Project Report: Team 7
Ghoul Pool

Aashray Arora, Sravanthi Pereddy, Varsha Paidi, and Udit Gupta

Department of Computer Science, Stony Brook University,
Stony Brook, NY 11794-4400
{aashray.arora,paidivarsha.mohan,sravanthi.pereddy,udit.gupta}@cs.stonybrook.edu
http://www.cs.stonybrook.edu/~skiena/591/projects

1 Challenge

A ghoul pool(also called death pool) is a game of prediction which involves guessing when someone will die [1]. Sometimes it is a bet where money is involved. A typical modern dead pool might have players pick out celebrities who they think will die within the year. Our challenge is to predict/model the risks of people dying. We are given 32 celebrities presented in figure 7, who belong to different professions and countries and are at a high risk of dying because of age or lifestyle. Our goal is to predict the death (probability) or remaining life for each of the personalities. Models that make such predictions are used heavily in the life insurance industry.

2 History/Background

In this section, we discuss why we may be interested about such models in general. Two important industries which rely on modeling the risks of people dying are,

- Health Care : Such predictive models are used in health care to help prevent death and maybe catch problems early. These models are usually specific to people with a certain disease or medial problems. For example Metha et. all [14] use logistic regression to predict death in patients with acute Type A aortic dissection and Farr et.all [15] used stepwise logistic regression to predict death in patients hospitalized for community-acquired Pneumonia . In another research, a "morality index" [4] has been defined for predicting the chances of dying in next 10 years for people older than 50 based on certain health factors. The 12-item index is for use by doctors in order to help them decide whether costly health screenings or medical procedures are worth the risk for patients unlikely to live 10 more years.
- Insurance : In order to price insurance products, insurance companies must develop projections of future insured events (such as death, sickness, and disability). To do this, they develop mathematical models of the rates and

timing of the events [3]. They do this by studying the incidence of these events in the recent past, and sometimes developing expectations of how these past events will change over time (for example, whether the progressive reductions in mortality rates in the past will continue) and deriving expected rates of such events in the future, usually based on the age or other relevant characteristics of the population. These are called mortality tables if they show death rates, and morbidity tables if they show various types of sickness or disability rates.

We feel ideas and principles used in the life insurance industry could help in our challenge. It is difficult to get medical data of famous people and hence modeling health parameters may not be an option.

Actuaries mathematically evaluate the probability of events and quantify the contingent outcomes in order to minimize the impact of financial losses associated with uncertain undesirable events, like death [5].

In actuarial science and demography, a life table (also called a mortality table or actuarial table) shows, for each age, the probability that a person of that age will die before his or her next birthday ("probability of death") [2]. From this starting point, a number of inferences can be derived, including:

- The probability of surviving any particular year of age
- Remaining life expectancy for people at different ages

We now explain some of the fields observed in life tables .

- q_x: is the probability that someone aged exactly x will die before reaching age (x+1)
- p_x: is the probability that someone aged exactly x will survive to age (x+1)
  $p_x = 1 - q_x$
- l_x: is the number of people who survive to age x
  Note that this is based on a radix, or starting point, of l_0 lives, typically taken as 100,000
  $l_{x+1} = l_x \,.\, (1 - q_x) = l_x \,.\, p_x$
  $\frac{l_{x+1}}{l_x} = p_x$
- d_x: is the number of people who die aged x last birthday
  $d_x = lx - l_{x+1} = lx \,.\, (1 - p_x) = l_x \,.\, q_x$
- ${}_tp_x$: is the probabiliaty that someone aged exactly x will survive for t more years, i.e. live up to atleast age x+t years
  ${}_tp_x = \frac{l_{x+t}}{l_x}$
- ${}_{t/k}q_x$: is the probability that someone aged exactly x will survive for t more years, then die within the following k years
  ${}_{t/k}q_x = {}_t\,p_x \cdot {}_kq_{x+t} = \frac{l_{x+t} - l_{x+t+k}}{l_x}$

Two main sources of life tables for our challenge were :

- The World Health Organization [6]
- The Social Security Agency [7]

Life tables give very useful demographic based data that we directly used in modeling our challenge. We discuss more on each of the above life tables in greater detail in section 4.1 .

3 Literature Review

To understand how age affects the probability of death for a person we looked at the paper "Age likes some years -A case study for ages more prone to death" [22].

In this article the author tried to answer the following questions. Are all ages susceptible to death? If not, which are the ages that are more, or less, susceptible? Is it possible that a trend, or a probability distribution, can be observed for the ages, which can be used for prediction and forecasting for the incidence of death? The author collected 5000 celebrities of different professions (business, medicine, academics, science, and law) from different sources like Encyclopedia Britannica (1993 - 1998), Encyclopedia Americana, Macmillan Encyclopedia, Encarta Multimedia Encyclopedia, Grolier Multimedia Encyclopedia and their birth and death dates are recorded to study ages more prone to death. These celebrities are then classified into 5 classes based on their profession and 26 strata are made according to the 26 alphabets. Then a simple random sample is selected from each stratum. Five stratified random samples of 60 celebrities of five different professions are the result.

The author observed that the age at death is indifferent to profession. The number of celebrities dying in different age groups are not the same and all classes of celebrities are corroborating this fact. Thorny peaks are observed for some age groups (56-60), (66-70), (76-80) and (86-90). These peaks are observed for approximately the same age groups in all the classes. There are some safe ages, after the thorny peaks. The ages 57, 68, 77 and 86 were claimed to be highly susceptible to deaths. Many reasons may be attributed for these four highly vulnerable ages, for example, genetic underpinnings, Physiological explanations etc.

We compared the results described in the paper with our results (on our dataset). We used equal number of people in each profession (15,000). The result is shown in figure 1. We did not observe the peaks mentioned by the author. The author has a very small sample. On a large dataset like ours, such claims would not hold.

Next we consider the research paper "smoothing and projecting age-specific probabilities of death by TOPALS" [24]. TOPALS is a tool for projecting age-specific rates using linear splines. The authors used a parameter to determine how fast death probabilities move towards the target values. This parameter gives the baseline scenario. Various scenarios are discussed in the paper and the author mentioned how they are formed. Also comparison between results from TOPALS and other smoothing and projection methods have been given.

This paper gave various ideas about how to choose baseline scenarios, base periods for projecting age-specific death probabilities. As per the paper, we also tried to use different base periods, i.e., using last n years, instead of last m years, would increase or decrease the life expectancy by few years. But since this method did not fetch satisfactory results with our dataset, we discarded it. This paper helped us to understand that factors like range of base period and the study of mortality trend over past decades can affect the death age

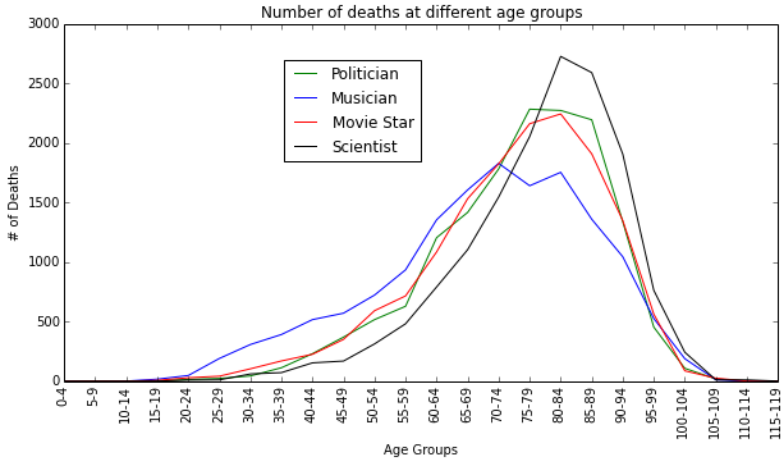

Fig. 1. Number of deaths at different age groups for different professions, to evaluate claim made in [22]. We do not observe the peaks mentioned by the author on larger dataset like ours.

prediction. Also, we used the idea that effects of health care (smoking behavior, as mentioned in paper) and changes in lifestyle can also play an important role in death age prediction.

The United States Life Tables [25] : presents complete period life tables for the United States by race, hispanic origin, and sex, based on age-specific death rates in 2008.

There are two types of life tables. A cohort life table is one which presents the mortality experience of a particular birth cohort. For example, all persons born in the year 1900. A Period life table is one which does not represent the mortality experience of an actual birth cohort but presents what would happen to a hypothetical cohort if it experienced throughout its entire life the mortality conditions of a particular time period. Life tables can also be classified in two ways according to the length of the age interval in which data are presented - Complete Life Tables which contains data for every single year of age and Abridged Life table which contains data by 5 or 10 year age intervals.

Further, it discuss the methodology used to construct these tables. The data used to prepare the U.S. life tables for 2008 was final number of deaths for the year 2008, post-censal population estimates for the year 2008, and age-specific death and population counts for medicare beneficiaries aged 66 to 99 for the year 2008 from the Centers for Medicare & Medicaid Services (CMS). Data from the Medicare program were used to supplement vital statistics and census data for ages 66 and over. The methodology used to estimate the U.S. life tables for 2008 was refined by modifying the smoothing technique used previously to estimate mortality at the oldest ages.

Then it describes about the life tables data and its format. The most frequently used life table statistic is life expectancy (e_x), which is the average number of years of life remaining for persons who have attained a given age (x) and survivors to specified ages. The next part gives an explanation of various columns present in the table, namely Age, Probability of dying, Number surviving, Number dying, Person-years lived, Total number of person-years lived and Expectation of life.

Further, the report discusses various results obtained through the tables. The most important were Life Expectancy (by Race, by Hispanic Origin and by Sex) and Survivorship in U.S.(by Race, by Hispanic Origin and by Sex). We selected period complete life tables are appropriate for our project.

In our challenge we have to predict the age at which a given celebrity will die. Since the output variable is not a class label, this is not a classification problem. The output variable - the age at which a celebrity will die - is a continuous value, hence it is clearly a regression problem. We wanted to understand how age can be used as an output variable in various machine learning regression models. Hence we reviewed "Demographic Prediction Based on User's Browsing Behavior" [29].

The authors tried to predict age and gender of a particular user based on demographic information obtained through web applications. The authors discuss three approaches. First, associating web page click-through data with gender and age. Second, predicting age and gender using demographic features of associated web pages using various machine learning algorithms. Third, using the fact that users with similar demographic information would visit similar web pages. For our project the second approach seems more appropriate. We have to predict age at which celebrity will die given his sex, gender, life expectancy, occupation, ethnicity, religion etc. Similar to author's second approach we can use these details as demographic features i.e., these are input variables used to predict age (which is the output variable).

The author used support vector machine regression for modeling. This paper gave us a good insight on preprocessing of dataset, constructing features and feeding the input into a machine learning algorithm.

4 Data Sets

In this section, we will present our research over relevant data sets appropriate for our challenge domain.

4.1 Data Sources for Life Tables

The Global Health Observatory : World Health Organization [6] - Our first data source for life tables has been taken from the WHO(World Health Organization). WHO provides life tables by country. Figure 2 shows one such table for USA, that mentions the life expectancy by different indicators in different years(for Male, Female and both) for various age groups. The indicator is part of a life

table, a set of tabulations that summarizes the mortality pattern that prevails across all age groups - children and adolescents, adults and the elderly. The table has been grouped into various indicators that may be of interest, like ${}_nM_x$: age-specific death rate between ages x and $x+n$, ${}_nq_x$: probability of dying between ages x and $x+n$ and few others. Any value of interest may be chosen based on the indicator of interest.

Details: off

Indicator	Age Group	1990			2000			
		Male	Female	Both sexes	Male	Female	Both sexes	Male
	85-89	0.153	0.107	0.120	0.150	0.11	0.122	0.129
	90-94	0.227	0.175	0.188	0.234	0.186	0.198	0.218
	95-99	0.307	0.259	0.268	0.328	0.279	0.289	0.316
	100+	0.419	0.385	0.391	0.457	0.418	0.424	0.460
nqx - probability of dying between ages x and x+n	<1	0.010	0.008	0.009	0.008	0.006	0.007	0.007
	1-4	0.002	0.002	0.002	0.002	0.001	0.001	0.001
	5-9	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	10-14	0.002	0.001	0.001	0.001	0.001	0.001	0.001
	15-19	0.006	0.002	0.004	0.005	0.002	0.003	0.004
	20-24	0.008	0.003	0.005	0.007	0.002	0.005	0.007
	25-29	0.009	0.003	0.006	0.006	0.003	0.005	0.006
	30-34	0.011	0.004	0.008	0.007	0.004	0.006	0.007
	35-39	0.014	0.006	0.010	0.010	0.006	0.008	0.008
	40-44	0.017	0.008	0.013	0.015	0.009	0.012	0.012
	45-49	0.024	0.013	0.019	0.023	0.013	0.017	0.020
	50-54	0.037	0.021	0.029	0.032	0.019	0.025	0.030
	55-59	0.059	0.034	0.046	0.048	0.03	0.039	0.043
	60-64	0.091	0.052	0.070	0.074	0.048	0.060	0.062

Fig. 2. Life Expectancy Table for the United States, provided by World Health Organization [6]

WHO also provides various informative data and statistics (in the form of fact sheets) on various risk factors. Some examples of the more important risk factors are underweight, unsafe sex, high blood pressure, tobacco and alcohol consumption, and unsafe water, sanitation and hygiene. A correct and intelligent observation over these data sets can provide a great insight over lifestyle of people of different countries.

The SSA Period Life Table [17] - The United State Social Security Administration Department also provides the period life tables for a particular year for the social security area population. The Social Security area population is comprised of (1) residents of the 50 States and the District of Columbia (adjusted for net census under count); (2) civilian residents of Puerto Rico, the Virgin Islands, Guam, American Samoa and the Northern Mariana Islands; (3) Federal civilian employees and persons in the U.S. Armed Forces abroad and their dependents; (4) non-citizens living abroad who are insured for Social Security benefits; and (5) all other U.S. citizens abroad.

A period life table is based on the mortality experience of a population during a relatively short period of time. For this table, the period life expectancy at a given age is the average remaining number of years expected prior to death for

a person at that exact age, born on January 1, using the mortality rates for that particular year over the course of his or her remaining life. The table also shows probability of dying within one year and number of survivors out of 100,000 born alive along with Life Expectancy for Male and Female grouped into exact age (ranging from 0 to 119). Figure 3 shows a part of this table as seen on the SSA website [7].

Period Life Table, 2010

	Male			Female		
Exact age	Death probability [a]	Number of lives [b]	Life expectancy	Death probability [a]	Number of lives [b]	Life expectancy
0	0.006680	100,000	76.10	0.005562	100,000	80.94
1	0.000436	99,332	75.62	0.000396	99,444	80.39
2	0.000304	99,289	74.65	0.000214	99,404	79.43
3	0.000232	99,259	73.67	0.000162	99,383	78.44
4	0.000172	99,235	72.69	0.000132	99,367	77.46
5	0.000155	99,218	71.70	0.000117	99,354	76.47
6	0.000143	99,203	70.71	0.000106	99,342	75.47
7	0.000131	99,189	69.72	0.000099	99,332	74.48
8	0.000115	99,176	68.73	0.000093	99,322	73.49
9	0.000096	99,164	67.74	0.000090	99,313	72.50
10	0.000082	99,155	66.74	0.000090	99,304	71.50
11	0.000086	99,147	65.75	0.000096	99,295	70.51
12	0.000125	99,138	64.76	0.000111	99,285	69.52
13	0.000205	99,126	63.76	0.000137	99,274	68.52
14	0.000319	99,106	62.78	0.000170	99,261	67.53
15	0.000441	99,074	61.80	0.000207	99,244	66.54
16	0.000562	99,030	60.82	0.000245	99,223	65.56
17	0.000690	98,975	59.86	0.000282	99,199	64.57
18	0.000820	98,906	58.90	0.000318	99,171	63.59
19	0.000949	98,825	57.95	0.000352	99,139	62.61
20	0.001085	98,731	57.00	0.000388	99,105	61.63
21	0.001213	98,624	56.06	0.000423	99,066	60.66
22	0.001304	98,505	55.13	0.000454	99,024	59.68
23	0.001345	98,376	54.20	0.000476	98,979	58.71
24	0.001350	98,244	53.27	0.000494	98,932	57.74
25	0.001342	98,111	52.34	0.000511	98,883	56.77

Fig. 3. Part of SSA Life table [6]

4.2 Data Sources for famous people

Notable Names Database (NNDB) [18] - The Notable Names Database (NNDB) is an online database of biographical details of over 40,000 people of note. NNDB describes itself as an intelligence aggregator of those it determines to be noteworthy, but mostly to identify connections between people. They have a google plus page full of obituaries of famous people. They have a database of dead people and details about them. The NNDB Mapper is a visual tool that allows to explore NNDB visually by graphing the connections between people.

Wikipedia [19] - is one the largest informative database available online about any topic. It is the sixth-most popular website and constitutes the Internet‘s largest and most popular general reference work. As of July 2014, the english version of wikipedia has data of about 1,445,000 persons.

DBpedia [20] - DBpedia is a project aiming to extract structured content from the information created as part of the Wikipedia project. This structured information is then made available on the World Wide Web. DBpedia allows users to query relationships and properties associated with Wikipedia resources, including links to other related data sets.

The DBpedia Ontology is a shallow, cross-domain ontology, which has been manually created based on the most commonly used infoboxes within Wikipedia. The ontology currently covers 685 classes which form a subsumption hierarchy and are described by 2,795 different properties. Based on ontology classes (person - in our case), we could get any information about the person like age, place, occupation, etc. Thus we could query DBpedia to get the data of 1,445,000 persons from wikipedia easily.

Freebase [10] - is a large collaborative knowledge base consisting of metadata composed mainly of its community members. It is an online collection of structured data harvested from many sources, including individual, user-submitted wiki contributions. Freebase aims to create a global resource which allows people (and machines) to access common information more effectively. It was developed by an American software company Metaweb and has been running publicly since March 2007. Metaweb was acquired by Google in a private sale announced July 16, 2010. Google's Knowledge Graph is powered in part by Freebase. Freebase currently has data of more than 2,200,000 people.

We analyzed all the above sources for gathering our final data, namely the Notable Names Database[8], Wikipedia, DBpedia[9] and Freebase[10]. Notable Names Database (NNDB) consists of data of more than 40,000 very famous people. But this was too small for our needs. We considered parsing wikipedia, but due to its unstructured nature, even a robust web crawler failed to get all the information we need easily. This led us to DBpedia. It exposes wikipedia as a structured database. This seemed good for our needs before we explored Freebase. Freebase is a community-curated database of well-known people, places, and things. Freebase consists of data from Wikipedia as well as some other sources, for example, biographies of people. The persons in Wikipedia is subset of the ones in Freebase. Using Freebase effectively increased our dataset size and hence we eventually used freebase for our dataset.

4.3 Freebase

Freebase [10] classifies objects using ontology. We are interested only in famous people. The persons class is useful in our case. Freebase has information of more than 2 million people. It was recently bought over by Google and we use Google APIs for Freebase to collect information from it.

Along with the APIs we use the Meta-web Query Language (MQL) for Freebase to query the Data we need. MQL is similar the Structured Query Language (SQL) to filter queries to return only the data that is necessary.

Figure 4 shows a block diagram of the information extractor we developed to gather our dataset. First we use MQL to make a query for all unique persons. We use the APIs to send this request and it returns us unique IDs for about 1000

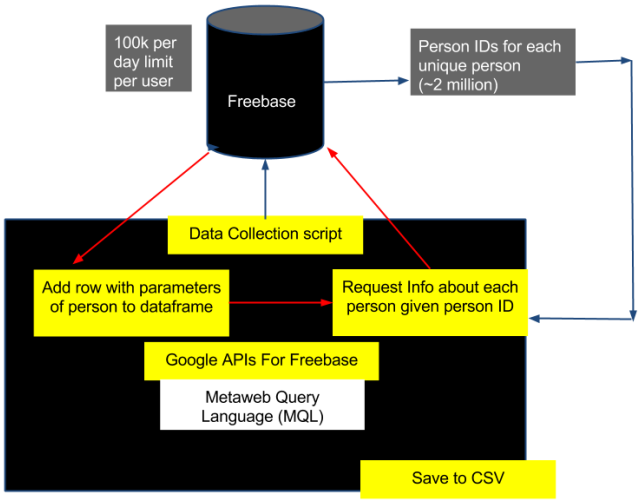

Fig. 4. Block Diagram : Information extraction from Freebase

people at a time. We use an iterator to get 1000 IDs at a time until we have all the IDs. We then send a request to get information about each person using the person's ID, one by one. We get fields like Name, Date of Birth, Date of Death (if dead), Age, Gender, Nationality, Cause of death, Profession, etc. Freebase imposes a limit of 100K queries per user per day, hence we used multiple user accounts over 2 weeks to collect information about each person.

4.4 Size of Dataset

We have a dataset of 2,274,435 rows (people) in total. We have collected many fields (columns) for each person. Overall our dataset has 63 columns. Some of them are Name, Date of Birth, Date of Death, Age, Birth Place, Nationality, Gender, Cause of death, Profession(s), Religion, Ethnicity, Spouse/Partner, Education, etc. Clearly, not all fields are relevant to our models, but since we had almost no extra cost in collecting them we have included them. This will help us in our next steps in case we find interesting ways to use them in our models. We remove the (currently) irrelevant columns in the data filtering phase.

4.5 Data Cleaning

We first did some basic cleaning on this dataset. We have deleted rows that do not contain a date of birth for a person. We have used one main profession

for people who were classified into many professions. We also performed some other minor cleaning up like removing null rows and random strings in numeric columns.

4.6 Preprocessing and Filtering

Next we did a lot of preprocessing on our dataset before it can be used on our machine learning models.

Firstly, we needed a way to specify the profession a person belongs to numerically. For this, we made a new column for each profession that we are interested in modeling and marked a 1 if the person belongs to that profession and a 0 if he/she does not.

Secondly, we needed a way to specify the nationality of a person numerically. For this we hashed each country to a unique numeric ID and created a new column for each person identifying the nationality_id he/she belongs to.

Third, we needed a way to specify the gender numerically. We used 0 to represent a male and 1 for a female.

Fourth, we decided to use the dead people in our dataset for training purposes. Hence, we filtered to include only dead people. The 2010 life tables would not depict good predictions for persons born very early in the timeline, say in 1800. Hence we further filtered to include dead people born after the year 1890.

Finally, after seeing incorrect results from models that predict the year a person would die and/or the age he/she would die. We decided to manipulate our dataset in a way that we can predict the number of remaining years in a person's life. For this we took different years into consideration, namely, 1910, 1920, 1930, 1935, 1945, 1955, 1960, 1970, 1980, 1985, 1995, 2005, 2010. For each person we calculate the remaining years from each of the aforementioned years. Thus we have the current age of the person at the year under consideration and the number of more years the person lived from his/her current age. Our code is robust enough to do this process for any other set of years, if necessary. This process effectively increases the size of our dataset by 13 times (the number of years taking into consideration).

We are quite satisfied with our data set, we have used an apt source given our project domain and the data set is large enough for us to train and predict using machine learning based models.

5 Observations

5.1 Evaluation of life tables on our data set

Our initial baseline model assigned the probability of death from life tables to each of the personalities based on their respective current age and gender. From the assigned probabilities, predictions were made as to which celebrity would have kicked the bucket first. This model used the probability of death obtained from World Health Organization (WHO) Life Table[6]. We evaluated

the WHO life tables as shown in figure 5 and realized that the values obtained were not falling in the expected probability ranges. Another point that could be misleading is that it gives a death probability of 1 for each person above the age of 100. Which is clearly not appropriate. Probably the WHO focuses on young people more than modeling the risks that come with aging.

We then moved from the WHO life tables to the Actuarial Life Table for 2010 given by the Social Security Agency [7]. These tables looked more like what an insurance company would be using. The probability of death for a 100 year old male given by them is around 0.35. Figure 6 shows the evaluation of SSA life tables on our data set.

For each person in the dataset, both living and dead, we calculated the death probability at each of the integral age until the person died or their current age. We also flagged that the person was alive for each of the integral ages and dead if the person died. Then the assigned probabilities were categorized into 17 categories. Starting from range less than 0.0002 all the way up to a range between 0.8 and 1. Subsequently, we calculated the number of alive people, number of dead people, total number of people and the ratio of number of dead people to total number of people for each of the probability categories. To calculate the number of alive people, we counted all the people who were alive for each of the probability range. Similarly, we calculated the number of dead people by summing up all those who were dead for each of the probability range. The total number of people for each of the probability range was calculated by summing the number of people alive and dead. Then the ratio of the number of dead people to total number of people was calculated. The obtained values after calculating the ratios were analyzed.

Table 5.1 shows the observed death probabilities on our data set using both SSA and WHO life tables. We had to use different probability ranges for each since the minimum and maximum probability value is different in both tables and so is the overall distribution.

We found that in SSA life tables, for higher probabilities, the ratio was within the probability ranges. But for some probability ranges there was a small deviation from the expected value. This is much better than what we observed with WHO life tables. Hence we decided to use SSA life tables.

5.2 Correlation Matrix

We plotted a correlation matrix (figure 7) on our dataset to identify the correlation between the various fields under consideration.

We observed high negative correlation between the expected remaining years from life tables ($life\_expectancy\_rem$) and current age ($curr\_age$) of a person. Similarly we observed high negative correlation between the actual remaining years ($remaining\_years$) and current age ($curr\_age$) as well. This is evident intuitively.

More importantly we observed a very high positive correlation between the expected remaining years from life tables ($life\_expectancy\_rem$) and the actual

WHO		SSA	
Probability	**Observed Probability**	**Probability**	**Observed Probability**
0.9 - 1.0	0.3442	0.4 - 0.6	0.36
0.5 - 0.9	0.2108	0.2 - 0.4	0.2406
0.3 - 0.5	0.12	0.1 - 0.2	0.1393
0.1 - 0.3	0.0614	0.06 - 0.1	0.0989
0.07 - 0.1	0.0376	0.04 - 0.06	0.0093
0.05 - 0.07	0.0259	0.02 - 0.04	0.0464
0.03 - 0.05	0.0144	0.01 - 0.02	0.0279
0.01 - 0.03	0.0082	0.006 - 0.01	0.0152
0.007 - 0.01	0.0030	0.004 - 0.006	0.0093
0.004 - 0.007	0.0023	0.002 - 0.004	0.0066
0.001 - 0.004	0.0011	0.001 - 0.002	0.0031
		0.0006 - 0.001	0.0015
		0.0004 - 0.0006	0.0011
		0.0002 - 0.0004	0.0016
		<0.0002	0.0004

Table 1. WHO vs SSA observed death probabilities on our dataset

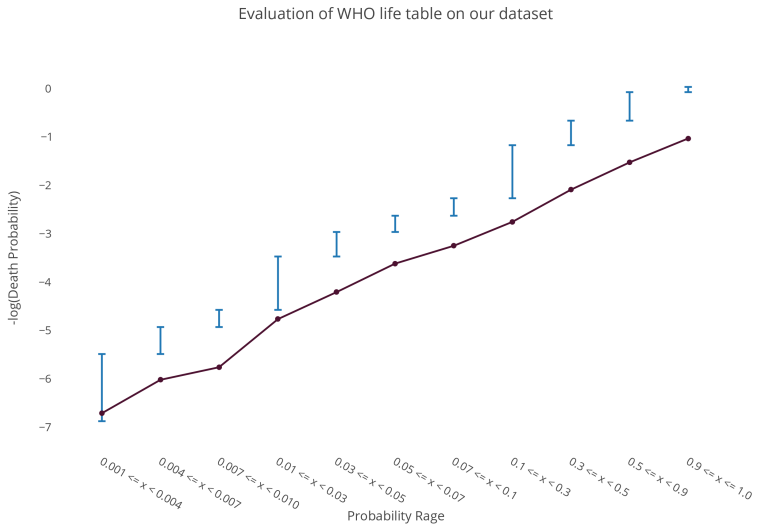

Fig. 5. Evaluation of WHO Life Tables : Range vs Observed Probability in our dataset

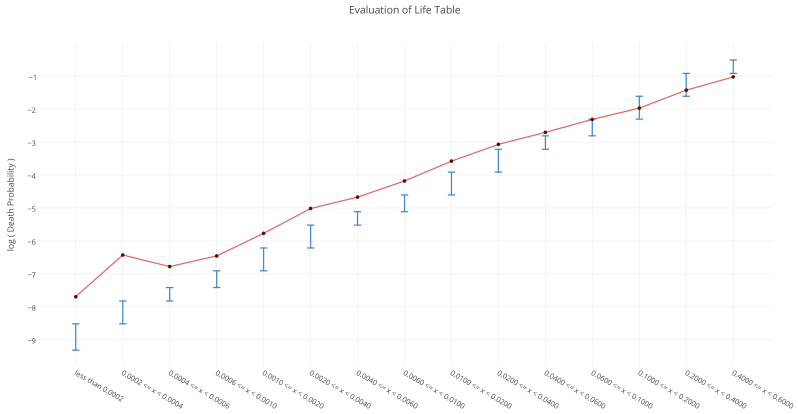

Fig. 6. Evaluation of SSA Life Tables : Range vs Observed Probability in our dataset

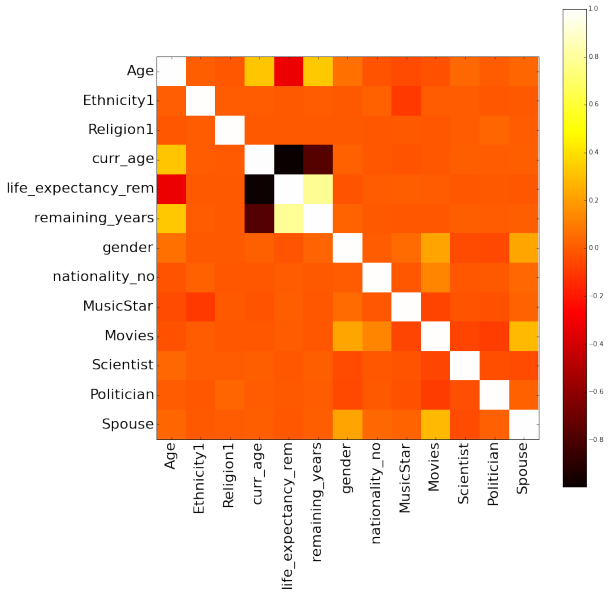

Fig. 7. Correlation between various attributes in our dataset showing high positive correlation between the expected remaining years from life tables ($life\_expectancy\_rem$) and the actual remaining years ($remaining\_years$)

remaining years (*remaining_years*). This shows that the life tables make good enough predictions and we could use it as a baseline.

5.3 Observations of death age on different professions

We plotted a box plot (figure 8) to make observations on the death age of persons belonging to various professions.

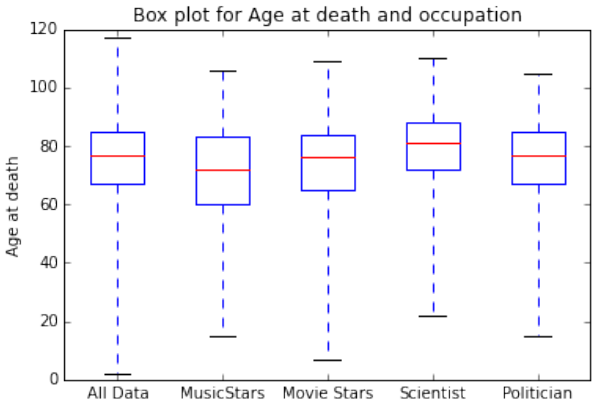

Fig. 8. Age at Death vs Occupation

On comparing the medians we observe that persons belonging to the music industry (rockstars), tend to have shorter lives than other famous persons. Scientists (including professors) tend to live longer than other famous persons.

5.4 Observations on Nationality

We plotted a data map (figure 9) to observe the life expectancy of the people in our dataset across different countries.

We observe that people of some African countries live shorter lives than others. We observe that people from the more developed countries like USA, Canada, U.K, etc. tend to live longer than others.

6 Baseline Model

We have used SSA life tables as the data source for our baseline model (figure 10). Our baseline model is a simple model which predicts the remaining life of a personality by a mere observation into the SSA life expectancy table. To predict the remaining life for a personality, we get the life table entry (expected remaining life) corresponding to that gender and age.

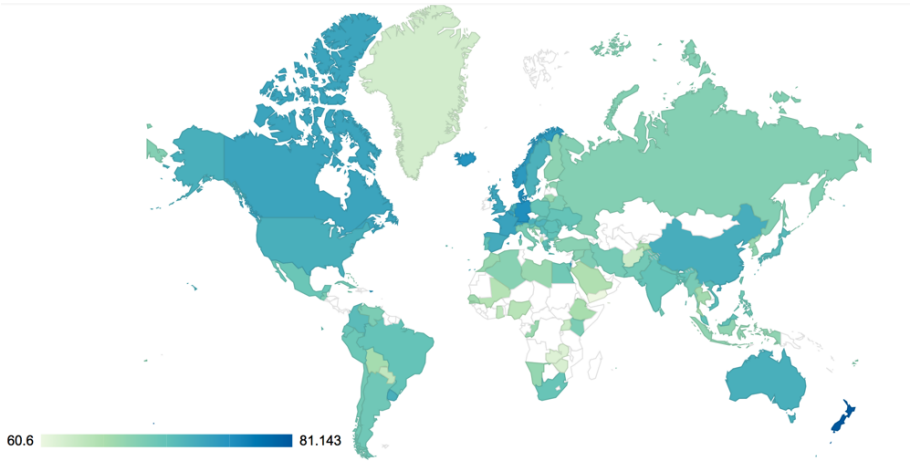

Fig. 9. Life Expectancy as per Nationality

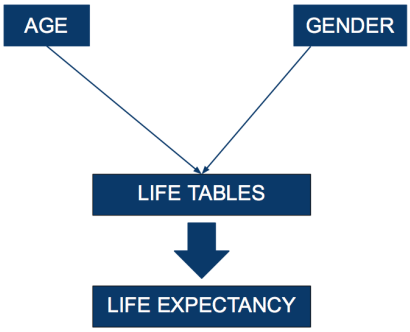

Fig. 10. Outline for baseline model

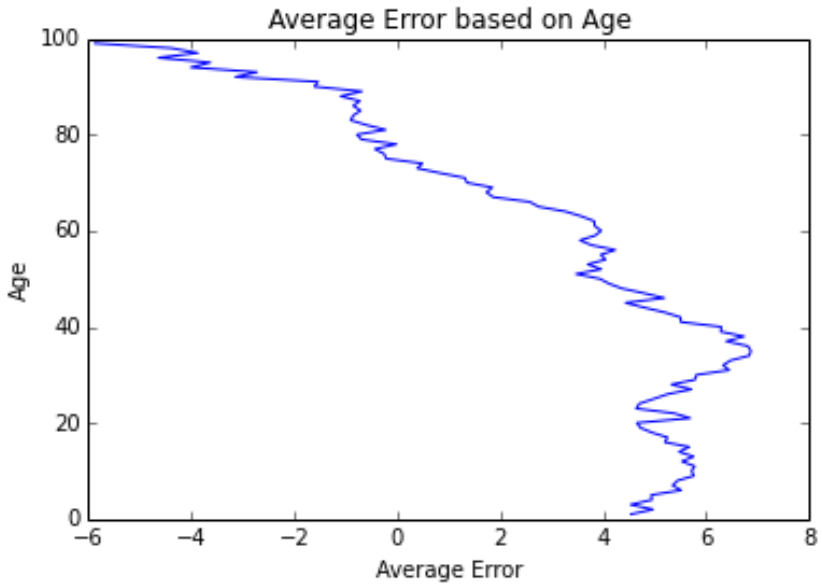

Fig. 11. Baseline Model Average Error **(in years)** for Age

We have validated our baseline model by plotting the scatter plot and error plot. We also generate evaluation statistics that help us understand it's performance. This simple model that uses life tables does a good enough job for a baseline model. However, it does not take into account the fact that we are dealing with old famous people and our advanced machine learning models exploit this fact to improve our predictions.

7 Advanced Models

All the machine learning methods were implemented using either the Scikit-learn python library [12] or the statsmodel python library [13].

7.1 Linear Regression

We used Linear Regression (also called Ordinary Least Squares, a type of supervised learning method) on our dataset. Linear Regression fits a linear model with coefficients w = (w_1, ..., w_p) to minimize the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation.

The coefficients generated by linear regression for each parameter are shown in table 2.

Parameter	Coefficient
Current Age	0.38
Expected remaining life (life tables)	1.28
Gender	2.56
Nationality	-0.0016
Music/Rock Star	-4.89
Movies	-1.94
Scientist	3.64
Politician	0.31
Spouse	1.28

Table 2. Linear regression coefficients of each parameter

7.2 Ridge Regression

We used Ridge Regression on our dataset. It addresses some of the problems of Ordinary Least Squares by imposing a penalty on the size of coefficients. Here, $\alpha \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of α, the greater the amount of shrinkage and thus the coefficients become more robust to collinearity. We have selected an α value of 1.

The coefficients generated by ridge regression for each parameter are shown in table 3.

Parameter	Coefficient
Current Age	0.008
Expected remaining life (life tables)	0.883
Gender	2.623
Nationality	-0.0007
Music/Rock Star	-4.78
Movies	-1.94
Scientist	3.70
Politician	0.28
Spouse	1.27

Table 3. Ridge regression coefficients of each parameter

7.3 K Nearest Neighbors (KNN)

Neighbors-based regression can be used in cases where the data labels are continuous rather than discrete variables. The label assigned to a query point is computed based on the mean of the labels of its nearest neighbors. The basic Nearest Neighbors Regression uses uniform weights: that is, each point in the

local neighborhood contributes uniformly to the classification of a query point. Under some circumstances, it can be advantageous to weight points such that nearby points contribute more to the regression than faraway points. This can be accomplished through the weights keyword. The default value, weights = 'uniform', assigns equal weights to all points. weights = 'distance' assigns weights proportional to the inverse of the distance from the query point. Alternatively, a user-defined function of the distance can be supplied, which will be used to compute the weights.

We have used k=100 to be the optimal value. During our testing as shown in figure 12 we observed the the performance stays almost constant after k=100. We tested till k=305 and did not find performance deteriorating yet.

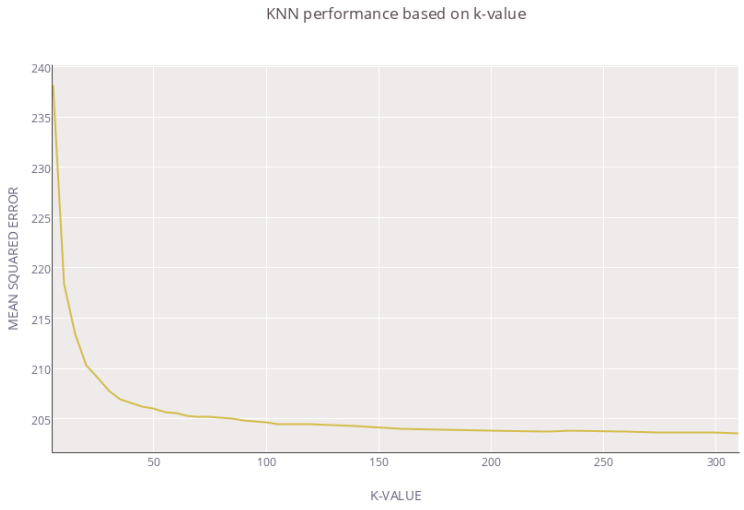

Fig. 12. Optimal k-value experiment : performance stays almost constant after k=100

7.4 Decision Trees

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. It works well for non-linear relationship between features and implicitly performs feature selection.

The main disadvantage with decision trees is they overfit the training data if the growth of the tree is not limited. Hence we did an experiment to observe the optimal value for the depth of a decision tree. Since it running on a very large dataset (it hangs the python kernel as values get very high) we compute the k

values until we start to observe a deterioration in performance. We observed that after depth = 8, the performance deteriorates and hence we choose the depth to be 8.

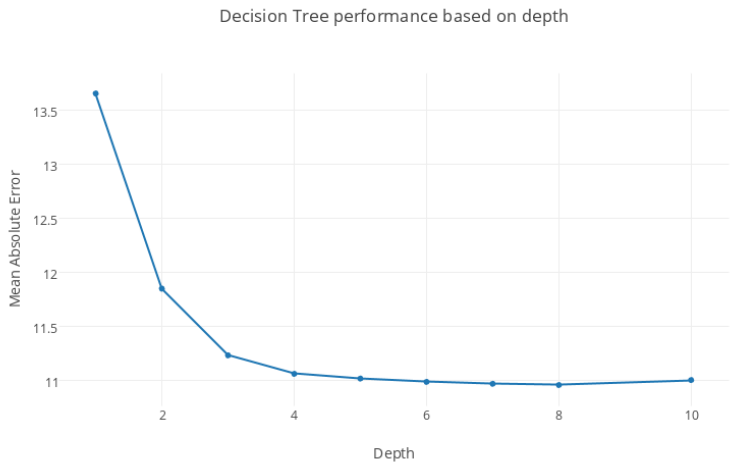

Fig. 13. Optimal depth for decision tree experiment : performance deteriorates after depth = 8

7.5 Random Forest

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

7.6 Summary of performance of advance models

Table 4 shows the error plots of all models, table 5 shows the scatter plots of all models and table 6 summarizes the performance statistics of each of these models. Ridge Regression performed the best (and liner regression is almost as good).

It is clear to us that overall our models are better than the baseline (if at all) by a very small margin. Our models achieved a maximum of 12.31% better than the baseline model for ridge and linear regression.

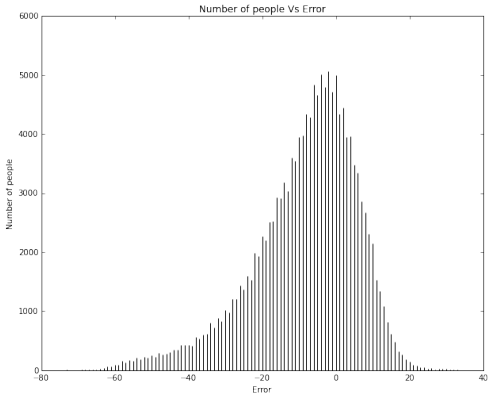

Fig. 14. Baseline

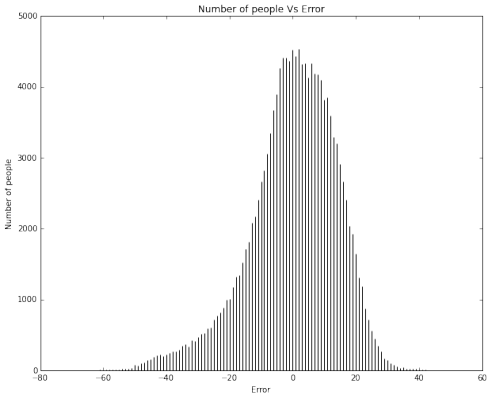

Fig. 15. Linear Regression

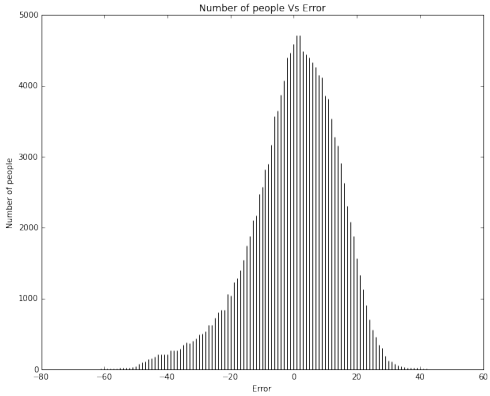

Fig. 16. Ridge Regression

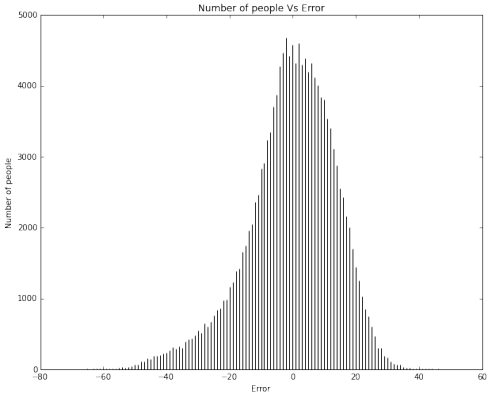

Fig. 17. K-Nearest Neighbor

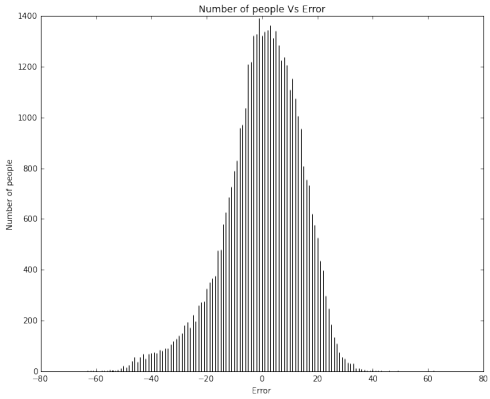

Fig. 18. Decision Trees

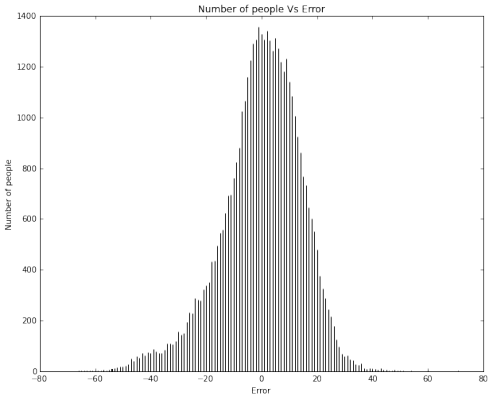

Fig. 19. Random Forest

Table 4. Error plots of baseline & advance models

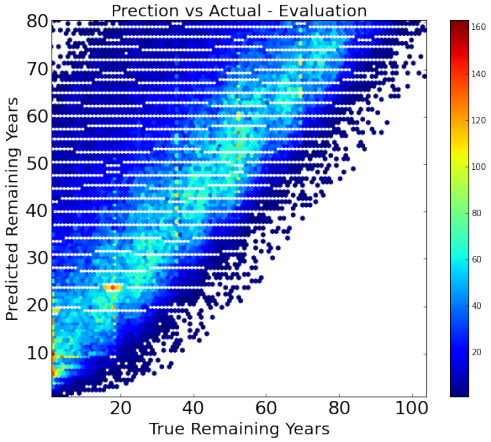

Fig. 20. Baseline

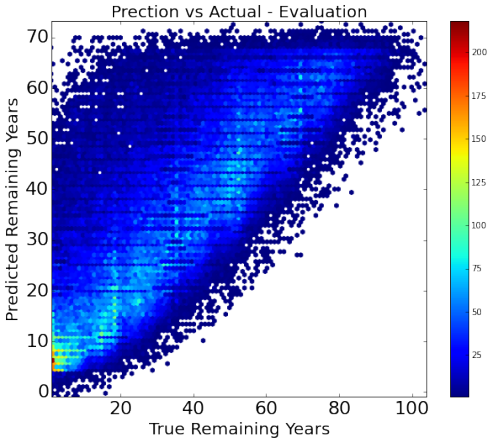

Fig. 21. Linear Regression

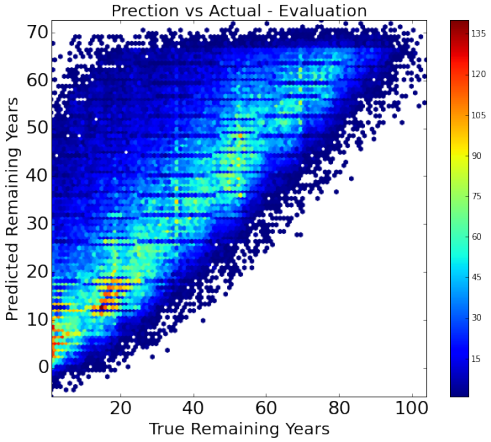

Fig. 22. Ridge Regression

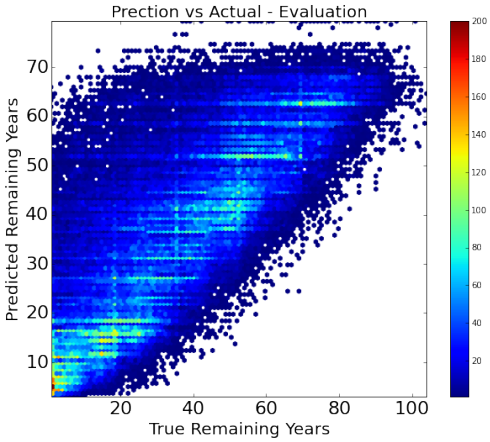

Fig. 23. K-Nearest Neighbor

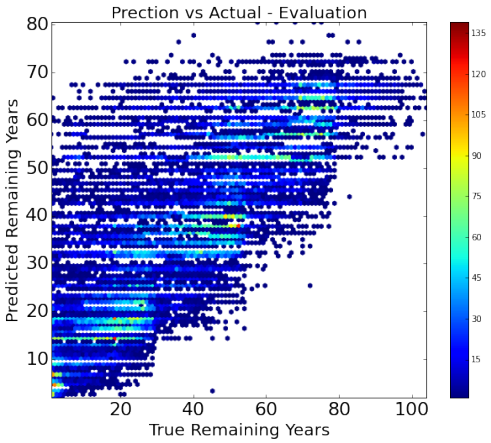

Fig. 24. Decision Trees

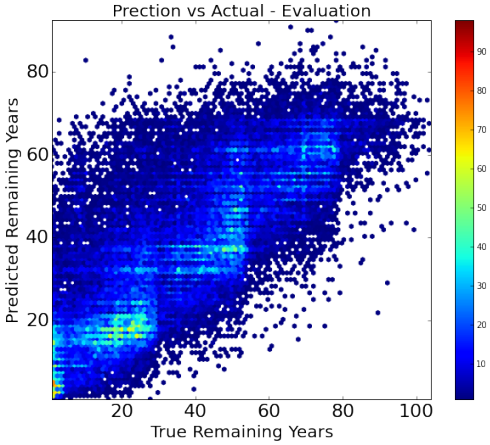

Fig. 25. Random Forest

Table 5. Scatter plots of baseline & advance models

Model	Mean Absolute Error(MAE)	Mean Squared Error	R^2 Error	% better than baseline in MAE
Baseline (life tables)	12.36	288.38	0.45	NA
Linear Regression	10.79	193.09	0.62	12.31%
Ridge Regression	10.79	193.42	0.62	12.31%
K-NN (k=100)	10.84	196.80	0.61	11.86%
Decision Trees (depth=8)	10.80	193.47	0.62	12.20%
Random Forest	11.09	206.36	0.59	9.40%

Table 6. Performance of baseline and advance models

8 Final Prediction and Conclusions

Our development environment makes the final predictions on each of the models. It also points out the best one in terms of Mean Absolute Error and gives the final predictions based on that model.

The current predictions as per the ridge regression model are shown in table 7.

These predictions do look good to us for a few reasons. Firstly, we observe that Rock/Music star and Movie stars are penalized by the model, predicting they would live a little lesser than others. Secondly, Scientists are observed to live a little longer than others as observed with Stephen Hawking.

We now discuss few difficulties that we faced during the course of our challenge and how we overcame them.

- Initially we were trying to predict the total number years a person would live. Our models were not up the mark for this and we observed incorrect results. We decide to model the challenge better by predicting the remaining life of a person rather than the total number of years he/she will live.
- To improve our models and have lesser error we added new fields to the model, for example we included the spouse field. We also tried our model with religion and ethnicity of persons (but we did not have enough data for these columns to cause significant improvements).
- Evaluation of life tables - we have discussed this in section 5.1. Calculating the integral ages for each person and modeling the probability of death at each of those ages took significant computation power and time.
- We split our dataset at various integral years as described in section 4.6. We observed that splitting at more number of years gives us a better prediction and removes the stair-case effect from our scatter plots.
- Fetching our dataset by scrapping Freebase was a time consuming process and involved some amount of sweat equity.

Acknowledgments. We sincerely thank Professor Skiena for meeting with us every single time we requested him. We had many meetings with Professor Skiena and his guidance at every step along with the valuable lectures ensured good progress for us.

Name	Age	Predicted remaining years	Expected remaining life (life tables)	Occupation
Olivia de Havilland	98	0.365081	2.76	Movies
Zsa Zsa Gabor	97	0.517778	2.94	Movies
Abe Vigoda	93	0.583961	3.20	Movies
Dick van Dyke	88	0.746971	4.64	Movies
Rev. Billy Graham	96	0.792318	2.62	Other
Helmut Schmidt	95	1.241866	2.79	Politician
Kirk Douglas	97	1.299100	2.47	Movies
John Glenn	93	1.355597	3.20	Other
Fats Domino	86	1.439223	5.38	MusicStar
Charles Manson	79	1.446387	8.63	MusicStar
Jerry Lee Lewis	79	1.446387	8.63	MusicStar
Pope Benedict	87	1.738238	5.00	Other
Henry Kissinger	91	1.823292	3.70	Politician
Shimon Peres	91	1.974032	3.70	Politician
Robert Mugabe	90	1.984550	3.99	Politician
Lee Kwan Yew	91	2.046084	3.70	Politician
Chuck Berry	88	2.107461	4.64	MusicStar
Clint Eastwood	84	2.158007	6.21	Movies
Jimmy Carter	90	2.375685	3.99	Politician
George Bush	90	2.375685	3.99	Politician
B.B. King	89	2.416726	4.30	MusicStar
Fidel Castro	88	2.813252	4.64	Politician
Doris Day	90	3.447550	4.80	Movies
Hosni Mubarak	86	3.553541	5.38	Politician
Rupert Murdoch	83	4.486244	6.65	Other
Mick Jagger	71	5.611970	13.40	MusicStar
Keith Richards	70	6.194624	14.07	MusicStar
Queen Elizabeth II	88	6.347887	5.56	Politician
Ali Khamenei	75	8.500730	10.89	Politician
Dick Cheney	73	9.580055	12.12	Politician
Stephen Hawking	72	13.532216	12.75	Scientist
Bashar al-Assad	49	24.966958	30.31	Politician

Table 7. Final Predictions based on best model : Ridge Regression

References

1. Wikipedia Entry on Death Pool, http://en.wikipedia.org/wiki/Dead_pool
2. Wikipedia Entry on Life Tables, http://en.wikipedia.org/wiki/Life_table
3. Wikipedia Entry on Actuary, http://en.wikipedia.org/wiki/Actuary
4. Predicting When You'll Die - weather.com, The Weather Channel, http://www.weather.com/health/predicting-your-death-20130306 (Accessed: October 21, 2014)
5. Actuarial science, wikipedia.com, https://en.wikipedia.org/wiki/Actuarial_science/ (Accessed:October 21, 2014)
6. World Heath Organization life tables, http://www.who.int/gho/mortality_burden_disease/life_tables/life_tables/en/
7. Social Security Agency life tables, http://www.ssa.gov/oact/STATS/table4c6.html
8. Notable Names Database, http://www.nndb.com/
9. DBPedia, http://dbpedia.org/
10. Freebase, https://www.freebase.com/
11. Distinguished Teaching Professor Skiena, Computer Science, Stony Brook University. http://www3.cs.stonybrook.edu/~skiena/
12. Scikit-learn: Machine Learning in Python, http://scikit-learn.org/stable/index.html
13. Statsmodels : a Python module that allows users to explore data, estimate statistical models, and perform statistical tests, http://statsmodels.sourceforge.net/
14. Mehta, Rajendra H., Toru Suzuki, Peter G. Hagan, Eduardo Bossone, Dan Gilon, Alfredo Llovet, Luis C. Maroto et al. "Predicting death in patients with acute type A aortic dissection." Circulation 105, no. 2 (2002): 200-206.
15. Farr, Barry M., Andrew J. Sloman, and Michael J. Fisch. "Predicting death in patients hospitalized for community-acquired pneumonia." Annals of internal medicine 115.6 (1991): 428-436.
16. Mortality and global health estimates, WHO, http://www.who.int/gho/mortality_burden_disease/en/ (Accessed: October 21, 2014)
17. social security, Actuarial Life Table, http://www.ssa.gov/oact/STATS/table4c6.html (Accessed:October 21, 2014)
18. NNDB: Tracking the entire world, NNDB: Tracking the entire world,http://www.nndb.com/ (Accessed: October 21, 2014)
19. Wikimedia Foundation, Wikipedia, Wikipedia, https://en.wikipedia.org/wiki/Wikipedia (Accessed:October 21, 2014)
20. wiki.dbpedia.org : About, wiki.dbpedia.org : About, http://dbpedia.org/About (Accessed:October 21, 2014)
21. DBpedia: Distributed Extraction, nileshc.com, http://nileshc.com/blog/2014/06/dbpedia_distributed_extraction/ (Accessed:October 21, 2014)
22. Siddiqi, Ahmed. "Age Likes Some Years." Scientometrics Vol. 69.2 (2006): 315-21. Akadmiai Kiad. Web
23. VIERCK, E., K. HODGES, Aging : Demographics, Health, and Health Services. Westport, Conn., Greenwood Press. (2003)
24. De Beer, Joop. "Smoothing and Projecting Age-specific Probabilities of Death by TOPALS." Demographic Research Vol. 27.20 : 543-92. Max Planck Institute for Demographic Research. Web. (2012)
25. Arias, Elizabeth. "United States Life Tables, 2008." National Vital Statistics Reports Vol. 61.3 Web. (2012)

26. Campbell, John, C. Diep, J. Reinken, and L. McCOSH. "Factors Predicting Mortality in a Total Population of the Elderly." Journal of Epidemiology and Community Health : 337-42. Web. (1985)
27. De La Croix, David, and Omar Licandro. "The Longevity of Famous People from Hammurabi to Einstein." Barcelona GSE Working Paper Series. Web. (2012)
28. Louis DZ, Robeson M, McAna J, et al. Predicting risk of hospitalisation or death: a retrospective population-based analysis. (2014)
29. Hu, Jian, et al. "Demographic prediction based on user's browsing behavior." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.