MONASH University

ETC3250/5250: Introduction to Machine Learning

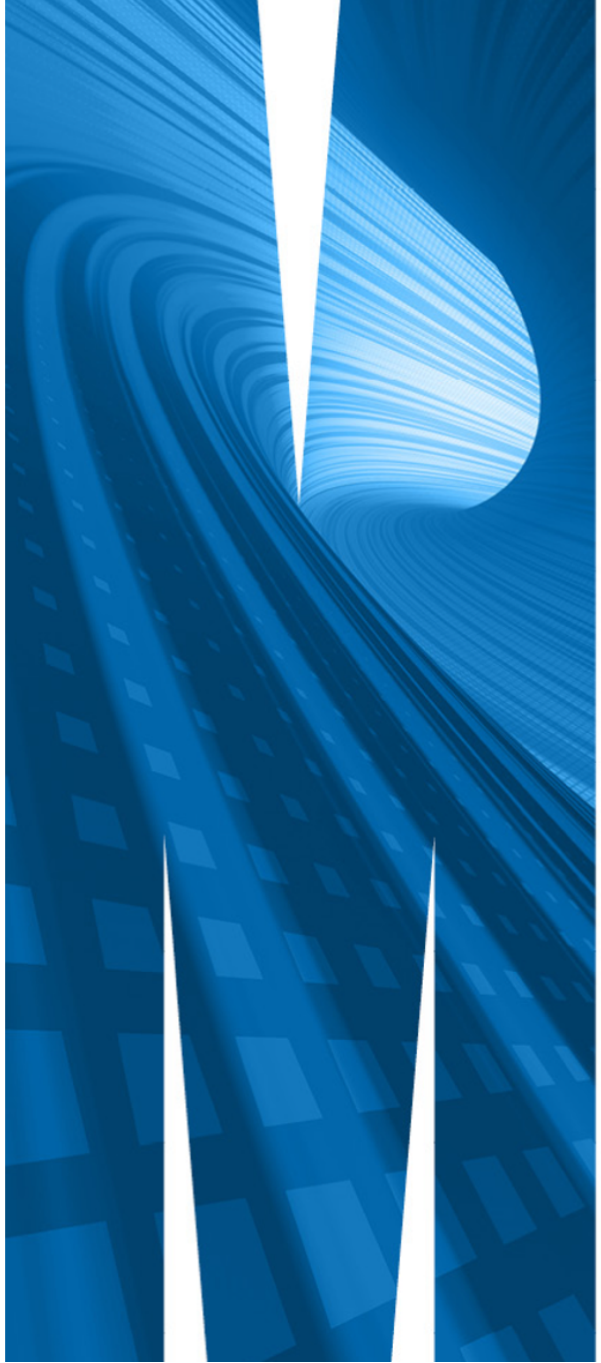
About the kaggle challenge

Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 12a



Patrick's honours thesis

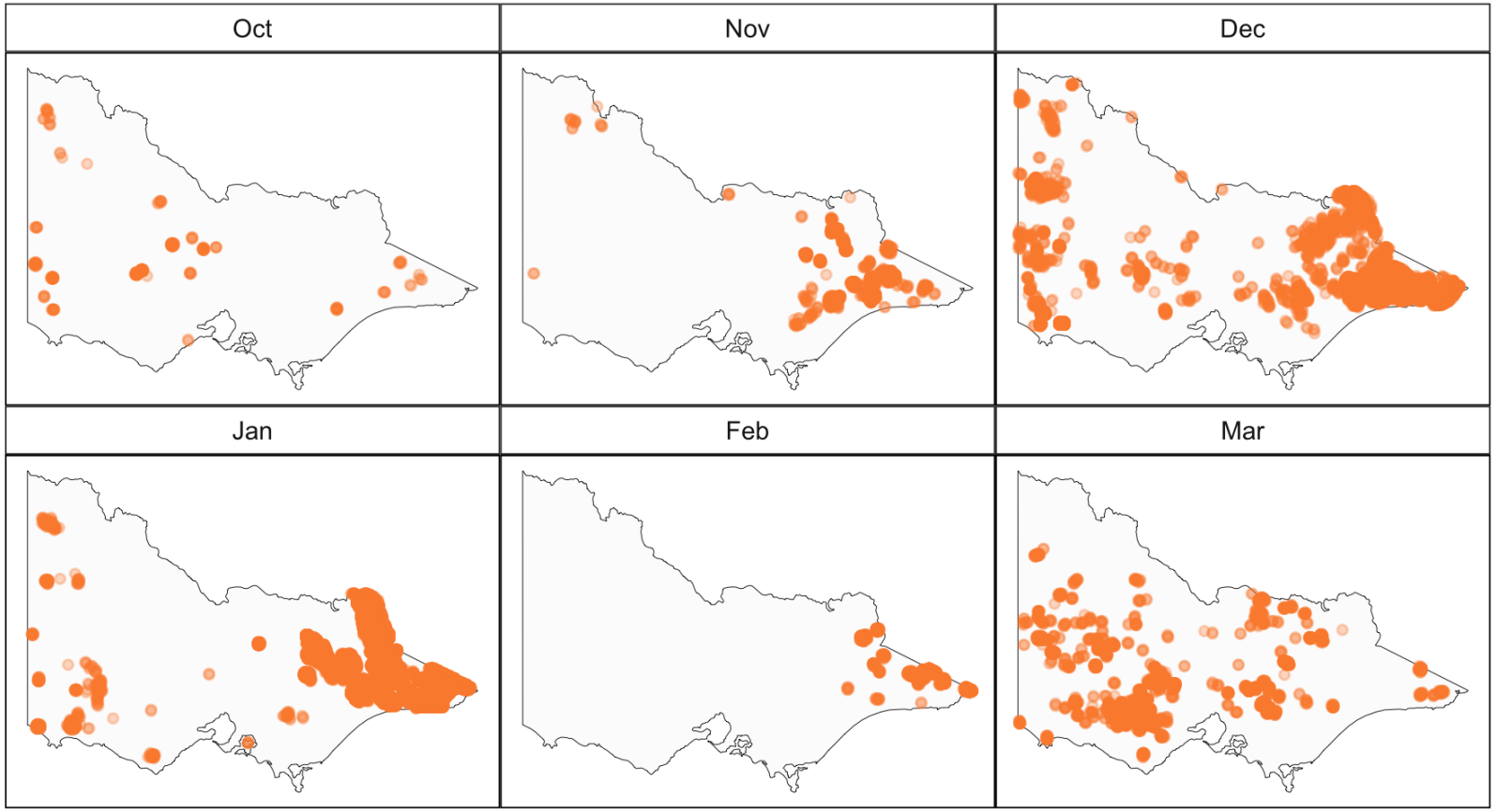
Weihao Li		Emily Dodwell
Monash EBS Honours		AT&T research

Motivation: Spatio-temporal visualisation and analysis of emergency call data. This is private so the bushfire data was collected because it has some similar form and structure.

⚡ Lightning or 🔥Arson?

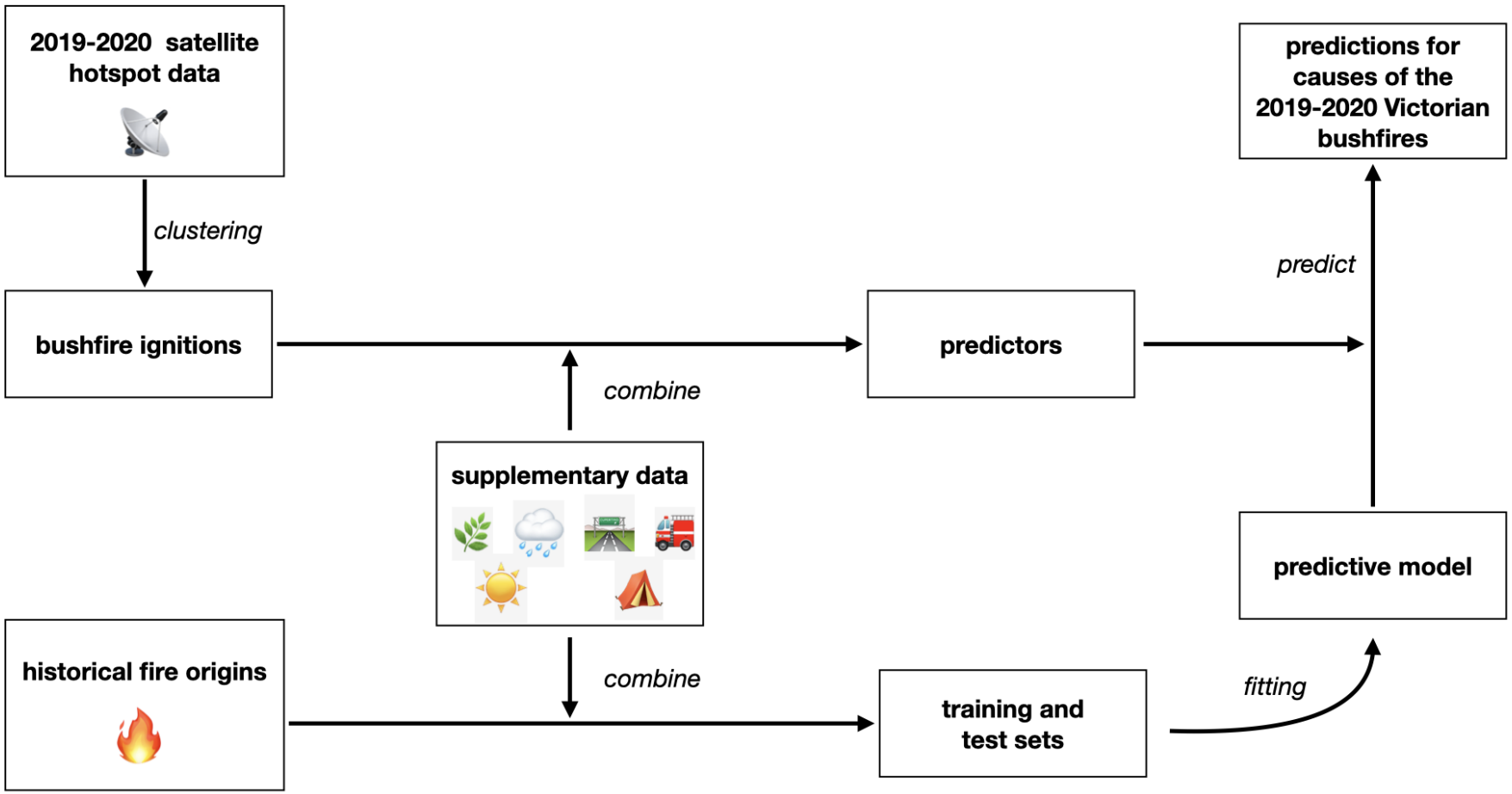
Remote sensing data

Japan Aerospace Exploration Agency provides a hotspot product (reflected energy from the earth) taken from the **Himawari-8** satellite.

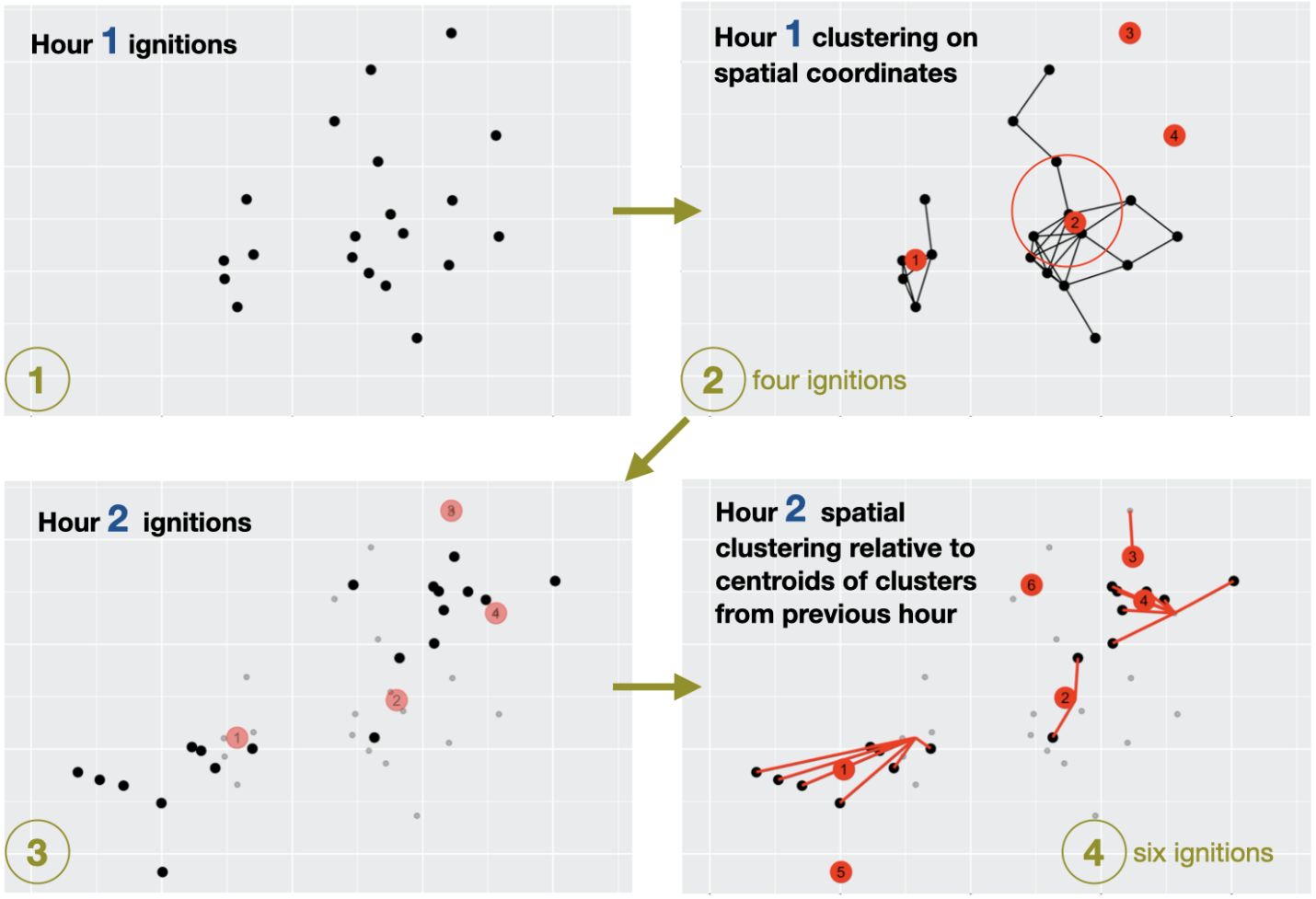


Example code to access data provided in a G. Williamson gist post

Data fusion


2019-2020 satellite hotspot data
clustering
bushfire ignitions
combine
supplementary data
predictors
predict
predictions for causes of the 2019-2020 Victorian bushfires
predictive model
historical fire origins
combine
training and test sets
fitting

Detect ignitions by clustering hotspot data



Estimated ignitions

76,000 hotspots reduced to 1,000 ignition sites.

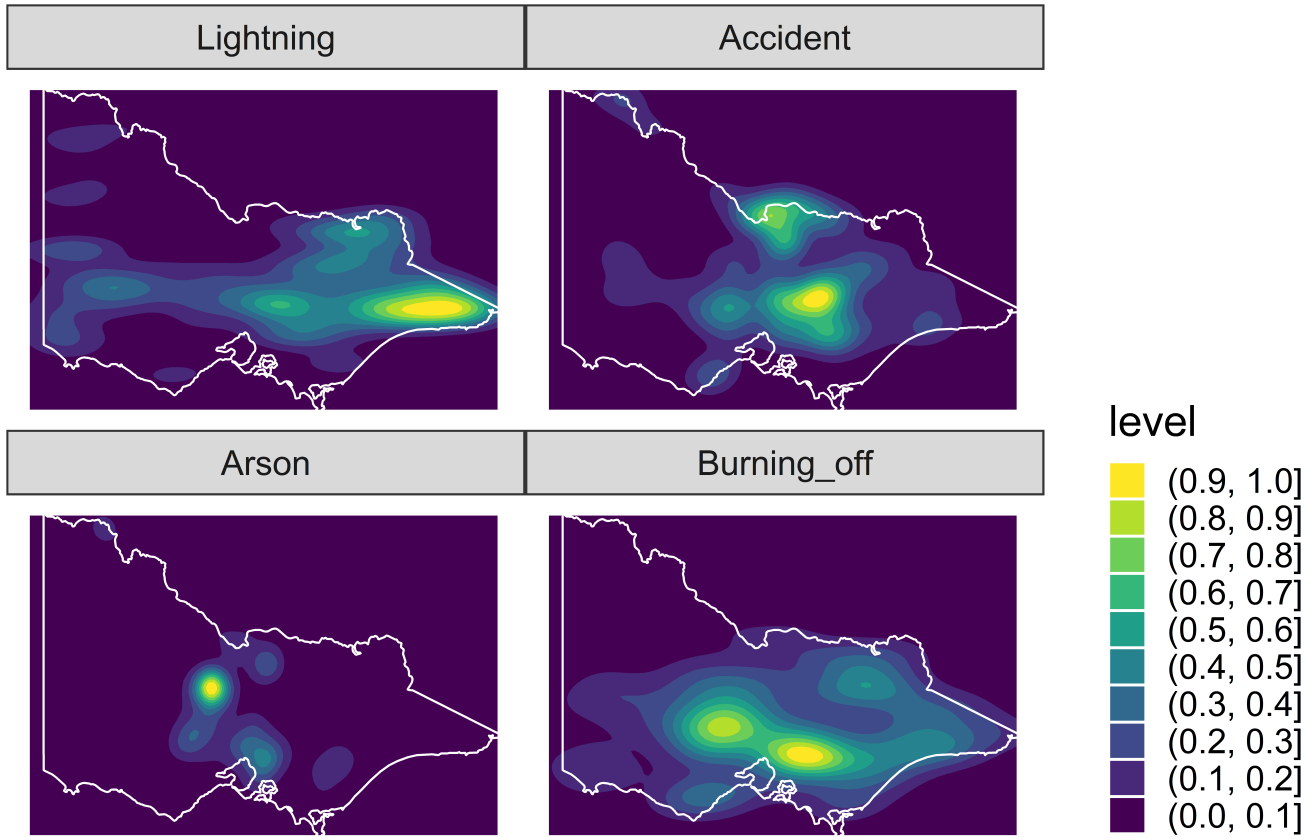
Exploratory analysis of historical fire origins

Text processing of 26 causes, reduced to four major causes. Lightning and accident were the two main sources of historical bushfire ignitions, which took up 41% and 34% respectively. There were 17% bushfires caused by arson.

Spatial distribution of historical fire origins

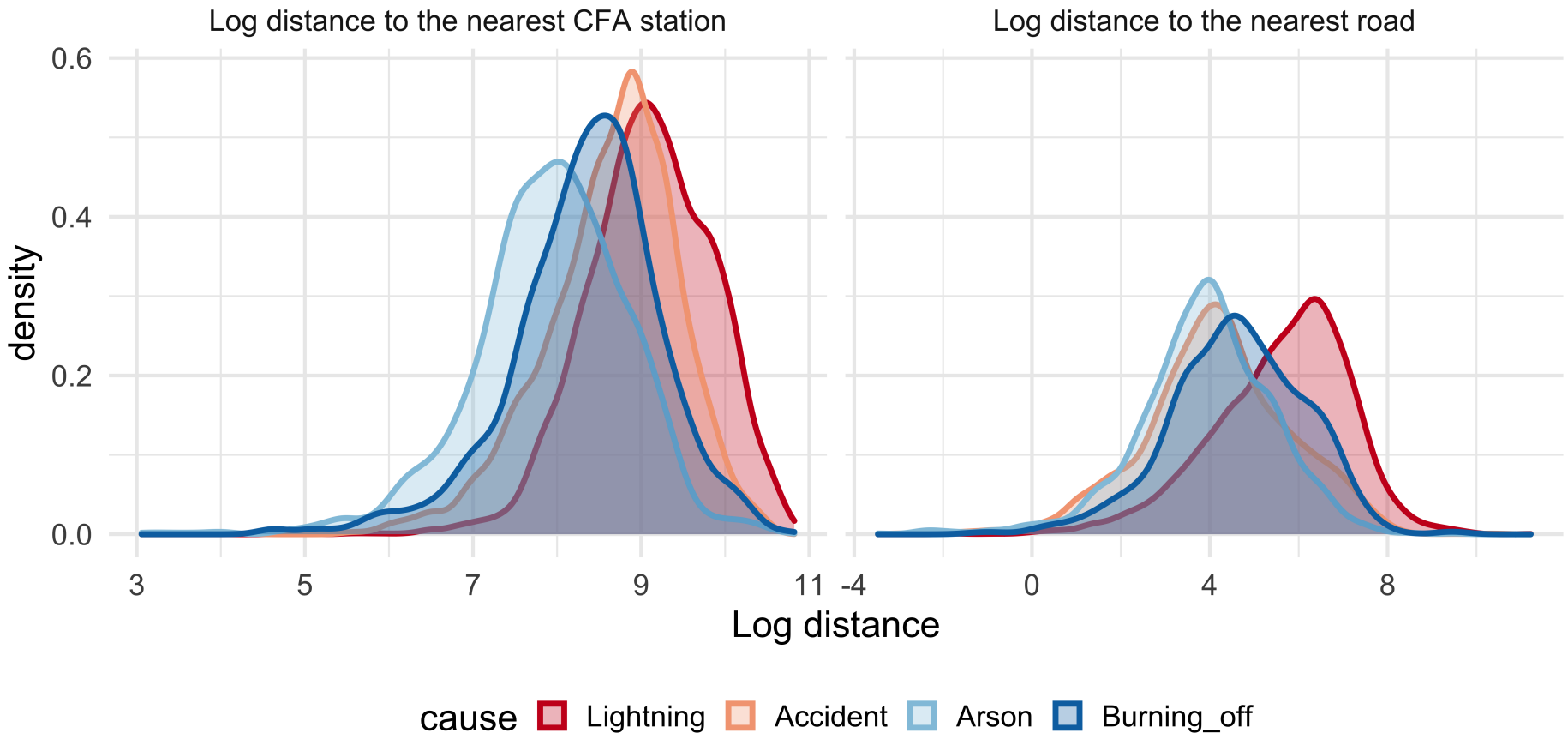
Roughly different spatial locations of ignition causes. Lightning bushfires were concentrated in the east of Victoria. Bushfires caused by arson were near Bendigo!

2D conditional density plot of historical bushfire ignitions



Proximity of historical fire origins

Lightning-caused bushfires were further away from the CFA stations and roads. In contrast, bushfires caused by arson were closer to CFA stations and roads.



Modelling

A **random forest** model outperformed other model choices to classify different causes of bushfire ignition.

80% of the data used as training set, 7497 observations, and the remaining 1872 observations was used as test set.

Model performance was compared using multi-class AUC (Hand and Till, 2001).

Model	Accuracy	Muti-class AUC
Multinomial logistic regression	0.53	0.74
GAM multinomial logistic regression	0.68	0.82
Random forest	0.75	0.88
XGBoost	0.74	0.88

Model performance

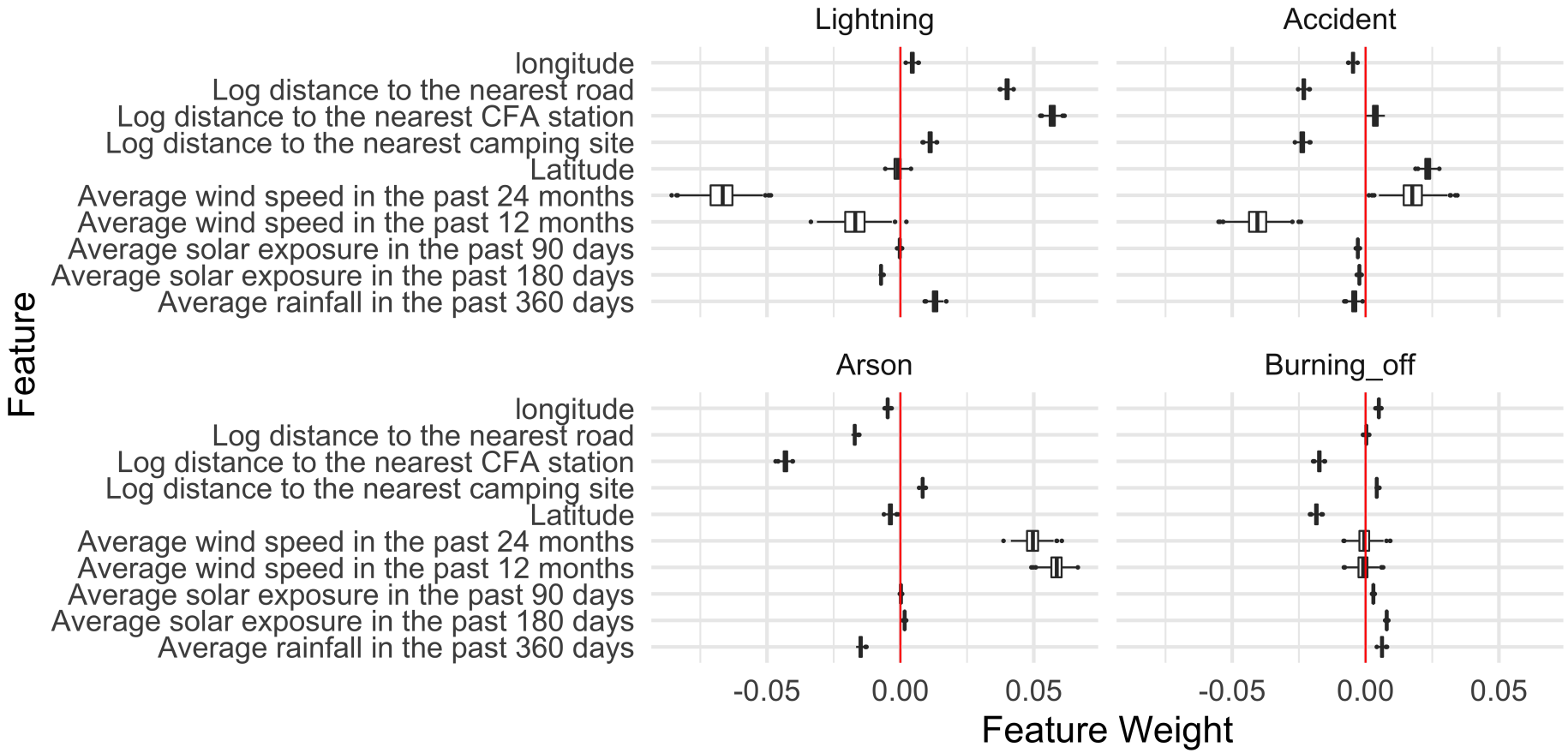
The overall accuracy of our model was 74.95%.

- High accuracy with lightning and accident ignitions.
- Less accurate predictions for arson and burning off.

	Lightning	Accident	Arson	Burning_off	Total
Prediction:Lightning	703 (0.9)	77 (0.12)	50 (0.15)	44 (0.32)	874
Prediction:Accident	51 (0.07)	494 (0.78)	89 (0.27)	38 (0.28)	672
Prediction:Arson	18 (0.02)	55 (0.09)	175 (0.54)	22 (0.16)	270
Prediction:Burning_off	5 (0.01)	8 (0.01)	11 (0.03)	32 (0.24)	56
Total	777	634	325	136	1872

Model interpretation

Variable importance assessed using Local Interpretable Model-agnostic Explanations (lime). Proximity to the nearest road, proximity to the nearest road and average wind speed had largest influence on the prediction.



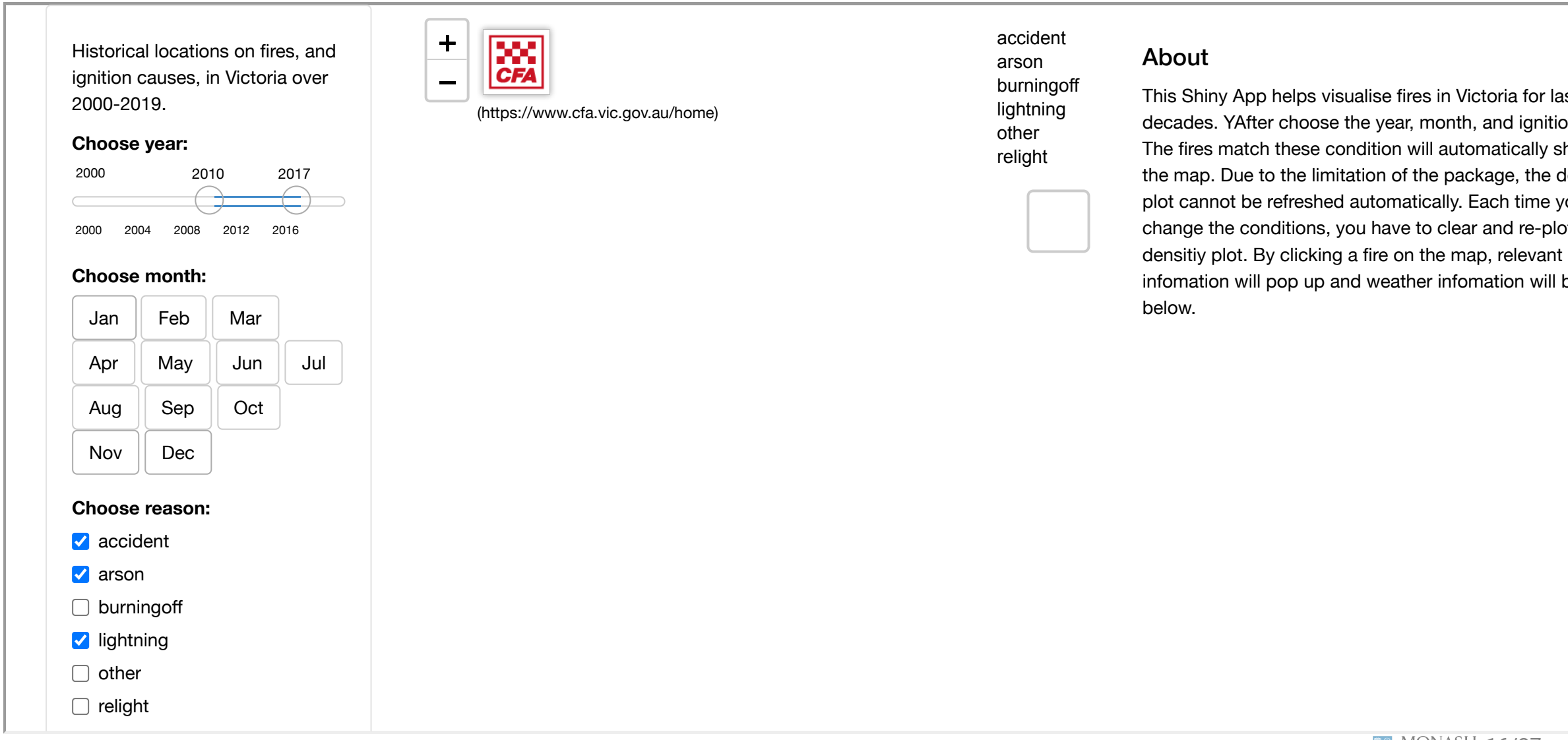
Prediction for 2019-2020 Australia bushfires

Summary of findings

- Majority of the bushfires in 2019-2020 season were caused by **lightning**.
- 138 bushfires caused by accidents which took up 14% of the total fires. Most of them were ignited in March.
- 37 bushfires were caused by arsonists, and over half of them were in March.
- Very few planned burns were predicted after October 2019 which suggests the correctness of our model.

Cause	Oct	Nov	Dec	Jan	Feb	Mar	Total
Lightning	19	57	315	266	32	149	838 (0.82)
Accident	3	8	34	13	0	80	138 (0.14)
Arson	2	2	10	2	0	21	37 (0.04)
Burning_off	7	0	2	0	0	0	9 (0.01)

Shiny app: https://ebsmonash.shinyapps.io/VICfire/



Your kaggle experience

In theory it should have been possible to perfectly predict the causes in the prediction set.

Baseline

```
set.seed(2021)
fires_rf <- rand_forest() %>%
  set_engine("randomForest",
             importance=TRUE) %>%
  set_mode("classification") %>%
  fit(cause~lon+lat+arf720+
        ase720+amaxt720+amint720+dist_cfa+
        dist_camp+dist_road+month+day, data=fires_tr_full)

fires_ts_true <- fires_ts_true %>%
  mutate(cause_p = predict(fires_rf,
                           fires_ts_true)$.pred_class)

bal_accuracy(fires_ts_true, cause, cause_p)
# bal_accuracy macro          0.702
```

Patrick's variables from actual predictions

```
set.seed(2021)
fires_rf <- rand_forest() %>%
  set_engine("randomForest",
             importance=TRUE,
             na.action=na.omit) %>%
  set_mode("classification") %>%
  fit(cause~log_dist_camp+log_dist_road+log_dist_cfa+lon+
        ase180+lat+aws_m24+aws_m12+arf360+ase90,
      data=fires_tr_full)

fires_ts_true <- fires_ts_true %>%
  mutate(cause_p = predict(fires_rf,
                           fires_ts_true)$.pred_class)

bal_accuracy(fires_ts_true, cause, cause_p)
# bal_accuracy macro          0.798
```

Modify forest parameters

```
set.seed(2021)
fires_rf <- rand_forest(mtry=1, trees=1000) %>%
  set_engine("randomForest",
             importance=TRUE,
             na.action=na.omit) %>%
  set_mode("classification") %>%
  fit(cause~log_dist_camp+log_dist_road+log_dist_cfa+lon+
        ase180+lat+aws_m24+aws_m12+arf360+ase90,
      data=fires_tr_full)

fires_ts_true <- fires_ts_true %>%
  mutate(cause_p = predict(fires_rf,
                           fires_ts_true)$.pred_class)

bal_accuracy(fires_ts_true, cause, cause_p)
# bal_accuracy macro          0.859
```

Filter only summer months

```
set.seed(2021)
fires_rf <- rand_forest(mtry=1, trees=1000) %>%
  step_filter(month %in% c(1,2,3,10,11,12)) %>%
  set_engine("randomForest",
             importance=TRUE,
             na.action=na.omit) %>%
  set_mode("classification") %>%
  fit(cause~log_dist_camp+log_dist_road+log_dist_cfa+lon+
        ase180+lat+aws_m24+aws_m12+arf360+ase90,
      data=fires_tr_full)

fires_ts_true <- fires_ts_true %>%
  mutate(cause_p = predict(fires_rf,
                           fires_ts_true)$.pred_class)

bal_accuracy(fires_ts_true, cause, cause_p)
# bal_accuracy macro          0.872
```

Use weights during model building

```
# Need to revert to old style code
set.seed(2021)
fires_tr_full_summer <- fires_tr_full %>%
  filter(month %in% c(1,2,3,10,11,12)) %>%
  drop_na()
fires_tr_full_summer %>% group_by(cause) %>% summarise(n=n()) %>% mutate(n/sum(n
fires_rf <- randomForest(cause~log_dist_camp+log_dist_road+
                log_dist_cfa+lon+ase180+lat+aws_m24+
                aws_m12+arf360+ase90,
        data=fires_tr_full_summer,
        mtry=1, trees=1000,
        importance=TRUE,
        classwt = c(0.414, 0.339, 0.174, 0.0730))

fires_ts_true <- fires_ts_true %>%
  mutate(cause_p = predict(fires_rf,
                           fires_ts_true, type="class"))
```

```
bal_accuracy(fires_ts_true, cause, cause_p)
# bal_accuracy macro          0.910

> conf_mat(fires_ts_true, cause, cause_p)
             Truth
Prediction    lightning accident arson burning_off
  lightning         820       21     1           1
  accident           17      112     3           1
  arson               1        4    32           0
  burning_off         0        1     1           7
```

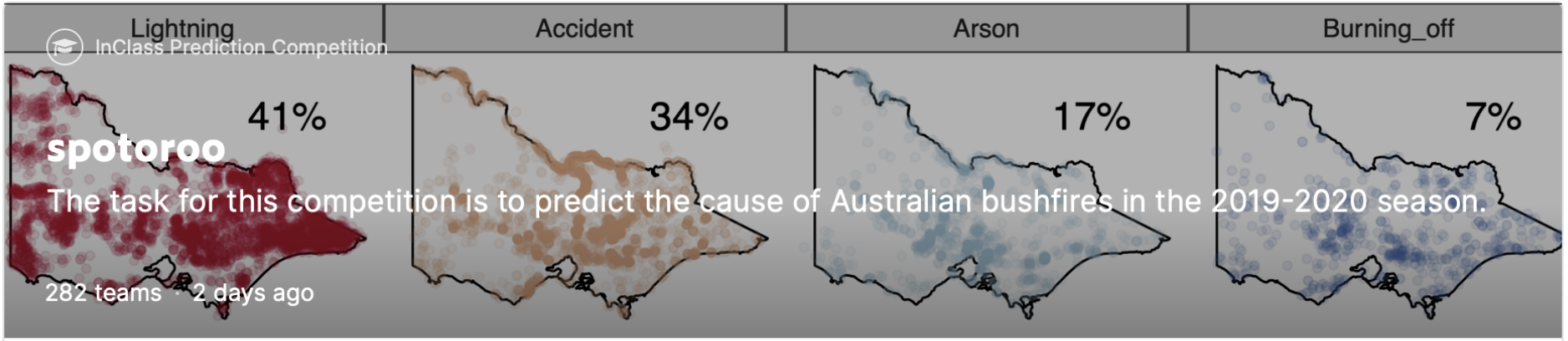
Upsample

```
set.seed(2021)
fires_rf <- rand_forest(mtry=1, trees=1000) %>%
  step_filter(month %in% c(1,2,3,10,11,12)) %>%
  themis::step_upsample(cause) %>%
  set_engine("randomForest",
             importance=TRUE,
             na.action=na.omit) %>%
  set_mode("classification") %>%
  fit(cause~log_dist_camp+log_dist_road+log_dist_cfa+lon+
        ase180+lat+aws_m24+aws_m12+arf360+ase90,
      data=fires_tr_full)

fires_ts_true <- fires_ts_true %>%
  mutate(cause_p = predict(fires_rf,
                           fires_ts_true)$.pred_class)

bal_accuracy(fires_ts_true, cause, cause_p)
```

Weighted accuracy

- lightning 0.02 (0.82)
- accident 0.05 (0.14)
- arson 0.18 (0.04)
- burning_off 0.75 (0.01)

	cause	n	p	invp	invp_sum1
	<fct>	<int>	<dbl>	<dbl>	<dbl>
1	lightning	838	0.820	0.00119	0.00814
2	accident	138	0.135	0.00725	0.0494
3	arson	37	0.0362	0.0270	0.184
4	burning_off	9	0.00881	0.111	0.758



Overview Data Code Discussion Leaderboard Rules Team Host My Submissions **Late Submission**

https://www.kaggle.com/c/spotoroo/host/all-submissions

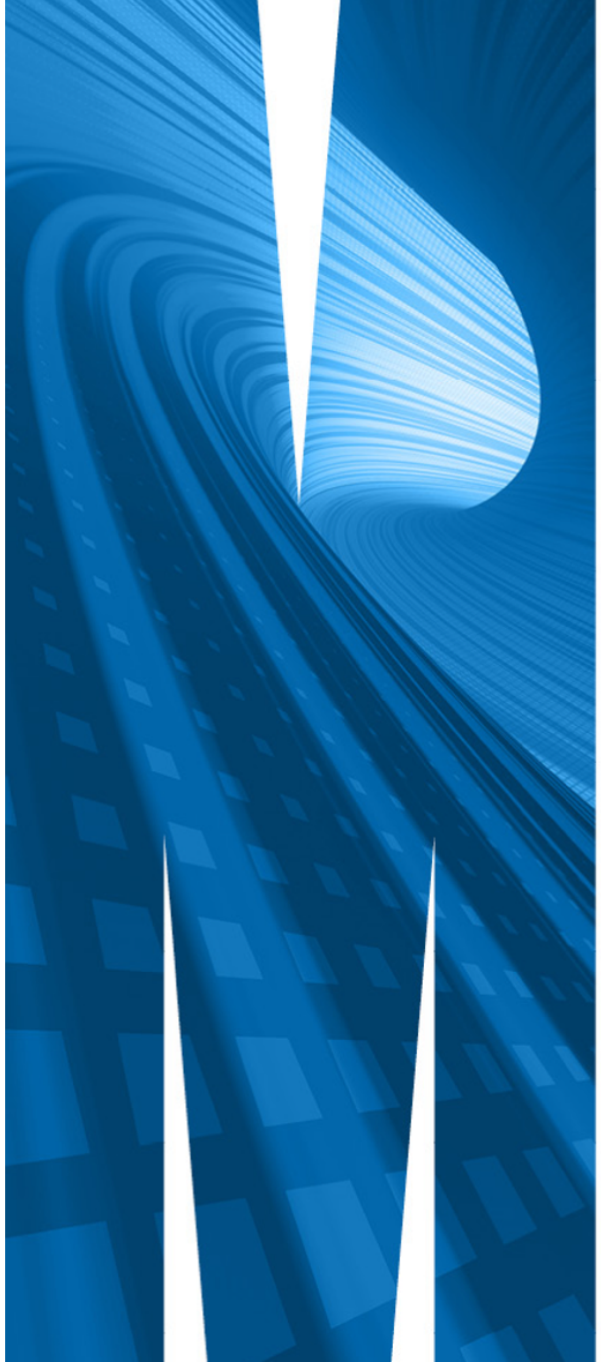
Host Controls

- Settings
- Images
- Privacy
- Manage Hosts
- Evaluation
- Sandbox Submissions
- All Submissions

282 teams , 3,257 submissions Sort By Rank Filter by team name...

#	Team Name	Entries	Public	Private	Models
1	[Deleted] 97a48b2...	4	1	1	
2	Admin	3	0.88172	0.88299	
3	Admin2	2	0.870968	0.866495	
4	jtho0048	16	0.793435	0.840722	
5	jwan0296	14	0.788342	0.832474	

MONASH University



Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 12a