Predicting Cause Of Death Data Analysis, (Spring 2016)

Ben Dykstra BFDYKSTRA@EMAIL.WM.EDU
College of William and Mary Williamsburg, VA 23187 USA

GwonJae Cho GCHO@EMAIL.WM.EDU
College of William and Mary Williamsburg, VA 23187 USA

Abstract

In this paper we use death records provided by the Center for Disease Control to classify the way that people die. We employ SVM, decision trees and K-Nearest Neighbors classification algorithms to classify an individuals manner of death, with the added challenge that many of the predictors are categorical variables.

1. Introduction

Each year, the CDC puts out the most detailed report on death in the United States. It is a record of every death in 2014, with details about the demographic and cause for each deceased individual. In the past, the US government has used this data to estimate life expectancy and generally understand trends in death in the United States. We used this data to try and predict the manner in which individuals died.

1.1. Motivation

The proportion of the aging population in the US is increasing every year. That also unfortunately means that a lot of those people will be dying. In order to help treat individuals, we first need to understand *how* they are dying. If we can predict how someone is likely to die based on their age, sex, race, marital status and many other variables, then we can start to move away from retroactive healthcare to preventative care.

Ideally, given information about an individual, we should be able to predict what they are at risk of, and help mitigate that risk. The effects of this can not be understated: Longer life span, better quality of life and more time spent with loved ones. Ultimately, death is a fact of life, and understanding that fact can help bring solace to many people.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

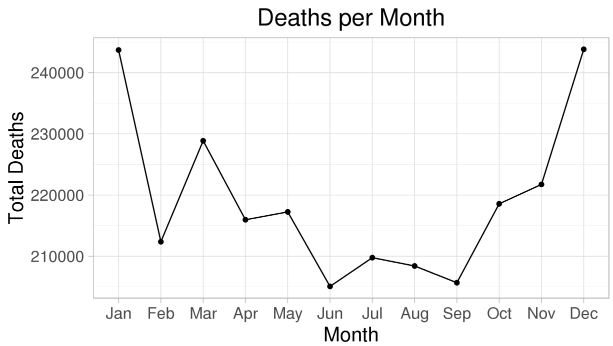
2. The Data

The data provided by the CDC is very rich. It includes information about the individuals race, sex, age, education level, marital status, whether they were a continental resident, the date of death, if they died while doing an activity, and of course a very detailed description of the underlying cause of death. Cause of death is reported according to the Icd10 standard. In 2014, there were 2,631,171 recorded deaths.

Lets take a look at some of the leading causes of death for 2014:

Cause	**Count**
Atherosclerotic Heart Disease	161961
Malignant neoplasm: Bronchus or lung, unspecified	154862
Unspecified dementia	122021
Acute myocardial infarction, unspecified	114107
Chronic obstructive pulmonary disease, unspecified	107836
Alzheimer disease, unspecified	91356
Stroke, not specified as haemorrhage or infarction	65578
Atherosclerotic cardiovascular disease, so described	60471
Congestive heart failure	60420

As expected heart problems, cancer and alzheimer's disease are the leading causes of death in the US. Next, lets look at the amount of deaths by month.



So, with the cold, also comes a lot of people dying. Now, lets look at the leading cause of death by age group. For ages 0 to 14, the top ten causes of death are

Cause, Age 0 to 14	**Count**
Extreme immaturity	2523
Sudden infant death syndrome	1457
Other ill-defined and unspecified causes of mortality	1208
Accidental Suffocation and strangulation in bed	819
Other preterm infants	736
Fetus and Newborn affected by premature rupture of membranes	553
Congenital malformation of heart, unspecified	547
Edwards syndrome, unspecified	399
Fetus and newborn affected by imcompetent cervix	376
Person injured in unspecified motor-vehicle accident	357

For ages 15 to 30, the predominant causes of death are by firearm, narcotic, car accident or suicide.

Cause, Age 15 to 30	**Count**
Assault by other and unspecified firearm discharge	4946
Accidental poisoning by and exposure to narcotics	3859
Person injured in unspecified motor-vehicle accident, traffic	3707
Accidental poisoning by and exposure to other unspecified drugs	3247
Intentional self-harm by hanging, strangulation and suffocation	3220
Intentional self-harm by other and unspecified firearm discharge	2149
Person injured in collision with other motor vehicles (traffic)	1079
Intentional self-harm by handgun discharge	926
Other ill-defined and unspecified causes of mortality	810
Pedestrian injured in traffic accident with other vehicle	671

The leading cause of death of young people in the United States is by firearm and narcotic overdose! It is one thing to hear about it in the news, however, it is another to see it in the raw data and the actual counts of deaths. Now, for people age 31 to 45, the causes begin to change:

Cause, Age 31 to 45	**Count**
Accidental poisoning by and exposure to other and unspecified drugs, medicaments and biological substances	6357
Accidental poisoning by and exposure to narcotics	5819
Intentional self-harm by hanging, strangulation and suffocation	3275
Assault by other and unspecified firearm discharge	2721
Intentional self-harm by other and unspecified firearm discharge	2606
Malignant neoplasm: Breast, unspecified	2545
Acute myocardial infarction, unspecified	2515
Person injured in unspecified motor-vehicle accident, traffic	2187
Hypertensive heart disease without (congestive) heart failure	1977
Atherosclerotic cardiovascular disease, so described	1891

Again, we see that drugs take an inordinate toll on our population as do suicides and guns. We take a look now at the middle aged population aged 46 to 65, which accounts for 21.23 % of our sample:

Cause, Age 46 to 65	Count
Malignant neoplasm: Bronchus or lung, unspecified	44949
Acute myocardial infarction, unspecified	27392
Atherosclerotic heart disease	22449
Atherosclerotic cardiovascular disease, so described	19872
Chronic obstructive pulmonary disease	16366
Malignant neoplasm: Breast, unspecified	15119
Malignant neoplasm: Pancreas, unspecified	11924
Malignant neoplasm: Colon, unspecified	11211
Hypertensive heart disease without (congestive) heart failure	10603
Alcoholic cirrhosis of liver	9379

As expected, cancer and heart disease overtake as leading causes of death. However, one should also note another disturbing trend in middle aged people.

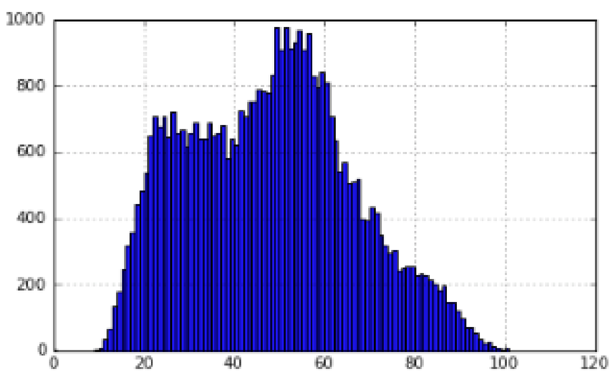


Figure 1. Suicides by Age

Not only do middle aged people suffer disproportionately from cancer and heart problems, they also commit suicide at a higher rate relative to the rest of the population. After about age 60, though the rates taper off greatly. The rest of the population, aged 65 and over accounts for 71.51 % of the deaths in the US.

Cause, Age 65 and up	Count
Atherosclerotic heart disease	137514
Unspecified dementia	120423
Malignant neoplasm: Bronchus or lung, unspecified	108354
Chronic obstructive pulmonary disease, unspecified	91112
Alzheimer disease, unspecified	90264
Acute myocardial infarction, unspecified	84025
Stroke, not specified as haemorrhage or infarction	58349
Congestive heart failure	55461
Atherosclerotic cardiovascular disease, so described	38555
Pneumonia, unspecified	37814

These data give a good indication about how age should affect manner of death. If the individual is over 60, chances are they died because of natural causes. If the individual is young, chances are that they died of a homicide, accident, or suicide. Here are the rest of the manners of death, by age.

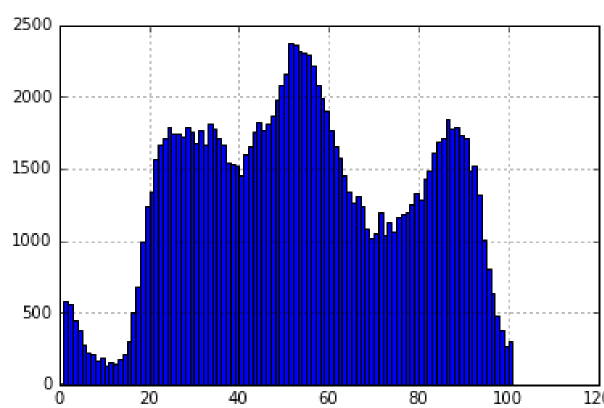


Figure 1. Accidental death by Age

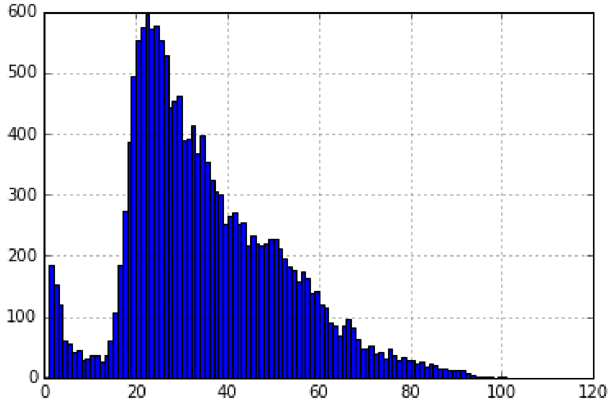


Figure 2. Homicides by Age

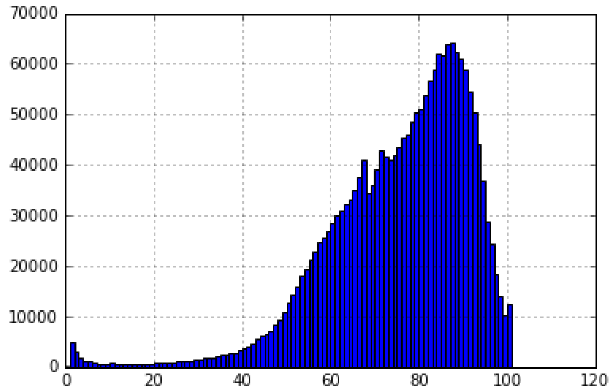


Figure 5. Natural Death by Age

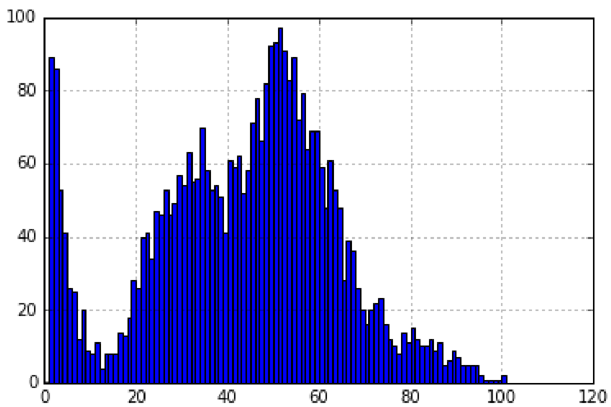


Figure 3. Pending Investigation by Age

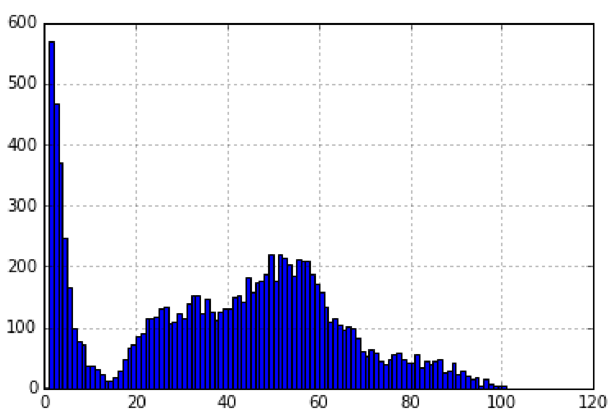


Figure 4. Could not Determine by Age

3. Methodology

In order to prepare our data for analysis, we first had to recode many of the categorical variables into binary variables. The variables that we recoded were sex, race, marital status, if they were injured at work and where they died. To randomly split up the data into testing and training sets, such that the training data comprised 70 % of the data and the test set was made up of the remainder of the data. For the decision tree algorithm, the data did not need to be recoded into binary variables because the algorithm is designed to handle categorical variables. We did not recode education level because they are ordinal and can be interpreted in relation to each other.

4. Result

We implemented four classification techniques in an attempt to predict the manner of death: K-Nearest Neighbor(KNN), Decision Tree, Logistic Regression and Support Vector Machine(SVM). In K-Nearest Neighbor, each test object follows the majority class of its k number of nearest objects, and we picked k=5 heuristically. The manner of death is consisted with 8 categories.

Code	Description
1	Accident
2	Suicide
3	Homicide
4	Pending Investigation
5	Could not determine
6	Self-Inflicted
7	Natural
0	Not Specified

Table 1. Manner of Death Code

Each code maps to a description of the manner of death. Each execution on the four techniques was ran on 400,000 rows(representing a person in each row), and each row contains a code of the manner of death.

Code(Manner of Death)	Num. of Classified
0	215,479
1	21,040
2	7,404
3	3,051
4	934
5	1,864
6	0
7	150,201
Total	400,000

Table 2. Total number of people classified in Manner of Death

After learning the train data, SVM returns the highest accuracy in classification, following with Decision Tree, KNN and Logistic Regression.

KNN	Decision Tree	Logistic	SVM
0.6877	0.7191	0.6868	0.7298

Table 3. Accuracy of test data each trained features with 8 code of Manner of Death

We inspect the counts of the classifications each technique fails to predict. K-Nearest Neighbor and Logistics Regression wrongly predicted proportionally to the total number. However, Decision Tree and SVM show leaning tendency to code 0 and less on 7 vise-versa.

Code	KNN	DecTree	Logistics	SVM
0	20,856	25,301	22,991	26,991
1	2,157	2,023	3,277	2,383
2	1,103	1,186	53	681
3	200	426	29	126
4	6	58	0	0
5	44	216	0	21
6	0	0	0	0
7	13,105	4,496	11,235	2,225
Total	37,471	33,706	37,585	32,427

Table 4. Incorrect classification categorized by the Manner of death Code

The test included all the codes of manner of death, but possible bias could have affected the result such as majority of the data is composed of code 0 and 7; they occupy 365,000 out of 400,000. Also, the description of the code 0 and 7 are Natural and Not Specified death. Therefore, we tested again without data that maps to 0 and 7 to investigate whether the trained data can predict the special cases.

Code(Manner of Death)	Num. of Classified
1	21,040
2	7,404
3	3,051
4	934
5	1,864
6	0
Total	34,293

Table 5. Total number of people classified in Manner of Death

Overall, the accuracy after removing data with code 0 and 7 decreased, but does not drop drastically than expected. The steepest decline was occurred in Decision Tree, whereas the least, KNN, drops 0.03.

KNN	Decision Tree	Logistic	SVM
0.6564	0.6215	0.6216	0.6675

Table 6. Accuracy of test data each trained features with 6 code of Manner of Death

The table 7 shows the counts of each algorithm wrongly classified data without code 0 and 7. By only looking at the sums and comparing them with the previous table 4, we can consider very little fluctuation of counts at code 1,2,3,6. However, the wrong predictions on code 4,5, which are Pending Investigation and Could not Determine, largely increased. Here, we can hypothesize that data with puzzling combination of features might be classified to code 4 or 5.

Code	KNN	DecTree	Logistics	SVM
1	2115	1953	3378	2486
2	923	1098	82	585
3	230	414	0	154
4	132	134	174	93
5	134	295	258	104
6	0	0	0	0
Total	3,534	3,894	3,892	3,422

Table 7. Incorrect classification categorized by the Manner of death Code

Conclusion

We have investigated the cause of death data, utilized four different classification techniques for predictions and analyzed the probabilities and incorrect classification counts. Decision Tree shows the highest accuracy on data with all the manner of death codes, but shrank to the lowest after taking Natural and Not Specified death data. After opted to code 1 through 6, we observed that counts in code 1,2,3,6 change very slightly while 4 and 5 increased drastically. Therefore, further studies on this peculiarity can be conducted in the future, and other techniques including neural network would be an option to enrich this studies.

References

[1] Murphy SL, Kochanek KD, Xu JQ, Arias E. *Mortality in the United States, 2014.*. NCHS data brief, no 229. Hyattsville, MD: National Center for Health Statistics. 2015.