**Suncorp Industry Sandbox Project Summary**

**Purpose of document**

According to the ICA Catastrophe Database, on average insurance companies have seen over $360m annually in normalised loss value, with over 780 claims on bushfires received annually in Australia over the past 5 decades. Given the financial implications of these natural hazards on insurance companies, bushfire event risk is a critical assumption for pricing, and will need to be assessed accurately to maintain long-term solvency.

**Data Preparation**

General data cleaning on internal and external data involved removing structurally missing or duplicate data, converting variable types to appropriate classes and filtering for relevant data. Additional pre-processing of data was performed to allow for the merging of datasets under a state-level spatial and daily temporal resolution. Historical catastrophe data was expanded such that observations were grouped by catastrophe event and state, with a filter applied to include only bushfire events. Monthly SOI and IOD data were converted to daily data to match the FWI dataset using linear interpolation as an approximation under the assumption that values given are for the start of each month.

**Solution**

Our chosen model for bushfire event risk is a Gradient Boosting Machine (GBM) with optimal tuning parameters chosen based on the cross validation error. Given the small sample size of data available, using 10-fold cross validation as a measure of model accuracy was more feasible than splitting the training set further for cross validation. Other models including Logistic GLM, Regularised Regression, Classification and Regression Tree (CART), Bagging and Random Forest were also considered but performed relatively poorly when compared across several metrics.

Given the model objective, there is a higher emphasis for predictive power over model interpretability. Further, the nature of bushfire claims which results in large insurance losses per bushfire event results in a higher preference for false positives over false negatives. Based on these criteria, Bagging and GBM outperform CART despite being less interpretable and having lower sensitivity as it has higher specificity in exchange for a slight reduction in sensitivity. When comparing Bagging with GBM, it is noted that accuracy is not as reliable when dealing with imbalanced class distributions and thus less weighting should be placed on this performance measure (Akosa, 2017). Given that there is only a slightly lower false positive rate using GBM, this model is selected for its better predictive power as indicated by its AUC performance.

The bushfire event risk score for each state was derived using the predicted probability of a bushfire event occurring using the historical mean of predictor values for each state under the GBM model. Using this measure of bushfire event risk, NSW was found to have the highest risk of 0.0798% with WA having the lowest risk of 0.0023%.

**Assumptions and Limitations**

The nature of bushfire events meant that experience was limited, and the grouping of observations by state further decreased the sample size of data available. Further, since imbalance in class distribution of the response variable reduces the accuracy of the models, particularly for ensemble models which rely on classification error to improve, the upsampling method was applied to the training set (O’Brien & Ishwaran, 2019). Whilst balancing the training set helps alleviate the previous issues, this is a manual adjustment to the dataset and therefore reduces the accuracy of our model.

**References**

Akosa, J. (2017) ‘*Predictive accuracy: A Misleading Performance Measure for Highly Imbalanced Data’.* Available at:

<https://support.sas.com/resources/papers/proceedings17/0942-2017.pdf> (Accessed: 11 November 2023).

O’Brien,R. and Ishwaran, H. (2019) ‘*A Random Forests Quantile Classifier for Class Imbalanced Data, Pattern recognition’*. Available at:  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6370055/> (Accessed: 11 November 2023).

LinkedIn (2023) ‘*How do you choose the appropriate level of granularity for your data models?’* Available at: <https://www.linkedin.com/advice/0/how-do-you-choose-appropriate-level-granularity-2e> (Accessed: 12 November 2023).

van Oldenborgh, G. J., Krikken, F., Lewis, S., Leach, N. J., Lehner, F., Saunders, K. R., van Weele, M., Haustein, K., Li, S., Wallom, D., Sparrow, S., Arrighi, J., Singh, R. K., van Aalst, M. K., Philip, S. Y., Vautard, R., and Otto, F. E. L (2021) ‘*Attribution of the Australian bushfire risk to anthropogenic climate change’*. Available at: <https://nhess.copernicus.org/articles/21/941/2021/> (Accessed: 12 November 2023).

Google (2023) ‘*Imbalanced Data’.* Available at: <https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data> (Accessed: 12 November 2023)

Nallamuth, N. (2020) ‘*Handling Imbalanced Data - Machine Learning, Computer Vision and NLP*’. *Analytics Vidhya*. Available at: <https://www.analyticsvidhya.com/blog/2020/11/handling-imbalanced-data-machine-learning-computer-vision-and-nlp/> (Accessed: 12 November 2023)

NT Police, Fire & Emergency Services (2023) ‘*History of the NT Fire and Rescue Service*’. Available at: <https://pfes.nt.gov.au/corporate/our-history/nt-fire-and-rescue-service> (Accessed: 12 November 2023)

Quilcaille, Y., Batibeniz, F., Ribeiro, A. F. S., Pardron, R. S., Seneviratne, S. I. (2022) ‘*Fire weather index data under historical and SSP projections in CMIP6 from 1850 to 2100*’. *Earth System Science Data*. Available at: <https://essd.copernicus.org/preprints/essd-2022-413/essd-2022-413.pdf> (Accessed: 12 November 2023)

Dowdy, A. J., Mills, G. A., Finkele, A. M. and de Groot, W. (2009) ‘*Australian fire weather as represented by the McArthur Forest Fire Danger Index and the Canadian Forest Fire Weather Index*’. *Australian Government Bureau of Meteorology*. Available at: <https://www.cawcr.gov.au/technical-reports/CTR_010.pdf> (Accessed: 12 November 2023)