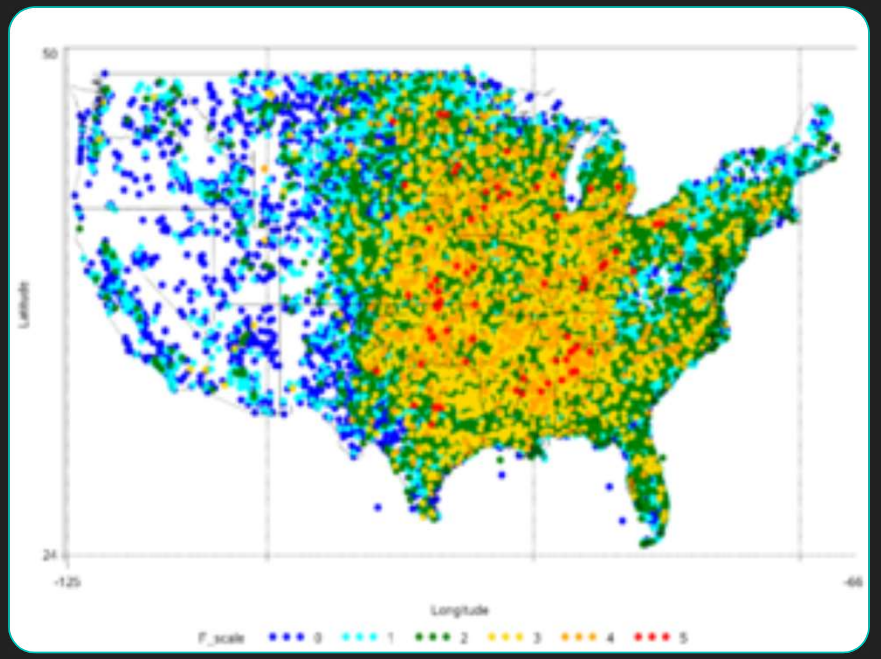
Predicting Tornado Impact

Travis Martin

Springboard Data Science Intensive Capstone Project

The Problem

- US impacted by more tornadoes than any other country by nearly 20x
- Even smallest tornadoes capable of destroying structures and causing loss of life
- Can existing data be leveraged to predict whether a tornado is likely to cause injuries and deaths?



The Data

- NOAA has compiled data on every tornado touchdown in the US 1950-2019
  - 66,389 rows and 29 columns

- Dataset Features
  - Tornado identifier
  - Time data (time and time zone)
  - Geographic data (state id, state number)
  - The number of that tornado in that state for that year
  - The magnitude of the tornado (on the F or EF-scale depending on year)
  - Human toll (fatalities, injuries)
  - Economic damage (property damage, crop damage)
  - Geographic track (starting and ending latitude and longitude, and length of the track in miles)
  - Width of the tornado in yards
  - Counties impacted, by FIPS code

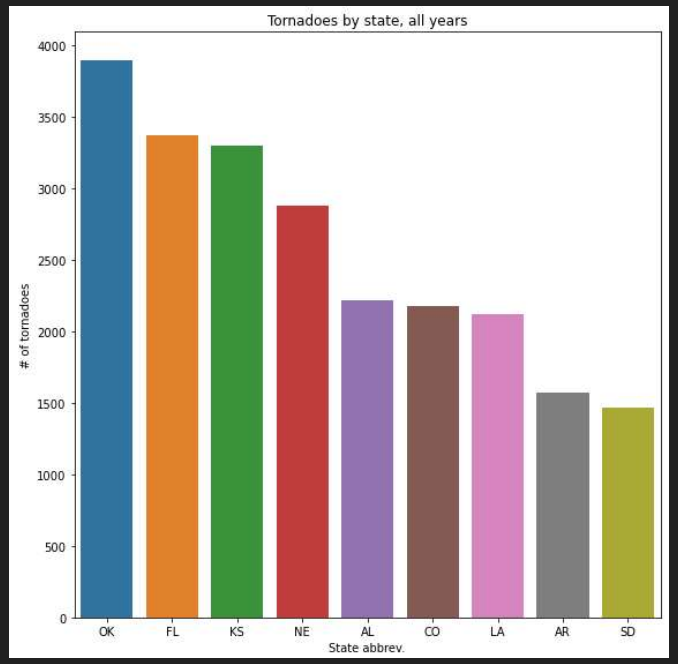
Data Wrangling

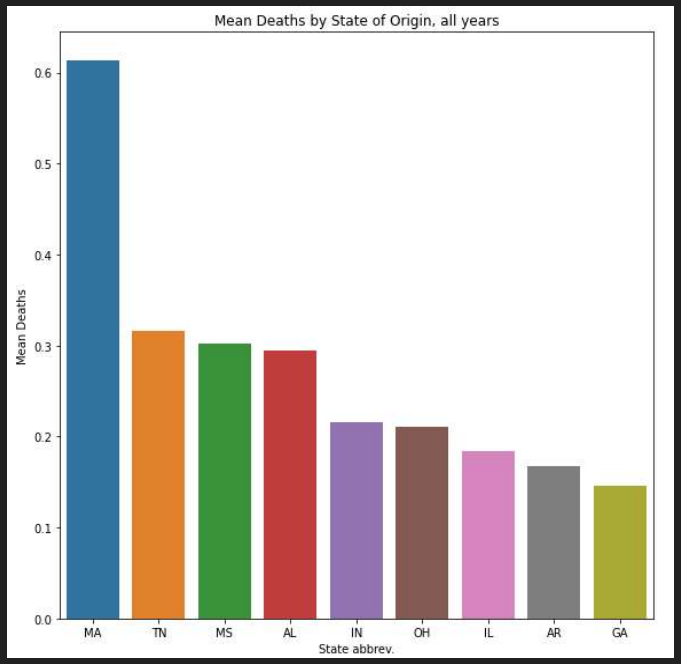
Issues Addressed

- Condensing all tornadoes to a single row with a single unique ID
- Removing duplicate county entries for same tornado
- Standardizing Property Damage (process changed in 1996)
- Filling null-values in magnitude and lat/lon columns
- Merging in county-level Land Area and Population Density data (source: Census data on data.gov)

Exploratory Data Analysis

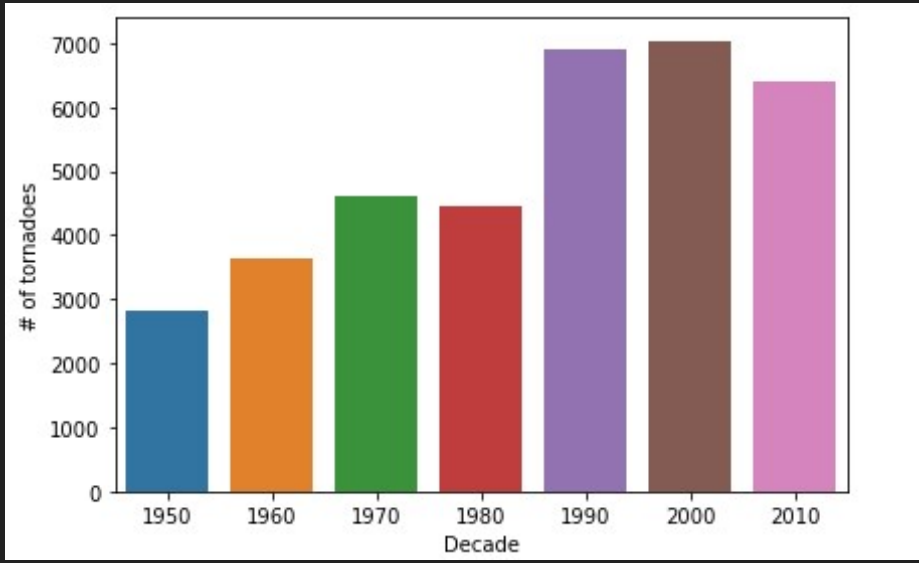
Do the states with the most tornadoes also have the deadliest?





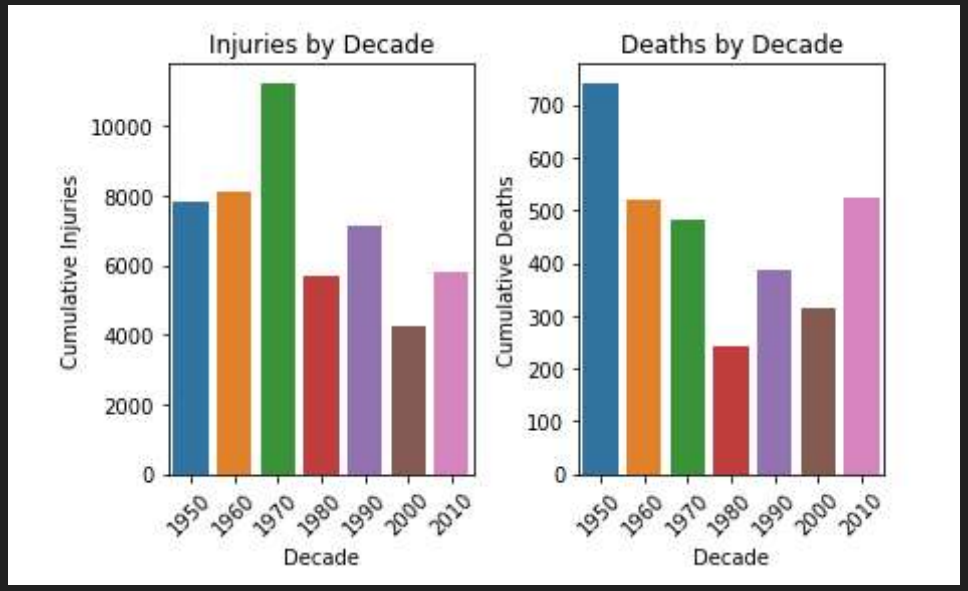
Exploratory Data Analysis

Are tornadoes becoming more frequent over time?



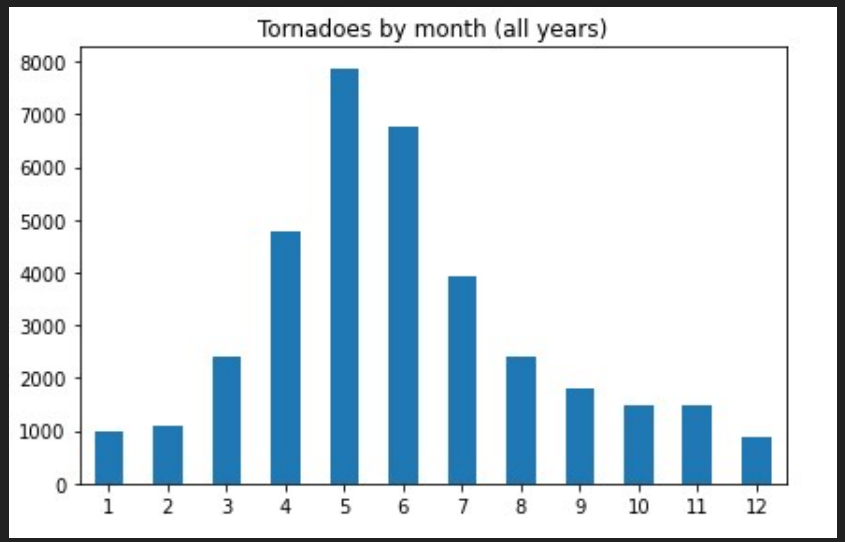
Exploratory Data Analysis

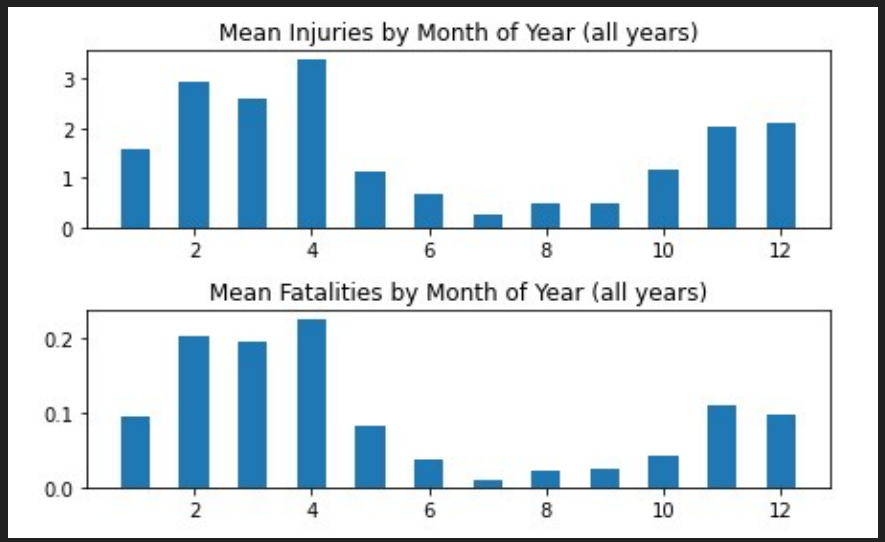
Have warning system improvements led to a decline in injuries and deaths?



Exploratory Data Analysis

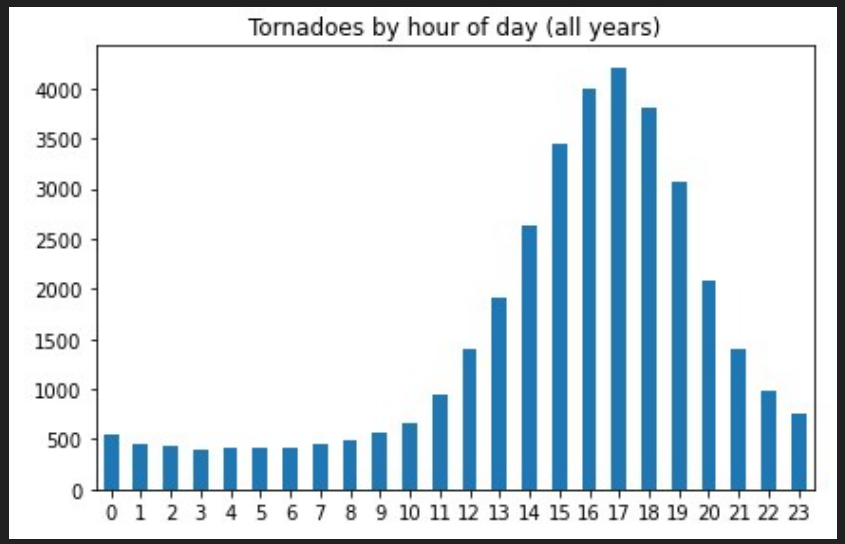
Is there a tornado season? Do injuries/deaths vary by month?

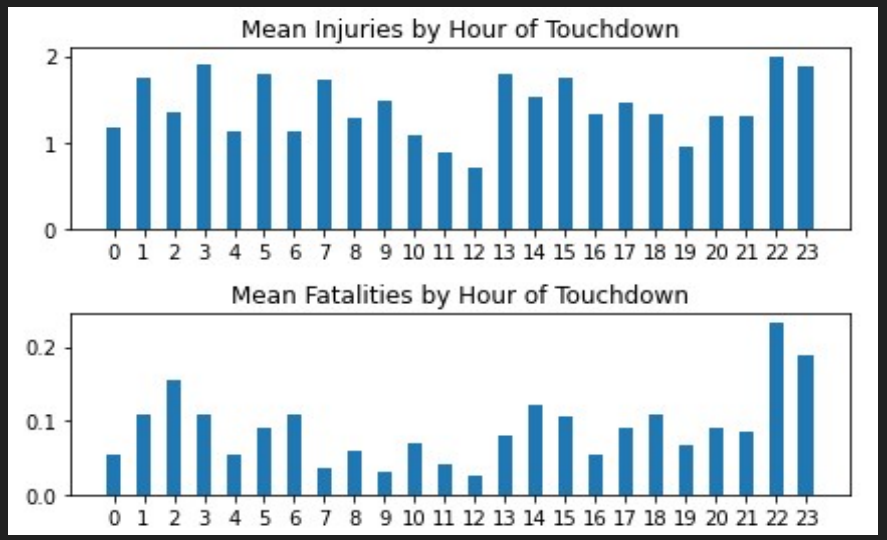




Exploratory Data Analysis

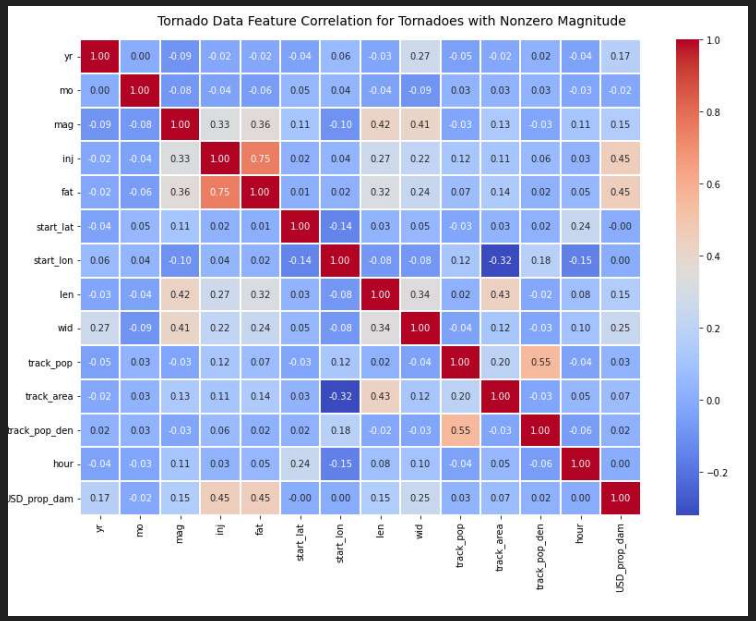
Do tornadoes tend to hit at a certain time of day? Do overnight tornadoes produce more injuries/deaths?





Exploratory Data Analysis

Feature Correlation Heatmap



Preprocessing

Target variable: Injuries, Deaths, or a combination?

- Harm – "Either/Or" blend of the two
  - Injuries OR Deaths OR Both — Harm = 1
  - Neither Injuries OR Deaths — Harm = 0

ML Modeling and Analysis

Final list of predictor variables:

- yr: The year in which the tornado occurred.
- mo: The month in which the tornado occurred.
- mag: Tornado's Magnitude on the EF-scale. This ranges from 1-5, with 5 being the largest and most destructive.
- len: The length of the tornado's on-ground track, measured in linear miles.
- wid: The width of the tornado funnel, measured in linear yards.
- no_counties: The number of counties through which the tornado traveled.
- track_pop: The total population for all counties in the tornado's track.
- track_area: The total land-area for all counties in the tornado's track.
- track_pop_den: The combined population density for all of the counties in the tornado's track.
- USD_prop_dam: The rough US-dollar value of property damage caused by the tornado. This measure is tiered.
- hour: The hour of day in which the tornado first touched down.
- start_lat: The latitude coordinate (in degrees) for the tornado's origination point.
- start_lon: The longitude coordinate (in degrees) for the tornado's origination point.

ML Modeling and Analysis

Supervised Learning Models

- Logistic Regression
- Decision Tree
- K-Nearest Neighbor (KNN)
- Support vector machine (SVM)
- Random Forest
- Gradient Boost

Hyperparameter Tuning

- RandomSearchCV
- 5-Fold Cross-Validation
- 100 trials

ML Modeling and Analysis

Final Preparation Steps

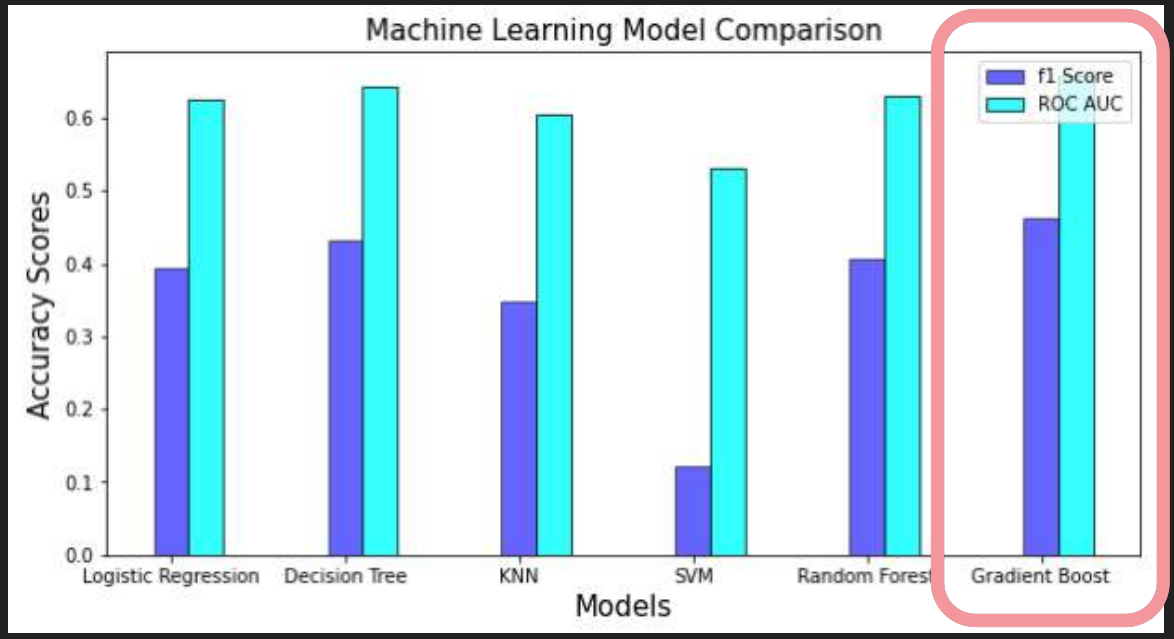
- Scaling the data with Scikit-Learn's Standard Scaler
- Splitting the data into Train/Test sets (80/20 Train/Test Split)

Model Comparison Metrics

- Accuracy not appropriate for unbalanced data
- Used instead **f1-Score** and **Area Under ROC Curve**

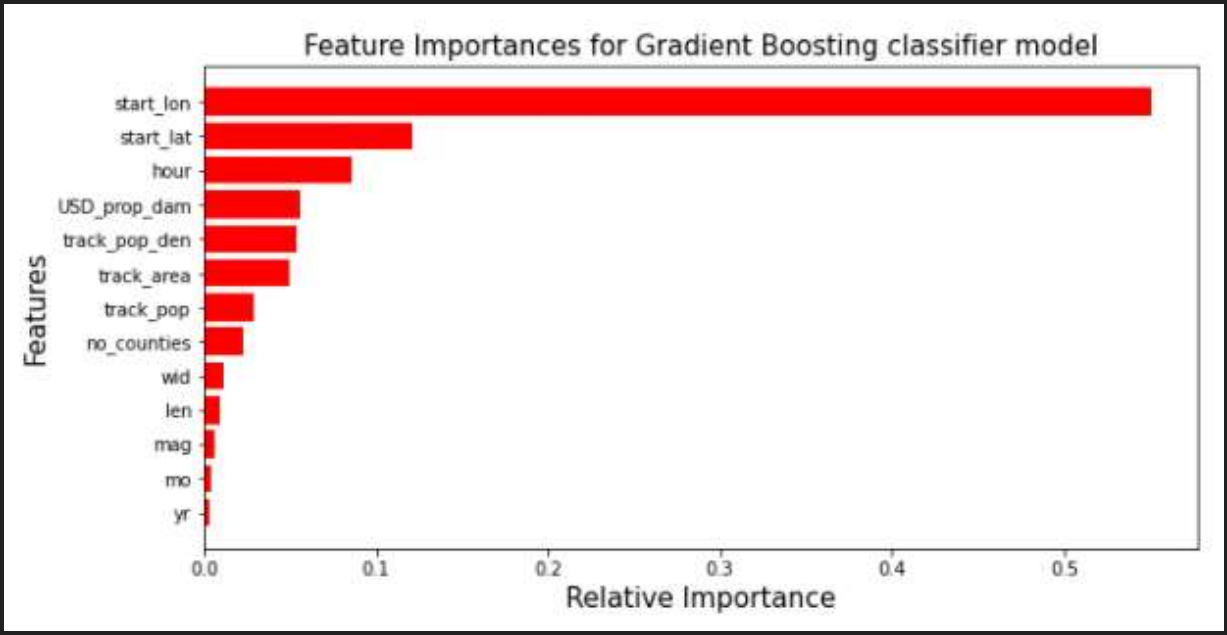
ML Modeling and Analysis

Results Comparison



ML Modeling and Analysis

Feature Importance for GB Model



Conclusion and Next Steps

Result: Strong model with noteworthy insights. But there is much room for improvement.

Opportunities for strengthening model by adding data:

- Historical climate data by county
- Home age data by county
- Long-term weather patterns (El Niño/La Niña)
- Data on prevalence of basements in homes