What is the Ultimate Price of Fame?

Does being famous shorten your lifespan and what factors may contribute?

William Long, April 10, 2018

Introduction

Being a celebrity comes with a lot of apparent perks. Fame. Money. Status.

But do the benefits of being a celebrity outweigh the ultimate cost - an early death?

Does a candle that burns brighter tend to burn out faster?

In this presentation I'll take a look at famous people and see if being a celebrity shortens your lifespan.

Step One:
Acquiring the Data

Data Acquisition

I used a dataset for Celebrity Deaths available on Kaggle.

This dataset was scraped from Wikipedia and contained information about every dead celebrity from 2006 to 2016.

The original dataset contained 21,458 total observations.

This dataset included features such as:

- Reason for fame
- Age at death
- Cause of death
- Nationality
- Fame score (# of Wikipedia citations)
- Birth year
- Death year

.

Step Two:
Preprocessing the Data

Preprocessing

The original dataset contained 21,458 total observations.

However, data was missing for some of the features that I wanted to use in order to make predictions.

Missing data for important features

Fame Score: 1,606 observations that had no values

Cause of Death: 12,484 observations that had no values

Since these were features that were going to be critical in my ability to predict overall lifespan for celebrities, I had to drop these rows since values could not be imputed for them.

This left me with 8,408 fully populated observations to work with.

Step Three:
Cleaning the Data

Cleaning

Of the dataset I had left after preprocessing, some of the features I wanted to use had a lot of extraneous information that needed to be stripped out to have a more standard format.

Data that needed to be cleaned

Famous For: baseball player (Oakland Athletics)

Cause of Death: cerebral hemorrhage [153]

Nationality: Hungarian-born

All of these features contained extra unnecessary information that didn't seem relevant and would hinder my ability to standardize the values that they contained.

Using .strip and PorterStemmer I was able to get these text features down to their common core values.

Step Four:
Feature Engineering

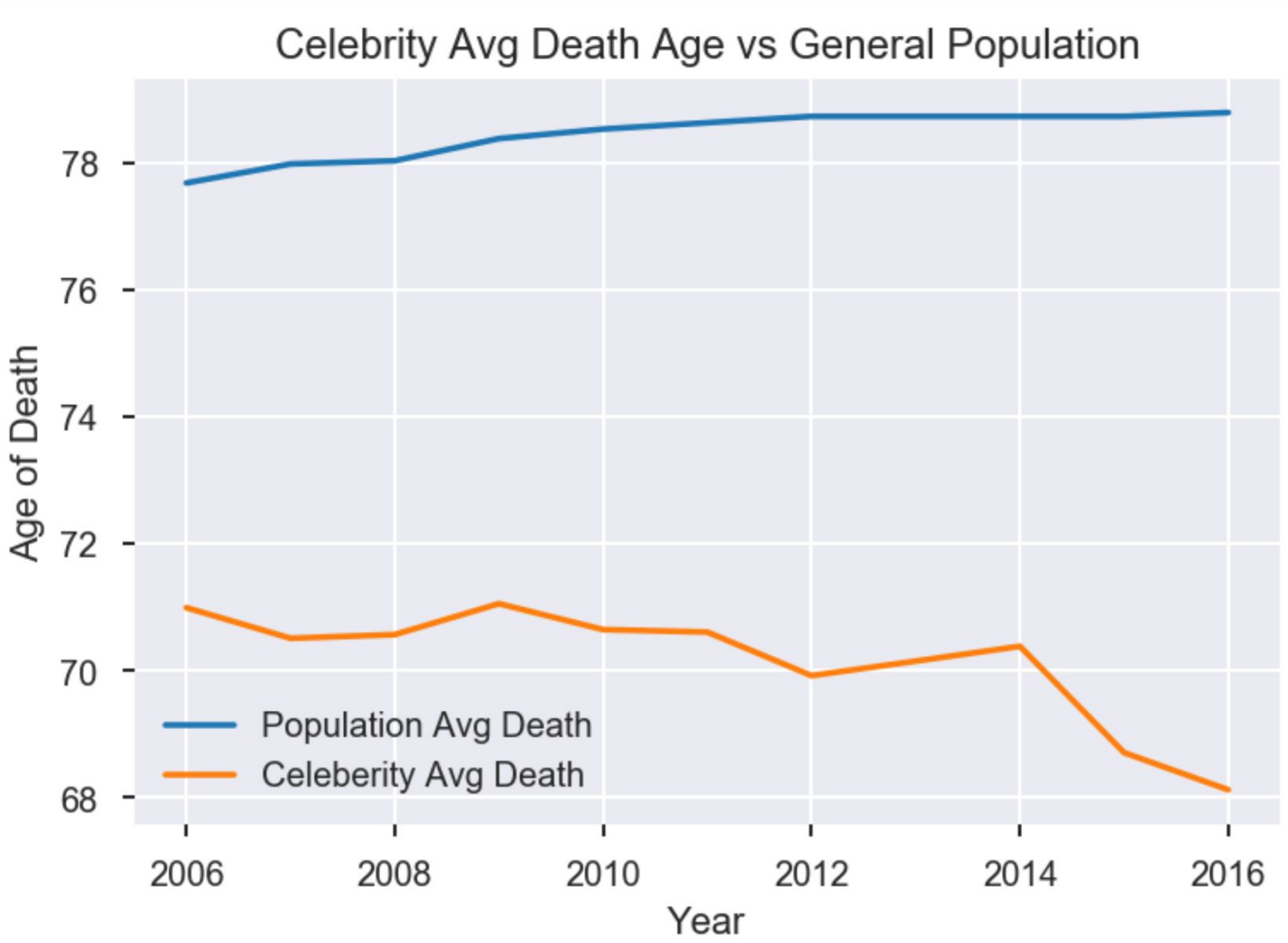
Feature Engineering

In order to try and predict whether celebrities died younger than average, I needed to bring in additional data from the WHO Mortality Database to determine the average age of death.

WHO mortality averages

By bringing in data from WHO, I established a baseline average death age for the general population of 78 years old.

This baseline age became my border to determine whether a celebrity died young or not. I created a new binary feature for died_young that I would base my predictions on.

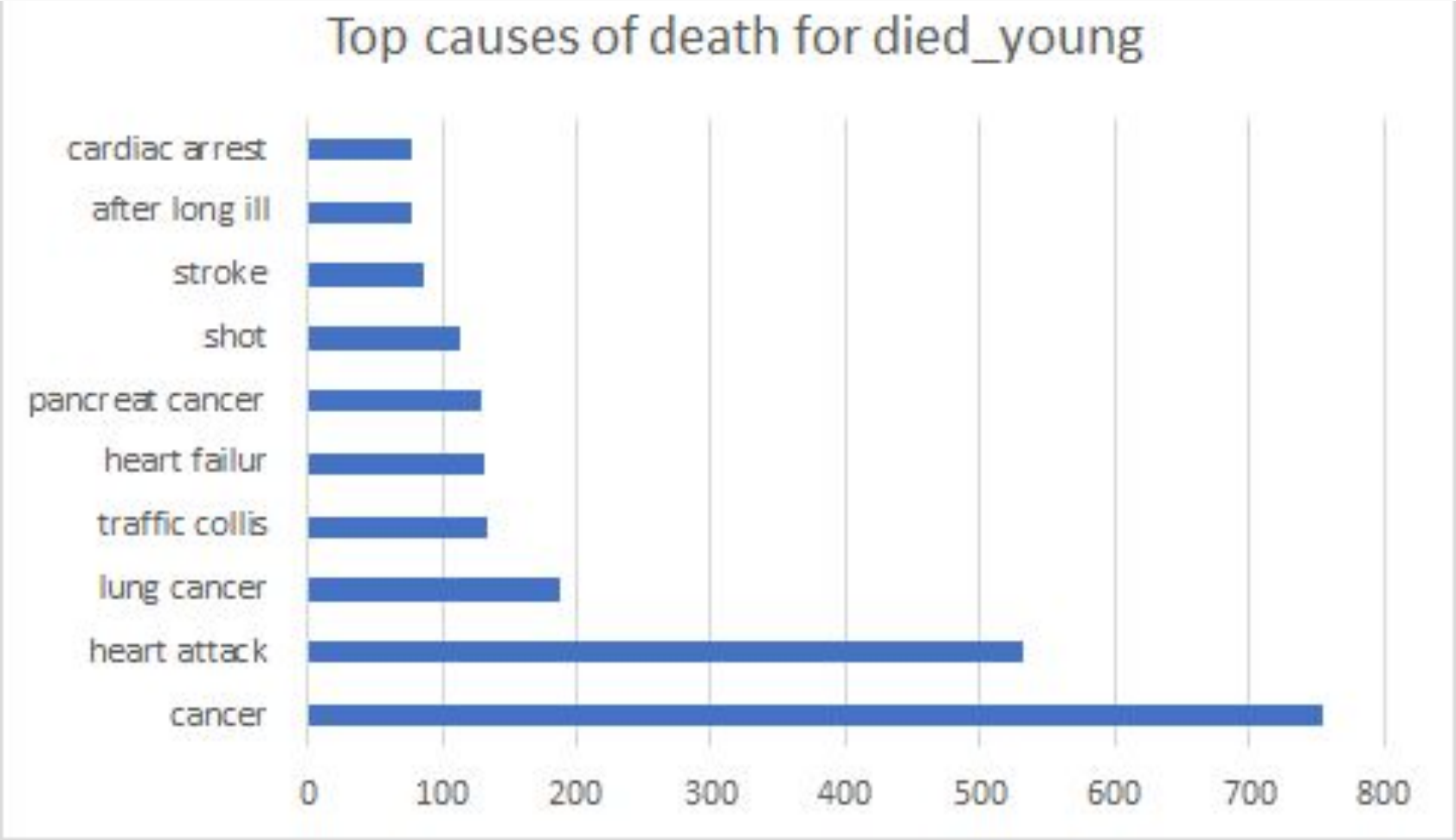

Celebrity Avg Death Age vs General Population
Age of Death
78
76
74
72
70
68
Population Avg Death
Celeberity Avg Death
2006
2008
2010
2012
2014
2016
Year

Feature Engineering

In order to try and predict whether celebrities died younger than average, I looked at the most frequent causes of death and created additional features for them.

Top causes of death for died_young

cause_of_death	count
cancer	754
heart attack	531
lung cancer	187
traffic collis	133
heart failur	132
pancreat cancer	129
shot	113
stroke	87
after long ill	78
cardiac arrest	77

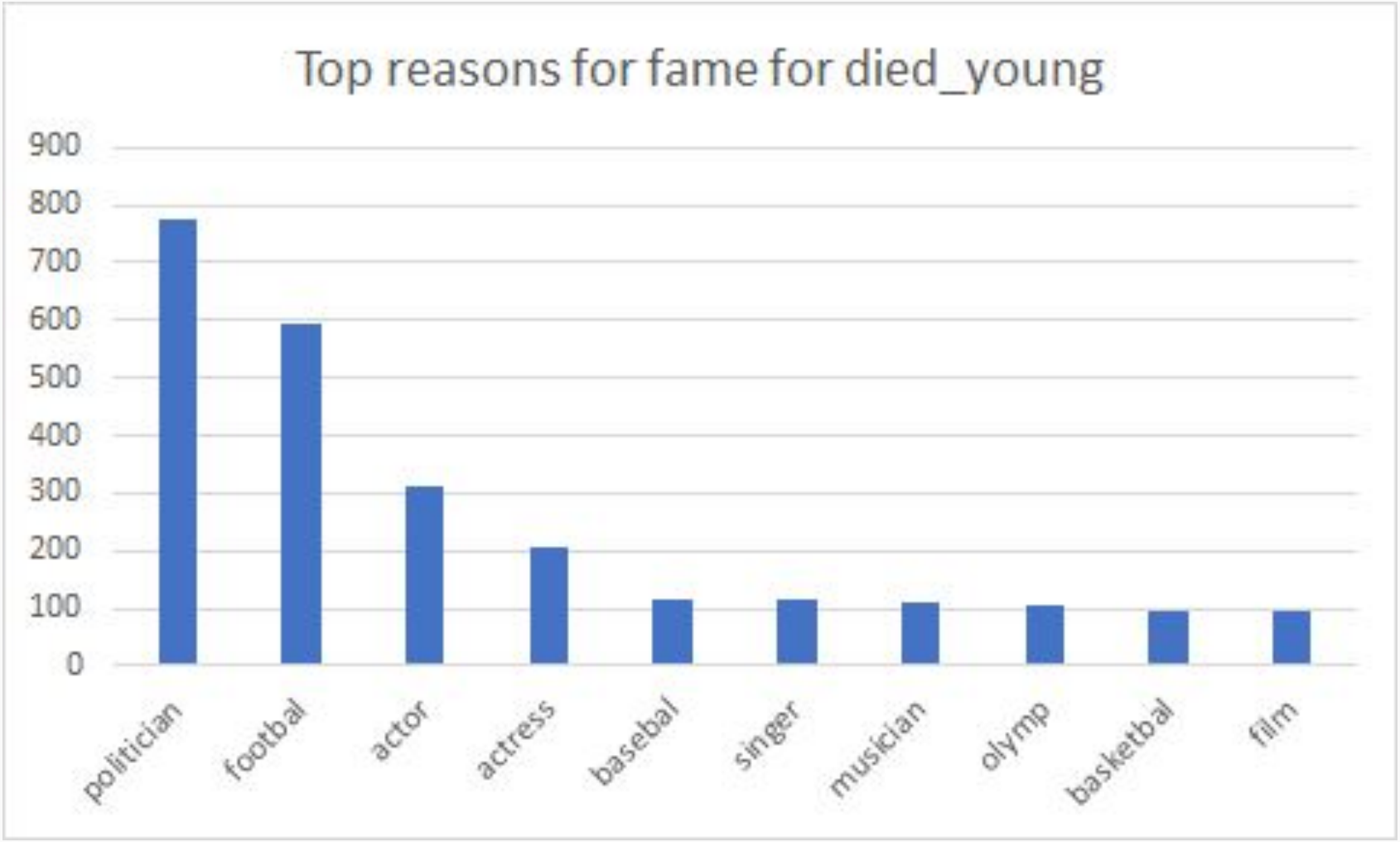

Top causes of death for died_young
cardiac arrest
after long ill
stroke
shot
pancreat cancer
heart failur
traffic collis
lung cancer
heart attack
cancer
0
100
200
300
400
500
600
700
800

Feature Engineering

In order to try and predict whether celebrities died younger than average, I looked at the reason for fame the celebrity had and created additional features for them.

Top reasons for fame for died_young

famous_for	count
politician	777
footbal	592
actor	314
actress	205
basebal	117
singer	113
musician	112
olymp	104
basketbal	96
film	96


Top reasons for fame for died_young
900
800
700
600
500
400
300
200
100
0
politician
footbal
actor
actress
basebal
singer
musician
olymp
basketbal
film

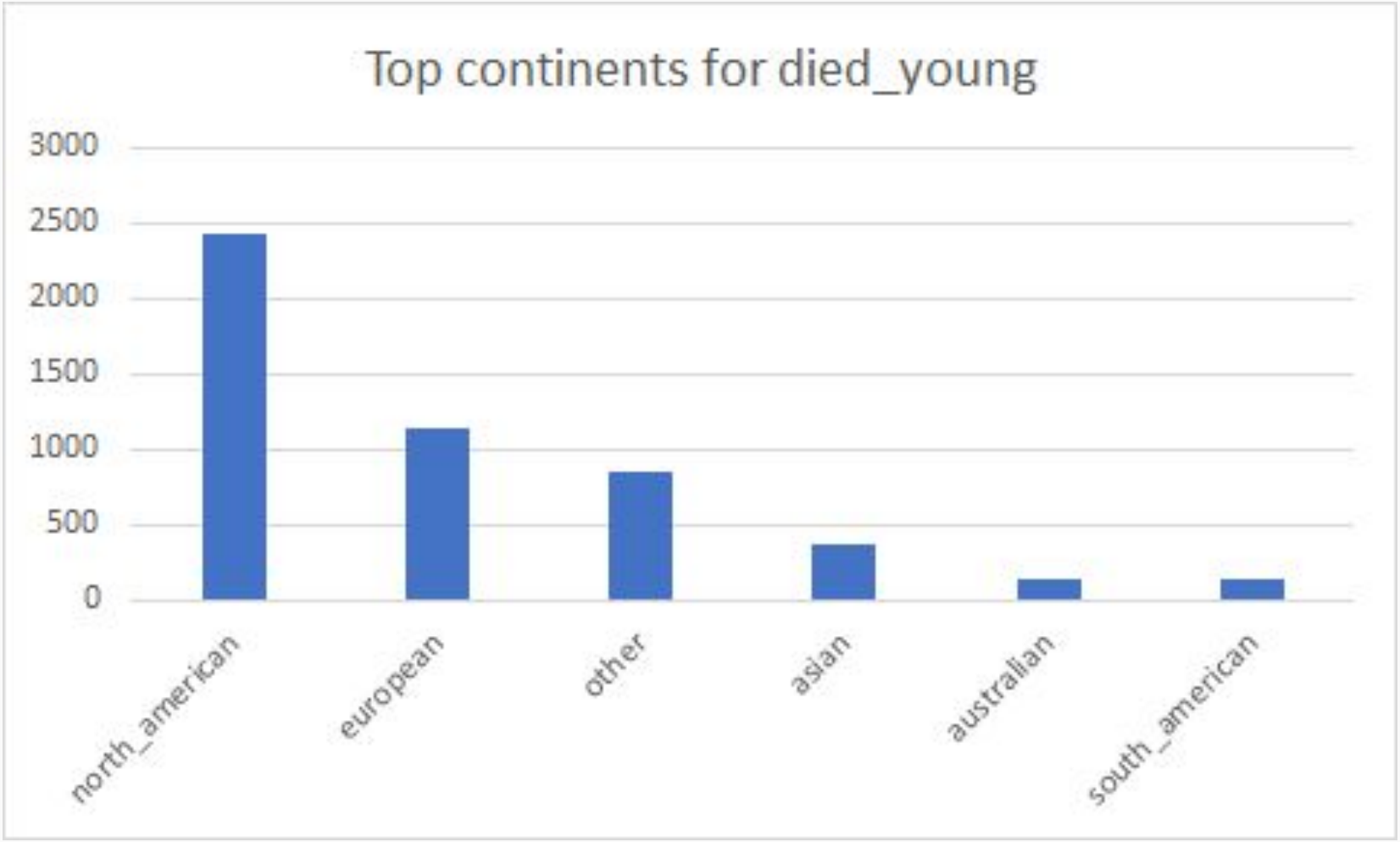
Feature Engineering

In order to try and predict whether celebrities died younger than average, I tried combining individual nationalities into a more standardized category of continent.

Top nationalities for died_young

nationality	count
american	2146.0
british	372.0
canadian	217.0
australian	136.0
english	111.0
indian	109.0
japanes	97.0
russian	91.0
brazilian	85.0
german	79.0

continent	count
north_american	2432
european	1127
other	846
asian	376
australian	136
south_american	128


Top continents for died_young
3000
2500
2000
1500
1000
500
0
north_american
european
other
asian
australian
south_american

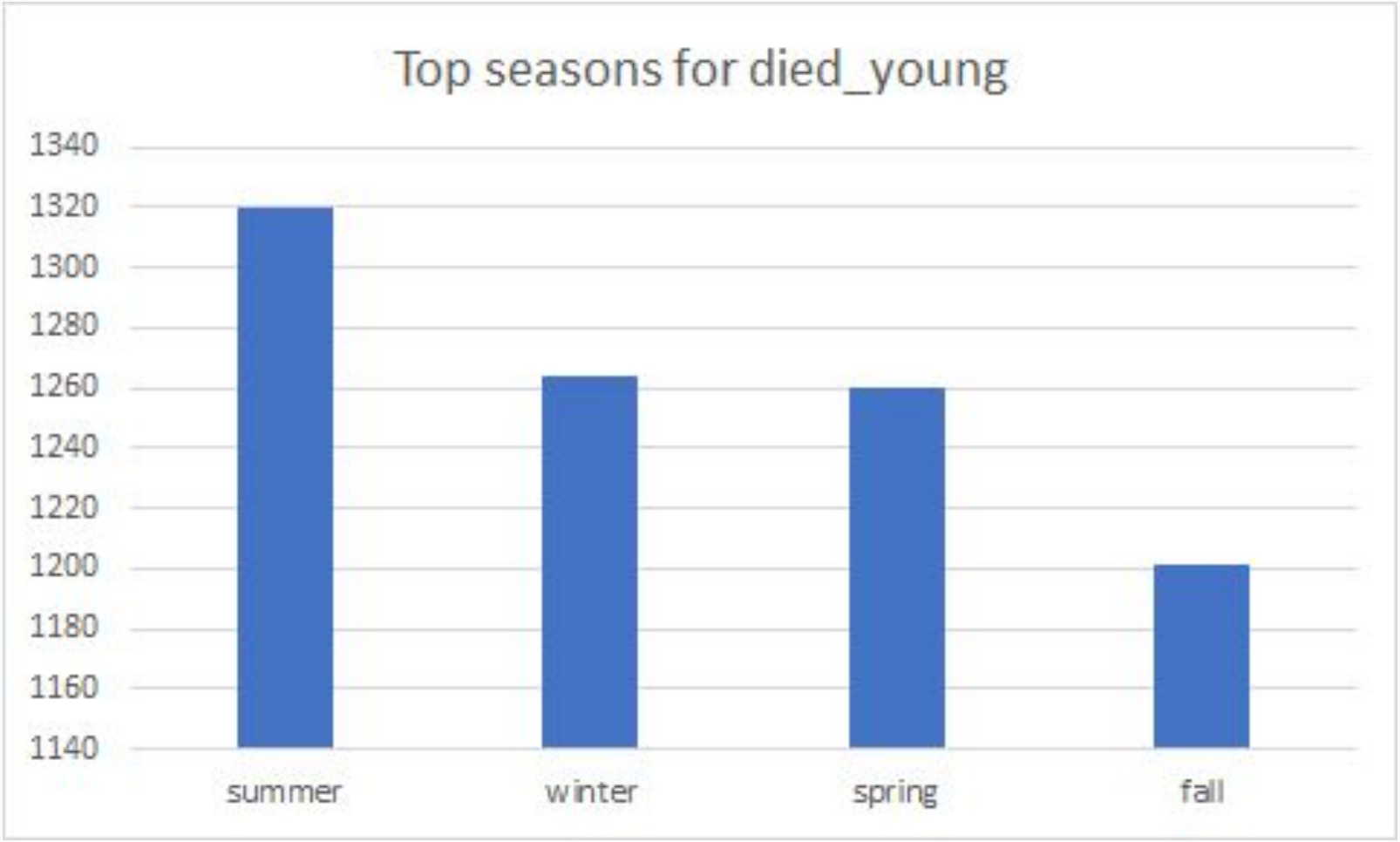
Feature Engineering

In order to try and predict whether celebrities died younger than average, I also compiled months of death into a separate season feature.

Top seasons for died_young

death_month	count
January	757
July	752
December	740
June	718
April	714
March	714
October	704
November	680
May	667
August	663
February	648
September	618

season	count
summer	1320
winter	1264
spring	1260
fall	1201


Top seasons for died_young
1340
1320
1300
1280
1260
1240
1220
1200
1180
1160
1140
summer
winter
spring
fall

Step Five: Modeling

Modeling

First I had to establish a baseline accuracy for died_young.

The WHO average age of death was 78.

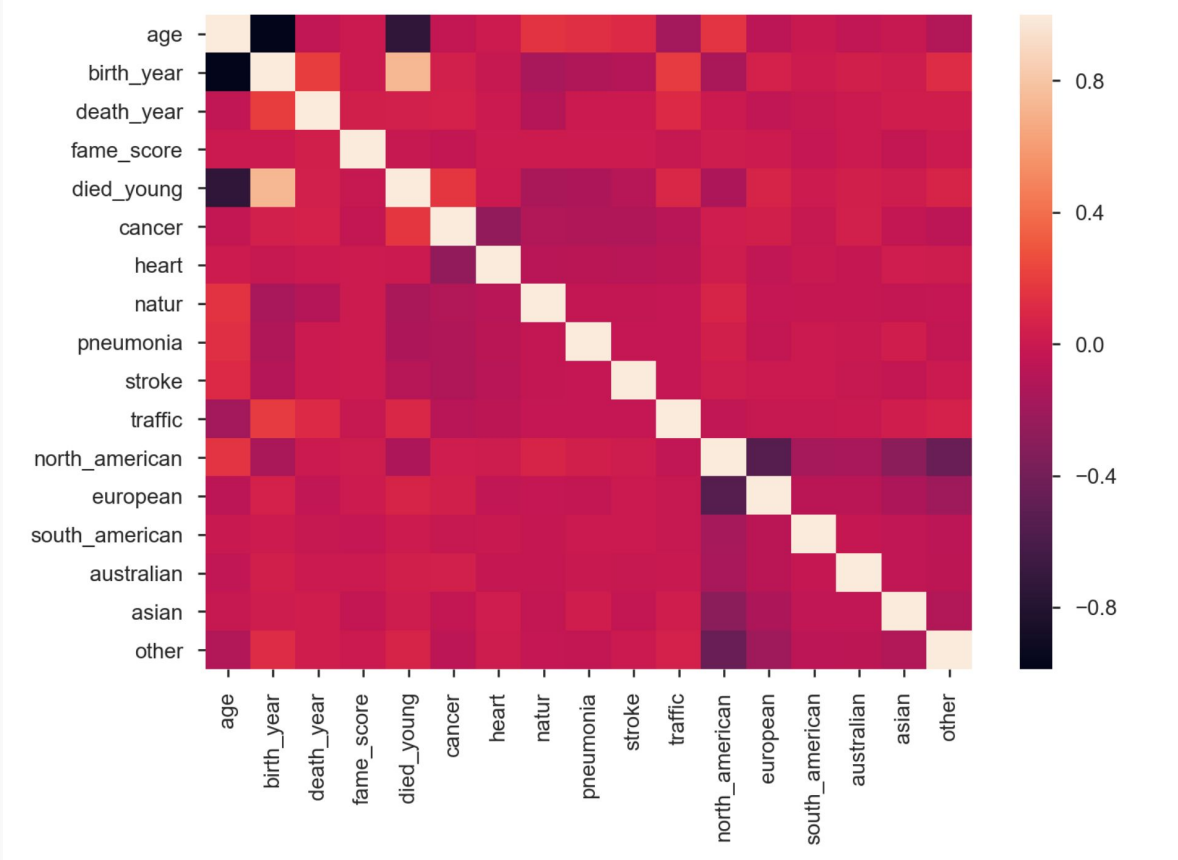
This would be my baseline to determine if a celebrity had died younger than average or not.

Populating my engineered feature for died_young based upon this, I established a baseline accuracy score of 60%.

older_than_average: 0.397611940299
younger_than_average: 0.602388059701
baseline_accuracy: 0.602388059701

Modeling

I took a look at a heatmap of some of my features to see if anything really stood out for modeling purposes.



Modeling

Since most of my data was textual, I decided to focus on classification models.

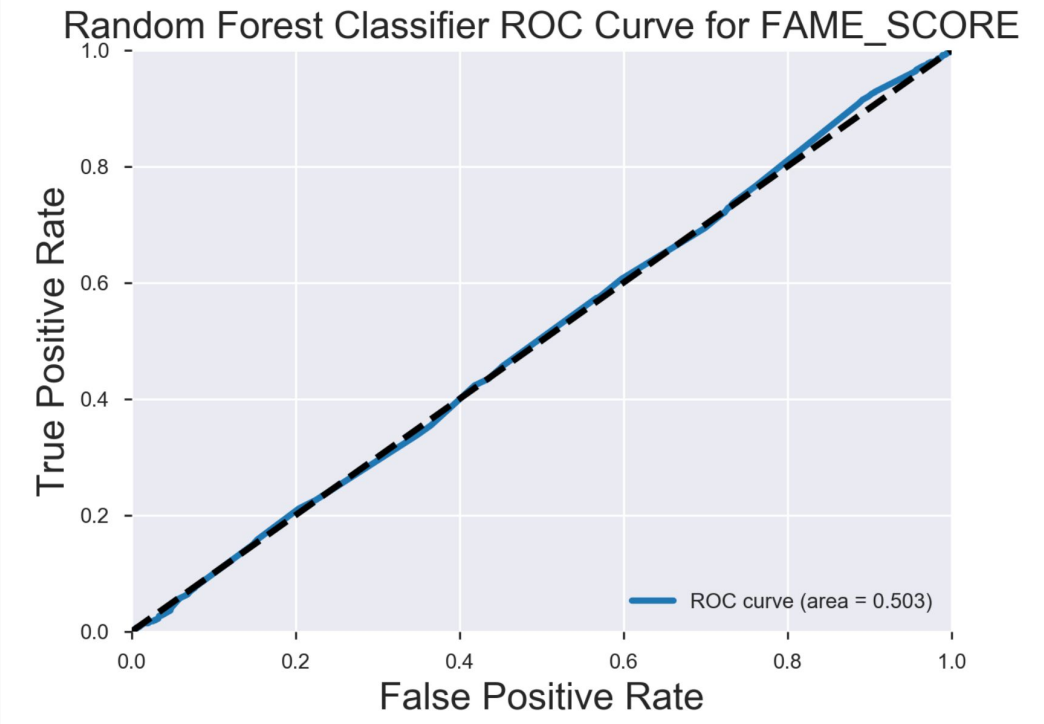
Using an evaluate_model function, I got some preliminary scores for various models.

- RandomForestClassifier
  - accuracy: 0.705133306805
- LogisticRegression
  - accuracy: 0.707918822125
- KNeighborsClassifier
  - accuracy: 0.654596100279
- GaussianNB
  - accuracy: 0.617986470354
- BernoulliNB
  - accuracy: 0.713489852766
- MultinomialNB
  - accuracy: 0.714285714286
- GradientBoostingClassifier
  - accuracy: 0.689614007163

Modeling

I plotted ROC/AUC scores for multiple features to see what combination might be the best predictor(s) for dying young.

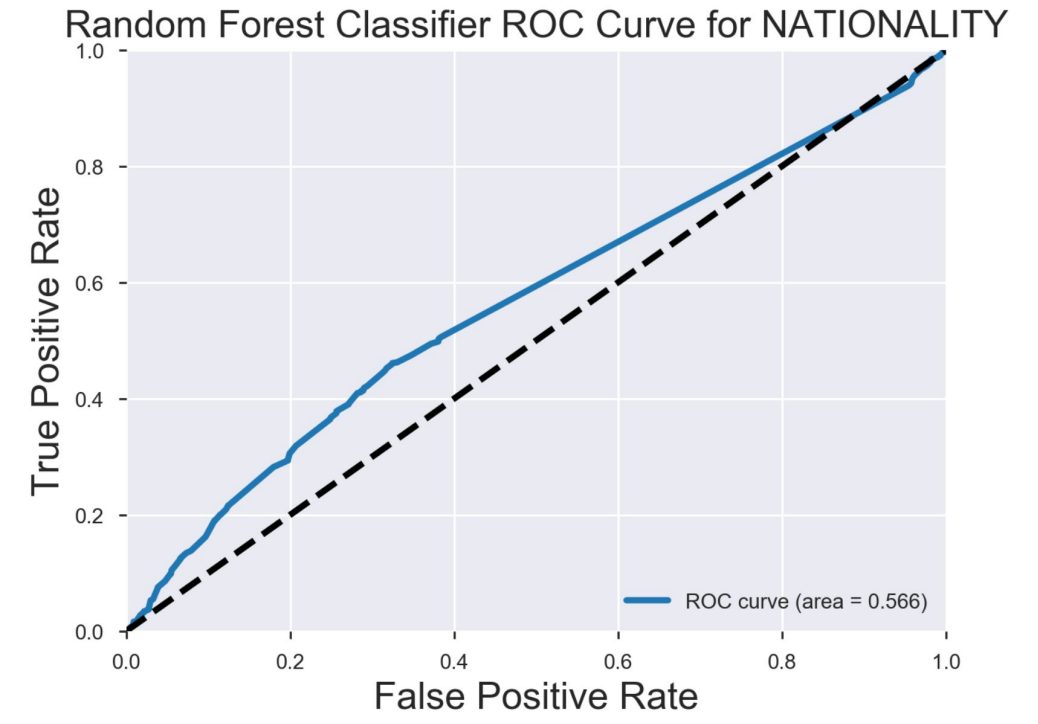
FAME_SCORE: Not very impressive.



Modeling

I plotted ROC/AUC scores for multiple features to see what combination might be the best predictor(s) for dying young.

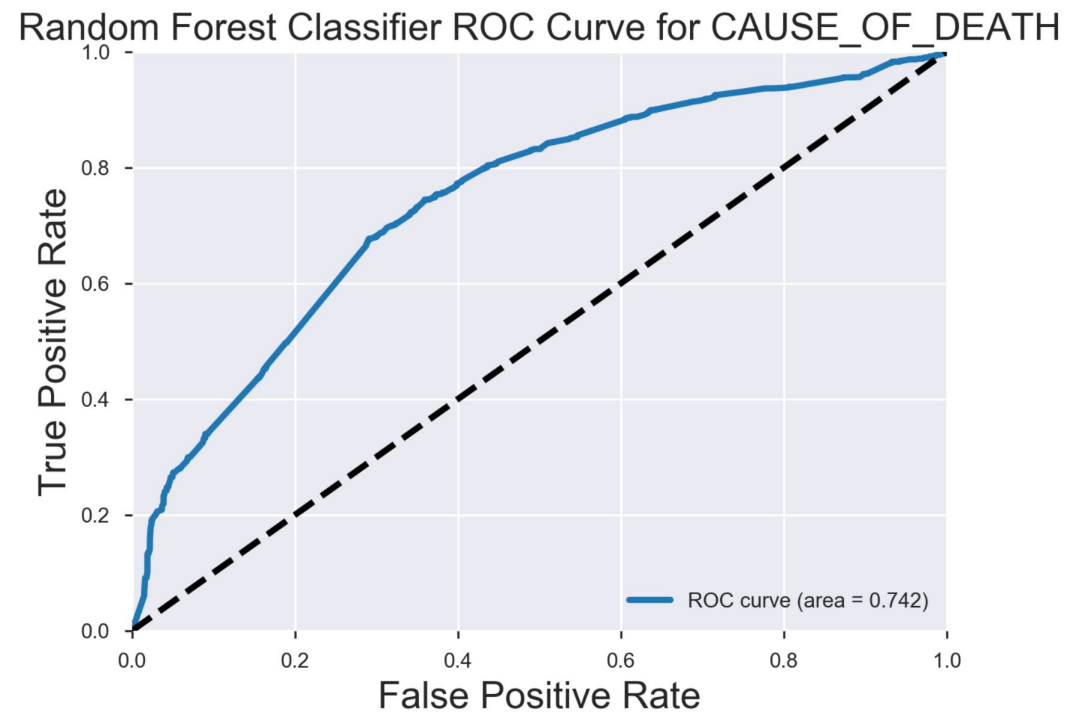
NATIONALITY: A little better but not great.



Modeling

I plotted ROC/AUC scores for multiple features to see what combination might be the best predictor(s) for dying young.

CAUSE OF DEATH: My best predictor of the bunch.



Evaluation / Future Steps

Evaluation

Overall, I was able to create a model that performed about 14% above baseline accuracy. More extensive tweaking of hyperparameters may have possibly been able to get those scores a little higher, but I saw diminishing returns once I hit around 74%.

Future Steps

Since I had to drop so many observations from my original dataset, in the future I might look for a more complete set of data or scrape my own.

I also thought it would be interesting to create a function that would allow a user to input different celebrity parameters and then predict the age of death based on my models, but I'll leave that for version 2.0.

‘Fun’ Factoid
Top Ten Celebrities Who Died Young

name	famous_for	fame_score	age	cause_of_death
Bobby Fischer	chess	695	64	kidney failur
Hugo ChÌÁvez	politician	488	58	heart attack
Michael Jackson	pop	466	50	acut propofol intox
Muammar Gaddafi	leader	403	69	shoot
Johan Cruyff	footbal	385	68	lung cancer
Whitney Houston	singer	381	48	accident drown
Ted Kennedy	politician	334	77	brain cancer
Jack Kemp	politician	330	73	cancer
David Bowie	singer-songwrit	327	69	liver cancer
Amy Winehouse	singer-songwrit	308	27	accident alcohol poison

Thanks!

Contact me:

William Long
312.213.6377
williamhlong@gmail.com

williamhlong.com
linkedin.com/in/williamhlong/
github.com/gnolw