**What Solutions Can Analytics Provide  
To The Worldwide Diabetes Epidemic?**

Team Members:

Bryan Snyder (kbsnyder@uncc.edu)

Dolly Tripathy (dtripat1@uncc.edu)  
Eric Johnson (jejohns3@uncc.edu)  
Jyoti Thakral (jthakral@uncc.edu)  
Neha Adhau (nadhau@uncc.edu)

Course: DSBA/MBAD 6211- Advanced Business Analytics

Instructor: Dr. Mousavi

Spring 2017

Date Submitted:

May 1, 2017

# Executive Summary:

## Business Problem Statement

Diabetes has become an international epidemic, anticipated to affect as many as 640 million people by 2040. Due to the expected increased impact around the world, the healthcare community foresees a significant challenge, both socially and economically, with overall expenditures increasing as a result of hospital care, medication and opportunity costs. Merck, one of the largest pharmaceutical companies in the world, has made significant investments in diabetes research & development and seeks to identify the countries that are most at risk of being impacted in terms of prevalence (common factors present in the population), number of cases and increased costs specifically related to Type 2 Diabetes in adults.

## Business Goal

The goal is to identify the countries that are most at risk of being impacted by diabetes so that Merck can best determine how/where to deploy its resources to try to positively affect the disease. In addition, Merck seeks to address the following questions:

1. How can Merck position itself to best combat diabetes in the selected countries?
2. Forecast diabetes prevalence for the selected countries through 2030, which we plan to achieve by doing a time series forecast.
3. Find the factors influencing the diabetes prevalence in the selected countries to develop strategies to address the issues. We plan to achieve this by implementing classification and regression techniques.

## Data Profile

This project draws insights from the following datasets:

**1. International Diabetes Federation (IDF) Diabetes Atlas** (http://www.diabetesatlas.org/resources/2015-atlas.html)

The IDF Diabetes Atlas contains regional diabetes prevalence information for the countries around the world for the year 2015.

|  |
| --- |
| IDF 2015 |
| 10 Columns |
| 228 Records |

Further details can be found in the appendix.

**2. USDA Food Environment** (https://www.ers.usda.gov/data-products/food-environment-atlas/)

This data source lists food environment factors, such as store/restaurant proximity, food   
prices, food and nutrition assistance programs and community characteristics across the US. This provides a view of the food choices and diet quality present in the region. The original dataset was provided in Excel and we selected the most important variables relevant to this project.

This dataset required a cleaning effort which consisted of the following steps:

1. Columns with missing values were dropped and replaced with the mean value
2. Columns related to adolescent factors were dropped as this project focuses on adults

|  |
| --- |
| USDA Food Environment |
| 28 Columns |
| 3,094 Records |

The target attribute “High” is a derived from variable PCT\_DIABETES\_ADULT\_10 which is an indicator of diabetes prevalence in adults for a particular county.

Further details can be found in the appendix.

**3. US Chronic Disease Dataset**

(https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-  
Diabetes/f8ti-h92k)

This is a comprehensive dataset that provides access to a wide range of indicators for chronic diseases, such as diabetes. This includes conditions and risk factors at the US state level and for select large metro areas. The data contains information about 50 states and five (5) island territories with male and female values.

|  |
| --- |
| US Chronic Disease Dataset |
| 11 Columns |
| 110 Records |

The target attribute is “prevdiab” which is the prevalence of diagnosed diabetes among adults ages >= 18 years old.

Further details can be found in the appendix.

## Results

This analysis concludes that the United States will have the highest prevalence of diabetes and related costs by 2030. While there are smaller countries with a high prevalence, the population density and scale of diabetes cases in the US prompts costs to skyrocket and thereby makes it the country where Merck can best utilize its resources to combat the disease.

The following factors could be taken into consideration in resolution of the problem:

* Abnormal results in tests for High BP, Cholesterol and Glycosylated hemoglobin could be an indicator of the person having undiagnosed diabetes
* Complications in flu could be another indicator of diabetes, so the vaccination should be done regularly
* Some high prevalence zones (areas with high hispanic/black population, areas with low income and areas with high obesity rates) should be targeted
* More resources should be directed towards SNAP, NSLP and WIC nutrition programs and these programs should be advertised and marketed to a larger audience, especially in the affected zones

# Project Report

## Introduction

Diabetes is a chronic disease in which the body does not produce or use insulin to maintain blood sugar levels. It can lead to kidney failure, blindness, limb amputation, strokes, heart attacks and many other health problems. The World Health Organization (WHO) projects diabetes will be the seventh-leading cause of death by 2030. Although the disease is classified into several categories, the CDC estimates between 90% and 95% are Type 2 Diabetes cases, which are largely linked to obesity and a sedentary lifestyle.

Unfortunately, the prevalence of diabetes across the globe has skyrocketed over recent years. The WHO reported that from 1980 to 2014, the number of diabetics grew from 108 million to 422 million. This trend has presented an enormous health and economic challenge to the world as diabetes is one of the most common metabolic disorders and its prevalence among adults has been increasing in recent decades.

Urban trends have driven major changes in lifestyle, particularly in developing countries. With these rapid transitions come accompanying increases in risk factors for noncommunicable diseases like Type 2 Diabetes. Estimates of the current and future burden of diabetes are important in order to appropriately allocate resources, drive health-promoting policies, and encourage action to prevent diabetes in future generations. It has been observed that as a result of a combination of a number of factors (under-performing health systems, low awareness among the general public and health professionals, and the often slow-onset of symptoms or progression), the condition may remain undetected for many years, during which time complications may develop.

The goal is to identify the country or region that is most at risk through 2030 so that Merck can best determine how/where to deploy its resources to try to positively affect the rise of the disease.

## Background

There has been a substantial amount of work done in the past to address the issues and concerns surrounding Type 2 Diabetes. Diabetes Research and Clinical Practices (a healthcare publication) has published numerous scholarly articles that cover the problem. Some of these studies have reported age-specific prevalence for diabetes in order to systematically select studies to generate estimates for 219 countries and territories.

Estimates for countries without available source data were modeled from various countries that were similar with respect to geography, ethnicity, and economic development. Logistic regression was applied to generate smoothed age-specific prevalence estimates for adults 20–79 years which were then applied to population estimates for 2015 and 2030.

Data reporting both diagnosed and previously undiagnosed diabetes were collected and selected according to the previously described IDF methodology for diabetes in adults (ages 20–79). Reported countries were segregated into 15 data regions based on their geographic IDF Region and World Bank income classification. The median UDM (Undiagnosed Diabetes) proportion was calculated using the data for each region. The number of UDM cases in 2013 was calculated from country, age and gender-specific estimates of known diabetes cases and region-specific data.

“Of 744 reviewed data sources, 88 sources representing 74 countries had sufficient information and were selected to generate estimates of UDM. Globally, 45.8%, or 174.8 million of all diabetes cases in adults are estimated to be undiagnosed, ranging from 24.1% to 75.1% across data regions. An estimated 83.8% of all cases of UDM are in low- and middle income countries. At a country level, Pacific Island nations have the highest prevalence of UDM.”

## Data

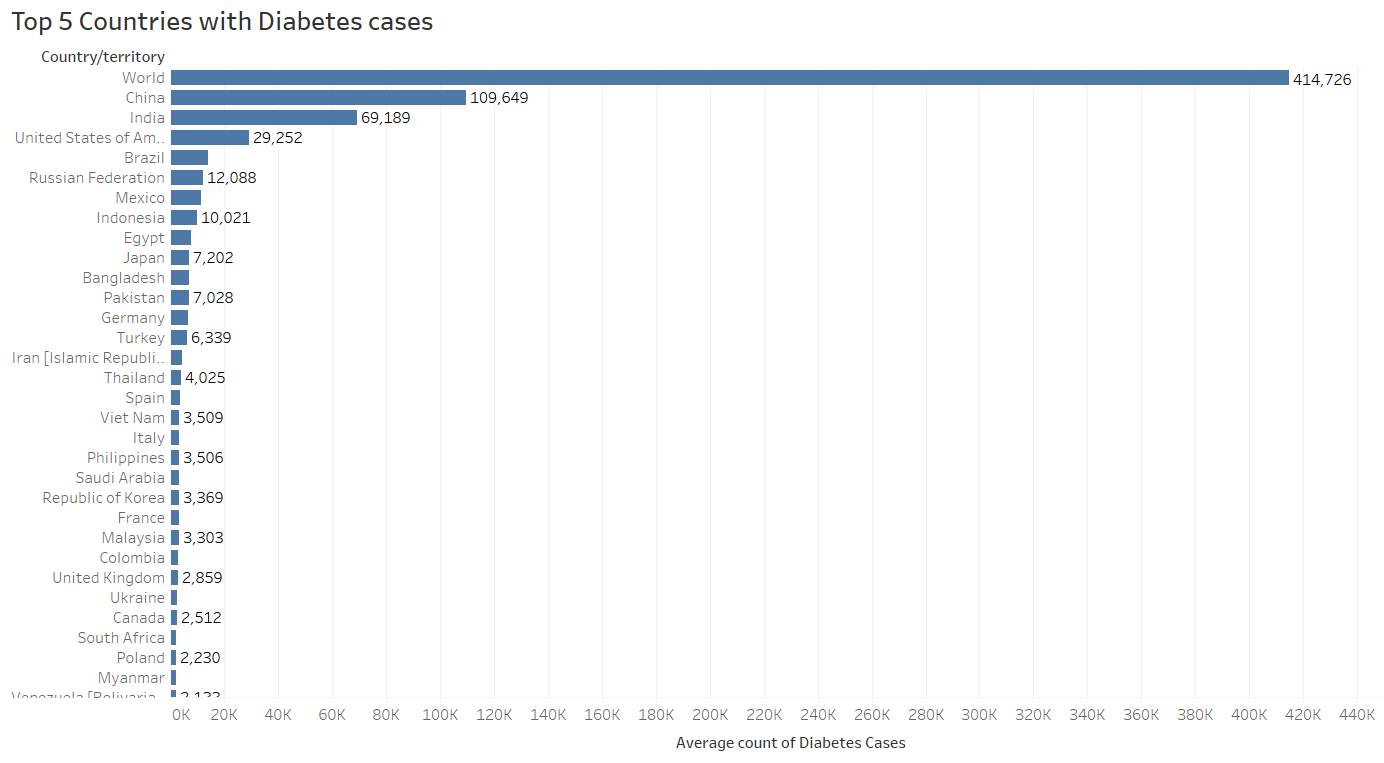
The initial dataset is sourced from the **International Diabetes Federation (IDF) Diabetes Atlas** which covers 2013 – 2015. Each year of data includes approximately 220 records with eight to twelve variables. A common view of the variables is illustrated below.  
  
*Table 2.1 – IDF Variables*

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| Country/territory | Country Name |
| Diabetes national prevalence [%] [uncertainty range] | Number of people in a given group or population who are reported to have a disease |
| Diabetes age-adjusted comparative prevalence [%] [uncertainty range] | Diabetes in comparison to an age |
| Adults with diabetes [20-79] in 1,000s [uncertainty range] | Number of Adults age(20-79) diagnosed with Diabetes |
| Adults with undiagnosed diabetes [20-79] in 1,000s [uncertainty range] | Number of Adults age(20-79) undiagnosed Diabetes |
| Mean diabetes-related expenditure per person with diabetes [R=2, USD] | Mean medical spending for people with diabetes in US dollars |
| Mean diabetes-related expenditure per person with diabetes [R=2, International Dollar] | Mean medical spending for people with diabetes in international dollars |
| Diabetes related deaths [20-79] | Number of Deaths in age (20-70) due to Diabetes |
| Number of children with type 1 diabetes [0-14] in 1,000s | Count of children with Type 1 Diabetes |
| Basis of estimate [OGTT, extrapolation, other\*] | OGTT=oral glucose tolerance test (OGTT) measures the body's ability to use a type of sugar, called glucose  Other = include 'Fasting blood glucose', 'Self-reported', 'Medical record or clinical diagnosis', and 'HbA1c' |

*Table 2.2 – IDF Summary Statistics  
Table below gives the statistical overview of the IDF variables with its range and data*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Country/ Territory | Prev | Diabetes national prevalence [%] [uncertainty range] | Diabetes age-adjusted comparative prevalence [%] [uncertainty range] | Adults with diabetes [20-79] in 1,000s [uncertainty range] | Adults with undiagnosed diabetes ]20-79] in 1,000s [uncertainty range] | Mean diabetes-related expenditure per person with diabetes [R=2, USD] | Mean diabetes-related expenditure per person with diabetes [R=2, International Dollar] | Diabetes related deaths [20-29] | OGTT |
| Length: 230 | Min.: 0.640 | 1.8 [1.3 – 4.2]: 2 | 13.1 [10.9 – 15.7]: 3 | 0.1 [0.1 – 0.2]: 1 | 0.1 [0.0 – 0.1]: 2 | Min.: 0.0 | Min.: 0.0 | Min.: 1 | Min.: 0.0000 |
| NA | Mean: 8.997 | 13.4 [11.2 – 16.0]:2 | 10.3 [9.0 – 11.8]: 2 | 1,020.0 [913.1 – 1,135.8]: 1 | 0.5 [0.3 – 0.9]: 1 | Mean: 1401.6 | Mean: 1612.8 | Mean: 73308 | Mean: 0.3091 |
| NA | Max.: 29.690 | (Other): 219 | (Other): 216 | (Other): 223 | (Other): 222 | Max.: 11851.1 | Max.: 10941.7 | Max.: 4960536 | Max.: 3.0000 |

*Graph 2.1: Adults with Diagnosed Diabetes aged(20-79) in 1000s   
Illustration of important variables within IDF dataset and insight to make initial hypothesis*



China 109,649

India 69.189

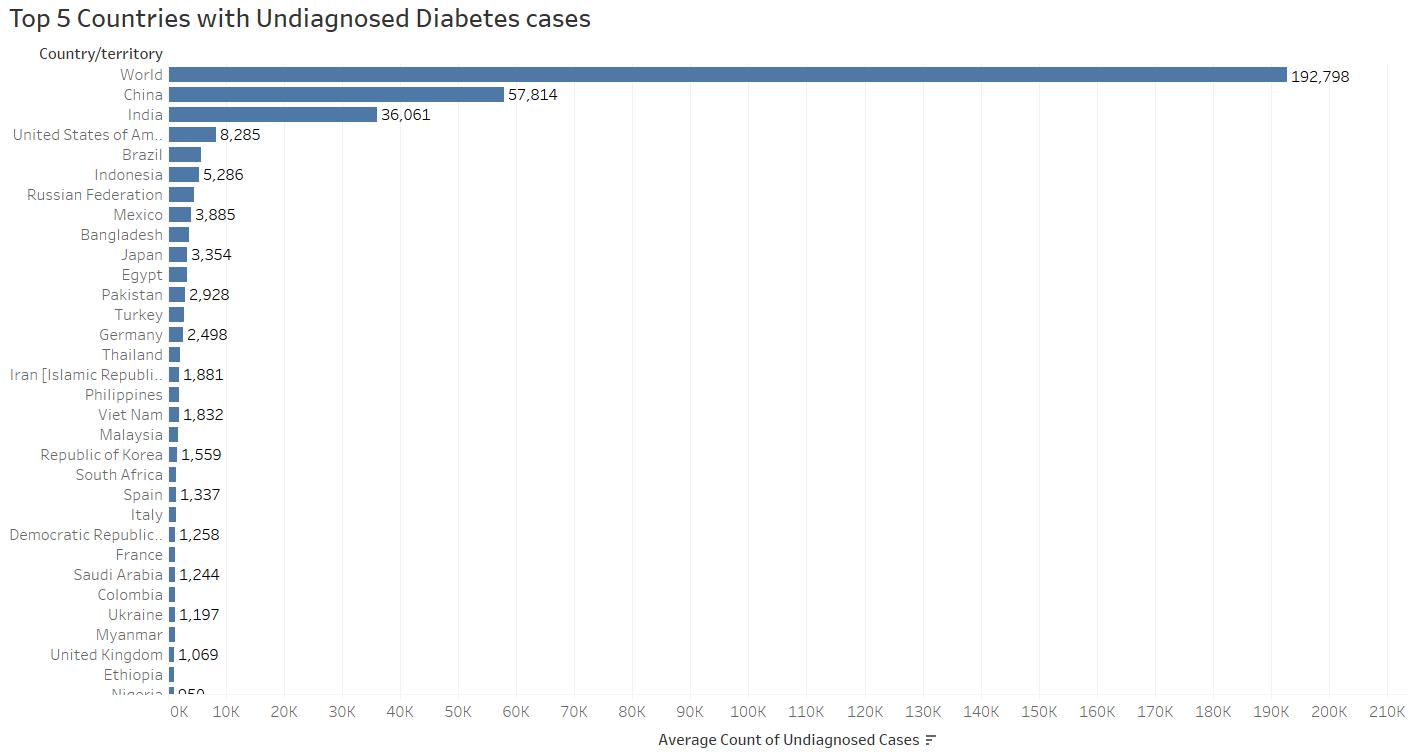
United States 29,252

Brazil 14251

Russian Federation 12088

* Top 5 HIGH risk countries contain 56.03% of the world’s diabetic population.

*Graph 2.2: Adults with Undiagnosed Diabetes aged(20-79) in 1000s*



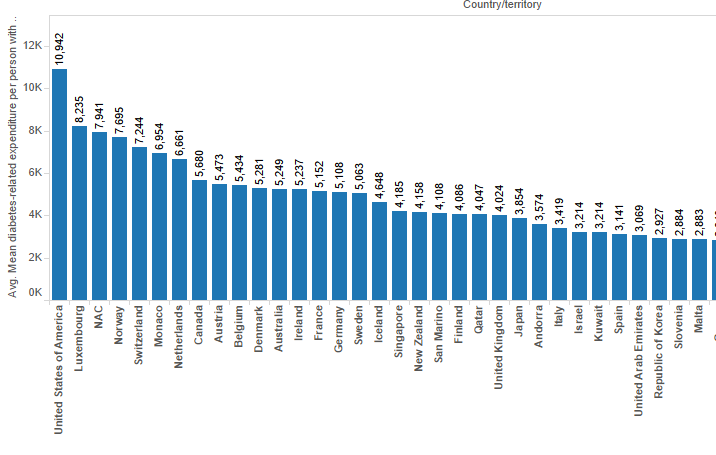
China 57,814

India 36,061

United States 8,285

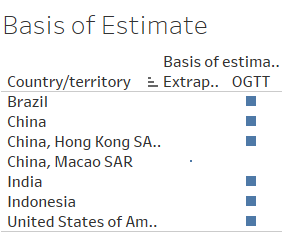
Brazil 5,724

*Graph 2.3: Mean diabetes-related expenditure per person with diabetes by country*



* Total global health expenditure per person for diabetes in America is highest among all the countries.

*Graph 2.4: Diabetes test type*



* These high risk countries use the basis of estimate as OGTT which could be improved with other sensitive and accurate methods.

*Graph 2.5: Diabetes Prevalence distribution in Urban and Rural Area*



* Most of the countries with diabetes prevalence are in medium and low income areas
* Among the top five high risk countries, three of them are MIC and prevalence is greater in urban areas
* America and China are HIC countries where China has greater prevalence in urban areas only and US has same prevalence in rural and urban areas

A dataset focused on the **USDA Food Environment** was also used and includes 3,094 records and 28 variables. Variable code and descriptions of each variables are listed below in Table 2.3

*Table 2.3 – USDA Food Environment Variables*

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| FIPS Code | Code for each county in United States |
| State | Name of the state |
| County Name | Name of the county |
| 2010 Census Population | County-wise population estimates for the year 2010 |
| PCH\_RECFAC\_07\_12 | Recreation & fitness facilities (% change), 2007-12 |
| CH\_FOODINSEC\_09\_12 | Household food insecurity (change %), 2007-09 to 2010-12\* |
| PCH\_FMRKT\_09\_13 | Farmers' markets (% change), 2009-13 |
| PCH\_GROC\_07\_12 | Grocery stores (% change), 2007-12 |
| PCH\_GROCPTH\_07\_12 | Grocery stores/1,000 pop (% change), 2007-12 |
| PCH\_SUPERCPTH\_07\_12 | Supercenters & club stores/1,000 pop (% change), 2007-12 |
| PCH\_SPECSPTH\_07\_12 | Specialized food stores (% change), 2007-12 |
| PCH\_SNAPS\_08\_12 | SNAP-authorized stores (% change), 2008-12 |
| PCT\_FREE\_LUNCH10 | Students eligible for free lunch (%), 2010 |
| PCT\_DIABETES\_ADULTS10 | Adult diabetes rate, 2010 |
| PCT\_NHWHITE10 | % White, 2010 |
| PCT\_NHBLACK10 | % Black, 2010 |
| PCT\_HISP10 | % Hispanic, 2010 |
| PCT\_NHASIAN10 | % Asian, 2010 |
| POVRATE10 | Poverty rate, 2010 |
| PCT\_OBESE\_ADULTS10 | Adult obesity rate (county), 2010 |
| MEDHHINC10 | Median household income, 2010 |
| PCH\_FMRKTPTH\_09\_13 | Farmers' markets/1,000 pop (% change), 2009-13 |
| ORCHARD\_ACRESPTH07 | Orchard acres/1,000 pop, 2007 |
| BERRY\_ACRESPTH07 | Berry acres/1,000 pop, 2007 |
| PCH\_SNAPSPTH\_08\_12 | SNAP-authorized stores/1,000 pop (% change), 2008-12 |
| PCH\_WICS\_08\_12 | WIC-authorized stores (% change), 2008-12 |
| PCH\_WICSPTH\_08\_12 | WIC-authorized stores/1,000 pop (% change), 2008-12 |
| PCH\_NSLP\_09\_14 | National School Lunch Program participants (change % pop), 2009-14\* |
| PCT\_REDUCED\_LUNCH10 | Students eligible for reduced-price lunch (%), 2010 |
| PCH\_CACFP\_09\_14 | Child & Adult Care (change % pop), 2009-14\* |
| PCT\_NHNA10 | % American Indian or Alaska Native, 2010 |
| PCT\_NHPI10 | % Hawaiian or Pacific Islander, 2010 |

The Food Environment dataset was uploaded and data wrangling methods were applied to remove invalid and missing values. The summary of each variable after the cleaning are mentioned below in Table 2.4

*Table 2.4 – USDA Food Environment Summary Statistics*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SUMMARY STATISTICS** | | | | |
| Variable | Observations | Min | Mean | Max |
| FIPS Code | 3094 | 1001.00 | 30317.00 | 56039.00 |
| 2010 Census Population | 3094 | 82.00 | 99143.00 | 9818605.00 |
| PCH\_RECFAC\_07\_12 | 3094 | -100.00 | -10.00 | 300.00 |
| PCH\_FMRKT\_09\_13 | 3094 | -100.00 | 31.00 | 1400.00 |
| PCH\_GROC\_07\_12 | 3094 | -100.00 | -4.09 | 300.00 |
| PCH\_GROCPTH\_07\_12 | 3094 | -100.00 | -6.96 | 335.53 |
| PCH\_SUPERCPTH\_07\_12 | 3094 | -100.00 | 13.12 | 479.70 |
| PCH\_SPECSPTH\_07\_12 | 3094 | -100.00 | -14.56 | 681.49 |
| PCH\_SNAPS\_08\_12 | 3094 | -100.00 | 40.12 | 1100.00 |
| PCT\_FREE\_LUNCH10 | 3094 | 0.00 | 43.04 | 99.40 |
| PCT\_NHWHITE10 | 3094 | 2.67 | 78.23 | 99.16 |
| PCT\_NHBLACK10 | 3094 | 0.00 | 8.80 | 85.44 |
| PCT\_HISP10 | 3094 | 0.00 | 8.36 | 95.75 |
| PCT\_NHASIAN10 | 3094 | 0.00 | 1.14 | 43.02 |
| POVRATE10 | 3094 | 2.00 | 105.10 | 331.00 |
| PCT\_OBESE\_ADULTS10 | 3094 | 13.10 | 30.58 | 47.90 |
| MEDHHINC10 | 3094 | 2.00 | 1486.10 | 2979.00 |
| PCH\_FMRKTPTH\_09\_13 | 3094 | -100.00 | 27.57 | 1485.88 |
| ORCHARD\_ACRESPTH07 | 3094 | 0.00 | 21.97 | 2913.07 |
| BERRY\_ACRESPTH07 | 3094 | 0.00 | 1.25 | 881.75 |
| PCH\_SNAPSPTH\_08\_12 | 3094 | -100.00 | 37.05 | 1131.42 |
| PCH\_WICS\_08\_12 | 3094 | -100.00 | -1.87 | 300.00 |
| PCH\_WICSPTH\_08\_12 | 3094 | -100.00 | -3.98 | 294.07 |
| PCH\_NSLP\_09\_14 | 3094 | -1.46 | -0.82 | 0.52 |
| PCT\_REDUCED\_LUNCH10 | 3094 | 0.00 | 8.51 | 57.42 |
| PCH\_CACFP\_09\_14 | 3094 | -0.38 | 0.03 | 0.97 |
| PCT\_NHNA10 | 3094 | 0.00 | 1.82 | 94.94 |
| PCT\_NHPI10 | 3094 | 0.00 | 0.06 | 11.33 |

The dataset focuses on **US Chronic Disease** and contains 110 records. Variable descriptions are listed below.

*Table 2.5 – US Chronic Disease Variables*

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| ***Independent Variables/Predictors*** | |
| **DiagAdult** | Adult with diagnosed diabetes(Age AdjPrev) |
| **eyeexam** | Dilated eye examination among adults aged >= 18 years with diagnosed diabetes |
| **footexam** | Foot examination among adults aged >= 18 years with diagnosed diabetes |
| **glymeasure** | Glycosylated hemoglobin measurement among adults aged >= 18 years with diagnosed diabetes |
| **prevdepre** | Prevalence of depressive disorders among adults aged >= 18 years with diagnosed diabetes |
| **influall** | Influenza vaccination among noninstitutionalized adults aged >= 65 years with diagnosed diabetes |
| **influadult** | Influenza vaccination among noninstitutionalized adults aged 18-64 years with diagnosed diabetes |
| **prevdepre** | Prevalence of diagnosed diabetes among adults aged >= 18 years |
| **HighBP** | Prevalence of high blood pressure among adults aged >= 18 years with diagnosed diabetes |
| **Highchole** | Prevalence of high cholesterol among adults aged >= 18 years with diagnosed diabetes |
| ***Dependent Variable/Target*** | |
| **prevdiab** | Prevalence of diagnosed diabetes among adults aged >= 18 years |

*Table 2.6 – US Chronic Disease Summary Statistics*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SUMMARY STATISTICS** | | | | |
| **Variable** | **Observation** | **Min.** | **Mean** | **Max.** |
| **DiagAdult** | 110 | 23.2 | 56.05 | 77.7 |
| **eyeexam** | 110 | 37.5 | 60.99 | 85.8 |
| **footexam** | 110 | 45.7 | 68.37 | 91.4 |
| **glymeasure** | 110 | 43.9 | 67.8 | 86.3 |
| **influall** | 110 | 26.4 | 66.18 | 84.9 |
| **influadult** | 110 | 13.1 | 44.73 | 74.2 |
| **prevdepre** | 110 | 9 | 31.79 | 56.6 |
| **prevdiab** | 110 | 5.6 | 9.282 | 14.9 |
| **HighBP** | 110 | 39.5 | 60.52 | 85.1 |
| **Highchole** | 110 | 36.8 | 58.74 | 74.2 |

The final dataset, **Diabetes Public Health Resource**, focuses on the number of U.S. adults aged 18 years or older with diagnosed diabetes (From 1980 through 2014). The data contains 35 records.

*Table 2.7 Diabetes Public Health Resource*

|  |  |
| --- | --- |
| **Dimensions/Attributes in dataset** | **Alias** |
| ***Independent Variables/Predictors*** | |
| Year | Year |
| Number (in Millions) of Civilian, Non-institutionalized Adults with Diagnosed Diabetes, United States, 1980-2014 | Number in Millions. |

*Table 2.8 Diabetes Public Health Resource Summary Statistics*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SUMMARY STATISTICS** | | | | |
| Variables | Observations | Min. | Mean | Max. |
| Year | 35 | 1980 | 1997 | 2014 |
| Number in Millions. | 35 | 5.5 | 11.62 | 22.2 |

## Method

2.4.1 – Analysis Methods

The following algorithms were used in this analysis:  
- Multiple linear regression  
- Random Forest (RF)  
- Gradient Boosting Machines (GBM)

**Dataset US Chronic Disease Indicator**

2.4.2 – Regression Analysis

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to the observed data. Every value of the independent variable *x* is associated with a value of the dependent variable *y*. As the US Chronic Disease Indicator dataset observations are continuous in nature we decided to use regression analysis for model evaluation. We fitted the dependent and independent variables in the regression equation and generated various models.

Regression OLS:

Variable Selection - Stepwise Regression

First we considered all 10 Independent variables in our regression model to find out highly influencing variable for diabetes prevalence.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STEP: AIC=81.34**     |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | DF | sum of SQ | RSS | AIC | F-Value | Pr(>F) | | influall | 1 | 0.1116 | 166.53 | 79.396 | 0.0443 | 0.833995 | | prevdepre | 1 | 0.1599 | 166.58 | 79.419 | 0.0634 | 0.801947 | | gender | 1 | 0.2278 | 166.65 | 79.45 | 0.0903 | 0.764704 | | eyeexam | 1 | 1.2087 | 167.63 | 79.902 | 0.4793 | 0.491149 | | footexam | 1 | 4.3230 | 170.74 | 81.319 | 1.7145 | 0.194948 | | <none> |  | 166.42 | 81.345 |  |  |  | | glymeasure | 1 | 4.4748 | 170.9 | 81.388 | 1.7747 | 0.187388 | | influadult | 1 | 8.0188 | 174.44 | 82.968 | 3.1802 | 0.079135 | | DiagAdult | 1 | 10.5189 | 176.94 | 84.064 | 4.1717 | 0.045104\* | | Highchole | 1 | 12.6261 | 179.05 | 84.976 | 5.0074 | 0.02862\* | | HighBP | 1 | 22.0995 | 188.52 | 88.945 | 8.7644 | 0.004264\*\* |   ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  **Step:  AIC=79.4**   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | DF | Sum of SQ | RSS | AIC | F-value | Pr(>F) | | prevdepre | 1 | 0.1282 | 166.66 | 77.455 | 0.0516 | 0.821030 | | gender | 1 | 0.3091 | 166.84 | 77.539 | 0.1244 | 0.725462 | | eyeexam | 1 | 1.3093 | 167.84 | 77.999 | 0.5267 | 0.470506 | | <none> |  |  | 166.53 | 79.396 |  |  | | footexam | 1 | 4.3886 | 170.92 | 79.399 | 1.7657 | 0.188428 | | glymeasure | 1 | 4.6596 | 171.19 | 79.521 | 1.8747 | 0.175512 | | influadult | 1 | 8.0673 | 174.6 | 81.039 | 3.2457 | 0.076111 . | | DiagAdult | 1 | 10.5796 | 177.11 | 82.139 | 4.2565 | 0.042980 \* | | Highchole | 1 | 15.4138 | 181.94 | 84.212 | 6.2014 | 0.015251 \* | | HighBP | 1 | 22.2262 | 188.76 | 87.043 | 8.9422 | 0.003895 \*\* |   ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  **Step:  AIC=77.46**   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | DF | sum of SQ | RSS | AIC | F-value | Pr(>F) | | gender | 1 | 0.9312 | 167.59 | 75.885 | 0.3800 | 0.539685 | | eyeexam | 1 | 1.4428 | 168.1 | 76.119 | 0.5887 | 0.445588 | | <none> |  | 166.66 | 77.455 |  |  |  | | footexam | 1 | 4.5618 | 171.22 | 77.535 | 1.8613 | 0.176973 | | glymeasure | 1 | 4.8237 | 171.48 | 77.653 | 1.9682 | 0.165192 | | influadult | 1 | 8.8188 | 175.48 | 79.426 | 3.5982 | 0.06209 | | DiagAdult | 1 | 10.6173 | 177.28 | 80.211 | 4.3321 | 0.041169\* | | Highchole | 1 | 15.3133 | 181.97 | 82.224 | 6.2481 | 0.014851\* | | HighBP | 1 | 22.1007 | 188.76 | 85.044 | 9.0175 | 0.003739\*\* |     ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  **Step:  AIC=75.88**     |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | DF | Sum of SQ | RSS | AIC | F-Value | Pr(>F) | | eyeexam | 1 | 1.3453 | 168.94 | 74.5 | 0.5539 | 0.459268 | | footexam | 1 | 4.2896 | 171.88 | 75.831 | 1.7661 | 0.188245 | | <none> |  | 167.59 | 75.885 |  |  |  | | glymeasure | 1 | 5.3874 | 172.98 | 76.321 | 2.2181 | 0.140958 | | influadult | 1 | 10.0430 | 177.63 | 78.366 | 4.1349 | 0.045856\* | | DiagAdult | 1 | 12.8589 | 180.45 | 79.577 | 5.2942 | 0.024423\* | | Highchole | 1 | 16.6262 | 184.22 | 81.168 | 6.8453 | 0.01091\* | | HighBP | 1 | 28.1696 | 195.76 | 85.848 | 11.5979 | 0.001104\*\* |   ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  **Step:  AIC=74.5**   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | DF | Sum of SQ | RSS | AIC | F-Value | Pr(>F) | | footexam | 1 | 3.2173 | 172.15 | 73.953 | 1.3331 | 0.25218 | | <none> |  | 168.94 | 74.5 |  |  |  | | glymeasure | 1 | 6.6190 | 175.56 | 75.459 | 2.7427 | 0.102179 | | influadult | 1 | 9.3053 | 178.24 | 76.629 | 3.8557 | 0.053549 | | DiagAdult | 1 | 11.7815 | 180.72 | 77.691 | 4.8817 | 0.030416\* | | Highchole | 1 | 16.2944 | 185.23 | 79.59 | 6.7517 | 0.011411\* | | HighBP | 1 | 27.3138 | 196.25 | 84.04 | 11.3177 | 0.001249\*\* |   ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  **Step:  AIC=73.95**   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | DF | Sum of SQ | RSS | AIC | F-Value | Pr(>F) | | <none> |  | 172.15 | 73.953 |  |  |  | | glymeasure | 1 | 4.905 | 177.06 | 74.116 | 2.0229 | 0.159318 | | influadult | 1 | 11.459 | 183.61 | 76.915 | 4.7259 | 0.033048\* | | Highchole | 1 | 15.790 | 187.94 | 78.71 | 6.5123 | 0.012867\* | | HighBP | 1 | 25.138 | 197.29 | 82.447 | 10.3673 | 0.001935\*\* | | DiagAdult | 1 | 26.755 | 198.91 | 83.076 | 11.0344 | 0.001415\*\* |     ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

**Based on the min AIC values, we removed insignificant variables and generated the final fit.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Residuals:**   |  |  |  | | --- | --- | --- | | Min | Median | Max | | -2.9954 | -0.2423 | 4.0905 |   **Coefficients:**   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Estimate | Std.Error | t-value | Pr(>|t|) | | (Intercept) | 3.44098 | 3.29501 | 1.044 | 0.29994 | | DiagAdult | -0.07312 | 0.02228 | -3.283 | 0.00161\*\* | | glymeasure | 0.03399 | 0.02497 | 1.361 | 0.17778 | | influadult | -0.03880 | 0.01792 | -2.166 | 0.03375\* | | HighBP | 0.08545 | 0.02658 | 3.215 | 0.00198\*\* | | Highchole | 0.06317 | 0.02824 | 2.237 | 0.02847\* |   ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1   |  |  | | --- | --- | | Residual standard error | 1.567 | | Multiple R-squared | 0.3849 | | Adjusted R-squared | 0.3322 | | F-statistic | 7.301 | | p-value | 4.42E-06 | |

Conclusion of Regression Analysis:  
Final Model as stated gives us best fit with most influencing variables. Below are the most important variables as per regression analysis:

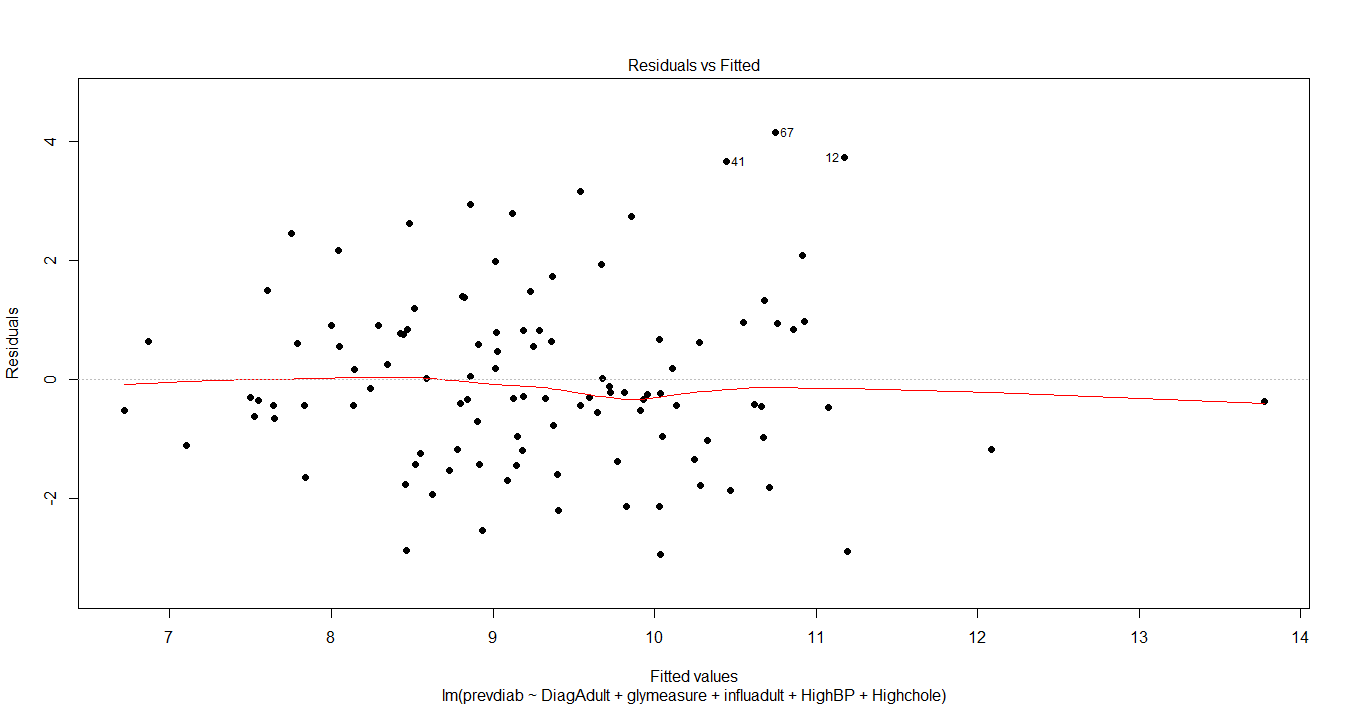
The most significant variables for target prediction are:

|  |
| --- |
| Adult with diagnosed diabetes(Age AdjPrev) |
| Glycosylated hemoglobin measurement among adults aged >= 18 years with diagnosed diabetes |
| Influenza vaccination among noninstitutionalized adults aged 18-64 years with diagnosed diabetes |
| Prevalence of high blood pressure among adults aged >= 18 years with diagnosed diabetes |
| Prevalence of high cholesterol among adults aged >= 18 years with diagnosed diabetes |

Accuracy Matrix

|  |  |  |
| --- | --- | --- |
| correlation\_accuracy | | |
|  | actuals | predicteds |
| actuals | 1 | 0.540617 |
| predicteds | 0.540617 | 1 |

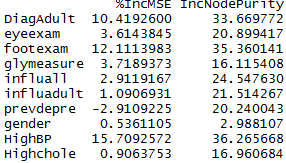
|  |
| --- |
| min\_max\_accuracy- 0.8962804  89.62% model is accurate on Test data. |

Residual plot for the fitted variable:  
The plot shows no systematic patterns and the variation in the residuals does not seem to change with the size of the fitted value.  
  
 *Graph 2.6– Residuals vs. Fitted for Regression Analysis*  


2.4.3 – Random Forest  
Random Forest is an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Based on the Random Forest prediction method we employed the ensemble learning technique for regression on our prediction variable and generated a new model.  
  
Optimal ntree:  
First we set the mtry to the default value (sqrt of total number of all predictors) and searched for the optimal ntree value. To find the number of trees that correspond to a stable classifier, we built random forest with different ntree values (10, 20, 30, …, 100). We built RF classifiers for each ntree value, recorded the OOB error rate and observed the number of trees where the out of bag error rate stabilized and reached its minimum. The best ntree value was 1000.

Optimal mtry:  
The optimal number of predictors selected for split is selected by observing which out of bag error rate stabilizes and reaches the minimum. Mtry =2 returns the best results in this case.

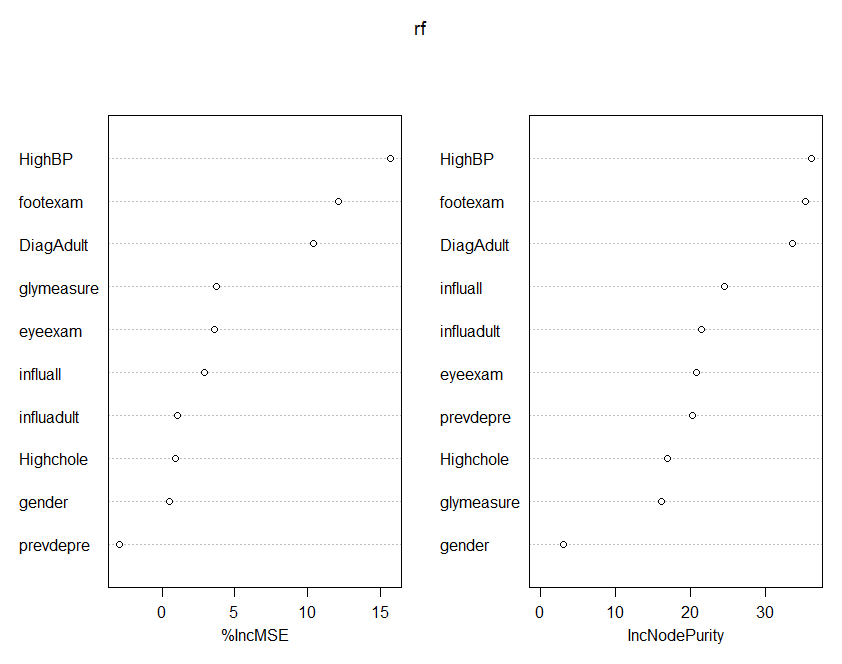
Variable importance:  
Given below are the ranked variables as per their importance**.** We see that removing “HighBP” Increases MSE of the model by 15.7%



Variable Plot:  
The higher the value of IncMSE, the higher the importance of the variable in the model. IncNodePurity relates to the loss function by which the best splits are chosen. The loss function is mse for regression and gini-impurity for classification.

**Top 5 important variables** obtained from Random Forest are as shown below.

1. Prevalence of high blood pressure among adults aged >= 18 years with diagnosed diabetes
2. Foot examination among adults aged >= 18 years with diagnosed diabetes
3. Adult with diagnosed diabetes(Age Adj Prev)
4. Glycosylated hemoglobin measurement among adults aged >= 18 years with diagnosed diabetes
5. Dilated eye examination among adults aged >= 18 years with diagnosed diabetes

*Graph 2.7 – Variable Importance in Random Forest*

Model Selection

The success of the MLRs can be assessed by evaluating the magnitude of the adjusted R2, the residual standard error (RSE) for the regression, the t test results for the individual predictor variables for the regression.   
  
The adjusted R Square for stepwise regression is about 33.22% while the variance explained by random forest is only 14.58%. The R(2) values for the MLRs were higher than the proportion of variance explained values for the RFRs. In general, MLRs seem to be superior to the RFRs in terms of predictive value and error. In the case of this dataset, MLR appears to be superior to RFR in terms of its explanatory value and error. This result suggests that MLR has advantages over RFR for prediction with this kind of dataset. So in this case, **Multiple Regression Model** is considered to be **best** model.

We see that Mean Squared Error obtained from Random forest is

**Mean of squared residuals: 2.777**

**% Var explained: 14.58**

And from Stepwise Multiple Linear Regression is :

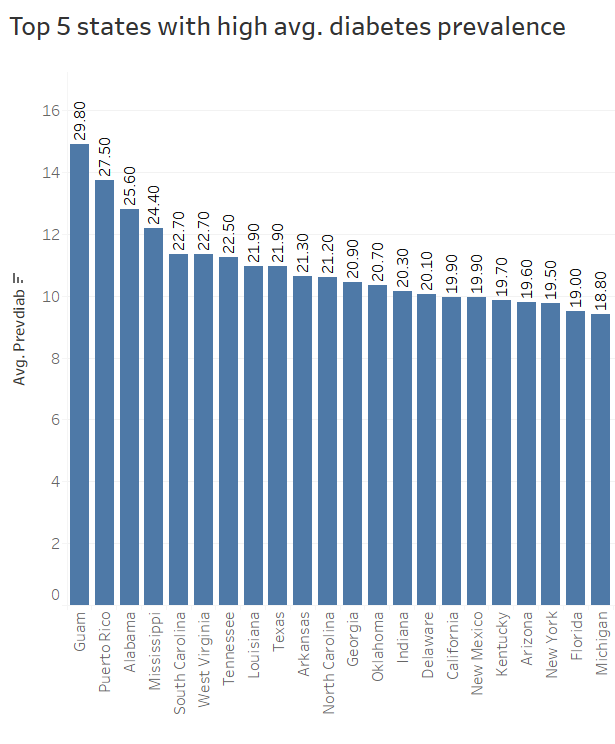
**Residual standard error: 1.567**

**Multiple R-squared: 0.3849,**

**Adjusted R-squared: 0.3322**

Visualizations:  
We analyzed the data and decided to study the top five (5) states having the highest prevalence of diagnosed diabetes.

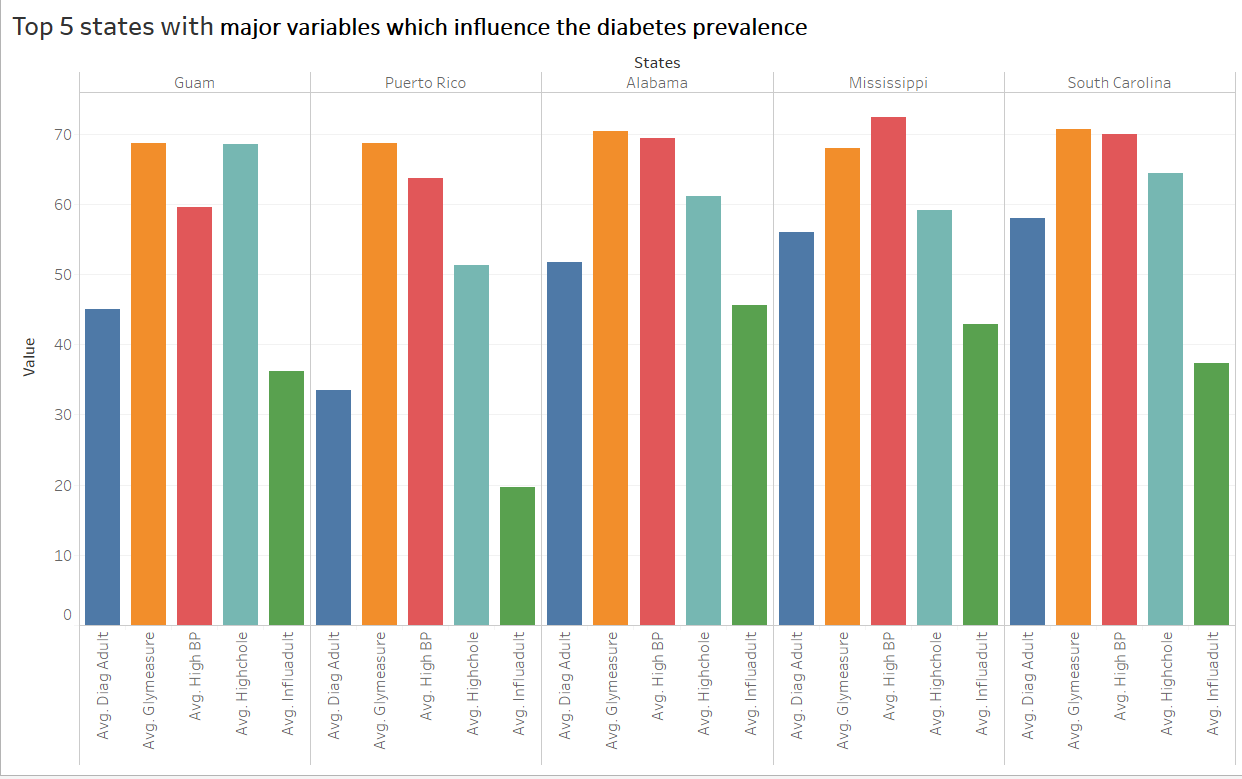
1. Guam – 29.8
2. Puerto Rico – 27.5
3. Alabama- 25.6
4. Mississippi – 24.4
5. South Carolina – 22.7

*Graph 2.8 – Top Five States with Highest Diabetes Prevalence*  


Based on our model from regression, the major variables which influence prevalence is plotted along with the top 5 states.

Among the top five diabetic states, glycosylated hemoglobin measurement is highest in Alabama, Puerto Rico and Guam. While the prevalence of high blood pressure among adults aged >= 18 years with diagnosed diabetes was highest in Alabama, Mississippi and South Carolina.

*Graph 2.9 – Top Five States with major variables which influence diabetes prevalence*



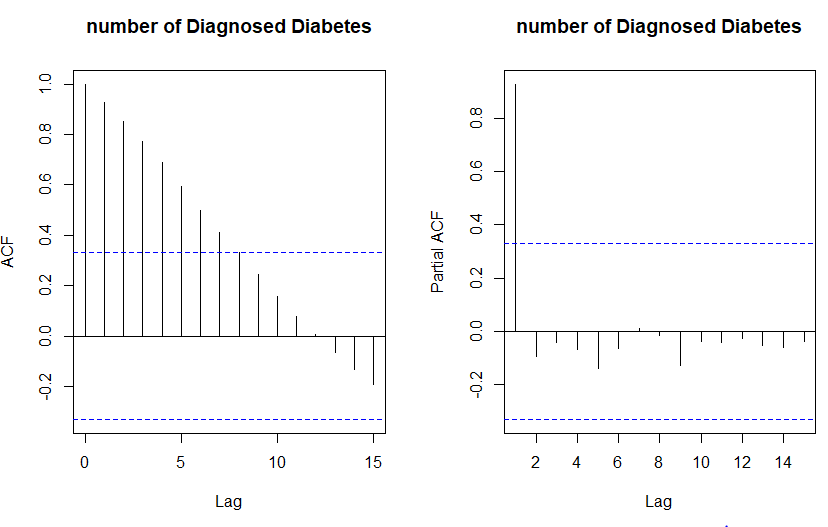
**Dataset Dataset Diabetes Public Health Resource**

### 2.4.4 Time series Forecast

ARIMA model to forecast future Diabetes Prevalence in US

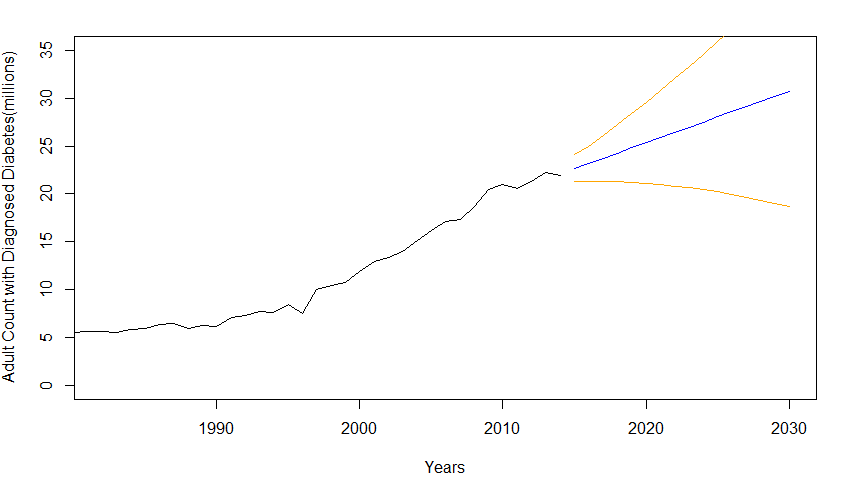
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ARIMA(2,2,2)                    : Inf  ARIMA(0,2,0)                    : 96.11724  ARIMA(1,2,0)                    : 84.50387  ARIMA(0,2,1)                    : 74.65333  ARIMA(1,2,1)                    : 74.4423  ARIMA(1,2,2)                    : Inf  ARIMA(2,2,1)                    : 75.83526  Best model: ARIMA(1,2,1)  Series: x  ARIMA(1,2,1)  Coefficients:   |  |  |  | | --- | --- | --- | |  | Ar1 | Ma1 | |  | -0.3164 | -0.7777 | | s.e. | 0.1823 | 0.1150 |  |  |  | | --- | --- | | sigma^2 estimated | 0.4625 | | log likelihood | -33.81 | | AIC | 73.61 | | AICc | 74.44 | | BIC | 78.1 |   Training set error measures:   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | |  | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 | | Training set | 0.07843026 | 0.6400395 | 0.4615632 | 0.875587 | 4.218382 | 0.7333247 | -0.1018803 | |

*Graph 2.12 ACF and Partial ACF plots*

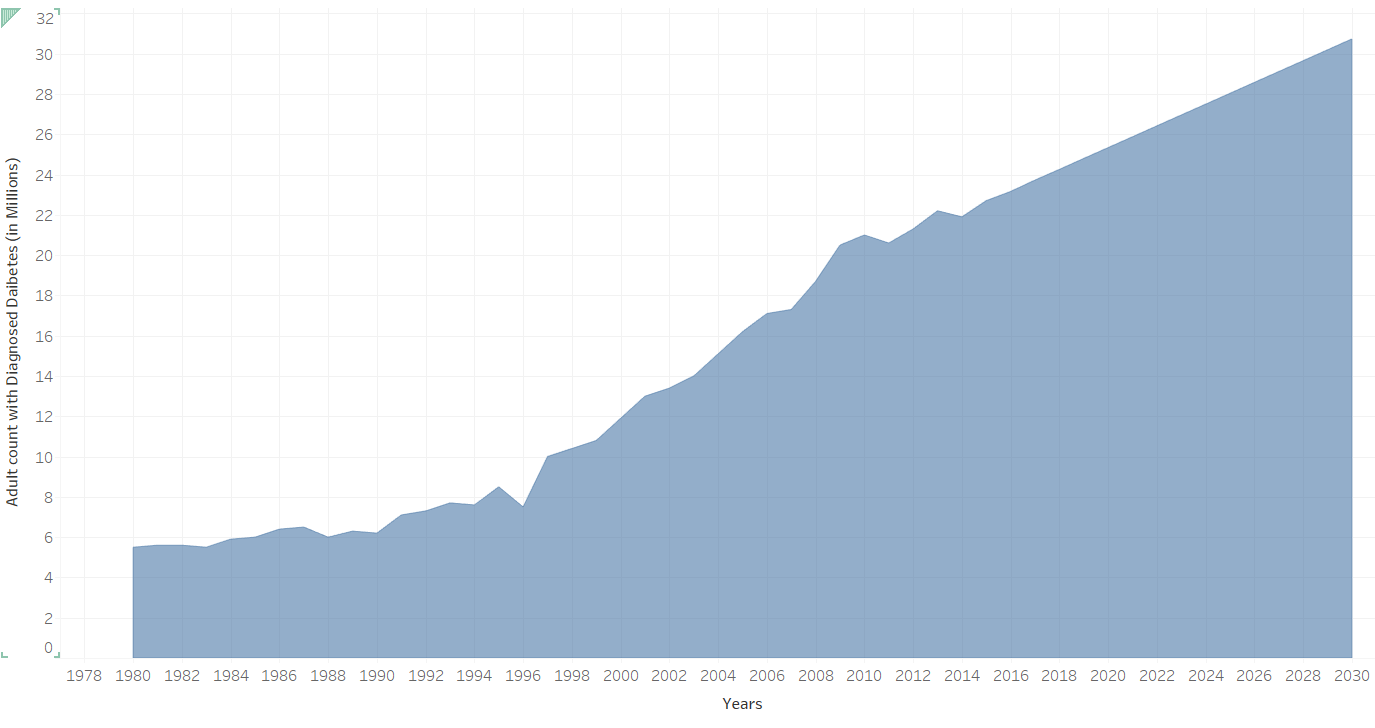


|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| $pred Time Series:  Start = 2015  End = 2030  Frequency = 1   |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  | | --- | --- | | 2015 | 22.70493 | | 2016 | 23.16022 | | 2017 | 23.72615 | | 2018 | 24.25706 | | 2019 | 24.79906 | | 2020 | 25.33755 | | 2021 | 25.87714 | | 2022 | 26.41639 | | 2023 | 26.95575 | | 2024 | 27.49507 | | 2025 | 28.03441 | | 2026 | 28.57374 | | 2027 | 29.11307 | | 2028 | 29.6524 | | 2029 | 30.19173 | | 2030 | 30.73107 | |  |   $se  Time Series:  Start = 2015  End = 2030  Frequency = 1  [1] 0.6800799 0.9176133 1.2091848 1.4980910 1.8039072 2.1222257 2.4545738 2.8003519 3.1594207 3.5314301 3.9160686  [12] 4.3130119 4.7219501 5.1425856 5.5746359 6.0178337 |

*Graph 2.13 -- Time Forecast for Number of Diagnosed Diabetes (from 2015 to 2030)*



*Graph 2.14 -- Tableau Graph for Diagnosed Diabetes Count from 1980-2030*



### **Dataset USDA Food Environment**

To address the problem of the factors which would help to predict if a region is susceptible to high or low diabetes prevalence, we decided to analyze the dataset using classification algorithms. We considered the following:

2.4.5 Random Forest

For all of the models a binary output column named High was derived from the column named **PCT\_DIABETES \_ADULT \_10.** The Median was taken for this column as a reference point. Any observation above the **median value was considered a “Yes” and the rest were considered a “No”.**

The model was trained with 60% of the data with number of tress value to be 80 for which best “mtry” was found from the following observations:-

mtry OOBError

2.OOB 2 0.2216066

3.OOB 3 0.2363804

4.OOB 4 0.2645429

5.OOB 5 0.2673130

7.OOB 7 0.3361034

|  |
| --- |
| The best mtry was found to be 2 with the lowest out of bag error. The positive class for the model was for value “Yes”. Model OOB error was 21.1% |

**CONFUSION MATRIX :-**  Below is the confusion matrix for the model Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
| Predicted |  | Yes | No |
| Yes | 456 | 195 |
| No | 23 | 254 |

.

2.4.6 Gradient Boosting Method (GBM):  
The GBM model was trained using 60% of the data and the ten most important variables was found to be as below in table 2.7

*Table 2.7- Variable Importance*

|  |  |  |
| --- | --- | --- |
| SL NO | VARIABLE | REL INFO |
| 1. | PCT\_OBESE\_ADULTS10 | 21.60240606 |
| 2. | PCT\_FREE\_LUNCH10 | 21.43491385 |
| 3. | PCT\_65OLDER10 | 8.24472657 |
| 4. | PCT\_HISP10 | 8.07165594 |
| 5. | MEDHHINC10 | 7.30187266 |
| 6. | PCT\_NHBLACK10 | 4.70311986 |
| 7. | PCT\_NHASIAN10 | 3.93257041 |
| 8. | PCH\_NSLP\_09\_14 | 3.61751194 |
| 9. | PCH\_SNAP\_09\_14 | 2.32963838 |
| 10. | PCH\_SPECSPTH\_07\_12 | 1.36939138 |

**CONFUSION MATRIX -** Below is the confusion matrix for model GBM

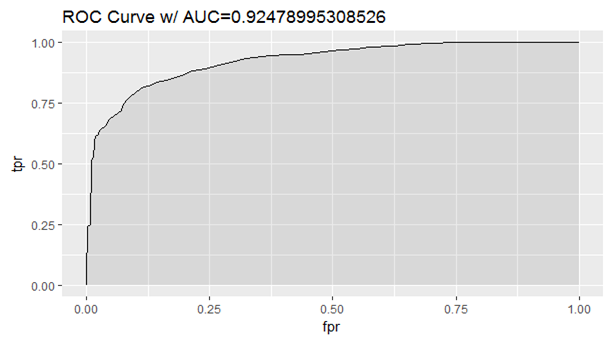
|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
| Predicted |  | Yes | No |
| Yes | 478 | 365 |
| No | 1 | 84 |

The models above were compared using Area Under the Curve (AUC) and other measures such as accuracy, sensitivity, and specificity as shown in table 2.8

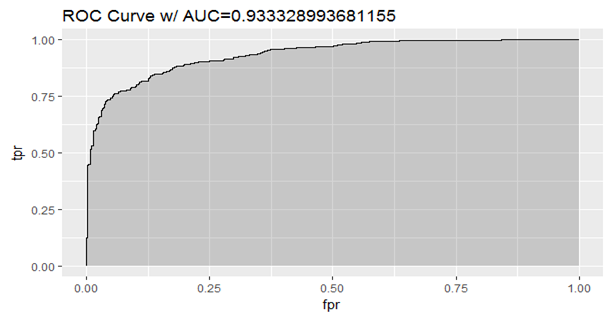
*Table 2.8 – Comparison for accuracy, sensitivity, and specificity for two models.*

|  |  |  |  |
| --- | --- | --- | --- |
| MODEL | ACCURACY | SENSITIVITY | SPECIFICITY |
| RANDOM FOREST | 76.51 | 95.20 | 56.57 |
| GRADIENT BOOSTING METHOD | 60.56 | 99.79 | 18.71 |

*Graph 2.10 – AUC for Random Forest*



*Graph 2.11 – AUC for Gradient Boosting Machine*



Model Selection:

For our problem predicting correctly if the region has more chances of having higher prevalence of diabetes is more important than correctly predicting if they don’t. Hence looking at the Confusion Matrix and AUC GBM does a better job than the other models. Hence **GBM** would be the best model to predict if a region has higher prevalence of diabetes for USDA food Environment dataset.

## Conclusion

Based on this analysis, Merck should target its resources within the United States in order to help combat the increased prevalence of Type 2 Diabetes by 2030. While the data does show that there are other countries that currently have a higher prevalence of the disease, when you combine this with economic considerations the potential scale in terms of cases and costs will be much more significant in the US. The most at-risk areas include Puerto Rico, Guam, Alabama, Mississippi and South Carolina.

We determined that the stepwise regression method is the best algorithm to predict if a region has a higher prevalence of diabetes as confirmed by the variance estimate and RMSE value. However, after pouring through the data to identify which country should be the area of focus for Merck, it was also important to determine the most influential factors in predicting the number of diagnosed Type 2 Diabetes over the next 13 years (until 2030). Then, we were able to combine the results to see how those factors were currently impacting the top five states. This now gives Merck specific feedback on how to tailor their response and organize their efforts.

In the news, Merck is already targeting its new drugs (Januvia and Janumet) in the US and, in fact, anticipates a contribution of $6 billion in sales to the company’s revenue stream. While there have been some setbacks related to the outcomes from patient studies, the company has managed to navigate these challenges and is currently conducting a study on a new drug that will help to reduce heart attacks and strokes among Type 2 Diabetes patients. There is tremendous competition in this area which will hopefully benefit millions of patients with the disease.

Data science continues to play a huge role in making sense of opportunities to improve our lives in many ways. Whether its business performance, technological discoveries or advancements in healthcare, the tools, methodologies and solutions provided by the data science community will no doubt help with the worldwide diabetes epidemic as well as many other causes in the future.

**Recommendations based on the results**

1. Merck should target its resources within the United States to help combat the increased number of diagnosed Type 2 Diabetes cases by 2030, which is predicted to increase by 41.31%.

2. Merck should invest in spreading awareness on diabetes in high prevalence states we identified, which can be achieved by targeting the following:

* High obesity rates in adults
* High population of 65 years or older
* High Hispanic population
* Low median household income
* High African American population

3. Initial analysis showed increased % of undiagnosed diabetes, hence the methods and symptoms chosen to detect diabetes should include:

* Glycosylated hemoglobin test should be done regularly. This test shows what a person's average blood glucose level was for the 2 to 3 months before the test. This can help determine how well a person's diabetes is being controlled over time.
* Flu complications (sometimes serious) and regular occurrence could be symptomatic of Diabetes. As a result, it becomes especially important for people with diabetes to get a flu vaccine.
* High BP patients should be tested for diabetes as this is sometimes the root cause of elevated BP
* Diabetes tends to increase bad cholesterol in the body, so high triglycerides or bad cholesterol in the body could be a symptom of diabetes

4. Resources should be directed towards existing programs such as:

* Supplemental Nutrition Assistance Program (SNAP)
* WIC Nutrition Program - Healthy food for your family
* National School Lunch Program (NSLP)

*References:*

# Shaw JE, Sicree RA, Zimmet PZ*.*2014*. “*Global estimates of the prevalence of diabetes for 2013 and 2035,*” Diabetes Res Clin Pract*(103:2),February, pp.137–149.

# Beagley Jessica, Guariguata Leonor, Weil Clara, Ayesha A. Motala*.*2014 *“*Global estimates of undiagnosed diabetes in adults*,” Diabetes Res Clin Pract*(103:2)February, pp.150–160.

1. **Appendix**  
     
   Clean-up for the **US Chronic Disease Indicator** dataset involved imputing missing values which was accomplished by inserting the mean.

|  |
| --- |
| newdata$DiagAdult[is.na(newdata$DiagAdult)] <- round(mean(newdata$DiagAdult, na.rm = TRUE))  newdata$eyeexam[is.na(newdata$eyeexam)] <- round(mean(newdata$eyeexam, na.rm = TRUE))  newdata$footexam[is.na(newdata$footexam)] <- round(mean(newdata$footexam, na.rm = TRUE))  newdata$glymeasure[is.na(newdata$glymeasure)] <- round(mean(newdata$glymeasure, na.rm = TRUE))  newdata$influall[is.na(newdata$influall)] <- round(mean(newdata$influall, na.rm = TRUE))  newdata$influadult[is.na(newdata$influadult)] <- round(mean(newdata$influadult, na.rm = TRUE))  newdata$prevdepre[is.na(newdata$prevdepre)] <- round(mean(newdata$prevdepre, na.rm = TRUE))  newdata$prevdiab[is.na(newdata$prevdiab)] <- round(mean(newdata$prevdiab, na.rm = TRUE))  newdata$HighBP[is.na(newdata$HighBP)] <- round(mean(newdata$HighBP, na.rm = TRUE))  newdata$Highchole[is.na(newdata$Highchole)] <- round(mean(newdata$Highchole, na.rm = TRUE)) |

Random Forest code

|  |
| --- |
| data\_csv <- read.csv("C:/Users/nehaa/Downloads/CDI2013.csv", header=TRUE, strip.white = TRUE, na.strings = c("NA","","NaN","?","."))  nrow(data\_csv)  colnames(data\_csv)  library(Hmisc)  newdata=data\_csv[,c(5,8,11,14,17,20,23,26,29,31,34)]  newdata <- setNames(newdata, c("DiagAdult","eyeexam","footexam","glymeasure","influall","influadult","prevdepre","prevdiab","gender","HighBP","Highchole"))  summary(newdata)  newdata$country<-data\_csv$LocationDesc  write.csv(newdata,"C:/Users/nehaa/Desktop/new\_CDI2013.csv")  library(pastecs)  stat.desc(newdata,c(1,4,5,9))  rcorr(newdata, type="pearson")  cor(newdata)$p.value  ncol(newdata)  nrow(newdata)  n<-nrow(newdata)  trainIndex = sample(1:n,  size = round(0.7\*n),  replace=FALSE) # We create an index for 70% of obs. by random  train\_data = newdata[trainIndex,] # We use the index to create training data  test\_data = newdata[-trainIndex,]  fit=lm(prevdiab ~ DiagAdult + prevdiab + eyeexam + footexam + glymeasure + influadult + influall + prevdepre + HighBP + Highchole + gender,data = train\_data)  step(fit, test="F")  final\_fit=lm(prevdiab ~ DiagAdult + glymeasure + influadult + HighBP + Highchole ,data = train\_data)  summary(final\_fit)  #pred <- pred(final\_fit, test\_data)  diaPred <- predict(final\_fit, test\_data)  diaPred  AIC(final\_fit)  actuals\_preds <- data.frame(cbind(actuals=test\_data$prevdiab, predicteds=diaPred)) # make actuals\_predicteds dataframe.  correlation\_accuracy <- cor(actuals\_preds)  correlation\_accuracy  head(actuals\_preds)  min\_max\_accuracy <- mean(apply(actuals\_preds, 1, min) / apply(actuals\_preds, 1, max))  min\_max\_accuracy  resid(final\_fit)  plot(density(resid(final\_fit)))  plot(fitted(final\_fit), residuals(final\_fit),xlabs="Predicted Scores", ylabs="Residuals") |

Random Forest Codes

|  |
| --- |
| library(randomForest)  mydata<-newdata  n<-nrow(mydata)  set.seed(32)  # Since the data is large, we sample the first 5k observations:  trainIndex = sample(1:n,  size = round(0.7\*n),  replace=FALSE) # We create an index for 70% of obs. by random  train\_data = mydata[trainIndex,] # We use the index to create training data  test\_data = mydata[-trainIndex,]  #train\_data <- head(mydata,n=75)  summary(train\_data)  nrow(train\_data)  rf <-randomForest(prevdiab~., data=train\_data, ntree=50, na.action=na.exclude, importance=T,  proximity=T)  print(rf)  nrow(train\_data)  # First, remove incomplete observations:  nrow(mydata)  train\_data <- mydata[complete.cases(mydata),]  nrow(train\_data)  mtry <- tuneRF(train\_data[-12], train\_data$prevdiab, ntreeTry=20,  stepFactor=1.5, improve=0.01, trace=TRUE, plot=TRUE)  best.m <- mtry[mtry[, 2] == min(mtry[, 2]), 1]  best.m  print(mtry)  print(best.m)  set.seed(32)  rf <-randomForest(prevdiab~., data=train\_data, mtry=best.m, importance=TRUE, ntree=20)  print(rf)  importance(rf)  varImpPlot(rf)  library(randomForest)  set.seed(32)  # Since the data is large, we sample the first 5k observations:  #data <- head(mydata, n = 6000)  mydata <- mydata[complete.cases(mydata),] # We only keep the observations with no missing values.  data<-mydata  n = nrow(data)  print(n)  trainIndex = sample(1:n, size = round(0.7\*n), replace=FALSE)  train\_data = data[trainIndex ,]  test\_data = data[-trainIndex ,]  summary(train\_data)  rf <-randomForest(prevdiab~., data=train\_data, ntree=5000, na.action=na.exclude, importance=T,  proximity=T)  print(rf)  mtry <- tuneRF(train\_data[-12], train\_data$prevdiab, ntreeTry=20,  stepFactor=1.5, improve=0.01, trace=TRUE, plot=TRUE)  best.m <- mtry[mtry[, 2] == min(mtry[, 2]), 1]  best.m  print(mtry)  importance(rf)  varImpPlot(rf)  library(caret)  predicted\_values <- predict(rf, test\_data,type= "response")  head(predicted\_values)  pred <- factor( ifelse(predicted\_values[,2] > threshold, 1, 0) ) |

The dataset **Diabetes Public Health Resource -** codes for Time series Forecast

|  |
| --- |
| data <- read.csv("C:/Users/nehaa/Downloads/us\_time.csv", header=TRUE, strip.white = TRUE,  na.strings = c("NA","","NaN","?","."))  install.packages("ts", repos = "https://cran.r-project.org")  install.packages("forecast", repos = "https://cran.r-project.org")  summary(data)  library(forecast)  x = ts(data[,2], start = c(1980,1), frequency=1)  plot(x)  z=log10(x)  plot(z)  y=diff(z)  plot(y)  par(mfrow = c(1,2))  acf(x,main='number of Diagnosed Diabetes')  pacf(x,main='number of Diagnosed Diabetes')  ARIMAfit = auto.arima(x, approximation=FALSE,trace=TRUE)  summary(ARIMAfit)  pred = predict(ARIMAfit, n.ahead = 16)  pred  par(mfrow = c(1,1))  plot(x,type='l',xlim=c(1980,2030),ylim=c(0,35),xlab = 'Year',ylab = 'Number')  lines((pred$pred),col='blue')  lines((pred$pred+2\*pred$se),col='orange')  lines((pred$pred-2\*pred$se),col='orange') |

Clean-up  and modeling for the US Food Environment dataset involved imputing missing values which was accomplished by inserting the mean followed by model planning on the same.

Data Preparation

|  |
| --- |
| setwd("E:/Spring 2017/Advanced BA/Project")  getwd()  data <- read.csv("Merged\_diabetes\_2010.csv",  sep=",", header=T, strip.white = T, na.strings = c("NA","NaN","","?"))  str(data)  summary(data)  data$Population.Estimate..2011 <- NULL  data$Population.Estimate..2012 <- NULL  data$CH\_FOODINSEC\_09\_12<- NULL  data$CH\_FOODINSEC\_09\_12 <- NULL  data$FOODINSEC\_CHILD\_03\_11 <- NULL  data$PCT\_OBESE\_ADULTS13 <- NULL  data$County.Name <- NULL  data$MEDHHINC10 <- as.numeric(data$MEDHHINC10)  data$POVRATE10 <- as.numeric(data$POVRATE10)  median(data$PCT\_DIABETES\_ADULTS10,na.rm=TRUE)  data$High <- ifelse(data$PCT\_DIABETES\_ADULTS10 >= 10.6, "Yes", "No")  data$PCT\_DIABETES\_ADULTS10 <- NULL  data$PCT\_DIABETES\_ADULTS10.1 <- NULL  data <- na.omit(data)  data$High <- as.factor(data$High)  summary(data)  set.seed(32)  data <- data[complete.cases(data), ] |

### Sampling test and train data

|  |
| --- |
| n = nrow(data)# n will be there number of obs. in data  trainIndex = sample(1:n,                     size = round(0.6\*n),                     replace=FALSE)  train\_data = data[trainIndex,]  test\_data = data[-trainIndex,] |

## Model 1 Random Forest

|  |
| --- |
| library(randomForest)  rf <-randomForest(High~., data=train\_data, ntree=80, na.action=na.exclude, importance=T,                   proximity=T)  print(rf)  mtry <- tuneRF(train\_data[-34], train\_data$High, ntreeTry=80,  stepFactor=1.5, improve=0.00, trace=TRUE, plot=TRUE, na.action=na.exclude)  best.m <- mtry[mtry[, 2] == min(mtry[, 2]), 1]  print(mtry)  print(best.m)  rf <-randomForest(High~., data=train\_data, mtry = 2, ntree=80, na.action=na.exclude, importance=T,                   proximity=T)  print(rf) |

##Check prediction

|  |
| --- |
| library(caret)  install.packages("plyr")  library(plyr)  predicted\_values <- predict(rf, test\_data,type= "prob")  head(predicted\_values)  threshold <- 0.2  pred <- factor( ifelse(predicted\_values[,2] > threshold, "Yes","No") )  head(pred)  levels(test\_data$High)[2]  confusionMatrix(pred, test\_data$High,positive = levels(test\_data$High)[2])  install.packages("e1071") |

### Plot

|  |
| --- |
| install.packages("ROCR")  install.packages("plyr")  library(ROCR)  library(ggplot2)  library(randomForest)  predicted\_values <- predict(rf, test\_data,type= "prob")[,2]  pred <- prediction( predicted\_values, test\_data$High)  perf <- performance(pred, measure = "tpr", x.measure = "fpr")  auc <- performance(pred, measure = "auc")  auc <- auc@y.values[[1]]  roc.data <- data.frame(fpr=unlist(perf@x.values),                        tpr=unlist(perf@y.values),                        model="RF")  ggplot(roc.data, aes(x=fpr, ymin=0, ymax=tpr)) +    geom\_ribbon(alpha=0.2) +    geom\_line(aes(y=tpr)) +    ggtitle(paste0("ROC Curve w/ AUC=", auc))  getwd()  importance(rf)  varImpPlot(rf) |

##Model 3 - GBM

|  |
| --- |
| set.seed(32)  gbm\_caret <- train(as.factor(High) ~ ., data = train\_data, method = "gbm", trControl = trainControl(method = "repeatedcv", number = 4, repeats = 4),verbose = FALSE)  summary(gbm\_caret)  predicted\_values <- predict(gbm\_caret, test\_data,type= "prob")  head(predicted\_values)  threshold <- 0.03  pred <- factor( ifelse(predicted\_values[,2] > threshold, "Yes","No") )  head(pred)  levels(test\_data$High)[2]  confusionMatrix(pred, test\_data$High,positive = levels(test\_data$High)[2])    predicted\_values <- predict(gbm\_caret, test\_data,type= "prob")[,2]  pred <- prediction( predicted\_values, test\_data$High)  perf <- performance(pred, measure = "tpr", x.measure = "fpr")  auc <- performance(pred, measure = "auc")  auc <- auc@y.values[[1]]  roc.data <- data.frame(fpr=unlist(perf@x.values),                        tpr=unlist(perf@y.values),                        model="GBM")  ggplot(roc.data, aes(x=fpr, ymin=0, ymax=tpr)) +    geom\_ribbon(alpha=0.2) +    geom\_line(aes(y=tpr)) +    ggtitle(paste0("ROC Curve w/ AUC=", auc)) |