Predicting Bushfires in Australia

Shelley Jones

02 November 2024

Agenda

- Context
- Business Question
- Key Data Dictionary
- Exploratory Data Analysis
- Data Pre-Processing and Visualisations
- Modelling & Model Evaluation
- Final Output
- Conclusions
- References
- Questions

Context

- The ability to predict a bushfire event and its intensity is crucial for several reasons:
  - Public safety
  - Resource management (intensity vs deployment of firefighting resources)
  - Environmental protection (guiding conservation efforts)
  - Economic impact (mitigate financial losses)
  - Climate adaptation (prepare for future risks)
  - Policy and planning (better land management and urban planning)
- Overall, effective prediction can save lives, protect the environment, and reduce economic costs associated with bushfires.

Business Question

“Using data that is publicly available from 2020-2024, can we develop a machine learning framework that can accurately predict the likelihood and intensity of bushfires in specific regions of Australia using environmental and climatic data, achieving an accuracy of at least 80% within the next four weeks?"

- **Specific**: Clearly defines the goal (predicting bushfire likelihood and intensity in specific regions of Australia)
- **Measurable**: Establishes a clear success criterion (achieving at least 80% accuracy) for evaluating the model's performance
- **Achievable**: With advancements in machine learning and access to relevant data (weather, vegetation, historical fire incidents), this goal is attainable.
- **Relevant**: The question addresses critical public safety and environmental concerns, aligning with broader community and governmental objectives to manage bushfire risks effectively.
- **Time-bound**: Specifies a timeframe (within the next four weeks) for developing the framework.

Key Data Dictionary

The VIIRS dataset contained 15 features in which 'frp' (fire intensity) would be the final target feature:

- Latitude: Center of nominal 375 m fire pixel.
- Longitude: Center of nominal 375 m fire pixel.
- Brightness: Channel 21/22 brightness temperature of the fire pixel measured in Kelvin.
- Scan: The algorithm produces approximately 375 m pixels at nadir. Scan and track reflect actual pixel size.
- Track: The algorithm produces approximately 375 m pixels at nadir. Scan and track reflect actual pixel size.
- Acquisition Date: Date of VIIRS acquisition.
- Acquisition Time: Time of acquisition/overpass of the satellite (in UTC).
- Satellite: N = Suomi National Polar-orbiting Partnership (Suomi NPP). N20 = NOAA-20 (JPSS1). N21 = NOAA-21 (JPSS2).
- Instrument: VIIRS
- Confidence: Quality of individual hotspot/fire pixels. Confidence values are set to low (l), nominal (n), and high (h).
- Version: Version (collection and source)
- Bright_t31: Channel 31 brightness temperature of the fire pixel measured in Kelvin.
- frp: Fire Radiative Power (MW)
- Daynight: D= Daytime fire, N= Nighttime fire
- Type: Inferred hot spot type. 0 = presumed vegetation fire, 1 = active volcano, 2 = other static land source, 3 = offshore detection

Exploratory Data Analysis: Fires

	latitude	longitude	suburb	state	brightness	scan	track	acq_date	acq_time	satellite	instrument	confidence	version	bright_t31	frp	daynight	type
9273352	-22.24022	145.56572	Townsville	QLD	338.66	0.48	0.64	2023-09-06	1615	N	VIIRS	n	2	290.42	2.31	N	0

- Had 10,673,377 rows!
- Reduced the timeframe to 2020-2024: 4,484,869 rows
- Mixture of data types
- No Null values
- No obvious callouts in descriptive statistics
- Categorical values made sense (no ‘rogue’ data)

```
df_fires.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 4484869 entries, 6188508 to 10673376
Data columns (total 15 columns):
 #   Column      Dtype
---  ------      -----
 0   latitude    float64
 1   longitude   float64
 2   brightness  float64
 3   scan        float64
 4   track       float64
 5   acq_date    object
 6   acq_time    int64
 7   satellite   object
 8   instrument  object
 9   confidence  object
 10  version     int64
 11  bright_t31  float64
 12  frp         float64
 13  daynight    object
 14  type        int64
dtypes: float64(7), int64(3), object(5)
memory usage: 547.5+ MB
```

Exploratory Data Analysis: Bushfire Prone Areas

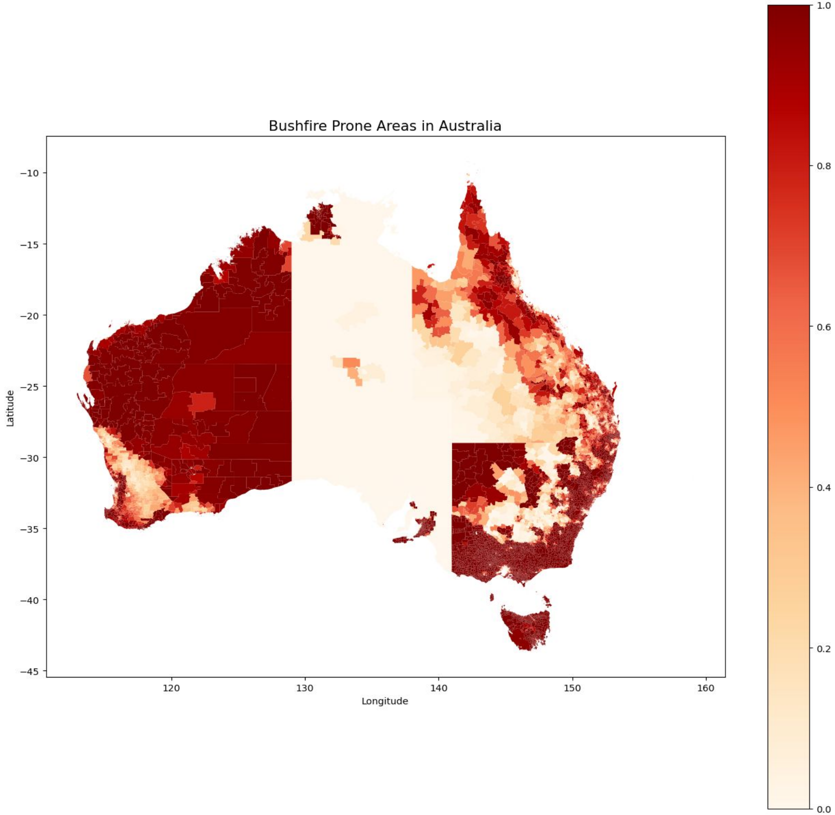
- Used a dataset of coordinates to identify each suburb and state
- Identified a dataset of ‘bushfire prone’ areas

	state_code	state	suburb_code	suburb	area	bf_area	bf_area_pct	cent_lat	cent_lon
12736	3	Queensland	30627	Clintonvale	33.778825	7.413885	0.219483	-28.094423	152.118128
11357	1	New South Wales	13792	Taylors Beach (NSW)	5.361416	5.146824	0.959975	-32.742624	152.068488

- area: Area of the suburb / locality
- bf_area: Area of suburb / locality deemed bushfire prone
- bf_area_pct: Bushfire prone area as a percentage of suburb / locality area
- cent_lat: Centroid (latitude)
- cent_lon: Centroid (longitude)

Exploratory Data Analysis: Bushfire Prone Areas

- Each State/Territory had their own guidance on identifying bushfire prone areas
  - Sudden cut offs e.g. WA and NT
- SA & NT thinks they're pretty safe!
- TAS - everywhere is dangerous!
- QLD and NSW seemed fairly aligned in their rating system



Data Pre-Processing and Visualisations

- Filtered for just NSW and QLD (all of Australia was killing my laptop!)
- Merged the ‘Fires’ data with the ‘Bushfire Prone Areas’ data
  - Led to problems with repeated, missing or very similar suburbs

```
Abbotsford (NSW)
Abbotsford (Qld)
Abercorn
Abercrombie
Abercrombie River
```

- Lots of cleaning, imputing missing suburbs using the nearest available suburb in the other dataset using geopy

Data Pre-Processing and Visualisations

- Identified weather data but it was structured across 3 separate datasets and the columns did not match my Fires dataset

	ClusterID	Datetime	TemperatureMean	TemperatureMax	TemperatureMin
0	100412	1999-12-31 00:00:00+00:00	4.318500	7.0785	3.3785

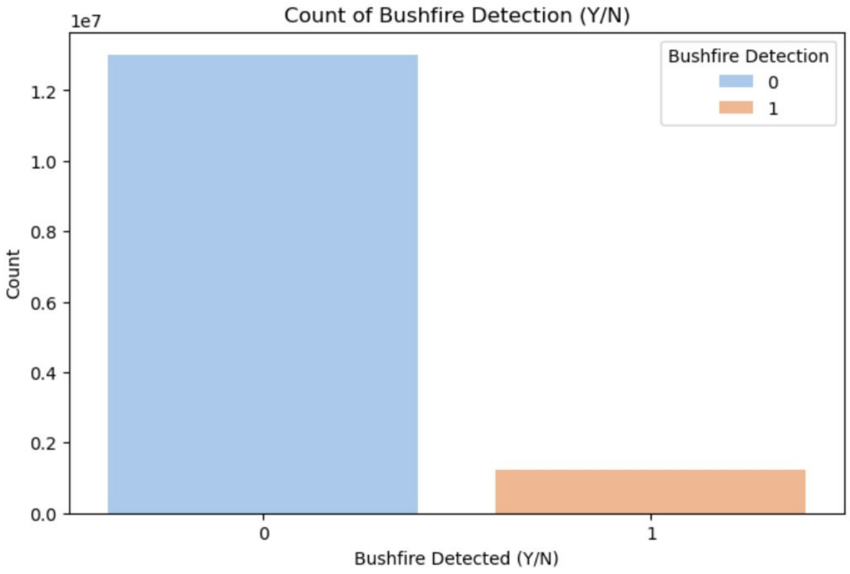
	OfficialNameSuburb	OfficialNameState	ClusterID
0	Adaminaby	NSW	100412.0

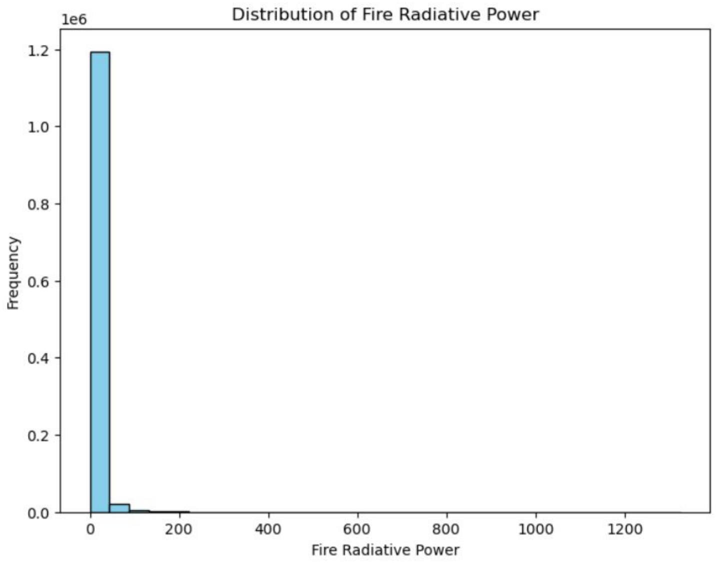
	latitude	longitude	suburb	state	brightness	scan	track	acq_date	acq_time	confidence	bright_t31	frp
0	-37.68088	148.33893	east jindabyne	NSW	337.38	0.33	0.55	2020-01-01	316	1	292.26	29.46

- Time to start cleaning and merging again!
- After 3 weeks I finally had a dataset that had the data I needed to progress

Data Visualisations

- Imbalanced / skewed target datasets





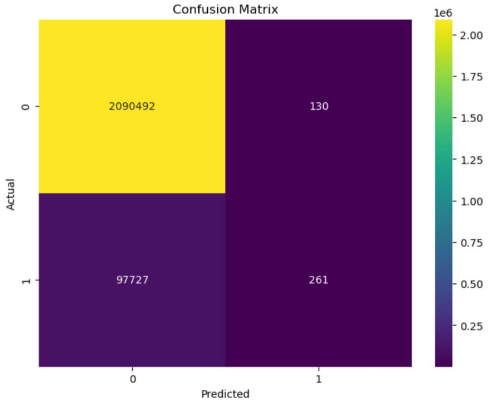
Data Visualisations

Correlation heatmap
- Selected features only
- Temps are collinear
- frp and brightness +vely correlated
- Some correlation between bushfire occurrence and temperatures, but low
- frp has almost no correlation with temperature

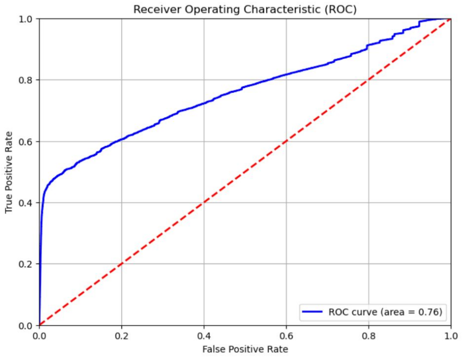
- Overall: Not much to go on

Modelling: To Predict Bushfire Probability

- Logistic Regression



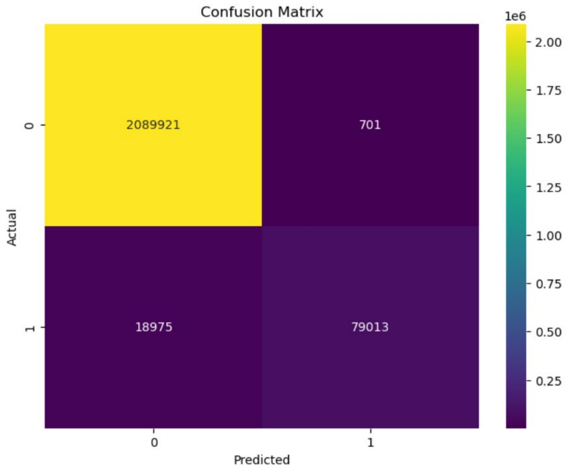
	precision	recall	f1-score	support
0	0.96	1.00	0.98	2090622
1	0.67	0.00	0.01	97988
accuracy			0.96	2188610
macro avg	0.81	0.50	0.49	2188610
weighted avg	0.94	0.96	0.93	2188610



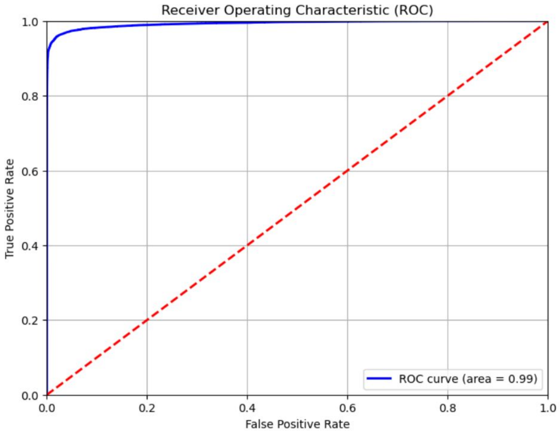
AUC: 0.76

Modelling: To Predict Bushfire Probability

- Random Forest Classifier



	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



AUC: 0.99

Modelling: To Predict Intensity (frp) of a Bushfire

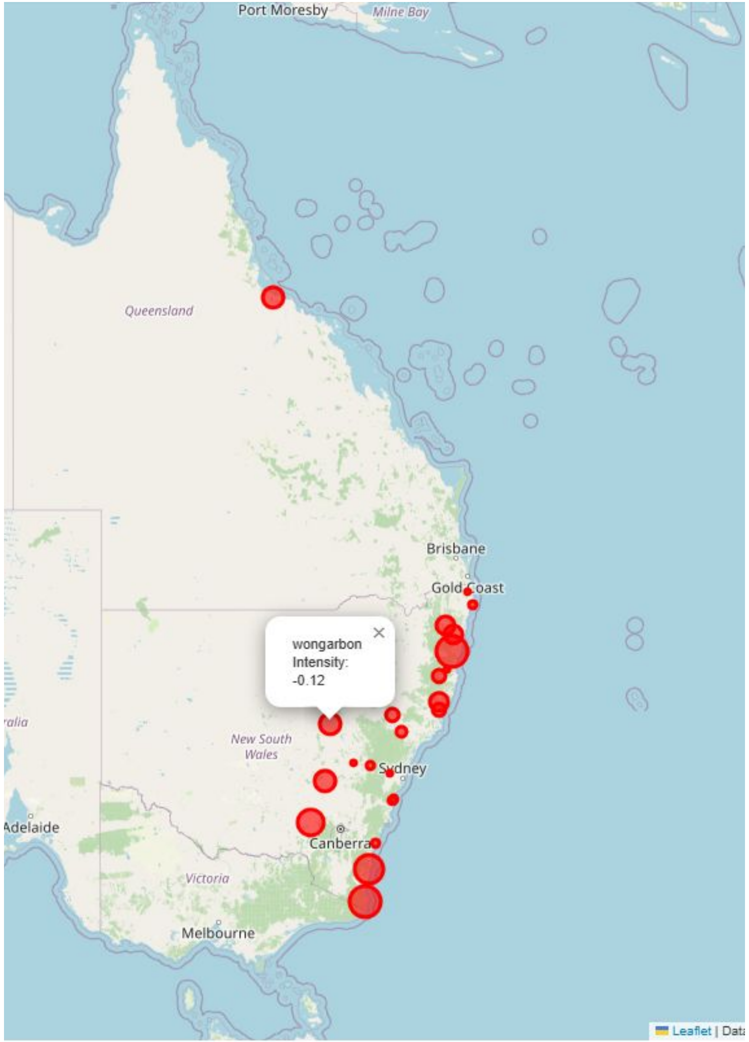
- Linear Regression and Random Forest Regressor

Method	R2	MSE
Linear Regression	Train: 0.28 Test: 0.26	Train:0.70 Test: 0.81
Random Forest Regressor	Train: 0.78 Test: 0.42	Train: 0.22 Test: 0.64

Overfitting?

Outcome

- With the 2 models, I could predict (for a given set of weather conditions) where a bushfire was likely to occur and its intensity



Conclusions

Headlines:

- The model failed to achieve its overall target of 80% accuracy:
    - Target variables (Bushfire Y/N and Fire Radiative Power) were skewed
    - Feature Selection: From the heatmap, no features had a strong relationship with the target features
    - Model Complexity: I couldn’t incorporate complex models due to computational expense and time
    - It was able to predict bushfire occurrence Y/N accurately
    - Performed poorly on the intensity data

Other Points to Consider / Next Steps:

- Increase the number of features e.g. relative humidity and wind speed
- Invest in more computing power (consider Google Colab for future computational expensive work)
- Research to see if there’s any data for the suburbs where a fire was not detected in the timeframe I selected
- Try more complex modelling techniques and hyperparameter tuning
- Convert risk to categorical data (e.g. low, medium, high) and display on map
- Expand the model to run Australia-wide

Finally:

- The model is not there yet but shows possibility given further investment in computing power and data acquisition.

References

Fires data:
This was provided to me as a result of a direct request to NASA for a download of their FIRMS Archive

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

Questions?