Final report 3250

Elvis Zhixiang Yang 30306396

5/23/2021

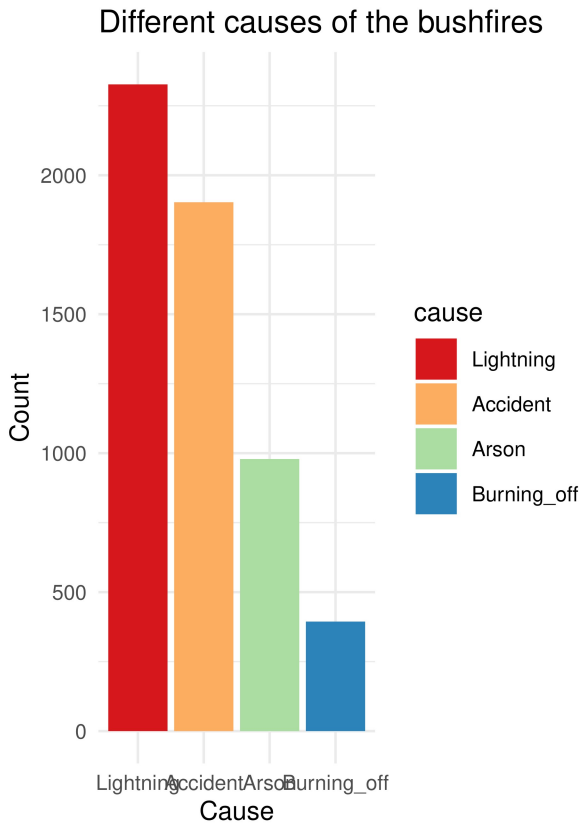
Data implementation changes

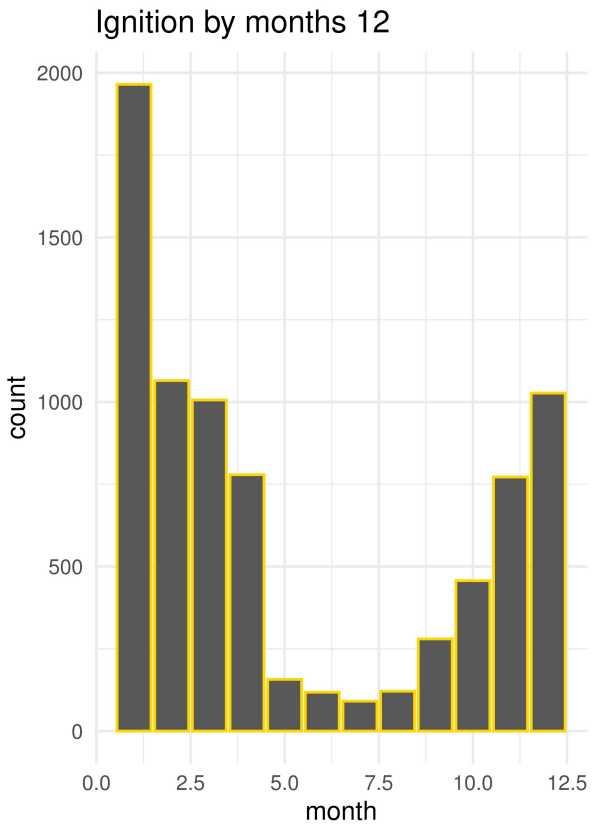
From following plots, We can get bushfires are more likely to start around the summer times of Victoria(In months 10,11,12,1,2,3).Here,to reduce the bias of data and fit the prediction dataset, we should only filter the significant months from dataset which are months 10,11,12,1,2,3.Meanwhile,among all 4 causes the lightning and Accident are most likely to happen. We use `outlier` package to replace the extreme value inside columns of the dataset of their column median, which reduce extreme values and bias. However, after we filter the extreme values, the Kaggle score is decreased while public increased. That means the data is overfitted or omitted so we couldn't remove outliers.

Modeling method	Random forest	XGboost
Before rm.outlier	0.84663	0.80412
After rm.outlier	0.85398	0.73969

Another issue is Categorical variables, the option in dummy variables don't have standardized interval scale, which respondents are not able to effectively gauge their options before responding(Blog, 2021). Thus,to avoid the interference when selecting significant variables in modeling, we here remove the categorical variable FOR_TYPE, FOR_CAT etc. **plot1**

PlotA, *histogram of bushfires cause and happend within one year*





Model selection RF model is used to predict causes for Random Forest is a practical way and regularly used in machine learning models. Random Forest Model have the variable selecting system (via boostraping) to decide the most significant tree and can reduce overfitting compared with decision tree.With that said, random forests are a strong modeling technique and much more robust comparing with many different methods. (Liberman, 2017) The model has a OOB of 26.29%, which reflects it's a good model to estimate.

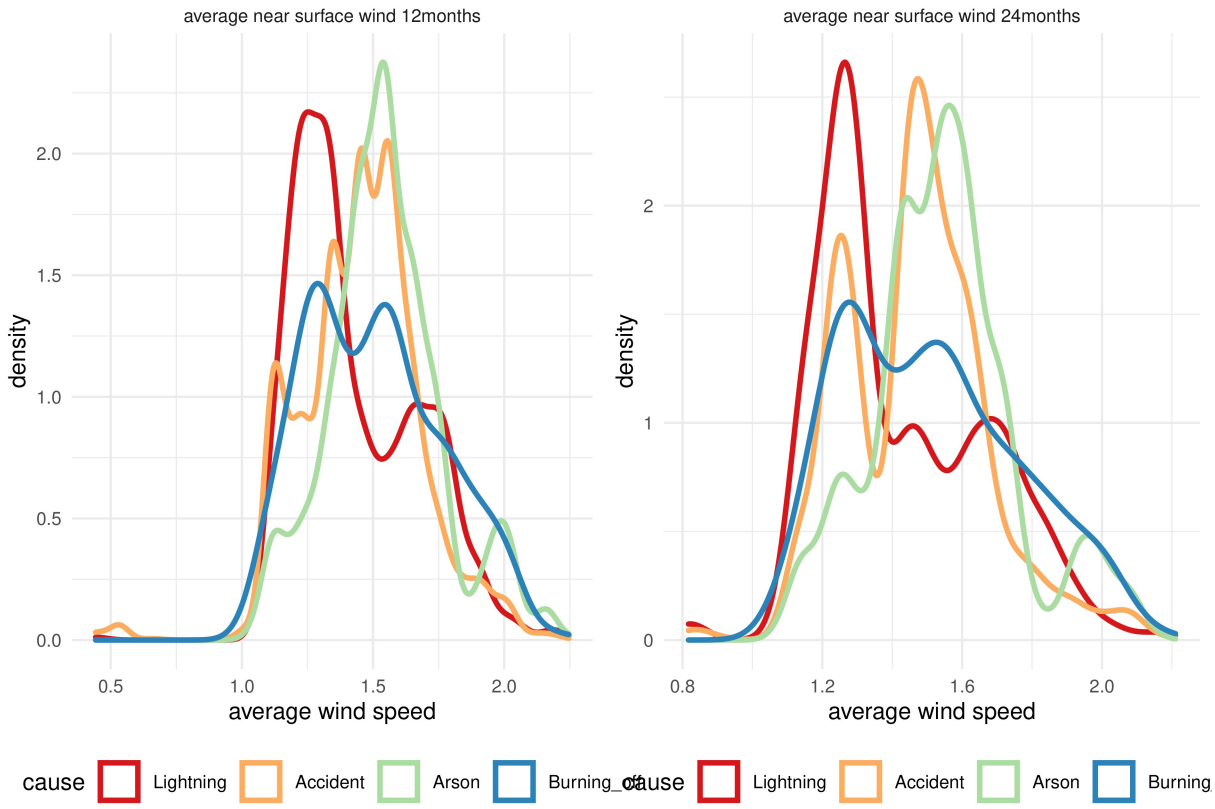
We then tested the XGBoost, it solves complex questions via building parallel computing for different trees. It can solve the classification and complex questions efficiently. However, after testing the XGBoost, we only got 0.74. Such bias may caused by overfitting, we cannot use XGBoost.

Modeling	fstvarkag	scndvarkag
Random Forest	0.85398	0.84663
XGBoost	0.74690	0.75618

Classification of the variables + Explaintory Data Analysis

Why Month ?

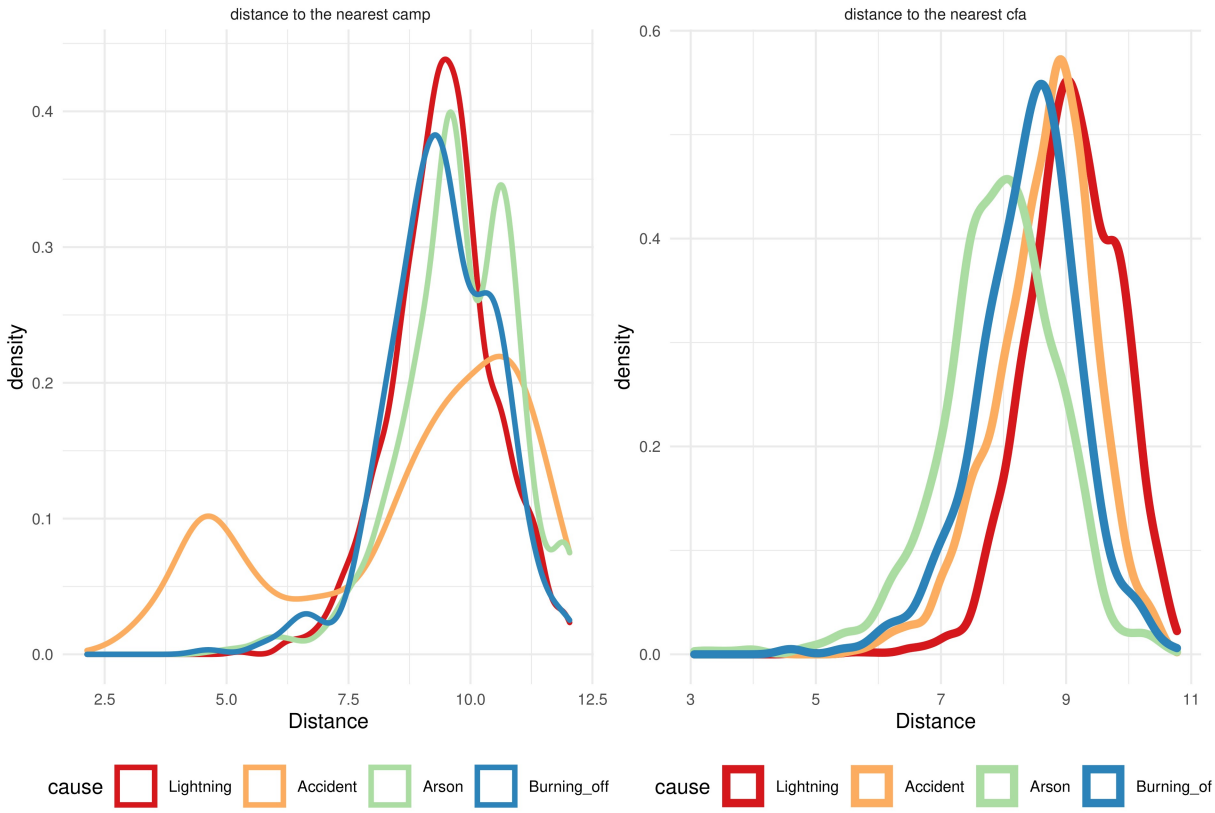
The data is Yearly, months are make more sense to track changes than a day(*observation dataset is large which may occur more bias*) and year(*observation dataset too small couldn't reflect the overall trend*). ### Why average wind speed(aws)? As we can conclude from following density plot, the bushfires are also positively correlated with the yearly wind speed. Based on plot and general rule we can say that the slower wind speed is, the less chance of getting fire lit and spread out.



Why distances(dist_road,cfa,camp)??

Also from the plot, there is a relationship between the fire ignition with distances to CFA, CAMP, ROAD. In summer times, the accident caused ignitions were close to camp, which means camp was a source of the bushfires. Moreover, we can also conclude that the closer to the road,more chance to get fired of accident. Also, different causes are showed different performance of the distance to CFA station and road. Such a reason maybe just because it's correlated with people's activities and density.

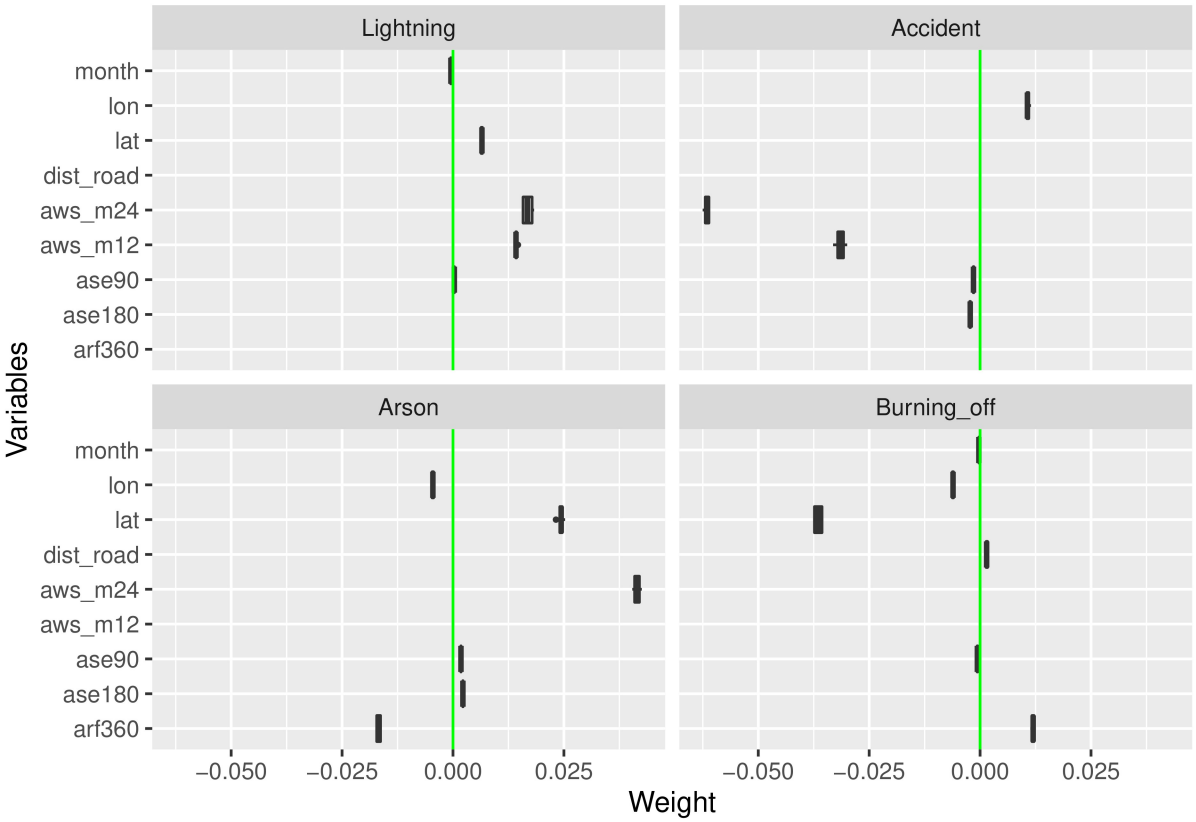
The distance density plot of the camp (After transformed the dist_road,dist_cfa,dist_camp)



Why ase and arf??

Moreover, based on the general role, we also found the Average solar exposure(ase) and Average Rainfall(arf) also did a great influence on the lighting. By observing the cause by ase and arf of **graph 1** . The exposure of solar is positively correlated with bushfires caused by lighting and Burning_off, while the rainfall is correlated with bushfires caused by lighting and burning_off. So we should choose these 2 classes as a reference because they are important when predicting these 2 causes(*majority among all 4 causes*).

In this graph, variable has a positive weight means it has a positive impact on the probability,vice versa . The magnitude shows the impact strength, which is the variable influence on a local point of view.



After the explanatory analysis,we already sort classes out, which are lon,lat,aws,arf,ase.By observing the data. Based on the mean decrease accuracy selection criteria performance on Kaggle, we found that if we use the global variable importance(MDA) in the model random forest is likely to be misleading and biased. So here we use Lime to get variable importance under the local aspects.Then after generating a new model, we'll test them on kaggle and compare their scores between different models. By repeating such a process, we sort out the significant variables.

Reference List

Liberman, N. (2017). Decision Trees and Random Forests. Retrieved 23 May 2021, from https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991 (Liberman, 2017)

Blog, F. (2021). Categorical Data: Definition + [Examples, Variables & Analysis]. Retrieved 23 May 2021, from https://www.formpl.us/blog/categorical-data

Packages

Thomas Lin Pedersen and Michaël Benesty (2021). lime: Local Interpretable Model-Agnostic Explanations. R package version 0.5.2. https://CRAN.R-project.org/package=lime

Lukasz Komsta (2011). outliers: Tests for outliers. R package version 0.14. https://CRAN.R-project.org/package=outliers

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.

Kuhn et al., (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686