Wildfire Structural Damage Prediction Executive Summary (Erdös Institute Data Science Bootcamp Spring 2025)

Team Members: Kevin Specht, Evan Ferguson, Andres Barei, Ruichen Kong, & Kiana Burton
Github: https://github.com/kbspecht/erdos-2025-wildfire-structural-damage

Overview:

For this project, we used data science to investigate the correlation between the structural features of various buildings impacted by wildfires in California since 2013 and whether or not those buildings are likely to be destroyed in the wildfires. Concisely, we developed a binary classification model to predict whether a structure impacted by a wildfire will be destroyed or not. The hope is that this will help determine both the level of danger those buildings face from wildfires from a structural perspective and the factors that most often lead to their structures being damaged in the fire.

Stakeholders:

- City planners, building inspectors, and other government workers determining how to make buildings and cities more durable against wildfires.
- Real estate/insurance agents appraising a building's value in the context of wildfire risk.
- Homeowners/renters determining their building's level of safety against wildfires.
- Emergency workers determining how wildfires are likely to impact the zones they're operating in.

KPI (Key Performance Indicator):

- Accuracy of predictions when classifying whether buildings are/aren't destroyed by wildfires.
- Precision, recall and f1 scores for the classification were also recorded.

Data Collection/Cleaning:

- Our source for wildfire damage data was the CAL FIRE Damage Inspection Program (DINS) database of structures damaged and destroyed by wildfires in California since 2013 by CAL FIRE and partnering agencies.
- We transformed the Damage column into a “Destroyed” column (a column consisting of two categories, Destroyed for Destroyed damage level & Not Destroyed for No Damage/Affected/Minor/Major damage levels) and transformed the Year Built column into an “Age” column.
- We excluded pre-2018 data since houses without damage were only logged from 2018 onwards (meaning including pre-2018 ones could bias the data).
- We removed various irrelevant or redundant features (ID, non-latitude/ longitude position information, incident names, features with mostly missing data, features with repeated information, etc).
- We fixed or removed observations with mislabeled or irrelevant (ex. Damage caused by earthquakes) data.
- We transformed some features ex. Damage Level became Destroyed or Not Destroyed, Year Built became Age, etc.

- We removed observations with missing values (future models could look into imputation or other methods for handling them).

EDA (Exploratory Data Analysis):

- Plotting histograms revealed possible correlations between most variables and target (no high impact from any single variable).
- K-Fold cross-validation used to highlight best features to use for different models.

Models:

- Baseline (Most Frequent Class)
- K-Nearest Neighbors Classifier
- Logistic Regression Classifier
- Categorical Naive Bayes Classifier
- Decision Tree Classifier
- Gradient Boosting Classifier
- Random Forest Classifier
- Extra Trees Classifier

Results:

- All models showed some improvement over baseline in classifying destruction.
- Ensemble methods (gradient boosting, etc.) outperformed regular methods.
- Most important predictive features were latitude, structure type, & exterior siding.
- Data does have small imbalances in Destroyed/Not Destroyed observations (40%/60%), dropping observations with missing values may introduce some bias, margin of error in observations due to fire damage/poor access/other factors, incident bias (ex. LA fires).

Future Work:

One fairly straightforward improvement we could make would be to collect more data in order to decrease bias caused by the missing values we had to remove or impute. We could also try additional models, imputation techniques (for the missing data), and data sources (geographical features of structure locations, additional sources of building or wildfire data etc.). We could also perform additional hyperparameter tuning and variable selection on our existing models to see if it improves performance across different KPIs.

With more data, we could also try to implement our original idea, which was to implement an ordinal multiclass classification of various damage levels for buildings affected by wildfires. Currently the dataset is heavily unbalanced with very few observations indicating a damage different from No Damage or Destroyed.

Acknowledgements:

We would like to thank Hatice Mutlu for her mentorship and feedback over the course of the project. We would also like to thank Steven Gubkin, Alec Clott, Roman Holowinsky, and the rest of the Erdös Institute staff for their instruction and support over the course of this bootcamp.