# Introduction

Bla

## 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 ignition, as it is suited to 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 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).

Oliveira and co-authors have applied both linear regression and Random Forest (based on regression trees) at a very large regional scale to model fire density in Mediterranean Europe (Oliveira et al., 2012). In their study they aim to predict the likelihood of fire occurrence for spatial units at a 10 resolution. They use a mixture of environmental and human-centric variables, concluding that Random Forest outperforms their regression model. They list precipitation, soil moisture, unemployment rate and density of roads as their most important predictors (Oliveira et al., 2012).

Cao and co-authors compare multiple methods of predictive modeling to predict wildfire ignition susceptibility in the South Chinese province of Yunnan (Cao et al., 2017). Unlike Oliveira and co-authors, they disregard anthropological factors including only topological, meteorological and vegetational predictor variables at a much smaller scale of a single Chinese province. They apply generalized linear models (both logit and probit), as well as Random Forest and Artificial Neural Networks to predict a binary wildfire ignition target variable. For model evaluation the employ a cost-sensitive misclassification metric, concluding that Random Forest performed the best in predicting the binary ignition target. Cao and co-authors stress the importance of data preparation and sampling methods to resolve issues of class imbalance, due to the overwhelmingly large number of non-ignition samples (Cao et al., 2017).

The Iberian peninsula has been the area of investigation for a multitude of studies, presumably due to being among the areas with the highest concentration of wildfires in Europe (Oliveira et al., 2012). Martínez and co-authors went for the opposite approach of Cao and co-authors by explicitly investigating only human-caused wildfire ignitions in Spain (Martínez et al., 2009). Their models are based exclusively on anthropological predictors, ignoring any topological, climatic or vegetational factors, in order to investigate human-centered factors of wildfire occurrences. They apply logistic regression to estimate their binary target variable (low risk / high risk), finding a high correlation (Spearman correlation of 0.71) between their risk estimates and the official ignition danger index that is used by the Spanish forest service. They find that agricultural landscape patterns, and rural development processes are the most important predictors for human-caused wildfire ignitions in Spain (Martínez et al., 2009).

A more recent study conducted by Rodrigues and de la Riva in 2014 expands on the work done by Martínez and co-authors by not only modeling human-caused wildfire ignitions in Span with logistic regression, but also Random Forest, boosted trees and Support Vector Machine (Rodrigues & de la Riva, 2014). In a similar setup to Martínez and co-authors they only include human-centered predictor variables in their training data. They find that logistic regression is outperformed by all other estimated models in terms of ROC-AUC (see section 3.4 for a discussion of various evaluation metrics in classification tasks). Due to their data preparation strategy resulting in a training set with balanced target classes (low occurrence / high occurrence) ROC-AUC is a legitimate metric for model evaluation. They conclude their analysis by recommending tree-based algorithms such as Random Forest and boosted trees for modeling wildfire ignition occurrences, 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 using logistic regression (Catry et al., 2009). They combine both environmental and anthropological variables, but explicitly make it a point to use 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). Due to the predictive performance of these variables and their frequent used in studies modeling wildfire ignitions all of them are used as predictors in the study at hand.

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

A recent study conducted by Malik and co-authors stays at the local scale but focuses on the area near Monticello and Winters in Northern California (Malik et al., 2021). Their approach is distinctly different to other discussed studies, as their models are trained on data that integrate meteorological data, remote terrain sensor data, data on landcover, as well as wildfire history and human variables, aiming to enable real-time wildfire risk predictions akin to weather forecasts. They compare two modeling strategies where in a first approach two separate Random Forest models are estimated for one data set containing data on weather, the presence of powerlines and terrain and a separate set of vegetation data. These two models are then combined into an ensemble with stacking where a boosted trees classifier (AdaBoost) is used for the final predictions. In a second modeling strategy the authors trained a single Random Forest model on the entirety of the combined training data, resulting in a superior model as evaluated with the ROC-AUC metric (Malik et al., 2021).

# Data

One of the primary challenges of this study was the compilation of a harmonized data set that was sourced from both GIS data (“geographic information system”) for all the spatial variables, as well as tabular data for the socio-economic, demographic, and political predictors. All GIS data was processed and joined in QGIS, whereas the tabular data was joined to the harmonized GIS data set in R, a 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 Canadian border. This diverse region is made up of 18 counties[[1]](#footnote-1) with a combined area size of roughly 113’380 . This region typically has a milder climate than the more southern counties with higher precipitation and more humid climate. Conifer forests, oak woodland and shrubland dominate the flora, although there is a considerable variation in the dominant vegetation pattern across the area. National forests and rugged mountain ranges dominate the landscape in this northern part of the state, as the foothills of the Sierra Nevada stretch up until 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 California have all erupted in the study area over the past three years. The Dixie fire, ignited in the Sierra Nevada in summer of 2021, has evolved into largest single-source fire in the history of the state (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 (CAL FIRE, 2021c). The deadliest wildfire in the history of California ignited in Butte County in 2018, with 85 recorded deaths (CAL FIRE, 2021b).

Ein Bild, das Karte enthält.

Automatisch generierte BeschreibungUsing the QGIS software I have divided the study area into a grid of squares with an area size of 4, making up the units of observation of this study. I have chosen 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 any further.

Figure 1: Study Area, with all recorded wildfires during the study period of 2010-2018

## 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 the occurrence of wildfire ignitions were obtained from the "Fire Perimeters" data set, compiled, and provided by the Fire and Resource Assessment Program (FRAP), a joint effort of the California Department of Forestry and Fire Protection (CAL FIRE), the United States Forest Service Region 5, the Bureau of Land Management, and the National Park Service of the United States (CAL FIRE, 2021a). “Fire Perimeters” is the most complete and frequently updated database on wildfire occurrences in California. This data set is provided as a shapefile and displays the perimeters of all recorded wildfire occurrences in the area, along with harmonized data such as the exact date of a wildfire’s discovery, as well as its extinguishment. I used a subset of this data set corresponding to the study area and period, including all recorded fires throughout each year.

The location accuracy of the recorded wildfire ignitions made this data well suited for spatial analysis. QGIS can access “Fire Perimeters” directly through the ArcGIS REST API, after which it must be projected to a suitable map projection for 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 the QGIS spatial analysis join algorithm was used to register all intersections of a wildfire perimeter and the grid made up of 4 squares, which serve as the units of observation of this study. The resulting table records all dates for which the 4 square elements of the grid have intersected with a fire perimeter. Note that this does not mean that a given 4 square element was completely covered by a wildfire perimeter (and hence was burned completely), merely that at least a single wildfire ignition has taken place and was recorded within the bounds of that specific 4 square.

In order to further process this data, this 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, the data was transformed so that each sample represented 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.

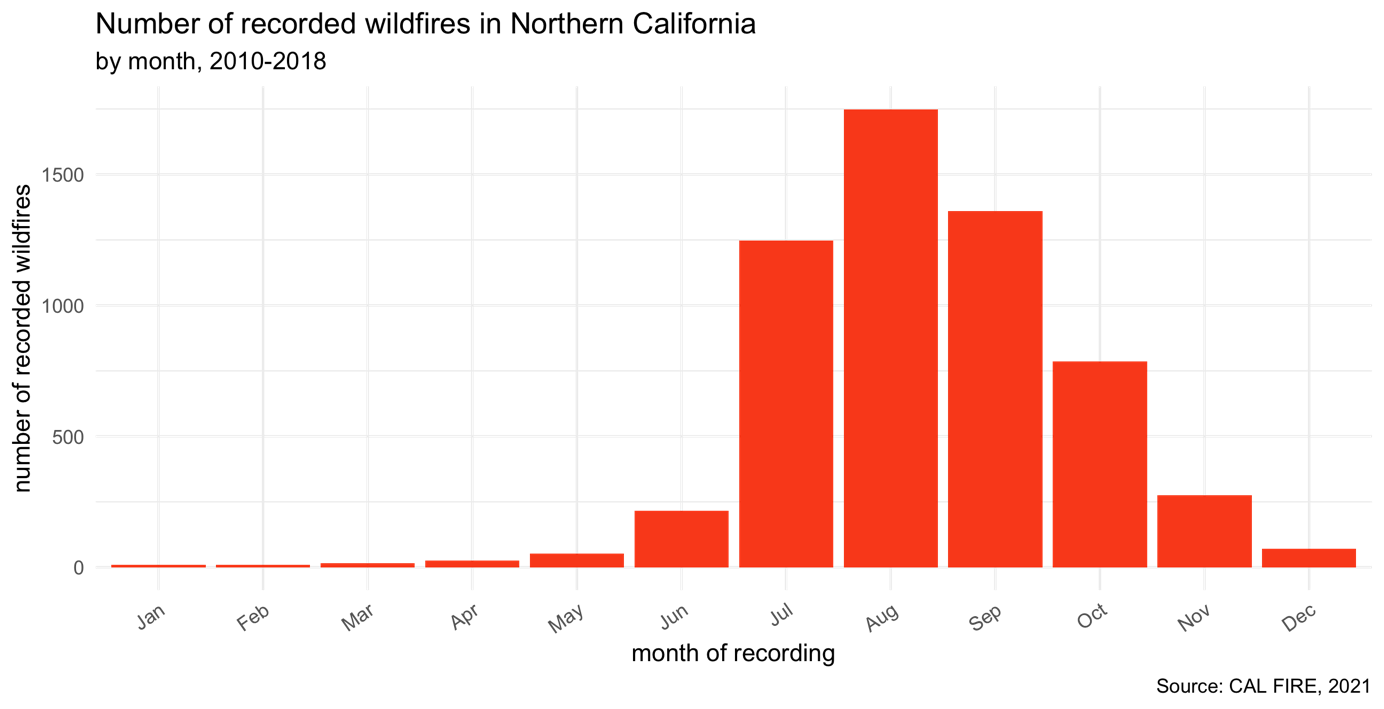


Figure : Monthly distribution of wildfire occurrences during the study period

The monthly distribution of wildfire ignition events shows a clear seasonality, as displayed in figure 2. The vast majority of recorded wildfires were registered as active during the summer and autumn months. This meant that the monthly data could be aggregated to a seasonal level, reducing the overall number of samples in the data set while preserving as much information on wildfire occurrence in Northern California as possible. Another motivation for this aggregation was the fact that many predictors were not available at the monthly level, making a data set at this level too granular for the variation contained in the predictor variables. In line with the study conducted by Tonini and co-authors, the period from May to October was assigned to the summer season, leaving 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 has 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. The implications of this circumstance on the modeling process and different strategies for addressing potential problems are discussed in depth in chapter 3 on the methods used for data pre-processing and model evaluation.

## Predictor Variables

Overall, I compiled a set of 54 predictor variables in total. Not all of these variables were ultimately used for modeling. Chapter 3 on data preprocessing and predictor selection discusses 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 variation 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 categories of predictor variables are environmental data (including both topographic, meteorologic and data concerning land cover), infrastructure data (both the proximity to human-made infrastructure, in addition to binary data indicating the presence of infrastructure in the units of observation), as well as demographic and socio-economic data for the study area.

### Environmental Predictors

Topographical features such as elevation are important predictors of spatial patterns of fire, as they account for local variations in climate, in addition to exerting influence on 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 has been compiled by the National Aeronautics and Space Administration (NASA) and the National Geospatial-Intelligence Agency (NGA) and 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 has to be aggregated to the 4 level of the grid elements 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, topographical data on the presence of major bodies of water, such as lakes and rivers, are added 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) These data are used twofold: In a first step, a binary variable is created indicating whether an object of observation is intersected by either a lake or a river. In a second step, the distance of each 4 square’s centroid to the nearest element of both the river and the lakes data set is calculated with the *v.distance* algorithm of the GRASS package for QGIS. This provides an additional indicator to the presence of water bodies, that is numeric and continuous, as opposed to the logical dummy variables created in the first step.

The 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 California. 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” system 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, this data can act as a proxy for the primary fuel type within the 4 units of observation. QGIS is used to determine the most frequent land cover type for each square.

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 fire ignitions (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 (corresponding to roughly 21) in raster format (Harris et al., 2014). Due to the large number of raster layers (monthly interval, eight-year study period, three data sets), these predictors are constructed iteratively, using QGIS’ python interface extract the mean values of minimum temperature, maximum temperature, and mean precipitation for each unit of observation from all raster layers. Ultimately these predictor variables are aggregated to the seasonal level, along with the target variable.

### Infrastructure Predictors

Access to roads has often been described as a driver of economic activity and a proxy for infrastructure development (SOURCE). 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 the response of 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 is processed, this data set is used to both create dummy variables indicating the presence of a major road for each 4 units and calculate the distance from each unit’s centroid to the nearest major road as well. The same process is repeated for a data set of powerlines, resulting in predictor variables indicating both their presence (binary) as well as the distance from each unit’s centroid to the nearest powerline. 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 only their presence is determined, as I do not expect these recreational structures to have any continuous effect if they’re not present – unlike powerlines, where larger distances function well as a proxy for a lack of economic development.

### Demographic 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). I obtained data on population density from SOURCE, which provide annual raster data at the LEVEL level (SOURCE). The average population density per 4 unit is calculated in QGIS. Additionally, the annual population growth at the county level is added to the data set. This data is supplied by SOURCE and acts as another proxy for economic development. 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.

In order to measure the political leanings of the inhabitants of the study area I include voting district level data on the share of registered Democrats and Republicans. Similarly, 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 (*Ballot Initiatives*, 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, 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 : Californian Ballot Initiatives relating to environmental issues during the study period

For each of these ballot initiatives the number of Yes-votes at county level are added to the data set.

### Socio-economic Predictors

Previous studies have found the rate of unemployment to be important predictors for wildfire occurrences (Martínez et al., 2009; Oliveira et al., 2012). Oliveira and co-authors mention two specific channels how 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). Besides the rate of unemployment, I also add the absolute number of unemployed adults and the monthly growth in unemployment rate as well. The data on monthly unemployment rates at county level are provided by the California department of employment and development (EDD), which are aggregated to the seasonal level along with the target variable (EDD, 2021b).

The EDD also provides monthly labor data denoting the share of employed workers by industry at county level (EDD, 2021a). These data are also added to the data set and aggregated to the seasonal level. Implicit zeros, which are appear in the data base as missing values, have to be introduced explicitly during the data cleaning process.

# Methods

The data is split into subsets of training and testing data. All data from the years 2010-2016 are used for training the models. The data from the years 2017 and 2018 are used for evaluating the model fits based on the testing data. This approach is chosen over the usual random split done in most Machine Learning use cases since it mirrors the process of forecasting wildfire risk on past data for future seasons.

The training set is 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). For each of these three algorithms models are estimated in three distinct steps:

1. A naïve model estimation strategy without resampling or hyperparameter tuning, using the heavily imbalanced training data set (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 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.

## 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 previously been widely used in cases that model wildfire occurrences, 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 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 : Mathematical notation of a logistic regression model

In the case of this study multicollinearity must be addressed during the preprocessing of the data, 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 these variables. The pre-processing steps are handled for each model separately, taking the individual features of each algorithm into account. Most of the pre-processing steps are similar, however.

For all models, the dummy variables indicating the presence of infrastructure (such as roads) or topological features (such as lakes) are 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 undergo 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 are removed, as are strongly correlated predictors (with a threshold of 0.75), similar to Oliveira and co-authors’ study (Oliveira et al., 2012). Categorical predictors are turned into dummy variables through one-hot encoding.

In order to balance out the classes in the target variable, models are estimated with both an upsampled, as well as a downsampled set of training data. Upsampling is done using the SMOTE algorithm, creating additional synthetic observations. For the downsampled training set the NearMiss 1 algorithm is 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 undergo 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 is intended to improve the classification boundary of the training data and the predictive power of the trained model.

During hyperparameter tuning I introduce both and regularization to logistic regression, resulting in an elastic net model. For this specific case all predictors undergo normalization. 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*) that I tune using a grid search strategy.

## Random Forest

*Random Forest* is a non-parametric ensemble learning algorithm, comprised of a set of *Decision Tree* models. These submodels aim to create splits in the training data that minimize the heterogeneity in the resulting subsets, learning the best rules to create these splits in the process (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 final ensemble predictions based on their own, highly variant predictions. The singular trees are made weak due to limited access to the training data, as each 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.

*Random Forest* does not share the assumptions of logistic regression, but fewer correlated predictors might improve the *variable importance* score estimates of tree-based algorithms (Kuhn & Johnson, 2019). Strongly correlated predictors indicating the presence of infrastructure or topological features are 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 has to be used for all *Random Forest* models except the naïve estimation. Subsampling is 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 that can be tuned, 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 for the ensemble (*trees*). During hyperparameter tuning the latter is kept constant at 500 trees due to computational restraints, while the two former parameters are tuned using grid-search.

## xgboost

*xgboost* is an implementation of the boosted trees method. Similar to *Random Forest* it is an ensemble algorithm 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, especially with the *xgboost* implementation that has taken 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 subsampling the number of samples 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 selection of variables is 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 implementation of *xgboost*.

*xgboost* shares the hyperparameters of *Random Forest* (*trees*, *min\_n* & *mtry*) and additionally 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 are 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 special precautions must be taken in order 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) curve is used to evaluate the ideal probability thresholds separating events from non-events (Kuhn & Johnson, 2013). ROC-curves for all model fits are displayed in the appendix. Similar to previous studies that use predictive modeling to predict wildfire ignition risk I use the *Area under the curve* (AUC) of ROC-curves as the primary metric for evaluation.

*Accuracy* is 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 blindly 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 : Computation of the F-measure

Additionally, not only the imbalance between the classes in the target variable are 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 took 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 implement 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 is used for evaluating both the estimated models and the hyperparameter combinations during tuning that 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

After the described data preparation steps, TABLE X shows the final selection of predictor variables used for modeling. All models are estimated based on this data set. 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 : 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 achieves a much higher recall value, 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 : 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 evaluation based on 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 disability 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 : 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 : 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 are trained on the downsampled training data set, which has proven to lead to comparable if not superior prediction results for the previously estimated models (see Table 3).

Not only is predictive performance increased for all models, the changes to the training and test split also affect these models differently. The Random Forest classifier now achieves 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 maximixe the ROC-AUC. Elastic net regression, which previously outperformed the other classifiers, profits the least from changes and now ranks last with a ROC-AUC of 0.809. Hyperparameter tuning has found that the ideal *penalty* value remains 0.0000000001, with the optimal *mixture* now being 0.5, which results in an equal mixture between LASSO and ridge penalties.

I ultimately rate the Random Forest as the strongest model. Not only does it achieve the largest ROC-AUC, but it outperforms elastic net regression and xgboost in terms of every other metric except for recall, which is still comparable. Random Forest achieves the highest accuracy, highest precision, and highest F-measure, despite the slightly lower recall. This combination of metrics indicates that the Random Forest model achieves the best classification boundary among these models while minimizing the penalized classification cost metric. 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. 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.

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

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

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

bla

## Variable Importance

bla

# Conclusion

bla

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)