DATA SCIENCE CASE STUDY

Counties with Increase Mortality Rate Prediction

Abstract

The objective of the document is present a noble way to predict the counties in United States which shall have a higher mortality rate compared to past

Gursewak Singh Sidhu

Sidhus234@gmail.com

Problem Statement:

Description

The Institute for Health Metrics and Evaluation (IHME) has published estimates of annual county-level age-standardized mortality rates for 21 mutually exclusive causes of death from 1980 to 2014. This data is available from http://ghdx.healthdata.org/record/united-states-mortality-rates-county-1980-2014 and you are free to supplement this dataset with any additional publicly available data you see fit.

Summary:

The report covers below key topics:

- Data Sources used for the solution
- Briefly explain the coding part
- Provide some insights on the model variable selection
- Details on aspects, that could be added, after discussion with business team

Index:

Data:

Three different sources of data was used for this assignment. The links and details about the data are given below:

1. Institute for Health Metrics and Evaluation (IHME) has published estimates of annual county-level age-standardized mortality rates for 21 mutually exclusive causes of death from 1980 to 2014. (http://ghdx.healthdata.org/record/united-states-mortality-rates-county-1980-2014)
2. County level Race by age group data (https://www2.census.gov/programs-surveys/popest/datasets/2010-2013/counties/totals/). Details about the data can be found here.
3. County level migration, death, birth data by US census (https://www2.census.gov/programs-surveys/popest/datasets/2010-2013/counties/). Details about the data can be found here.

Exploratory Data Analysis:

The jupyter notebook, "00 Mortality Analysis (By Gender, state and cause).ipynb" covers the basic exploratory data analysis part. In the code, we look at:

- Mortality rates by gender
- Top 5 reasons which contribute the most to mortality in 2014 (Population and gender level)
- Top 5 reasons which saw the most increase or decrease in mortality rate compared to 1980 (Population and gender level)
- The above two points are then analysed at the state level.

Data Manipulation:

This section covers the next three jupyter notebooks, shared as html files:

1. "**01 Data Prep - Mortality at county level by different causes.ipynb**": This notebook manipulates the IHME data to get variables at a county level. The final dataset has 462 features.
2. "**02 Data Prep - Mortality at county level by different causes by Gender.ipynb**": This notebook also manipulates the IHME data to get variables at a (gender x year x cause_type) level. This gives us additional 2590 features.
3. "***03 Data Prep County level population data.ipynb***": This notebook, manipulates the raw data which gives birth rate, deaths, migration (internal, external) at a county level. It gives us additional 60 variables
4. "***04 Data Prep 04.ipynb***": This notebook, creates 8323 features based on race type, and population change at county level.

Data Consolidation:

"***05 Data Prep for Model.ipynb***" covers the code to consolidate the above 4 data sources to one final dataset. Key changes in this code are:

- All variables for the period of 2013, and 2014 have been dropped.
- The objective of the predictive model is to predict "US Counties", which will have a higher mortality compared to 2014. Hence, no data from 2013, and 2014 can be used as independent variable.
- The structure of the dependent variable is as below:

2010	2011	2012	2013	2014
Observation Period			Lag Period	Prediction period

Dimensionality Reduction:

The final dataset has 3,197 rows, and 11,264 columns. The number of features is almost 4 times that the number of observations. To reduce the features:

- Information Value approach: A cut off of 0.10 was applied on information value. (To know more about information value, please refer: http://www.mwsug.org/proceedings/2013/AA/MWSUG-2013-AA14.pdf)
- After the IV, we are left with 3173 variables. To reduce them further, we did the principal component analysis and derived 20 new orthogonal features. There are different techniques that could have been applied as:
  - Factor Analysis
  - Variable Clustering
  - Correlation between independent variables
  - Multi-collinearity reduction
  - Variable Importance in Random Forest/GBM model
- Due to shortage of time, PCA was preferred as it gives orthogonal components.

Model Selection and Performance:

Random Forest:

Feature Selection:

Information value and PCA was done, as explained in logistic regression model. Once, we have 20 variables, we do a grid search by varying the hyper-parameters. And based on the performance of the models, we select a 6 variable model

Performance:

The performance of the model is:

Metric	Train Sample	Test Sample
KS	44.6%	42.6%
Accuracy	79%	78%

Comments:

The final variable importance and the partial dependence plots are not given, as the feature space if Principal components. It is difficult to draw a cause affect relationship. For variable reduction, we used IV and PCA, though a random forest has internal "variable importance measure". To reduce variables by RF variable importance, an important rule of thumb is "No of trees = 10 times no of independent variables", which would have given us 100,000+ trees. Considering the time constraint and resources, we used IV and PCA to reduce variables.

Gradient Boosting Machines:

Feature Selection:

Information value and PCA was done, as explained in logistic regression model. Once, we have 20 variables, we do a grid search by varying the GBM hyper-parameters. And based on the performance of the models, we select 8 variable model

Performance:

The performance of the model is:

Metric	Train Sample	Test Sample
KS	52.4%	44.6%
Accuracy	81%	78%

Comments:

Due to shortage of time, the variable selection process was based on IV, and PCA. As mentioned before in the document, it is not the best way to reduce features for a GBM model. Each modelling technique has its own criteria to select/utilize the information of the model. In case of GBM (Tree based models), it provides an internal measure of Variable importance which could have been used. The shortcoming of this measure is that the variable importance is hyper-parameter dependent (If we vary the hyper-parameters, the variable importance will change). To overcome that, we select a base hyper-parameter and reduce variables only using that configuration.

Index 1: Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin

CC-EST2013-ALLDATA-[ST-FIPS]: Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2013

File: 7/1/2013 County Characteristics Resident Population Estimates

Source: U.S. Census Bureau, Population Division

Release Date: June 2014

Sort order of observations: COUNTY, YEAR, and AGEGRP within STATE

Data fields (in order of appearance):

VARIABLE DESCRIPTION

- SUMLEV: Geographic Summary Level
- STATE: State FIPS Code
- COUNTY: County FIPS Code
- STNAME: State Name
- CTYNAME: County Name
- YEAR: Year
- AGEGRP: Age group (See code below)
- TOT_POP: Total population
- TOT_MALE: Total male population
- TOT_FEMALE: Total female population
- WA_MALE: White alone male population
- WA_FEMALE: White alone female population
- BA_MALE: Black or African American alone male population
- BA_FEMALE: Black or African American alone female population
- IA_MALE: American Indian and Alaska Native alone male population
- IA_FEMALE: American Indian and Alaska Native alone female population
- AA_MALE: Asian alone male population
- AA_FEMALE: Asian alone female population
- NA_MALE: Native Hawaiian and Other Pacific Islander alone male population
- NA_FEMALE: Native Hawaiian and Other Pacific Islander alone female population
- TOM_MALE: Two or More Races male population
- TOM_FEMALE: Two or More Races female population
- WAC_MALE: White alone or in combination male population
- WAC_FEMALE: White alone or in combination female population
- BAC_MALE: Black or African American alone or in combination male population
- BAC_FEMALE: Black or African American alone or in combination female population
- IAC_MALE: American Indian and Alaska Native alone or in combination male population
- AAC_MALE: Asian alone or in combination male population
- IAC_FEMALE: American Indian and Alaska Native alone or in combination female population.
- AAC_FEMALE: Asian alone or in combination female population
- NAC_MALE: Native Hawaiian and Other Pacific Islander alone or in combination male population

- NAC_FEMALE: Native Hawaiian and Other Pacific Islander alone or in combination female population
- NH_MALE: Not Hispanic male population
- NH_FEMALE: Not Hispanic female population
- NHWA_MALE: Not Hispanic, White alone male population
- NHWA_FEMALE: Not Hispanic, White alone female population
- NHBA_MALE: Not Hispanic, Black or African American alone male population
- NHBA_FEMALE: Not Hispanic, Black or African American alone female population
- NHIA_MALE: Not Hispanic, American Indian and Alaska Native alone male population
- NHIA_FEMALE: Not Hispanic, American Indian and Alaska Native alone female population
- NHAA_MALE: Not Hispanic, Asian alone male population
- NHAA_FEMALE: Not Hispanic, Asian alone female population
- NHNA_MALE: Not Hispanic, Native Hawaiian and Other Pacific Islander alone male population
- NHNA_FEMALE: Not Hispanic, Native Hawaiian and Other Pacific Islander alone female population
- NHTOM_MALE: Not Hispanic, Two or More Races male population
- NHTOM_FEMALE: Not Hispanic, Two or More Races female population
- NHWAC_MALE N:ot Hispanic, White alone or in combination male population
- NHWAC_FEMALE: Not Hispanic, White alone or in combination female population
- NHBAC_MALE: Not Hispanic, Black or African American alone or in combination male population
- NHBAC_FEMALE: Not Hispanic, Black or African American alone or in combination female population
- NHIAC_MALE: Not Hispanic, American Indian and Alaska Native alone or in combination male population
- NHIAC_FEMALE: Not Hispanic, American Indian and Alaska Native alone or in combination female population
- NHAAC_MALE: Not Hispanic, Asian alone or in combination male population
- NHAAC_FEMALE: Not Hispanic, Asian alone or in combination female population
- NHNAC_MALE: Not Hispanic, Native Hawaiian and Other Pacific Islander alone or in combination male population
- NHNAC_FEMALE: Not Hispanic, Native Hawaiian and Other Pacific Islander alone or in combination female population
- H_MALE: Hispanic male population
- H_FEMALE: Hispanic female population
- HWA_MALE: Hispanic, White alone male population
- HWA_FEMALE: Hispanic, White alone female population
- HBA_MALE: Hispanic, Black or African American alone male population
- HBA_FEMALE: Hispanic, Black or African American alone female population
- HIA_MALE Hispanic, American Indian and Alaska Native alone male population
- HIA_FEMALE Hispanic, American Indian and Alaska Native alone female population
- HAA_MALE Hispanic, Asian alone male population
- HAA_FEMALE Hispanic, Asian alone female population
- HNA_MALE Hispanic, Native Hawaiian and Other Pacific Islander alone male population
- HNA_FEMALE Hispanic, Native Hawaiian and Other Pacific Islander alone female population
- HTOM_MALE Hispanic, Two or More Races male population

- HTOM_FEMALE Hispanic, Two or More Races female population
- HWAC_MALE Hispanic, White alone or in combination male population
- HWAC_FEMALE Hispanic, White alone or in combination female population
- HBAC_MALE Hispanic, Black or African American alone or in combination male population
- HBAC_FEMALE Hispanic, Black or African American alone or in combination female population
- HIAC_MALE Hispanic, American Indian and Alaska Native alone or in combination male population
- HIAC_FEMALE Hispanic, American Indian and Alaska Native alone or in combination female population
- HAAC_MALE Hispanic, Asian alone or in combination male population
- HAAC_FEMALE Hispanic, Asian alone or in combination female population
- HNAC_MALE Hispanic, Native Hawaiian and Other Pacific Islander alone or in combination male population
- HNAC_FEMALE Hispanic, Native Hawaiian and Other Pacific Islander alone or in combination female population

The key for SUMLEV is as follows:

50 County and/or Statistical Equivalent

The key for the YEAR variable is as follows:

- 1 = 4/1/2010 Census population
- 2 = 4/1/2010 population estimates base
- 3 = 7/1/2010 population estimate
- 4 = 7/1/2011 population estimate
- 5 = 7/1/2012 population estimate
- 6 = 7/1/2013 population estimate

The key for AGEGRP is as follows:

- 0 = Total
- 1 = Age 0 to 4 years
- 2 = Age 5 to 9 years
- 3 = Age 10 to 14 years
- 4 = Age 15 to 19 years
- 5 = Age 20 to 24 years
- 6 = Age 25 to 29 years
- 7 = Age 30 to 34 years
- 8 = Age 35 to 39 years
- 9 = Age 40 to 44 years
- 10 = Age 45 to 49 years
- 11 = Age 50 to 54 years
- 12 = Age 55 to 59 years
- 13 = Age 60 to 64 years
- 14 = Age 65 to 69 years
- 15 = Age 70 to 74 years
- 16 = Age 75 to 79 years
- 17 = Age 80 to 84 years

- 18 = Age 85 years or older

Note: "In combination" means in combination with one or more other races. The sum of the five race groups adds to more than the total population because individuals may report more than one race. The estimates are based on the 2010 Census and reflect changes to the April 1, 2010 population due to the Count Question Resolution program and geographic program revisions. Hispanic origin is considered an ethnicity, not a race. Hispanics may be of any race. Responses of "Some Other Race" from the 2010 Census are modified. This results in differences between the population for specific race categories shown for the 2010 Census population in this file versus those in the original 2010 Census data. For more information, see http://www.census.gov/popest/data/historical/files/MRSF-01-US1.pdf. For population estimates methodology statements, see http://www.census.gov/popest/methodology/index.html.

Index 2: Annual Resident Population Estimates, Estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for States and Counties

CO-EST2013-alldata: Annual Resident Population Estimates, Estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for States and Counties: April 1, 2010 to July 1, 2013

File: 7/1/2013 County Population Estimates

Source: U.S. Census Bureau, Population Division

Release Date: March 2014

Sort order of observations: Counties within State in FIPS code sort

Data fields (in order of appearance):

- SUMLEV Geographic summary level
- REGION Census Region code
- DIVISION Census Division code
- STATE State FIPS code
- COUNTY County FIPS code
- STNAME State name
- CTYNAME County name
- CENSUS2010POP 4/1/2010 resident total Census 2010 population
- ESTIMATESBASE2010 4/1/2010 resident total population estimates base
- POPESTIMATE2010 7/1/2010 resident total population estimate
- POPESTIMATE2011 7/1/2011 resident total population estimate
- POPESTIMATE2012 7/1/2012 resident total population estimate
- POPESTIMATE2013 7/1/2013 resident total population estimate
- NPOPCHG2010 Numeric change in resident total population 4/1/2010 to 7/1/2010
- NPOPCHG2011 Numeric change in resident total population 7/1/2010 to 7/1/2011
- NPOPCHG2012 Numeric change in resident total population 7/1/2011 to 7/1/2012
- NPOPCHG2013 Numeric change in resident total population 7/1/2012 to 7/1/2013
- BIRTHS2010 Births in period 4/1/2010 to 6/30/2010
- BIRTHS2011 Births in period 7/1/2010 to 6/30/2011
- BIRTHS2012 Births in period 7/1/2011 to 6/30/2012
- BIRTHS2013 Births in period 7/1/2012 to 6/30/2013
- DEATHS2010 Deaths in period 4/1/2010 to 6/30/2010
- DEATHS2011 Deaths in period 7/1/2010 to 6/30/2011
- DEATHS2012 Deaths in period 7/1/2011 to 6/30/2012
- DEATHS2013 Deaths in period 7/1/2012 to 6/30/2013
- NATURALINC2010 Natural increase in period 4/1/2010 to 6/30/2010
- NATURALINC2011 Natural increase in period 7/1/2010 to 6/30/2011
- NATURALINC2012 Natural increase in period 7/1/2011 to 6/30/2012
- NATURALINC2013 Natural increase in period 7/1/2012 to 6/30/2013

- INTERNATIONALMIG2010 Net international migration in period 4/1/2010 to 6/30/2010
- INTERNATIONALMIG2011 Net international migration in period 7/1/2010 to 6/30/2011
- INTERNATIONALMIG2012 Net international migration in period 7/1/2011 to 6/30/2012
- INTERNATIONALMIG2013 Net international migration in period 7/1/2012 to 6/30/2013
- DOMESTICMIG2010 Net domestic migration in period 4/1/2010 to 6/30/2010
- DOMESTICMIG2011 Net domestic migration in period 7/1/2010 to 6/30/2011
- DOMESTICMIG2012 Net domestic migration in period 7/1/2011 to 6/30/2012
- DOMESTICMIG2013 Net domestic migration in period 7/1/2012 to 6/30/2013
- NETMIG2010 Net migration in period 4/1/2010 to 6/30/2010
- NETMIG2011 Net migration in period 7/1/2010 to 6/30/2011
- NETMIG2012 Net migration in period 7/1/2011 to 6/30/2012
- NETMIG2013 Net migration in period 7/1/2012 to 6/30/2013
- RESIDUAL2010 Residual for period 4/1/2010 to 6/30/2010
- RESIDUAL2011 Residual for period 7/1/2010 to 6/30/2011
- RESIDUAL2012 Residual for period 7/1/2011 to 6/30/2012
- RESIDUAL2013 Residual for period 7/1/2012 to 6/30/2013
- GQESTIMATESBASE2010 4/1/2010 Group Quarters total population estimates base
- GQESTIMATES2010 7/1/2010 Group Quarters total population estimate
- GQESTIMATES2011 7/1/2011 Group Quarters total population estimate
- GQESTIMATES2012 7/1/2012 Group Quarters total population estimate
- GQESTIMATES2013 7/1/2013 Group Quarters total population estimate
- RBIRTH2011 Birth rate in period 7/1/2010 to 6/30/2011
- RBIRTH2012 Birth rate in period 7/1/2011 to 6/30/2012
- RBIRTH2013 Birth rate in period 7/1/2012 to 6/30/2013
- RDEATH2011 Death rate in period 7/1/2010 to 6/30/2011
- RDEATH2012 Death rate in period 7/1/2011 to 6/30/2012
- RDEATH2013 Death rate in period 7/1/2012 to 6/30/2013
- RNATURALINC2011 Natural increase rate in period 7/1/2010 to 6/30/2011
- RNATURALINC2012 Natural increase rate in period 7/1/2011 to 6/30/2012
- RNATURALINC2013 Natural increase rate in period 7/1/2012 to 6/30/2013
- RINTERNATIONALMIG2011 Net international migration rate in period 7/1/2010 to 6/30/2011
- RINTERNATIONALMIG2012 Net international migration rate in period 7/1/2011 to 6/30/2012
- RINTERNATIONALMIG2013 Net international migration rate in period 7/1/2012 to 6/30/2013
- RDOMESTICMIG2011 Net domestic migration rate in period 7/1/2010 to 6/30/2011
- RDOMESTICMIG2012 Net domestic migration rate in period 7/1/2011 to 6/30/2012
- RDOMESTICMIG2013 Net domestic migration rate in period 7/1/2012 to 6/30/2013
- RNETMIG2011 Net migration rate in period 7/1/2010 to 6/30/2011
- RNETMIG2012 Net migration rate in period 7/1/2011 to 6/30/2012
- RNETMIG2013 Net migration rate in period 7/1/2012 to 6/30/2013

The key for SUMLEV is as follows:

040 = State and/or Statistical Equivalent
050 = County and /or Statistical Equivalent

The key for REGION is as follows:

1 = Northeast
2 = Midwest
3 = South
4 = West

The key for DIVISION is as follows:

1 = New England
2 = Middle Atlantic
3 = East North Central
4 = West North Central
5 = South Atlantic
6 = East South Central
7 = West South Central
8 = Mountain
9 = Pacific

Note: Total population change includes a residual. This residual represents the change in population that cannot be attributed to any specific demographic component. See Population Estimates Terms and Definitions at http://www.census.gov/popest/about/terms.html.

Net international migration in the United States includes the international migration of both native and foreign-born populations. Specifically, it includes: (a) the net international migration of the foreign born, (b) the net migration between the United States and Puerto Rico, (c) the net migration of natives to and from the United States, and (d) the net movement of the Armed Forces population between the United States and overseas.

The estimates are based on the 2010 Census and reflect changes to the April 1, 2010 population due to the Count Question Resolution program and geographic program revisions. See Geographic Terms and Definitions at http://www.census.gov/popest/about/geo/terms.html for a list of the states that are included in each region and division. All geographic boundaries for these population estimates are as of January 1, 2013. For population estimates methodology statements, see http://www.census.gov/popest/methodology/index.html.