Applied Competitive Lab in Data Science - Project Report

Ma'abada & Mika

Submitted by

Dor "Wildfire Whisperer" Berenstein

Ilan "Inferno Insight" Vysokovsky

Mika "Pyro Predictor" Rokach Cohen

Natalie "Flare Finder" Gal

Introduction

Our mission in this project was to develop an accurate and interpretable model to predict the causes of wildfires in the USA. We worked as a team of four, split into pairs. One pair focused on the data preprocessing and feature engineering aspects, while the other pair concentrated on the model development and optimization.

By splitting the work of the data preprocessing and model development to two teams, we were able to deliver an improved analysis of the wildfire cause prediction problem. The collaborative nature of this project allowed us to dive more thoroughly into each part of the project, ultimately leading to a more comprehensive understanding of the wildfire causes.

Data cleaning, Feature Extraction and Engineering

Feature Selection

As a first step, we dropped features that were found to be irrelevant for the prediction task. This was done to focus the data and the models on the most important factors influencing the target variable, the fire cause.

We removed features that were either ID-related, had too many null values, or were suspected to cause data leakage. This process helped prepare the data for the prediction task.

Feature Classification and Engineering

Next, we classified the remaining features into categorical, ordinal, and continuous types. This understanding of feature types allowed us to apply appropriate transformations and encoding techniques.

Following literature on wild fires, we created new features to capture relevant information, such as:

- Holidays when fires are more likely to occur.
- Seasonal information (Spring, Summer, Fall, Winter).
- Climate zones based on the Köppen-Geiger climate classification system.
- Reporting source, whether by a nature-related unit or not.
- Land ownership (US owned, private, tribal, etc.)

These engineered features were intended to ease the model's prediction task and capture potentially important factors influencing fire causes.

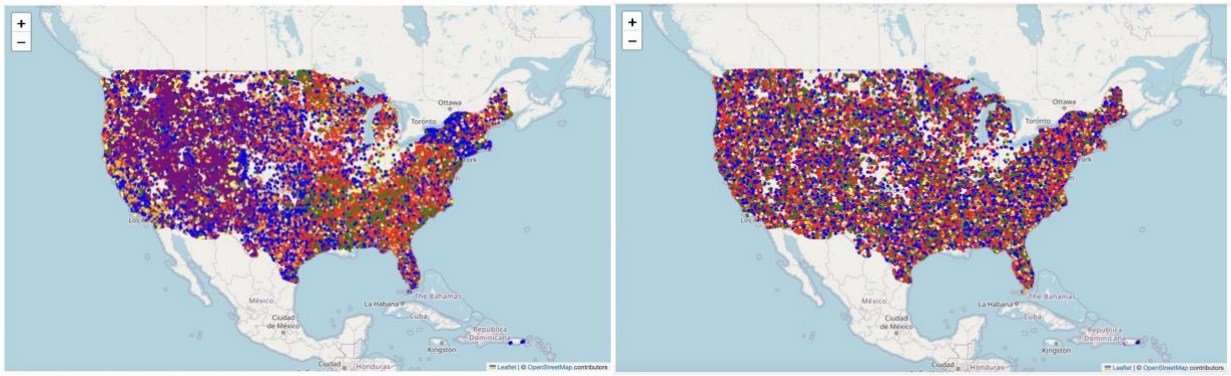
In our initial exploration, we considered the 4th of July holiday as a potential feature, hypothesizing that it might be associated with increased fire risk due to activities like fireworks. However, after evaluating the importance of this feature in the prediction task, we found that using only this single holiday did not provide sufficient data to meaningfully capture temporal patterns in the fire causes. To address this, we expanded the holiday feature to include additional key dates, such as Memorial Day, Labor Day, Halloween, and New Year's Eve. By incorporating these various holidays, we aimed to reflect the changes in human behavior during holidays that could influence wildfire occurrences.

In addition to the holiday feature, we also sought to represent the geographic diversity of the United States in a more nuanced way than simply using state boundaries. Recognizing that environmental conditions play a crucial role in fire risk, we leveraged the Köppen-Geiger climate classification system to categorize the states into distinct climate zones. This allowed us to investigate how environmental factors, beyond just location, might contribute to the prediction of fire causes across the country.

Additionally, there was discussion surrounding the potential impact of the land ownership feature on the prediction task. It was suggested that land ownership could serve as a valuable indicator of area characteristics. After going through the prediction task, the land ownership features showed a high significance for the prediction.

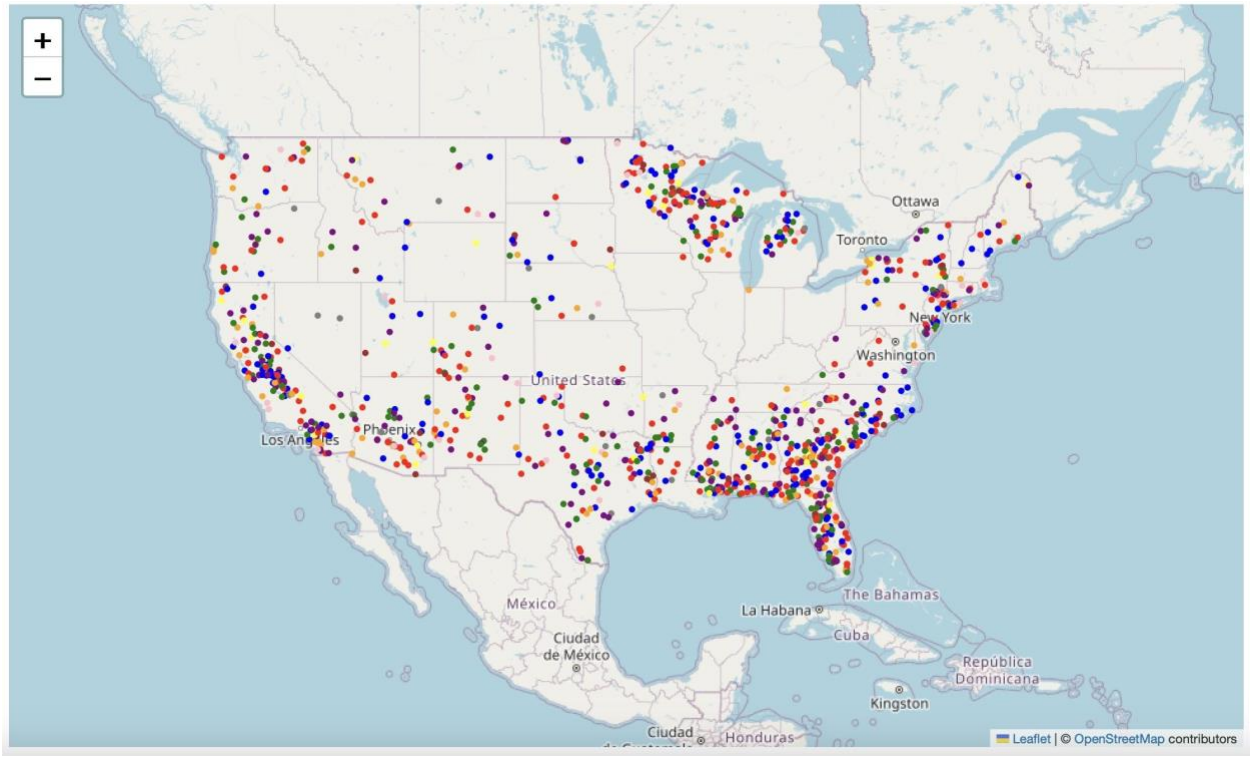
Exploratory Data Analysis

To gain insights into the data, we performed several visualizations and analyses:



We plotted the fire events on a map of the United States, coloring them by the fire cause. On the left we see the complete data. On the right is the Non-Holiday spatial representation of the data.

The complete data representation shows the dominance of lightning-caused fires in the West and debris burning in other regions. The visualization of non-holiday data surprisingly shows a uniform distribution of fire causes across the nation.



Holidays spatial representation: The holiday data shows a higher concentration of fires in densely populated areas like California, New York, and Florida.

Data Rescaling

Finally, we rescaled the numeric features. This helped ensure that the features were on a similar scale, which is important for tasks like correlation analysis and model training.

Model Development and Optimization

Baseline Model Evaluation

We first trained three baseline models on the fire cause dataset. Two of the models - Random Forest and Decision Tree, were chosen for their inherent interpretability, allowing us to analyze the feature importance in predicting fire causes. The third model, XGBoost, was selected for its powerful capabilities and potential for higher accuracy.

We evaluated the baseline models using the requested metric - Weighted ROC AUC Score. We also used two other metrics that are suitable for evaluating multi-class models prediction.
The best performing baseline model was the Random Forest. The Decision Tree had the lowest score for the baseline.

Hyperparameter Optimization

To further improve the models' performance, we performed hyperparameter tuning. This process aimed to find the optimal configuration of hyperparameters for each model, such as the number of estimators, tree depth, regularization, and learning rate.

For the Random Forest and Decision Tree models, we leveraged the Optuna library to optimize the hyperparameters. This allowed us to efficiently explore the parameter space and suggest new configurations based on the model's performance.

In the case of the XGBoost model, we could not directly use Optuna as it is not natively supported by models that are not part of the sci-kit learn library. Instead, we opted for a RandomizedSearchCV approach, which allowed us to randomly sample from a defined grid of hyperparameters. We ran the Random Search Process multiple times. Each time narrowing down the parameter grid to focus on the most promising ranges. In order to improve run-time of the XGBoost optimization, we used Google Colab with GPU. It led us to explore different options to run our jupyter notebook, and how to tune the configurations of the model to utilize the GPU.

Model Evaluation and Comparison

After finding the best-performing hyperparameters, we evaluated the models using a comprehensive set of metrics. This analysis provided a more detailed understanding of the models' strengths and weaknesses in handling the class imbalance problem.

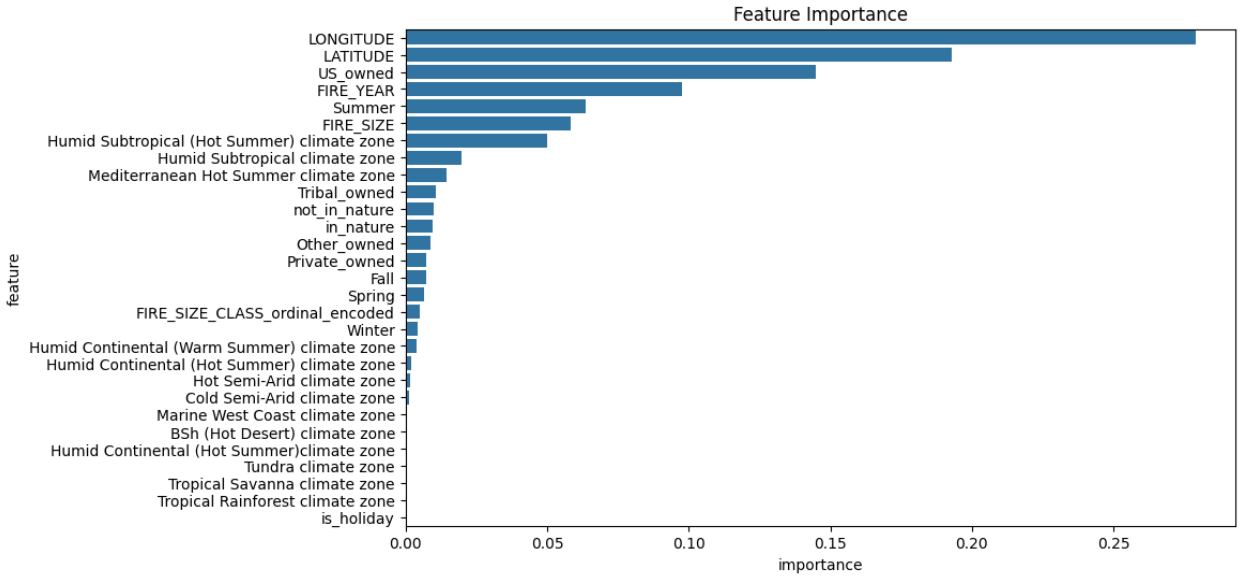
The top-performing model was the Random Forest. It demonstrated the strongest overall performance, predicting the class with the highest instance count and showing decent performance for some of the more prevalent classes. The Decision Tree and XGBoost models showed a lower performance compared to the Random Forest, with varying degrees of success in predicting the different classes. We also noted that all three models struggled to accurately classify instances of the less frequent classes.

All three models showed improved performance after the hyperparameter tuning process. Especially the Decision Tree, which had the most notable improvement compared to the baseline model.

Feature Importance Analysis

To better understand the key factors influencing the models' predictions, we analyzed the feature importance scores. This analysis was made possible due to the interpretable models we chose.

Feature importance for the Random Forest Model is given in the following graph:



The analysis reveales that location and timing features were the most critical factors across all models, as expected. While some features maintained high importance across all the models, there were also notable differences in the specific features each model prioritized for prediction. For instance, the models diverged in their emphasis on features such as land ownership type and environmental conditions.

Addressing Class Imbalance

To address class imbalance, we experimented with assigning lower weights to the most frequent classes during model training. This approach encouraged the models to be more cautious when predicting these classes, leading to slight improvements in performance metrics.

Conclusion

Our collaborative investigation into wildfire cause prediction led to a deeper understanding of machine learning concepts and techniques. We dived into both the pre-processing steps and the model's development and optimization, while gaining knowledge in the domain of wildfires and their causes in the US.

The ongoing work during the semester helped us reach better insights about the data and the domain. One of which was the feature engineering process, which proved pivotal, as the addition of temporal, geographic, and environmental variables significantly enhanced the models' ability to capture the nuances of wildfire causes. The inclusion of holiday information, climate zone classifications, and indicators of land ownership type allowed us to uncover patterns that would have been missed by relying solely on the baseline dataset.

While the models exhibited varying degrees of success in accurately predicting fire causes, particularly in the face of class imbalance, the overall findings provide a solid foundation for future improvements. The feature importance analysis revealed the critical role of location, timing, and environmental conditions on the cause of fire. We were surprised to some extent that a relatively simple model such as the Random Forest proved to be the best out of the ones we explored.

Moving forward, potential areas for improvement include exploring alternative additional data sources, implementing more advanced techniques for handling class

imbalance, and investigating the combination of multiple models or ensemble methods to leverage their complementary strengths. By continuously refining the predictive models and incorporating new data and methodologies, we can work towards a more accurate and comprehensive understanding of the complex factors underlying wildfire causes.