Red Hot

Team 152: Sierra Pitman, Christopher Holmes, Victor Tchervenobrejki, and Gabriel Glazer

1. Introduction

In this research project, we aim to evaluate the probability of a wildfire occurring in the state of Georgia and to visualize the results with a Tableau choropleth map. To achieve this, we will build predictive algorithms in Python, which use publicly available meteorological data (API-based) and wildfire data from an SQLite database. The significance of this research is understood by the $18 billion (about $55 per person in the US) in losses attributed to wildfire damage in 2021[1]. These damages include deteriorating air quality, respiratory, and cardiovascular effects [3][4]. Leveraging big data analytics provides an opportunity to mitigate these costs, as demonstrated by *Yu et al. (2018),* who offer insights into big data's applications in disaster management and *Taylor et al. (2013),* who emphasize the importance of statistical approaches in predicting wildfires. Our research primarily targets the mitigation stage of disaster management following *Altay, et al. (2006)* framework, where the stages of disaster management are defined as mitigation, preparedness, response, and recovery. If successful, authorities can use our model to proactively allocate firefighting resources to high-risk areas or alert communities of fire dangers in advance. The benefits could be quantified in reduced casualties, property damage, and air pollution. Deployment in a region would represent a natural experiment, so the impact could be evaluated through a difference-in-difference estimate.

2. Formal problem definition

Our aim is to develop xgboost and random forest classification algorithms that will use publicly available weather forecast data from open-meteo.com, to cross-sectionally evaluate future fire probabilities within the state of Georgia. The granularity of the dataset is at the city-day level. The map GUI will incorporate interactive elements to switch between regions and future days.

3. Related Works

Since 1972, the United States has utilized the National Fire Danger Rating System, which "estimates today's and tomorrow's fire danger for a given area [5][6]." This model uses "fuels, weather, topography and risk [6]" to predict fire danger but does not capture geo-temporal context.

Linardos et al. (2022)'s systematic literature review quantifies the current research landscape of machine learning within disaster management. Based on their results, only 9.09% of relevant sources addressed "disaster and hazard prediction" and only 2.9% of relevant sources addressed wildfires.

Jain er al. (2020) provides an overview of the machine learning algorithms that are prevalent in wildfire science and management literature. From their systematic literature review, only 6% of applicable journal papers focused on fire weather.

Malik et al. (2021) Utilizes data parameters uncommon in other wildfire prediction models, including powerlines, terrain, and vegetation. They use an ensemble of algorithms to increase robustness: Adaboost, Decision trees, Gradient descent, Multi-layered perceptron, Random Forest Tree (RF), and Long Short-Term Memory (LSTM).

Sakr et al. (2010) presents an SVM-based approach to predicting a day's fire hazard level based on historical weather data, including daily average humidity. They aim to reduce the number of features, to eliminate sources of erroneous error.

Stojanova et al. (2012) presents a methodology that employs both single and ensemble algorithms. They consider numerous factors including weather conditions, historical fire data, and GIS data.

Sakr et al. (2011) sought to create an economically viable fire prediction system for developing nations. The authors prioritized feature selection, narrowing their research to relative humidity and cumulative annual precipitation. Their research was intentionally confined to artificial neural networks and support vector machines.

Singh et al. (2021) pursued a more technical approach to their fire prediction analysis, aiming for computational efficiency using parallel computing and illustrating their use of Hadoop, Python, Spark, and PySpark. Their research centers on the implementation of parallel SVM models that simultaneously process weather data.

Crimmins (2006) Used a weather typing model to associate different weather patterns with fire likelihood. They were able to identify a few factors that strongly correlated to previous high fire days.

4. Proposed Method

4.1 Using KDTree to enable big data pulls in view of API restrictions.

Our first data source is an SQLite database from Kaggle that has 1.88 million geo referenced wildfire records across the US, between 1992 to 2015. This data detailed the precise latitude and longitude of each fire's origin. This translated to 111,672 distinct fire geo-locations for the 660 cities in Georgia. Weather data was obtained from open-meteo.com which permits API calls for individual geo-locations but restricts users to 3 requests per hour. To overcome this challenge, we also procured a dataset from simplemaps.com that associates each US city with a unique set of coordinates. By employing k-dimensional tree based nearest neighbor searches in Python, we matched these coordinates with the wildfire data. This reduced our API call requirement for weather data from 111,672 requests to a more manageable 660.

4.2 Dealing with dataset imbalance.

Dataset imbalance can significantly impact the performance of predictive models. In our dataset, out of 3.8 million observations, only 100k represented fire incidents in Georgia. To address this, we employed under sampling on the majority class (non-fire days) to achieve a 1-to-2 ratio between fire and non-fire days. This ensured that our models were exposed to a balanced representation, allowing for a more unbiased prediction of fire occurrences.

4.3 Leveraging the superiority of ensemble algorithms.

XGBoost is an advanced implementation of gradient boosting algorithm. What sets it apart from previously used algorithms (e.g., SVN, AdaBoost) is its computational efficiency and parallel processing, which is paramount given our dataset size. Moreover, we utilized its regularized framework which performed variable selection and reduced overfitting. We also used its built-in feature importance, to evaluate the impact of various metrics on the likelihood of fires. On the other hand, random forest is renowned for its simplicity and interpretability. By comparing the out-of-sample AUC of both the xgboost and the random forest, we aim to select the best model ensuring the highest predictive accuracy.

4.4 Hyperparameter tuning with Bayesian Optimization and k-fold cross-validation.

To optimize model performance, we integrated Bayesian Optimization for hyperparameter tuning. This method uses probability to optimize an objective function, making it more computationally efficient than grid search or random search. In our specific case, the objective function maximizes the out-of-sample ROC AUC based on 5-fold cross-validation. The cross-validation ensures that our model is robust and is not prone to overfitting. The ROC AUC metric provides insights into the model's discriminative power between fire and non-fire days. It incorporates the TPR and FPR which are both important to our project. A high TPR would ensure accurate fire resource allocation, while a low FPR would minimize resource waste.

The specific hyperparameters that we fine-tuned for the xgboost model include: max_depth, gamma, colsample_bytree, learning_rate, n_estimators, reg_alpha, and reg_lambda. Each of them plays a crucial role in how the model learns. For instance, max_depth controls the depth of the trees, affecting the model's complexity, while learning_rate impacts how quickly the model adapts to errors. Variable selection was performed through the alpha and lambda regularization hyperparameters. Similarly, the hyperparameters that we tuned for the random forest model include: n_estimators, max_depth, min_samples_split, and min_samples_leaf. A higher number of trees (n_estimators) typically lead to a better-performing and more robust model, at the expense of a higher computational cost. Deeper trees (max_depth) can capture more intricate data relationships at the expense of overfitting. The 'min_samples_split' and 'min_samples_leaf' parameters govern the minimum number of samples required to split a node, which prevents overfitting.

4.5 Visualizing the results in Tableau.

Leveraging the visualization power of Tableau, we built dynamic choropleth maps that reflect our forecasts across different days, counties and cities. Upon importing our processed data, Tableau identifies geographical fields and plots the results from our models. By representing wildfire risk through color gradients, we will offer a visual distinction between the various levels of risk across the cities. To enhance user interaction, we will add interactive elements for location-based filtering and forecast day selection.

5. Summary of innovations

The method described above provides multiple innovations, improving on previous research. Here is a summary:

- *Sakr et al. (2011)* performs variable selection based on a subjective evaluation of measurability and association with the dependent variable, potentially resulting in a suboptimal covariate subset. In our methodology, we plan to mitigate this through elastic net regularization.
- The SVM models developed *(Sakr et al. (2010), Sakr et al. (2011) and Singh et al. (2021)*) represent single models outside an ensemble. These do not capture non-linear functional forms. Thus, we are using xgboost and random forest instead.
- *Malik et al. (2021)* and *Stojanova et al. (2012)* use adaboost algotythms which do not allow for regularization and hyperparameter flexibility. Both of these steps are incorporated with our xgboost algorithm.
- By limiting the number of latitude and longitude locations through KDTree, we allow for data processing of a much wider geographic area and time-horizon.

6. Experiments & Evaluation: Questions our experiments are designed to answer (description of testbed):

Q1: Which model performs better in predicting wildfires – XGBoost or Random Forest?

Understanding the intricacies of wildfire prediction necessitates a robust and meticulous evaluation of different predictive algorithms. The selected model will form the backbone of our predictive analytics.

Q2: Is the model accurate enough to be used by public institutions?

Bearing in mind the immense consequences of wildfires, it is important to employ a holistic set of metrics that would indicate the model's real-world relevance and generalizability.

Q3: What are the key factors that influence wildfire risk?

Understanding the feature importance is crucial for several reasons. First, identifying the most influential features can guide future research which aims to optimize similar models. Second, understanding key factors can inform policy decisions.

7. Experiments and Evaluation: Detailed description of experiments and observations

7.1 Choosing the best model

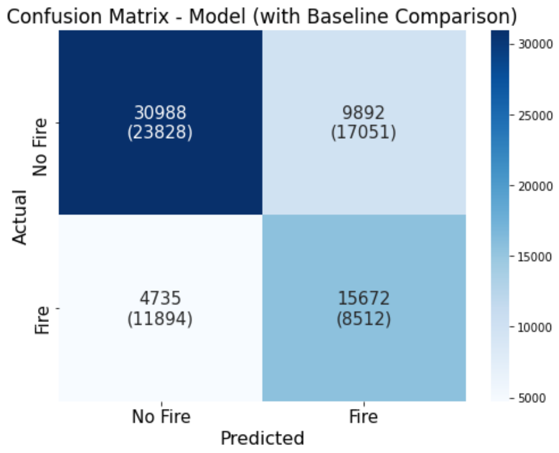
Our first objective was to determine the best performing model for predicting wildfires, between XGboost and Random Forest. After performing Bayesian optimization to tune hyperparameters of both models based on cross-validation, their ROC AUC value was compared on a test set. Their performance was nearly identical, with the xgboost model having a slight edge: 0.845 versus 0.835.

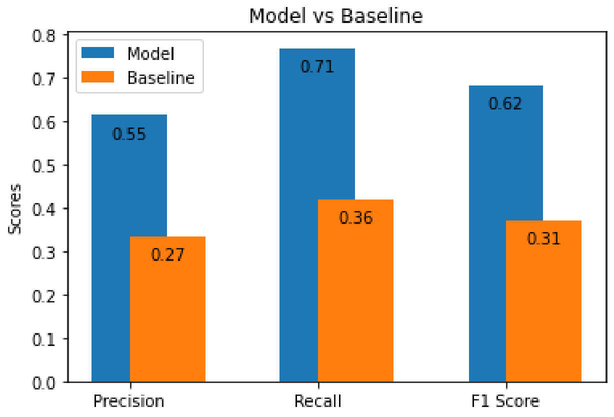
7.2 Model accuracy

To assess the performance of our wildfire prediction model, we used a 20% test-split and compared our model with a baseline model using the following evaluation metrics:

- Logloss: Quantifies how close our predicted probabilities are to the actual occurrences. This ensures a model that is not just correct in classification, but also confident in its predictions on an interval scale.
- **ROC AUC:** Gauges the model's ability in differentiating between fire and non-fire days.
- Precision: Falsely predicting a fire would lead to resource waste, making resources scarce at locations where they are required. This metric ascertains the proportion of predicted fire events that were actual fires.
- Recall: Equally important, missing a real wildfire event can have dire consequences. Recall measures how many of the actual fire days were successfully predicted by the model.
- **F1 Score:** Combining precision and recall, the F1 score serves as a harmonic mean between the two. Since both false negatives and false positives carry significant weight, F1 is important.

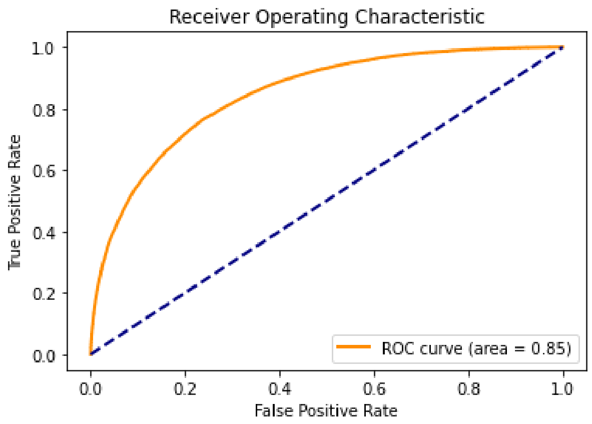
First, model performance was evaluated at a fixed threshold level that determines the point at which probabilities are classified as positive (fire) or negative (non-fire) days. It was selected to maximize the f1 score from a range of 100 potential thresholds, spanning from 0% to 100% in 1% increments. The baseline model in this approach is a random classifier, of which the number of positive/negative predictions match the number of positive/negative predictions of the actual model. The following 2 plots show the results:

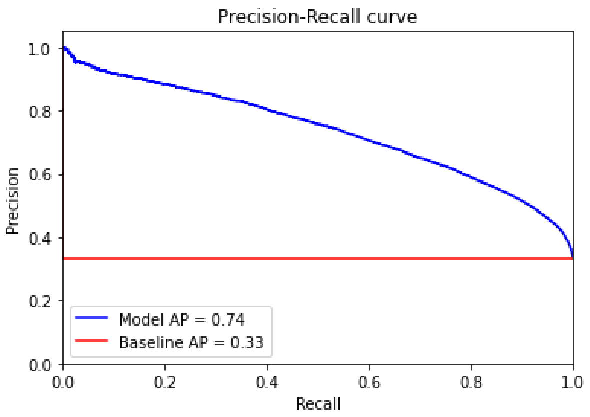




The confusion matrix shows that our model significantly outperforms the baseline in accurately identifying both true positives (fire days) and true negatives (non-fire days). Consequently, both false positives and false negatives of the model are substantially lower than the baseline. The accompanying bar chart reinforces these findings. The model's precision, recall and f1 scores are about twice as high as the corresponding baseline values.

Our second approach was to evaluate the model without fixing the threshold level. In this approach, the baseline model produces a constant probability prediction, equal to the average positive rate in the balanced dataset. The following two plots illustrate the results:

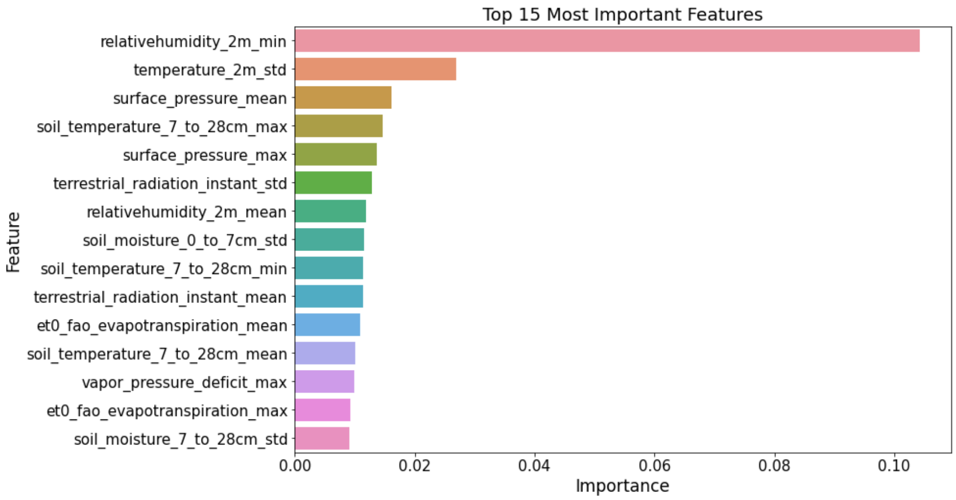




The ROC curve exhibits a high degree of concavity with an AUC equal to 0.85. The precision-recall curve shows that precision is much higher than baseline, at any fixed recall level. Moreover, our model produced a logloss of 0.64 compared to a baseline of 0.46. These results further accentuate the superiority of our model over a baseline.

7.3 Feature importance

Hourly data was aggregated by day, to produce a structured dataset with observations for each city and each day between 2000-01-01 and 2015-12-31. The final weather dataset includes 178 covariates and 3.8 million rows. It encompasses variables related to temperature, humidity, dewpoint, precipitation, snowfall, cloud cover, visibility, evapotranspiration, wind metrics, soil conditions, UV index, and solar radiation. Below are the top 15 features as according to the feature importance metrics employed:



The most significant factors of wildfire risk include relative humidity 2 meters above the ground, temperature and soil-temperature related features, surface pressure, evapotranspiration and terrestrial radiation. These are indicative of the complex interplay between atmospheric and soil conditions in determining wildfire risk and portray the importance of a holistic approach in wildfire prediction.

8. Conclusions and discussion

Our research contributes to the field of wildfire prediction by leveraging advanced machine learning techniques and large-scale data analytics. By integrating multiple disparate data sources into our algorithms, we have established a robust model that efficiently, accurately, and fairly predicts wildfire occurrences.

Our model performed approximately twice as good as a baseline model, according to metrics such as: logloss, precision, recall, f1, accuracy and auc. Such performance indicates that our product could potentially help guide efficient resource allocation and alert communities on the event of high risk. It also highlights the potential of ensemble algorithms such as xgboost and random forest in enhancing wildfire management strategies.

Our findings regarding future importance can guide government policy. Given the importance of soil moisture and temperature, the government can implement management policies that help maintain optimal soil conditions. Activities related to minimizing soil disturbance, maximizing soil cover or expanding biodiversity could also be addressed and incentivized.

We aimed to make the product visually intuitive, compelling and understandable for users through the dynamic choropleth map in Tableau. An important area for future improvement lies in conducting user studies to verify whether this was achieved. These studies will evaluate the usability and clarity of our final product, ensuring that it effectively communicates its intent to its users.

Implementing KDTree paid off by making our data processing more efficient, enabling us to work around the API limitations. However, one key limitation of our study is the geographical scope of our dataset, which is confined to the state of Georgia. This limitation was dictated by practical constraints. Given the limitations of 3 API requests per hour, pulling data for all 30,845 cities in all states using 4 computers would take around 107 days (about 3 and a half months). This was outside the time-scope of the current project.

Looking ahead, future work should explore more computationally efficient, economically viable and streamlined methods of data collection that would allow live streaming and establishing CI/CD pipelines. Nevertheless, our research lays a solid foundation for future endeavors in this domain and offers a scalable framework that can be adapted and expanded upon. Most of the issues can be tackled with a budget for massive extraction using an API. The ETL pipeline would have to change to create different models per state. While the map in Tableau would have to be adjusted as well, future work could tailor the various filters and views for different organizations’ requirements.

All team members contributed an equal amount to the project.

References

1. Zandt, F. (2023), How Much Damage Do Wildfires Cause? [Webpage]. Statista. https://www.statista.com/chart/30602/estimated-losses-caused-by-wildfires-heat-and-drought-in-the-us/#:~:text=Natural%20disasters&text=Natural%20catastrophes%20often%20connected%20to,providers%20Aon%20and%20Munich%20Re
2. Martin, S. (2023). 2023 U.S. Wildfire Statistics. Bankrate. https://www.bankrate.com/insurance/homeowners-insurance/wildfire-statistics/
3. Editorial Staff, "How Wildfires Affect Our Health," https://www.lung.org/blog/how-wildfires-affect-health
4. United States Environmental Protection Agency, "Health Effects Attributed to Wildfire Smoke," https://www.epa.gov/wildfire-smoke-course/health-effects-attributed-wildfire-smoke
5. National Fire Danger Rating System, https://www.fs.usda.gov/detail/cibola/landmanagement/resourcemanagement/?cid=stelprdb5368839
6. NWCG Fire Danger Working Team (Compiled by Schlobohm, P. & Brian, J.).(n.d.). Gaining an Understanding of the National Fire Danger Rating System. [PDF Document]. https://gacc.nifc.gov/rmcc/predictive/nfdrs_gaining_understanding.pdf
7. Linardos, V., Drakaki, M., Tzionas, P., & Karnavas, Y. L. (2022). Machine Learning in Disaster Management: Recent Developments in Methods and Applications. Machine Learning and Knowledge Extraction, *4*(2), 446–473. https://doi.org/10.3390/make4020020
8. Jain, P., Coogan, S.C.P., Subramanian, S.G., Crowley, M., Taylor, S., & Flannigan, M.D. (2020). A review of machine learning applications in wildfire science and management. Environmental Reviews. 28(4): 478-505. https://doi.org/10.1139/er-2020-0019
9. Taylor., S. W. , Woolford, D. G., . Dean., C. B., Martell, D. L. (2013). Wildfire Prediction to Inform Fire Management: Statistical Science Challenges. Statist. Sci. 28(4) 586 – 615. https://doi.org/10.1214/13-STS451
10. Malik, A., Rao, M. R., Puppala, N., Koouri, P., Thota, V. A. K., Liu, Q., Chiao, S., & Gao, J. (2021). Data-Driven Wildfire Risk Prediction in Northern California. *Atmosphere*, *12*(1), 109. https://doi.org/10.3390/atmos12010109
11. Sakr, G. E., Elhajj, I. H., Mitri G., & Wejinya, U. C. (2010). Artificial intelligence for forest fire prediction 2010. In IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Montreal (pp. 1311-1316). Montreal, QC, Canada. doi: https://ieeexplore.ieee.org/document/5695809
12. Stojanova, D., Kobler, A., Ogrinc, P., & et al. (2012). Estimating the risk of fire outbreaks in the natural environment. Data Min Knowledge Discovery 24(3), 411–442. https://doi.org/10.1007/s10618-011-0213-2
13. Yu, M., Yang, C., & Li, Y. (2018). Big Data in Natural Disaster Management: A Review. Geosciences, 8(5), 165. https://doi.org/10.3390/geosciences8050165
14. Taylor, S. W., Alexander, M. E. (2006). Science, technology, and human factors in fire danger rating: the Canadian experience. International Journal of Wildland Fire, 15, 121-135. https://doi.org/10.1071/WF05021
15. Lee, W.; Kim, S.; Lee, Y.-T.; Lee, H.-W.; Choi, M. (2017). Deep Neural Networks for Wildfire Detection with Unmanned Aerial Vehicle. In Proceedings of the 2017 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 8–10 January 2017. https://ieeexplore.ieee.org/document/7889305
16. Altay, N.; Green, W.G. (2006). OR/MS research in disaster operations management. European Journal of Operational Research, 175, 475–493. https://doi.org/10.1016/j.ejor.2005.05.016

17. Crimmins M.A. (2006). Synoptic Climatology of Extreme Fire-Weather Conditions Across the Southwest United States. International Journal of Climatology, 26(8): 1001–1016. https://doi.org/10.1002/joc.1300
18. Sakr, G. E., Elhajj, I. H., Mitri G. (2011). Efficient forest fire occurrence prediction for developing countries using two weather parameters. Engineering Applications of Artificial Intelligence, 24(5), 888-894. https://doi.org/10.1016/j.engappai.2011.02.017
19. Singh, K.R., Neethu, K.P., Madhurekaa, K., Harita, A., & Mohan, P. (2021). Parallel SVM model for forest fire prediction. Soft Computer Letters, 3, 100014. https://doi.org/10.1016/j.socl.2021.100014