Capstone Project

Predicting Bushfires in Australia

Shelley Jones

02 November 2024

Table of Contents

Problem Statement

The project investigates the challenge of accurately predicting the likelihood and intensity of bushfires in specific regions of Australia using publicly available environmental and climatic data from 2020-2024. Given the increasing frequency and severity of bushfires due to climate change, effective prediction could significantly mitigate risks to life, property, and ecosystems.

Addressing the problem of bushfires is valuable for several reasons. Firstly, it significantly enhances public safety; early predictions can facilitate timely evacuations and effective resource allocation, potentially saving lives. Secondly, by reducing the frequency and impact of bushfires, we can lessen the economic burden on governments and communities, which includes costs associated with firefighting, recovery efforts, and insurance claims. Additionally, accurate predictions play a crucial role in environmental protection, helping to safeguard biodiversity and ecosystems that are affected by these devastating events. Finally, improved predictive models provide essential information for policymakers and local governments, enabling them to identify risk areas and make informed decisions regarding land use planning and resource allocation.

Currently, many regions in Australia struggle with bushfires that are difficult to predict, resulting in several significant challenges. One major issue is unsatisfied customers; communities often feel inadequately informed and protected during bushfire seasons, leading to frustration and anxiety. Additionally, lost revenue is a critical concern, as property and agricultural losses due to bushfires can be devastating, adversely affecting local economies. Moreover, emergency services frequently experience strain during severe bushfire events, as firefighting resources become overstretched, which ultimately impacts response times and overall effectiveness in managing these crises.

Ideally, the goal would be to establish a robust machine learning framework that effectively predicts bushfire likelihood and intensity with at least 80% accuracy in forecasts. This framework would enable proactive measures by providing actionable insights for emergency services and communities, allowing them to prepare for potential bushfire events. Additionally, it would improve resource allocation, assisting in the more effective deployment of firefighting resources based on predicted risks. Lastly, the framework would enhance public awareness by informing the community about bushfire risks in a timely manner, fostering a more prepared and resilient population.

The problem has been addressed by various research projects, yielding mixed outcomes. Some studies have focused on predictive modelling, developing models that use historical data on weather, vegetation, and past fire occurrences. These models often employ machine learning techniques such as random forests and neural networks, achieving varying levels of accuracy between 70% and 90%, which reflects the work in my project. Other projects have concentrated on risk assessment frameworks that integrate social, environmental, and climatic data, although they sometimes lack real-time applicability. Additionally, geospatial analysis has been utilised to identify high-risk areas, but these approaches have often fallen short in predicting fire intensity.

Overall, while there has been progress, challenges remain in achieving high accuracy and generalising models across different regions and conditions. My project aimed to build on this foundation by refining predictive capabilities specifically for Australian bushfires, focusing on utilising recent, publicly available data to enhance both reliability and applicability.

Industry / Domain

The project falls within the **environmental science** and **data science** domains, specifically focusing on **wildfire management** and **climate analytics**. This industry includes government agencies, environmental organisations, and private sector companies involved in disaster management, environmental monitoring, and climate impact assessment.

Increased frequency and severity of wildfires due to climate change are major concerns, alongside data overload from vast amounts of information that can be overwhelming. Resource constraints also limit funding for firefighting and predictive technologies, while emerging startups create competition with innovations in predictive analytics and AI. Additionally, integration issues often hinder the effective use of diverse data sources (as I have experienced in my project).

The industry's value chain includes key stages such as data collection, which involves gathering environmental and meteorological data; data analysis, utilising analytics and machine learning to process this information; model development, where predictive models for fire likelihood and intensity are created; and decision support, providing actionable insights to emergency services and policymakers. Implementation follows, executing emergency plans based on predictions, and post-event analysis evaluates responses to refine models for future use.

Key concepts within the industry encompass predictive analytics for forecasting bushfire risks, remote sensing for environmental monitoring, geospatial analysis to identify high-risk areas, climate change adaptation strategies, and emergency management for planning responses. This project is also relevant to various sectors, including insurance, where predictive models enhance risk assessment; agriculture, aiding farmers in protecting crops; urban planning, incorporating fire risk into development decisions; and climate research, contributing to broader environmental understanding.

Stakeholders

Government agencies play a critical role in bushfire management, with State and Federal fire services responsible for fire response and community safety. Environmental protection agencies monitor the impact of fires on biodiversity and ecosystem health, while emergency management organisations coordinate disaster responses and resource allocation. Their overarching goal is to protect public safety and effectively manage disaster responses while safeguarding the environment.

Local communities, particularly residents in fire-prone areas, are deeply concerned about safety and property protection. Community groups engage in advocacy and preparedness initiatives to enhance resilience. The agricultural sector is also significantly affected by bushfires, with farmers seeking to protect their crops and livestock, while agricultural associations advocate for improved risk management and support for affected farmers.

Other stakeholders include insurance companies, which require accurate risk assessments for fair pricing and claims management, and environmental NGOs that prioritise the protection of natural habitats. Academia and research institutions are focused on advancing fire prediction models, while technology companies seek to innovate in predictive analytics and monitoring. Utilities and infrastructure providers are concerned about protecting their assets and minimising liabilities.

Business Question

The primary question is how accurately we can predict the likelihood and intensity of bushfires in specific regions of Australia using historical environmental and climatic data from 2020 to 2024.

Answering this question could lead to significant cost savings. If predictive models can reduce severe bushfire events by 20%, this would save approximately AUD 6 million per event in firefighting costs. Additionally, preventing the loss of 100 homes and 500 acres of crops annually could result in about AUD 25 million in property and crop protection savings. For insurance companies, a 15% reduction in fire-related claims could save AUD 200 million annually. Improved predictions would also enhance community resilience, contributing to long-term economic stability in affected regions.

To make a meaningful impact, a prediction accuracy of at least 80% is necessary, ensuring stakeholder confidence in the forecasts and effective resource allocation. However, false positives could waste resources and erode public trust, while false negatives could lead to severe consequences, including loss of life and increased financial burdens. Ultimately, achieving this accuracy could provide substantial economic value, estimated at AUD 231 million annually, while improving community safety and resilience. Addressing the risks of false predictions is essential for stakeholder confidence and the overall success of the predictive framework.

Data Question

The key question is: What historical environmental and climatic factors most significantly influence the likelihood and intensity of bushfires in specific regions of Australia from 2020 to 2024?

To answer this question at its basic level, several types of data are needed:

1. **Bushfire Incident Data**: This includes geospatial data indicating the locations of bushfires, temporal data detailing when fires were detected and ideally their duration, and measures of fire intensity, such as the heat released.

2. **Environmental Data**: Relevant information including how prone to bushfire an area is taking into account vegetation type, land cover and vegetation characteristics (e.g., forest types, grassland), and fuel load data, providing insight into available combustion fuel, including moisture content.

3. **Climatic Data**: This encompasses historical temperature records, precipitation data, relative humidity levels, wind speed and direction.

Data

Data was sourced from multiple locations:

1. **Bushfire Incident Data:** Sourced from NASA via their online portal for access to archived VIIRS data. VIIRS (Visible Infrared Imaging Radiometer Suite) is an advanced sensor used on the Suomi National Polar-orbiting Partnership (Suomi NPP) and NOAA-20 satellites. It captures a wide range of data across various spectral bands, which allows it to monitor and provide valuable insights into several environmental and atmospheric phenomena such as active fires and hotspots. For the years 2020-2024, over 10 million rows of data were available across 17 columns.

   https://firms.modaps.eosdis.nasa.gov/data/download/DL_FIRE_M-C61_525305.zip

latitude	longitude	brightness	scan	track	acq_date	acq_time	satellite
-34.46005	150.88739	337.26	0.42	0.38	2015-01-01	402	N

confidence	version	bright_t31	frp	daynight	type
n	2	300.09	6.08	D	2

2. **Environmental Data**: Each State and Territory in Australia maintains its own data on bushfire-prone areas, however not all governments make this information easily accessible or available for direct download. An online report by 360info included a link to a GitHub repository where the datasets used for reporting on bushfire-prone areas were provided. This repository contained valuable information, such as the percentage of bushfire-prone areas for each suburb.

   The dataset consisted of 9 columns and 15,345 rows.

   https://github.com/360-info/report-bushfire-prone-land

	STE_CODE21	STE_NAME21	SAL_CODE21	SAL_NAME21	area	bf_area	bf_area_pct	cent_lat	cent_lon
12007	1	New South Wales	14442	Worrowing Heights	13.518403	12.983327	0.960419	-35.076977	150.640627
4242	2	Victoria	20401	Bungador	34.920689	34.920689	1.000000	-38.424096	143.328603
3074	5	Western Australia	50934	Menzies (WA)	9567.963491	8514.606422	0.889908	-29.710359	121.132402

3. **Climatic Data:** A dataset was sourced from Kaggle, where a contributor had compiled three separate datasets from 2000-2024. These primarily consisted of information related to suburbs, states/territories, and historical daily recordings of maximum, mean

and minimum temperatures, as well as rainfall. However, humidity and wind speed data were unfortunately not included in this compilation.

https://www.kaggle.com/datasets/nadzmiagthomas/australia-weather-data-2000-2024

The datasets were as follows:

a. Suburbs and their associated ClusterID (suburbs were clustered together by location e.g. Local Government Area)

	OfficialNameSuburb	OfficialNameState	ClusterID
0	Adaminaby	NSW	100412.0
1	Albury	NSW	101407.0

b. Each ClusterID had associated weather in a second dataset

ClusterID	Datetime	TemperatureMean	TemperatureMax	TemperatureMin	RainSum
100412	1999-12-31 00:00:00+00:00	4.318500	7.0785	3.3785	0.0
100412	2000-01-01 00:00:00+00:00	8.793083	13.0785	4.1785	0.0

c. The third dataset consisted of ClusterID, Latitude and Longitude. This dataset was an Excel file and I didn't need to use it as I already had the geographical data below.

ClusterID	Latitude	Longitude
100412	-35.85585586	148.64864864864865
92411	-32.97297297	148.28828828828827
101407	-36.21621622	146.84684684684683

4. **Geographical Data:** A dataset of Australian suburbs, their associated state, latitude and longitude was identified by internet searching. It consisted of 140 rows and 4 rows.

https://www.peter-johnson.com.au/AustraliaPlaces

	Acacia Park	NSW	-30.5403962	151.6804555
0	Akuna Bay	NSW	-33.648331	151.2344447
1	Albion Heights	TAS	-42.958049	147.3215466

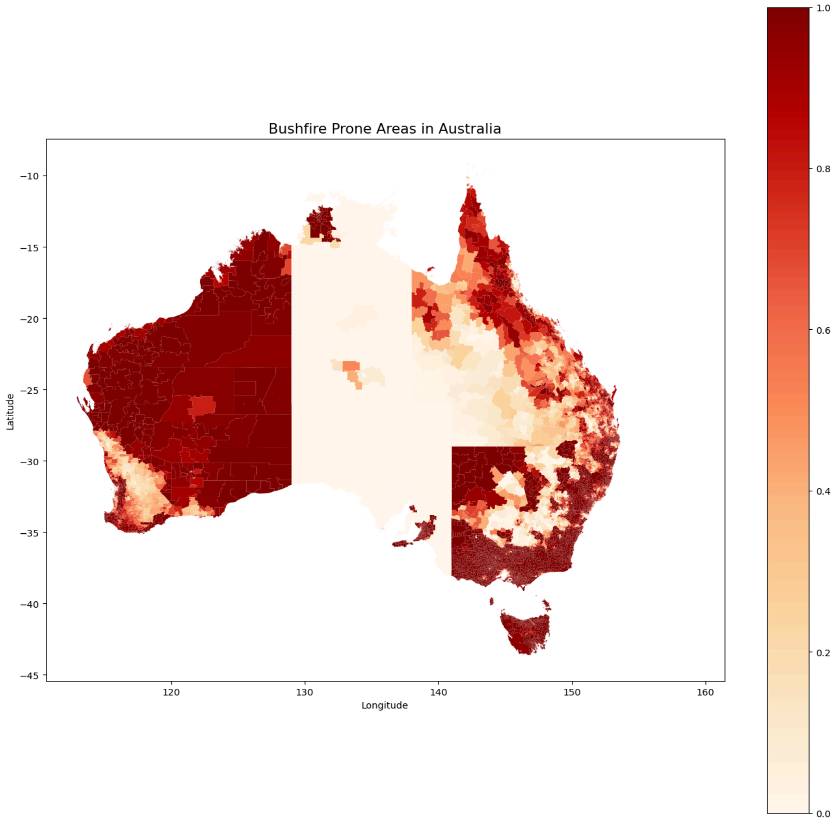
Data sourced from official government agencies and reputable institutions is generally regarded as reliable, with historical data typically well-documented. As a result, the overall quality of this data is considered good to high.

Data Science Process

Data Analysis

Bushfire Prone Areas

To visualise bushfire-prone areas on a map, I downloaded a shapefile from the Australian Bureau of Statistics and merged it with my existing bushfire-prone areas data using 'suburb' as the common key. After cleaning the data, I generated the following plot:



Several insights emerged from the analysis:

- Each State and Territory has its own criteria for identifying bushfire-prone areas, evident from the distinct 'cut-offs' in risk levels, particularly between Western Australia and the Northern Territory/South Australia.
- South Australia and the Northern Territory exhibited similar approaches, although only a limited number of areas were classified as bushfire prone.
- Tasmania categorised the majority of the state as high bushfire prone. While this may reflect the state's vegetation cover, the overall risk of bushfires in Tasmania is generally lower due to its comparatively lower mean temperatures than the rest of Australia.
- New South Wales and Queensland appeared to have aligned rating systems.

Based on these observations, I decided to focus further analysis exclusively on New South Wales and Queensland, due to the computational limitations of modelling Australia-wide data on a laptop.

Bushfire Incident Data

The location data in this dataset was limited to latitude and longitude coordinates, so I needed to determine the corresponding suburbs. To achieve this, I constructed a 'suburb tree' using the cKDTree function from scipy.spatial to identify the nearest suburb and its associated state for each set of coordinates in the bushfire incident dataset.

To merge the bushfire incident data with the bushfire prone areas data, I identified suburbs in the bushfire incident dataset that were absent from the bushfire prone dataset. I replaced these missing suburbs with the nearest suburb from the bushfire prone dataset, utilising the great_circle function from geopy.distance for this purpose. I set a maximum distance of 10 km, as distances greater than this could introduce inaccuracies.

Some suburbs were not located, and since I lacked data for them, I chose to exclude them from further analysis. The dataset remained substantial, so I focused on bushfire incidents classified as either 'nominal' or 'high' confidence. By removing rows with 'low' confidence, I reduced the dataset to 1,224,891 records, maintaining a robust sample size without compromising data quality.

After completing the mapping process and subsequent data cleaning and encoding, I developed a dataset that included both bushfire incident data and the percentage of bushfire prone area for each suburb.

Climatic Data

The initial step involved merging the two climatic datasets using 'ClusterID' as the unique key. This process provided the suburb, state, and daily weather data for the period from January 1, 2020, to July 31, 2024 (the cutoff date for the bushfire incident dataset). I then created a unique key in both the climatic dataset and the merged bushfire incidents/bushfire prone areas dataset by concatenating the 'suburb' and 'date' columns. This unique key facilitated the merging of all data into a comprehensive working dataset.

Throughout this process, I used a test suburb (Allambie Heights) to verify that the merging functioned as intended. Once I was confident in the integrity of the working dataset, I saved it as a 'pickle file' to allow continuation in a new Jupyter notebook without the need to rerun all previous code.

Final Steps

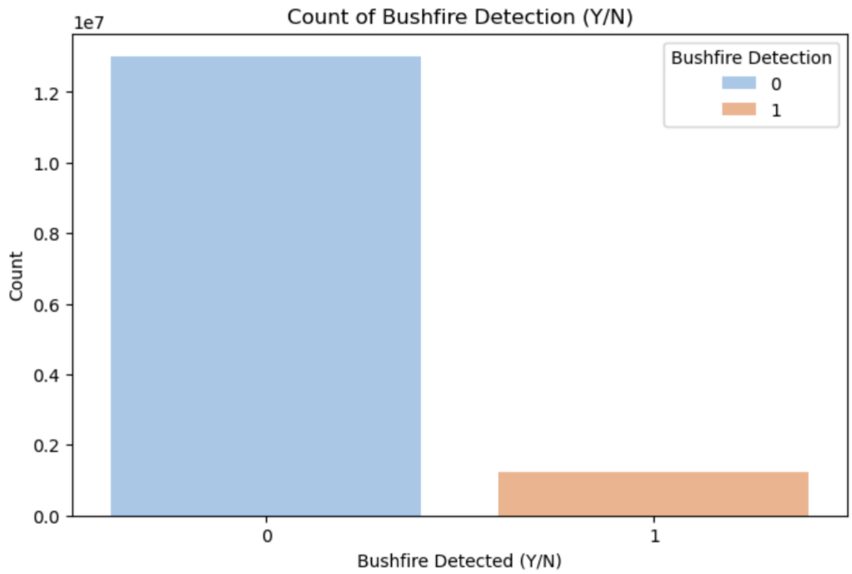
With the working dataset established, I created a new feature (bushfire y/n) and populated this column for each row by referencing the 'acquisition time.' Specifically, if a row was null, it indicated that no bushfire occurred in that suburb on that day.

I noted that there were 1,881,576 rows where the bf_area_pct (percentage of bushfire prone land) was null. While I could have examined the nearest surrounding suburbs with valid bf_area_pct values and imputed the mean, time constraints during this project prevented me from doing so. This could be a potential avenue for future work, or I could consider reaching out to Local Government Area councils to obtain the necessary data.

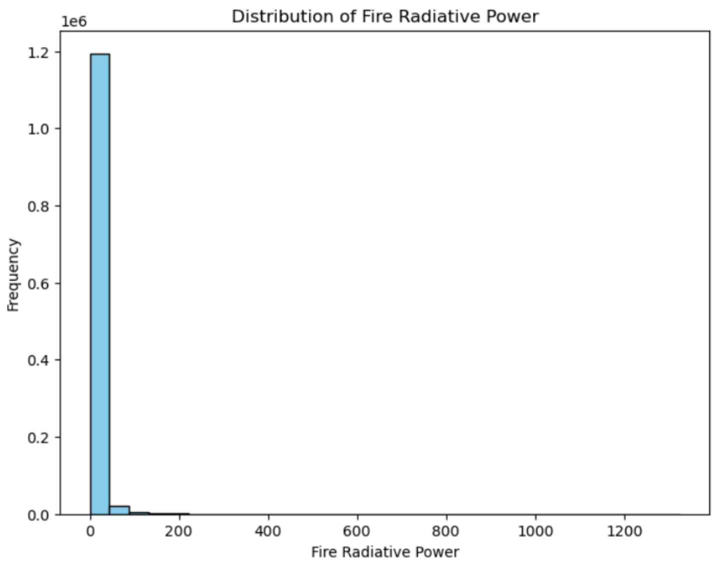
Ultimately, the missing data represented approximately 13% of the entire dataset, so I chose to drop these rows and proceeded with the analysis without them.

Exploratory Data Analysis and Visualisations

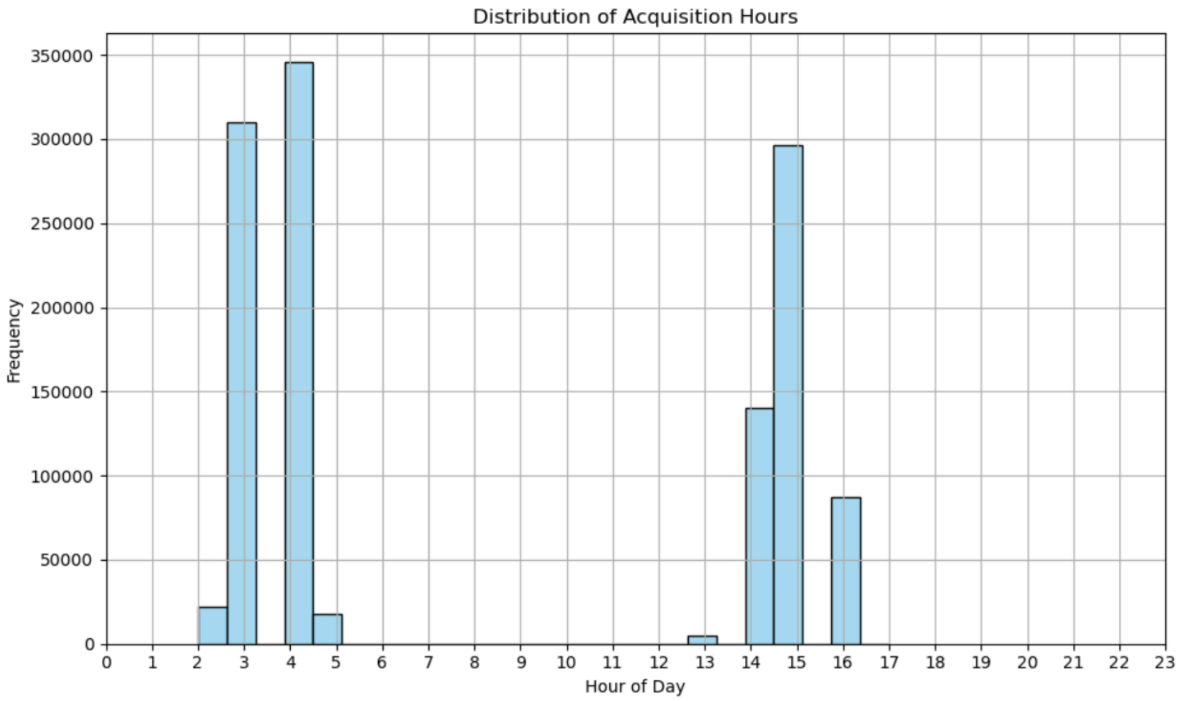
The target variables (bushfire y/n and Fire Radiative Power) exhibited significant imbalances, with the bushfire y/n target constituting only 8.6% of the dataset.



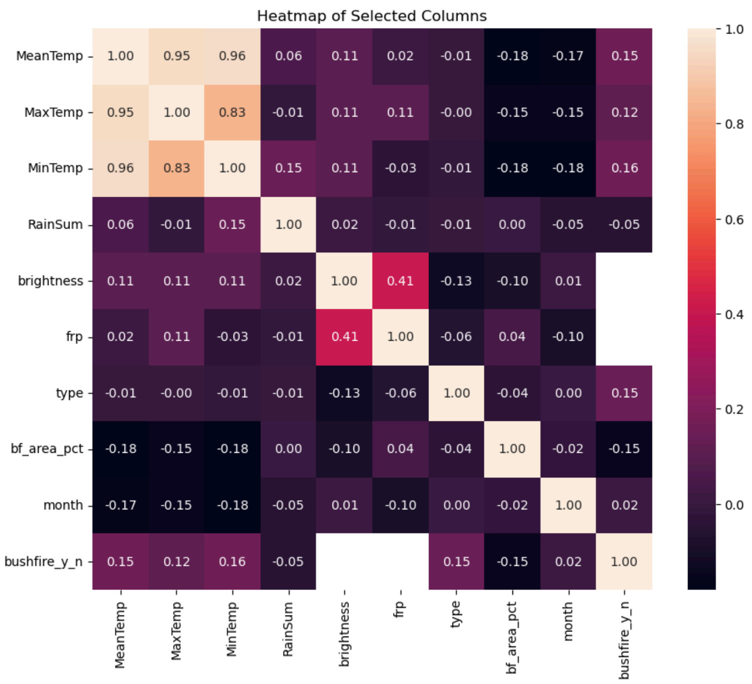
Fire Radiative Power (frp) was also highly skewed, necessitating careful handling during modelling, particularly through scaling and stratification during data splitting.



Interestingly, image acquisition time was divided into two main blocks, likely reflecting the satellite's overpass times over Australia. As a result, the ‘hour’ feature was deemed not particularly useful for modelling, as it only indicated when the satellite captured the image, rather than the actual time of day a bushfire occurred.



To explore correlations with bushfires, I plotted a heatmap using selected features, given the dataset's size.



The correlation heatmap revealed minimal correlations with maximum, mean, and minimum temperatures, as well as the type of fire (notably vegetation type fires), and the percentage of bushfire prone area. There was no significant correlation between bushfire y/n and either rainfall or month. The variables related to brightness and frp were blank due to missing values; however, this would not affect the ability to predict the occurrence of bushfires at this stage of modelling.

I found that temperatures were largely collinear, leading me to drop the MeanTemp feature from the modelling process. I was also surprised to observe that frp, which indicates fire intensity, had minimal correlation with maximum temperature.

Pipeline

With some modifications, the pipeline is designed to be reusable and modular, facilitating easy adjustments to accommodate future data. For instance, if new variables are introduced, corresponding feature engineering steps can be seamlessly integrated. Additionally, changes to data sources—such as incorporating weather APIs—can be made without the need to redesign the entire pipeline.

Modelling

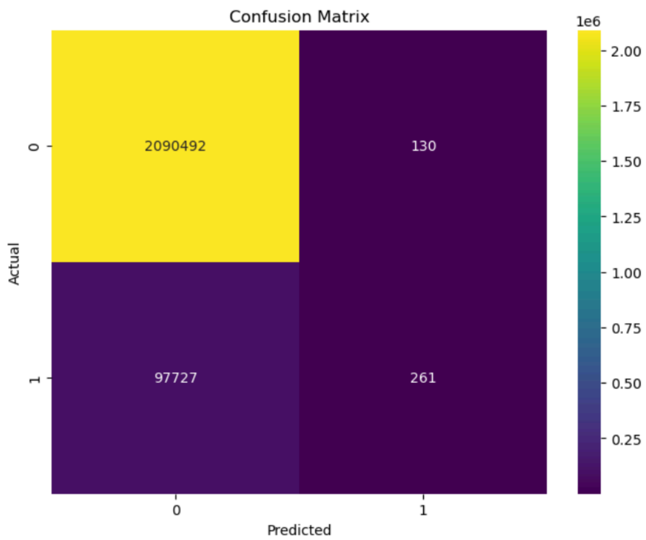
Module 1: To Predict Bushfire Probability

For this module, the selected features included 'suburb_encoded', 'MaxTemp', 'MinTemp', 'RainSum', 'bf_area_pct', and 'month'. The target feature was 'bushfire y/n'.

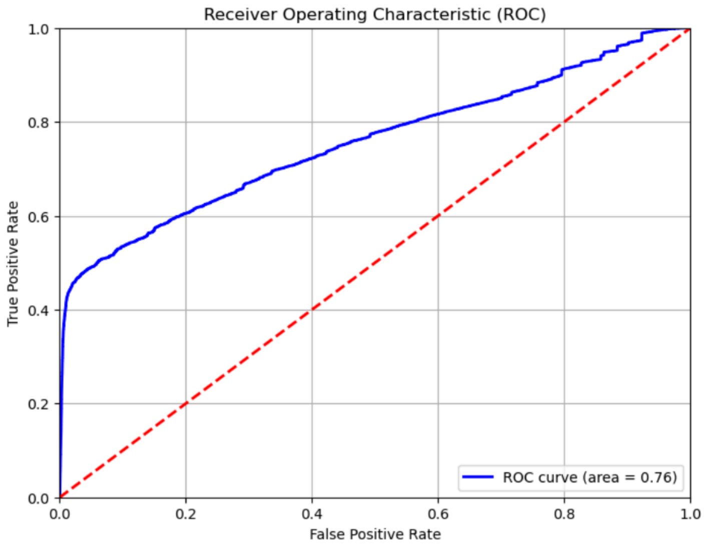
Two models were evaluated: Logistic Regression and Random Forest Classifier. The evaluation metrics employed included training accuracy, testing accuracy, precision, recall, ROC curve, and AUC.

Logistic Regression

The model took approximately 20 minutes to fit and generate predictions. The results were as follows:



	precision	recall	f1-score	support
0	0.96	1.00	0.98	2090622
1	0.67	0.00	0.01	97988
accuracy			0.96	2188610
macro avg	0.81	0.50	0.49	2188610
weighted avg	0.94	0.96	0.93	2188610



When predicting the bushfire y/n probability, the model achieved an accuracy of 0.9553 for both the training and testing datasets. This outcome suggests a few possibilities:

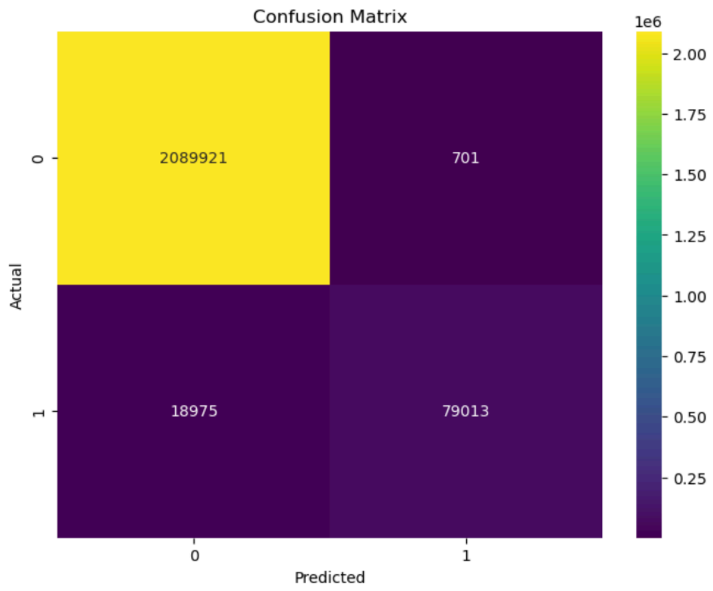
- The model may be generalising well to unseen data
- It does not appear to be overfitting
- The dataset might be too straightforward or clean, enabling high accuracy without requiring complex modelling.

Given the data's significant imbalance, even though I stratified the training and test sets, the model could have a tendency to predict 'no fire'. However, the ROC curve indicated that this was not the case.

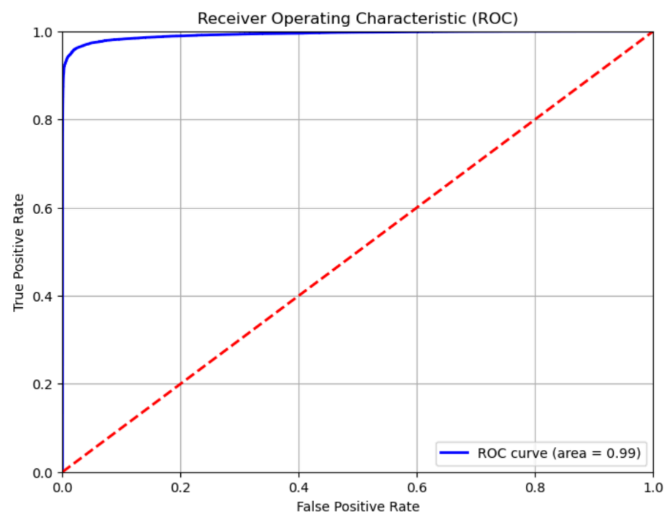
While the model performs well in identifying non-events, it struggles considerably with predicting actual events. Implementing class weighting could potentially enhance event detection.

Random Forest Classifier

I defined the hyperparameter space and identified the optimal parameters for the model. The total time taken to fit the model and generate predictions was approximately 2 hours. The results were as follows:



	precision	recall	f1-score	support
0	0.99	1.00	1.00	2090622
1	0.99	0.81	0.89	97988
accuracy			0.99	2188610
macro avg	0.99	0.90	0.94	2188610
weighted avg	0.99	0.99	0.99	2188610



When predicting the bushfire y/n probability, the model achieved an accuracy of 0.9911 for the training set and 0.9910 for the test set.

At first glance, the Random Forest Classifier demonstrated a significantly improved ability to identify both classes, reflected in its very high overall accuracy. However, without cross-validation, I could not confirm that this high performance would be consistent across different data subsets, rather than specific to this split. More computing power would have been required to conduct cross-validation simultaneously.

For this reason, I opted to use Logistic Regression in the first module of my final model.

Module 2: To Predict Intensity (frp) of a Bushfire

In this module, all features from the cleaned bushfire incident dataset (excluding 'frp') were utilised as input features, with the target feature being 'frp'. Two models were evaluated: Linear Regression and Random Forest Regressor. The evaluation metrics included Mean Squared Error (MSE) and R^2 score.

Method Summary

Model	R^2 (Train)	R^2 (Test)	MSE (Train)	MSE (Test)
Linear Regression	0.28	0.26	0.70	0.81
Random Forest Regressor	0.78	0.52	0.22	0.64

The Linear Regression model produced relatively low R^2 values, indicating that it explains only a small portion of the variance in the target variable. The increase in Mean Squared Error (MSE) from training to testing suggests some overfitting, as the model does not generalise well to the test data.

Conversely, the Random Forest model performed significantly better on the training set, reflected by the high R^2 value (0.78) and low MSE (0.22). However, the substantial drop in R^2 (to 0.42) and the increase in MSE (to 0.64) on the test set indicate overfitting; while the model captures the training data effectively, it struggles to generalise to unseen data.

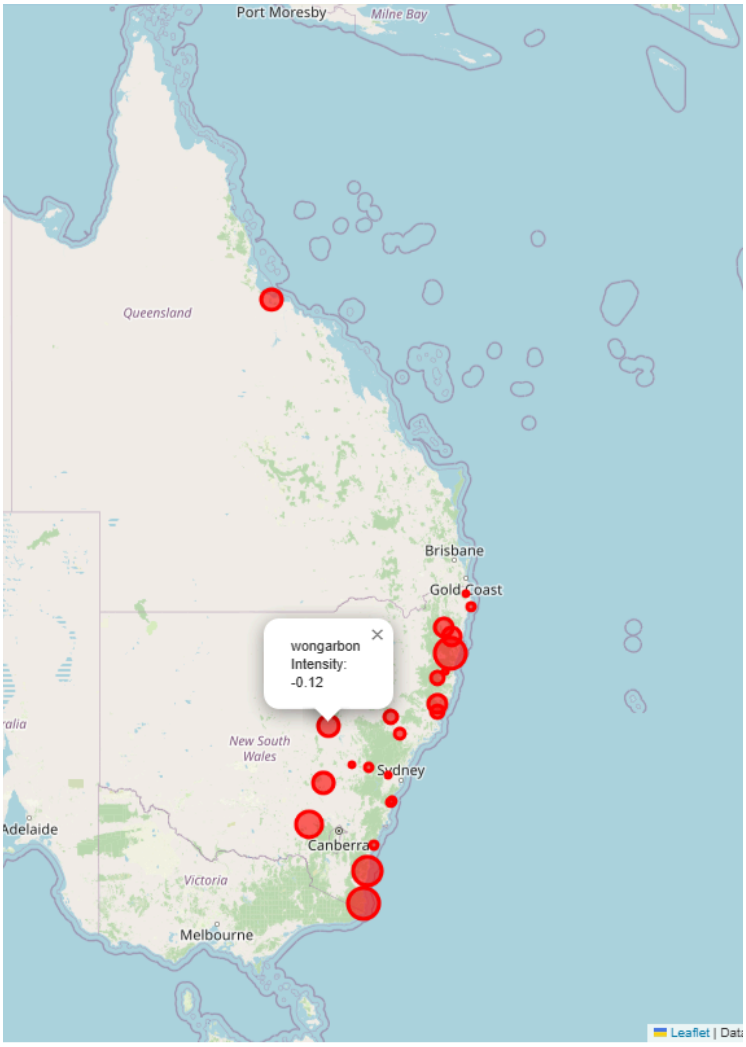
For these reasons, I decided to use Linear Regression in the second module of my final model.

Using Both Models to Predict Fire Activity and Intensity

With my two selected models ready, I created a new dataset representing 'dummy' weather conditions. I passed this data to the first model to predict which suburbs were likely to be affected by bushfires based on these hypothetical conditions.

Next, I utilised the bushfire incident data to calculate the monthly mean intensity for each predicted suburb. I then combined this data to generate a plot illustrating the predicted fire locations and their intensity.

The output is shown below:



Outcomes

The modelling process revealed that the Logistic Regression model achieved a high accuracy of 95.53% in predicting bushfire occurrences, though it struggled with data imbalance, mainly identifying non-events. In contrast, the Random Forest Classifier showed excellent training accuracy (99.11%) but suffered from overfitting, leading to poorer performance on the test set. For predicting fire intensity (FRP), Linear Regression produced low R^2 values, indicating limited explanatory power, while the Random Forest Regressor also exhibited overfitting.

Both models faced challenges due to the high imbalance in bushfire occurrence data, which impacted their ability to predict actual fire events. Key features like temperature and bushfire-prone area percentage were identified as relevant for predicting bushfire activity, although the relationship between weather conditions and fire intensity requires further investigation. Future improvements could include incorporating class weighting, exploring additional data sources, and applying advanced modelling techniques.

Implementation

To effectively implement the model in production, several key considerations must be addressed. First, ensuring continuous access to high-quality, up-to-date data sources is essential for maintaining model accuracy. Regular performance monitoring should be established to track metrics like accuracy and precision, while also checking for concept drift that may affect outcomes over time.

Scalability is crucial, especially during peak bushfire seasons, so the model architecture should leverage cloud resources for processing power. Additionally, optimising inference time is important for providing timely predictions. A user-friendly interface would be necessary for stakeholders to access insights easily, complemented by clear visualisations of risk levels and fire intensities. Finally, thorough documentation and feedback mechanisms will help refine the model, while robust disaster recovery plans would ensure data protection.

Data Answer

The key data question focused on identifying which historical environmental and climatic factors influence the likelihood and intensity of bushfires in specific regions of Australia from 2020 to 2024. Essential data types included bushfire incident data (locations, timings, and intensity measures) and climatic data (historical temperature and rainfall).

The analysis indicated that while certain factors, such as temperature and bushfire-prone area percentage, were relevant to predicting bushfire activity, challenges like data imbalance and overfitting affected model performance. The confidence level in the findings is medium, supported by the approach taken, but limited by issues like overfitting and the lack of cross-validation. Future efforts could enhance confidence through improved data quality and additional feature exploration.

Business Answer

The analysis of historical environmental and climatic data has yielded important insights into the prediction of bushfire likelihood and intensity. Although the models demonstrated high accuracy in predicting bushfire occurrences, with the Logistic Regression model achieving 95.53% accuracy, challenges such as data imbalance and overfitting were present. These factors indicate that while the model can effectively identify non-events, its performance in predicting actual fire events may be limited.

For predicting fire intensity, the results revealed that the models did not adequately capture the complexities of the data, with issues of overfitting observed in the Random Forest Regressor. Overall, while some predictive capability exists, the findings suggest that further refinement of the models is needed to enhance accuracy. Incorporating additional features, improving data quality, and conducting cross-validation would likely lead to more robust predictions.

In conclusion, while the historical data provides a foundation for understanding bushfire patterns, the current models need improvements to increase reliability in predicting both the likelihood and intensity of bushfires across Australia.

Response to Stakeholders

To effectively predict and manage bushfire risks, it is crucial to prioritise access to high-quality, up-to-date data sources, such as weather APIs and bushfire incident reports. Continuous investment in data quality is essential for maintaining the accuracy of predictive models. Additionally, enhancing current modelling approaches by exploring additional features and advanced techniques will help capture the complexities of bushfire activity and intensity more effectively.

Investing in increased computing power or cloud resources will further enable the implementation of more complex modelling, allowing for better predictions and more informed decision-making in bushfire management. By focusing on these areas, stakeholders can improve their predictive capabilities and overall response strategies.

End-to-End Solution

The end-to-end solution for utilising the developed bushfire prediction model begins with establishing reliable data sources, such as weather APIs and bushfire incident reports, ensuring continuous access to high-quality data. Preprocessing the data involves cleaning and handling missing values, followed by feature engineering to create relevant predictors based on significant historical environmental and climatic factors. Once the data is ready, appropriate modelling techniques are applied, and the models are trained and evaluated for performance to ensure robustness.

After training, the models are deployed in a production environment with an intuitive user interface that allows stakeholders to input data and receive predictions on bushfire likelihood and intensity. Continuous monitoring of model performance is essential to detect any changes over time, while regular updates and retraining ensure ongoing accuracy. User training and feedback mechanisms will facilitate effective interpretation of model outputs, promoting iterative improvements.

Finally, the model outputs should be used to generate regular reports that summarise predictions and trends related to bushfire risks. These insights can support decision-making processes in emergency management, helping stakeholders allocate resources more effectively and enhance overall bushfire risk management strategies.

References

Fires data:

https://firms.modaps.eosdis.nasa.gov/data/download/DL_FIRE_M-C61_525305.zip

Coordinates data:

https://www.peter-johnson.com.au/AustraliaPlaces

Bushfire Prone Areas:

https://github.com/360-info/report-bushfire-prone-land

Australian Weather Data:

https://www.kaggle.com/datasets/nadzmiagthomas/australia-weather-data-2000-2024