GEORGETOWN UNIVERSITY
School *of* Continuing Studies
Center for Continuing and Professional Education

Cohort 12
Mortality Probability Team
Final Project Report

Introduction

The US Center for Disease Control has long emphasized that social factors are contributing elements in one's mortality at any given age (Song *et al.* 2011). Researchers indicate that socio-economic factors such as income, education, race and gender, as well as medical technology (Case and Deaton, 2017 and Cutler, Huang and Lleras-Muney, 2016), help determine the age at which individuals die. Although improving medical technology across time attenuates impact of illnesses, social factors play a great role in exposure to illnesses and how one can access this health improving technology. It is not hard to imagine that socio-economic factors impact health and thereby mortality: income can help access health insurance; one's location shapes the difficulty of accessing health care during life threatening emergencies; one's income and social status, determined through such factors as race, education or occupation, may induce dietary and other habits detrimental to one's health.

Since 1980, motivated by on-going research on socio-economic determinants of mortality, the US Census Bureau has provided data that follows up a sample of individuals from the census to match death certificates of individuals in case death occurs within a particular time period. Socio-economic data provided for each person is from the census year; and no further socio-economic data is collected after that year. These data have been made available as public use microdata sample (or PUMS). The National Longitudinal Mortality Study (NLMS) provides the PUMS data as a "specialized extract of the full NLMS that is designed to provide easy access to an extract of the main NLMS while protecting the confidentiality of those who have responded to the original surveys on which the NLMS is based" (US Census, 2013). The first set of NLMS data made available as a PUMS provided basic data from the 1980 census and matched the death certificates of individuals dying within the subsequent eleven years. In the cases where death occurred NLMS provided the figure for elapsed number of days between the day the individual was interviewed and death; it also provided a cause of death. To the census questions an added question indicated the status of enrolment into a health insurance (US Census Bureau, 2013, and CDC 2015).

NLMS data was intended primarily to identify distal and proximal causes of death; this type of information would lead to establishing both social policies as well as priorities for medical research. For insurers, from this data set, it may also be possible to show how the distal socio-economic causes can yield predictive probabilities for death of a given person. As one of the primary purposes of data analytics is to generate predictions, the PUMS data on mortality was used towards predicting the probability of dying given the socio-economic profile and the age of the person.

Hypothesis

To explain an occurrence of death amid a population, an event binomially distributed, most authors use a logistic relation where person's age figures prominently. Age is a primary factor in determining the probability of death; it is almost the basic notion of risk in demography: for example, CDC reports that average age a person living in the US died in 1950 was 68, while it was nearly 80 by 2007 (CDC, 2008). While the US population death rate for the overall population is 0.8%, for those above 55 it starts to reach 1.3% and becomes 15% at age 85. Age is a primary factor in being able to predict the probability of person's death.

The standard socio-economic factors that are likely to shape mortality are education, income, social status determined by the type of occupation and one's ethnic status, conditions of one's birth and childhood rearing and access to medical care. Aside from generating income, education provides understanding that help adopt healthy behavior such as exercising and dietary habits. Income enables one to have more suitable housing, access health care and afford healthier dietary habits. It has been noted that childhood conditions can affect health in later years (Case, Fertig and Paxon, 2005); as where one is born may determine some childhood condition, varying situations in one's birth country may contribute to future health.

Place of residency impacts health through availability of medical care, crime, community cohesion and pollution. PUMS version of NLMS did not provide finer details of location; it provided residency status of a person by state. Using state as a feature one can capture the locality-level factors such as average access to medical care, social cohesion and some occupational factors. For example, many coastal states have fishing industries or off-shore drilling which are among the most dangerous of all occupations. It is possible that states within a region are similar with respect to a multiple number of factors such as political structure that shape social policy and the composition of the population.

Health insurance gives access to medical care and should be a factor in improving health. However, ascertaining the impact of health insurance even on usage of care, let alone health, has been extremely difficult through statistical methods (see Finkelstein, 2017). Changing medical technology when used at a mass scale contributes to health of the population; usually this is seen as a statistical control for missing variable when socio-economic factors are used to explain mortality. Various methods have been used for this approach; for example, a time factor can be used to note technological change.

Obviously, health condition of a person at a particular time is a strong predictor of death probability in the near future. In order to reduce missing variables' impact, one should include health conditions at the time of survey when analyzing contribution of socio-economic factors to explain mortality. However, data on obtaining health information in most surveys are difficult. Self-reporting in health is usually biased; and collecting objective information is difficult (Murray and Chen, 1992).

We, members of the project team, tested the central hypothesis: *One can predict individual mortality through finding a relation between whether a person died or not within a period of 11 years and factors or features such as poverty, education, employment status, martial status, place of birth, being in the labor market, state residency and health insurance status.*

PUMS version of NLMS data set provided all these variables. We approximated being in the labor market (not necessarily employed or currently seeking employment) through having a social security number, this also shows the level at which a person is established in the US. The citizenship information was incomplete, as this may not have been a question always asked during a census. Our presentation shows how these features were used. We had no features representing evolving medical technology.

Deaths occurred at different times following the initial census interview and we ascertained that across the eleven years there are visible signs of age patterns in mortality rate, we could not incorporate time as a proxy for technology in our analysis. If a temporal factor such as follow-up time till death were to be used, it would create a spurious relation as mortality increases with time due to people in the sample getting older. Most likely to measure the impact of technology in constructing probability of death would require a different approach than we were able to use for this project. Information on health conditions at the time of survey was incomplete in the data set. Correlated with health condition are indicators that describe healthy behavior. No such data was available. These limitations, however, do not prevent us from providing a way to estimate a probability measure for one's death using mostly socio-economic factors.

Application and Motivation

Our intention was to provide a method that can be used to predict probability of death over an eleven year period for a given person. Due to the data acquired for this project, we incorporated primarily socio-economic status of an individual to predict mortality of a given person within the next 11 years. The intention was not to limit us completely to socio-economic factors; however, this was a data constraint that could not be changed by extracting more data. To clarify, we note that our result estimates the probability of dying within the next eleven years as opposed to estimating the probability of dying within the 11 years given that one has not died at a particular point of time.

We incorporated one of the standard methods from epidemiology: to consider the target variable, probability of death, to be distributed as a logistic relation. This entailed that our first attempt was to use logistic regression. Thus, our motivation was to see how models can compete with the logistic relation to yield results that report probability of death. The logistic relation resembles the linear probability relation if the proportion of outcome of interest in the data set is away from 0 or 1. In our case the proportion dying in the data set was around 0.12. Thus, our basic model is logistic regression. The logistic regression and other types of models we consider predicts the death or non-death occurrence within a subsequent period of 4018 days through use of basic socio-economic data collected at date 0. It is possible the socio-economic data may have changed substantially for some people during this time.

Within the data set, younger people die infrequently within the eleven years. It is also possible for those under 40 years of age, one would see a great deal of social and economic changes within the eleven years. Further, for certain factors considerable amount of data were not collected for those under 21. In most instances we did not include the age group below 21. Thus, we predict mortality probability for those above 21.

Of course, the current exercise is an initial attempt. Perhaps an improved model predicting death can become a user-facing application. Further motivation involves, through incorporating the chosen model for the project, the initial development of an APP that predicts mortality for a given age and a socio-economic profile.

Project Pipeline

Data Ingestion, Wrangling and Exploratory Analysis

Note: for the purposes of the following discussion, the terms attributes, factors and features will be used interchangeably to describe the independent variables used to predict mortality.

Given that the 11-year data used in our project was already in .csv format—and with a detailed library defining all variables—data ingestion was fairly simple and straightforward. The un-wrangled .csv data had already been sanitized to address privacy implications and contained 1,835,072 records with 43 descriptive attributes (or variables, or features). Six of the attributes were related specifically to tobacco usage and contained no entries, because the issue of smoking-related deaths constituted a separate, more focused study of NLMS cohorts. Two other attributes—underlying cause (“cause113”) & follow-up time (“follow”) were only used in the case of death during the 11-year period; these two variables represented “leakage” and were also discarded prior to entering wrangling activities. At this point, the data was very imbalanced: 1,674,322 non-deaths, 160,750 deaths (~ 91%/9%). The attributes also reflected a mix of continuous, binary, hierarchical and categorical data.

Data wrangling efforts included the following activities: (1) all binary variables were converted to (0,1); (2) bins were created for variables with multiple values; (3) variables were dropped for features with missing data, bias and/or low correlation with the dependent variable (“inddea” or death indicator); (4) rows with missing data were dropped; (5) dummy variables for the remaining five hierarchical variables were generated. The final wrangled dataset ended up with 40 dependent variables and 934,809 records. The contribution of each remaining variable was then verified using feature selection for machine learning: univariate selection, principal component analysis and feature importance all indicated that “age” was the most important variable for the purposes of predicting mortality.

Wrangling was approached differently in order to present different types of data analysis results. For example, showing the density of death in the days following the census interviews was displayed via the Kaplan-Maier Method using un-wrangled data. For formulation of the hypotheses—mostly by visualization—descriptive analysis was carried out using wrangled data.

Prior to computation and modeling, the data set was split in 80/20 ratio for training and testing purposes.

Computation and Modeling

Insights which we gained during exploratory analysis led us to proceed with feature analysis and subsequent feature selection. Due to the class imbalance hurdle we faced with the dataset, we were forced to drop some features with high rates of missing data and keep as many underrepresented positive (death) class records as possible. In addition to the class imbalance issue, we realized that our data was very prone to bias due to including such sensitive features as race, ethnicity, income etc. Having recently read “Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy” by Cathy O’Neil, as well as consulted some members of the faculty, it was decided to drop those sensitive variables. Subsequent feature selection utilized such methods as correlation coefficients, Lasso, Ridge and Elastic Net, in combination with Kaplan-Meier models to gain deeper understanding as to which features were determinants of individual mortality. Selected features were transformed into categorized features and used in the predictive models as discussed below.

Having addressed data bias, missing variables and hierarchical features, we now had to overcome class imbalance issue, and for that purpose it was decided to attempt taking two routes:

1. Use regression and classification models while balancing the classes (`class_weight=”balanced”`).
2. Use “`imblearn.combine`” methods to combine over-sampling and under-sampling of under-represented and over-represented classes respectively.

Obtained results showed that the first method yielded high recall, accuracy and f1 scores in combination with little time spent on model training, while second route took significantly more time for data over- and under-sampling while having score results comparable to the first one.

For model evaluation purposes we focused on recall score, as well as f1 score with weighted average. Since our positive class was smaller, i.e. underrepresented class with death outcomes, and the ability to detect correctly positive samples (predicting death when death did occur) was our primary focus (correct detection of negatives examples was less important to the problem) recall proved to be a better model assessment metric.

In-depth Model Analysis and the learning curve

With various options available in model selection, choosing the right one has become ever important. To venture on this topic, we decided to venture on a popular one on Kaggle, XGBoost to establish a process for selection. Considering the size of our dataset at 1.8 million records, speed was an important criteria. While there are other implementations of gradient boosting like sklearn, we were intrigued by the recent success and popularity of XGBoost that had sizeable speed advantage over others based on benchmarks conducted.

```
Orginal Features Considered:
 ['ssnyn', 'vt', 'povpct', 'Cause of Death', 'Occupation', 'isMale', 'Age_Bins',
'Education_Bins', 'BirthCountry', 'MarriedStatus', 'HomeOwnerStatus',
'EmploymentStatus', 'PovertyPct_Bins', 'Regions', 'IsDead']
After OneHotEncoding Features:
 ['ssnyn', 'vt', 'povpct', 'isMale', 'IsDead', 'Cause of Death_Alive', 'Cause of
Death_Cerebrovascular Diseases', 'Cause of Death_Chronic Liver Disease', 'Cause of
Death_Diabetes', 'Cause of Death_Diseases of Heart', 'Cause of Death_Influenza and
Pneumonia ', 'Cause of Death_Malignant Neoplasms', 'Cause of Death_Nephritis', 'Cause
of Death_Other Causes', 'Cause of Death_Respiratory Diseases', 'Cause of
Death_Septicemia', 'Cause of Death_Suicide', 'Occupation_Disabled',
'Occupation_Homemaker-Student', 'Occupation_No Info', 'Occupation_Retired',
'Occupation_Unemployed', 'Age_Bins_18-21', 'Age_Bins_22-30', 'Age_Bins_31-45',
'Age_Bins_46-55', 'Age_Bins_56-65', 'Age_Bins_66-75', 'Age_Bins_76-85',
'Age_Bins_Above 85', 'Education_Bins_2', 'Education_Bins_3', 'Education_Bins_4',
'Education_Bins_5', 'BirthCountry_Asia', 'BirthCountry_Eastern Europe',
'BirthCountry_Islands', 'BirthCountry_Latin America', 'BirthCountry_Mexico',
'BirthCountry_Missing', 'BirthCountry_NA_Not_US_Mexico', 'BirthCountry_Outside_US',
'BirthCountry_South America', 'BirthCountry_USA', 'BirthCountry_Western Europe',
'MarriedStatus_Divorced', 'MarriedStatus_Married', 'MarriedStatus_Never Married',
'MarriedStatus_Separated', 'MarriedStatus_Widowed', 'HomeOwnerStatus_Rent',
'EmploymentStatus_Disabled', 'EmploymentStatus_Employed',
'EmploymentStatus_Retired_HomeMaker', 'EmploymentStatus_Student',
'EmploymentStatus_Unemployed', 'PovertyPct_Bins_1', 'PovertyPct_Bins_2',
'PovertyPct_Bins_3', 'Regions_Northeast', 'Regions_South', 'Regions_West']
```

We wanted to get an understanding of each of the parameters used in the model. Grid Search with 3-fold cross validation was conducted to see which parameters performed the best.

```
cv_params = {'max_depth': [3,5,7], 'min_child_weight': [1,3,5]}

ind_params = {'learning_rate': 0.1, 'n_estimators': 10, 'seed':0, 'subsample': 0.8,
'colsample_bytree': 0.8,

             'objective': 'binary:logistic'}
```

```
optimized_GBM = GridSearchCV(xgb.XGBClassifier(**ind_params),
                             cv_params,
                              scoring = 'accuracy', cv = 3, n_jobs = 2)
```

```
Grid Scores:
[mean: 0.92667, std: 0.00073, params: {'max_depth': 3, 'min_child_weight': 1},
 mean: 0.92667, std: 0.00073, params: {'max_depth': 3, 'min_child_weight': 3},
 mean: 0.92667, std: 0.00073, params: {'max_depth': 3, 'min_child_weight': 5},
 mean: 0.93036, std: 0.00020, params: {'max_depth': 5, 'min_child_weight': 1},
 mean: 0.93036, std: 0.00020, params: {'max_depth': 5, 'min_child_weight': 3},
 mean: 0.93036, std: 0.00020, params: {'max_depth': 5, 'min_child_weight': 5},
 mean: 0.93055, std: 0.00025, params: {'max_depth': 7, 'min_child_weight': 1},
 mean: 0.93053, std: 0.00024, params: {'max_depth': 7, 'min_child_weight': 3},
 mean: 0.93054, std: 0.00026, params: {'max_depth': 7, 'min_child_weight': 5}]
```

The hyperparameter combination of max_depth at 7 and min_child_weight at 1 performed the best. We also tried other combinations of hyperparameters, learning_rate and sub_sampling and obtained the following scores.

```
[mean: 0.92624, std: 0.00091, params: {'learning_rate': 0.1, 'subsample': 0.7},
 mean: 0.92625, std: 0.00091, params: {'learning_rate': 0.1, 'subsample': 0.8},
 mean: 0.92625, std: 0.00091, params: {'learning_rate': 0.1, 'subsample': 0.9},
 mean: 0.92452, std: 0.00228, params: {'learning_rate': 0.01, 'subsample': 0.7},
 mean: 0.92453, std: 0.00228, params: {'learning_rate': 0.01, 'subsample': 0.8},
 mean: 0.92453, std: 0.00228, params: {'learning_rate': 0.01, 'subsample': 0.9}]
```

Such tests at the early stage of our project made us aware of the implications of parameter tuning. The model was run with the optimized parameters:

Feature Importance:

```
[('MarriedStatus_Never Married', 291),
 ('ssnyn', 250),
 ('povpct', 199),
 ('Age_Bins_66-75', 150),
 ('Age_Bins_76-85', 140),
 ('isMale', 137),
 ('Age_Bins_56-65', 124),
 ('EmploymentStatus_Retired_HomeMaker', 119),
 ('Age_Bins_31-45', 109),
 ('Education_Bins_2', 108),
 ('EmploymentStatus_Disabled', 102),
 ('Age_Bins_46-55', 97),
 ('Age_Bins_Above 85', 88),
 ('vt', 86),
 ('Age_Bins_22-30', 84),
 ('Education_Bins_5', 77),
 ('MarriedStatus_Widowed', 75),
 ('Age_Bins_18-21', 69),
 ('MarriedStatus_Married', 62),
 ('EmploymentStatus_Employed', 59),
 ('Education_Bins_3', 59),
 ('BirthCountry_USA', 54),
 ('Education_Bins_4', 53),
 ('HomeOwnerStatus_Rent', 50),
 ('Regions_Northeast', 40),
 ('Regions_South', 30),
 ('BirthCountry_Outside_US', 29),
 ('BirthCountry_Missing', 28),
 ('MarriedStatus_Divorced', 28),
 ('Regions_West', 26),
 ('BirthCountry_Mexico', 25),
 ('BirthCountry_Latin America', 25),
 ('EmploymentStatus_Student', 22),
 ('MarriedStatus_Separated', 20),
 ('EmploymentStatus_Unemployed', 19),
 ('BirthCountry_South America', 18),
 ('BirthCountry_Asia', 17),
 ('BirthCountry_Western Europe', 15),
 ('BirthCountry_NA_Not_US_Mexico', 11),
 ('PovertyPct_Bins_2', 8),
 ('PovertyPct_Bins_3', 8),
 ('BirthCountry_Eastern Europe', 7),
 ('PovertyPct_Bins_1', 5),
 ('BirthCountry_Islands', 1)]
```

```
Accuracy:
accuracy_score(y_pred, y_test), 1-accuracy_score(y_pred, y_test)

(0.9309728485211776, 0.06902715147882243)
```

While we were thrilled with our initial accuracy scores that models predicted right out of the box, we quickly became aware of the dangers of interpreting the machine learning models incorrectly. This lead us to make a sincere attempt at understanding the inner working of the models, even though it was outside of the scope of this project. However, we did look at the options for measuring the feature importance that the XGBoost model presented: Weight – number of times a feature is used to split the data across all trees; Cover – The number of times a feature is used to slept the data across all trees weighted by the number of training data points that go through those splits; and Gain – The average training loss reduction gained when using a feature for splitting. Of these, weight was the default option.

Based on our initial research, we had identified the key points of parameter tuning for any model as:

- Control Overfitting
- Deal with imbalanced data
- Trust cross-validation

We also learned that there are model specific parameter tuning. XGBoost presented parameters related to controlling the model complexity using max_depth, min_child_weight and gamma; Robust to noise using sub_sample, colsample_bytree; ranking order using scale_pos_weight; detect continuously being worse on test set using early.stop.round; etc.

These early experiments provided us with the inputs required for steering the project. In the process, we also learned that the success of any data science related project requires in-depth knowledge from multiple resources who can contribute in their areas of specialty.

Results

Among all applied models, Logistic Regression model with balanced classes yielded the best combination of recall, f1 and accuracy scores, combined with fast model training, which needs to be taken into account when training data contains large number of instances.

Models which used over- and under- sampled data prior to applying predictive algorithms, showed very comparable performance, but took much longer to run due to the class balancing being a very time-consuming operation.

Selected logistic regression model was "pickled" and deployed in final data product – web application for calculating 11-year mortality probability.

Final Data Product

Final data product produced at the end of the project was a logistic regression model based web application, designed to calculate individual 11-year mortality probability based on user input. Application was written in Python and utilized such open source technologies as Flask web development platform and Heroku PaaS. Web application can be access at this URL:

https://mortality-predictor.herokuapp.com/

Conclusion

As it relates to our hypothesis, we found that there is sufficient evidence at the alpha level of significance to support the claim that mortality is affected by socioeconomic causal factors including, income, education and marital status. Based on our H1, we can support our original claim that mortality can be predicted based on these factors and additional factors. These additional factors including health insurance type, place of birth, presence of SSYN, etc. were captured in our feature selection process based on correlation coefficient testing and feature selection methods (e.g. Feature Importance).

While domain knowledge is critical in most data science product applications, our team was able to leverage the data's pre-built documentation and learned evaluation techniques as the primary driver for our decision-making process from ingestion to our final front-end Flask application. As it relates to ingestion, with one csv. file, pulling in the data was relatively simple. While, wrangling and munging had some missing feature information (e.g. occupation and tobacco use) which could have been useful to telling more to the story, the user, and targets, most of the data was already well structured and annotated. During computation and modeling, class imbalance presented a very real-world data science problem, however performance measures and time expenditure generated the most efficient results. Finally, through the use of scikit-learn tools such as Yellowbrick and Matplotlib visualization to diagnose and report, the information we report makes our results and finding intuitive.

With regards to our team lessons learned, we understood first and foremost that communication and team task management is critical to success. Additionally, it is important to note early on in the team forming stage where individual team members want to build on their weak points (e.g. wrangling, machine learning, programming, statistics, etc.) and how to leverage individual strengths at the same time. Moreover, using a team process agreement early on can be useful for setting goals, expectations and mitigating team issues internally.

GitHub Repository: https://github.com/georgetown-analytics/Mortality-Probability

References

A. Case, A. Fertig, and C. Paxson, (2005). The lasting impact of childhood health and circumstances, *Journal of Health Economics*. 24(2005), 365-389.

Case, A. and A. Deaton, (2017). "Mortality and morbidity in the 21st century", Brookings Pap Econ Act: 397–476.

Center for Disease Control (2008). https://www.cdc.gov/nchs/data/hus/2010/022.pdf

Center for Disease Control (2015). https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm

Cutler, D., W. Huang, A. Lleras-Muney. (2016). Economic conditions and mortality: evidence from 200 years of data. National Bureau of Economic Research Working Paper 2260. September.

Finkelstein, A., K. Baicker, H. Allen and B. Wright (2017). "The Effect of Medicaid on Medication Use Among Poor Adults: Evidence from Oregon" *Health Affairs*, 36(12)

Murray C. and L. Chen, (1992). "Understanding Mortality Change", *Population and Development Review*, 18(3): 481-503.

O'Neil , K., (2016). "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy". Broadway Books, New York.

Rmidi, A., (2018). "Build, Develop and Deploy a Machine Learning Model to predict cars price using Gradient Boosting".

Song R., H, Hall, K. Harrison, T. Sharpe, L. Lin and H. Dean. (2011). "Identifying the Impact of Social Determinants of Health on Disease Rates Using Correlation Analysis of Area-Based Summary Information", *Public Health Report,* 126 issue: 3_suppl, page(s): 70-80.

Statista (2016). https://www.statista.com/statistics/241572/death-rate-by-age-and-sex-in-the-us/

United States Census Bureau, (2015.
https://www.census.gov/did/www/nlms/publications/docs/NLMSPublic-useFileReferenceManual.pdf

Steinweg-Woods, J. (2016, June 5). A Guide to Gradient Boosted Trees with XGBoost in Python. Retrieved June 06, 2018, from https://jessesw.com/XG-Boost/