Memento Mori

Predicting Natural vs. Unnatural Death

Christian Cofoid, Mark Ronnenberg, and Ramazan Yol

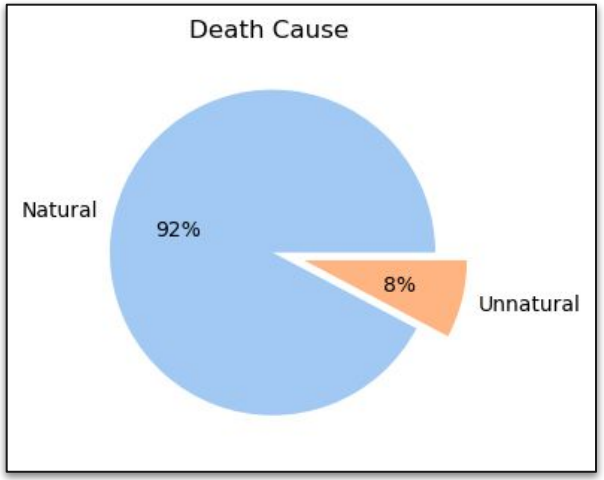
Predicting Natural vs. Unnatural Death

- **Goal:** predict whether or not a U.S. citizen will die of a **natural** or **unnatural** cause.
- **Method:** build a **binary classifier** utilizing some simple features
- **Why?**
  - **Unnatural deaths are relatively uncommon in the U.S.**
  - **Life insurance companies need to be able to offer competitive premiums while reducing sticker shock**

Data

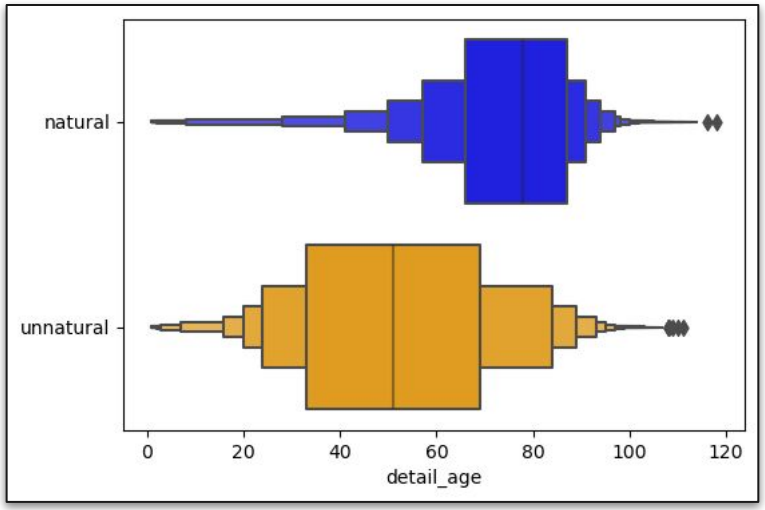
CDC Data obtained at: https://www.kaggle.com/datasets/cdc/mortality

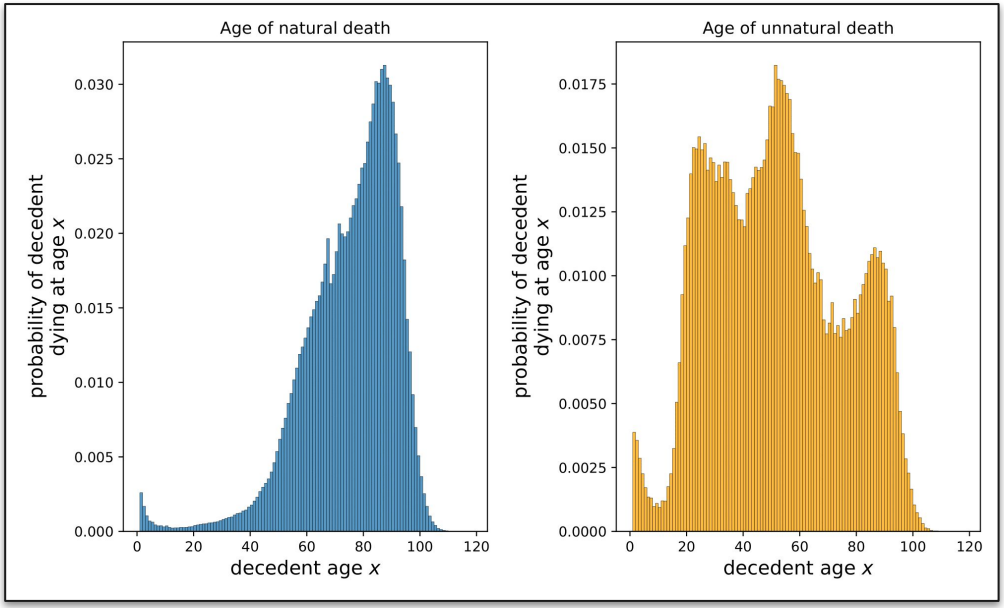
- Focus on data from 2014
- CSV containing the data, and json file containing data codes
- About 2.7 million data entries
- 77 features
- Select features: **age, sex, marital status, education**
- Target: **unnatural**
    - **One-hot encode:**
        - **Unnatural = 1**
        - **Natural = 0**



Cleaning and EDA: detail_age

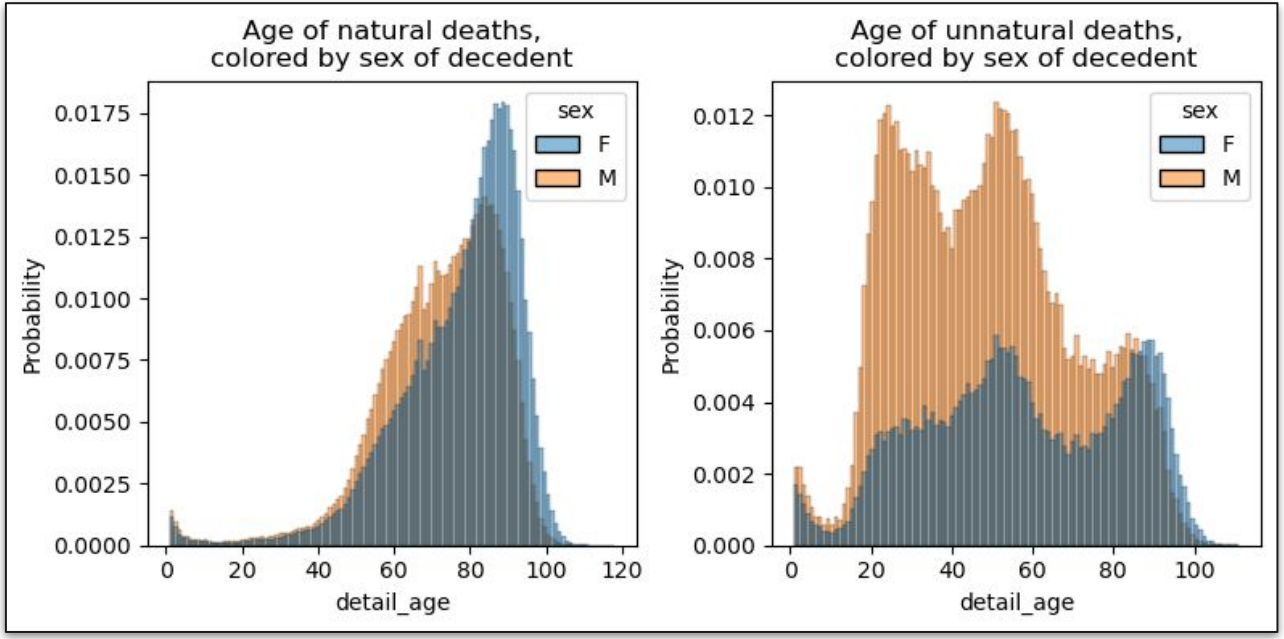
- Records age of decedent





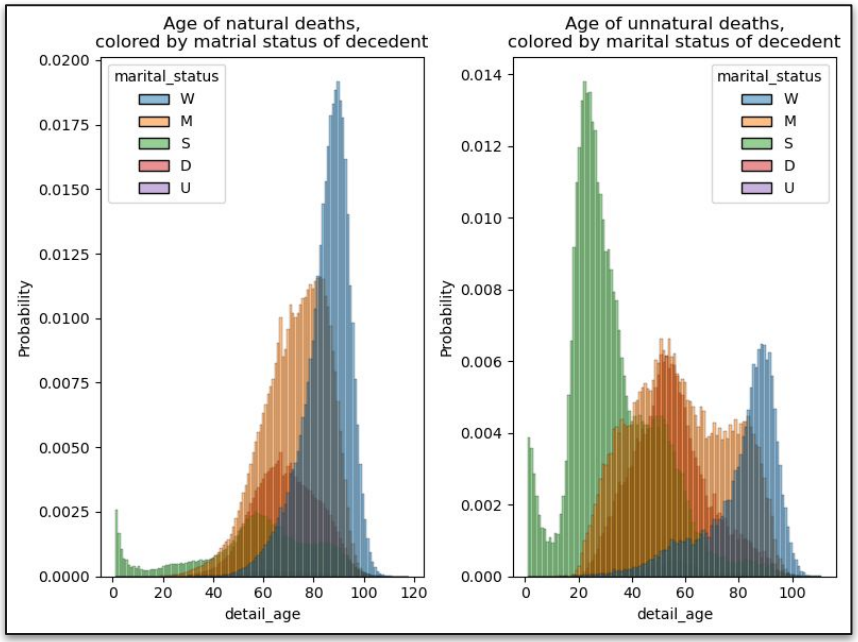
Cleaning and EDA: sex

- Records sex of the decedent



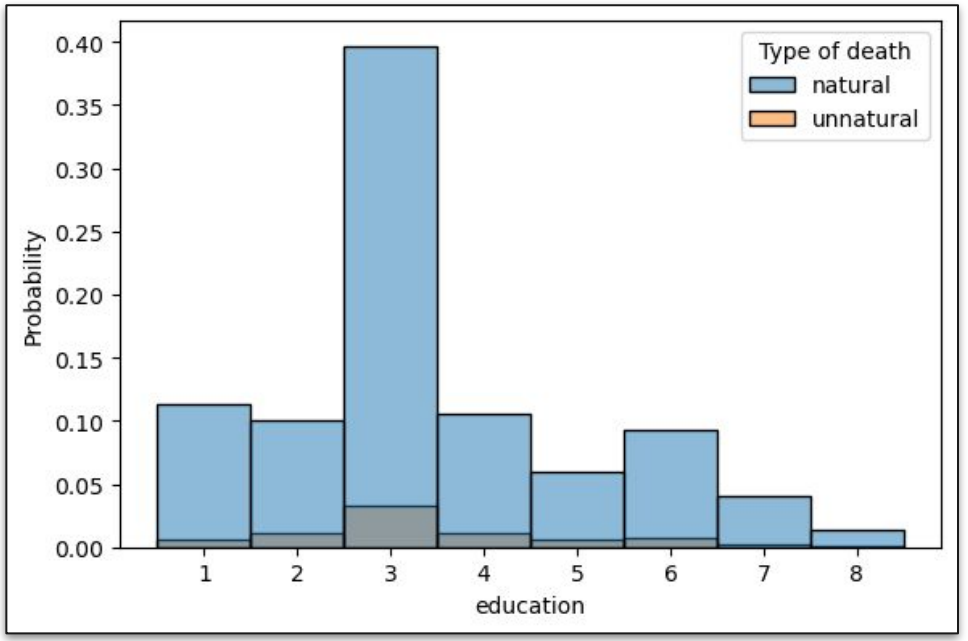
Cleaning and EDA: martial_status

- Records marital status of decedent
- Codes:
  - W = widowed
  - M = married
  - S = single, never married
  - D = divorced
  - U = unknown



Cleaning and EDA: education

- Records education level of decedent
- Created this feature by combining the *education_2003_revision* and *education_1989_revision* features
- Codes:
  - 1 = 8th grade or less
  - 2 = 9-12th grade, no diploma
  - 3 = high school graduate or GED
  - 4 = some college, no degree
  - 5 = Associate's degree
  - 6 = Bachelor's degree
  - 7 = Master's degree
  - 8 = doctorate or professional degree



Models and Results

- **Metrics:** recall and macro f_1 score
- **Baselines:**
  - Guess all instances as natural
  - Bernoulli random variable $p \approx 0.08$
  - Default logistic regression
- **Models Tested:**
  - **Logistic regression**
    - **With scaled continuous inputs**
    - **With class weights**
  - **Random forest**
  - **Random forest with weights**
  - **Naive Bayes classifier**
  - **XGBoost**
- Ran Stratified K-Fold Cross Validation

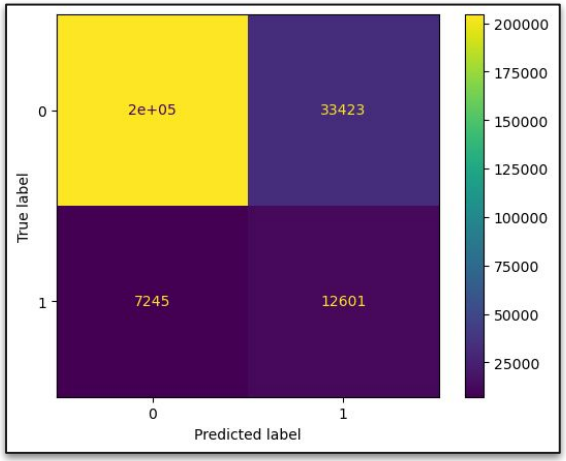
Models and Results

	recall_0	recall_1	recall_2	recall_3	recall_4	recall_mean	f1_macro_0	f1_macro_1	f1_macro_2	f1_macro_3	f1_macro_4	f1_mean
Logistic_Regression_w_Scaling+Weights	0.706304	0.704596	0.709711	0.711082	0.711027	0.708544	0.592173	0.592195	0.592806	0.594345	0.592696	0.592843
Logistic_Regression_w_Weights	0.706304	0.704736	0.710047	0.711055	0.711278	0.708684	0.592173	0.592057	0.592439	0.594420	0.592562	0.592730
Random_Forest_Classifier	0.264304	0.262876	0.260728	0.262884	0.267811	0.263721	0.672305	0.672460	0.671463	0.671903	0.674992	0.672624
Random_Forest_Classifier_w_Weights	0.649003	0.645280	0.653445	0.651373	0.652325	0.650285	0.639499	0.639035	0.637376	0.642980	0.640481	0.639874
Gaussian_Bayes	0.373222	0.371655	0.372617	0.374605	0.375416	0.373503	0.641470	0.640687	0.641728	0.642864	0.643004	0.641951
XGboost	0.999860	0.999860	0.999804	0.999664	0.999804	0.999798	0.072345	0.072018	0.072161	0.072248	0.072208	0.072196
Random_Guessing	0.076279	0.077315	0.076449	0.078213	0.074910	0.076633	0.499379	0.500050	0.500151	0.500320	0.498525	0.499685
Guess_Everything_Natural	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.479982	0.479982	0.479982	0.479982	0.479982	0.479982
Logistic_Regression	0.116588	0.115888	0.116871	0.113764	0.117123	0.116047	0.573193	0.572397	0.573646	0.571309	0.573607	0.572830

Final Model and Results

- Random Forest Classifier with weights
  - Grid search cross validation to tune hyperparameters

	feature	importance_score
0	detail_age	0.803768
4	marital_status_S	0.131128
6	sex_F	0.029270
2	marital_status_D	0.014232
1	education	0.013158
3	marital_status_M	0.008385
5	marital_status_U	0.000059



	precision	recall	f1-score	support
0	0.97	0.86	0.91	237933
1	0.27	0.63	0.38	19846
accuracy			0.84	257779
macro avg	0.62	0.75	0.65	257779
weighted avg	0.91	0.84	0.87	257779

```
f1 macro test score : 0.6460830296921457
recall test score : 0.6349390305351205
Prediction Percentages :
 0    0.821459
1    0.178541
dtype: float64
```

Stakeholders

- Life insurance companies
  - Account for the likelihood of unnatural death for a particular individual
  - Our model predicts about 65% of all unnatural deaths
  - Our model is a significant improvement over random guessing
  - Competitive initially-advertised premiums
  - Prevent insurer from offering artificially low rates

Linkedin Profiles

Christian Cofoid: https://www.linkedin.com/in/christiancofoid/

Mark Ronnenberg: https://www.linkedin.com/in/mark-ronnenberg-079b20229/

Ramazan Yol: https://www.linkedin.com/in/ryol/

Thank you!