FEMA Disaster Cost Forecasting

Michael Garber

Introduction and Problem

- The Federal Emergency Management Agency's (FEMA) primary purpose is to coordinate responses to disasters in the United States.
- As one might expect, the funding necessary for these responses can be quite costly.
- The goal of this project is to use the given data from the agency's online published dataset, OpenFEMA, to forecast reasonable cost estimates.

Source Data

I used two data sets from **OpenFEMA**, noted below.

1. The “**Declarations**” set generally describes the type of disaster and the type of assistance.
2. The “**Summaries**” set provides the financial assistance amounts, number of applications, and distinguishes between public assistance, individual assistance, and hazard mitigation.

OpenFEMA Dataset: FEMA Web Disaster **Declarations** - v1

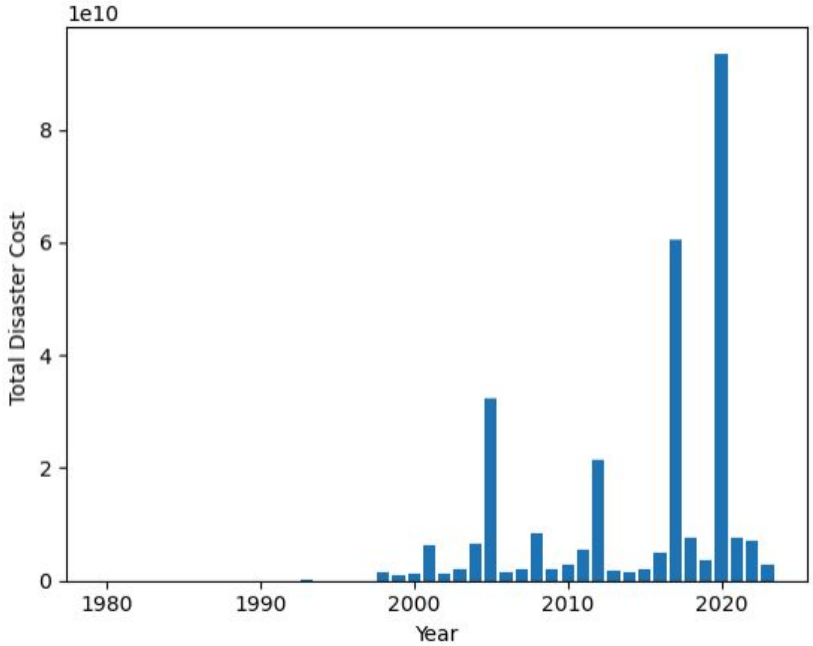
- info https://www.fema.gov/openfema-data-page/fema-web-disaster-declarations-v1
- data https://www.fema.gov/api/open/v1/FemaWebDisasterDeclarations.csv

OpenFEMA Dataset: FEMA Web Disaster **Summaries** - v1

- info https://www.fema.gov/openfema-data-page/fema-web-disaster-summaries-v1
- data https://www.fema.gov/api/open/v1/FemaWebDisasterSummaries.csv

Data Wrangling

- The two **data sets** (“Declarations” and ”Summaries”) were **joined** into a single master data set with the “disasterNumber” field used as the index
- **NULL values** in the financial and number of applications columns we **set to zero**
- **Engineered** new feature/field ‘**totalDisasterCost**’ which was the sum of Public (“totalObligatedAmountPa”), Individual (“totalAmountIhpApproved”), and Hazard mitigation (“totalObligatedAmountHmgp”) funding as the **target variable**
- **Filtered** the **data** to a finite time range of **1980 to 2023** (explanation for time range on next slide)
- **Created annual summary table** for yearly analysis (see total below)



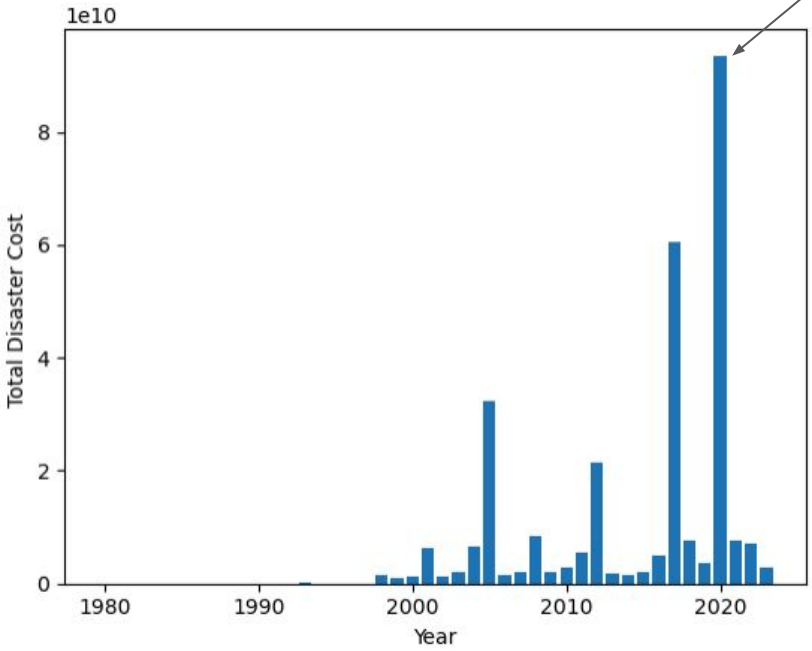
EDA - Data Range

incidentYear	totalDisasterCost
1953	0.000000e+00
1954	0.000000e+00
1955	0.000000e+00
1956	0.000000e+00
1957	0.000000e+00
1958	0.000000e+00
1959	0.000000e+00
1960	0.000000e+00
1961	0.000000e+00
1962	0.000000e+00
1963	0.000000e+00
1964	0.000000e+00
1965	0.000000e+00
1966	0.000000e+00
1967	0.000000e+00
1968	0.000000e+00
1969	0.000000e+00
1970	0.000000e+00
1971	0.000000e+00
1972	0.000000e+00
1973	0.000000e+00
1974	0.000000e+00
1975	0.000000e+00
1976	0.000000e+00
1977	0.000000e+00
1978	0.000000e+00
1979	0.000000e+00
1980	1.312700e+05
1981	0.000000e+00
1982	0.000000e+00
1983	0.000000e+00
1984	0.000000e+00
1985	0.000000e+00
1986	0.000000e+00
1987	0.000000e+00
1988	4.029932e+06
1989	0.000000e+00
1990	2.816831e+06
1991	0.000000e+00
1992	9.284427e+06
1993	1.296857e+08
1994	4.004084e+07
1995	1.655565e+07
1996	3.827911e+07
1997	6.722975e+05
1998	1.521569e+09
1999	1.103999e+09
2000	1.307119e+09
2001	6.256020e+09
2002	1.224330e+09
2003	1.918860e+09
2004	6.505514e+09
2005	3.218994e+10
2006	1.636244e+09
2007	2.127157e+09
2008	8.307189e+09
2009	2.169473e+09
2010	2.735658e+09
2011	5.478104e+09
2012	2.162758e+10
2013	1.792630e+09
2014	1.528293e+09
2015	2.093004e+09
2016	4.950492e+09
2017	6.054809e+10
2018	7.643331e+09
2019	3.693169e+09
2020	9.358846e+10
2021	7.678739e+09
2022	7.026751e+09
2023	2.889871e+09
2024	3.644366e+09

<u>Observation</u>
All costs are zero until 1980.

<u>Why?</u>
FEMA established in 1979

EDA - Data Insight



Q: 2020 was an outlier and the most costly year. Why?

A: Incident type “Biological” dominated. COVID-19 emergency declared.

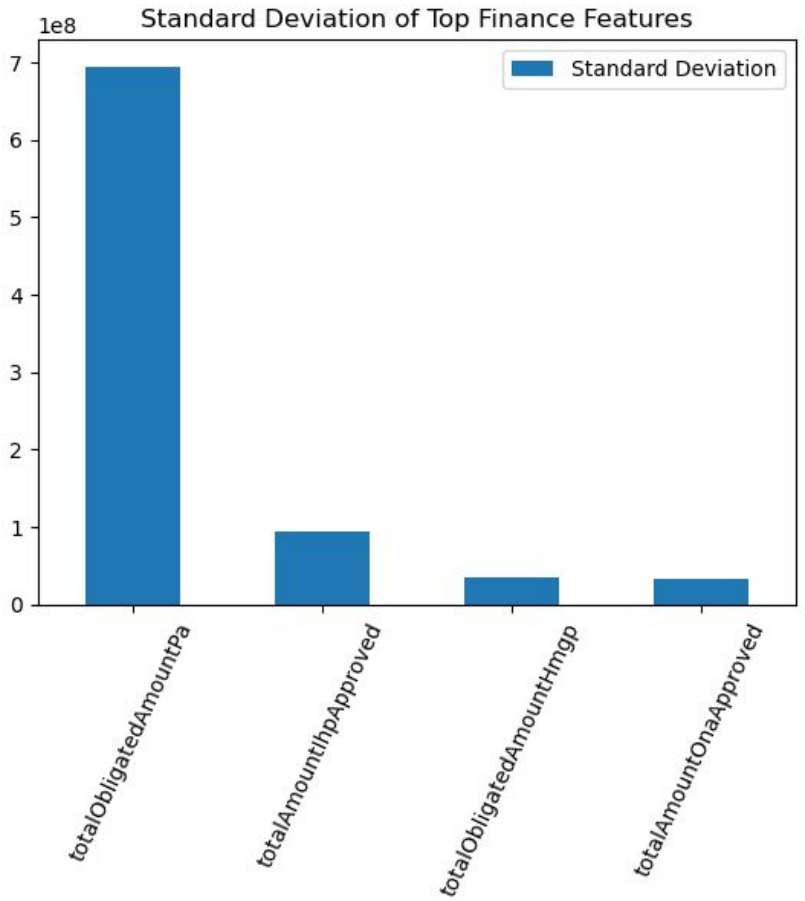
COVID-19 Stats for 2020	
# of FEMA declared COVID incidents	165
Average cost per incident	$520.3 million
Most costly single incident	$15.5 billion
Cumulative Sum of costs	$85.86 billion

Raw data sample

	declarationDate	incidentType	disasterName	totalDisasterCost
660	2020-03-20T00:00:00.000Z	Biological	COVID-19 PANDEMIC	15537706641.50
655	2020-03-25T00:00:00.000Z	Biological	COVID-19 PANDEMIC	15324259971.42
658	2020-03-22T00:00:00.000Z	Biological	COVID-19 PANDEMIC	10860797912.51
654	2020-03-25T00:00:00.000Z	Biological	COVID-19 PANDEMIC	3294448727.64
652	2020-03-25T00:00:00.000Z	Biological	COVID-19 PANDEMIC	3165835425.05

EDA - Feature Selection

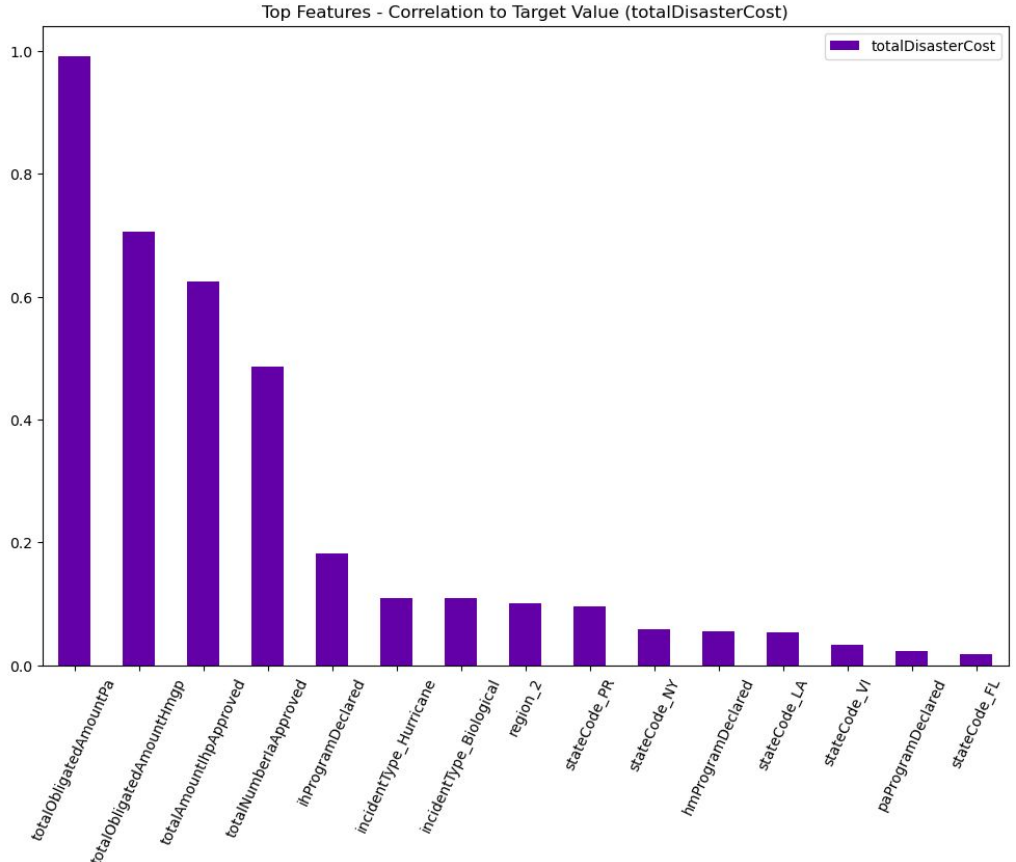
	Standard Deviation
totalObligatedAmountPa	6.943666e+08
totalAmountIhpApproved	9.426586e+07
totalObligatedAmountHmgp	3.585852e+07
totalAmountOnaApproved	3.257539e+07



EDA - Feature Selection

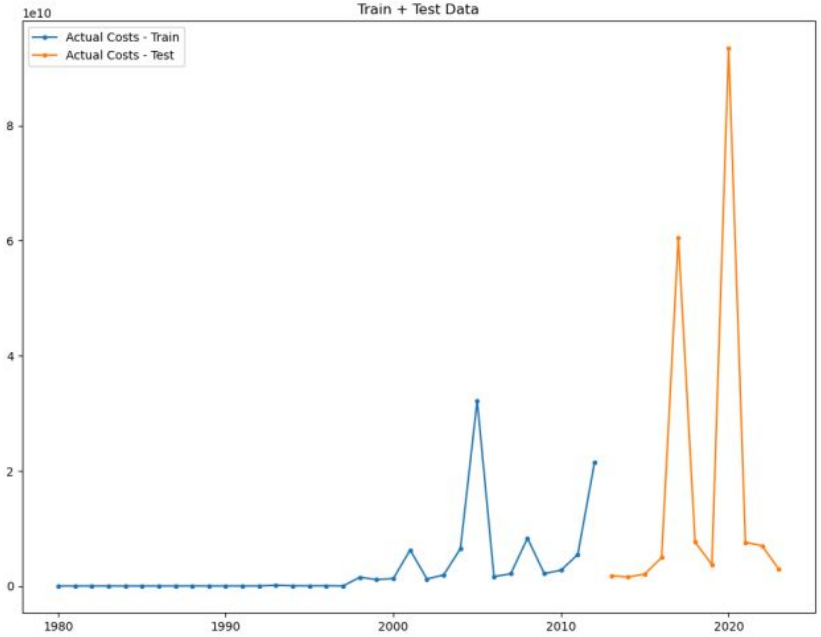
- Feature correlation to target value

	totalDisasterCost
totalObligatedAmountPa	0.991038
totalObligatedAmountHmgp	0.705335
totalAmountIhpApproved	0.623800
totalNumberIaApproved	0.486875
ihProgramDeclared	0.182474
incidentType_Hurricane	0.109729
incidentType_Biological	0.109476
region_2	0.101300
stateCode_PR	0.096434
stateCode_NY	0.058935
hmProgramDeclared	0.054482
stateCode_LA	0.052970
stateCode_VI	0.033389
paProgramDeclared	0.022607
stateCode_FL	0.017756



Approach

Once EDA and data wrangling was complete, the annualized data was standard scaled and train\test split. The target value “totalDistasterCost” is plotted below from 1980 to 2023.



Three models were chosen for evaluation.

1. Linear Regression
    a. using Robust scaled data as the model is sensitive to outliers (also tested with unscaled data)
2. Lasso Regression
    a. Chosen because the algorithm performs built-in feature selection and is less likely to overfit
3. Random Forests
    a. Chosen as an alternative to the other two linear-based models and because it’s resistant to overfitting
    b. Hyperparameters tuned: n_estimators, max_features, max_depth

Results

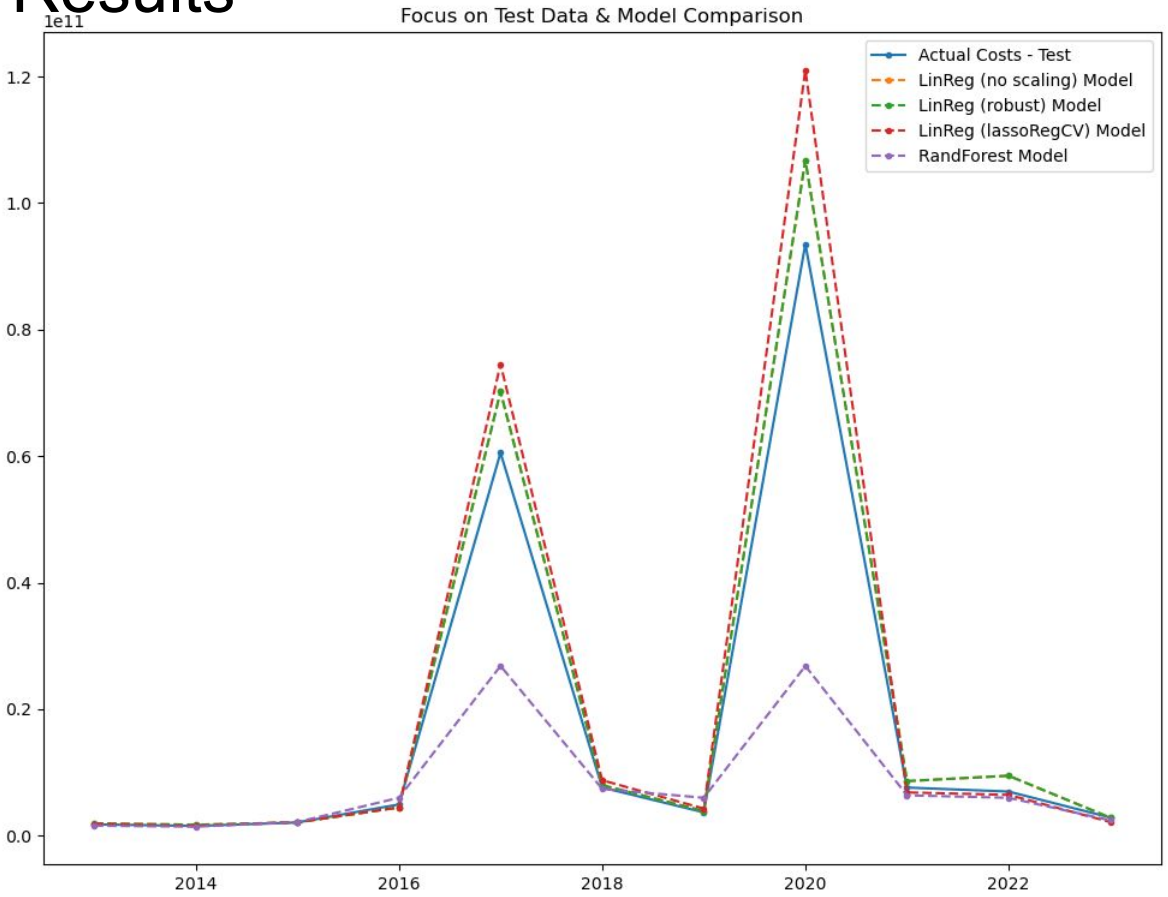
- Evaluation Metric used
  - Mean Absolute Percentage Error (MAPE)

- Justification for choosing MAPE?
  - For FEMA's high variance data set, MAPE provides consistent, representative results
  - Some other metrics like RMSE overly penalize errors whereas MAPE analyzes them proportionally

- MAPE results for each model
  - Linear Regression (robust) : 0.10993907091761934
  - Linear Regression : 0.10993907091761944
  - Linear Regression (Lasso) : 0.12913890487107163
  - Random Forest : 0.25576845002626825

- Best model
  - Linear Regression (with robust scaled data) minimized the error (MAPE)
  - The target value "totalDistasterCost" plus model predictions are plotted below from 2013 to 2023.

Results



MAPE results for each model

- Linear Regression (robust) : 0.10993907091761934
- Linear Regression : 0.10993907091761944
- Linear Regression (Lasso) : 0.12913890487107163
- Random Forest : 0.25576845002626825

Further Research and Improvements

- Use weather prediction modeling to forecast future disasters. Use this as input for forecasting cost
  - Climate Predictions from Climate Risk and Resilience Portal (ClimRR)
    - https://climrr.anl.gov/

- I was not aware that the FEMA data set had such granular financial values which included well as constituent values (e.x. “totalAmountOnaApproved” or the “Other Needs Assistance (ONA)” makes up a portion of the IHP (Individual and Households Program) - “totalAmountIhpApproved”. This causes multicollinearity of the financial data and must be handled carefully

- The data set has a column that lists all other incident types (e.x. hurricane, fire, biological, etc) for each individual disaster supporting one incident to many disaster type cardinality, but using this would further complicate analysis

- Scope creep - in the cost forecasting objective as the final model delivered is designed to use current data to predict the target data of the same time period rather than the next year. This represents some drifting of objective