NASA Satellite Data of Australia Wildfires

-- Zesheng Jia

Introduction

This data analysis report is based on the dataset from NASA Satellite Data of Australian wildfires (Wong, 2022). As we know, wildfires occur frequently in Australia during the summer. However, most of the time, these fires begin in remote areas. When they become wildfires, the damage can be significant. For example, the three deadliest blazes in Australia caused numerous deaths and extensive property damage. The Black Saturday in Victoria in 2009 resulted in 173 fatalities and 2,000 lost homes. The Ash Wednesday fires in Victoria and South Australia in 1983 claimed 75 lives and destroyed nearly 1,900 homes. The Black Friday fires in Victoria in 1939 led to 71 deaths and the destruction of 650 houses (Wikipedia contributors, 2023). With the help of satellites, these fires can no longer remain hidden. Even in the most remote corners of the globe, sensors in the sky can detect their heat signatures. In our case, we aim to use this dataset to discover the link between the early stages of satellite information and its potential numerical Fire Radiative Power, which can indicate the future damage caused by wildfires.

Data Exploration

In our dataset, we have 183593 instances and 14 features. Here we write down the description for each feature before we go into the data exploration section.

Attribute	Short Description	Long Description
Latitude	Latitude	Center of nominal 375 m fire pixel
Longitude	Longitude	Center of nominal 375 m fire pixel
Bright_ti4	Brightness temperature I-4	VIIRS I-4 channel brightness temperature of the fire pixel measured in Kelvin.

Scan	Along Scan pixel size	The algorithm produces approximately 375 m pixels at nadir. Scan and track reflect actual pixel size.
Track	Along Track pixel size	The algorithm produces approximately 375 m pixels at nadir. Scan and track reflect actual pixel size.
Acq_Date	Acquisition Date	Date of VIIRS acquisition.
Acq_Time	Acquisition Time	Time of acquisition/overpass of the satellite (in UTC).
Satellite	Satellite	N= Suomi National Polar-orbiting Partnership (Suomi NPP), 1=NOAA-20 (designated JPSS-1 prior to launch)
Confidence	Confidence	This value is based on a collection of intermediate algorithm quantities used in the detection process. It is intended to help users gauge the quality of individual hotspot/fire pixels. Confidence values are set to low, nominal and high. Low confidence daytime fire pixels are typically associated with areas of sun glint and lower relative temperature anomaly (<15K) in the mid-infrared channel I4. Nominal confidence pixels are those free of potential sun glint contamination during the day and marked by strong (>15K) temperature anomaly in either day or nighttime data. High confidence fire pixels are associated with day or nighttime saturated pixels. Please note: Low confidence nighttime pixels occur only over the geographic area extending from 11deg E to 110 deg W and 7 deg N to 55 deg S. This area describes the region of influence of the South Atlantic Magnetic Anomaly which can cause spurious brightness temperatures in the mid-infrared channel I4 leading to potential false positive alarms. **These have been removed from the NRT data distributed by FIRMS.**
Instrument	Instrument	NASA Satellite Instruments, MODIS or VIIRS
Version	Version (Collection and source)	Version identifies the collection (e.g. VIIRS Collection 1) and source of data processing: Near Real-Time (NRT suffix added to collection) or Standard Processing (collection only). "1.0NRT" - Collection 1 NRT processing. "1.0" - Collection 1 Standard processing
Bright_ti5	Brightness temperature I-5	I-5 Channel brightness temperature of the fire pixel measured in Kelvin.

FRP	Fire Radiative Power	FRP depicts the pixel-integrated fire radiative power in MW (megawatts). FRP depicts the pixel-integrated fire radiative power in MW (megawatts). Given the unique spatial and spectral resolution of the data, the VIIRS 375 m fire detection algorithm was customized and tuned in order to optimize its response over small fires while balancing the occurrence of false alarms. Frequent saturation of the mid-infrared I4 channel (3.55-3.93 µm) driving the detection of active fires requires additional tests and procedures to avoid pixel classification errors. As a result, sub-pixel fire characterization (e.g., fire radiative power [FRP] retrieval) is only viable across small and/or low-intensity fires. Systematic FRP retrievals are based on a hybrid approach combining 375 and 750 m data. In fact, starting in 2015 the algorithm incorporated additional VIIRS channel M13 (3.973-4.128 µm) 750 m data in both aggregated and unaggregated format.
Type	Type of fire	Different types of fire that indicates the scale and potential damage.

Table 1 Feature Descriptions

From table 1, we can see that this dataset contains spatial information (Longitude and Latitude), satellite information (scan, track, satellite, instrument, and version), fire information (bright_ti4, bright_ti5, frp, confidence), and time series information (acq_data, acq_time). In our case, we want to predict the FRP (Fire Radiative Power) to find out whether a fire is potentially destructive.

```
    latitude         longitude       bright_ti4         scan            track             acq_date        acq_time      satellite
 Min.   :-43.080   Min.   :114.1   Min.   :208.0   Min.   :0.3200   Min.   :0.3600   2019-09-09:  7450   Min.   : 229.0   N:184778
 1st Qu.:-28.163   1st Qu.:131.4   1st Qu.:310.9   1st Qu.:0.3900   1st Qu.:0.3800   2019-09-10:  6882   1st Qu.: 415.0
 Median :-16.207   Median :141.6   Median :335.9   Median :0.4400   Median :0.4500   2019-09-06:  6817   Median : 511.0
 Mean   :-19.685   Mean   :139.7   Mean   :331.6   Mean   :0.4629   Mean   :0.4875   2019-09-07:  6691   Mean   : 899.2
 3rd Qu.:-13.544   3rd Qu.:150.2   3rd Qu.:346.7   3rd Qu.:0.5200   3rd Qu.:0.5700   2019-09-08:  6186   3rd Qu.:1528.0
 Max.   : -9.394   Max.   :153.5   Max.   :367.0   Max.   :0.8000   Max.   :0.7800   2019-09-14:  4838   Max.   :1827.0
                                                                                     (Other)   :145914
 instrument     confidence    version    bright_ti5         frp               type
 VIIRS:184778   h: 13597   Min.   :1   Min.   :253.4   Min.   :  0.100   Min.   :0.00000
                l: 13256   1st Qu.:1   1st Qu.:288.2   1st Qu.:  1.800   1st Qu.:0.00000
                n:157925   Median :1   Median :296.4   Median :  4.500   Median :0.00000
                           Mean   :1   Mean   :297.6   Mean   :  8.632   Mean   :0.06489
                           3rd Qu.:1   3rd Qu.:306.8   3rd Qu.:  9.000   3rd Qu.:0.00000
                           Max.   :1   Max.   :380.0   Max.   :595.200   Max.   :3.00000
```

Table 2 Dataset summary

And from table 2, we can see that there are three features: *instrument*, *version, and satellite*, only have 1 type of input. Because this dataset only contains 1 kind of NASA satellites. For the other satellites, we need to discover different patterns of those specifically. Here, we only use the dataset that contains the VIIRS satellite. Hence, we delete those three features from our dataset. In the meantime, we change feature, acq_date to a date format, then convert to numerical value for further predicting.

Then, we checked the dataset, and there are no missing values in every predictor as shown in Table 3.

latitude	longitude	bright_ti4	scan	track	acq_date	acq_time	confidence	version	bright_ti5	frp	type
0	0	0	0	0	0	0	0	0	0	0	0

Table 3 Missing values report

Here, we plot the pairplot that can show the relation between each predictor. And we can also use it to find the outliers.

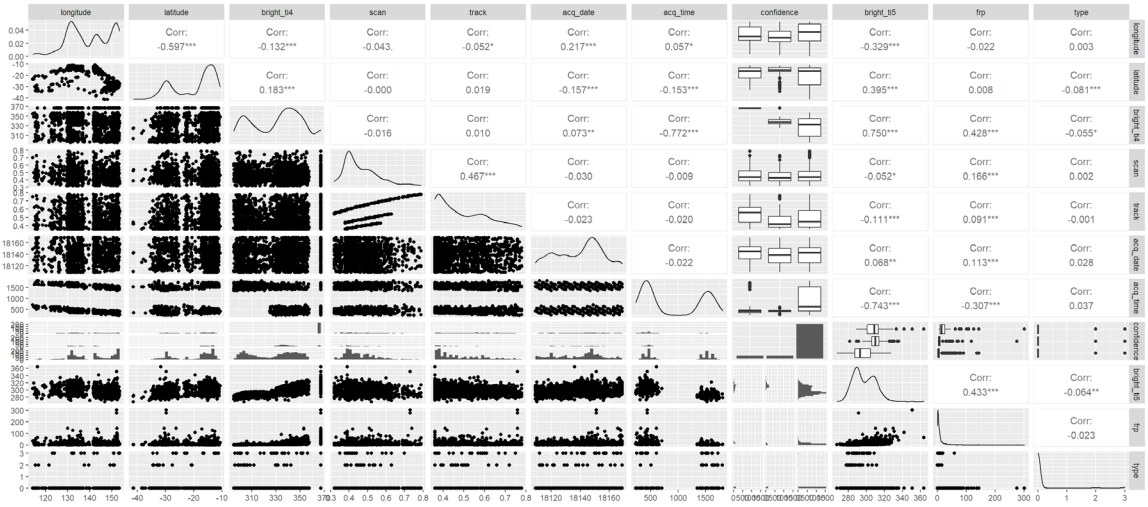

Figure 1 Pairplots of our dataset

From Figure 1, we can see that there are several predictors with non-ideal data distributions. "Scan," "track," "FRP," "bright_ti5", and "type" have long tails on the right side, indicating many outliers on that end. To bring these features closer to the center, we can perform a log transformation. Additionally, "bright_ti5" has a positive correlation with our response variable, "FRP," suggesting it may be an important

predictor for our predictions. Regarding spatial information, the scatter plot of longitude versus latitude provides a rough representation of Australia's shape, and we can observe that most fires occurred on the contour of Australia. More details on this will be discussed in the Data Visualization section. Concerning the "Confidence" predictor, most predictors have the same level of confidence except for "bright_ti4." Further investigation is needed to understand the relationship between these two features. For the time series predictors, "acq_time" and "acq_date," we can observe that "acq_time" points are spread in two directions, corresponding to Day and Night when the satellite passes over Australia. As the satellite is not solely dedicated to Australia, this explains the two directions.

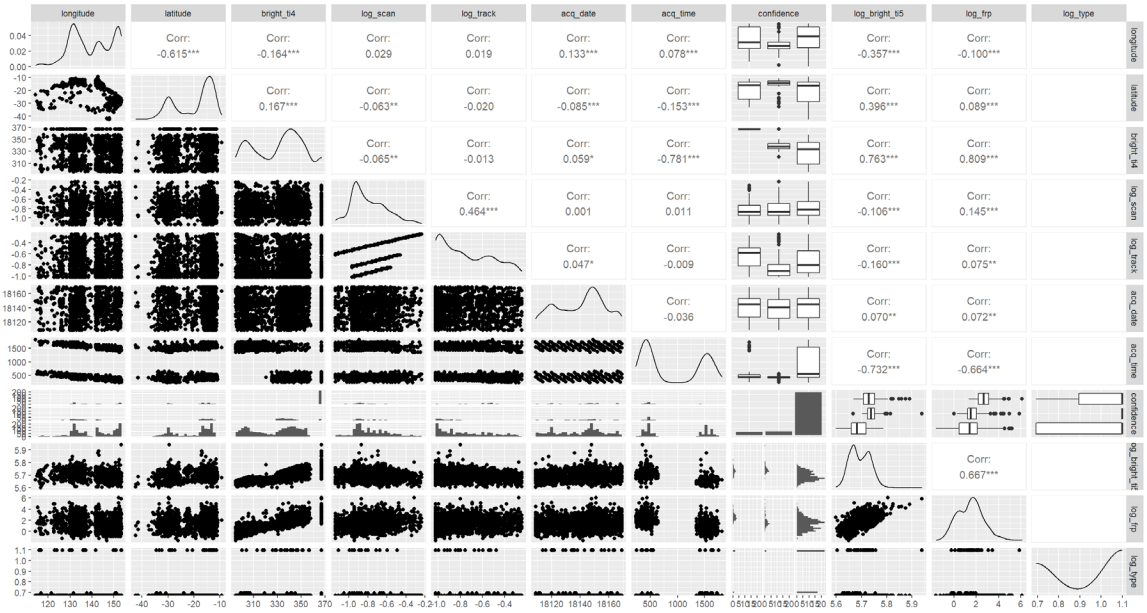

Figure 2 Log transformed features

We can see that by applying the log transformation, the "log_frp" shows a normal distribution shape as the most of points are located in the center. And the "log_bright_ti5" shows a very strong positive correlation with "log_frp". The shapes of "log_scan" and "log_track" are not improved by much. We keep that in this way before training the model. In case of those two features are not very important during the prediction.

Data Visualization

First, let's investigate the Spatial information of Australia wildfires.

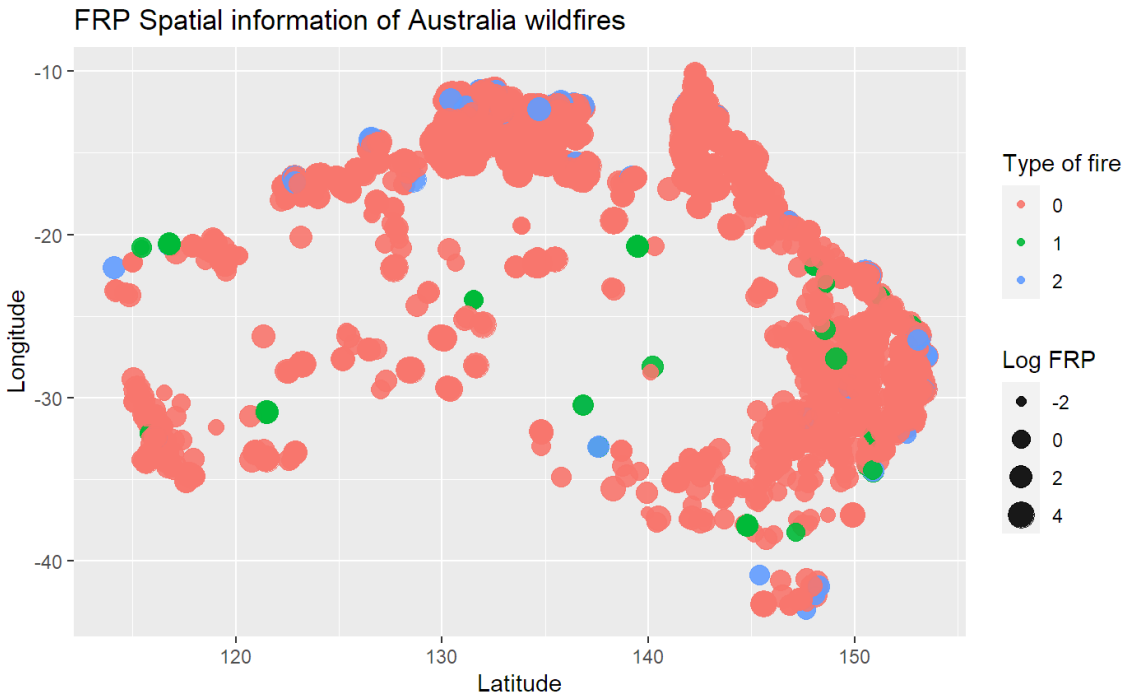

Figure 3 Spatial representation of wildfires

From Figure 3, we can see that most of the fires that have occurred in Australia are classified as type 0. Type 1 and type 2 fires tend to have much larger logarithmic Fire Radiative Power (FRP) compared to type 0 fires. Most type 2 fires occur along the contour of Australia, while the central region has more type 0 and type 1 fires.

To provide a clearer view of the previous plot, we created a new plot where only the border lines of the points are displayed instead of solid circles. As shown in Figure 4 (on the next page), many fires occur in the northern part of Australia, and most of them are classified as type 0 or type 2. In the central region of Australia, fires tend to have very high FRP, indicating that caution should be exercised when fires occur in this area. Type 1 fires often occur in the west, east, and middle regions of Australia, and there are

relatively few of these fires in the north.

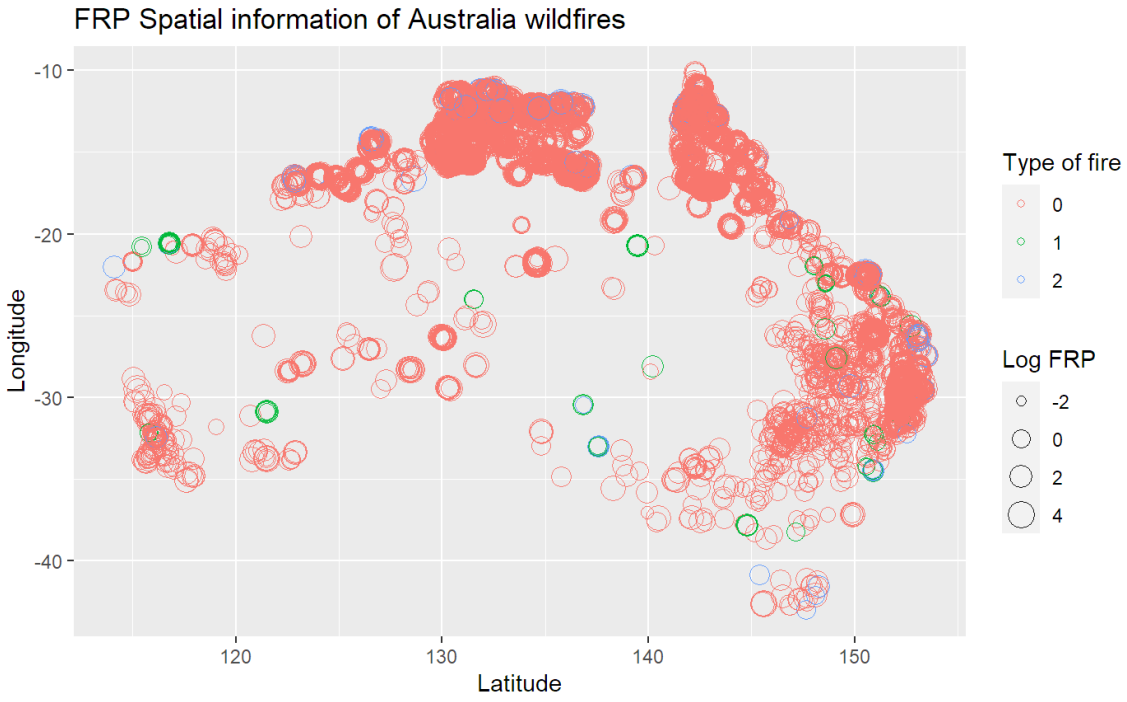

Figure 4 Border plot of previous plot

Now, we move on to the most correlated two predictors, log_bright_ti5 and log_FRP.

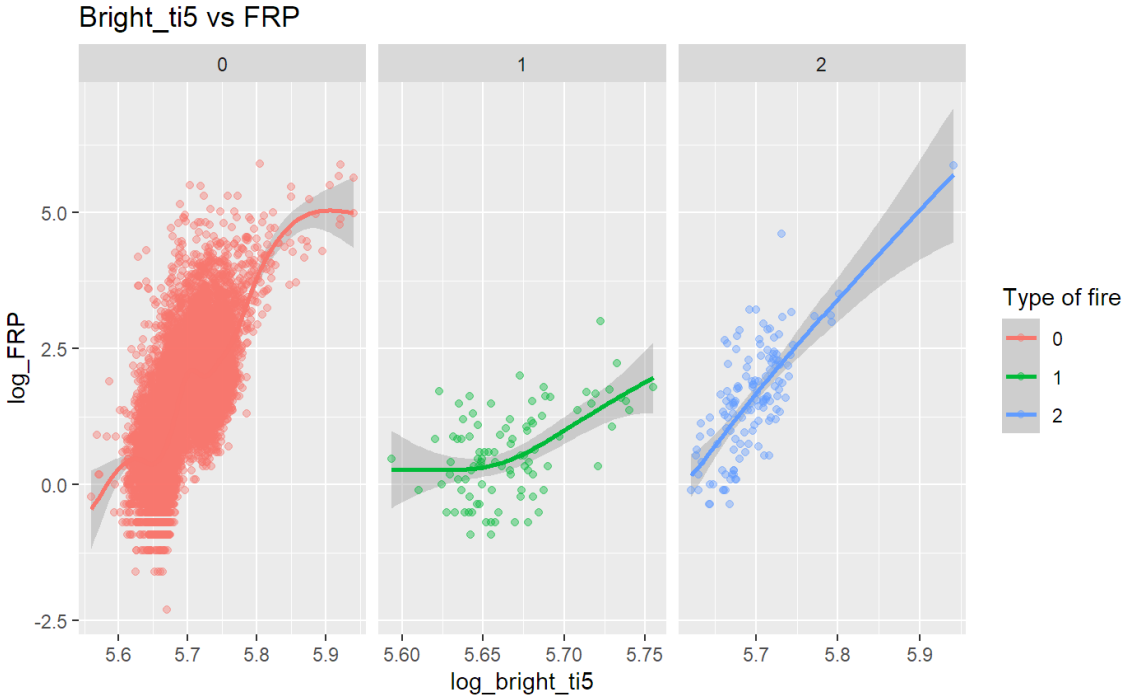

Figure 5 Bright_ti5 vs FRP

We separate three different types of fire and plot the log transformation of bright_ti5 on the x-axis, and log transformation of Fire radiative power on the y-axis. We can see that there are a very strong and intense positive correlation between those two features. When the bright_ti5 increases, the FRP will also increase. And different type of fires could have different ratio of increasing speed.

For time series information, in order to show the details of different acq_date, we convert all dates into month and day as two separated columns. And we found that this dataset only contains the August and September satellite information. Since those two months have the most devastated damage.

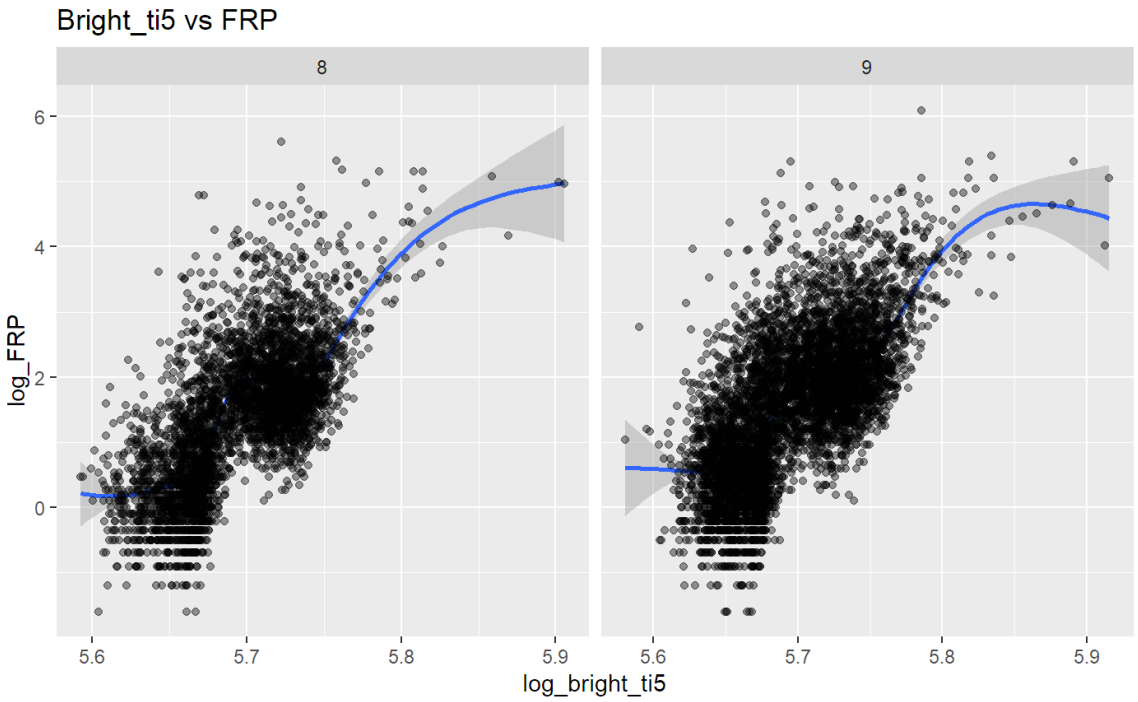

Figure 6 Bright_ti5 vs FRP of August and September

From Figure 6, we can see that the relation between bright_ti5 and FRP in two different months are similar. Hence, we will focus on days for now on. From Figure 7 (in the next page), there are more fires that happen between 6th to 15th in each month. It is worth to be care of those days for the fire departments.

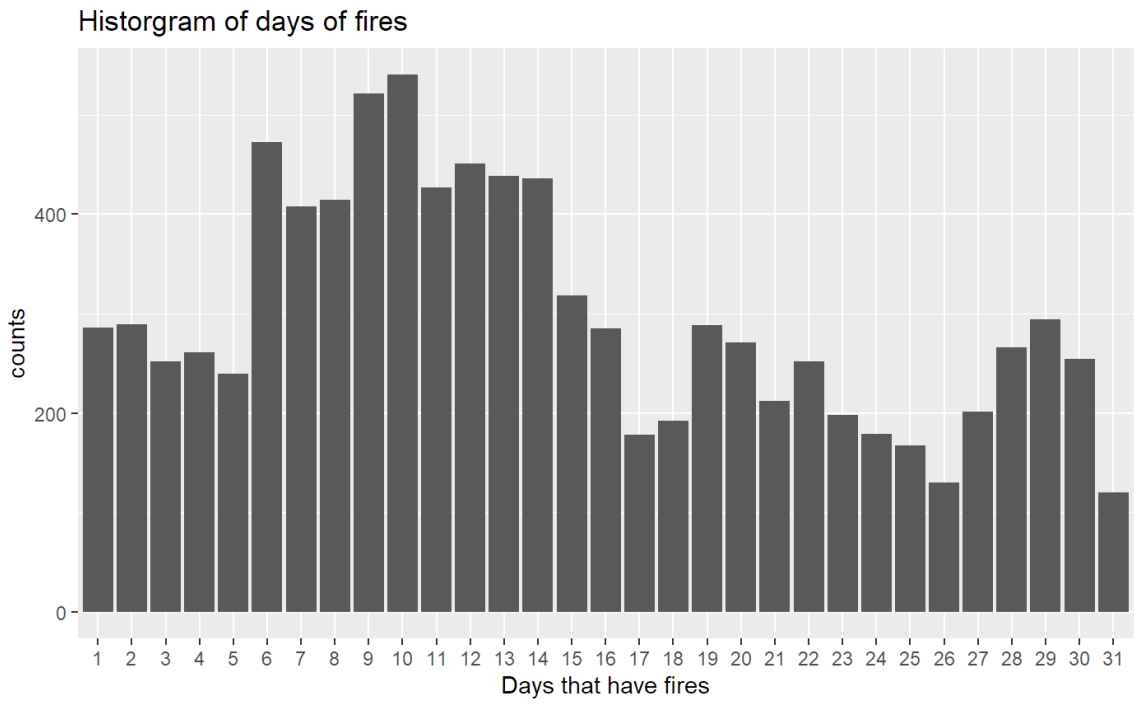

Figure 7 Histogram of days of fires

Based on the finding, we also plot the Barplot of days vs log_FRP, as shown in Figure 8.

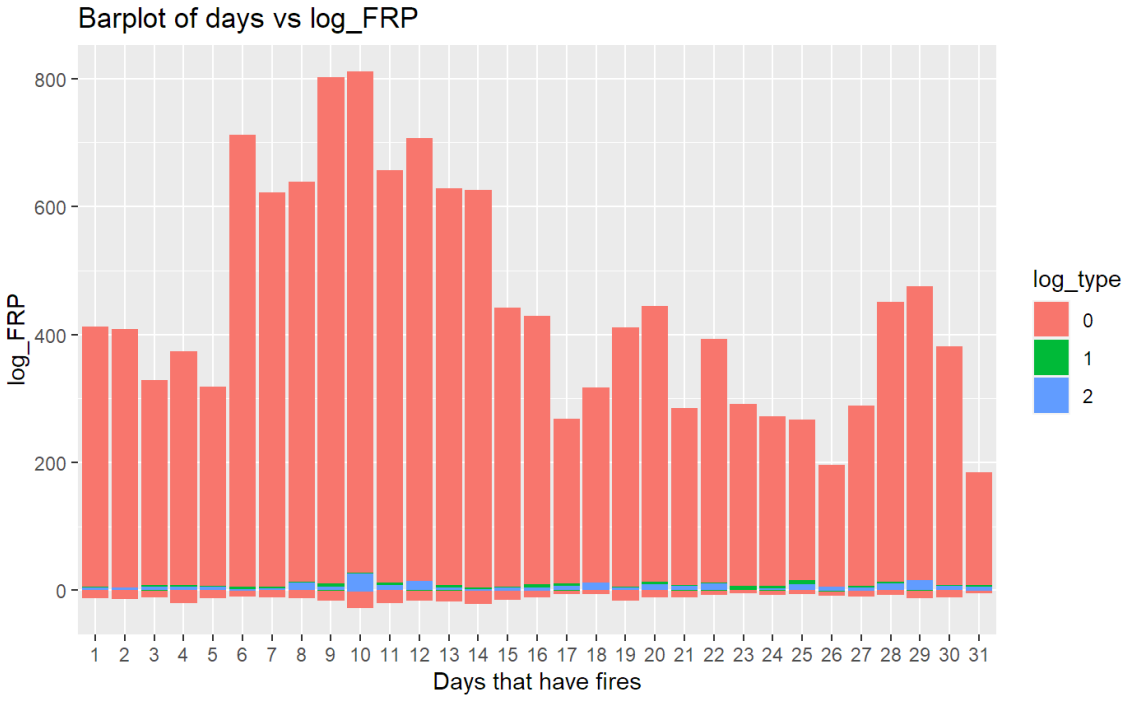

Figure 8 Barplot of days vs log_FPR

From Figure 8, in each day in a month, there are very few type 1 and type 2 fires. And combine with the data exploration section's finding, we assume type of fires could be an important predictor. We also try to discover the relation between FRP and confidence, so we create Figure 9.

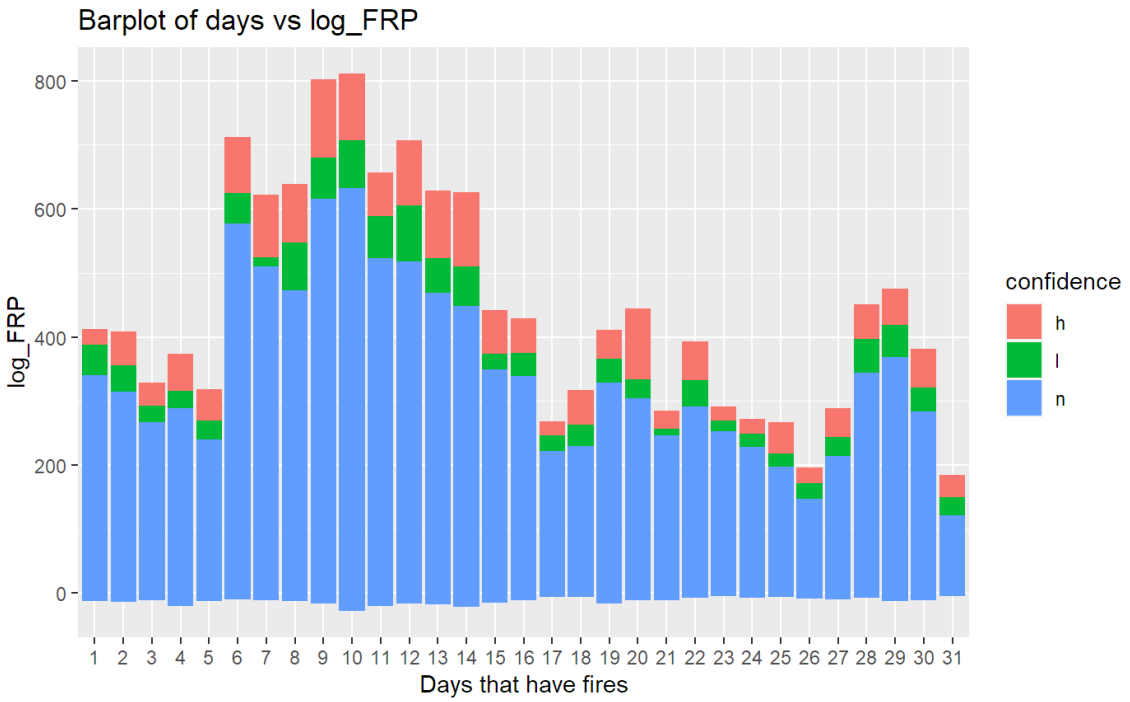

Figure 9 Days vs FRP fill by confidence

Here, we restate the definition of confidence in our dataset.

"Confidence values are set to low, nominal and high. Low confidence daytime fire pixels are typically associated with areas of sun glint and lower relative temperature anomaly (<15K) in the mid-infrared channel I4. Nominal confidence pixels are those free of potential sun glint contamination during the day and marked by strong (>15K) temperature anomaly in either day or nighttime data. High confidence fire pixels are associated with day or nighttime saturated pixels."

We can observe that the majority of fires that occur each day are classified as nominal confidence, indicating that they are free of potential sun glint contamination during the day. Only a very small number of fires are classified as high confidence or low confidence. To investigate the spatial relationship between the fires, we created Figure 10 (on the following page), which shows that fires with nominal confidence are distributed throughout the entire country.

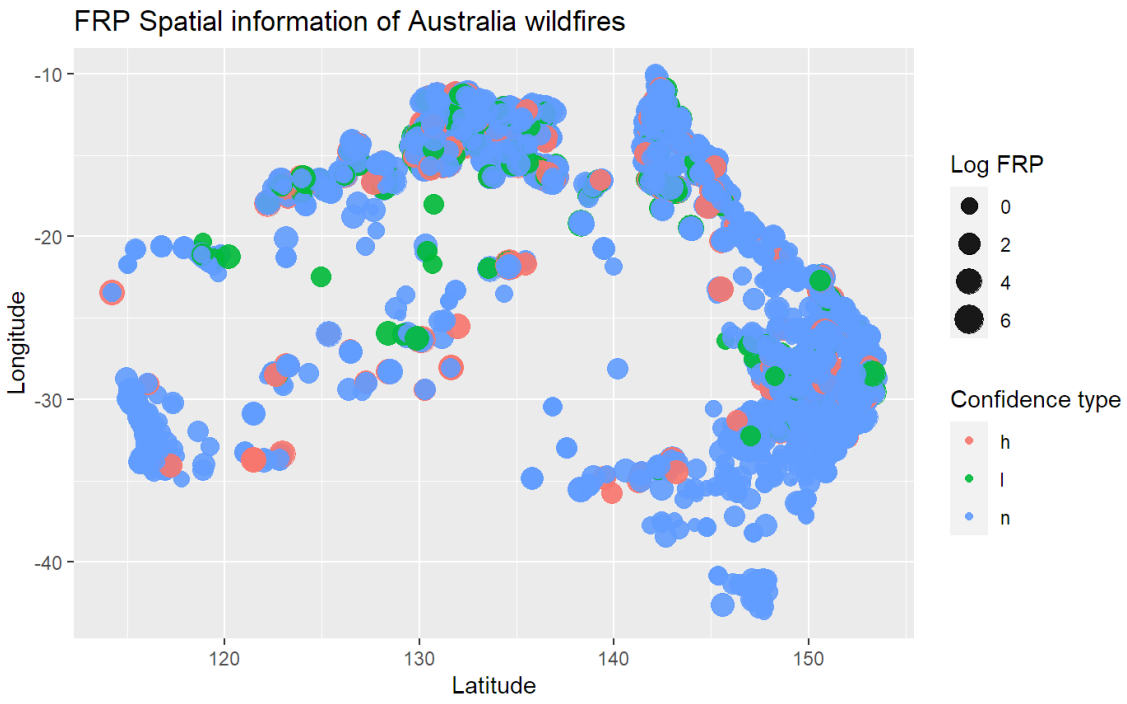

Figure 10 Confidence type on Spatial Map

In case to find out the relation of confidence between bright_ti5 and FRP, we create Figure 11.

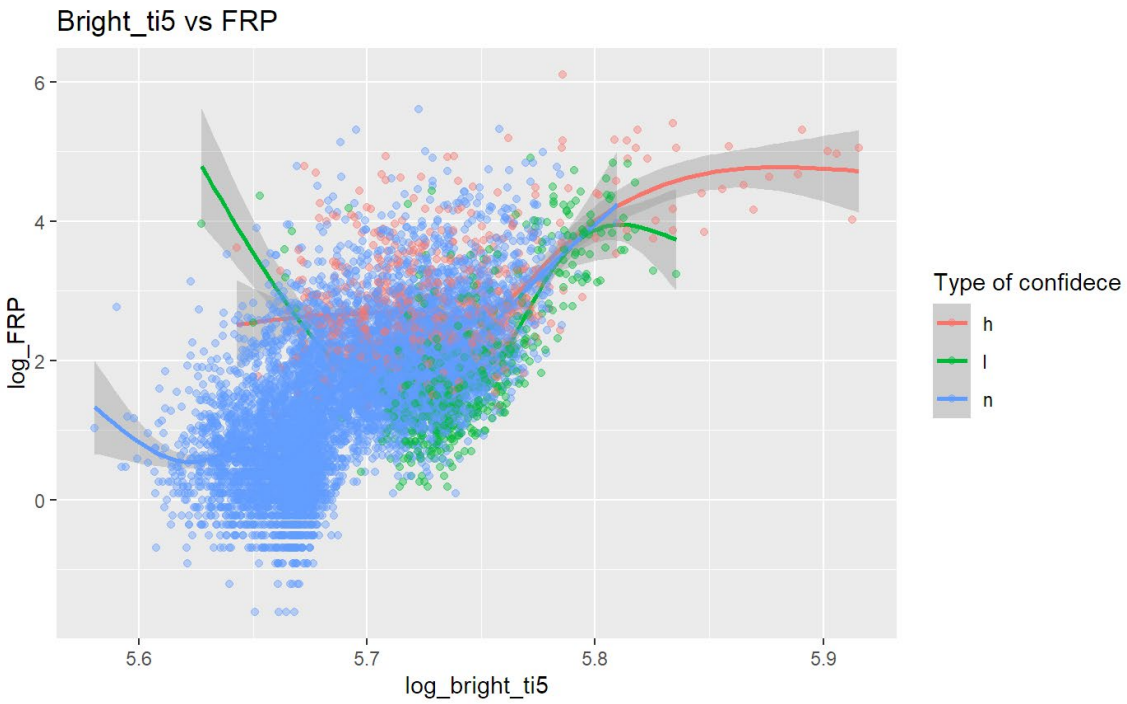

Figure 11 Confidence on Bright_ti5 vs log FRP

We can see that the high confidence fires have higher FRP and Bright_ti5 than the other two types of confidence in average speaking. Hence, the type of confidence also maybe an important predictor of the Fire Radiative Power.

For predictor “bright_ti4”, we can see there are a few outliers at the left hand in Figure 12. And we choose to remove every instance that is less than 285. And we found there are only 3 instances are deleted.

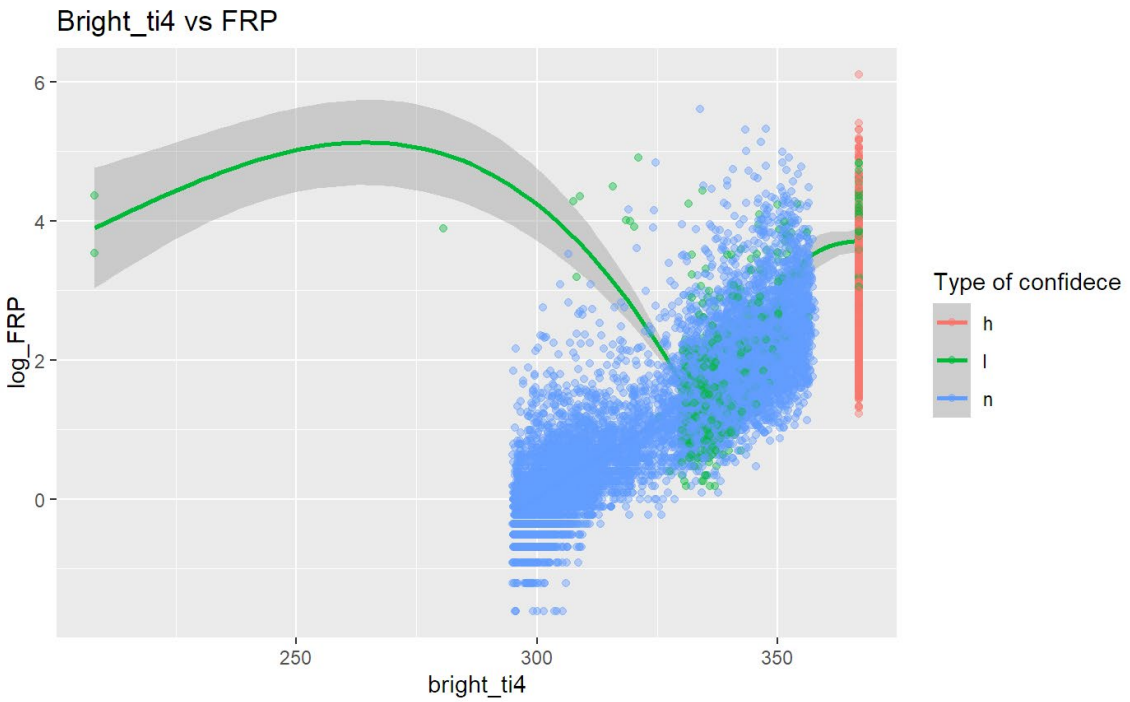

Figure 12 Log bright_ti4 and bright_ti5

After deleting those outliers, we have Figure 13.

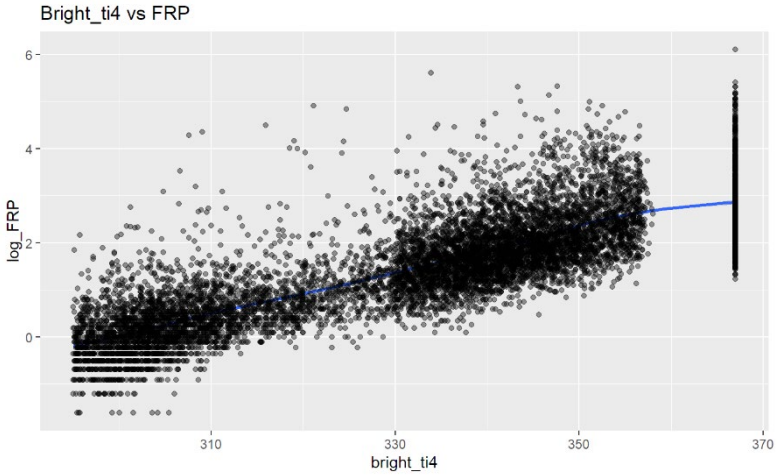

Figure 13 After deleting left outliers

We can see that birght_ti4 have a very strong positive correlation of birght_ti5. Hence, we don't do the visualization of bright_ti4 again. And about the right hand outliers, we can see that at same range of bright_ti4, it has a lot of different values of bright_ti5. It may cause by some error or range of recording limit. We ignore that for now, since there is nothing we can do about it.

The last part of Data Visualization is Scan vs Track. Based on the dataset information, scan and track together are related to the area of satellite visible view block.

Scan	Along Scan pixel size	The algorithm produces approximately 375 m pixels at nadir. Scan and track reflect actual pixel size.
Track	Along Track pixel size	The algorithm produces approximately 375 m pixels at nadir. Scan and track reflect actual pixel size.

Table 4 Scan and Track

Here, we create Figure 13, we can see although the area of those visible view blocks are different, they have all range of FRP and confidence types. It could mean that those two features are not very important.

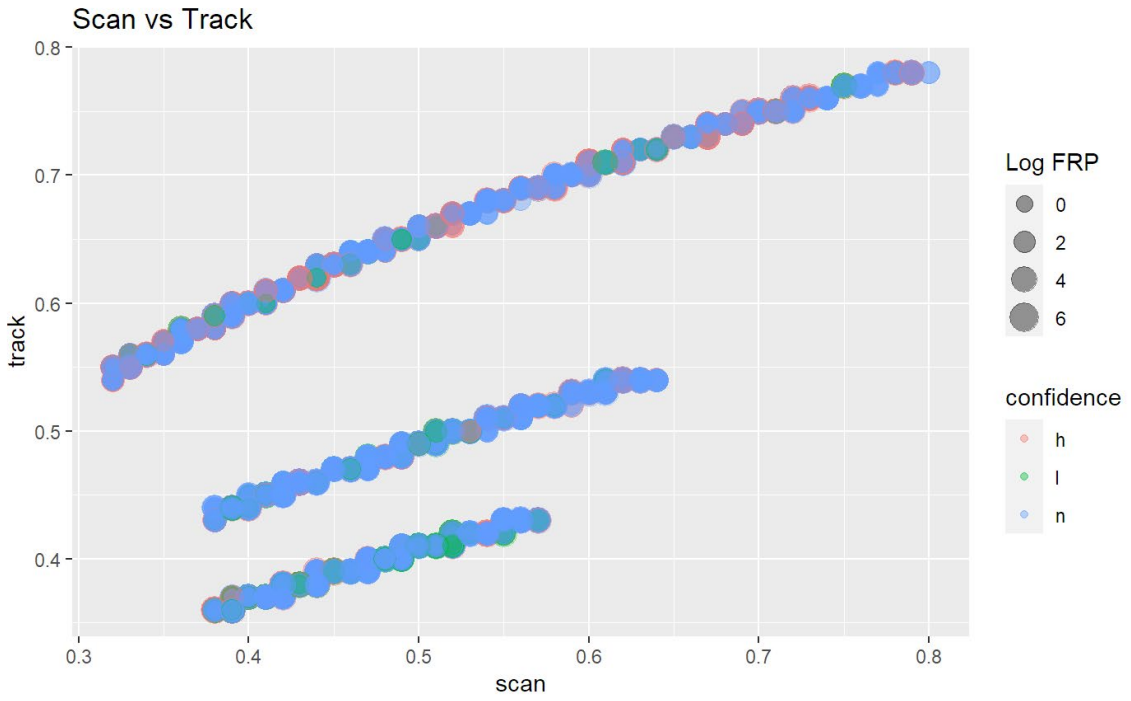

Figure 14 Scan vs Track

Based on the discoveries made in the Data Exploration and Data Visualization sections, we can conclude that certain predictors can provide more information for predicting Fire Radiative Power. These predictors include bright_ti5, fire_type, confidence, and acq_day. Fires with higher values of bright_ti5 tend to have higher FRP. Additionally, the

majority of fires are classified as type 0, while types 1 and 2 are rare but occur daily. Furthermore, fires with higher levels of confidence tend to have higher FRP. Using these findings, we can now proceed to create our models.

Predictive Models

We tried 4 different models that are linear model, decision tree, random forest, and lightgbm model. First, we fitted the linear model and plot its Autoplot.

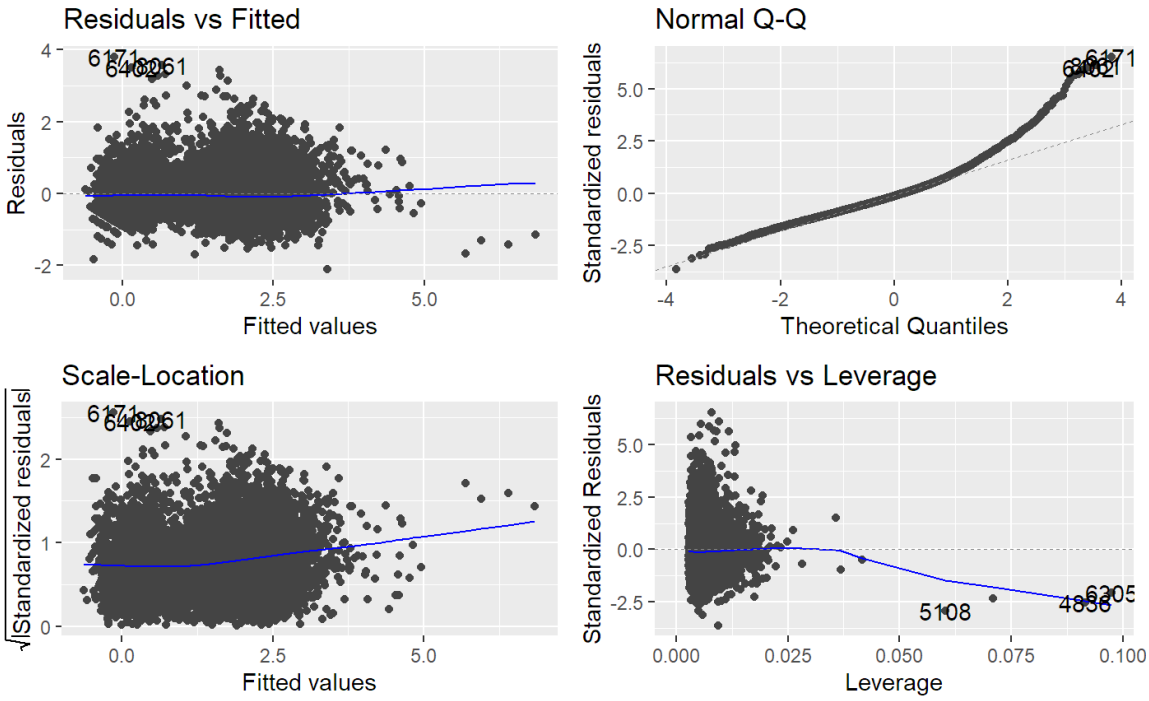

Figure 15 Autoplot of Linear model

From Figure 15, it is evident that our fitted values and residuals do not follow a normal distribution, with a significant right tail visible in the Normal Q-Q plot. Additionally, there is a considerable variance in the residuals observed across nearly all fitted values. The Residuals vs Leverage graph reveals the presence of several leverage points, which may have a significant impact on the final predictions. Similarly, the scale-location plot indicates that large variances still exist, which is not desirable.

In order to solve this issue, we tried GAM and quadratic terms, which doesn't have a significant improve than the normal linear model. Their Root mean square error is around 11.4209.

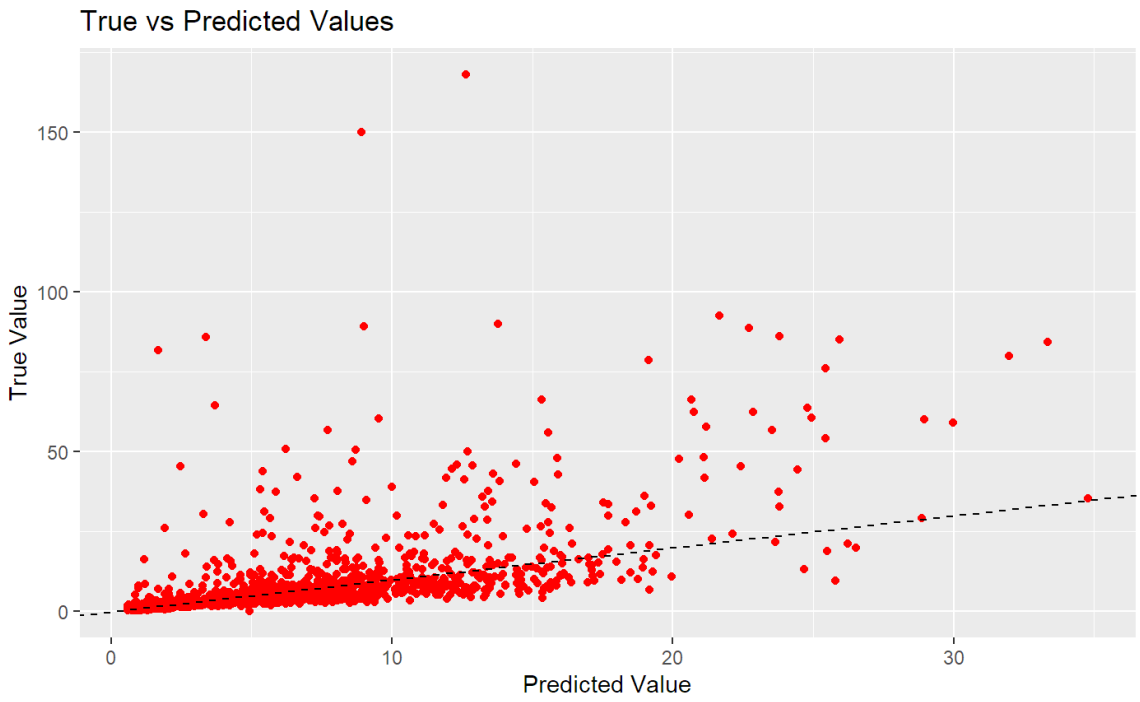

Figure 16 Predicted value vs True value of linear model

As the Figure 16 shows, although most of points lie in the y = x line as the ground truth, there are still a lot of points are far away this line. It means some of true FRP is much larger than the prediction. It fits the autoplot' result we showed before.

In order to improve the performance, we fitted a decision tree model with largest depth = 15, we have the following tree split as Figure 17 (in the next page) shows. We can see that the most informative split at the top is bright_ti4 < 328. From the data exploration part, we know that the birght_ti4 numerical range is from 208 to 367. The second split is bright_ti4 < 307 and 328 < birght_ti4 < 346. When we go down of our decision tree, we can see that there are splits about "Scan", "acq_time", and "bright_ti5", which fits our assumption in Data Visualization section. In this model, we have test set RMSE as 11.8. That is worse than the linear model, but it gives us some explanation of important features.

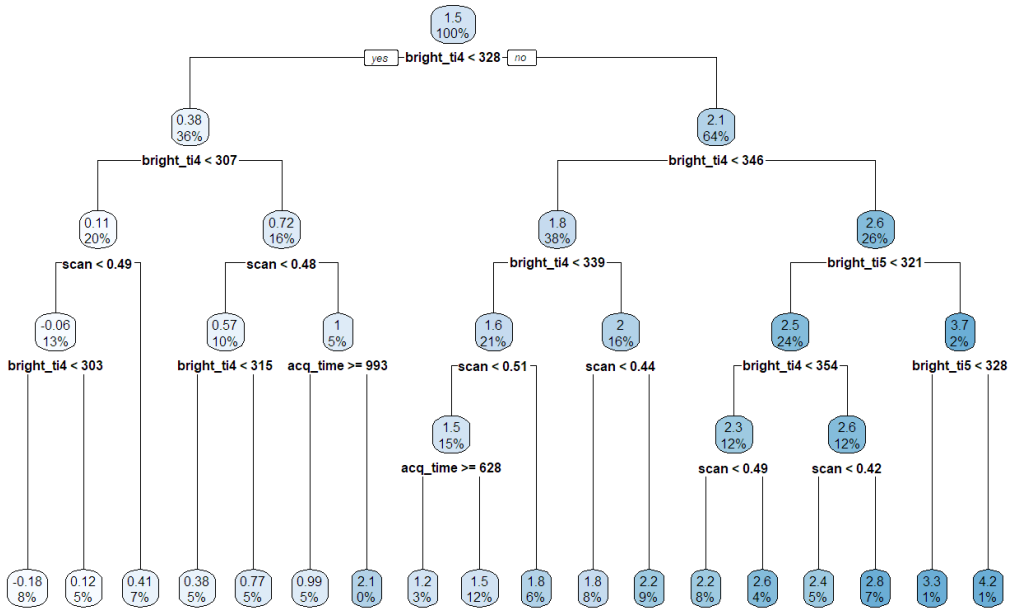

Figure 17 Decision Tree model split

By doing the same True vs False plot, we have Figure 18.

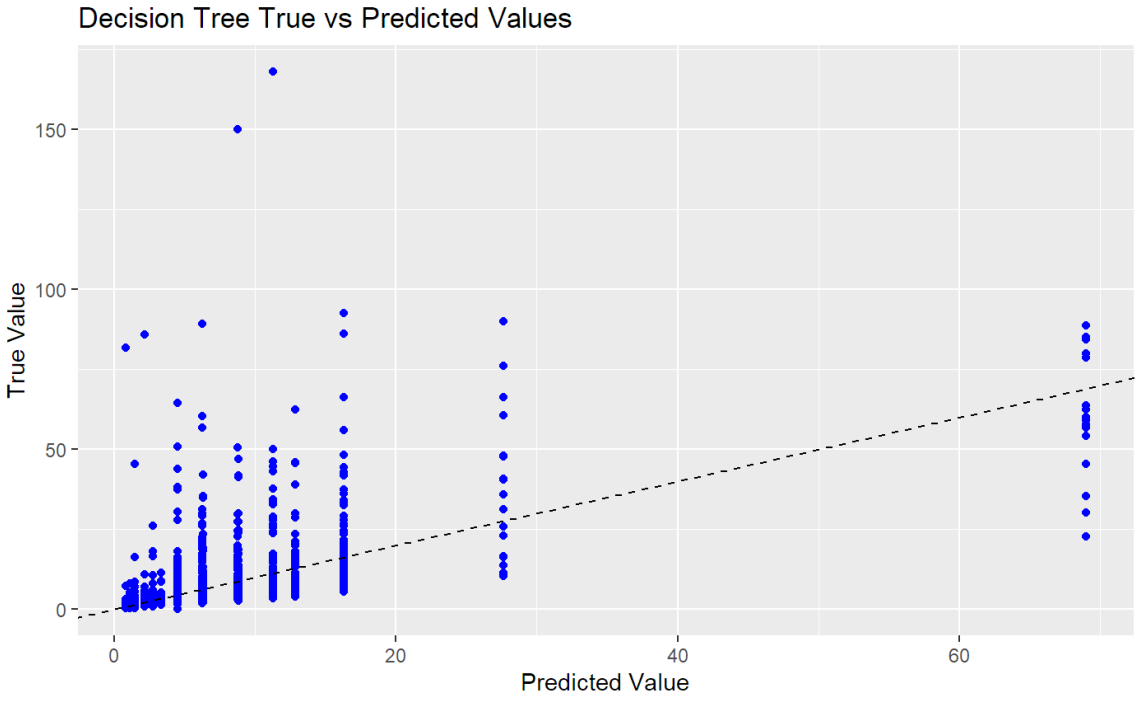

Figure 18 Decision Tree True vs Prediction

In Figure 18, we can see that despite the drawbacks of the decision tree model, which can only predict a limited number of outcomes, most of the points lie on the y=x line, and there are not too many points with large variances. In the meantime, we can see that the decision tree's highest prediction range is larger than the linear models.

Hence, we then fitted a Random Forest model to see whether it can improve its performance. This model has test set Root mean square error as 8.442, which is better than the previous two models. As shown in Figure 19, Random Forest's prediction has

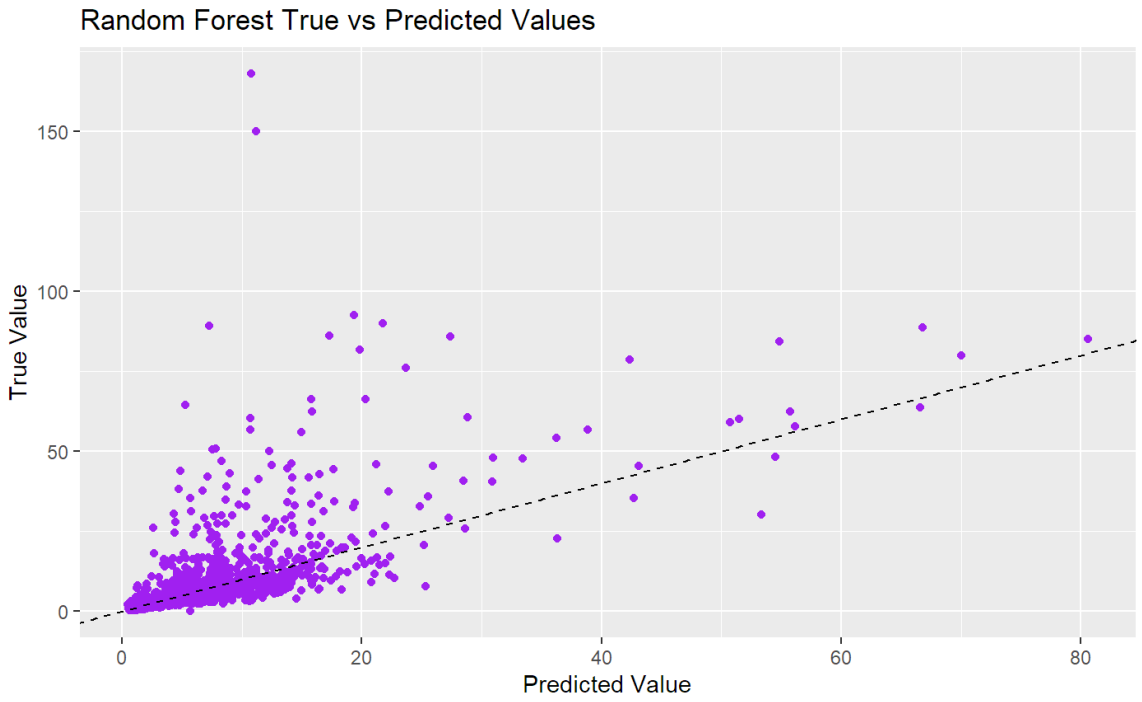

Figure 19 Random Forest True vs Pred

larger prediction range, and much less variance. Most of points lie on the y=x line. And in our case, there are around 3 outliers whose prediction is small but true value is very large around 150 plus. We could say that they may cause by very extreme case of fire or just recording errors. Since Random Forest is not explainable. We then plot the feature importance to see whether this model also fit our assumptions.

As Figure 20 shown, Random Forest take log_bright_ti4, bright_ti4, acq_time, bright_ti5, log_bright_ti5, and acq_date as the most important features.

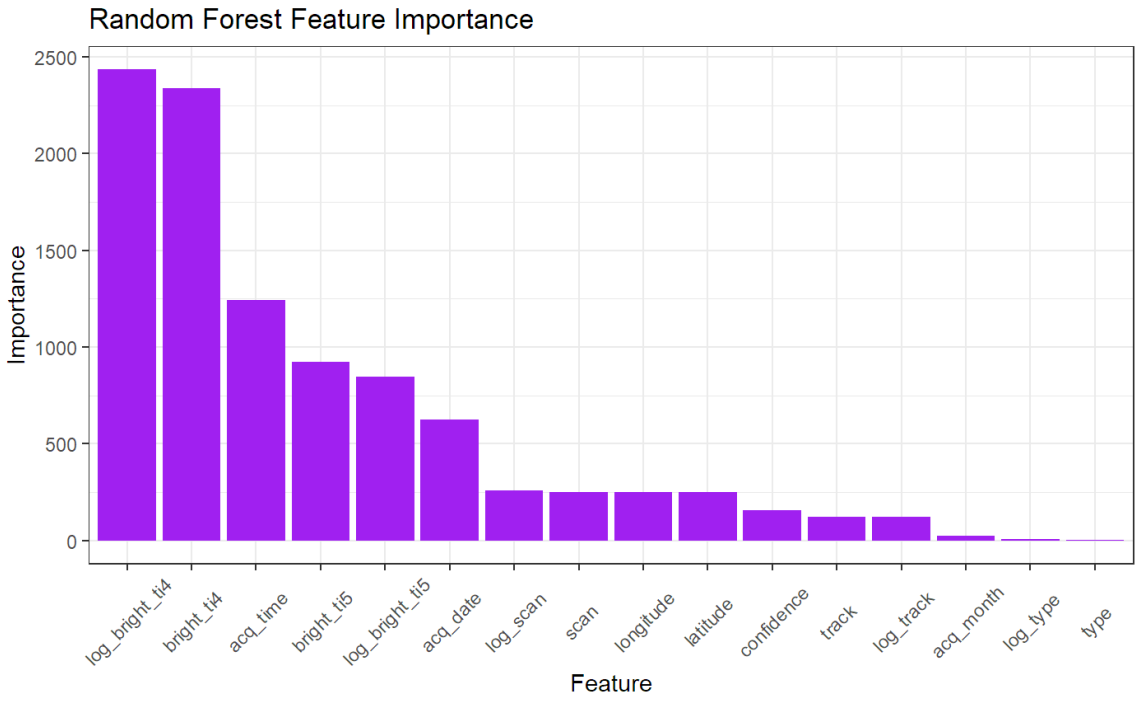

Figure 20 Random Forest feature importance

Since the nature of Random Forest only take a few predictors to create a tree, we would say the log_bright_ti4/5 and bright_ti4/5 are similar feature. That could be the reason why the previous models don't use those. And acq_time also play an important part in decision tree model. The only difference is that the random forest doesn't use "Scan" as a very important feature as the decision tree did. We need one further investigation on the feature importance by the lightGBM model.

By fitting a lightGBM model, we have Root mean square error around 10.42. It is lightly worse than the random forest model. So we also plot the feature importance and True vs Prediction plot of lightGBM model. As shown in Figure 21 and Figure 22.

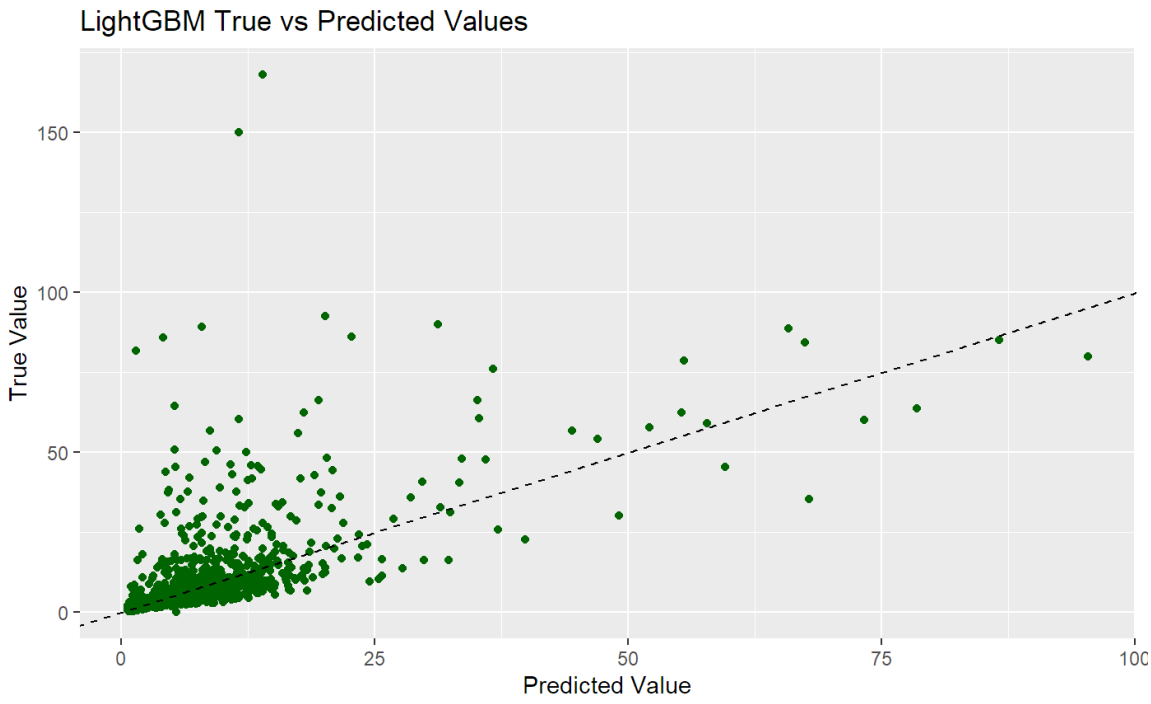

Figure 21 LightGBM True vs Pred

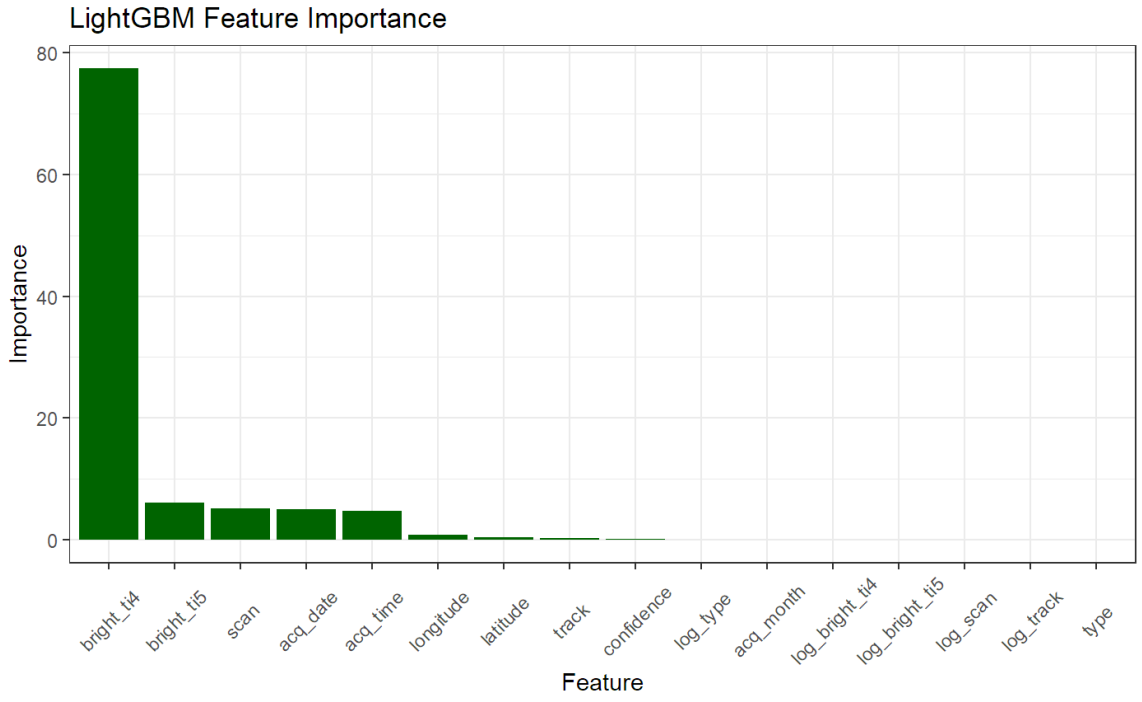

Figure 22 LightGBM Feature Importance

References:

Wong, M. (2022). VIIRS I-Band 375 m Active Fire Data. Earthdata. https://www.earthdata.nasa.gov/learn/find-data/near-real-time/firms/viirs-i-band-375-m-active-fire-data

Wikipedia contributors. (2023). Bushfires in Australia. *Wikipedia*. https://en.wikipedia.org/wiki/Bushfires_in_Australia