# Introduction

Wildfires present a major environmental hazard in Northern California. In recent years Northern California has seen a stark increase in dangerous wildfire ignitions with the state’s three largest wildfires to date having all ignited in the region since 2018. The 2020 wildfires have grown to a record-breaking scope with an area the size of almost 8500 being burnt by the end of the season. This makes the 2020 wildfire season the largest and most devastating in the recorded history of the state (Yan et al., 2020). There is mounting evidence for climate change creating environmental conditions that favor wildfire ignitions and reduce forest resilience to wildfires (Stevens-Rumann et al., 2018).

Understanding which factors contribute to wildfire ignitions is thus essential for policymakers. Statistical modeling with regression methods has long been used to predict wildfire susceptibility of affected areas. More recently, advances in Machine Learning and data collection have made a new array of modeling methods and databases available for predictive modeling. Additionally, these techniques also facilitate a better understanding of the various intersecting factors that contribute to wildfire ignitions. The goal of this study is to make use of both statistical modeling (logistic regression) and machine learning methods (Random Forest and Gradient Boosting) to predict local wildfire susceptibility in Northern California. This study aims to answer the following questions:

* *How do different algorithms perform at modeling wildfire occurrence in Northern California?*
* *Which areas in Northern California are predicted to be the most susceptible to wildfire ignition risk?*
* *Which predictors contribute the largest predictive power for modeling wildfire occurrence in Northern California?*

The following section 2 introduces a collection of previous studies that investigate wildfire modeling using a similar methodology in the context of different study areas and periods. Section 3 details the data collection and preparation process, after giving a brief overview over the chosen study area of Northern California. Section 4 elaborates this study’s approach to modeling wildfire occurrence and discusses the chosen classification algorithms and evaluation methods. Section 5 reports the obtained results while section 6 discusses their implications and provides an interpretation. Section 7 concludes.

# Literature

Previous studies have used statistical modeling and machine learning methods to estimate fire susceptibility for various areas and time periods at both a local and a regional scale. Logistic regression has been a popular method to model wildfire occurrence. It is suited to predict binary variables and estimates a probabilistic output, which can be interpreted as the estimated risk value associated with a given sample (Cao et al., 2017; Catry et al., 2009; Martínez et al., 2009; Rodrigues & de la Riva, 2014; Syphard et al., 2008). Non-parametric Machine Learning algorithms such as tree-based ensemble models like Random Forest and boosted trees have seen a rise in popularity and are often being compared to baseline predictions based on regression models (Cao et al., 2017; Oliveira et al., 2012; Rodrigues & de la Riva, 2014).

One example of the combined use of statistical modeling and Machine Learning in the field of wildfire modeling is the study focused on Mediterranean Europe performed by Oliveira and co-authors (Oliveira et al., 2012). They used the average density of fire occurrences at the NUTS3-level of provinces as the dependent variable, as a proxy of fire ignition. Oliveira and co-authors applied both linear regression and Random Forest (based on regression trees) at a very large regional scale. In their study they aim to predict the probability of fire occurrence for spatial units at a 10 resolution. They use a mixture of environmental and human-centric variables for modeling with predictors that cover topography, land cover, climate, infrastructure, demography, and socioeconomics. They find that Random Forest outperforms their regression model. They list precipitation, soil moisture, unemployment rate and road density as their most important predictors, asserting the importance of both environmental and anthropogenic factors on wildfire occurrence (Oliveira et al., 2012).

A more focused look at the Mediterranean was taken by Tonini and co-authors, focusing on this single region of Liguria (Tonini et al., 2020). They leveraged data on burned areas from a 20-year period (1997-2017) to assess wildfire susceptibility of the region, dividing the data into subsets for the winter season (November-April) and the summer season (May-October). Unlike the study conducted by Oliveira and co-authors, who used the continuous variable of average density of fire occurrences as the target variable, Tonini and co-authors designed their study as a binary classification case, distinguishing between *burnt* and *unburnt* areas. Since pixels on their created GIS-map are treated as individual samples, this binary approach allows for the creation of very detailed fire susceptibility maps, since for every individual pixel a probabilistic value for being *burnt* is calculated by the model. The authors used variables describing the topography, land cover and the distance to anthropogenic features such as urban areas and roads as predictors. No variables capturing demographics or socioeconomics were used in this study. All models were estimated using Random Forest. Models were evaluated by predicting a set of test data consisting of the last 6 years of analysis (25% of the entire data). The final model showed good global predictive power in both seasons, discriminating between burnt and unburnt areas within the 75th percentile. The authors stress the importance of using cross validation to mitigate spatial autocorrelation with resampling (Tonini et al., 2020).

Cao and co-authors compare multiple modeling methods to estimate wildfire susceptibility in the South Chinese province of Yunnan (Cao et al., 2017). They disregard anthropogenic factors and include only environmental predictors at the scale of a single Chinese province. They apply both logistic and probit regression, Random Forest and Artificial Neural Networks to model wildfire risk. For model evaluation the authors use a cost-sensitive misclassification metric, concluding that Random Forest performed the best at predicting the binary ignition target. Cao and co-authors stress the importance of subsampling methods to resolve issues of class imbalance by balancing out the large number of non-ignition samples (Cao et al., 2017).

The Iberian peninsula has been investigated in a multitude of studies, being among the areas with the highest concentration of wildfires in Europe (Oliveira et al., 2012). Martínez and co-authors’ approach explicitly investigates only human-caused wildfires in Spain (Martínez et al., 2009). Their models are trained exclusively with anthropogenic predictors, ignoring any environmental factors. They apply logistic regression to estimate their binary target variable, finding that agricultural landscape patterns and rural development processes are the most important predictors for human-caused wildfire ignitions in Spain (Martínez et al., 2009).

Rodrigues and de la Riva expand on the work done by Martínez and co-authors by estimating models for human-caused wildfire occurrences in Spain with Random Forest, boosted trees and Support Vector Machine in addition to logistic regression (Rodrigues & de la Riva, 2014). Like Martínez and co-authors they only include anthropogenic predictors in their data. They find that logistic regression is outperformed by all other classifiers in terms of ROC-AUC (see section 3.4 for a discussion of various evaluation metrics in classification tasks). They conclude by recommending tree-based algorithms such as Random Forest and boosted trees while calling SVM less adequate due to the time-consuming optimization process and comparably worse predictive power (Rodrigues & de la Riva, 2014).

Catry and co-authors use a small set of predictor variables to predict the spatial patterns of wildfire ignitions in Portugal with logistic regression (Catry et al., 2009). They combine both environmental and anthropogenic variables, but explicitly make a point of using as few predictor variables as possible. They find that logistic regression provides satisfying prediction results (ROC-AUC of ~0.85) despite only using population density, land cover, distance to roads and elevation as predictors (Catry et al., 2009).

California has previously been the area of analysis in wildfire modeling studies (Malik et al., 2021; Syphard et al., 2008). Syphard and co-authors took a regional approach by investigating wildfire ignitions and frequency in the Santa Monica mountains in Southern California (Syphard et al., 2008). They used topological and climatic predictors, as well as distance to infrastructure. Their data is of very high granularity with a maximum resolution of 10m in their data set. They used logistic regression and found that while wildfire ignitions were best predicted with anthropogenic predictor variables, environmental data was better at predicting ignition frequency (Syphard et al., 2008).

A large selection of algorithms were employed by Ghorbanzadeh and co-authors in their study on wildfire susceptibility for Amol County in Iran (Ghorbanzadeh et al., 2019). They created their own data set based on the MODIS database, which provides daily shapefiles on global active wildfires. Predictor variables describing topography, meteorology, vegetation, and anthropogenic factors (such as distance to inhabited areas and roads) were used for modeling. This study too disregards demographic or socioeconomic features. A variety of models were trained on fire perimeter data from 2012-2017 at pixel level. Separate models were estimated using an Artificial Neural Network, a Support Vector and Random Forest. Overall, the RF model was found to achieve the highest prediction performance. Wildfire susceptibility maps were created based on the pixel-level probability of being classified as a burnt area. The authors conclude by recommending their workflow for other areas of interest comparing the performances of multiple algorithms tuned using cross validation (Ghorbanzadeh et al., 2019).

# Data

The aforementioned studies have demonstrated a high predictive power of environmental and anthropogenic variables for wildfire occurrences. This study follows the findings of the discussed literature by modeling wildfire occurrence with a combination of topological, meteorological, demographical, socioeconomical and political data, as well as data indicating the presence of and distance from the nearest anthropogenic infrastructure.

One of the primary challenges of this study was the compilation of a harmonized data set sourced from both GIS data (“geographic information system”) for all the spatial variables, as well as tabular data for the socioeconomic, demographic, and political predictors. All GIS data was processed and joined in the open-source software QGIS, whereas the tabular data was joined to the QGIS output in the R software designed for statistical computing.

## Study Area

The study area encompasses the northernmost counties of California, stretching from Sutter County north of Sacramento, up to the border of Oregon. This diverse region is made up of 18 counties[[1]](#footnote-1) with a combined area size of roughly 113’380 . The area typically has a milder climate with higher precipitation and more humid weather than the southern counties. Conifer forests, oak woodland and shrubland are the most common flora, although there is a considerable variation in the dominant vegetation patterns across the area. National forests and rugged mountain ranges dominate the landscape in this northern part of the state, with the foothills of the Sierra Nevada stretching up to the very north of the state.

This area has seen a large number of devastating wildfires in recent years. The three largest wildfires in the history of the state have all erupted in the study area since 2018. The Dixie fire, ignited in the Sierra Nevada in summer of 2021, has evolved into the largest single-source fire in the recorded history of California (Bermel, 2021). It rivals the August Complex wildfire of 2020 in size and has grown to almost twice the size of the Mendocino Complex fire of 2018 (California Department of Forestry and Fire Protection, 2021g). The deadliest wildfire in the history of California ignited in Butte County in 2018, with 85 recorded deaths (California Department of Forestry and Fire Protection, 2021f).

Ein Bild, das Karte enthält.

Automatisch generierte BeschreibungUsing the QGIS software I divided the study area into a grid of squares each with an area size of 4, making up the units of observation of this study. I chose this area size for the individual units due to some of the environmental predictor data only being available at this resolution, as well as the implications on computational cost when increasing the granularity.

Figure 1: Study Area, with all recorded wildfire perimeters during the study period  
(California Department of Forestry and Fire Protection, 2019, 2021d)

## Target Variable

The occurrence of wildfire in the geospatial units during the study period serves as the target variable for this study. The period between the years 2010 and 2018 constitutes the period of analysis.

Data on wildfire occurrence were obtained from the *Fire Perimeters* data set, compiled, and provided by the Fire and Resource Assessment Program (FRAP), an initiative led by the California Department of Forestry and Fire Protection (California Department of Forestry and Fire Protection, 2021e). *Fire Perimeters* is the most complete and frequently updated database on wildfire occurrence in California. This data set is provided as a shapefile and displays the perimeters of all recorded ignitions in the area, along with harmonized data such as the exact date of a fire’s discovery, as well as its extinguishment. I created a subset of this data corresponding to the study area and period, including all recorded fires throughout each year. The location accuracy of the recorded wildfire ignitions makes this data well suited for spatial analysis as long as a suitable map projection is chosen before further processing. For this project I chose to use the “NAD 1983 California (Teale) Albers (Meters)” projection, which is recommended for statewide datasets of California due to its property of having the coordinate system’s origin at the center of the state (Patterson, 2021).

After projection I used the QGIS spatial analysis join algorithm to register all intersections of wildfire perimeters and the grid of 4 squares, which serve as the units of observation. The resulting table records all dates for which the 4 grid units have intersected with a fire perimeter. Note that this does not mean that a given 4 unit was completely covered by a wildfire perimeter (and hence was burned completely), but that at least a single wildfire ignition took place and was recorded within the bounds of that specific 4 square.

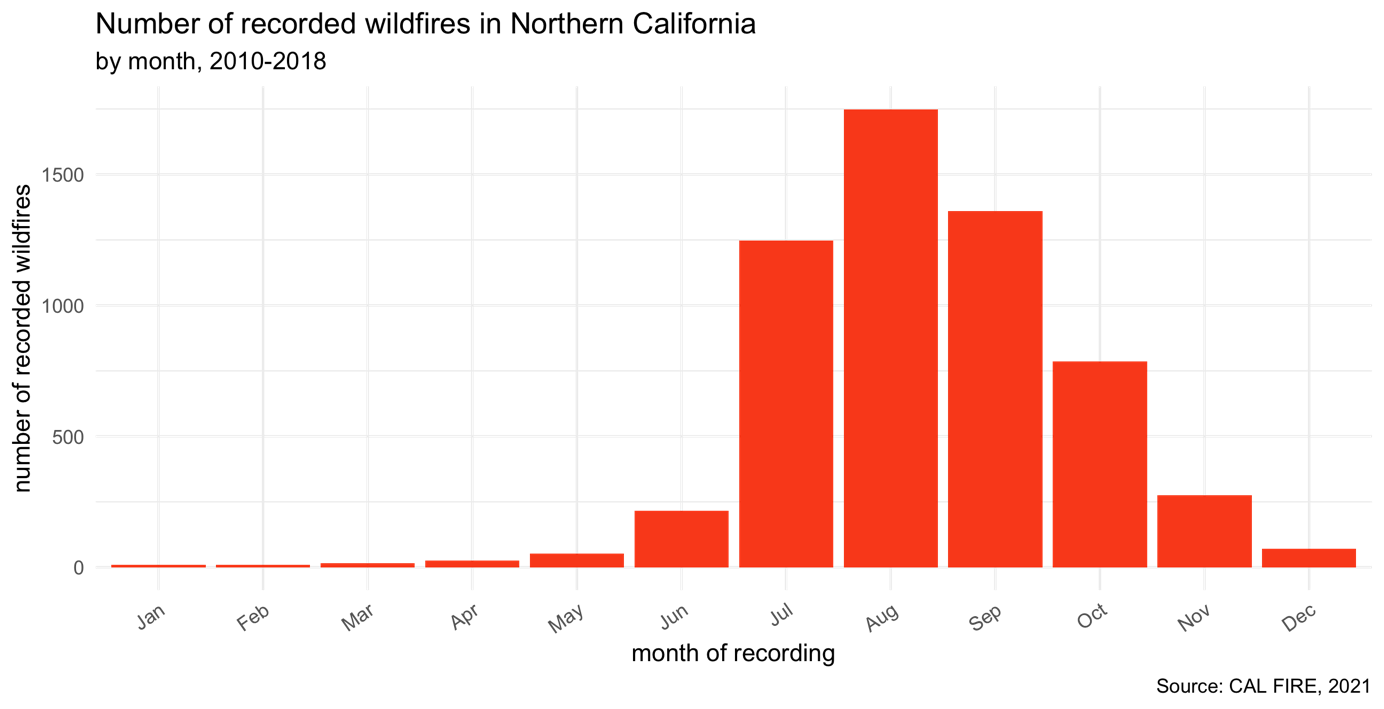
To further process this data, the resulting table had to be transformed. It is not the date of a wildfire ignition that is of interest for this study per se, but the wildfire ignition status of the grid elements during the observed intervals of the study period. To represent this within the data set, I transformed the data so that each sample represents the wildfire ignition status of a 4 square for each month of the study period of 2010 to 2018. This binary variable called *fire*, with the possible values of *fire* and *none*, serves as the target variable for all predictive models estimated for this study.

Figure 2: Monthly distribution of wildfire occurrences during the study period

The monthly distribution of wildfire ignition events shows a clear seasonality, as displayed in figure 2. The vast majority of active wildfires were registered during the summer and autumn months. I aggregated the monthly data to the seasonal level accordingly, reducing the overall number of samples while preserving as much information on wildfire occurrence as possible. Additional motivation for this aggregation was the fact that many predictors were not available at the monthly level, making the data set too granular for the variation contained within the predictor variables. In line with the study conducted by Tonini and co-authors, I assigned the period from May to October to the summer season and the period from November to April to the winter season (Tonini et al., 2020).

The number of recorded events (*fire*) and non-events (*none*) have proven to be strongly imbalanced, with non-events making up the vast majority of all samples in the data set. The aggregation to the seasonal level slightly improved this circumstance, increasing the share of samples reporting an active wildfire from 0.72% to 2.11%. Despite this, the seasonal data set still musters a high imbalance in the target variable’s values. This is common in cases of extreme-events prediction, as Cao and co-authors have also addressed (Cao et al., 2017). The implications of this circumstance on the modeling process and different strategies for addressing potential problems are discussed in depth in chapter 3 which details the methods I used for data pre-processing and model evaluation.

## Predictor Variables

I compiled a set of 54 predictor variables in total. Not all of these variables were ultimately used for modeling. Chapter 3 discusses predictor selection and why some of these variables were left out of the modeling process. These predictor variables are chosen both for their documented use in previous studies on wildfire modeling, as well as their availability for the study area of Northern California during the study period of 2010 - 2018.

These data were acquired at the highest available granularity in order to introduce as much variance into the final data set as possible. I included predictors of multiple categories, all of which are relevant to the occurrence of both human-caused wildfire ignitions and naturally occurring wildfires, similar to the study conducted by Oliveira and co-authors (Oliveira et al., 2012). The included predictor variables are environmental data (including both topographic, meteorologic and data concerning land cover), infrastructure data (both the proximity to human-made infrastructure and dummy variables indicating the presence of infrastructure in the units of observation), as well as demographic, socioeconomic and political data for the study area.

### Environmental Predictors

Topographical features such as elevation are important predictors of spatial patterns of fire. They account for local variations in climate, in addition to influencing ground flammability through their impact on soil and fuel moisture and the vegetational distribution of land cover (Oliveira et al., 2012; Syphard et al., 2008; Whelan, 1995). A digital elevation map of California at a 90m resolution based on satellite imagery was compiled by the National Aeronautics and Space Administration (NASA) and the National Geospatial-Intelligence Agency (NGA). It is distributed as a raster band data set, where each pixel of the map corresponds to a numeric elevation value (NASA & NGA, 2000). This very high resolution means that the data had to be aggregated to the 4 level of the grid units of this study. The QGIS software provides tools to process raster data and toolsets to calculate the zonal statistics such as the average elevation value for each of the 4 squares of the grid.

Furthermore, I added topographical data on the presence of major bodies of water, such as lakes and rivers, to the map. Bodies of water act as natural fire barriers and directly influence soil moisture and vegetation in their vicinity. These data are provided as shapefiles by the California Department of Fish and Wildlife (California Department of Fish & Wildlife, 2015, 2018). I used these data two create two kinds of predictor variables: First I created dummy variables indicating whether an object of observation is intersected by either a lake or a river. In a second step, I calculated the distance of each 4 unit’s centroid to the nearest river and lake with the *v.distance* algorithm as implemented in QGIS. This provides an additional indicator to the presence of water bodies, that is numeric and continuous, as opposed to the logical dummy variables.

Local vegetation and land cover are often cited as being associated with fire occurrences, both natural and caused by humans (Martínez et al., 2009; Oliveira et al., 2012; Syphard et al., 2008). Due to the strong local variations in climate, land cover not only indicates the naturally occurring fuel types, but also the various biomes found in Northern California. I hence included categorical data on the land cover and land use of the study area. The data was compiled by the Department of Geography at the University of California as a single shapefile, depicting the canopy dominant vegetation species for the entire state (Department of Geography UC Berkeley, 2014). The *California Wildlife Habitat Relationships* data set provides a detailed classification of tree dominated, shrub dominated, herbaceous dominated, aquatic, developed and non-vegetated habitats, each with their own subcategories. Due to this highly detailed breakdown of the dominant land cover, I used this data as a proxy for the primary fuel type within the 4 units of observation. I used QGIS to determine the most frequent land cover type for each unit.

Meteorological factors are well documented predictors of wildfire occurrence, as they affect fuel accumulation and ground moisture, creating the conditions that may favor or hinder wildfire ignitions and their spread (Oliveira et al., 2012; Syphard et al., 2008; Vilar et al., 2010). The WorldClim database offers monthly historical temperature and precipitation data at spatial resolution of 2.5 minutes (roughly 21) in raster format (Fick & Hijmans, 2017). Due to the large number of raster layers (monthly interval, eight-year study period, three data sets), these predictors had to be constructed iteratively, using QGIS’ python interface to extract the mean values of minimum temperature, maximum temperature, and mean precipitation for each unit of observation from all raster layers.

### Infrastructure Predictors

Access to roads is often seen as a driver of economic activity and a proxy for economic development due to facilitating access to markets and enabling migration[[2]](#footnote-2). In the context of fire occurrence, road access and the distance to roads are frequently used predictor variables, since these factors also determine the speed of response for a given fire containment strategy (Martínez et al., 2009; Oliveira et al., 2012). The Californian road system is well documented and provided as a shapefile containing all major roads (MTFCC codes S1100 and S1200) by the U.S. Census Bureau (US Census Bureau, 2015). Similar to how the GIS data on rivers and lakes was processed, I used this data set to both create dummy variables indicating the presence of a major road for each unit and calculate the distance from each unit’s centroid to the nearest major road. The same process was repeated for data set displaying powerlines and fire stations, resulting in predictor variables indicating both their presence (binary) as well as the distance from each unit’s centroid (California Department of Fish & Wildlife, 2020; California Department of Forestry and Fire Protection, 2021d). Powerlines are a frequently cited cause of wildfire ignitions and have been used to construct predictor variables in similar studies (McFall-Johnsen, 2019; Oliveira et al., 2012; Texas Wildfire Mitigation Project, 2014).

For recreational routes, campgrounds, picnic sites and state parks I only determined their presence, as I do not expect these recreational structures to have any continuous effect if they’re not present – unlike powerlines, where larger distances can function well as a proxy for a lack of economic development and human activity (California Department of Parks and Recreation, 2019a, 2019b, 2019c, 2021). I also calculated the distance to the nearest city boundary for each unit’s centroid (Spidle, 2019).

The FRAP database maintained by CAL FIRE offers a variety of shapefiles indicating institutional measures taken to address wildfire risk. I created a variable indicating a unit’s status of belonging to an area of state, local or federal emergency response responsibility, as provided by FRAP (California Department of Forestry and Fire Protection, 2021c). Similarly, I added data showing wildland fire protection areas and created categorical variables for each unit indicating the responsible agencies and to what level of government these agencies belong (California Department of Forestry and Fire Protection, 2021b). The *firewise* data set reports the date at which communities have formulated a “community wildfire protection plan” which I turned into a binary variable indicating whether such a protection plan had been put in place for a unit of observation (California Department of Forestry and Fire Protection, 2018). FRAP also offers a data set based on the *Forest Fuels and Species Conservation* program, which reports the area and duration of active fuel removal activity, which I turned into a dummy variable indicating active fuel removal for each unit (California Department of Forestry and Fire Protection, 2021a). Finally, I created dummy variables indicating whether a unit was part of an area protected for open space purposes or registered for land easement (Green Info Network, 2021b, 2021a).

### Demographic and Political Predictors

Population density is an especially important predictor for human-caused wildfires, as it describes the distribution of potential “causative agents” for fire ignitions (Oliveira et al., 2012: 121). I obtained data on population density from the WorldPop database, which provides annual raster data at a 1 resolution in raster format (WorldPop, 2020). I then calculated the average population density per 4 unit using QGIS. Additionally, I added the annual population growth at the county level to the data set as a proxy for local economic opportunity (California Department of Finance, 2012, 2021). Martínez and co-authors’ study on human-caused forest fires in Spain included a variety of housing data as their predictors (Martínez et al., 2009). In a similar fashion, I include annual data on the average vacancy rate at the county-level, as well as the average number of people per household at the county level as predictors (California Department of Finance, 2012, 2021).

To measure the political leanings of the inhabitants of the study area I included data on the share of registered democrats and republicans at the voting precinct level, as recorded for each biennial statewide general election (Statewide Database, 2020). California’s direct democratic instruments allow for operationalization of political leanings beyond the mere party spectrum by taking vote shares on referendums into account. Californians are able to propose laws and constitutional amendments by way of so-called “ballot initiatives”, even without the support of the Governor or the Legislature (Bonta, 2011). Out of all the ballot initiatives that were voted on during the study period I selected a subset of four initiatives that pertained to environmental issues specifically, gauging the environmental sensibilities at the county level:

|  |  |  |
| --- | --- | --- |
| Proposition number | *Year* | *Description* |
| 21 | 2010 | Vehicle License Fee Increase, dedicated to state parks and wildlife programs |
| 23 | 2010 | Suspension of GHG- emissions reduction law, until California's unemployment rate decreases to 5.5% |
| 65 | 2016 | Dedication of Revenue from Disposable Bag Sales to Wildlife Conservation Fund |
| 67 | 2016 | Ban on sale of plastic bags. |

Table 1: Californian Ballot Initiatives relating to environmental issues during the study period

For each of these ballot initiatives I added the share of Yes-votes at county level to the data set, as published by the California Secretary of State platform as part of the statement of vote (California Secretary of State, 2010, 2018).

### Socioeconomic Predictors

Previous studies have found the rate of unemployment to be an important predictor for wildfire occurrences (Martínez et al., 2009; Oliveira et al., 2012). Oliveira and co-authors mention two specific channels through which the rate of unemployment may factor into wildfire risk: Both as a proxy for social conflict, which can cause increases in arson as part of generic vandalism, or arson as a deliberate strategy to increase the local demand for firefighters, thus furthering a unemployed person’s chance of finding employment (Oliveira et al., 2012: 125). Besides the rate of unemployment, I also constructed the monthly growth in unemployment rate as an additional predictor. The data on monthly unemployment rates at county level are provided by the California department of employment and development (California Department of Employment & Development, 2021b).

The same department also provides monthly labor data denoting the share of employed workers by industry at county level (California Department of Employment & Development, 2021a). I also added these data as predictor variables. Implicit zeros, which appear in the data base as missing values, had to be introduced explicitly during the data cleaning process.

# Methods

Following the examples from the discussed literature the aim of this study is to model wildfire occurrence in Northern California as a binary classification case. The resulting probabilistic output was used to create fire susceptibility maps of the study area.

The data were split into subsets of training and testing sets. All data from the years 2010-2016 were used for training the models. The data from the years 2017 and 2018 were used for evaluating the estimated model by predicting the target variable for the testing data. This procedure follows the example set by Tonini and co-authors who split their data into training and test set based on year as it mirrors the process of forecasting wildfire risk on past data for future seasons (Tonini et al., 2020).

The training set was used to estimate increasingly complex and sophisticated models, using three algorithms that are appropriate for binary classification cases to model wildfire occurrence: logistic regression (GLM), *Random Forest* (RF) and *xgboost* (XGB). These classifiers have a proven to be popular and effective modeling techniques for classification cases, as seen in the widespread application in the discussed literature on wildfire modeling and the broader Machine Learning community at large. For each of these algorithms models are estimated in four distinct steps:

1. A naïve model estimation strategy without resampling or hyperparameter tuning, using the heavily imbalanced training data (no subsampling).
2. A more informed model estimation strategy where the training data is resampled using 5-fold cross validation and the imbalance in the training set (split on year) is addressed with both upsampling and downsampling methods.
3. A modeling strategy focused on maximizing predictive performance by selecting hyperparameters from a tuning grid using grid search, again using resampling and subsampling methods to ensure stable results.
4. A final set of models with tuned hyperparameters and target class imbalance addressed by downsampling, but with the training and test data assigned randomly.

For this final step I chose to also estimate a final set of models for comparison with the data assigned to training and test set at random, as is common in Machine Learning evaluation.

## Logistic Regression

Logistic regression models the relationship between a binary variable and a set of independent predictor variables. Due to using the logit transformation of the binary target variable, this parametric, linear model can be used for estimating the probability of event occurrences as is done in this study.

Regression models have been widely used in cases that model wildfire occurrence, especially logistic regression (Catry et al., 2009; Martínez et al., 2009; Oliveira et al., 2012; Syphard et al., 2008). Since logistic regression uses maximum likelihood estimation, many assumptions of linear regression (which uses ordinary least squares estimation instead) do not apply to logistic regression. Assumptions of logistic regression include independent errors, the absence of multicollinearity among the predictors, linearity in the logit for continuous variables, and a lack of strongly influential outliers (Stoltzfus, 2011). Logistic regression can be expressed as follows:

Equation 1: Mathematical notation of a logistic regression model

In the case of this study multicollinearity had to be addressed during data preprocessing, as data exploration identified many predictors as strongly correlated. Data exploration also revealed that the distributions of the distance predictors are heavily skewed. A testing of multiple methods showed a power transformation to be the most effective in creating a more normal-like distribution for those variables. The preprocessing steps are handled for each model separately, taking the individual characteristics of each algorithm into account. Most of the preprocessing steps were similar, however.

For all models, the dummy variables indicating the presence of infrastructure (such as roads) or topological features (such as lakes) were removed, as long as there is a corresponding distance variable present in the data set. These variable pairs are highly correlated, and the distance variables provide more information and introduce more variance into the data compared to the dummies. These distance variables underwent a power-transformation for the regression models only, as logistic regression might benefit from predictors that have a more normal-like distribution (Kuhn & Silge, 2021). In a next step predictors with zero variance were removed, as were strongly correlated predictors (with a threshold of 0.75), similar to Oliveira and co-authors’ study (Oliveira et al., 2012). Categorical predictors were turned into dummy variables through one-hot encoding.

In order to balance out the classes in the target variable, models were estimated with both an upsampled, as well as a downsampled set of training data. Upsampling was done using the *SMOTE* algorithm, creating additional synthetic observations. For the downsampled training set the *NearMiss 1* algorithm was used, which retains observations of the majority class with the smallest distance to the k-nearest neighbors of the minority class. The training data for both the upsampled and the downsampled models underwent an additional step of *Tomek’s Links* removal, a procedure that removes majority class observations that are the nearest neighbor of an observation belonging to the minority class. This procedure was intended to improve the classification boundary of the training data and hence the predictive power of the trained model.

During hyperparameter tuning both and regularization were introduced to logistic regression, resulting in an Elastic Net model. For this specific case all predictors underwent normalization, as is necessary for regularized logistic regression. Elastic Net combines both the Ridge and *LASSO* penalties to shrink the estimated coefficients. Both the size of these penalties as well as the mixture between the two are tunable hyperparameters (*penalty* and *mixture*) whose optimal values I determined using a grid search strategy.

## Random Forest

Random Forest is a non-parametric ensemble learning algorithm, comprised of a set of multiple Decision Tree models. These submodels aim to create decision rules that divide the data into homogenous subsets, by splitting based on Gini impurity at each decision node (Kuhn & Johnson, 2013). This algorithm is known to suffer from large variance, meaning that the smallest changes to the training data can have a large influence on the overall model fit. Random Forest is built upon the idea of exploiting this property by letting its *weak classifiers* vote on the ensemble’s predictions based on their own, highly variant mode fits. The singular trees are made weak due to limited access to the training data, as each Decision Tree is only trained on a set number of bootstrapped predictor variables (Kuhn & Johnson, 2013). This approach is intended to counteract overfitting on the training data as no tree is trained on the entire data set. This strategy uses the bias-variance-tradeoff in Machine Learning to drive down bias at the expense of increased variance, intended to ultimately result in more accurate and stable predictions (Kuhn & Johnson, 2013).

Random Forest does not share the assumptions of logistic regression, but fewer correlated predictors might improve the *variable importance* *scores* commonly reported by tree-based algorithms (Kuhn & Johnson, 2019). Strongly correlated predictors indicating the presence of infrastructure or topological features were hence removed from the training data for Random Forest as well, as are predictors with a correlation coefficient exceeding 0.75 and zero variance predictors.

Unlike logistic regression and *xgboost*, Random Forest does not require one-hot encoding of categorical variables. Despite this, due to the used software requiring exclusively numeric predictors for subsampling, one-hot encoding still had to be used for all Random Forest models except the naïve estimation. Subsampling was conducted in the same way as for logistic regression, with *SMOTE* and *NearMiss 1* being used in combination with *Tomek’s links* removal. Neither Random Forest nor *xgboost* require any additional preprocessing, highlighting this inherent strength of these non-parametric, tree-based models.

Random Forest has multiple hyperparameters available for tuning, such as the minimal number of data points required at each node to qualify for further splitting (*min\_n*), the number of bootstrapped predictors at each split (*mtry*), as well as the total number of trees grown in the ensemble (*trees*). During hyperparameter tuning the latter was kept constant at 500 trees due to computational restraints, while the two former parameters were tuned using grid-search.

## xgboost

*xgboost* is an optimized implementation of the boosted trees algorithm. Similar to Random Forest it is an ensemble model that relies on using multiple Decision Trees as weak classifiers(with a predictive power marginally better than random classification)to ultimately combine them into a strong ensemble classifier(Kuhn & Johnson, 2013). Despite this similarity in concept boosted trees algorithms do not grow independent trees in parallel, as Random Forest does. Instead the Decision Trees grown by boosted trees are fit sequentially with each subsequent tree attempting to minimize the loss of its preceding trees by placing more weights on their misclassified samples, updating the predicted values by adding the previous tree’s predictions to the predicted values of the current tree (Kuhn & Johnson, 2013). This approach has been wildly successful and popular in classification tasks. The *xgboost* implementation has been especially successful by taking gradient boosting to the extreme with parallelization, fast optimization and convergence and overall computational efficiency. Like Random Forest *xgboost* also creates weak classifiers by limiting the access of Decision Trees to training data, but unlike Random Forest it does so by sampling the number of observations accessible during each boosting iteration instead of bootstrapping predictor variables.

Like Random Forest, *xgboost* does not share the assumptions of logistic regression. Nevertheless, the same steps for predictor selection was used to remove highly correlated predictors from the training data and ensure reliable variable importance scores. The pre-processing steps for *xgboost* are the same as for Random Forest, with the exception of one-hot encoding of categorical predictors being uniformly applied all models, due to the requirements of the *xgboost* implementation.

*xgboost* shares the hyperparameters of Random Forest (*trees*, *min\_n* & *mtry*). Additionally it offers tuning the depth of each tree in the ensemble by setting the maximum number of possible splits (*tree\_depth*), the reduction in the loss function required for further splits (*loss\_reduction*), regularization of the individual decision trees by weighting the corrections made by each subsequent tree (*learning\_rate*), as well as the sample size of the data used for modeling within each boosting iteration (*sample\_size*). During hyperparameter tuning all of these were tuned using grid-search, except for the number of trees grown, due to the high computational cost associated with large ensembles of trees.

## Evaluation

Due to the heavy imbalances in the data set’s target variable specific precautions had to be taken to evaluate the estimated models properly. Since the models used in this study don’t directly predict classes, but class probabilities, the *Receiver Operator Characteristic* (ROC) curves were created to visualize predictive performance at varying classification thresholds (Kuhn & Johnson, 2013). ROC-curves for all model fits are displayed in the appendix. Similar to previous studies that use predictive modeling to estimate wildfire ignition risk, I used the *Area under the curve* (AUC) of ROC-curves as the primary metric for evaluation (Catry et al., 2009; Malik et al., 2021; Rodrigues & de la Riva, 2014).

*Accuracy* was disregarded as a primary metric for evaluating model quality, as models can achieve very high accuracy scores when predicting imbalanced data if all testing samples are uniformly predicted to belong to the majority class. In these cases, the *F-measure*, defined as the harmonic mean between *recall* and *precision*, is a more reliable score due to taking both the true positive rate as well as the reliability of positive predictions into account. The *F-measure* is defined as follows, where *TP* denotes true positives, *FP* denotes false positives and *FN* denotes false negatives:

Equation 2: Computation of the F-measure

Additionally, not only the imbalance between the classes in the target variable had to be considered, but also their associated classification cost. There is a higher cost associated with misclassifying wildfire ignitions as non-events than predicting a high probability for wildfire occurrence when in truth no fire ignition takes place. This circumstance can be built into model evaluation by weighting misclassifications in such a way that false negatives are penalized more heavily than false positives. I implemented such a metric by customizing the *classification cost* function with a cost matrix that penalizes false negatives twice as much as false positives.

ROC-AUC was used for evaluating both the finalized models and the hyperparameter combinations during tuning in order to maximize predictive power. The *F-measure* is also listed among the performance metrics for all model fits, along with its components *precision* and *recall*, as well as the *penalized classification cost* and *accuracy*. The performance of the best estimated models is also visually represented with confusion matrices, displaying a cross-tabulation of the predicted and observed classes. The complete collection of ROC-curves and confusion matrices can be found in the appendix.

# Results

TABLE X shows the final selection of predictor variables used for modeling, as present after the described data preparation steps. All models were estimated using these predictors, with some model-specific transformations applied as described in chapter 4. Recall that the data was split into training and testing subsets based on the year instead of random sampling.

## Resampling

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| .metric | GLM (down) | GLM  (up) | RF  (down) | RF  (up) | XGB (down) | XGB  (up) |
| accuracy | 0.77 | 0.719 | 0.873 | 0.986 | 0.827 | 0.904 |
| Classification cost | 0.276 | 0.307 | 0.234 | 0.047 | 0.234 | 0.158 |
| F-measure | 0.09 | 0.081 | 0.162 | 0.429 | 0.124 | 0.159 |
| precision | 0.048 | 0.042 | 0.09 | 0.585 | 0.067 | 0.092 |
| recall | 0.735 | 0.795 | 0.794 | 0.339 | 0.79 | 0.584 |
| ROC-AUC | 0.827 | 0.83 | 0.916 | 0.933 | 0.89 | 0.866 |

Table 2: Performance metrics based on 5-fold CV

Performance metrics based on resampling the training data with 5-fold cross validation show promising results for all model specifications. There is a clear model hierarchy with logistic regression being outclassed by all tree-based ensemble models in terms of ROC-AUC. Random Forest ranks highest in terms of ROC-AUC, regardless of the subsampling method used to balance out the classes of the target variable. Both the upsampled and the downsampled versions of the training data lead to ROC-AUC values beyond the 0.9 threshold when using the Random Forest classifier. RF-estimation based on the upsampled data achieves an average ROC-AUC of 0.933 over all five folds and a mean ROC-AUC of 0.916 when using the downsampled training data. The xgboost classifier falls short of surpassing the 0.9 threshold with a mean of 0.89 ROC-AUC over all five folds based on the downsampled data, outperforming the upsampling workflow which achieves a 0.866 mean ROC-AUC. Both subsampling methods lead to almost identical performance when used with logistic regression, achieving 0.83 mean ROC-AUC with upsampling and 0.827 mean ROC-AUC with downsampling.

Out of these preliminary results I rate the Random Forest classifier based on the downsampled training data as the most promising, despite it not having the largest ROC-AUC out of all tested model specifications. The much lower recall value of the specification using upsampling hints towards a tendency to classify most samples as non-events, which is not desirable in the case of wildfire risk prediction. The downsampled specification achieved a much higher recall value and lower classification cost, indicating a better ability to identify true positives, even if the ROC-AUC is comparably smaller.

## Hyperparameter Tuning

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| .metric | GLM  (down) | GLM  (up) | RF  (down) | RF  (up) | XGB  (down) | XGB  (up) |
| accuracy | 0.606 | 0.658 | 0.845 | 0.952 | 0.785 | 0.958 |
| Classification cost | 0.441 | 0.338 | 0.29 | 0.204 | 0.266 | 0.094 |
| F-measure | 0.14 | 0.15 | 0.182 | 0.059 | 0.166 | 0.01 |
| precision | 0.077 | 0.084 | 0.116 | 0.148 | 0.098 | 0.12 |
| recall | 0.789 | 0.745 | 0.425 | 0.036 | 0.526 | 0.005 |
| ROC-AUC | 0.734 | 0.755 | 0.735 | 0.729 | 0.728 | 0.722 |

Table 3: Performance metrics after tuning, predictions for test set

The performance metrics evaluated through 5-fold cross validation are not reflected in the results obtained after hyperparameter tuning and prediction of the test set. Performance is expected to increase after hyperparameter tuning, due to the most favorable hyperparameter combinations being selected. Despite ROC-AUC being used as the metric for hyperparameter selection, all ROC-AUC values are worse when evaluating performance by predicting the test set. Based on ROC-AUC the regularized elastic net regression model trained on upsampled data shows the best performance. The best hyperparameters for this model are the very small *penalty* of 0.0000000001 in combination with a *mixture* of 1, resulting in LASSO regression which shrinks the coefficients of unneeded predictor variables to zero.

The elastic net regression models are the only model fits with a recall larger than 0.53, hinting towards a general inability to identify true positives within the testing data. While the elastic net models achieve acceptable recall, they also come with the lowest precision values out of all model fits. This indicates that these models achieve their comparably high recall by overestimating the number of samples where fire ignitions occurred, leading to a large number of false positives among the predicted classes as seen in the confusion matrix depicted in Table 4. This fact is reflected in the low F-measure of these models, which is hampered by low precision despite comparably high recall.

|  |  |  |
| --- | --- | --- |
|  | Truth | |
| Prediction | Fire | none |
| Fire | 899 | 9858 |
| none | 308 | 18663 |

Table 4: Confusion Matrix of tuned Elastic Net Model (upsampled)

## Stratified random split

Clearly, these models are unable to satisfyingly classify fire ignitions and non-events in the data set designated for testing. This comes as a surprise after the performance estimates based on resampling the training data showed promising results for all model specifications. The superior performance achieved through resampling hints toward the reason for the drop in performance when predicting the test set: All classifiers appear to achieve much better results when evaluated on testing data that was not split based on year, but random sampling as is done during cross validation.

For this reason, I estimated another set of models with the exact same modeling specifications, except that the training and testing split being conducted at random and not based on year.

|  |  |  |  |
| --- | --- | --- | --- |
| .metric | GLM | RF | XGB |
| accuracy | 0.642 | 0.86 | 0.807 |
| Classification cost | 0.388 | 0.238 | 0.244 |
| F-measure | 0.092 | 0.201 | 0.155 |
| precision | 0.048 | 0.115 | 0.086 |
| recall | 0.833 | 0.812 | 0.82 |
| ROC-AUC | 0.809 | 0.914 | 0.894 |

Table 5: Performance metrics when using a randomized training-/test split and downsampled training data

Changing from time-based splitting to a randomized training and test split dramatically improves the performance metrics for all estimated models, as seen in Table 5. These models were trained on the downsampled training data set, which has proven to lead to comparable if not superior prediction results for the previously estimated models (as seen in Table 3).

Not only was predictive performance increased for all models, the changes to the training and test split also affected these models differently. The Random Forest classifier achieved the highest performance with a ROC-AUC of 0.914, closely followed by *xgboost* with a ROC-AUC of 0.894. Hyperparameter tuning determined the best combination of parameters to be *mtry* of 23 and *min\_n* of 4, given 500 grown decision trees in the ensemble. For *xgboost* a similar *mtry* with a value of 22 and a much larger *min\_n* of 29, combined with *tree\_depth* of 8, a *learn\_rate* of 0.0637, *loss\_reduction* of 0.000000916 and a *sample\_size* of 82.6% were determined to maximize the ROC-AUC. Elastic Net regression, which previously outperformed the other classifiers, profited the least from this change and ranked third with a ROC-AUC of 0.809. Hyperparameter tuning determined the ideal *penalty* value to remain at 0.0000000001 with an optimal *mixture* of 0.5, which results in an equal combination of LASSO and ridge penalties.

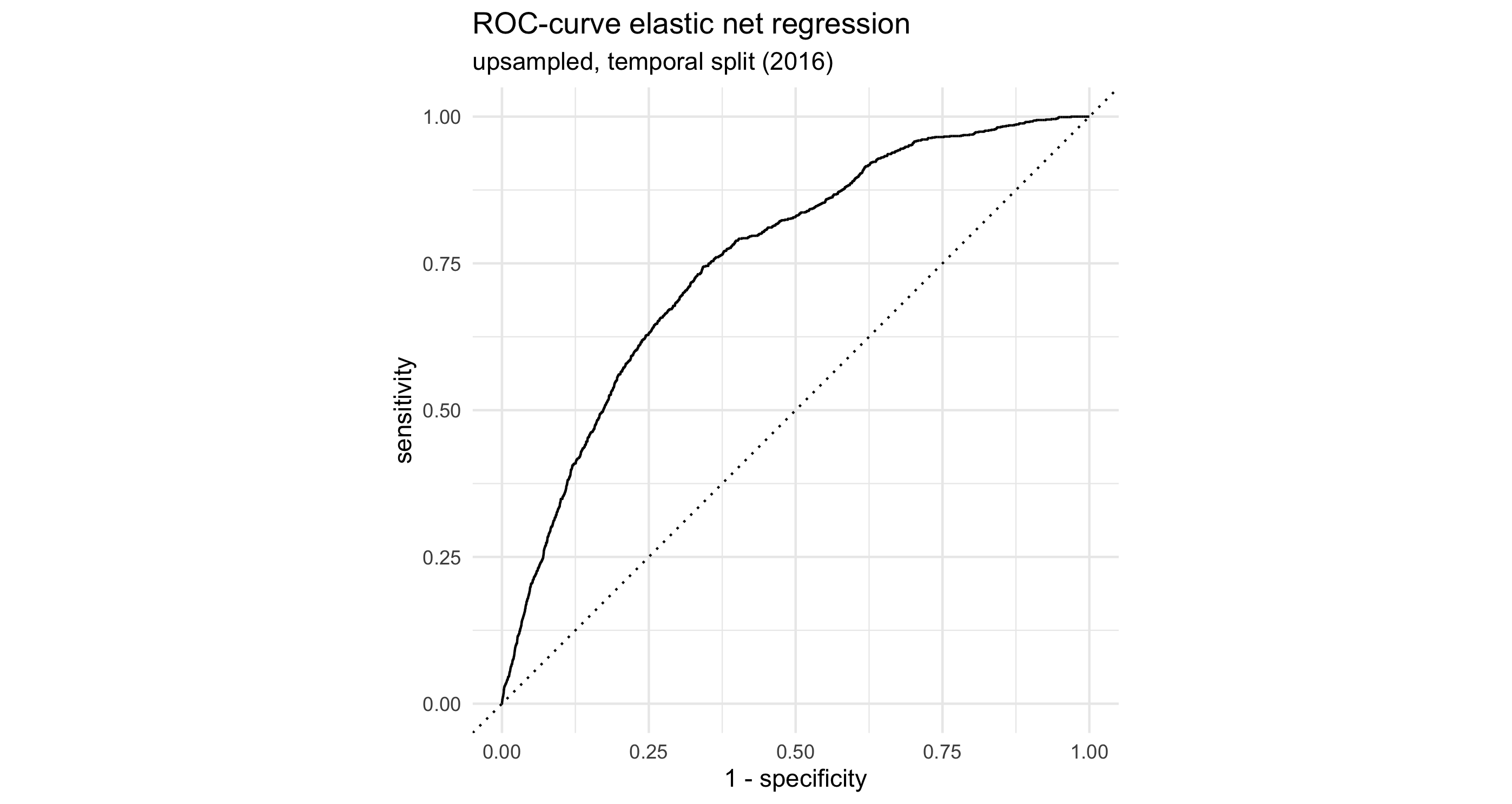
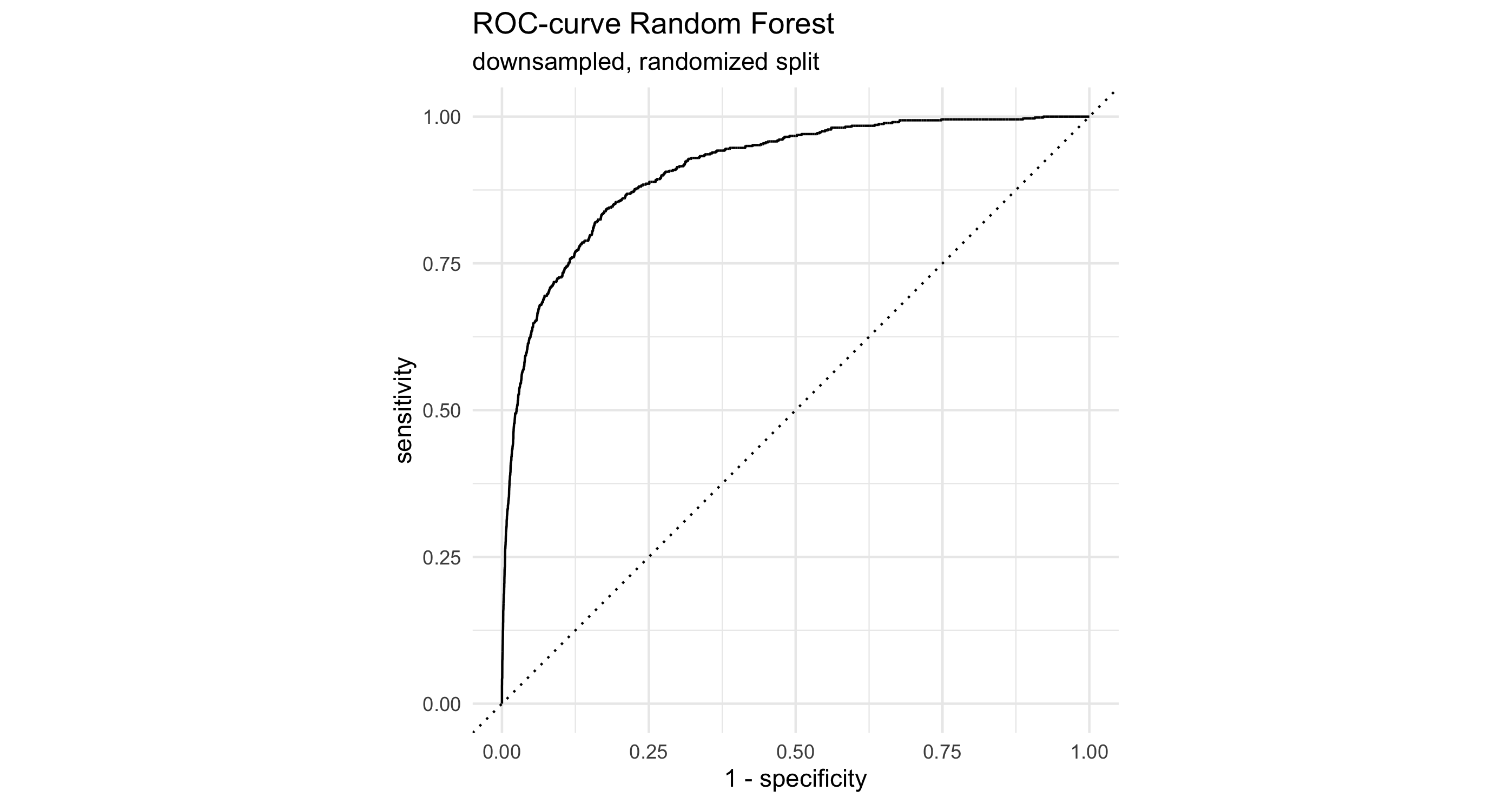
I ultimately rate the Random Forest as the strongest model. Not only did it achieve the largest ROC-AUC, but it also outperformed Elastic Net regression and *xgboost* in terms of every other metric except for recall, which still remained comparable. Random Forest achieved the highest accuracy, highest precision, and highest F-measure, despite the slightly lowered recall. This combination of metrics indicates that the Random Forest model achieved the best classification boundary among these models while minimizing the penalized classification cost metric.

|  |  |  |
| --- | --- | --- |
|  | Truth | |
| Prediction | Fire | none |
| Fire | 524 | 4038 |
| none | 121 | 25045 |

Table 6: Confusion Matrix of Random Forest Model (random sampling split)

This is reflected in the confusion matrix of the Random Forest model, as seen in Table 6. This confusion matrix also demonstrates that despite the optimized classification boundary this model still overestimates the number of fire ignitions, leading to a large number of false positives.

Figure 3: ROC curves for best performing models under temporal and randomized split



Compared to the best performing Elastic Net model trained on the time-based data split depicted in Table 4 the false positive rate is much smaller, resulting in larger specificity. Figure 3 illustrates the difference in terms of ROC-AUC between the best performing model under a temporal training / test split (Elastic Net regression) and the best performing model with a randomized split (Random Forest). These results are much more in line with the performance estimates based on resampling as seen in Table 2, both in terms of the individual metrics as well as the ranking of the classifiers.

# Discussion

The results of this study suggest that the risk for wildfire ignitions in northern California is distributed in a spatial PATTERN that XXX.

## Model performance

Comparison of the modeling results shows that these models struggle to predict wildfire ignition risk if there were observations no from the predicted time period in the training set. Conversely, predictive performance is high if data from the predicted time period was present in the training data. This also holds true without hyperparameter tuning, as the performance metrics based on resampling show that predictive performance is high if the data is split into training and validation sets at random.

In the setting of the time-based data split upsampling the training data to address the target variable’s class imbalance provided little benefit. Both the upsampled Random Forest and *xgboost* ended up predicting almost exclusively the majority class leading to high accuracy but very low recall. Logistic regression outperformed its non-parametric competitors no matter the subsampling method. Even when downsampling the majority class instead, the tree-based ensemble models achieved such low *recall* that in the best case only around half of the ignition samples were predicted correctly. Despite Random Forest and *xgboost* taking a clear lead in predictive power once the training and test sets are assigned at random, they fail at reliably predicting ignitions if the split is time-based.

This has multiple implications on forecasting wildfire ignition risk. The difference in performance between the two data splitting methods implies that there is an annual component to wildfire occurrence that cannot be sufficiently explained with the predictor variables I collected. The tree-based ensemble models appear to be prone to overfitting on the training data if the data split is conducted based on year, as these models failed to generalize patterns leading to wildfire ignitions that can be applied to subsequent years.

This problem is rectified once the training and test data sets are assigned at random. Both Random Forest and *xgboost* are shown to generalize well based on the training data, both classifying over 80% of the ignition samples in the test set correctly. This is also the case for the Elastic Net regression in this modeling setting, although its significantly lower precisionscore means that these true positive predictions come at a price of many more false positives. Even if false positives are not as expensive as false negatives, they are still best avoided if possible.

The fact that the non-parametric models outperform logistic regression in the setting of a randomized data split also implies that there are non-linear relationships between the predictors and the target variable, as has been noted in multiple previous studies on modeling wildfire ignition risk (Oliveira et al., 2012; Syphard et al., 2008; Vilar et al., 2010).

## Variable Importance

Regression outputs are disregarded in this section due to penalized estimation making standard errors unreliable for meaningful interpretations of regression coefficients and their significance. The best performing Random Forest and xgboost models in the temporal split setting share 13 predictors among their 15 most important predictor variables. Even among their five most important predictors they have four in common: the average temperature, average precipitation, the current season, and the distance to the nearest city. Random Forest also lists unemployment growth at county level among its five most important predictors, while *xgboost* includes the percentage of registered democrats at voting district level. Both of these predictors are found among the top 15 predictors of the other classifier respectively. Population density and population growth only rank in the top 15 for *xgboost*, but not for Random Forest, despite being often cited as important predictors in previous studies.

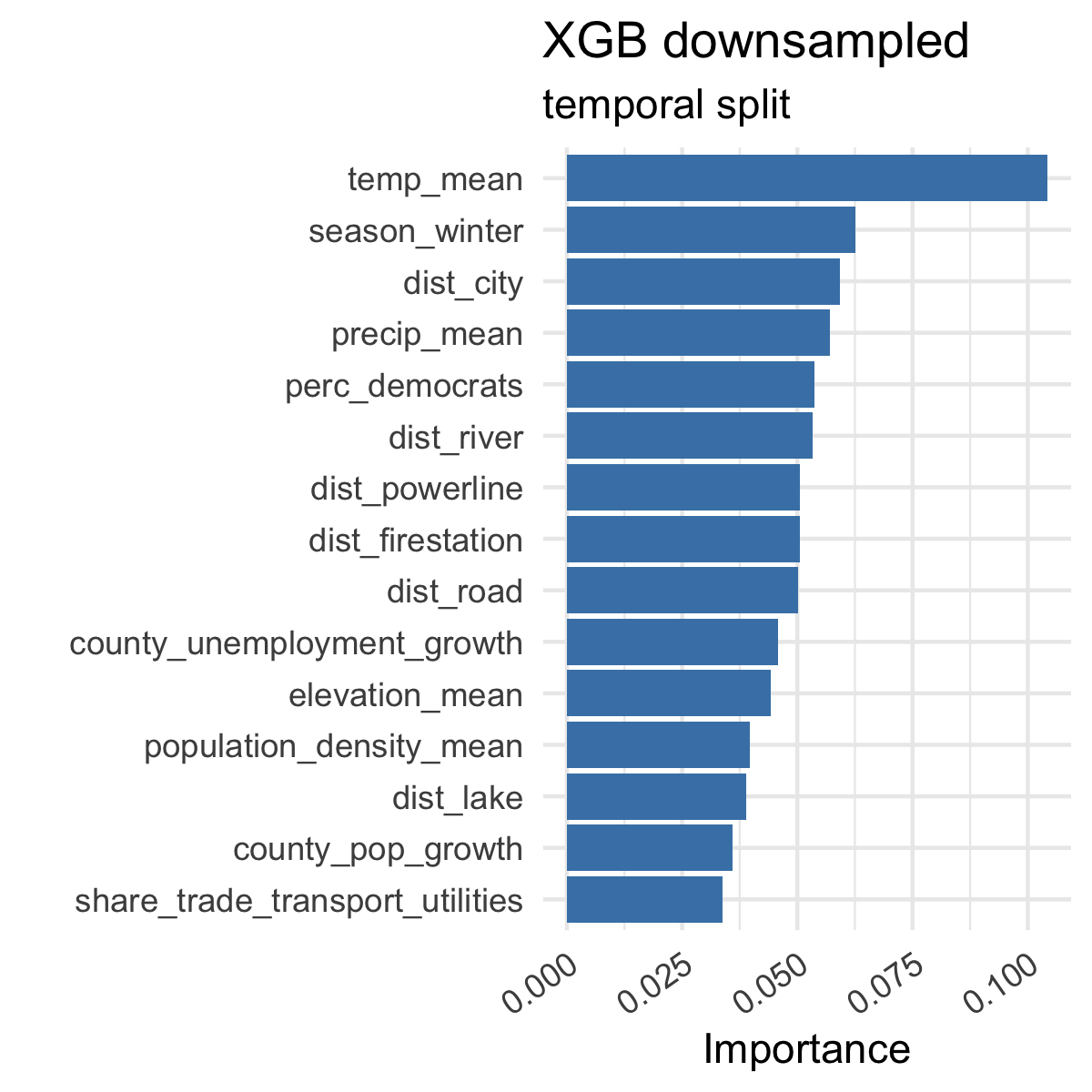
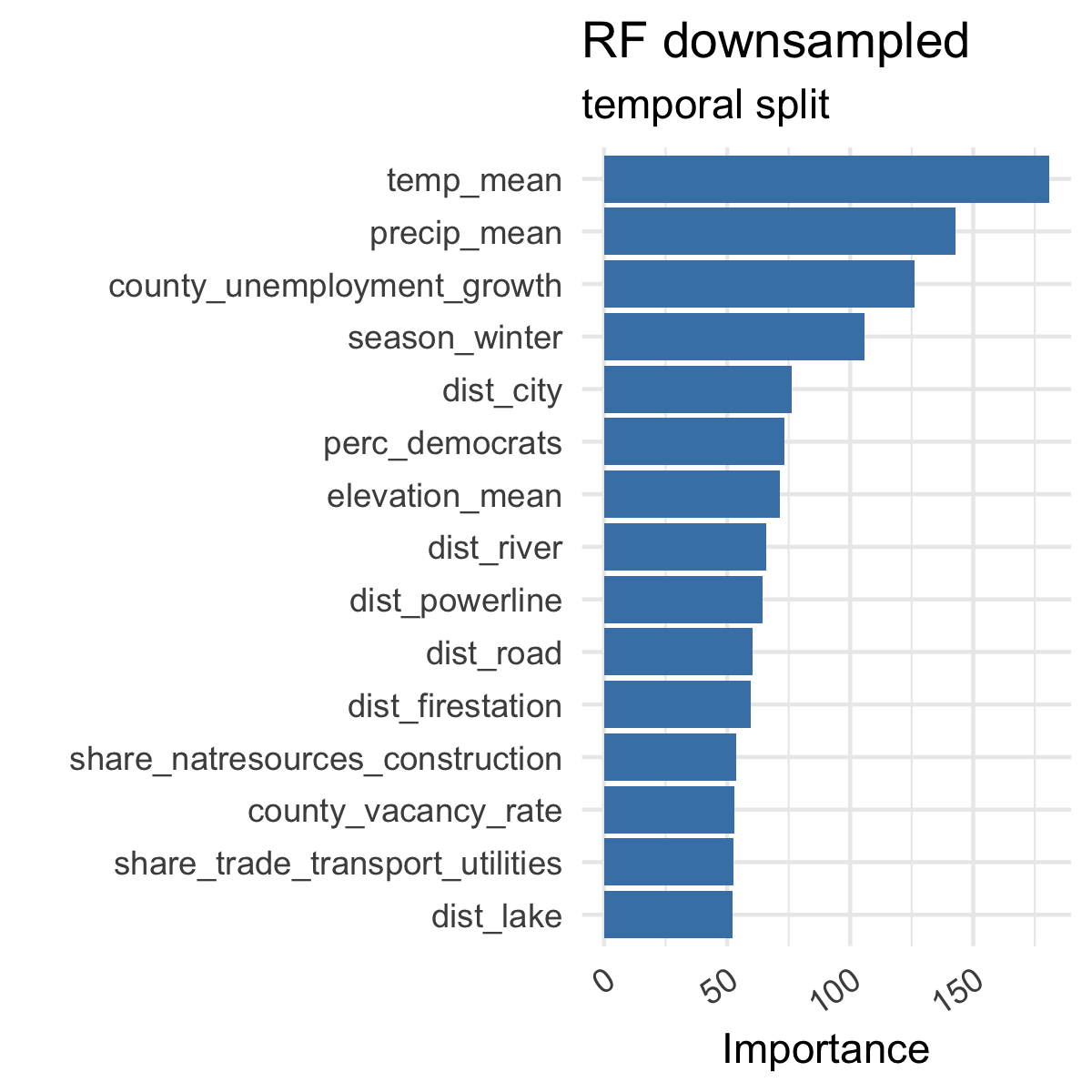
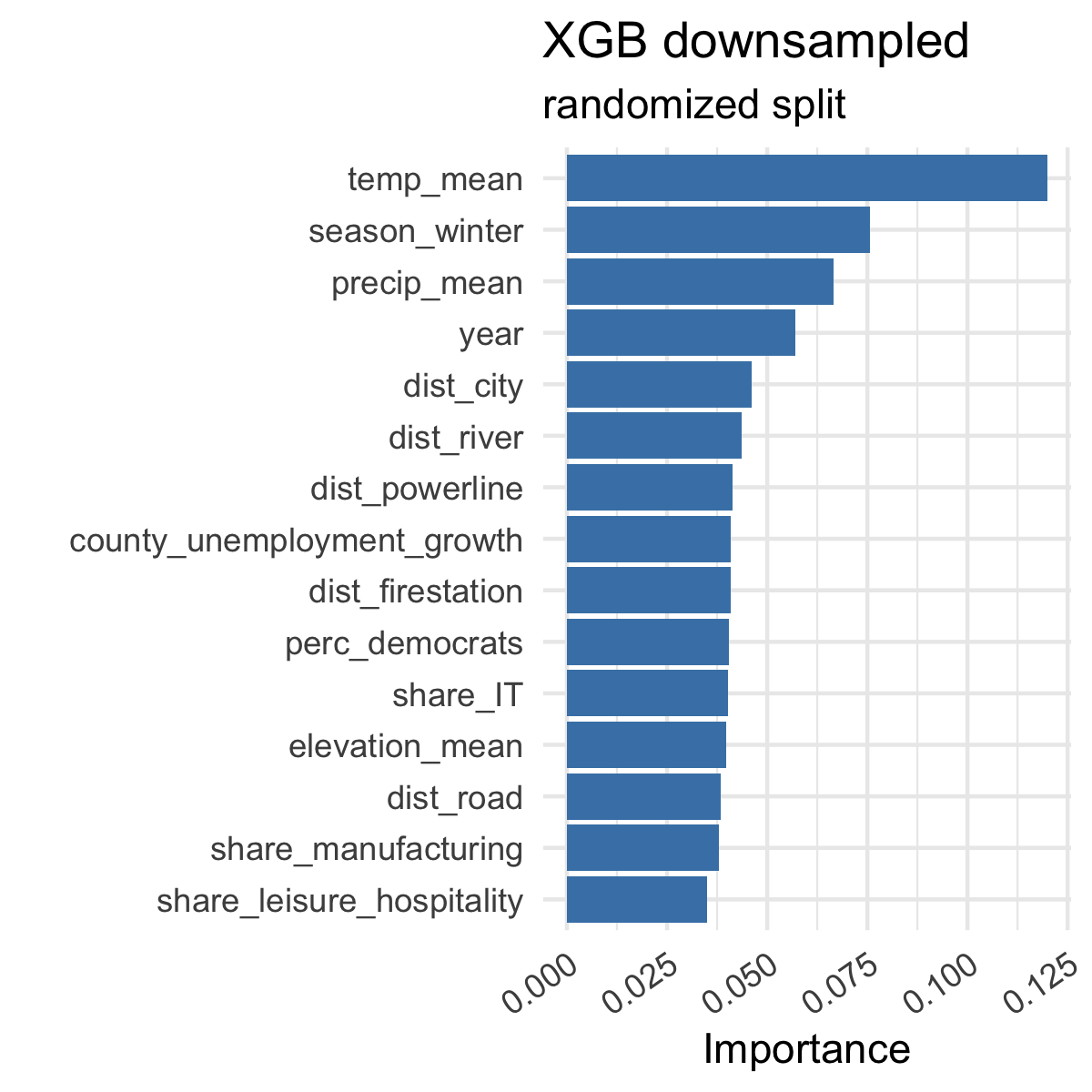
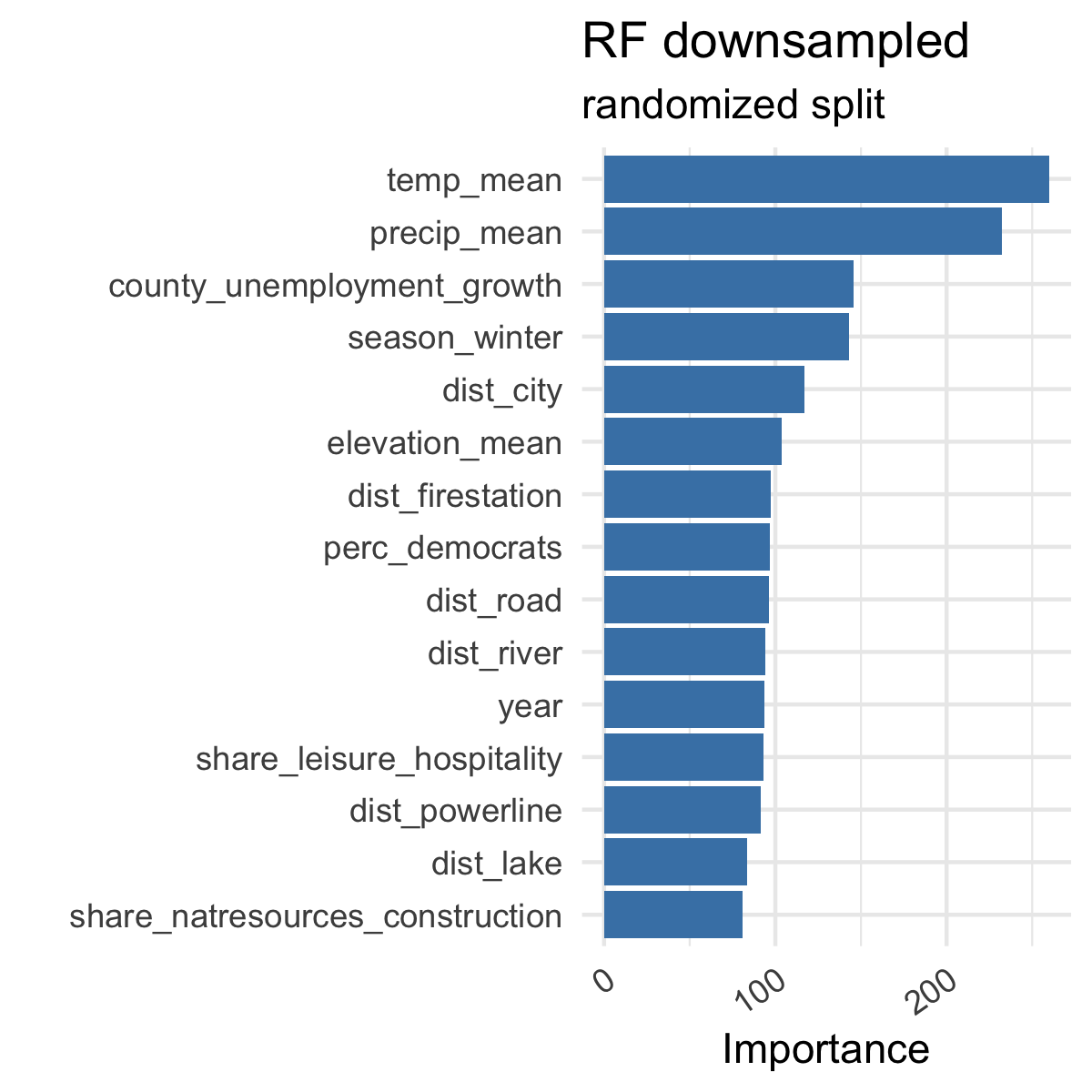


Figure 4: Comparison of variable importance for best models with temporal data split

A similar pattern arises for the models estimated with training data assigned at random. Random Forest and *xgboost* again share 13 variables among their top 15 most important predictors. Variable importance remains largely the same with one notable difference of the year indicator only taking a prominent role as a predictor if the training data is assigned at random.

Figure 5: Comparison of variable importance for best models with randomized data split



From these variable importance scores it can be inferred that wildfire ignition risk is influenced by a variety of meteorological, demographic and anthropogenic factors with meteorological factors being the most important. The average temperature has proven to be the most important predictor across all models, highlighting the impact of rising temperatures caused by progressing climate change on wildfire risk. Temperature being an important predictor is an expected outcome, due to temperature influencing on soil und fuel moisture, which are often cited predictors of wildfire risk on their own. Temperature also acts through other indirect channels such as increased popularity of outdoor activities and fire hazards such as campfires and fireworks during warm weather.

Average precipitation ranks second, being the second most important predictor for both final Random Forest models and among the top five predictors for both *xgboost* models. Like temperature, precipitation manifests its impact both onto environmental factors (soil and fuel moisture) as well as human-activity (decreased popularity of outdoor activities and related fire hazards).

Anthropogenic predictors are consistently ranked among the most important predictors. The distance from the nearest city ranks among the top five predictors across all models. The association of road infrastructure and fire risk as found in similar studies is confirmed by the importance of distance from the nearest road as a predictor (Syphard et al., 2008). Surprisingly, the distance to the nearest powerline does not rank as highly as anticipated, as powerlines have the potential to directly spark ignitions if damaged or downed. The distance to the nearest fire station does not affect wildfire ignitions per se but manifests its impact by shortening time until extinguishment, thus combating the spread of fire.

These results confirm the finding from previous studies that unemployment acts as an important socioeconomic predictor of wildfire risk, as unemployment growth at county level ranks as the third most important predictor for both Random Forest models and among the top 10 predictors for *xgboost*. Topological features such as the average elevation and the distance to the nearest river are prominent predictors for all estimated models, presumably due to their role in shaping the conditions that either favor or discourage wildfire ignitions and spread.

The local share of registered democrats also proves to be an important predictor across all models. I interpret this as party registrations acting as a proxy for local environmental policy such as forest management, fire control and investment in local wildfire prevention measures. There is no clear pattern indicating whether the composition of the labor force is an important predictor. Also, variables indicating local responsibility areas, fuel removal campaigns and protected areas are wholly absent from the top 15 most important predictors, as are the predictors related to landcover.

## Interpretation

These findings provide insight into the multidimensionality of factors that contribute to wildfire risk in Northern California. They highlight the importance of considering predictor variables that capture both the environmental as well as the anthropogenic factors of wildfire risk. Even if human activities are cited as the most common vector of wildfire ignition, the meteorological and topological conditions dictate the spread of fire and its perimeter.

These are important findings for local environmental policy, as fire prevention measures ought to be focused on areas where the most important environmental and anthropogenic predictors intersect. This leads to the inference that preventive measures would be most effective in peripheral communities with high average temperatures, low average rainfall and rapidly increasing unemployment. The robustness of the most important predictors across all models indicates the validity of these findings beyond the individual classifiers and modeling procedures.

Model performance also suggests that neither classifier achieves high performance when forecasting wildfire risk for a period unseen in the training data. While logistic regression performed marginally better than Random Forest and *xgboost* at this task, the tree-based ensemble models generally outperformed logistic regression once they had access to samples from the predicted period during training. One possible strategy to address this problem is regular retraining of models during the winter season once enough data to improve predictions for the summer season becomes available.

## Limitations

# Conclusion

Northern California has seen a stark increase in areas burnt by wildfires since 2017. Wildfire occurrences are not evenly distributed as they most often occur XXX with a combination of environmental and anthropogenic factors contributing to the probability of wildfire ignitions and spread. I aimed to depict these intersecting factors by compiling a data set that includes both meteorological, topological, demographic, socioeconomic, political and infrastructure data. I then used these data as predictors for local wildfire occurrence at a 4 resolution in Northern California within an 8-year study period from 2010 to 2018. I modeled the likelihood of wildfire occurrence by using several statistical and Machine Learning algorithms: logistic regression, Random Forest and *xgboost*. The estimated models predict a 4 unit’s probability of seeing wildfire occurrence during the winter and summer seasons of 2017 and 2018. Comparison of the modeling results led me to the conclusion that assigning observations into training and test set based on their year of occurrence leads to lower predictive performance than expected. While preliminary performance metrics obtained through resampling the training data (2010-2016) showed promising results, final predictions of the test set (2017-2018) showed a general inability to generalize patterns of wildfire occurrence from the training years for predicting the testing years.

I estimated another set of models accordingly where the training and test data were assigned at random. This increased predictive performance for all classifiers, but especially Random Forest and *xgboost*. Logistic regression, which previously scored the highest in terms of ROC-AUC with training and test data assigned by year, profited the least from this change. The regression model ranked third behind Random Forest and *xgboost* after splitting the data at random. This led me to the conclusion that tee-based ensemble models are more likely to overfit on the training data if there is a substantial annual variation in the modeled patterns. It also showed the classification power of non-parametric models that require only minimal data pre-processing while eclipsing the predictive performance of linear statistical models if observations from all years of the data set are made available for training.

Regardless of the used data splitting method, variable importance rankings of all estimated tree-based ensemble models show a consistent image of the factors that contribute the most to wildfire occurrences in Northern California. Environmental and anthropogenic factors in combination were consistently ranked as the most important predictors, suggesting that preventive measures would be most effective in peripheral communities with high average temperatures, low average rainfall and rapidly increasing unemployment. Such qualitative insights provide policy makers with valuable guidance to design forest management and wildfire prevention initiatives more efficiently. It also highlights how predictors of wildfire occurrences cannot be reduced to purely environmental or human-related factors.

Potential shortcomings of this study include the large resolution of this analysis, due to the unavailability of most predictor variables at a more granular resolution at regular intervals. Available data sets for the environmental conditions in Northern California are highly fragmented with important predictors of wildfire ignitions either only being available at a large resolution global scale or only at the very granular, local level. One key objective for future research is thus the creation of granular state-level data sets that combine the various types of important predictors into harmonized data bases fit for modeling local differences in wildfire risk. Furthermore, future research could focus on improving the predictive performance of classifiers trained on temporally-split data, as this setting most closely mirrors the real-life application of environmental risk prediction of future seasons based on data of the past.

# References

1. Included counties: Butte, Colusa, Del Norte, Glenn, Humboldt, Lake, Lassen, Mendocino, Modoc, Nevada, Plumas, Shasta, Sierra, Siskiyou, Sutter, Tehama, Trinity, Yuba [↑](#footnote-ref-1)
2. See (Banerjee et al., 2020) for a detailed discussion of road construction, economic growth and causality [↑](#footnote-ref-2)