Predicting Burned Area of Forest Fires Using Meteorological Data

Taylor Callahan

Department of Data Science, Bellevue University

DSC 680: Applied Data Science

Professor Catherine Williams

June 13, 2023

Business Problem

The increasing occurrence, spread, and impact of forest fires presents a pressing problem in the world today. While early detection systems exist, accurately predicting the extent of forest fire damage remains a challenge. In Portugal, a country heavily affected by forest fires, over 2.7 million hectares of forest area were destroyed from 1980 to 2005. This places Portugal among the top three highest danger countries in the EU (European-Commission). To address this problem, this analysis aims to examine meteorological factors in conjunction with burned area data from forest fires in Portugal. The objective is to create predictions for the impact of forest fires and identify high-risk areas. This analysis holds the potential to enhance early detection, support fire management decisions, and effectively warn the public of fire risk.

Background/History

Forest fires have long been recognized as a natural phenomenon with potential environmental benefits. However, in recent years, there has been a notable increase in their occurrence, spread, and impact, affecting both human and natural ecosystems. Studies have shown that while early detection systems, such as sensors and monitoring, exist, accurately predicting the extent of forest fire damage remains a challenge (Clarke et al., 2022). Additionally, the role of meteorological factors, including temperature, relative humidity, wind, and rain, in contributing to the occurrence and extent of forest fires has often been overlooked.

Prior research suggests that large and severe forest fires are associated with warm and dry conditions, which occur with a warming climate. Based on projections, current trends, and simulation modeling, warmer and drier conditions will lead to longer fire seasons, increasing the frequency and extent of fires (Halofsky et all., 2020).

Data Explanation

The forest fire dataset used in this analysis is derived from "A Data Mining Approach to Predict Forest Fires using Meteorological Data" by P. Cortez and A. Morais (Cortez & Morais, 2007). The dataset captures information from the Montesinho natural park in Portugal, spanning from January 2000 to December 2003. It includes various attributes such as spatial coordinates (X and Y), month and day of the fire, meteorological data (temperature, relative humidity, wind speed, rainfall), and indices from the Forest Fire Weather Index (FWI) system. The dataset comprises 517 observations.

There are certain limitations to consider. Firstly, the dataset is limited to a specific time period and geographical location, potentially limiting its generalizability. Additionally, the dataset does not provide information on other potential factors influencing forest fires, such as human activities or topographical features.

The data is defined in Table 1 below, where “Area” is the outcome variable to be considered.

Table 1: Data Dictionary

Variable	Description
X	X-axis spatial coordinate within the Montesinho park map
Y	Y-axis spatial coordinate within the Montesinho park map
Month	Month of the year (jan to dec)
Day	Day of the week (mon to sun)
FFMC	Fine Fuel Moisture Code from the FWI system
DMC	Duff Moisture Code from the FWI system
DC	Drought Code from the FWI system
ISI	Initial Spread Index from the FWI system
Temp	Temperature in Celsius degrees
RH	Relative humidity in percentage
Wind	Wind speed in km/h
Rain	Rainfall in mm/m^2
Area	Burned area of the forest (in hectares)

The data preparation involved several steps to ensure compatibility with the modeling process. First, the variables X and Y were converted to numeric format to represent the spatial coordinates accurately. Next, the variables month and day were converted to factors to categorize them appropriately for analysis. Oversampling was performed by duplicating instances where the area was not equal to 0, effectively increasing the representation of non-zero area cases in the data. This oversampling step was necessary to address the issue of imbalanced data, as the majority of observations had an area value of 0, while the non-zero area cases were relatively scarce. By duplicating these instances, the oversampling technique helps to provide a more balanced representation and greater variability, enabling the models to learn and generalize better. This was necessary to avoid bias towards the majority class and improve the model's ability to capture patterns.

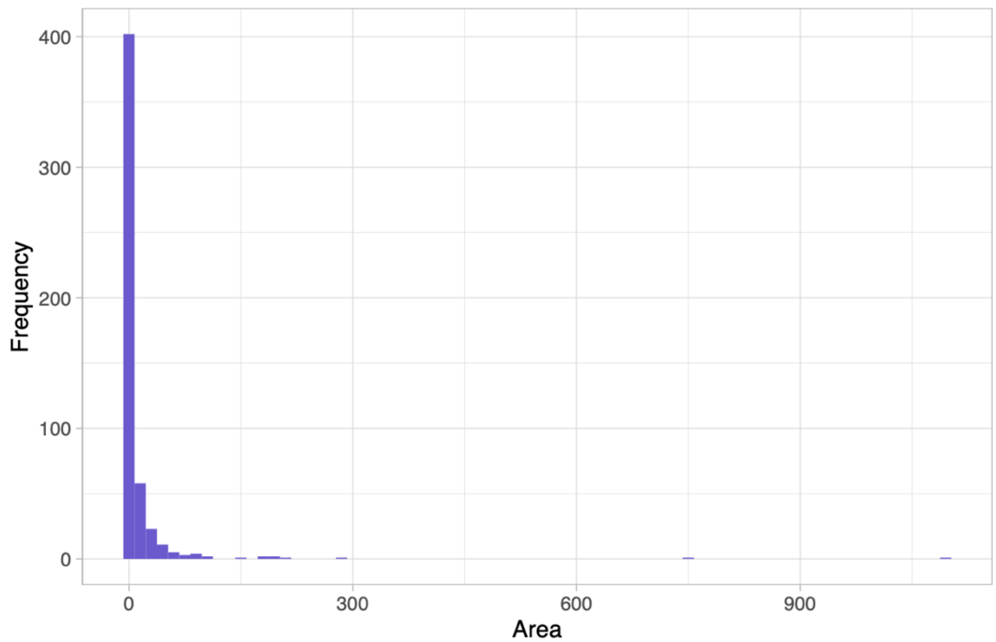
To address the challenge of missing values during resampling, a constant value of 1 was added to all numeric variables. This was done to prevent the occurrence of missing values that could arise due to the large number of zeros in the outcome variable, which was producing NAs in the Random Forest and Neural Network models.

Methods and Analysis

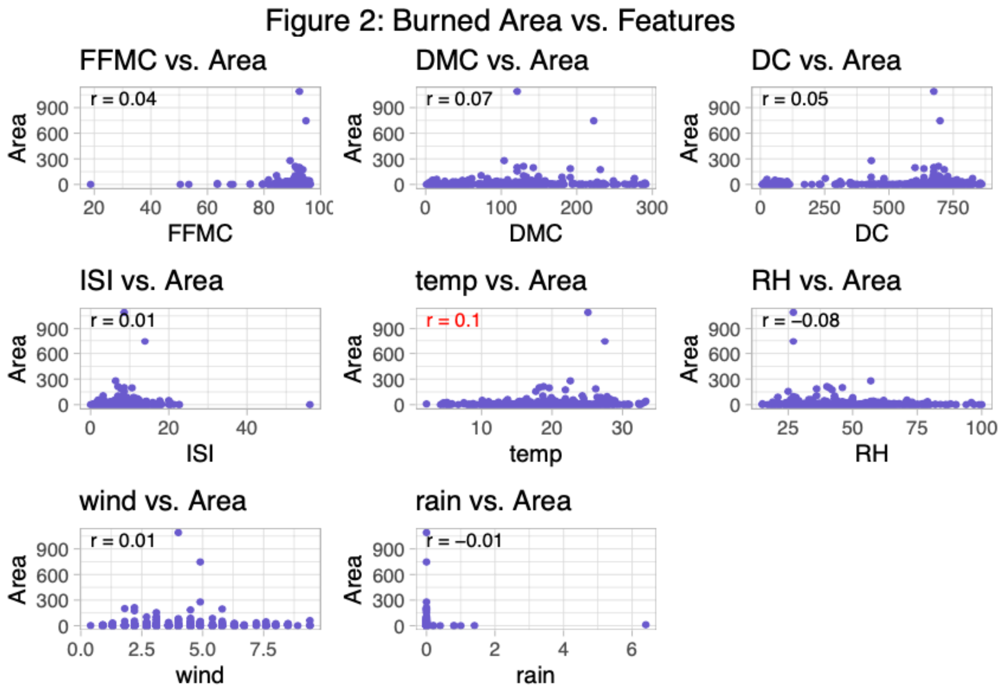
Exploratory data analysis (EDA) was performed prior to modeling. First, it was determined that there were no missing values in the dataset, which ensured the analysis could be conducted on the complete set of observations.

To understand the distribution of the target variable "area," a histogram was created, revealing a significant imbalance towards values close to zero. This indicated that the majority of instances in the dataset had no or very minimal burned area. Additionally, the distribution of the "area" variable appeared to be right-skewed. This can be seen in Figure 1 below.

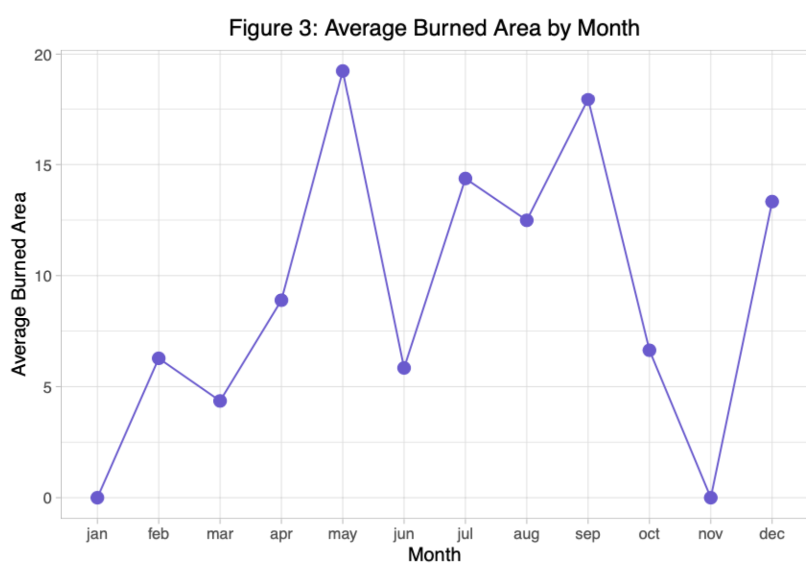
Figure 1: Distribution of Burned Area



Scatterplots were constructed to examine the linear relationship between each numeric feature and the "area" variable. None exhibited a strong linear correlation with "area." The "temp" feature showed the strongest linear correlation, albeit with a low correlation coefficient (r = 0.1). This indicated that there might be some relationship between temperature and burned area, although it was not strictly linear. It is important to note that low linear correlation does not necessarily imply the absence of correlation; there could be other non-linear or complex relationships at play. These scatterplots can be seen in Figure 2.



An analysis of the average burned area across different months was performed, where a graph was created to visualize the average burned area by month. This revealed distinct peaks in May, September, and December, while the lowest values were observed in January and November. This temporal pattern suggested that the occurrence and severity of forest fires might be influenced by seasonal factors. This is seen in Figure 3.



The high imbalance and right-skewed distribution of the "area" variable indicate that accurately predicting instances with significant burned area might be challenging due to the scarcity of such cases in the dataset. The weak linear correlation observed between the numeric features and "area" suggests that a linear modeling approach may not capture the full complexity of the relationship. Additionally, the temporal patterns in burned area suggest the importance of considering seasonal factors in the modeling process.

Random Forest, Neural Network, and Support Vector Machine (SVM) were selected as potential candidates for modeling the data. These models were chosen due to their ability to handle complex relationships, capture non-linear patterns, and effectively model high-dimensional data. Random Forest is known for its robustness against overfitting and ability to handle interactions between variables. Neural Network excels in learning intricate patterns and relationships through interconnected layers of artificial neurons. SVM, with its ability to find optimal hyperplanes in high-dimensional spaces, can capture complex decision boundaries (Osisanwo et al., 2017).

To train the models, the dataset was split into training and testing sets. The hyperparameters were tuned to optimize performance. For Random Forest, parameters such as the number of trees, maximum depth, and split rule were tuned using a grid search approach. Fr Neural Network, parameters such as the number of hidden layers, size of hidden layers, and decay rate were tuned. SVM hyperparameters, including the choice of kernel, cost parameter, and gamma, were optimized to find the best configuration.

Model evaluation was performed using key metrics to assess the performance of each model. RMSE (Root Mean Squared Error) and R-Squared were chosen as evaluation metrics. RMSE provides a measure of the average prediction error, taking into account both the magnitude and direction of the errors. R-Squared, also known as the coefficient of determination, represents the proportion of the variance in the target variable that is explained by the model. These metrics help gauge the accuracy and goodness-of-fit of the models.

After evaluating the models, it was observed that SVM exhibited the highest R-Squared value, indicating that it explains a larger portion of the variance in the forest fire area. Additionally, SVM had the lowest RMSE, suggesting that its predictions were closer to the true values. Table 3 below shows the evaluation results of the three models. Based on these evaluation metrics, it can be concluded that SVM performed the best among the considered models in predicting the forest fire damage area.

Table 2: Model Evaluation Metrics

Model	RMSE	R-squared
Random Forest	33.464	0.866
SVM	21.518	0.934
Neural Network	82.047	0.003

Conclusion

The analysis aimed to address the problem of accurately predicting and assessing the impact of forest fires in Portugal. The EDA revealed the imbalance and right-skewed distribution of the "area" variable, indicating the challenge of predicting instances with significant burned area due to their scarcity in the dataset. The scatterplots showed a weak linear correlation between numeric features and "area," with temperature ("temp") exhibiting the strongest but still low correlation. Additionally, the temporal pattern analysis demonstrated distinct peaks in May, September, and December for burned area, suggesting the influence of seasonal factors.

To tackle the complex relationships and non-linear patterns in the data, Random Forest, Neural Network, and SVM were selected as modeling candidates. After evaluation, SVM emerged as the best-performing model, exhibiting the highest R-Squared value and the lowest RMSE. Therefore, SVM was identified as the most suitable model for predicting the extent of forest fire damage based on the considered evaluation metrics.

The results of this analysis have the potential to enhance early detection systems, support fire management decision-making processes, and effectively communicate fire risk to the public. By accurately predicting the impact of forest fires and identifying high-risk areas, stakeholders can take proactive measures to mitigate the damage caused by these fires and protect both human and natural ecosystems. The good fit of the SVM model suggest that these meteorological factors play a large role in how much impact a forest fire has, thus confirming the need for further research and examination of such factors in predicting forest fires.

Assumptions

It was assumed that the dataset used is representative of the broader population of forest fires in Portugal. While the dataset covered a specific time period and location, it may not capture all the potential factors influencing forest fires, such as human activities or topographical features. Additionally, it was assumed that the relationships between the predictor variables and the target variable remain stable over time and are not subject to significant changes or external influences. Furthermore, it was assumed that the selected models (Random Forest, Neural Network, and SVM) are appropriate for capturing the complex relationships and patterns in the data. These assumptions provide a foundation for interpreting the results and making informed decisions based on the analysis.

Limitations and Challenges

The dataset is limited to a specific time period (January 2000 to December 2003) and geographical location (Montesinho natural park in Portugal). This narrow scope may limit the generalizability of the findings to other regions or time periods. Additionally, the dataset does not include information on potential factors that may influence forest fires, such as human activities (e.g., arson) or topographical features (e.g., slope, elevation). The absence of these variables may overlook important contributing factors to the occurrence and spread of forest fires.

Another limitation is the imbalanced nature of the target variable, with a majority of instances having no or minimal burned area. This can pose challenges in modeling and accurately predicting instances with significant burned areas. Balancing techniques, such as oversampling, were employed to address this challenge, but it is important to be cautious of potential biases introduced through such techniques. While the selected models (Random Forest, Neural Network, and SVM) were chosen based on their ability to handle complex relationships, there is no guarantee that these models will capture all the underlying patterns in the data. Non-linear or complex relationships may not be fully captured by the chosen modeling approaches. The optimal hyperparameter configuration may vary depending on the dataset and modeling technique, and finding the right combination can be a time-consuming and computationally intensive task.

Future Uses/Additional Applications

The analysis of forest fire data and the development of predictive models have broader implications beyond the scope of this study. The models trained on the forest fire dataset can serve as a foundation for future research and practical applications. One potential future use is the integration of real-time meteorological data and satellite imagery to create a dynamic forest fire prediction system. By incorporating up-to-date weather information and monitoring fire hotspots, the models can provide timely warnings and help in proactive fire management. The models can also be extended to other geographical regions and ecosystems prone to forest fires, allowing for the identification of high-risk areas and the allocation of resources for prevention and firefighting efforts.

The analysis can be expanded to include other variables and factors that may influence forest fire behavior, such as topographical features, land use patterns, and human activities. This would provide a more comprehensive understanding of the complex dynamics of forest fires and contribute to the development of more accurate and robust predictive models. The models and insights gained from this analysis can contribute to the development of effective strategies for forest fire prevention, mitigation, and management, leading to enhanced safety, environmental conservation, and preservation of natural resources.

Recommendations and Implementation

It is recommended to improve the data collection process by incorporating additional variables and factors that may impact forest fire behavior. Variables such as land use patterns, topographical features, and human activities can provide valuable insights and enhance the accuracy of predictive models. Integrating these data sources can contribute to a more comprehensive understanding of the underlying dynamics of forest fires. Another recommendation is to enhance the spatial and temporal resolution of the data. Collecting data at a finer spatial and temporal scale can capture more localized and dynamic patterns of forest fires.

More advanced modeling techniques beyond the ones considered in this study could be explored. While Random Forest, Neural Network, and Support Vector Machine models were effective in capturing complex relationships, other techniques such as gradient boosting, ensemble methods, or deep learning architectures may yield even better results. These methods have shown promising outcomes in various domains and can potentially enhance the accuracy and robustness of forest fire prediction models.

It is essential to continually update and refine the models as new data becomes available. The field of forest fire prediction is dynamic, and incorporating up-to-date information can improve the models' performance. Regular model evaluation and retraining should be conducted to ensure their effectiveness over time.

If the above recommendations are taken and a model is refined and validated, the next step is to integrate it into practical applications for forest fire management and mitigation. This can involve collaborating with relevant stakeholders, such as forest management agencies, firefighting departments, and local communities. The model can be incorporated into early detection systems, decision support tools, and risk assessment frameworks to aid in proactive fire management strategies. Training programs and workshops can be conducted to familiarize stakeholders with the model and its applications, enabling them to make informed decisions and take appropriate actions in response to predicted fire risks. Regular communication and collaboration with experts in the field, researchers, and policymakers can further enhance the model's impact and facilitate knowledge sharing.

Ethical Assessment

The development and implementation of a forest fire prediction model raise important ethical considerations. It is crucial to ensure the responsible and ethical use of data, which involves obtaining informed consent from individuals whose data is collected, ensuring data privacy and confidentiality, and adhering to relevant data protection laws and regulations. Potential biases and fairness must also be considered. Biases can arise from the data used to train the model, leading to discriminatory outcomes or perpetuating existing inequalities. It is important to carefully select and preprocess the data to mitigate biases and ensure fairness in the predictions. Regular monitoring and auditing of the model's performance should be conducted to identify and rectify any biases that may arise during its operation.

Transparency and interpretability of the model are crucial ethical considerations. Stakeholders, including policymakers, fire management agencies, and the public, should have access to information about the model's underlying algorithms, data sources, and assumptions. Clear documentation and explanations of the model's functioning can help foster trust and enable stakeholders to understand and question its predictions.

The predictions provided by the model should be used as tools to inform decision-making, but not as absolute determinants. Human expertise, local knowledge, and contextual factors should also be considered in the decision-making process. Regular evaluation and validation of the model's performance against real-world outcomes can help identify and mitigate any unintended consequences.

The equitable distribution of resources and support for fire management and mitigation efforts must also be considered. Disadvantaged communities or regions should not be disproportionately burdened or excluded from the benefits of the model's predictions and interventions. Efforts should be made to ensure equitable access to resources, information, and support for all communities affected by forest fires.

Questions

1. What is the significance of accurately predicting the extent of forest fire damage?
2. How does this research contribute to enhancing early detection systems and supporting fire management decisions?
3. What meteorological factors were found to be influential in predicting the impact of forest fires?
4. How does this research address the issue of imbalanced data in forest fire occurrences?
5. What are the limitations of the dataset used in this analysis, and how might they affect the generalizability of the findings?
6. How do the selected models (Random Forest, Neural Network, SVM) differ in their ability to capture complex relationships and patterns?
7. Can the findings from this analysis be applied to other regions or countries that experience forest fires?
8. How reliable are the evaluation metrics (RMSE, R-Squared) in assessing the performance of the predictive models?
9. Were there any unexpected or surprising findings during the exploratory data analysis or modeling process?
10. How might the results and insights from this research be used to inform future fire management strategies or policies?

References

Clarke, H., Nolan, R.H., De Dios, V.R. *et al.* Forest fire threatens global carbon sinks and population centres under rising atmospheric water demand. *Nat Commun* **13**, 7161 (2022). https://doi.org/10.1038/s41467-022-34966-3

European-Commission. Forest Fires in Europe. Technical report, Report N-4/6, 2017.

Halofsky, J.E., Peterson, D.L. & Harvey, B.J. Changing wildfire, changing forests: the effects of climate change on fire regimes and vegetation in the Pacific Northwest, USA. *Fire ecol* 16, 4 (2020). https://doi.org/10.1186/s42408-019-0062-8

Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O., & Akinjobi, J. (2017) Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology (IJCTT)-*, 48(3). https://doi.org/ISSN:2231-2803

P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978 989-95618-0-9.

Code Appendix

```r
library(gridExtra)
library(ggplot2)
library(kableExtra)
library(grid)
library(dplyr)
library(caret)
library(ranger)
library(rpart)
library(nnet)
library(e1071)
library(kernlab)

# forest fire data
df <- read.csv("forestfires.csv")

# Check for missing values in the dataset
missing_values <- colSums(is.na(df))

# Print the variables with missing values
print(missing_values[missing_values > 0])

# Data Dictionary
data_dict <- data.frame(
  Variable = c("X", "Y", "Month", "Day", "FFMC", "DMC", "DC", "ISI", "Temp",
               "RH", "Wind", "Rain", "Area"),
  Description = c("X-axis spatial coordinate within the Montesinho park map",
                  "Y-axis spatial coordinate within the Montesinho park map",
                  "Month of the year (jan to dec)",
                  "Day of the week (mon to sun)",
                  "Fine Fuel Moisture Code from the FWI system",
                  "Duff Moisture Code from the FWI system",
                  "Drought Code from the FWI system",
                  "Initial Spread Index from the FWI system",
                  "Temperature in Celsius degrees",
                  "Relative humidity in percentage",
                  "Wind speed in km/h",
                  "Rainfall in mm/m^2",
                  "Burned area of the forest (in hectares)"),
  stringsAsFactors = FALSE
)

# Format the Data Dictionary table using kableExtra
data_dict_table <- kable(data_dict, format = "latex", booktabs = TRUE,
                         caption = "Data Dictionary") %>%
  kable_styling(latex_options = c("striped", "hold_position"),
                full_width = FALSE) %>%
  row_spec(1, bold = FALSE) %>%
  collapse_rows(columns = 2, valign = "top") %>%
  column_spec(1, border_right = FALSE)

# Print the Data Dictionary table
```

```r
print(data_dict_table)

# Create a histogram of the "area" variable
ggplot(df, aes(x = area)) +
  geom_histogram(fill = "#6A5ACD", binwidth = 15) +
  labs(title = "Figure 1: Distribution of Burned Area",
       x = "Area",
       y = "Frequency") +
  theme_light() +
  theme(plot.title = element_text(hjust = 0.5))

# Numeric variables
numeric_vars <- c("FFMC", "DMC", "DC", "ISI", "temp", "RH", "wind", "rain")

# Calculate correlations
correlations <- sapply(numeric_vars, function(var) {
  corr <- cor(df[[var]], df$area)
  paste0(round(corr, 2))
})

corr <- as.numeric(correlations)

# Create scatterplots
scatterplots <- lapply(seq_along(numeric_vars), function(i) {
  var <- numeric_vars[i]

 ggplot(df, aes_string(x = var, y = "area")) +
    geom_point(color = "#6A5ACD", fill = "#6A5ACD", size = 1) +
    annotate("text", x = min(df[[var]]), y = max(df$area),
             label = paste("r =", round(cor(df[[var]], df$area), 2)),
             hjust = 0, vjust = 1,
             color = ifelse(abs(corr[i]) == max(abs(corr)), "red", "black"),
             size = 3) +
    labs(title = paste(var, "vs. Area"),
         x = var,
         y = "Area") +
    theme_light()
})

# Arrange scatterplots in a custom grid layout
scatter_grid <- grid.arrange(
  scatterplots[[1]], scatterplots[[2]], scatterplots[[3]],
  scatterplots[[4]], scatterplots[[5]], scatterplots[[6]],
  scatterplots[[7]], scatterplots[[8]],
  nrow = 3, ncol = 3,
  widths = c(1, 1, 1),
  heights = c(3, 3, 3)
)

# Create the title grob
title <- textGrob("Figure 2: Burned Area vs. Features",
                  gp = gpar(fontsize = 14))
```

```r
# Arrange the scatterplot grid and the title using gridExtra
grid.arrange(title, scatter_grid, nrow = 2, heights = c(0.2, 3.8))

# Calculate the average burned area by month
monthly_avg <- aggregate(area ~ month, df, mean)

# Define the order of months
month_order <- c("jan", "feb", "mar", "apr", "may", "jun", "jul", "aug",
                 "sep", "oct", "nov", "dec")
monthly_avg$month <- factor(monthly_avg$month, levels = month_order)

# Create the line plot
ggplot(monthly_avg, aes(x = month, y = area, group = 1)) +
  geom_line(color = "#6A5ACD") +
  geom_point(color = "#6A5ACD", size = 3) +
  labs(title = "Figure 3: Average Burned Area by Month",
       x = "Month",
       y = "Average Burned Area") +
  theme_light() +
  theme(plot.title = element_text(hjust = 0.5))

# Set seed for reproducibility
set.seed(123)

df <- read.csv("forestfires.csv")

# Subset the data where area is not equal to 1
subset_data <- df[df$area != 0, ]

df <- rbind(df, subset_data)
df <- rbind(df, subset_data)

# Drop NA values from df
df <- na.omit(df)

# Add 1 to numeric features
numeric_vars <- c("FFMC", "DMC", "DC", "ISI", "temp", "RH", "wind", "rain")
df[, numeric_vars] <- df[, numeric_vars] + 1

# Add 1 to the "area" outcome
df$area <- df$area + 1

# Convert 'X' and 'Y' to numeric
df <- df %>%
  mutate(X = as.numeric(X),
         Y = as.numeric(Y))

# Convert 'month' and 'day' to factors
df <- df %>%
  mutate(month = as.factor(month),
         day = as.factor(day))

# Split data into training and testing sets
```

```r
train_indices <- createDataPartition(df$area, p = 0.6, list = FALSE)
train_data <- df[train_indices, ]
test_data <- df[-train_indices, ]

ctrl <- trainControl(method = "cv", number = 10)

# Define parameter grids for tuning
param_grid_rf <- expand.grid(mtry = seq(2, 8, by = 2),
                             min.node.size = c(1, 5, 10),
                             splitrule = "variance")
param_grid_svm <- expand.grid(
  C = c(0.1, 1, 10),
  sigma = c(0.1, 1, 10)
)
param_grid_nn <- expand.grid(size = seq(5, 20, by = 5),
                             decay = c(0.001, 0.001, 0.0001))

# Train the random forest model using caret and ranger
model_rf <- train(area ~ ., data = train_data, method = "ranger",
                  trControl = ctrl, tuneGrid = param_grid_rf)

# Train the decision tree model using caret and rpart
model_svm <- train(
  area ~ .,
  data = train_data,
  method = "svmRadial",
  trControl = trainControl(method = "cv", number = 5),
  tuneGrid = param_grid_svm
)

# Train the neural network model using caret and nnet
model_nn <- train(area ~ ., data = train_data, method = "nnet",
                  trControl = ctrl, tuneGrid = param_grid_nn)

# Make predictions on test data
predictions_rf <- predict(model_rf, newdata = test_data)
predictions_svm <- predict(model_svm, newdata = test_data)
predictions_nn <- predict(model_nn, newdata = test_data)

# Calculate RMSE
rmse_rf <- RMSE(predictions_rf, test_data$area)
rmse_svm <- RMSE(predictions_svm, test_data$area)
rmse_nn <- RMSE(predictions_nn, test_data$area)

# Calculate R-squared
r_squared_rf <- R2(predictions_rf, test_data$area)
r_squared_svm <- R2(predictions_svm, test_data$area)
r_squared_nn <- R2(predictions_nn, test_data$area)
# Create a data frame with model names and evaluation metrics
evaluation_table <- data.frame(
  Model = c("Random Forest", "SVM", "Neural Network"),
  RMSE = c(rmse_rf, rmse_svm, rmse_nn),
  R_squared = c(r_squared_rf, r_squared_svm, r_squared_nn)
```

```
)

# Highlight the lowest RMSE and highest R-squared values in red
evaluation_table$RMSE <- ifelse(evaluation_table$RMSE
                                == min(evaluation_table$RMSE),
                                sprintf("\\textcolor{red}{%.3f}",
                                        evaluation_table$RMSE),
                                sprintf("%.3f", evaluation_table$RMSE))

evaluation_table$R_squared <- ifelse(evaluation_table$R_squared
                                     == max(evaluation_table$R_squared),
                                     sprintf("\\textcolor{red}{%.3f}",
                                            evaluation_table$R_squared),
                                  sprintf("%.3f", evaluation_table$R_squared))

# Set the column names
colnames(evaluation_table) <- c("Model", "RMSE", "R-squared")

# Print the table using kable with LaTeX formatting
kable(evaluation_table, format = "latex", booktabs = TRUE, escape = FALSE,
      caption = "Model Evaluation Metrics") %>%
  kable_styling()
```