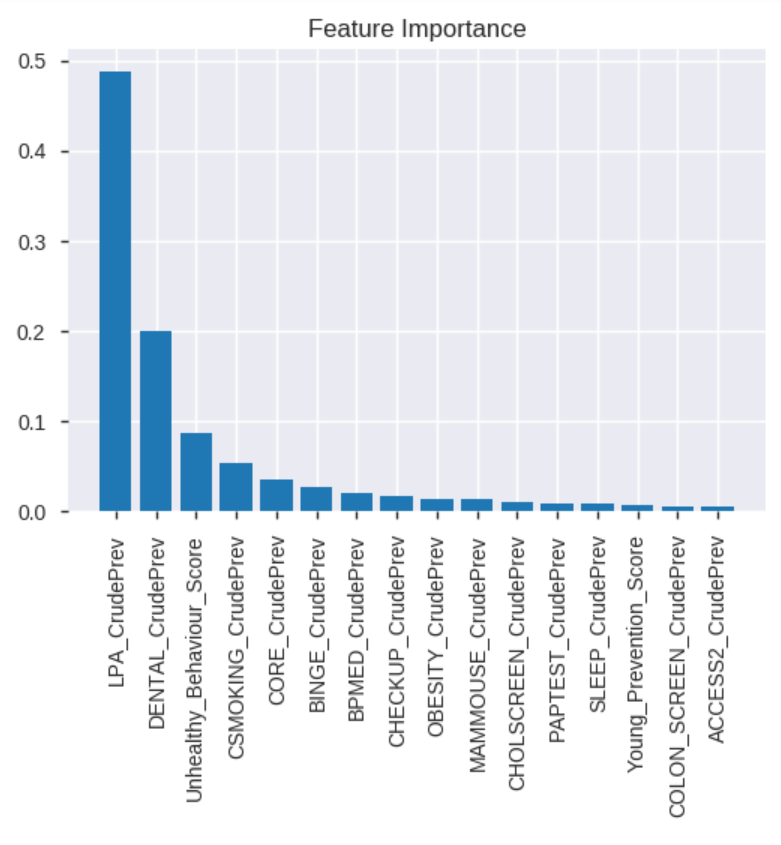
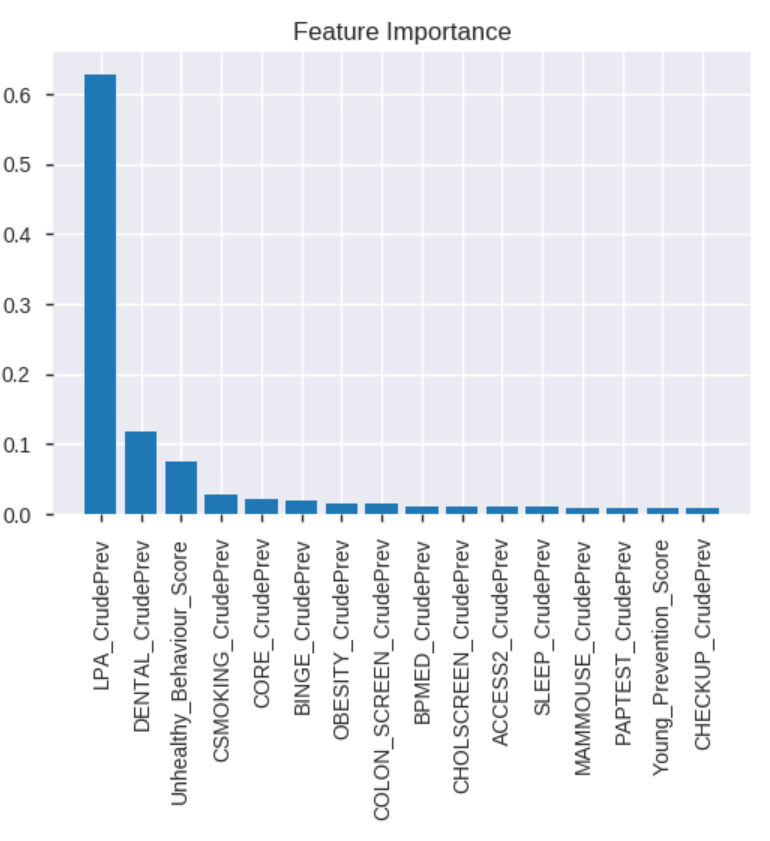
1.No leisure-time physical activity among adults aged ≥18 years

2.Visits to dentist or dental clinic among adults aged ≥18 years

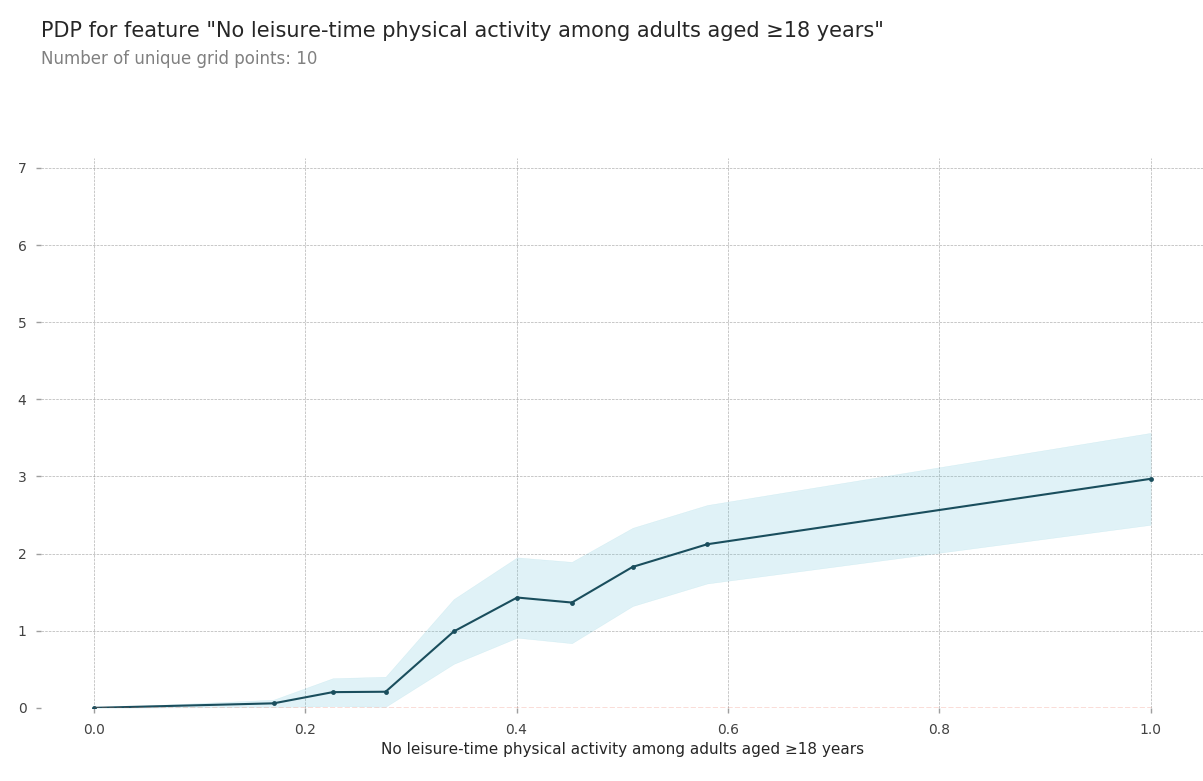
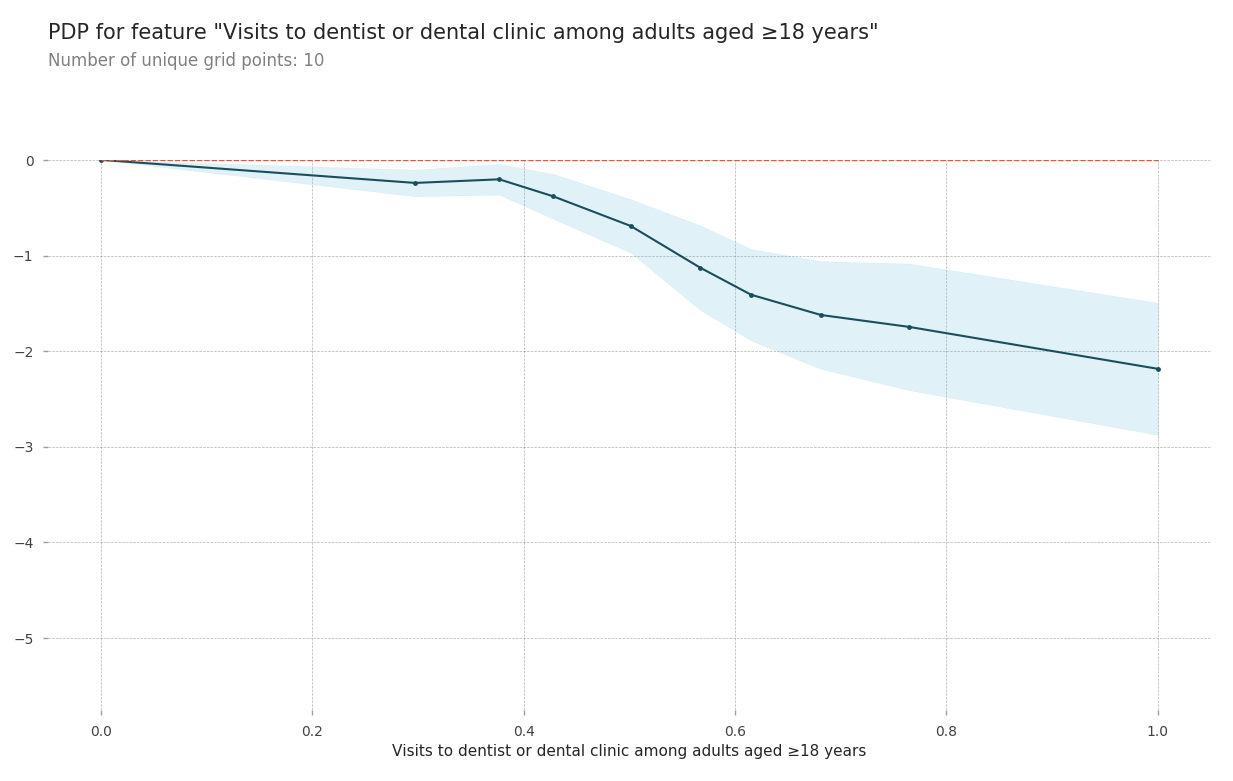
3.Unhealthy Behavior Score

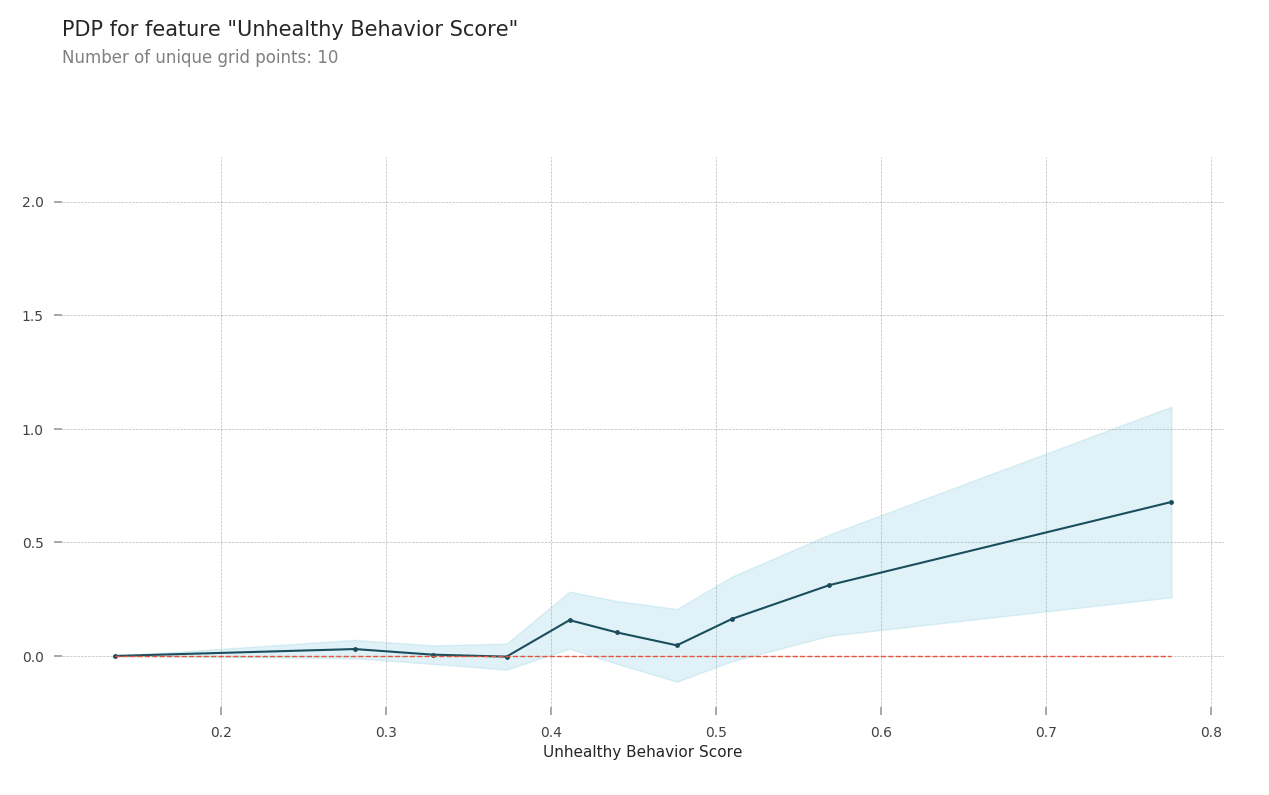
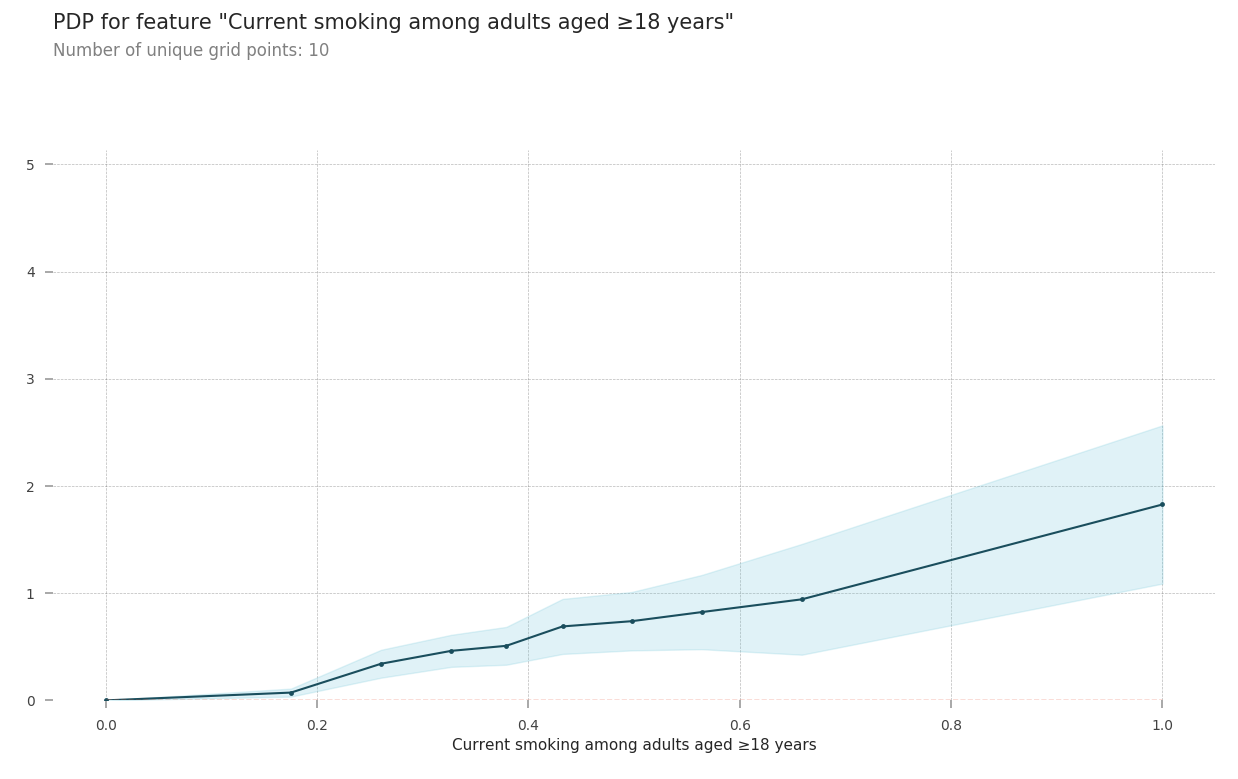
4.Current smoking among adults aged ≥18 years

Feature Importance of RF: Feature Importance of GBM:



The partial dependence plots for these variables are shown below. For variables, LPA, Unhealthy behavior score and current smoking habit, we can see that as the value of the variable increases, the predicted output “physical health not good” increases but as the number of visits to the dentist increases, the predicted output decreases which indicates that people who are cautious and preventive are less likely to get sick.

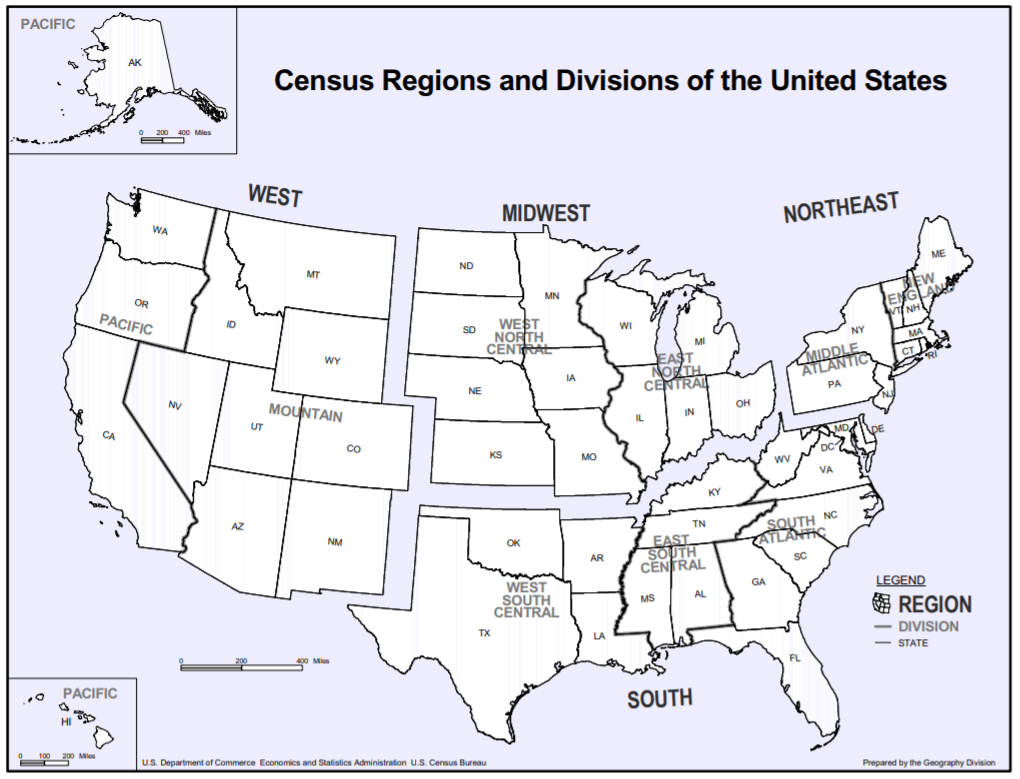
 

**Discussion:**

As we know that physical health and mental health are closely coupled, we might want to try including mental health into the model. We might also include variables like healthy diet index and personal hygiene. As the dataset is obtained from survey that is subjective, to make it more reliable, if annual health check-up results include such surveys to be taken with their doctor, it will make the data more reliable for policy implementations. If the statistics of respondents to the survey are available, we might want to consider and build new variables based on the statistics.

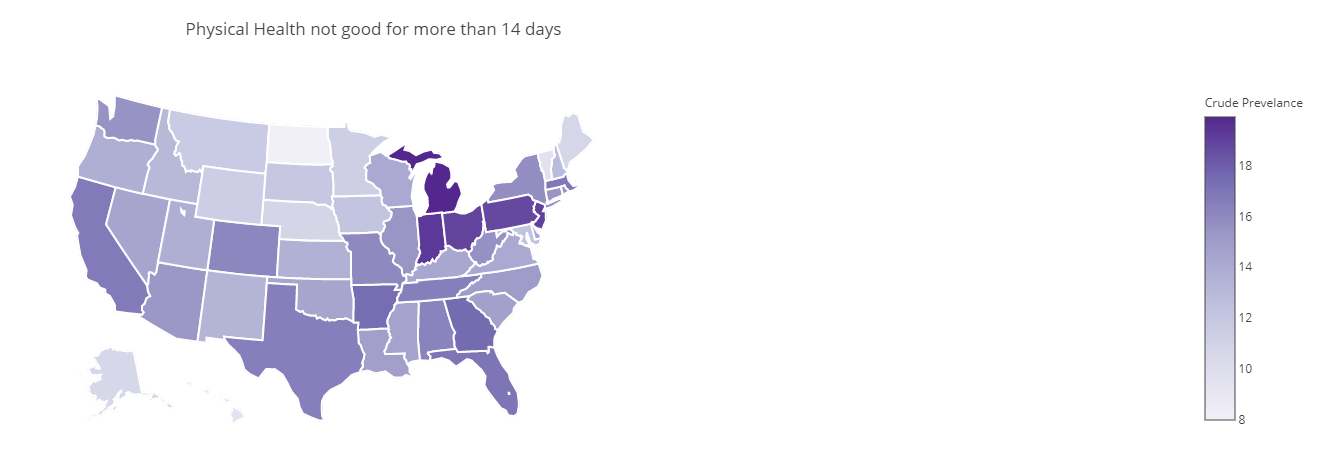
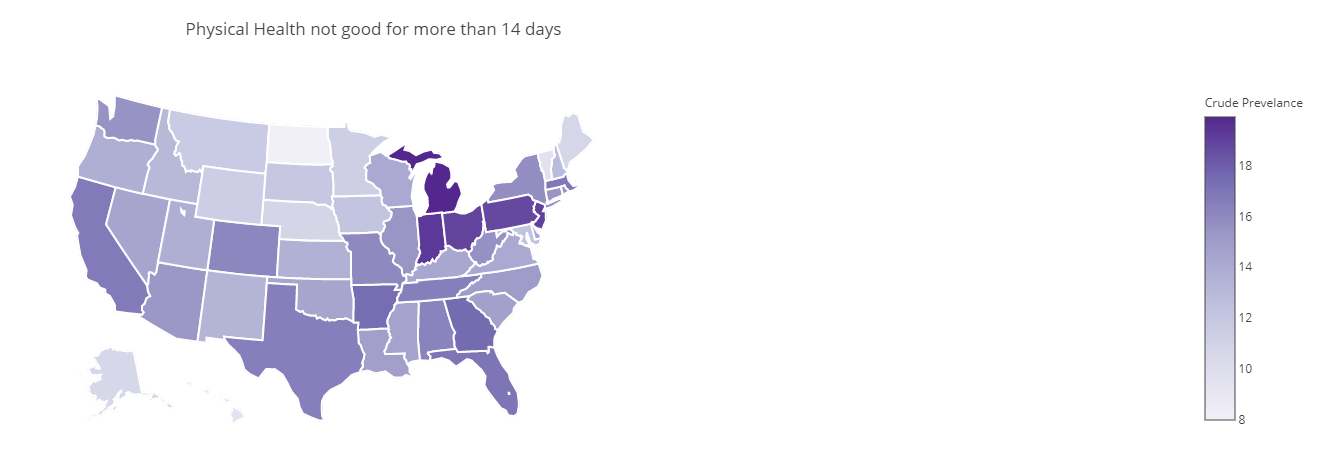
**Regional Analysis:**

To look at the regional results, we tried dividing all the states based on census data.[2](https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf) All the states are divided into four regions – West, Midwest, Northeast and South. Alaska was grouped into region West.

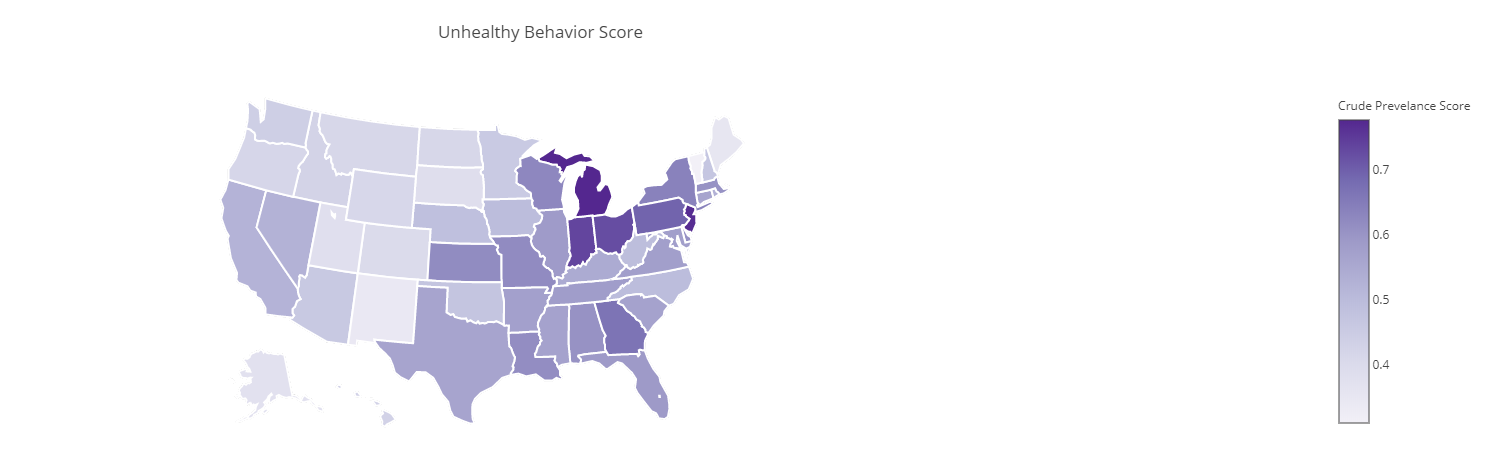
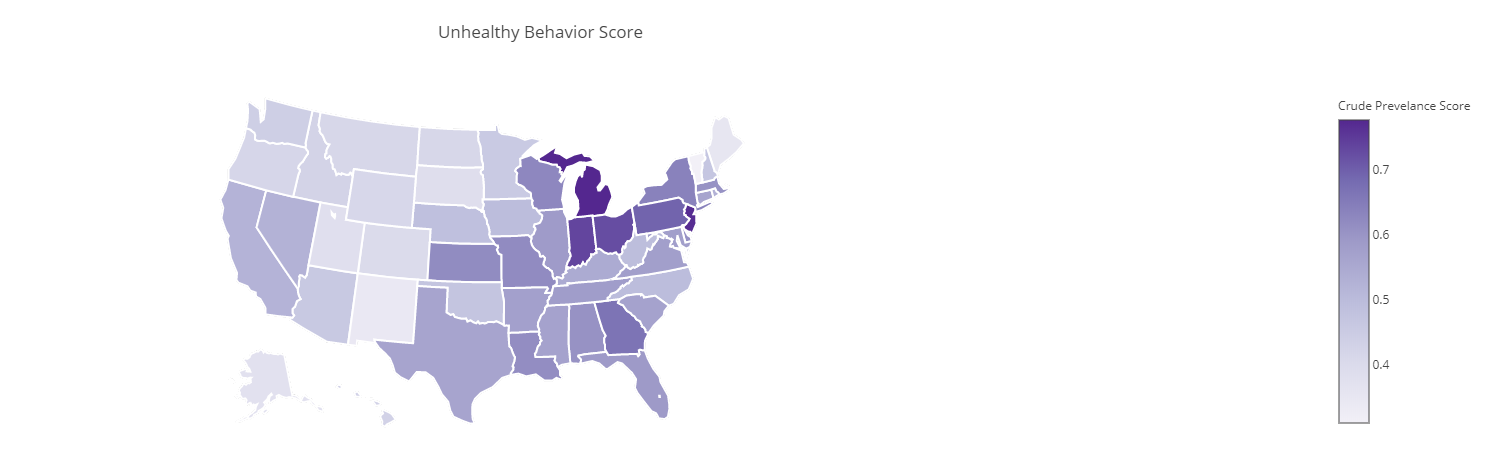


If we look at the trends, in terms of choropleth map based on the highest values of physical health not being good for more than 14 days, we can see the following trends. Live chart available on Colab file. The east north central part looks to be the unhealthiest. West north central turned out to be healthiest region in terms of physical health. Some unhealthy states that fell in other regions were Pennsylvania and New jersey.

1. <https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf>



This trend is very similar to the unhealthy behavior feature engineered column. Below is the plot based on unhealthy behavior score. The pattern is so like the pattern of physical health not being good.



Based on the trends in the dataset, the eastern states are unhealthier compared to the western states.

**Conclusion:**

These models can be used by insurance companies to offer premiums based on the health of the person. Furthermore, if we can get demographic classification on data in order to make better prediction, we can classify and predict physical health based on Age, Gender, Race, etc. We might also try including the job roles and try to see the influence of each job role, depression and stress levels into consideration. These can be categorized based on mental health and behavior patterns. We can also take into consideration the environmental conditions. Last but not the least, water intake, bowel movements can be added to the survey and they can be a valuable predictor of physical health. We will need more information about all these aspects in order to make a better prediction of physical health not being good for more than 14 days.