Predicting Earthquake Fatalities

Bradley Marx
Brown University
12/06/2022

https://github.com/bradmarx112/Data1030_Project

Recap

Predicting Earthquake Fatalities

- **Question:**

  "Given the <u>location</u>, <u>magnitude</u>, and <u>focal depth</u> of an earthquake, along with <u>local population density</u> and <u>infrastructural development</u>, can we predict the resulting death toll *magnitude* which will occur?"

  Not considering deaths from tsunamis

- **Purpose:**

  To inform earthquake preparedness efforts via sensitivity analysis of earthquake-vulnerable regions

- **Problem Type:**

  <u>Regression</u> on an <u>ordinal variable</u>

Recap

Data Sources

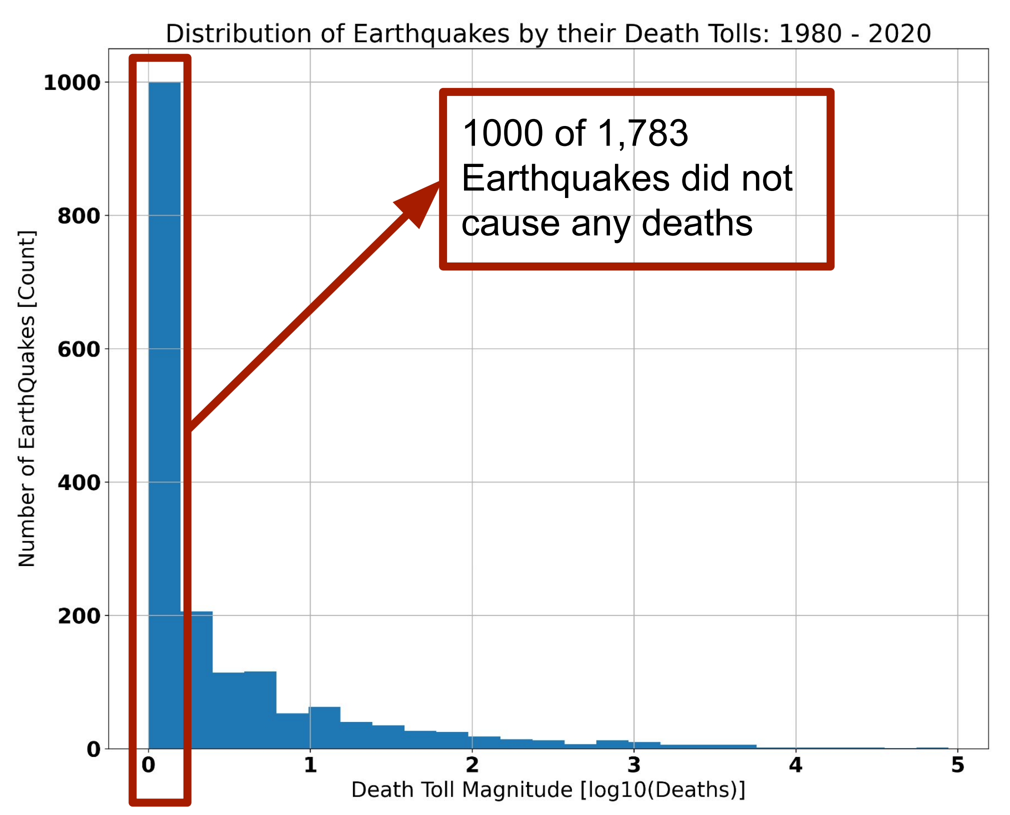
1. Record of Historic Earthquakes and Resulting Death Toll: CSV File
    a. > Magnitude 3
    b. 1980 - 2020

2. 2015 Global Population Estimates: Raster File
    a. 30 Km^2 resolution

3. 2015 Global HDI Estimates: Raster File
    a. Subnational-level resolution

Recap

Target Variable Analysis

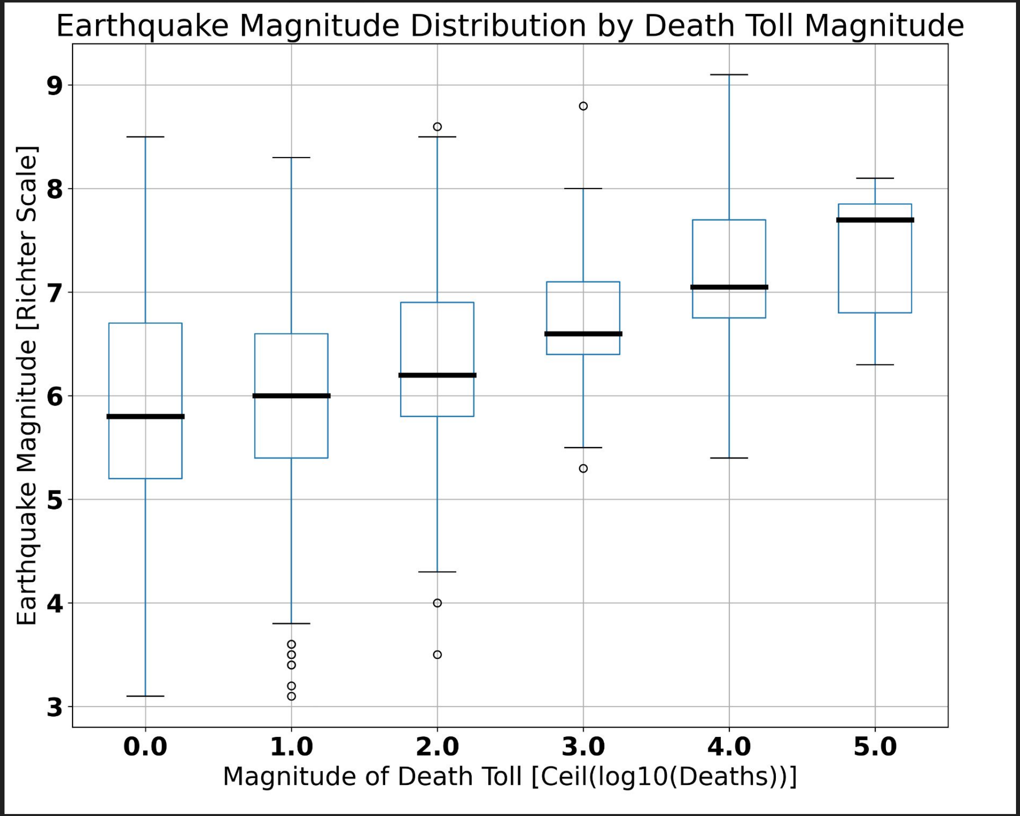


Death Toll

- Extremely long tail
- Even logged, values are highly concentrated at 0 with long tail
- 56% of earthquakes have 0 fatalities

Recap

Effect of Magnitude on Death Toll



Grouped by Death Toll Magnitude

Median earthquake magnitude increases along with Death Toll Magnitude

Zero-death earthquakes have much wider distribution

Preprocessing

- **39 Total Features after Preprocessing (20 before)**
- **1,798 Total Data Points**

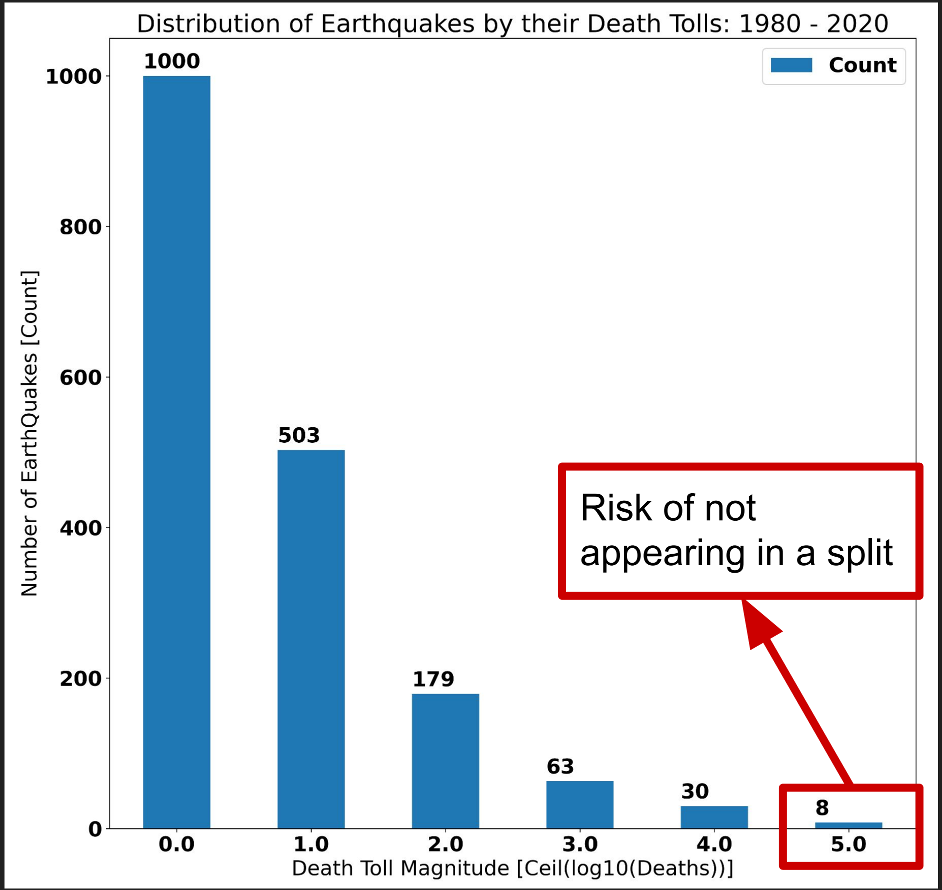
- **One-Hot Encoding**
    - Region Code

- **Standard Scaler**
    - Magnitude
    - All socioeconomic indicators
    - Geospatial mapping metadata (avg. distance between quake and population/HDI point, etc)

Data Splitting

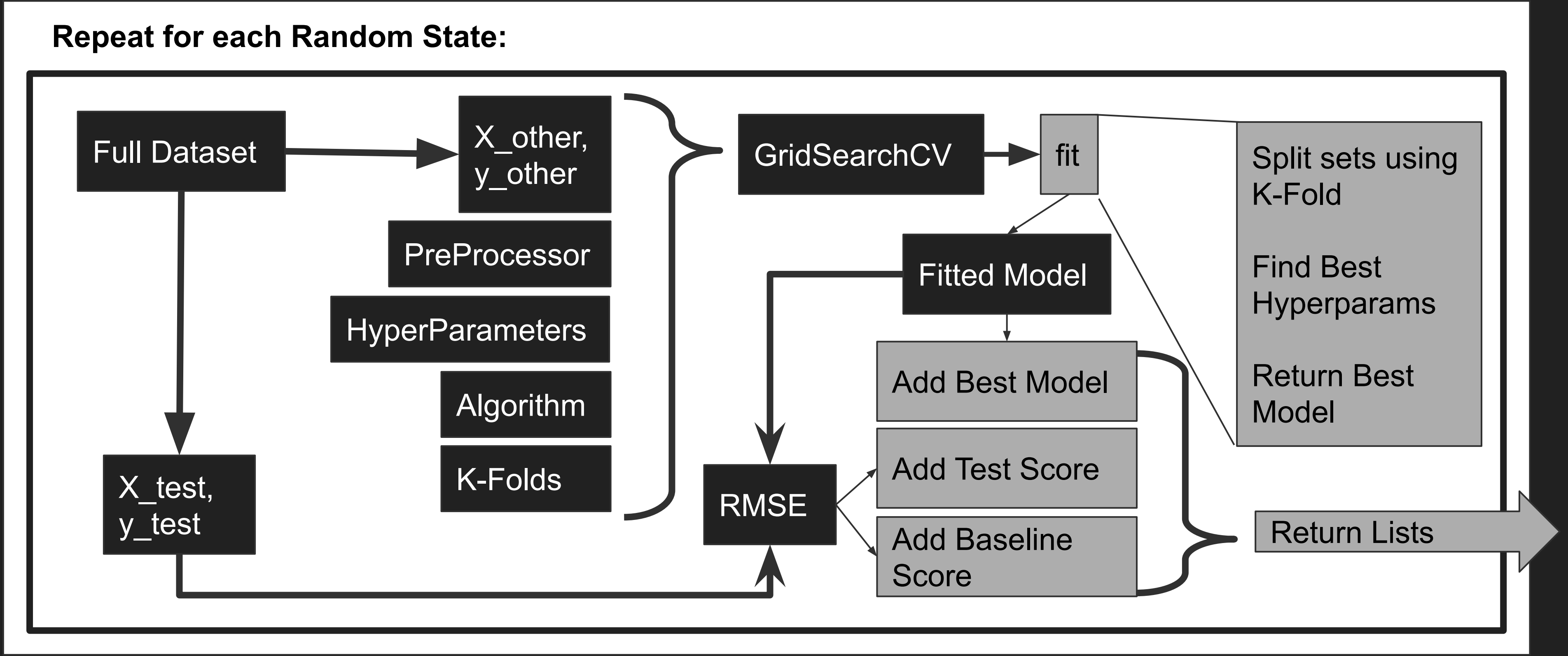
- Stratified K-Fold Splitting

- Want to ensure all categories in each split



Cross Validation

ML Pipeline



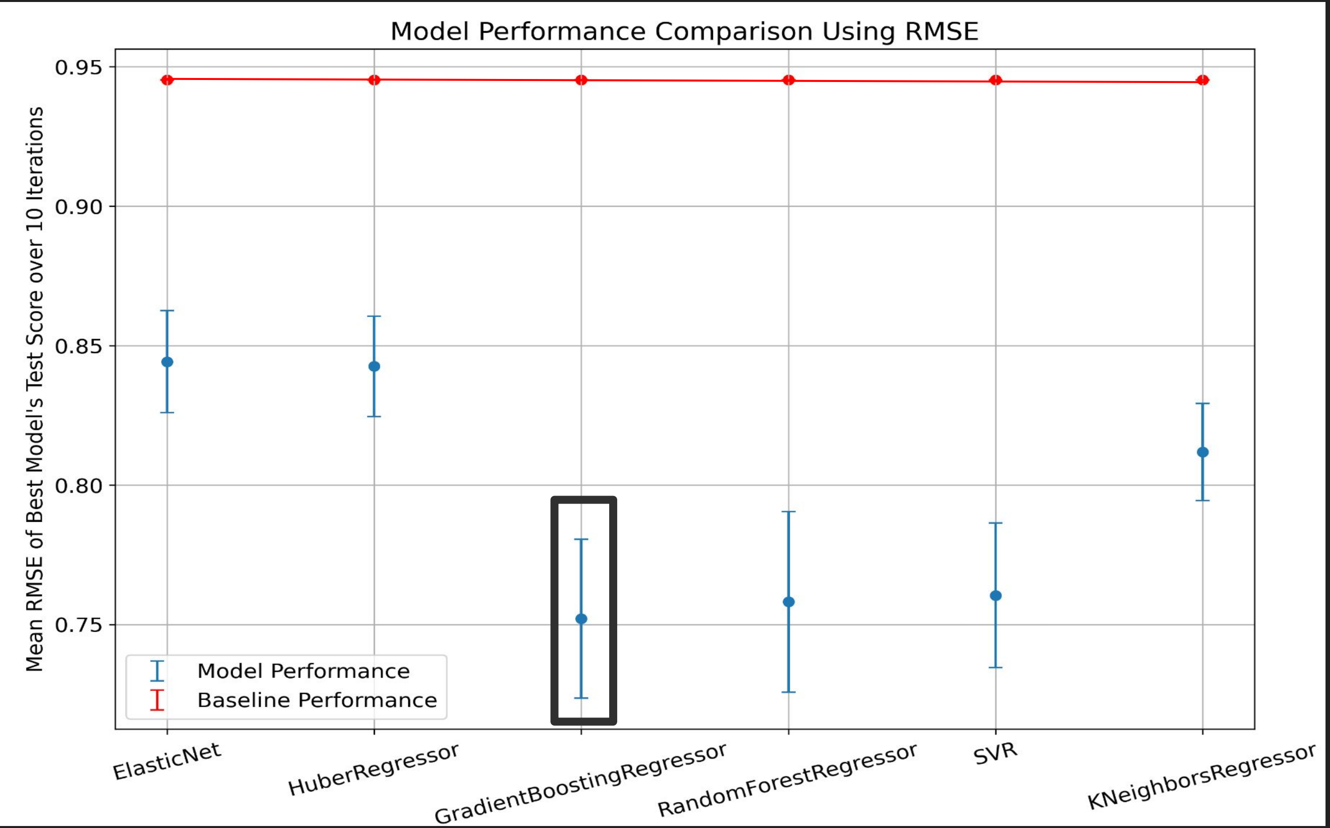
Cross Validation

Algorithms Tested

Supervised Algorithm	Linear / Non-Linear	Tuned Parameters and Values
Gradient Boosting Regressor	Non-Linear	**Learning Rate**- np.logspace(-3, 0, 10) **Max Depth**- [3, 4, 5, 6] **Num. Estimators**- [100, 300, 1000]
Random Forest Regressor	Non-Linear	**Max Depth**- [10, 20, 25, 30, 50, 100] **Max Features** - np.linspace(0.1, 1.0, 10)
Support Vector Regression	Non-Linear	**C**- np.logspace(-1, 2, 20) **Gamma**- np.logspace(-4, 2, 20)
K-Neighbors Regressor	Non-Linear	**Num. Neighbors**- [3, 5, 9, 10, 11, 20, 30, 40, 80, 100]
ElasticNet Regression	Linear	**Alpha**- np.logspace(-4, 1, 15) **L1 Ratio**- np.linspace(0.011, 0.99, 11)
Huber Regression	Linear	**Alpha**- np.logspace(-2, 2, 10) **Epsilon**- np.logspace(0.01, 1.8, 10)

Results

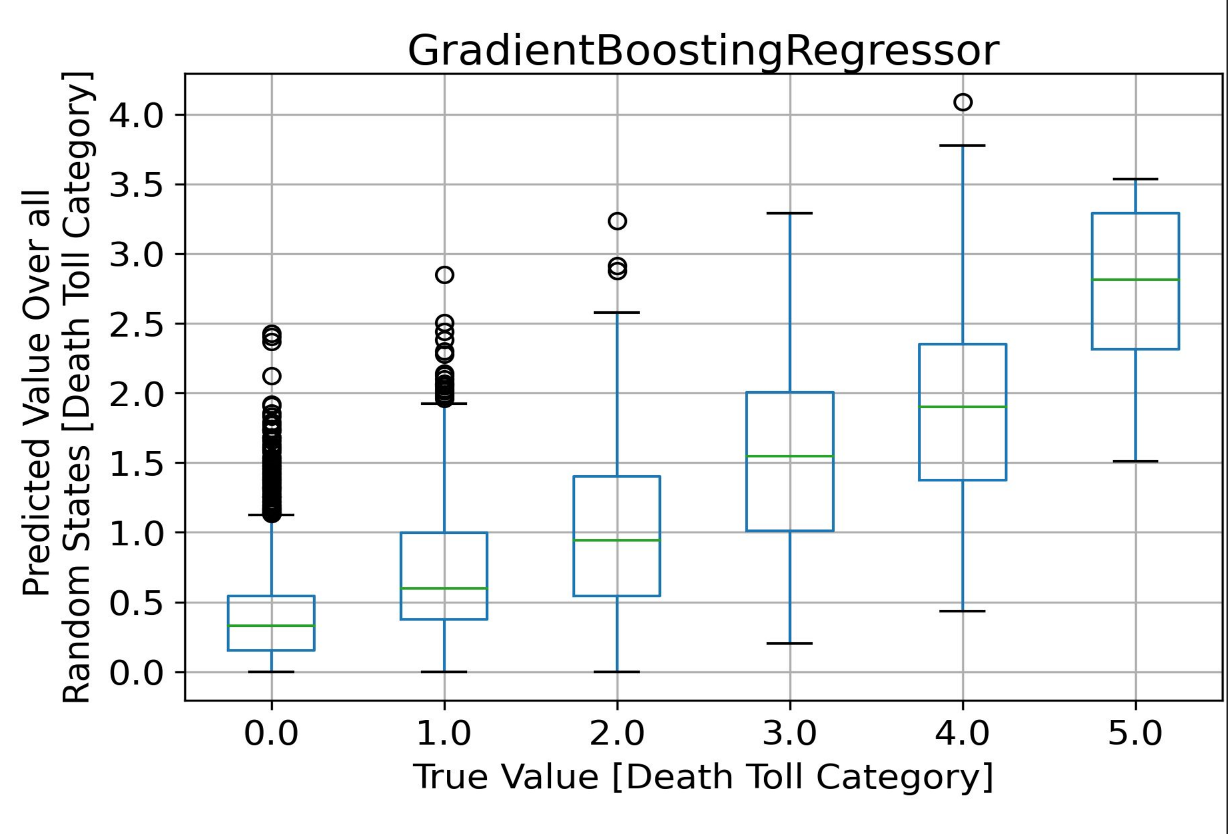
Test Scores



- **Gradient Boosting Regressor** performed best
  - 0.752 RMSE
- All models outperformed baseline

Results

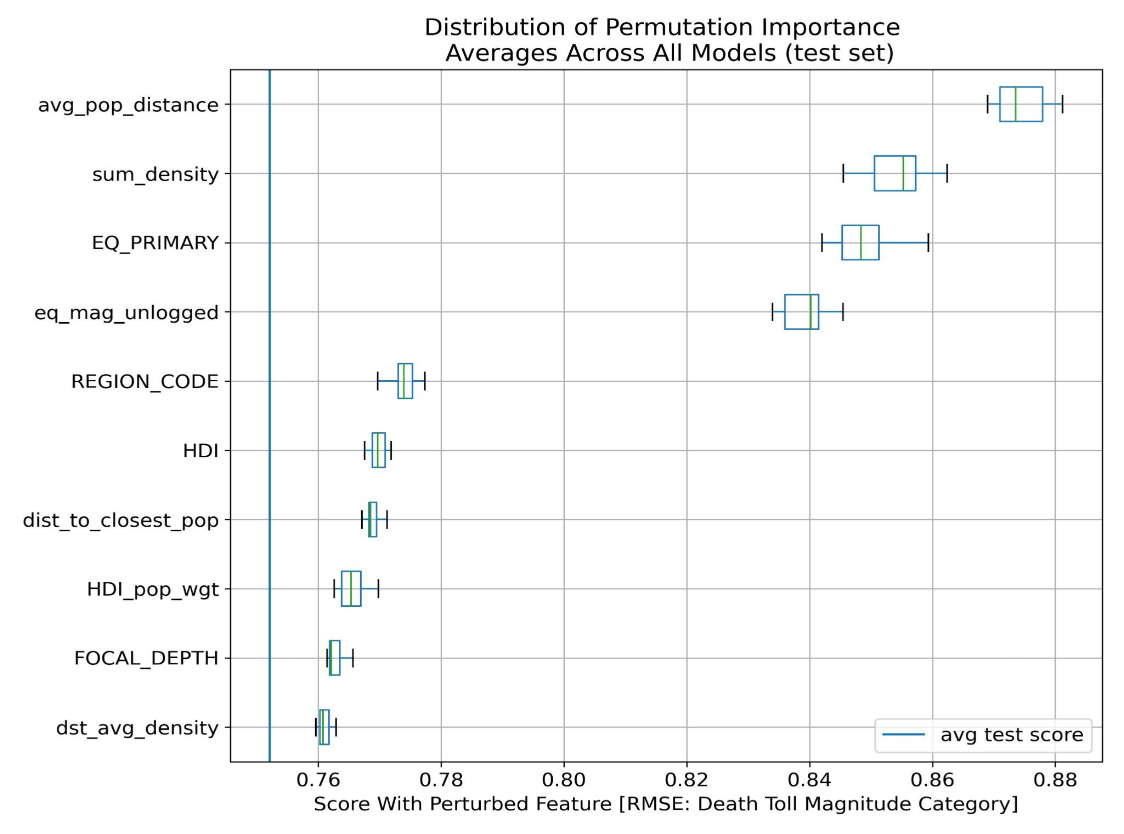
Prediction Behavior



- Predictions tended to be below actual value for categories above 0.
- Result of category imbalance?
  - Over half of of earthquakes have no death toll…

Results

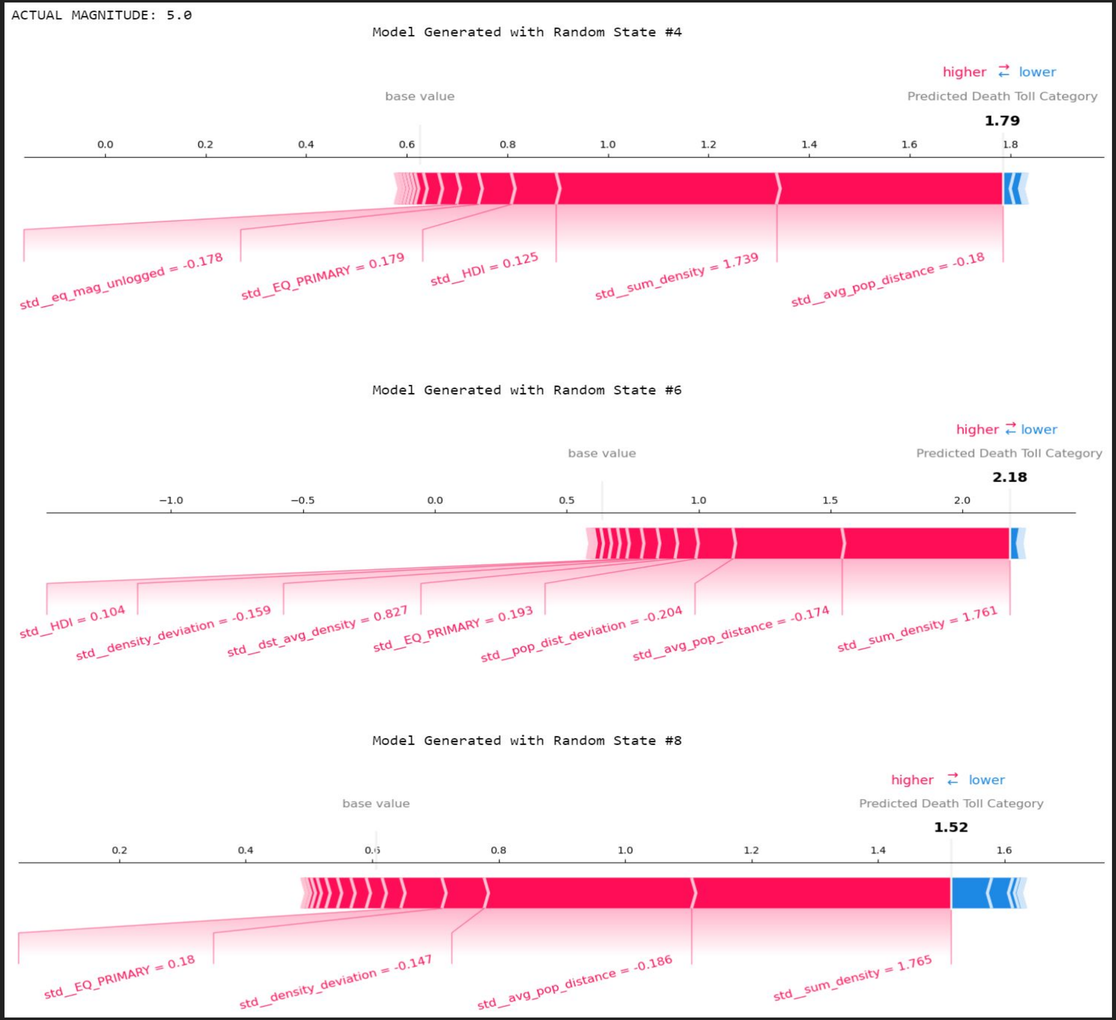
Feature Importance



- Large gap between top four most important features and rest
- Local population size and proximity have large impact on predictions!
- Earthquake magnitudes also have significant impact
- HDI much less significant

Results

Local Importance Example: Large Error



Future Improvements

- **Data**
  - Collect multiple years of gridded HDI and Population data instead of relying on one year (2015)
  - Incorporate gridded soil density data as new feature(s)
  - More robust distance calculation for geospatial mapping
    - Use Vincenty distance instead of Euclidean

- **Decompose Problem**
  - First: model binary classification problem predict if an earthquake will result in any fatality
  - Second: Perform ordinal regression on data points predicted to have fatalities

Questions

Appendix

Results

Test Scores

Rank	Model	Mean Test Score (RMSE)	Test Score Std Dev (RMSE)	Mean Baseline (RMSE)
1	Gradient Boosting Regressor	0.752121	0.028509	0.945257
2	Random Forest Regressor	0.758145	0.032391	0.945257
3	SVR	0.760472	0.025891	0.945257
4	K-Neighbors Regressor	0.811875	0.017423	0.945257
5	Huber Regressor	0.842605	0.018011	0.945257
6	ElasticNet	0.844238	0.018250	0.945257