Assessing the Feasibility of Multiple Machine Learning Models in Predicting Wildfire Likelihood in California

By Coffy Andrews-Guo and Vyanna Hill

Abstract 1
Keywords 2
Problem Statement 2
Literature review 2
Background on Parameters 2
Background on Algorithm Selection 3
Methodology 5
Data Collection 5
Data Exploration 5
Data Preparation 8
Data Analysis 9
Results and Discussion 11
Summary 14
Future Applications 16
Code Appendix 16
Bibliography 17

Abstract

The large spikes in wildfires in California can disrupt the food availability for US residents. The research project explores the viability of multiple machine-learning algorithms based on historical and geological data of the state for its accuracy on wildfire presence. The project compared six algorithms in its classification of wildfire on its model performance: Distributed Random Forest, Gradient Boosted Decision Tree, Naive Bayes, AutoML, Supervised Learning, and Deep Learning. Feature selection for the analysis was structured around each model's interpretation with the importance of the response variable. The goal was based on the most optimal Area Under the Curve (AUC) score of the training set.

The study's results suggest that using machine-learning algorithms can make wildfire prediction feasible, as some of the models achieved accuracies and sensitivities of over 90%. The most effective algorithms identified were Random Forest, Gradient Boosting, and AutoML, collectively exhibiting a 99% AUC. AutoML performed exceptionally well, with 95% accuracy, 93% precision, and 98% True Positive Rate (TPR) on the test set. AutoML's success can be attributed to its stacked ensemble, a model trained on multiple algorithms, which employed a layered training approach, resulting in an R^2 of 85%. Other models, like Gradient Boosting, achieved a 93% precision, demonstrating their ability to rectify errors from preceding trees. The

findings suggest incorporating stacked algorithms in model tuning could significantly improve wildfire prediction research.

Keywords

Feature Selection, Machine Learning, Classification, Sensitivity, Spatial Data

Problem Statement

Prevention of human and environmental loss is dependent on the detection of wildfire warning signs in California. Forest services require a system that can alert rangers of areas in California with significant potential for a wildfire. The project seeks multiple algorithms on their performance in wildfire predictions and their success. Success will be measured by its test accuracy, sensitivity (True Positive Rate), and its measure of accounting for variance in the model. The project's main goal is the assessment of multiple machine learning algorithms on California's historical and geological data and their credibility in model performance.

Literature review

The literature review involved an analysis of the parameters and algorithms planned for employment in the project. The team determined the project's parameters by analyzing previous research and identifying the parameters that exhibited strong performance. Across all the journals: precipitation, topography, and climate were found to be the most effective parameters.

Background on Parameters

Precipitation was particularly influential in the models the two research teams developed. In *Spatial patterns and drivers for wildfire ignitions in California*, the annual snow precipitation was the best feature in their lighting-made fire prediction model. Chen and Jin (2022) concluded that annual snowfall increased fuel moisture, which aided in the spike in human-incident wildfires. Similarly, The researchers of Syphard et al. (2018) saw if the winter season did not see the expected rainfall, then the summer/fall season saw increased wildfire incidents.

Elevation and slope were the most identified features included in multiple models. Syphard et al (2018) found slopes in steep terrains had a higher influence on the prediction model, as the terrain interacted with the distribution of the water runoff. The interaction between slope and water distribution captured again in Sadrabadi et al (2023), where the distance from the nearest water source is a feature.

Climate provides additional context and supports other features. For example, Syphard et al. (2018) noted changes in annual minimum temperature in the Californian winter season caused a decrease in water reservations for the summer. This interaction can be missed if weather is not included in the model build. *Calhoun et al. (2021)* research journal provides additional support on the effects of the climate variable. The Mediterranean climate of California, characterized by dry summers and mild winters, makes it highly prone to wildfires, a problem further exacerbated by climate change.

Vegetation in the research papers was of the area's soil type and flora. It is a complex feature due to its handling in the data collection and its individual effects on a model's accuracy. In Syphard et al. (2018), the research team expressed concern about how researchers treat vegetation as static and cannot examine real-time "feedback" with vegetation and its future behaviors (pg.2). Their team explored multiple models with and without vegetation but could not determine which dataset had the most feasible projection. The team concluded any wildfire prediction model would have more complications without the vegetation feature included as vegetation features brought more features not seen in water and climate.

In the research led by Jiang et al. (2021), the researchers chose to incorporate the "Normalized Difference Vegetation Index (NDVI)" as a predictor. This index evaluates tree density, which supports forest resilience against wildfires. Their findings highlighted that the high-risk zones were predominantly within forested regions. Nevertheless, as they fine-tuned the NDVI to reflect healthier tree conditions, the level of risk exhibited a gradual reduction. Notably, the researchers' predictions featured fewer false positives than the official outlook from the Northwest Large Fire Interactive Map. In contrast, Lydersen et al. (2014) blended vegetation features in a singular model. Their team saw differences in accuracy dependent on the vegetation type. They noticed low fuel had the highest accuracy as the high-fuel areas saw an overprediction in true positives.

Background on Algorithm Selection

Classification Models for Susceptibility Assessment:

Jaafari et al. (2018) introduced another significant approach to wildfire prediction using classification models. The authors employed various classifiers, including ADT, LMT, CART, FT, and NBT, to evaluate wildfire susceptibility. The primary objective was to assess the probability of wildfires occurring in specific regions based on a combination of explanatory variables. The ADT classifier emerged as the most effective, showcasing its superiority in predictive capabilities. This classification approach is critical in identifying areas with a high potential for fire outbreaks, allowing proactive measures to be taken in those regions.

Machine Learning and Ensemble Models:

Malik et al. (2021) brought machine-learning models to the forefront of wildfire prediction. Different machine-learning models were used for weather and remote sensing data. Support Vector Machine, XGBoost, Random Forest, Multi-Layer Perceptron, Convolutional Neural Network, and Long Short-Term Memory models were applied to analyze data and predict fire risk. An ensemble model combined the results from these various models, resulting in a more reliable and accurate prediction. Machine learning techniques have demonstrated their capability to capture complex relationships between variables, thus enhancing the precision of fire risk prediction.

Decision Trees and Random Forest:

In Sadrabadi and Innocente (2023), decision trees and random forests were introduced as robust machine-learning methods. Decision Trees are universal function approximators, and Random Forest leverages bagging and random subspace to construct accurate models. These models are essential for mapping out the potential fire risk in different regions and identifying the key predictors contributing to this risk.

Gradient Boosting Techniques:

Gradient Boosting Decision Tree (GBDT) and Extreme Gradient Boosting (XGB) were highlighted by Sadrabadi and Innocente (2023) as techniques for improving predictive models. GBDT iteratively creates functions to enhance predictions, while XGB combines multiple decision trees for more robust models. These techniques help refine wildfire prediction by considering complex interactions and non-linear relationships among variables.

Stepwise Regression:

Yue et al. (2013) researched the relationships between weather factors and fire indexes and observed wildfires in three regions using stepwise regression models. They analyzed current and past data on temperature, humidity, rainfall, and fire weather indicators. The study indicates that larger fires are more likely to occur in some areas when the weather is hotter and drier. However, in the Sierra Nevada region, an unusual pattern emerged with a negative correlation between higher temperatures two years before and smaller fires. This phenomenon may be attributed to soil moisture and plant growth. The regression models used in the study effectively explained a significant portion of the variability in wildfire size.

Wildfire prediction aims to offer precise and timely evaluations of potential fires occurring at specific places and times. The literature reviewed in this study encompasses various models and techniques, including statistical methods like classification models, machine learning methods, ensemble approaches, and stepwise regression. For the wildfire prediction efforts in

California, this project will utilize the classified learning approach proposed by Jaafari et al. (2018) and Malik et al. (2021).

Methodology

Data Collection

The team gathered multiple resources for the features mentioned in the literature review. The wildfire likelihood data was from FEMA, which contained the national risk index of wildfire likelihood for 2022. FEMA's National Risk Index dataset contains other weather phenomena such as strong winds, heat waves, and lighting with its historical frequencies and annual events. These additional data points are beneficial as strong winds were a risk component called out in Yue's Journal. The tract level granularity allowed California only observations in the train and test sets. The team aimed for other features at the tract level, which was available for slope and vegetation data from LandFire.

LandFire provided the slope and vegetation data via raster files. The team used the QGIS platform to retrieve the raster data seen on the Californian tract shapefile and exported it into a CSV file for analysis. Granularity was a pain point in data collection as different agencies had county level as the lowest availabile. California's elevation provided from the US Geographical Survey with its granularity at the county level. NOAA Climate at a Glance provided the annual precipitation and temperature. The final data set combined all data by its tract and county census id.

Data Exploration

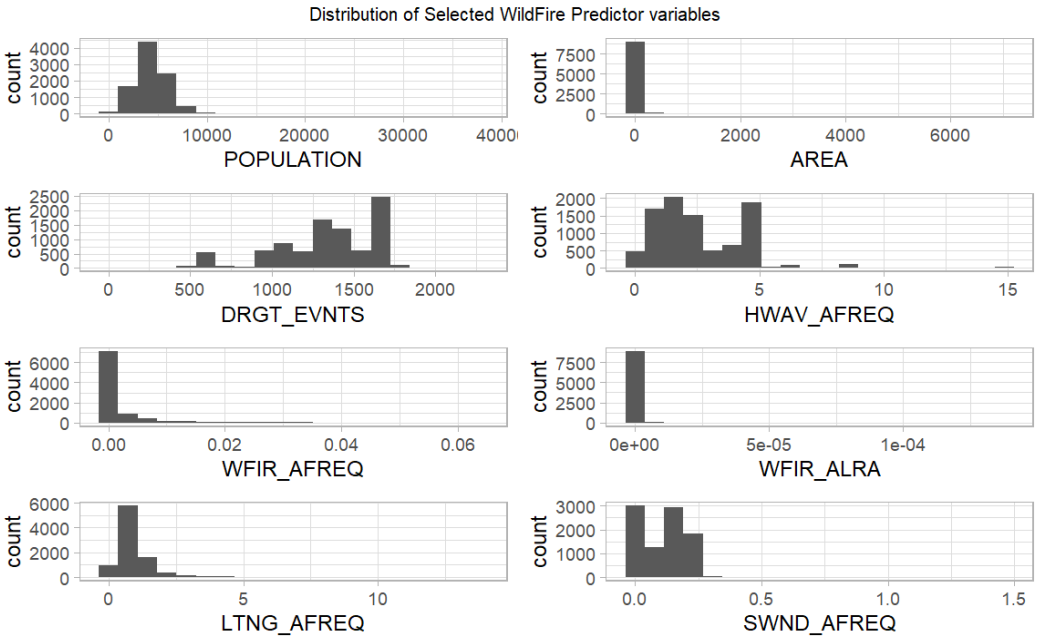
The data set observed 9,089 Californian tracts with 29 variables in the dataset. The response variable is "WFRI_R", which is the binary response of the tract's likelihood of a wildfire. The predictor variables are seen in the list below.

<u>Predictor Variables</u>

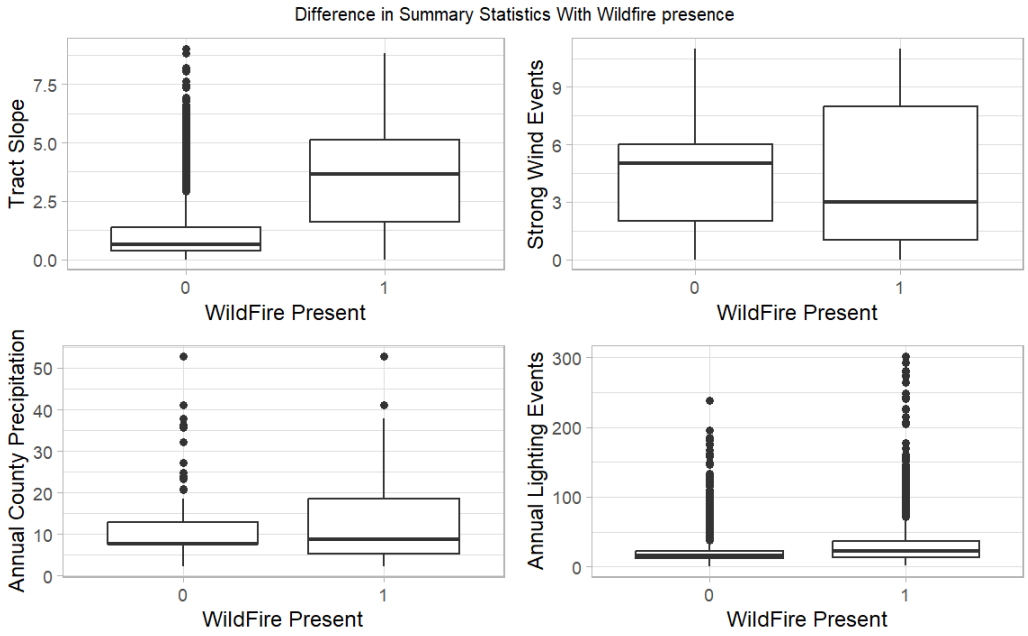
- POPULATION: Population (2020)
- AREA: Area (sq mi)
- DRGT_AFREQ: Drought - Annualized Frequency
- DRGT_HLRA: Drought - Exposure - Agriculture Value
- HWAV_EVNTS: Heat Wave - Number of Events
- HWAV_AFREQ: Heat Wave - Annualized Frequency
- HWAV_HLRA: Heat Wave - Historic Loss Ratio - Agriculture
- LTNG_EVNTS: Lightning - Number of Events
- LTNG_AFREQ: Lightning - Annualized Frequency
- SWND_EVNTS: Strong Wind - Number of Events
- SWND_AFREQ: Strong Wind - Annualized Frequency
- SWND_HLRA: Strong Wind - Historic Loss Ratio - Agriculture

- WFIR_AFREQ: Wildfire - Annualized Frequency
- WFIR_HLRP: Wildfire - Historic Loss Ratio - Population
- WFIR_HLRA: Wildfire - Historic Loss Ratio - Agriculture
- WFIR_ALRA: Wildfire - Expected Annual Loss Rate - Agriculture
- TRCT_WAREA: Tract Water Area
- TRCT_SLOPE: Tract Slope
- CNTY_ELEV: County Elevation
- CNTY_TEMP: County Annual Temperature (2022)
- CNTY_PRECIP: County Annual Precipitation (2022)
- TRCT_VEGLF: Tract Average Vegetation Lifeform

Variable Statistics



The team reviewed a selected group of predictor variables' density distributions. The majority of features appear non-normal with a left-skewed distribution. This skewness is the result of the majority of observations with low frequencies. Drought events were a feature of right-skewed distribution, as droughts are a common occurrence in California.



The box plots above display the differences between the likelihood of a wildfire. The team noticed the variables tract slope and strong wind events have the most noticeable difference in distribution. It appears that tracts with a higher slope have a higher likelihood of a wildfire than lower slope tracts. Tracts with the likelihood of a wildfire see more strong wind events, as its maximum hits around seven annual events. The team theorized from the boxplots that the features mentioned above have a strong influence on feature importance. Another observation is there are outliers in a few features. The models in the project will have to handle them.

Imbalanced data

The team checked the distribution of the wildfire response cases. The team noticed that the data set is heavily imbalanced as the majority of the cases do not have the likelihood of a wildfire. This imbalance can affect the accuracy of the models, as the models are predominately trained on non-wildfire cases and may falsely predict positive wildfire cases.

Data Preparation

Missing Values

In the data set, there were fourteen tracts with missing values. As these values are not missing with a pattern, the team can assume that the missingness is at random. In the preservation of all Californian tracts, the team used MICE imputation for the missing values. MICE uses regression in its prediction of the missing values and fills in the null value. Unlike the median used for imputation on the missing, this imputation offers another way for less biased imputed values in the data set.

Categorical Variables

Second, wildfire response and vegetation life form variables were categorical variables. Binary variables replaced their categorical counterparts. The response transformed from WFIR_RISKV to its binary variable WFRI_R. The vegetation feature had nine unique values, which created eight new features (ex: ".isShrub").

Downsampling

From the data exploration, the team saw a heavily imbalanced data set. The data set was downsampled for an equal balance between wildfire and non-wildfire cases. Downsampling helped the machine-learning models see more wildfire cases. This change also helped the possible overfitting in the previous process, as older models had accuracies of 99% in the training set.

Multi-collinearity

The team wanted to leave feature selection to the models as different algorithms may prioritize certain features based on their subsets. The priority was checking if the predictor variables were not correlated with each other, as it can affect the weighting of the coefficients. The correlation matrix revealed drought events and drought frequency were highly correlated. Only drought frequency made the feature list from these results.

Data Analysis

Wildfire Risk Prediction Models

We used the H2o package in R for the prediction models. We used six models from the package: Distributed Random Forest (DRF), Automatic Machine Learning (AutoML), Deep Learning, Gradient Boosting Machine (GBM), Naïve-Bayes, and Support Vector Machine (SVM).

Distributed Random Forest (DRF):

DRF creates multiple decision trees using a random subset of the dataset and features at each node. In the first version, the team created the maximum amount of trees to a thousand to see where the model produced without the default max of 50 trees. In addition, there was a stopping metric included in the model that checks the AUC score every five iterations. The stopping metric prevents the model from spitting further if the model does not see an improvement in the AUC score. Then, cross-validation was enacted in the training set of the model. The model broke its training set into five folds, which tested the performance of the training set separately. The results of the first model produce a tree of 50 trees with a max depth of 20 branches.

In the final version of the model, the team optimized based on the first model's performance. From the first model, the top ten features based on their gini coefficient were subseted into a new feature list. The max trees were shortened to 25 trees as fewer trees limit the possibility of overfitting. These improvements to the second model saw a training accuracy of 96% with an average R^2 of 85% across the cross-validated sets.

Automatic Machine Learning (AutoML):

H2o's AutoML function created multiple machine-learning models against the data set. The model's max model limit was five as the algorithm traded off the quantity of the produced models with the model's development. Unlike DRF, the optimization is developed by the algorithm decision process with the training set. The algorithm ran its produced models against a cross-validated set for its own performance measure. The top model was a stacked ensemble, which is a model with two machine-learning algorithms. The stacked model saw the highest cross-validated accuracy of 96%.

Deep Learning:

The presented code, *"dl <- h2o.deeplearning(x = features_v2, y = response, distribution = "AUTO", hidden = c(1), epochs = 1000, train_samples_per_iteration = -1, reproducible = TRUE, activation = "Tanh", single_node_mode = FALSE, balance_classes = FALSE, force_load_balance = FALSE, seed = 23123, score_training_samples = 0, score_validation_samples = 0, training_frame = train.h2o,*

stopping_rounds = 0, keep_cross_validation_predictions = TRUE)," exemplifies implementing a deep learning model using the H2O platform. Deep learning, as a subset of machine learning, employs artificial neural networks with multiple layers to autonomously discern patterns in data. This code specifies various parameters to configure the deep learning model, including the choice of the activation function ("Tanh"), the number of hidden layers, and the total number of training epochs. The reproducibility parameter ensures consistency in results across runs. Unlike traditional methods, deep learning avoids the need for manual feature selection by leveraging backpropagation to adjust internal settings autonomously.

Gradient Boosting Machine (GBM):

GBM creates new trees based on the previous tree's errors with a leaf-to-stump approach. Version one of the model had the following features: a learning rate of 0.10, a stopping metric at AUC, and a cross-validation fold at five. The learning rate of the model was set to 0.10, as the team noticed with slower learning rates the average R^2 was 41%. The team theorized with the slower rate, the model was overfitting the model with more tree subsets.

For the final version of the model, the team included the top ten features of the previous version and increased the stopping interval. The increased stopping interval checks the AUC score more frequently for review, as optimizations are limited on the model. The average R^2 from the cross-validated set was 84% and accuracy was 95%.

Naïve Bayes:

A Naive Bayes classification model is constructed using the H2O machine learning framework in the given code snippet. The code, *"pros_nb <- h2o.naiveBayes(x = features_v2, y = response, training_frame = train.h2o, laplace = 0, nfolds = 5, seed = 1234, keep_cross_validation_predictions = TRUE),"* employs explicitly the Naive Bayes algorithm to predict the response variable based on the features provided. The *"laplace = 0"* hyperparameter indicates the absence of Laplace smoothing, while *"nfolds = 5"* specifies the number of cross-validation folds for model evaluation. The seed parameter ensures reproducibility and the option to retain cross-validation predictions allows for further model performance analysis. Naive Bayes, as described, operates on the assumption of independence between features given the class label, simplifying computations.

Support Vector Machine (SVM):

The provided code snippet, *"svm_model <- h2o.psvm(gamma = 0.01, rank_ratio = 0.1, y = response, training_frame = train.h2o, disable_training_metrics = FALSE, seed = 1),"* encapsulates the implementation of a Support Vector Machine (SVM) model using the H2O machine learning framework. The parameters employed in this code, such as *"gamma"* and *"rank_ratio,"* are crucial in shaping the SVM's behavior. The *"gamma"* parameter influences the kernel function, determining the shape of the decision boundary, and *"rank_ratio"* controls the number of support vectors used in the model, impacting its generalization ability. The code ensures a comprehensive SVM model by specifying the response variable training data and incorporating options like *"disable_training_metrics = FALSE"* for tracking training metrics. SVMs excel in handling classification tasks by

identifying optimal hyperplanes that effectively separate different class data points while maintaining a margin.

According to the *Model Accuracy with ML Algorithms* table, the AutoML and Gradient Boosting Machine (GBM) algorithms provided the most accurate results among all other selected algorithms.

Model Accuracy with ML Algorithms Table

Model <chr>	Accuracy <dbl>	Unique_Hyperparameters <chr>
AutoML (Automatic Machine Learning)	0.9558	max_models = 5
Deep Learning	0.9531	Hidden = c(1), activation = 'Tanh', epochs = 1000
Distributed Random Forest (DRF)	0.9487	balance_classes = FALSE
Gradient Boosting Machine (GBM)	0.9549	learn_rate = 0.1, ntrees = 1000
Naïve-Bayes (NB)	0.9142	
Support Vector Machine (SVM)	0.5566	Gamma = 0.01, rank_ratio = 0.1

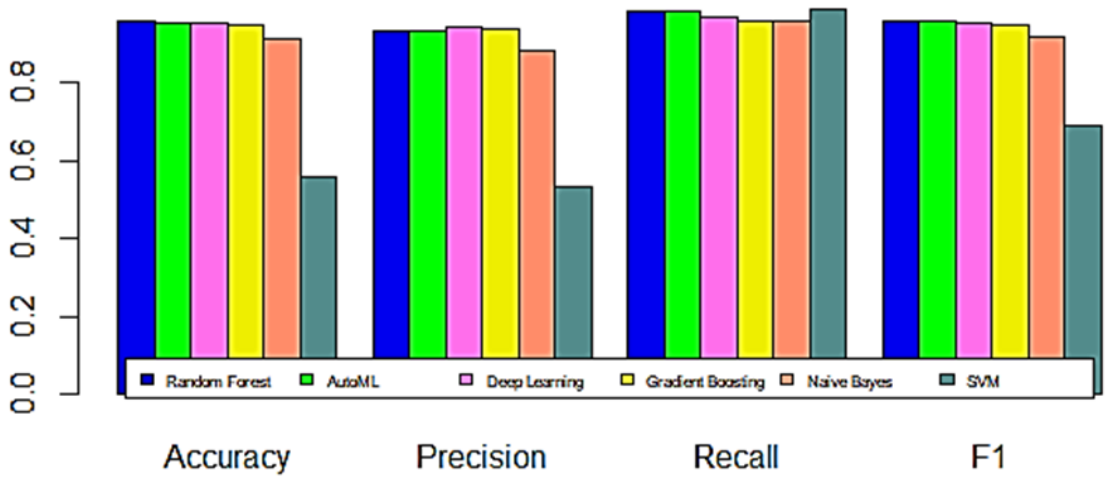
Results and Discussion

The evaluation metrics for different machine learning models, namely accuracy, precision, recall, and F1 score, are presented in the Table, *Performance Metrics "Confusion Matrix."* The AutoML model achieved the highest overall performance, with an accuracy of 0.9558, precision of 0.9328, recall of 0.9823, and an F1 score of 0.9569. The Gradient Boosting Machine model came in a close second, demonstrating competitive results across all metrics. The Deep Learning model exhibited strong precision and reasonable recall. The Distributed Random Forest model also performed well, with its high precision and recall. However, the Support Vector Machine model exhibited lower performance across all metrics, indicating potential challenges in accurately predicting positive instances. These metrics offer a comprehensive overview of the models' effectiveness in predicting accurately, with AutoML emerging as the top performer.

Performance Metrics "Confusion Matrix" Table

	Accuracy <dbl>	Precision <dbl>	Recall <dbl>	F1 <dbl>
AutoML	0.9558	0.9328	0.9823	0.9569
Gradient Boosting Machine	0.9549	0.9327	0.9805	0.9560
Deep Learning	0.9531	0.9399	0.9681	0.9538
Distributed Random Forest	0.9487	0.9393	0.9593	0.9492
Naive Bayes	0.9142	0.8811	0.9575	0.9177
Support Vector Machine	0.5566	0.5304	0.9894	0.6905

A graphical representation of the performance of the selected machine learning models on the modified values dataset.



Visualization of Model Performance Metrics

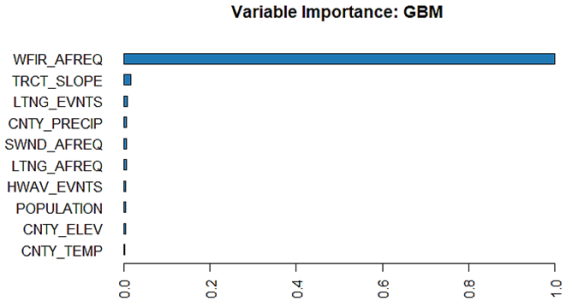
The table of Receiver Operating Characteristic (ROC) analysis presents the outcomes of assessing and contrasting the performance of different binary classification machine learning models. Random Forest (RF) exhibited robust performance, with an R^2 of 0.844, low MSE and RMSE values, and a high AUC, indicating accurate predictions and effective classification ability. Gradient Boosting (GB) showed similar excellence, with a slightly higher R^2 of 0.833 and comparable metrics. The AutoML Stack Ensemble fared just as well as RF and GB, achieving an impressive R^2 of 0.830 low MSE and RMSE values. Deep Learning (DL) and Naive Bayes also performed satisfactorily but were outperformed by RF, GB, and AutoML. However, the R^2 score of -0.91 for the Supervised Vector Machine indicates that the model cannot detect useful patterns or relationships between the variables. This implies that the SVM model is unsuitable for the data because the variables included do not significantly affect the outcome. The AutoML Stack Ensemble emerged as the leading predictive model with a 95.58 % performance accuracy based on ROC and the Confusion Matrix metrics.

Receiver Operating Characteristic (ROC) Analysis Metrics

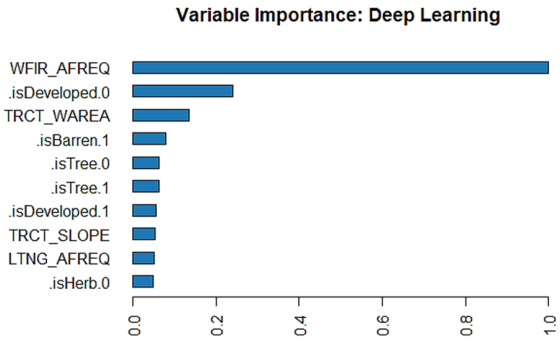
	R2	MSE	RMSE	AUC	classError
Random Forest	0.8442478	0.03893805	0.1973273	0.9914371	0.03893805
Gradient Boosting	0.8336283	0.04159292	0.2039434	0.9901966	0.04159292
automl- stack ensemble	0.8300885	0.04247788	0.2061016	0.9902984	0.04247788
Navie Bayes	0.5752212	0.10619469	0.3258753	0.9855776	0.10619469
Deep Learning	0.7911504	0.05221239	0.2285003	0.9855776	0.05221239
Supervised Learning Model	-0.9150442	0.47876106	0.6919256	0.9855776	0.47876106

To gain insights into decision-making processes, we evaluated the interpretability of models. Different approaches use unique methodologies to assess feature importance across three models - Gradient Boosting Machine (GBM), Deep Learning, and Distributed Random Forest (DRF).

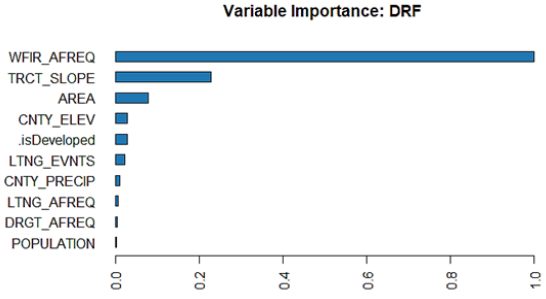
GBM creates a sequence of decision trees, iteratively correcting errors, and assigns feature importance based on the frequency of features used in tree splits that improve overall model performance. A higher importance score in GBM suggests a more significant impact on the model's predictions.



In contrast, with its layered neural network architecture, Deep Learning applies layer-wise relevance propagation or sensitivity analysis techniques to evaluate feature importance. The greater sensitivity or relevance indicates a more significant effect of the feature on the model's output.



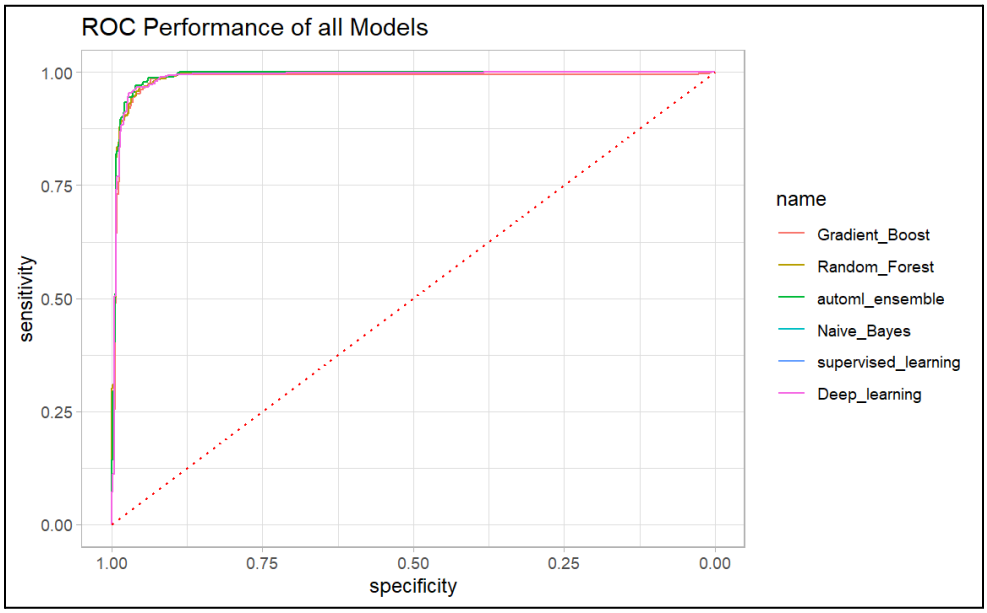
On the other hand, DRF constructs an ensemble of decision trees independently and calculates feature importance by measuring the reduction in impurity each feature brings to decision trees. Higher feature importance in DRF implies a more significant influence on decisions across the ensemble.



The GBM, Deep Learning, and DRF models agree that "Wildfire Annualized Frequency" is the most important predictor for predicting wildfires. This shared emphasis indicates that it significantly affects the likelihood of wildfires. Additionally, while GBM and DRF prioritize

"Tract Slope" in their predictions, there is a subtle difference between them. However, Deep Learning diverges in its feature selection, prioritizing ".isDeveloped.0" as a secondary influential factor. These differences highlight each model's unique perspective, with GBM and DRF focusing on the timing of wildfires and Deep Learning highlighting the affected area. Understanding these model-specific preferences helps us grasp the factors influencing wildfire predictions and choose models that suit the dataset and analysis goals.

Even when downsampling is employed to address class imbalance in our dataset, the ROC (Receiver Operating Characteristic) curve and AUC (Area Under the Curve) metrics are still important for evaluating a classification model's performance. These metrics consider the trade-offs between true positive and false positive rates at different classification thresholds. While downsampling helps balance the dataset, the ROC and AUC metrics provide valuable insights into the model's generalizability, regardless of class distribution. The selected models in the ROC plot demonstrate a higher AUC, closer to 1, indicating their ability to balance sensitivity and specificity across various thresholds. This highlights the models' proficiency in distinguishing between positive and negative classes, reinforcing their reliability in wildfire prediction scenarios.

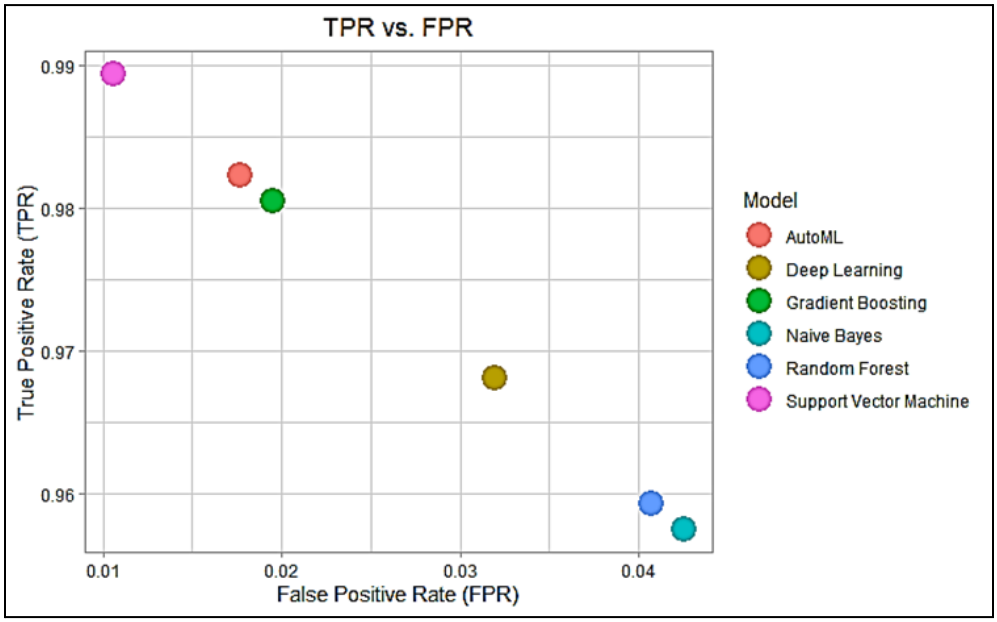


Summary

In our analysis of wildfire prediction models, we found that Gradient Boosting Machine (GBM), Deep Learning, and Distributed Random Forest (DRF) all agree that "Wildfire Annualized Frequency" is the most important predictor. While GBM and DRF focus on temporal aspects, Deep Learning emphasizes spatial dimensions and model-specific insights to guide the selection process based on the dataset and analysis goals. Although downsampling can address class imbalance, evaluating model performance using ROC and AUC metrics remains critical. The higher AUC values in Figure 3 confirm that the selected models can balance sensitivity and specificity across different thresholds, making them reliable for wildfire prediction scenarios.

However, downsampling produced biased results in the SVM model, indicating a moderate accuracy of 0.5566, with a trade-off between precision (0.5304) and recall (0.9894). The high recall suggests a solid ability to capture positive instances, but this comes at the cost of lower precision and an increased likelihood of false positives—the F1 score (0.6905) balances precision and recall. The elevated true positive rate (0.9894) highlights the model's effectiveness in identifying positive instances. However, the overall complexity suggests that further analysis and potential model refinement are necessary to optimize performance in the specific application context.

Model <chr>	TPR <dbl>	FPR <dbl>
Support Vector Machine	0.9894	0.0106
AutoML	0.9823	0.0177
Gradient Boosting	0.9805	0.0195
Deep Learning	0.9681	0.0319
Random Forest	0.9593	0.0407
Naive Bayes	0.9575	0.0425



The historical spatial and temporal data was used to test the efficacy of supervised machine learning models in predicting and controlling California wildfires. The AutoML model had the highest true positive rate (TPR) at 0.9823, followed by GB at 0.9805, DL at 0.9681, RF at 0.9593, and NB at 0.9575 (excluding SVM with the lowest model accuracy). These findings demonstrate the effectiveness of different models and emphasize the significance of selecting the appropriate model to tackle this complex task.

Future Applications

For this project, there are a few areas that can be expanded upon with additional resources and time. The items below can be further explored in future research on wildfire prediction:

- The availability of tract-specific meteorological data
- Expansion of stacked ensemble models and different combinations of base models
- Examining the feasibility of including real-time data in the model
- Fine-tuning the weighting of wildfire frequency and the effects on feature importance

Code Appendix

Please see the executed code in the Rpubs linked provided here!

Bibliography

Data resources

"Data Resources." Data Resources | National Risk Index, hazards.femHUD USPS ZIP Code

Crosswalk Files, www.huduser.gov/apps/public/uspscrosswalk/home. Accessed 2 Dec. 2023.a.gov/nri/data-resources#csvDownload. Accessed 2 Dec. 2023.

NCEI.Monitoring.Info@noaa.gov. "County Mapping: Climate at a Glance." County Mapping | Climate at a Glance | National Centers for Environmental Information (NCEI), www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/mapping/4/tavg/202212/12/rank. Accessed 2 Dec. 2023.

Geographic Names Information System, edits.nationalmap.gov/apps/gaz-domestic/public/search/names. Accessed 2 Dec. 2023.

"LANDFIRE ." LANDFIRE Program: Data Products - Vegetation, landfire.gov/vegetation.php. Accessed 2 Dec. 2023.

Works Cited

Calhoun, K. L.; Chapman, M.; Tubbesing, C.; McInturff, A.; Gaynor, K. M.; Van Scoyoc, A.; Wilkinson, C. E.; Parker‐Shames, P.; Kurz, D.; & Brashares, J. (2021). Spatial overlap of wildfire and biodiversity in California highlights gap in non‐conifer fire research and management. Diversity and Distributions, 28(3), 529–541. https://doi.org/10.1111/ddi.13394

Chen, Bin, and Yufang Jin. "Spatial Patterns and Drivers for Wildfire Ignitions in California." Environmental research letters 17, no. 5 (2022): 55004–

Syphard, Alexandra D, Timothy Sheehan, Heather Rustigian-Romsos, and Kenneth Ferschweiler. "Mapping Future Fire Probability Under Climate Change: Does Vegetation Matter?" PloS one 13, no. 8 (2018): e0201680–e0201680.

Tavakol Sadrabadi, Mohammad, and Mauro Sebastián Innocente. "Vegetation Cover Type Classification Using Cartographic Data for Prediction of Wildfire Behaviour." Fire (Basel, Switzerland) 6, no. 2 (2023): 76–.

Malik, Ashima, Megha Rajam Rao, Nandini Puppala, Prathusha Koouri, Venkata Anil Kumar Thota, Qiao Liu, Sen Chiao, and Jerry Gao. "Data-Driven Wildfire Risk Prediction in Northern California." Atmosphere 12, no. 1 (2021): 109–.

Lydersen, Jamie M., Brandon M. Collins, Carol M. Ewell, Alicia L. Reiner, Jo Ann Fites, Christopher B. Dow, Patrick Gonzalez, David S. Saah, and John J. Battles. “Using Field Data to Assess Model Predictions of Surface and Ground Fuel Consumption by Wildfire in Coniferous Forests of California: Fuel Characterizations and Wildfire.” Journal of geophysical research. Biogeosciences

Jaafari, A., Zenner, E. K., & Pham, B. T. (2018). Wildfire spatial pattern analysis in the Zagros Mountains, Iran: A comparative study of decision tree based classifiers. Ecological Informatics, 43(January 2018), 200-211. https://doi.org/10.1016/j.ecoinf.2017.12.006

Jiang, T., Bendre, S. K., Lyu, H., & Luo, J. (2021). From static to Dynamic prediction: Wildfire risk assessment based on multiple environmental factors. 2021 IEEE International Conference on Big Data (Big Data). https://doi.org/10.1109/bigdata52589.2021.9672044