FEMA Disaster Cost Forecasting
Final Project

Problem Statement

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.

Data

I used two data sets from OpenFEMA, noted below. The "Declarations" set generally describes the type of disaster and the type of assistance. 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

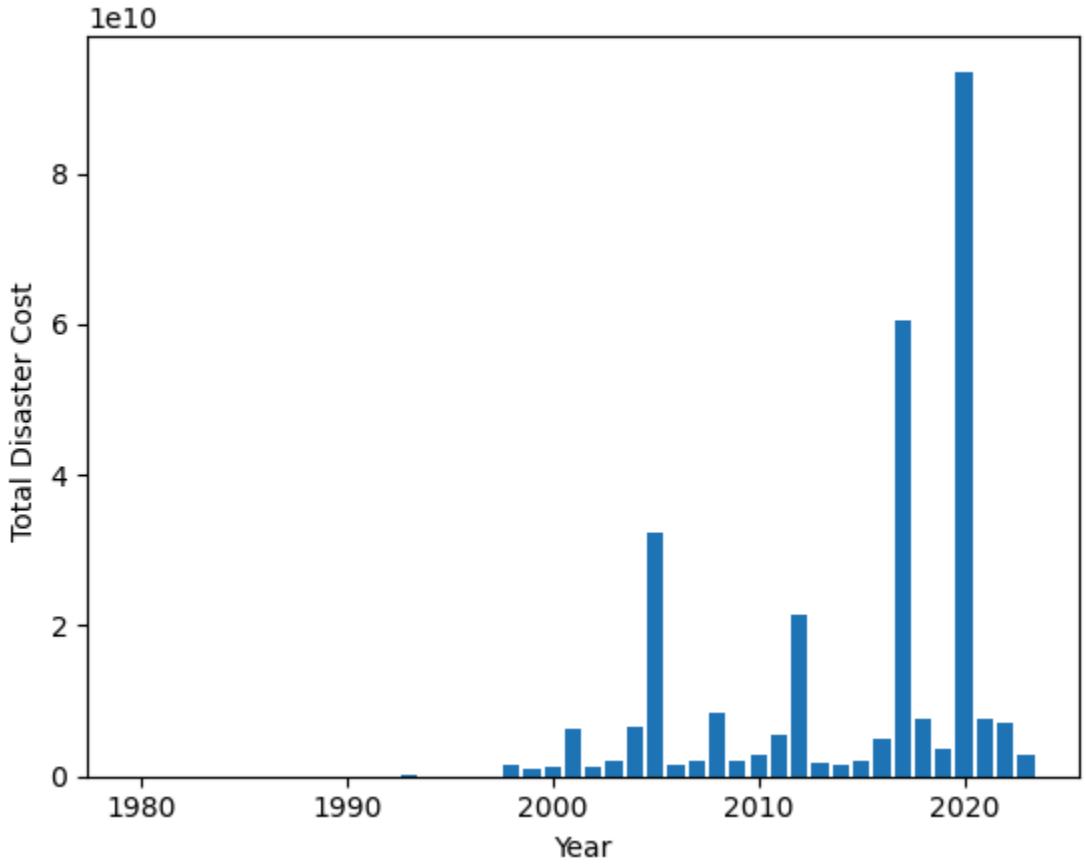
- info https://www.fema.gov/openfema-data-page/fema-web-disaster-declarations-v1
- data https://www.fema.gov/api/open/v1/FemaWebDisasterDeclarations.csv

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

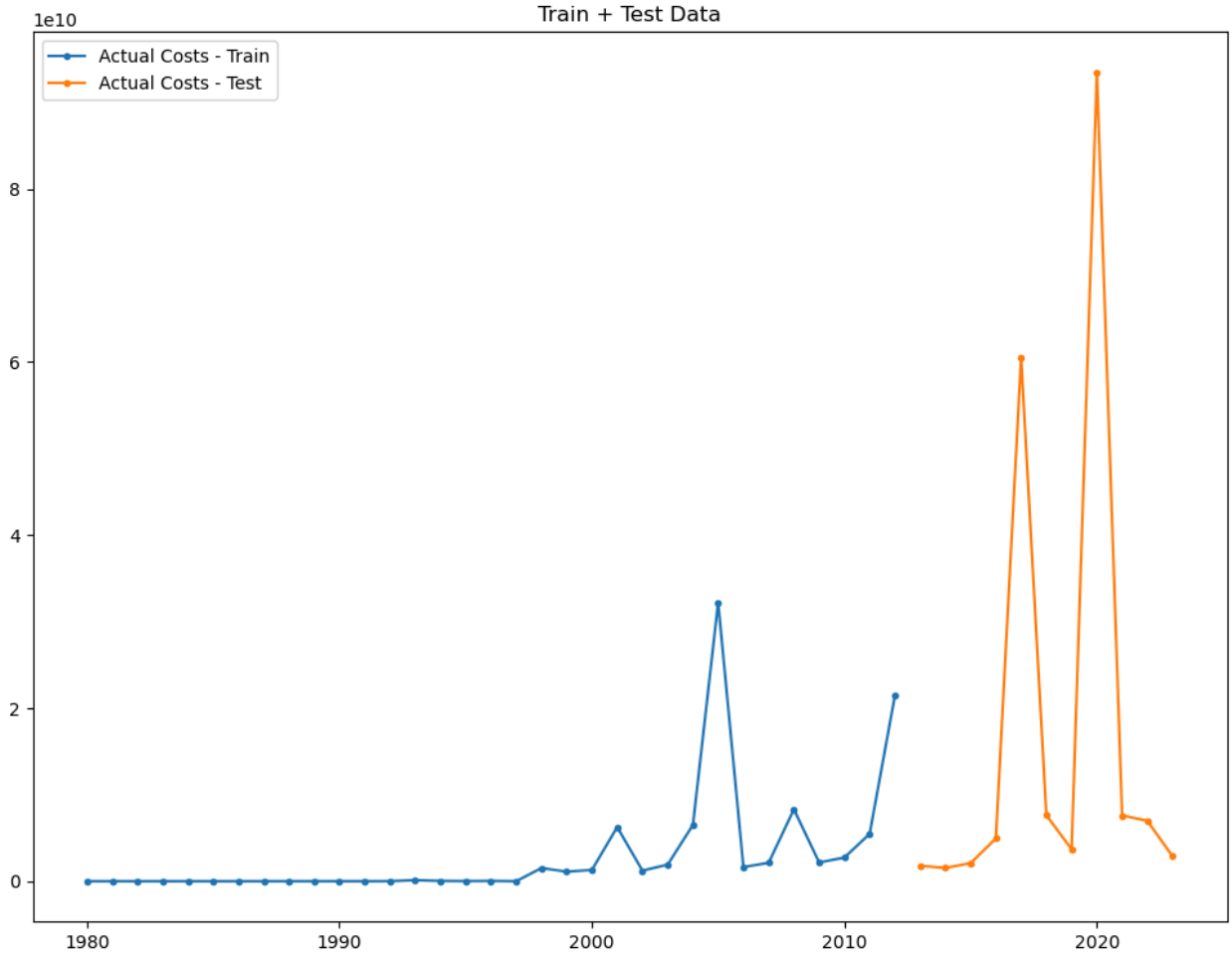
Data Wrangling

- The two data sets (“Declarations” and ”Summaries”) were joined into a single master data set with the disasterNumber field 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, Individual, and Hazard mitigation funding as the target variable
- Filtered the data to a finite time range of 1980 to 2023
- Created annual summary table for yearly analysis (see Totalbelow)



Approach

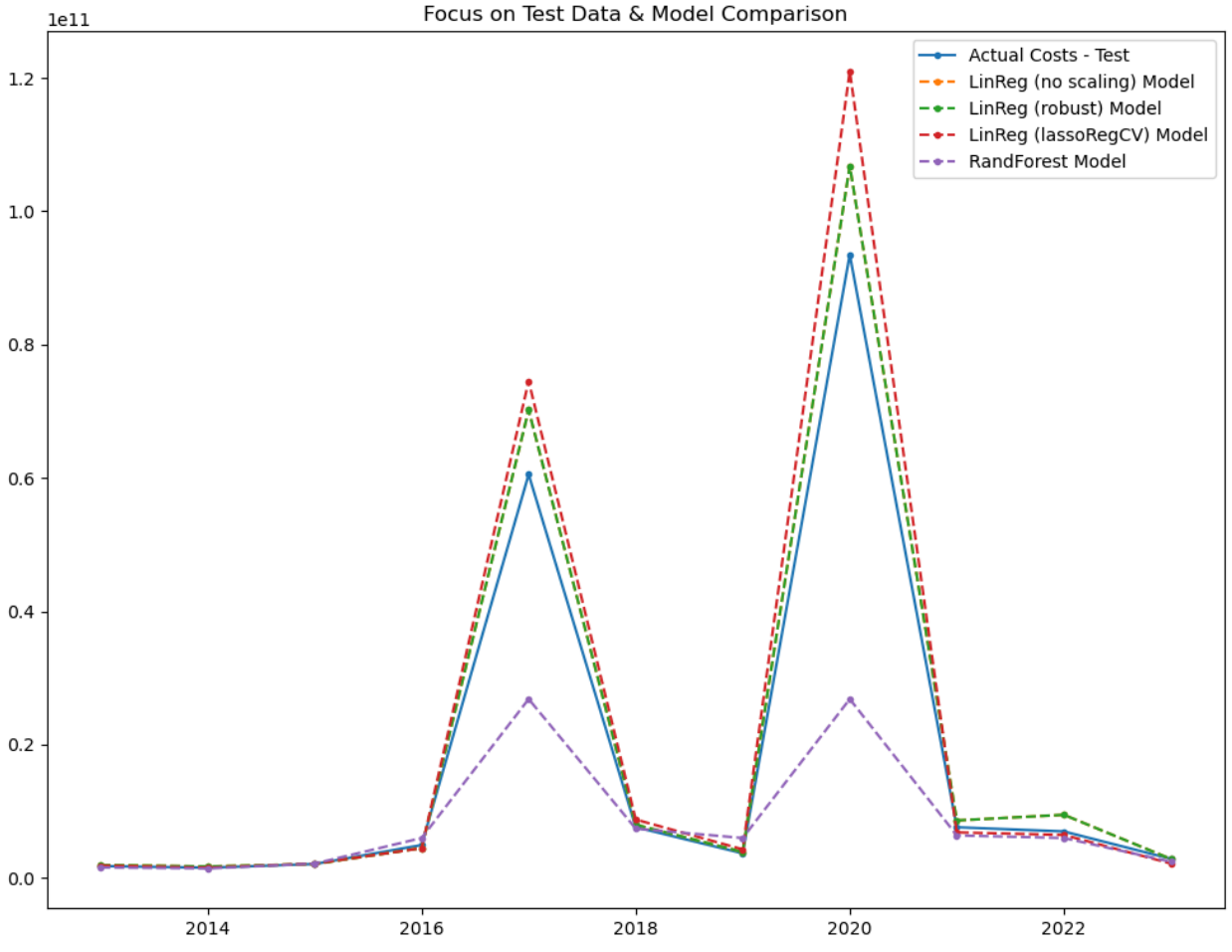
Once EDA and data wrangling was complete, the annualized data was standard scaled and train\test split. The target value “totalDistasterCost” is plotted below from 1980 to 2023.



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

1. The mean absolute percentage error (MAPE) was used to evaluate these models.
2. MAPE was chosen as it provides a consistent metric for comparing results on what is a high variance data set with regards to FEMA disaster cost values.
    a. MAPE results for each model
        i. Linear Regression (robust) : 0.10993907091761934
        ii. Linear Regression : 0.10993907091761944
        iii. Linear Regression (Lasso) : 0.12913890487107163
        iv. Random Forest : 0.25576845002626825
3. Best model: Linear Regression (with robust scaled data) minimized the error
    a. The target value “totalDistasterCost” plus model predictions are plotted below from 2013 to 2023.
4. 

Ideas for Further Research

- Since the financial data in the “Summaries” data set was so multi-collinear and especially with a single feature “totalObligatedAmountPa” accounting for an outsized portion of the model’s performance, it may be more interesting to use only the “Declarations” data to estimate disaster costs. This is the data set which includes categorical fields like type of disaster type (“hurricane”, “fire”, etc) and geographic region where the disaster incident occurred.

Recommendations for how to use the data

1. The info and model can be used as an aid for estimating disaster costs
2. The analysis (especially during EDA steps) provides overall insights including the most costly disaster types and regions and how different disaster types are spread over out over different regions
3. The Lasso Regression and Random Forest models can be used to perform feature selection to answer various questions about how import a specific feature is for predicting a given target variable