A SIMPLIFIED MACHINE LEARNING BASED WILDFIRE IGNITION MODEL FROM INSURANCE PERSPECTIVE

Yaling Liu, Son Le, Yufei Zou, Mojtaba Sadeghi
Kettle Reinsurance
83 Norwood Ave, Kensington
Kensington, California, USA
yaling@ourkettle.com

Yang Chen
Department of Earth System Science
University of California
Irvine, California, USA

Pierre Gentine
Dept. Earth & Environmental Engineering
Columbia University
New York, USA

Niels Andela
Cardiff University
Cardiff CF10 3AT, UK

ABSTRACT

In the context of climate change, wildfires are becoming more frequent, intense, and prolonged in the western US, particularly in California. Wildfires cause catastrophic socio-economic losses and are projected to worsen in the near future. Inaccurate estimates of fire risk put further pressure on wildfire (re)insurance and cause many homes to lose wildfire insurance coverage. Efficient and effective prediction of fire ignition is one step towards better fire risk assessment. Here we present a simplified machine learning-based fire ignition model at yearly scale that is well suited to the use case of one-year term wildfire (re)insurance. Our model yields a recall, precision, and the area under the precision-recall curve of 0.69, 0.86 and 0.81, respectively, for California, and significantly higher values of 0.82, 0.90 and 0.90, respectively, for the populated area, indicating its good performance. In addition, our model feature analysis reveals that power line density, enhanced vegetation index (EVI), vegetation optical depth (VOD), and distance to the wildland-urban interface stand out as the most important features determining ignitions. The framework of this simplified ignition model could easily be applied to other regions or genesis of other perils like hurricane, and it paves the road to a broader and more affordable safety net for homeowners.

1 INTRODUCTION

Wildfires have caused many fatalities and are devastating large regions, with far-reaching impacts on the environment, ecosystems, climate and human health[1]. In California, wildfires have been increasing for decades, driven by climate-change-induced warming and decreases in precipitation[2]. Furthermore, land and fire management have exacerbated the hazards[3]. Finally, population and socioeconomic growth, particularly at the wildland-urban interface, has dramatically increased the exposure of communities to wildfires and expanded the fire niche into regions and seasons with wetter fuels that are less conducive to fire activity[3-6]. The combined result of accelerating wildfire risk is clearly visible in California, which experienced enormously damaging fires in 2017, 2018 and 2020. In 2018 alone, the total damage due to wildfires in California amounted to approximately $148.5 billion (about 1.5% of California's annual GDP) [2].

While insurers are suffering from a surge in homeowner insurance claims for wildfires, reinsurers in California had raised their prices by 600% during 2015-2020 to make up for lost profits and rising risks[7], due to an over self-correction and inadequate assessment of fire risks. As a result, the passed through expenses of reinsurance have made insurance unaffordable for many homeowners in California. These increased premiums may also exacerbate financial pressures on residents. The California Department of Insurance (CDI) reported a 6% increase in insurer-initiated homeowner policy non-renewals between 2017 and 2018[8].

Recent years have seen significant progress in wildfire risk assessment framework [9-12], wherein ignition is a critical component. For ignition component of those frameworks, it usually either randomly sample ignitions within a Fire Planning Unit (FPU) [9, 13], which are organized to address fire suppression over a geographic area and divided by region (e.g., California currently has 21 Fire Planning Units), or just sample ignitions relying on historical observations[11, 12]. Next, all the sampled ignitions, will be fed into a separate spread model for wildfire spread simulations. Afterwards, wildfire simulations will be used to calculate burn probabilities, and then will be combined with exposures and building vulnerabilities to estimate wildfire risks. Nonetheless, the ignition component is not adequately represented within those earlier frameworks. Specifically, randomly sampling within a FPU ignores the spatial heterogeneity of ignition propensity within a particular FPU, and sampling solely based on historical ignition observations may cause no or inadequate sampling in areas without historical ignitions but are still prone to have ignitions in the future. Either of above cases will cause significant biases or uncertainties in ignitions, which will be cascaded to downstream components and will undermine the final wildfire risk assessment. Therefore, a better representation of the ignition component is imperative.

In light of surging wildfire (re)insurance prices which put many homes under non-renewal risk, an accurate estimate of wildfire risks is essential for affordable (re)insurance for homeowners and insurance carriers. In addition, from a (re)insurer's perspective, the risk estimation should be practical in business and agile in production. To meet those pressing and practical needs, here we present a simplified machine learning (ML) based ignition model, which is a critical component of our wildfire risk assessment framework that is currently used for the wildfire (re)insurance in California. We model the ignition probability in a simplified manner at yearly scale, as opposed to daily or monthly scale, mainly for practical reasons. First, a (re)insurance policy is typically written or renewed for a one-year term, so an annual estimate of wildfire risk is good from (re)insurance perspective. Thus, there is no pressing needs to model ignition at monthly or daily scale from this regard. Second, more than 80% of wildfire ignitions in the US are caused by human activities[3], suggesting that fast-changing natural factors such as weather may play a less important role in ignitions. Meanwhile, the anthropogenic factors used in our model (Table A2 in the Appendix), which are used as proxies for human activities, do not change rapidly at seasonal scale (e.g., population density, distance to the wildland-urban interface), thus yearly data would be sufficient to approximate their gradual changes. Third, from operational perspective, building, maintaining, and refining a ML-based yearly scale model is more easy, lightweight and agile for production and operation. In this work, we mainly elaborate how we build the model (Section 2) and how well the model performs in California (Section 3), and we conclude with final takeaways (Section 4).

2 METHODS AND DATA

2.1 METHODS

2.1.1 BASELINE APPROACH

Based on historical ignition observations, we build the baseline model by using frequency of ignition occurrence. Specifically, we first split the training and testing data (i.e., ignition observations) in a temporally continuous manner (see details in Section A1 of the Appendix). If the average ignition frequency during the training stage is ≥ 1 ignition per year within a grid, then during the testing stage it is assumed that grid will continue to have ignition occurrence every year. On the other hand, those grids with ignition frequency < 1 ignition per year during the training stage will be assumed to have no ignition occurrence during testing. The historical frequency of ignition occurrence during the training stage (1992-2014) is shown in Figure A1 of the Appendix.

There are two caveats for the baseline approach. First, the predicted ignitions will be static for the testing period and the future, any inter-annual variability in ignitions will not be captured. This could cause significant bias in wildfire risk assessment, especially when climate change is ever-increasing and is projected to intensity in the future[14]. Second, since this approach is solely based on historical observations of ignition occurrence, there is no way to understand what factors determine ignition occurrence with this approach, which is critical in actual practices of wildfire (re)insurance underwriting.

2.1.2 MACHINE LEARNING APPROACH

To address the two abovementioned caveats of the baseline model, and also to potentially improve the baseline model, the extreme gradient boost (XGBoost[15]) machine learning method is introduced to model the yearly ignition probability across California at a spatial resolution of 0.1-degree. We frame the task as a classification problem, and the learning objective of the XGBoost model uses logistic regression for binary classification and outputs ignition probability, where the logistic loss is used as the loss function. The model treats fire ignition as target: if there is an observed fire ignition in a grid for a specific year, then the target will be class 1, if not the target will be class 0. The XGBoost model takes the input data (see Section 2.2) and predicts the ignition probability for each grid cell in each year of the historical period 1992-2020. A probability threshold of 0.5 is used to decide if the model predicts a fire ignition (when probability is ≥ 0.5) or not (when probability is < 0.5). The training and testing data split and hyper parameter tuning are detailed in Section A1 and A2 of the Appendix, respectively.

Note that the ignition probability here is not "natural" probability that directly represents the probability of fire ignition occurrence from a statistics perspective, but rather a term to indicate the spatial variability in ignition propensity from a relative aspect. In other words, we only use the ignition probability as a reference for ignition sampling at spatial locations in our simulation framework, i.e., grids with higher probability will be sampled more often. By doing so, the spatial variations in ignition propensity is well captured, which is an improvement to ignition component of some earlier wildfire risk assessment frameworks which just randomly sample ignitions within a FPU[9, 13] or just rely on historical observations[11, 12]. Therefore, despite the caveat of being not a "natural" probability, the practicality in the use case of wildfire risk assessment for (re)insurance justifies the existence and success of our simplified ML-based ignition model.

2.2 DATA

We have conducted careful feature engineering and extensive literature review to deploy available data that best suit the problem of ignition modeling. For the data source of wildfire ignitions, we use the FPA_FOD dataset developed by the US Forest Service[16], which collects and compiles ground truth ignition data from different agencies from 1992-2020. It is the most comprehensive and up-to-date wildfire ignition data source, including fires of all sizes across the states. In our use case, we only use the ignition data for wildfires in California. The input features of the model include biogeophysical and anthropogenic drivers of fire ignition [17-19], classified into the following four categories: 1) climate; 2) vegetation; 3) topography; and 4) anthropogenic factors. The detailed input variables, temporal aggregation and data sources are listed in Table A2 of the Appendix. The preparation of all data is documented in Section A3 of the Appendix.

3 RESULTS

3.1 EVALUATIONS OF THE BASELINE MODEL

The baseline model is evaluated on the testing set which spans from 2015 to 2020. The evaluation presents a precision, recall, and area under precision-recall curve (AUC-PR) of 0.73, 0.75 and 0.74 (see Figure A2 in the Appendix), respectively. Note that from the perspective of (re)insurance, usually a higher recall is more desirable, this is because the miss detection of wildfires may cause catastrophic financial loss to the (re) insurers and they would try to avoid it. In this context, the recall of the baseline model could be further improved by seeking alternative approaches.

3.2 EVALUATIONS OF THE ML-BASED IGNITION MODEL

With the same training and testing period, the ML-based ignition model reports a precision of 0.69 and a recall of 0.86, with an AUC-PR of 0.81 (see Figure A3a in the Appendix), suggesting a good performance of the ignition model even for the extreme years such as 2017, 2018 and 2020 when wildfire intensity is high. With a much higher recall and overall better performance (as indicated by higher AUC-PR), the ML-based ignition model is clearly more desirable than the benchmark model.

In addition, we also trained an ignition model in the same way but only with data from the populated area (with population density $\geq$ 10km^2). We conduct this additional experiment because, from a wildfire (re)insurance perspective, the insured is usually home properties. Thus, it is the populated areas that matter most, although ignitions in unpopulated areas could spread to populated areas. We find that the precision, recall, and AUC-PR of the populated-area-only model are 0.82, 0.90, and 0.90, respectively (see Figure A3b in the Appendix), significantly higher than those of the model for the entire California. This is probably because wildfire ignitions are disproportionally more located in populated area. Based on the FPA_FOD dataset[16], 60% of the ignitions occur in the populated area (population density $\geq$ 10/km2) which only accounts for $\sim$1/4 of California territory. Thus, the ignition model for the populated area tends to have relatively more training sample and the model could learn the ignition mechanisms better. Nonetheless, for complete spatial coverage, we still present results from the ignition model for the entire California in following sections.

3.3 Feature importance analysis

To better understand which input features dominate wildfire ignitions, we conduct feature importance analysis for the ML-based ignition model of the entire California based on the SHapley Additive exPlanations (SHAP) value[20]. The analysis reveals that the top six important features are power line density, enhanced vegetation index (EVI), vegetation optical depth (VOD), distance from wildland-urban interface (wui-distance), vapor pressure deficit (VPD) and relative humidity (RH) (Figure 1). While power line density and WUI-distance are proxies for intensity of human-related activities, EVI and VOD reflect the vegetation biomass and vegetation water content, respectively, and VPD and RH represent atmospheric aridity and humidity, respectively. These results are in line with earlier studies on the key drivers of human activities, fuels and moisture deficit [3, 17, 18, 21, 22].

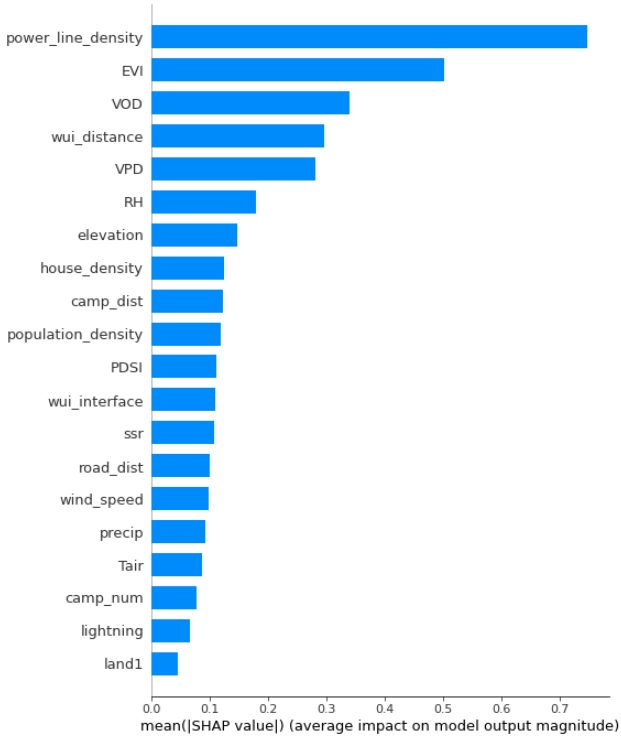

Figure 1: Feature importance of the ML-based ignition model. EVI=enhanced vegetation index, VOD= vegetation optical depth, RH=relative humidity, VPD=vapor pressure deficit, PDSI=Palmer Drought Severity Index, ssr =surface shortwave radiation, Tair= air temperature, land1 = evergreen needleleaf forests. See Table A2 in Appendix for full description of all features.

3.4 Spatiotemporal variations in ignition probabilities

The ignition probabilities across California vary from area to area due to differences in natural and anthropogenic features like weather, topography, land cover, population density, etc., presenting distinct spatial patterns (Figure 2a-b). The ignition probabilities are relatively low in the sparsely populated areas such as the desert area in southeast California, and in agricultural areas of the Central Valley, due to low vegetation density or relatively low human activities. On the other side, the ignition probabilities are much higher in the densely populated areas such as the coastal region of California, especially the southern coast, and the east of the Central Valley, due to intense human activities. The ignition probabilities in the northern mountain area are in-between, mostly driven by lightning and dense vegetation[17]. These spatial patterns align with earlier studies[17]. Although the spatial patterns of ignition probabilities seem similar from year to year, the probabilities still present clear inter-annual variabilities in both magnitudes and spatial patterns. Specifically,

more devastating wildfire years present overall higher ignition probabilities (e.g., 2018 vs. 2012 in Figure 2). In addition, some areas tend to change more drastically in ignition probabilities than other areas (e.g., Figure 2c). Apparently, the inter-annual differences in their spatial patterns (Figure 2c) will result in differences in spatial sampling of wildfire ignitions for different years in our wildfire risk assessment framework, which may lead to significant implications for the outcome of wildfire risk assessment for different years.

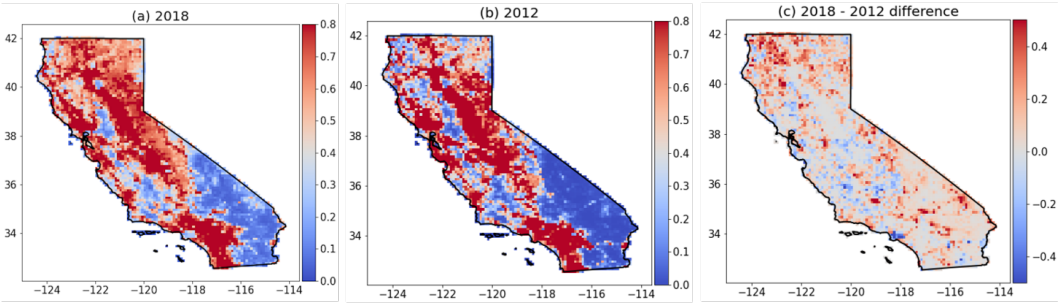

Figure 2: Modeled ignition probabilities in 2018 (a) and 2012 (b) and their differences (c) in California.

4 CONCLUSIONS

In light of the escalating frequency, intensity and extent of wildfires in the Western US and their subsequent catastrophic socioeconomic impacts, efficient and accurate estimation of fire ignitions is imperative for any preventive measures of wildfire risk. Wildfire (re)insurance provides a safety net for potential wildfire induced loss to homeowners, which is critical in safeguarding home properties and socioeconomic stability. At the same time, the nature of randomness in wildfire ignitions poses great challenge to the modeling of wildfire ignitions. To strike a balance among model complexity, performance, and practicality for (re)insurance, we use a simplified ML-based ignition model at yearly scale in our (re)insurance business, which is easy-to-build, lightweight, efficient, and has been proven to work well in wildfire (re)insurance in California. In addition, this framework could be easily applied to other regions, and to other perils such as hurricanes, which involve similar processes (genesis and moving track) like wildfires (genesis and contagion). The ignition model will continue to be refined (see Section C in the Appendix) to best support our wildfire (re)insurance activities, and in turn will help provide a broader and more affordable safety net for homeowners.

References

[1] World Meteorological Organization (WMO). "Drought and heat exacerbate wildfires," 01/24/2023.

[2] D. Wang, D. Guan, S. Zhu, M. Mac Kinnon, G. Geng, Q. Zhang, H. Zheng, T. Lei, S. Shao, and P. Gong, "Economic footprint of California wildfires in 2018," *Nature Sustainability*, vol. 4, no. 3, pp. 252-260, 2021.

[3] J. K. Balch, B. A. Bradley, J. T. Abatzoglou, R. C. Nagy, E. J. Fusco, and A. L. Mahood, "Human-started wildfires expand the fire niche across the United States," *Proceedings of the National Academy of Sciences*, vol. 114, no. 11, pp. 2946-2951, 2017.

[4] H. A. Kramer, M. H. Mockrin, P. M. Alexandre, S. I. Stewart, and V. C. Radeloff, "Where wildfires destroy buildings in the US relative to the wildland–urban interface and national fire outreach programs," *International journal of wildland fire*, vol. 27, no. 5, pp. 329-341, 2018.

[5] H. A. Kramer, M. H. Mockrin, P. M. Alexandre, and V. C. Radeloff, "High wildfire damage in interface communities in California," *International journal of wildland fire*, vol. 28, no. 9, pp. 641-650, 2019.

[6] V. C. Radeloff, D. P. Helmers, H. A. Kramer, M. H. Mockrin, P. M. Alexandre, A. Bar-Massada, V. Butsic, T. J. Hawbaker, S. Martinuzzi, and A. D. Syphard, "Rapid growth of the US wildland-urban interface raises wildfire risk," *Proceedings of the National Academy of Sciences*, vol. 115, no. 13, pp. 3314-3319, 2018.

[7] J. Kauflin, "Fintech's Wildfire Fighter: This Reinsurer Is Using A.I. To Make California Insurance Affordable," *Forbes*, Forbes, 2020.

[8] CDI, "New Data Shows Insurance is Becoming Harder to Find as a Result of Wildfires," California Department of Insurance Press, 2019.

[9] D. E. Calkin, A. Ager, M. P. Thompson, M. A. Finney, D. C. Lee, T. M. Quigley, C. W. McHugh, K. L. Riley, and J. M. Gilbertson-Day, "A comparative risk assessment framework for wildland fire management: the 2010 cohesive strategy science report," 2011.

[10] J. H. Scott, M. P. Thompson, and D. E. Calkin, "A wildfire risk assessment framework for land and resource management," 2013.

[11] E. J. Kearns, D. Saah, C. R. Levine, C. Lautenberger, O. M. Doherty, J. R. Porter, M. Amodeo, C. Rudeen, K. D. Woodward, and G. W. Johnson, "The construction of probabilistic wildfire risk estimates for individual real estate parcels for the contiguous United States," *Fire*, vol. 5, no. 4, pp. 117, 2022.

[12] J. Scott, K. Short, and M. Finney, "FSim: The Large Fire Simulator Guide to Best Practices," Pyrologix LLC. Available online: `https://pyrologix.com/wp-content/uploads/2019/11/FSimBestPractices_0`, vol. 3, 2018.

[13] M. A. Finney, C. W. McHugh, I. C. Grenfell, K. L. Riley, and K. C. Short, "A simulation of probabilistic wildfire risk components for the continental United States," *Stochastic Environmental Research and Risk Assessment*, vol. 25, no. 7, pp. 973-1000, 2011.

[14] V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, and M. Gomis, "Climate change 2021: the physical science basis," Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change, vol. 2, 2021.

[15] T. Chen, and C. Guestrin, "Xgboost: A scalable tree boosting system." pp. 785-794.

[16] K. C. Short, "Spatial wildfire occurrence data for the United States, 1992-2020 [FPA_FOD_20221014]," 2022.

[17] B. Chen, and Y. Jin, "Spatial patterns and drivers for wildfire ignitions in California," *Environmental Research Letters*, vol. 17, no. 5, pp. 055004, 2022.

[18] N. Faivre, Y. Jin, M. L. Goulden, and J. T. Randerson, "Controls on the spatial pattern of wildfire ignitions in Southern California," *International Journal of Wildland Fire*, vol. 23, no. 6, pp. 799-811, 2014.

[19] G. Di Virgilio, J. P. Evans, S. A. Blake, M. Armstrong, A. J. Dowdy, J. Sharples, and R. McRae, "Climate change increases the potential for extreme wildfires," *Geophysical Research Letters*, vol. 46, no. 14, pp. 8517-8526, 2019.

[20] S. M. Lundberg, and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[21] A. P. Williams, R. Seager, M. Berkelhammer, A. K. Macalady, M. A. Crimmins, T. W. Swetnam, A. T. Trugman, N. Buenning, N. Hryniw, and N. G. McDowell, "Causes and implications of extreme atmospheric moisture demand during the record-breaking 2011 wildfire season in the southwestern United States," *Journal of Applied Meteorology and Climatology*, vol. 53, no. 12, pp. 2671-2684, 2014.

[22] F. Sedano, and J. Randerson, "Vapor pressure deficit controls on fire ignition and fire spread in boreal forest ecosystems," *Biogeosciences Discussions*, vol. 11, no. 1, pp. 1309-1353, 2014.

[23] F. Mesinger, G. DiMego, E. Kalnay, K. Mitchell, P. C. Shafran, W. Ebisuzaki, D. Jović, J. Woollen, E. Rogers, and E. H. Berbery, "North American regional reanalysis," *Bulletin of the American Meteorological Society*, vol. 87, no. 3, pp. 343-360, 2006.

[24] J. T. Abatzoglou, "Development of gridded surface meteorological data for ecological applications and modelling," *International Journal of Climatology*, vol. 33, no. 1, pp. 121-131, 2013.

[25] R. E. Orville, "Development of the national lightning detection network," *Bulletin of the American Meteorological Society*, vol. 89, no. 2, pp. 180-190, 2008.

[26] D. Sulla-Menashe, and M. A. Friedl, "User guide to collection 6 MODIS land cover (MCD12Q1 and MCD12C1) product," *Usgs: Reston, Va, Usa*, vol. 1, pp. 18, 2018.

[27] L. Moesinger, W. Dorigo, R. de Jeu, R. van der Schalie, T. Scanlon, I. Teubner, and M. Forkel, "The global long-term microwave vegetation optical depth climate archive (VODCA)," *Earth System Science Data*, vol. 12, no. 1, pp. 177-196, 2020.

[28] Z. Jiang, A. R. Huete, K. Didan, and T. Miura, "Development of a two-band enhanced vegetation index without a blue band," *Remote sensing of Environment*, vol. 112, no. 10, pp. 3833-3845, 2008.

[29] J. J. Danielson, and D. B. Gesch, "Global multi-resolution terrain elevation data 2010 (GMTED2010)," 2011.

[30] V. Radeloff, M. Dubinin, N. Coops, A. Allen, T. Brooks, M. Clayton, G. Costa, C. Graham, D. Helmers, and A. Ives, "The dynamic habitat indices (DHIs) from MODIS and global biodiversity," *Remote Sensing of Environment*, vol. 222, pp. 204-214, 2019.

Appendix

A Implementation Details

In this part, we provide supplementary materials for the methods and data used in this work.

A.1 Training and testing data split

We split the training and testing set with a ratio of 80% and 20% for the historical period of 1992-2020 across California, i.e., the first 80% of the time series of 1992-2020 (including both input and target data) are used for training and the remaining are used for testing. Here we do it in a temporal continuous manner rather than shuffling the training data, because we want the model to be able to capture the trend of increasing wildfire frequency and intensity in California in latest years. If the model performs well during the test set period (2015-2020), it suggests that the model captures that trend well. This will help in a case if future extreme climate tends to be more frequent and intense.

A.2 Hyper parameter tuning

We use the Python package Hyperactive for hyperparameter tuning, and five-fold cross validation is used to avoid over-fitting. The tuned hyperparameters and optional values for each hyper parameter are listed in Table A1. Recall is used as the scoring metric for hyperparameter tuning because we need to best avoid false negative, in which case the model fails to predict a real fire and then it may result in catastrophic loss to the (re)insurer. The best set of hyper parameters is selected by choosing the one that achieves the highest recall during cross-validation, and they are also listed in Table A1.

Table A1: Hyperparameters and corresponding options for tuning, and final selected value after tuning (please see documentation at `https://xgboost.readthedocs.io/` for description of each hyperparameter)

Hyperparameter	**Option values**	**Selected value**
n_estimators	200, 300, 400, 500	500
max_depth	3, 5, 6, 8, 10	8
min_child_weight	2, 3, 4	2
learning_rate	0.005, 0.01, 0.05, 0.1	0.1
gamma	0.2, 0.4, 0.6, 0.8	0.2
subsample	0.6, 0.8, 1	0.8
colsample_bytree	0.6, 0.8, 1	0.8
reg_alpha	0.01, 0.05, 0.1	0.1

A.3 Data preparation

The spatial domain of this study is California, and we prepare the data in a spatial resolution of 0.1-degree at grid level, with a total of 4432 grids in California. All the related data are processed to 0.1-degree and yearly scale, except those static input variables are kept static along time. Specifically, for any data that are finer than 0.1-degree spatial resolution (or yearly temporal resolution), they are aggregated to 0.1-degree (or yearly average or yearly total as shown in Table A2). For data that are coarser than 0.1-degree, the nearest neighbor is used to obtain 0.1-degree gridded data. For data that have coarser temporal resolution than yearly, linear interpolation is used for period when data are available, and the nearest neighbor is used for period when data are unavailable. In addition, categorical input data, specifically land cover type, are transformed to binary using one-hot encoding. For continuous input data, they are scaled using standard scaler function in sklearn package before feeding into the ignition model.

B Supplementary results

The supplementary figures A1, A2 and A3 are included in this section.

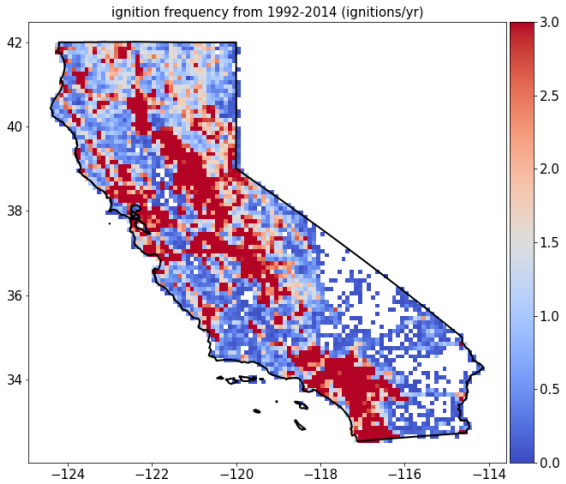

Figure A1. Spatial distribution of historical frequency of ignition occurrences during 1992-2014.

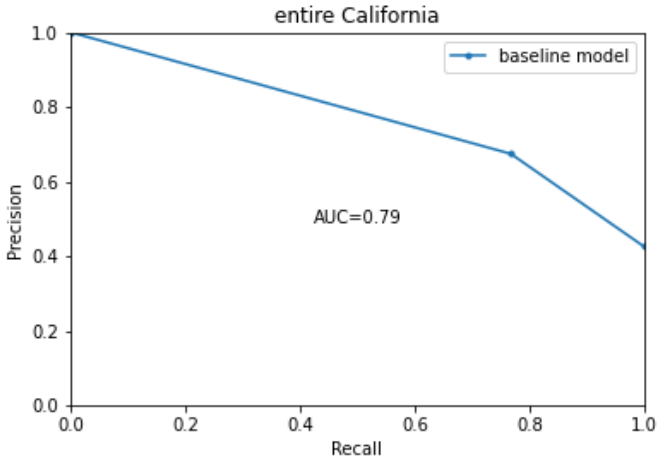

Figure A2. Precision-recall curve of the baseline model for entire California. AUC = area under curve.

C Limitations and Future Work

C.1 Limitations

The randomness of wildfire ignitions, complexity in causes, and limited historical wildfire records make estimating the "natural" probability of wildfire ignitions challenging. Constrained by that challenge and the limitations of the XGBoost method itself, a caveat of the ignition probability derived by our simplified ignition model is that it is not "natural" probability which directly represents the probability of fire ignition occurrence from statistics perspective. Nonetheless, the spatial variations of the ignition probabilities from our ignition model are used as a reference for spatial sampling of ignitions, which is the main purpose of this simplified ignition model and it well suits our wildfire simulation framework in our practical use case of wildfire (re)insurance. In addition, the possible omission or redundancy of wildfire ignitions in the FPA_FOD dataset could cause potential biases in our model results.

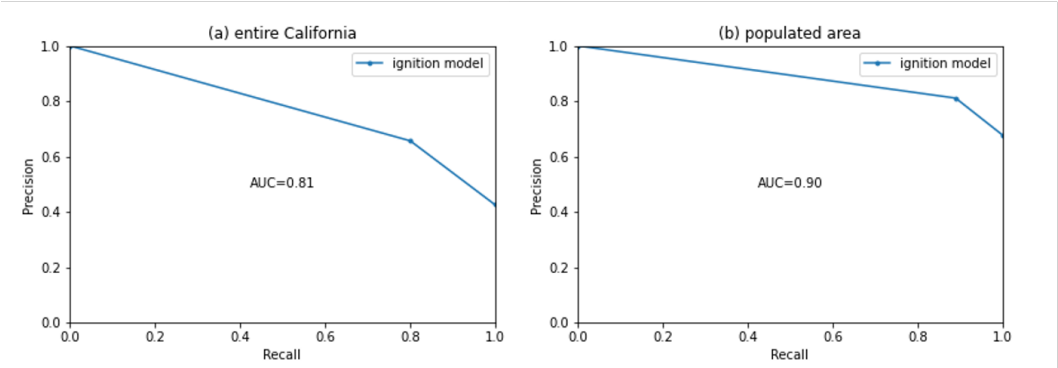

Figure A3. Precision-recall curve of the ignition model for entire California (a) and the populated area (b). AUC = area under curve.

C.2 FUTURE WORK

Future and ongoing work includes but not limited to: 1) examining the mechanisms of big fire ($\geq$ 1000 ha) and small fire ($<$1000 ha) ignitions; 2) adding more observations of fire ignitions from different sources across the US; 3) finding and adding more related data like nearest distance to power line, secondary properties of power line (e.g., intensity of poles, quality and age of power lines) to the ignition model.

Table A2: Wildfire ignition data and input features for the ML-based wildfire ignition model

Category	variables	Temporal aggregation	Data sources
Wildfires	ignitions	yearly total (binary, if total>0 treat it as 1, otherwise 0)	FPA_FOD[16]
Climate	air temperature (Tair)	yearly mean	NARR (North America Regional Reanalysis)[23]
	precipitation (precip)	yearly total	NARR (North America Regional Reanalysis)[23]
	surface solar radiation (ssr)	yearly mean	NARR (North America Regional Reanalysis)[23]
	relative humidity (RH)	yearly mean	NARR (North America Regional Reanalysis)[23]
	wind speed (wind_speed)	yearly mean	NARR (North America Regional Reanalysis)[23]
	Palmer Drought Severity Index (PDSI)	yearly mean	GRIDMET[24]
	lightning strike density (lightning)	yearly total	NLDN[25]
Vegetation	land cover type (land0 to land15, representing 16 different types)	static	MCD12C1[26]
	vegetation optical depth (VOD)	yearly mean	VODCA[27]
	enhanced vegetation index (EVI)	yearly mean	EVI2[28]
Topography	elevation (elevation)	static	GMTED2010[29]
Anthropogenic	power line density (power_line_density)	static	California Public Utilities Commission
	population density (population_density)	static within a year, changing yearly	SILVIS Lab[30]
	average housing unit density (house_density)	static within a year, changing yearly	SILVIS Lab[30]
	fraction of area classified as wildland-urban interface – interface (wui_interface)	static within a year, changing yearly	SILVIS Lab[30]
	distance from wildland-urban interface (wui_dist)	static within a year, changing yearly	SILVIS Lab[30]
	distance to nearest camp (camp_dist)	static	US Campground Directory
	number of campgrounds (camp_num)	static	US Campground Directory