California Wildfire Predictor

By: Aanvi Goel

ABSTRACT

The goal of this project was to build a Linear Regression model to predict the size of wildfire for occurrences in the state of California based on weather conditions, in order to facilitate wildfire containment by the USDA Forest Services and local fire departments. I worked with historic data available from USDA Forest Services to extract past fire occurrence reports and scraped the weather data for each occurrence location and date combination. Using Feature Engineering, Linear Regression and Regularization techniques, I improved the regression models by reducing the R2 score.

DESIGN

Nearly every year, California experiences major wildfires that can be devastating to the environment and communities throughout the state. Just in 2021 there were 5267 wildfires started across the state of California, burning 204,921 acres of land as per stats from CAL FIRE.

The data has been obtained from USDA Forest Services, which keeps records of all reported Fire incidents across US counties along with containment information including the **Fire Size**. Creating a Linear Regression model for predicting the size of wildfire would help the fire departments and the Forest Services in deploying appropriate resources right at the time of fire discovery, leading to quicker containment times and reduced environmental and infrastructural damage.

DATA

Wildfire occurrence data:

- ❖ United States wildfire data: The database file has all the wildfire reported occurrences from the year 1992-2018
- ❖ The SQLite database file was loaded to Jupyter notebook and converted to a Dataframe. The Dataframe was further filtered for State = “CA” (California) and for Fire_Year = 2018
- ❖ Final *ca_fire_df* has 5566 rows and 11 columns. **Column: “FIRE_SIZE”** has been used as the target/dependent variable for the project

Weather historic data (Web Scraped):

- Weather data: The site is web scarped using Selenuim and BeautifulSoup to obtain the weather features (such as Temp_max, Precipitation, Cloud_cover, Wind_speed etc) or each location and date combination from *ca_fire_df*
- Final *weather_df* has 3830 rows and 13 columns

Final Dataframe:

- The above two Dataframes are combined to get *combined_data_df.* After removing duplicates and dropping some irrelevant columns we get the final df with **3657 fire occurrence points x 19 columns (18 Features, 1 Target)**

ALGORITHM

Feature Engineering:

- Converted categorical feature: NWCG_CAUSE_CLASSIFICATION, to encoded numerical values

Y Tranformations:

- Log, SQRT and BoxCox transformations are evaluated for improving R2 value to reduce the heavy right skew of dependent variable (FIRE_SIZE) and to handle non-normal residual distribution

Polynomial Transformation:

- Dependent features are transformed using PolynomialFeatures() to better fit for non-linear relationships between independent features and dependent target

Regularization:

- Lasso, Ridge and ElasticNet models are evaluated for reducing the variance in the baseline linear regression model

Model Evaluation and Selection:

- KFold cross validation is used to validate the trained models across 5 diverse dataset folds
- R2 score is evaluated and used as the key performance metric to reduce the variance in the regression model
- Mean absolute error, Mean squared error and RMSE are used for comparing accuracy of models

TOOLS

1) Data Acquisition -
    a) SQL and sqlite3 module in Pandas for querying from DB
    b) Selenium and BeautifulSoup for Web Scraping
    c) Pickle module to save the parsed data serially
2) Pandas and Numpy for data cleaning and manipulation
3) Matplotlib and Seaborn for data plotting
4) Statsmodels and Sklearn for modeling and testing

COMMUNICATION

A slidedeck and Jupyter notebook code are included along with this write-up as part of the project on Github: https://github.com/aanvigoel/Metis/tree/main/LRWS