Sandbox Project: Prediction of Building Claims Costs in SMEs

SUNCORP

UNSW SYDNEY Business School


Queenie Huang


Melantha Wang


Michael Jacinto


Kevin Li


Harry Peng

Background to the project

Project

The (challenging) task of predicting building claims costs in SMEs (Small and Medium Enterprises) across Australia

Issues

Insufficient claims experience per occupation to understand occupation-level risks

High heterogeneity between policies → aggregate modelling would be inefficient

More granular risk modelling ⟷ Less differentiating risk modelling

Goals

1) Develop occupational rating scheme both accurate and consistent with domain knowledge

2) Build and test models for predicting working claims cost of SME building insurance

Impact

- Fast growing SME sector → market potential
- More accurate model → sustainable and competitive pricing → edge over other insurers

Agenda

1. **Data Preparation**

   Impute missing values and handle invalid data.

2. **Exploratory Data Analysis**

   Identify patterns in the data, group high cardinal variables.

3. **Occupation Grouping**

   Cluster occupations by claim size and proportion of fire claims, rank them by claim size.

4. **Claims Cost Model**

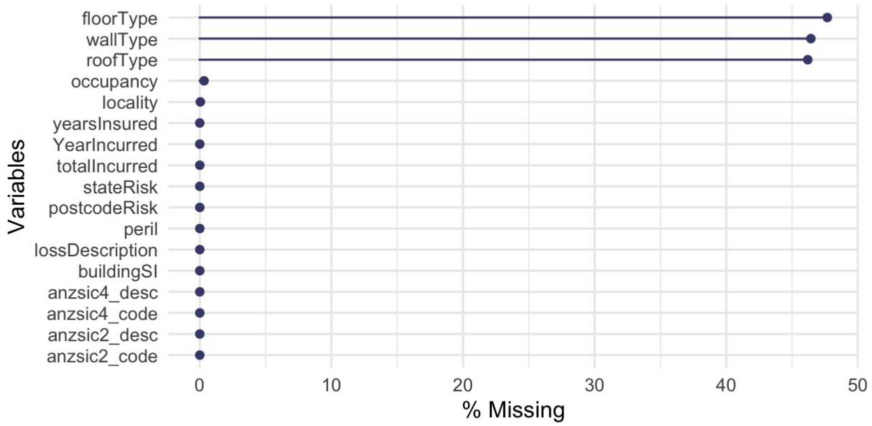
   Propose a GLM-XGBoost hybrid model, demonstrate how it can be applied in practice.

5. **Limitations and Next Steps (including Appendix)**

   Considered limitations and ways of improvements, further findings

Data Preparation

Building materials was plagued with **missing data**



floorType, wallType and roofType are 45-47% missing

Imputation not feasible as too many values missing

Building materials was plagued with **missing data**

locality	roofType	wallType	floorType	fireprotecAlarm	fire
Officefrst	Alloysteel	Brick	Concrete	0	
Officefrst	Alloysteel	Brick	Concrete	0	
Indust	Alloysteel	Brick	Concrete	0	
Officefrst	Alloysteel	Brick	Concrete	0	
Conversion	Asbestos	Brick	$\boxed{\text{Z-Missing}}$	0	
Conversion	Alloysteel	Brick	Concrete	0	
Conversion	Alloysteel	Concretsla	Concrete		
Substfrst	Othmixed	Brick	Othmixed	0	
Mainrd	Alloysteel	Brick	Concrete	0	

floorType, wallType and roofType are 45-47% missing

Imputation not feasible as too many values missing

Solution: code the missing values as a separate factor

Invalid data was treated depending on the variable

Anomalies (invalid data)

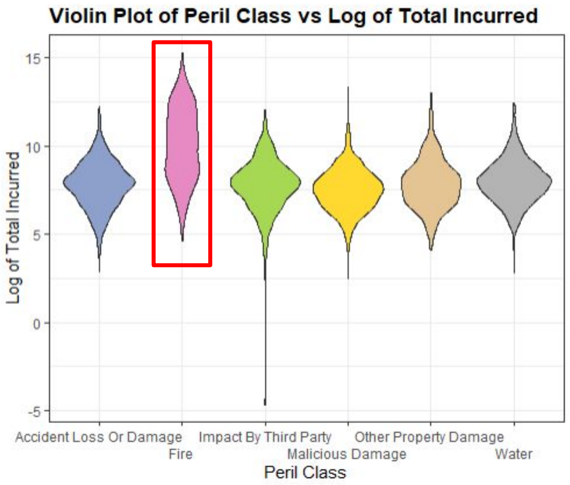
Out of the **6,746** claims from the data,

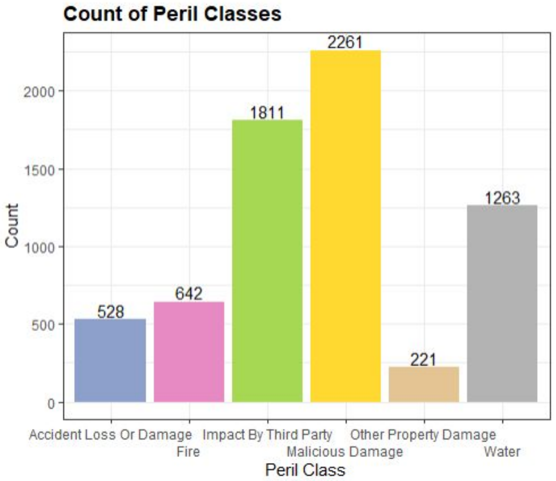
20	35	11	4	Occupancy
Claims with a negative incurred amount	Claims with incurred amount > sum insured	Policies with negative tenure	Policies with tenure > 100 years	Has a single level: "PropOwner"
Remove	**Keep**	**Change**	**Change**	**Remove**

Data preparation → EDA → Occupation grouping → Claims cost models → Limitations & next steps

Exploratory Data Analysis

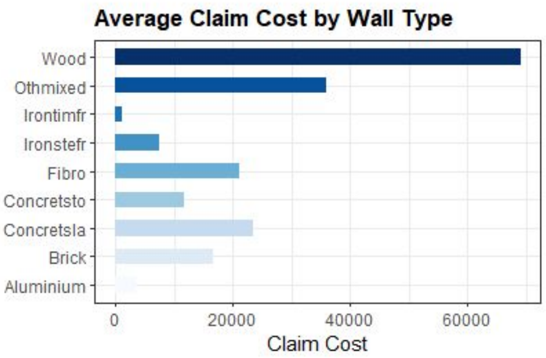
Fire claims show different behaviours to other claim types

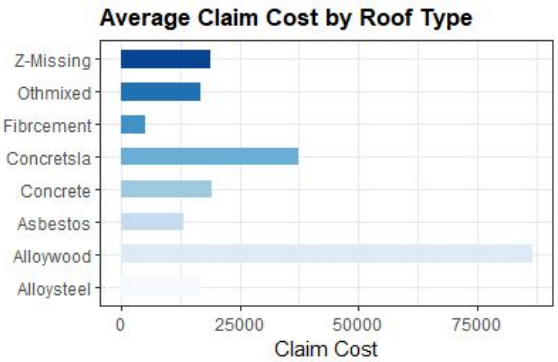


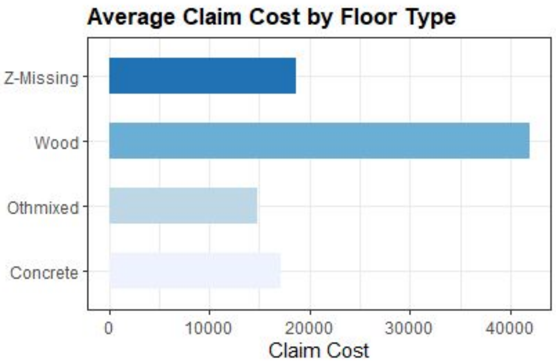


Fire claims, although less frequent, tend to be more severe than all other perils

Building materials have a significant impact on claim size



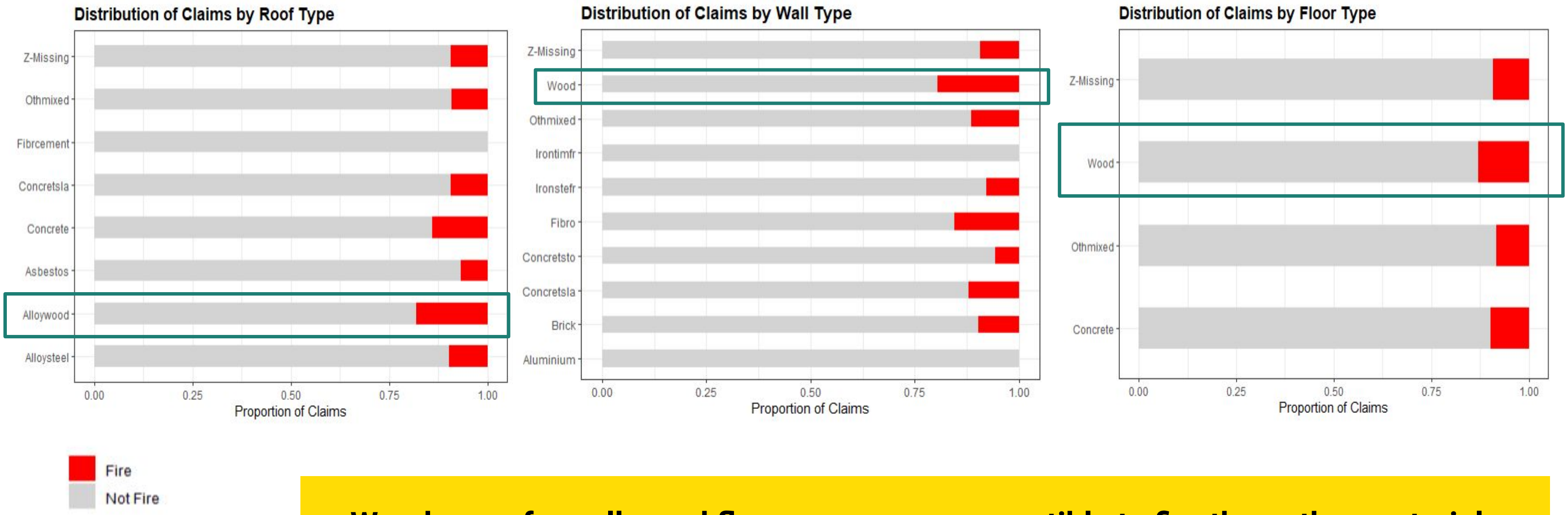




Alloywood and wood have higher average claims costs than other materials

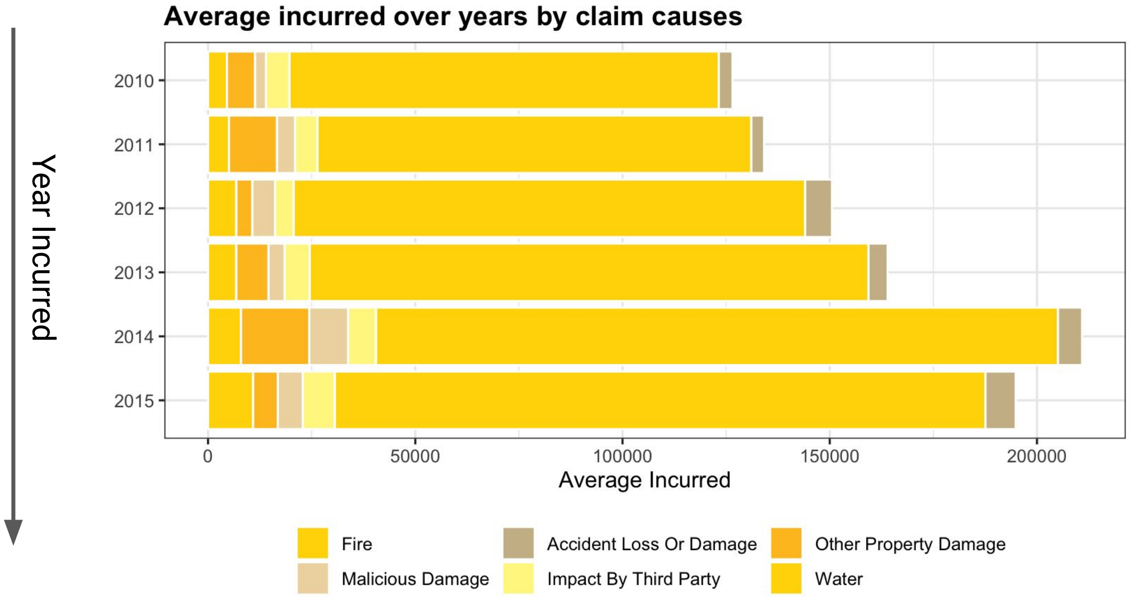
Data preparation → **EDA** → Occupation grouping → Claims cost models → Limitations & next steps

Building materials have a significant impact on claim size



Wooden roofs, walls, and floors are more susceptible to fire than other materials

Later modelling must account for **superimposed inflation**

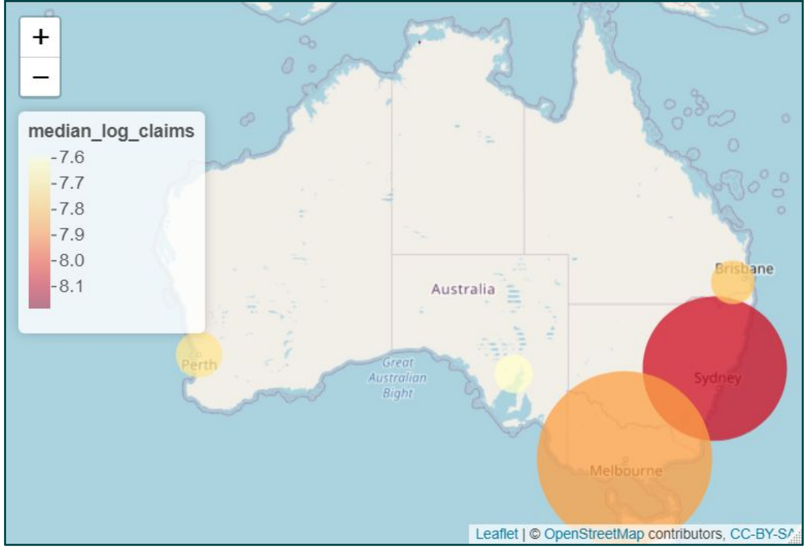


Steady rise in average incurred across years

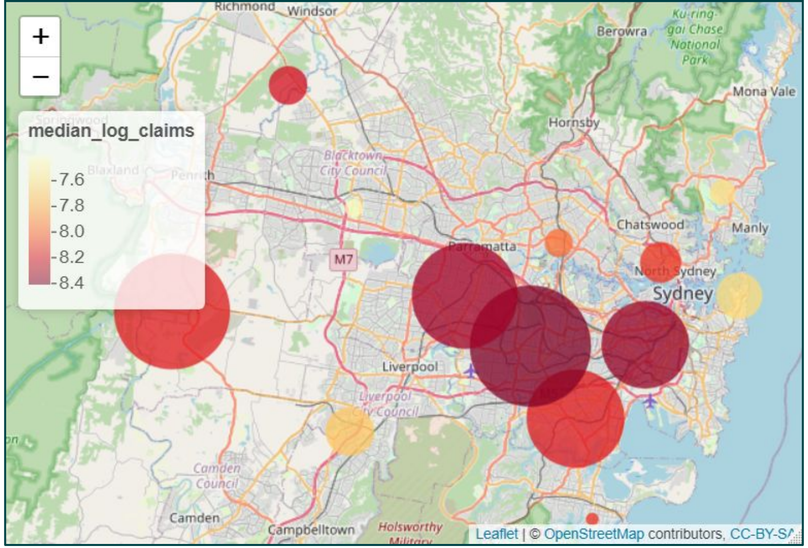
Potential inflation impact

Postcode risk was **regrouped into Statistical Area (level 4)** to reduce feature cardinality

Median total incurred* by states is too broad (while postcode is too granular)



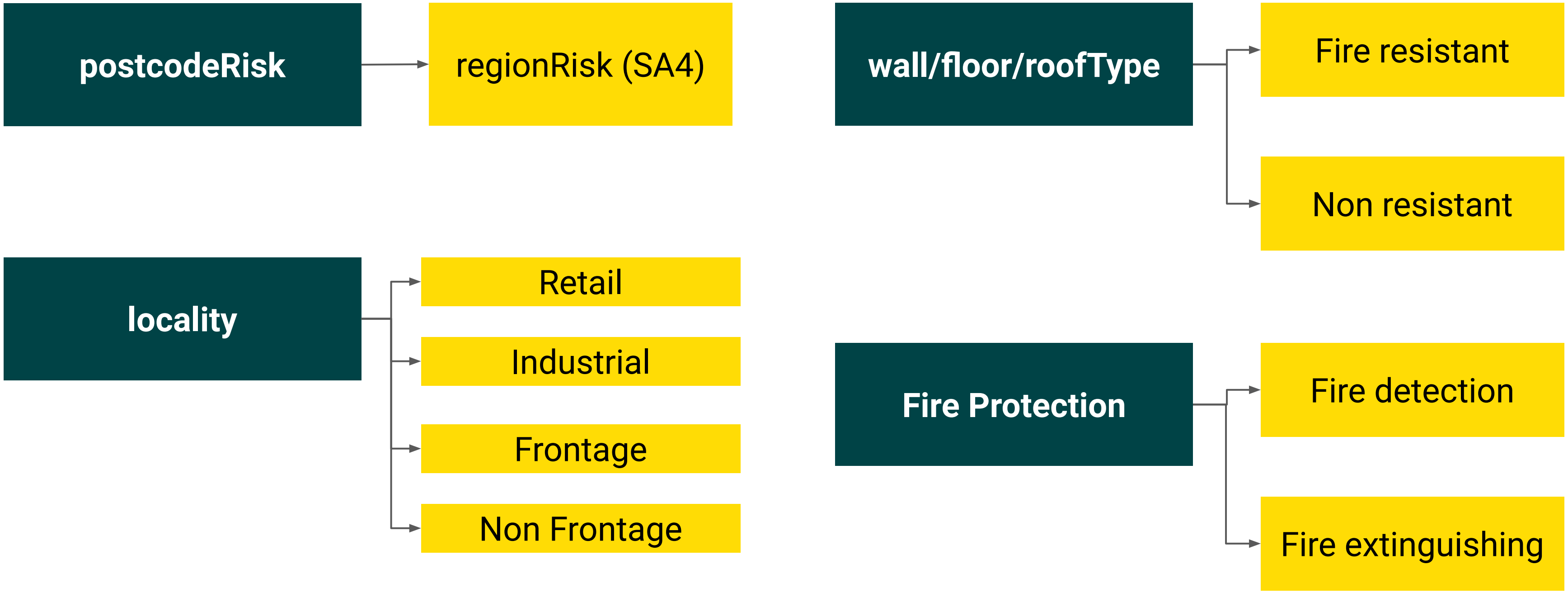
Median total incurred by Statistical Area Level 4



*size of circles indicates number of claims

Data preparation → **EDA** → Occupation grouping → Claims cost models → Limitations & next steps

Features were **grouped or combined** to simplify the dataset



Occupation Grouping

The data was **aggregated by ANZSIC2** code

ANZSIC4...	**ANZSIC2...**
...has **254** values	...has **62** values
Each occupation has on average **26.5** claims	Each occupation has on average **108.4** claims

Although ANZSIC2 has a small sample size, the analysis will be robust if we ***use 5 or less variables (Formann, 1984)***

Aggregation is more robust and is hence preferred

Claim **severity**, **frequency** and **fire proportion** were used to quantify risk

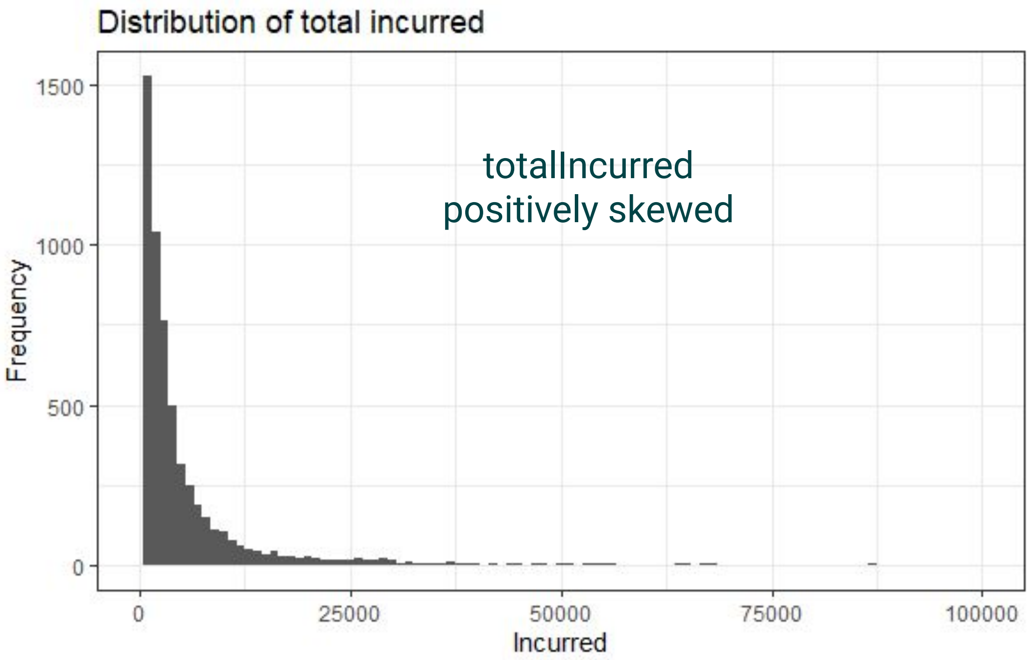
Building insurance risk

- Claim severity
  - Mean of totalIncurred
  - Std dev. of totalIncurred
- Claim frequency
  - Claims per 1000 businesses, w/ ABS

Fire claims showed different behaviour to non-fire claims

- Proportion of fire claims

Variables were **normalised** to minimise the impact of outliers



Take logarithm of totalIncurred

Variables were **normalised** to minimise the impact of outliers

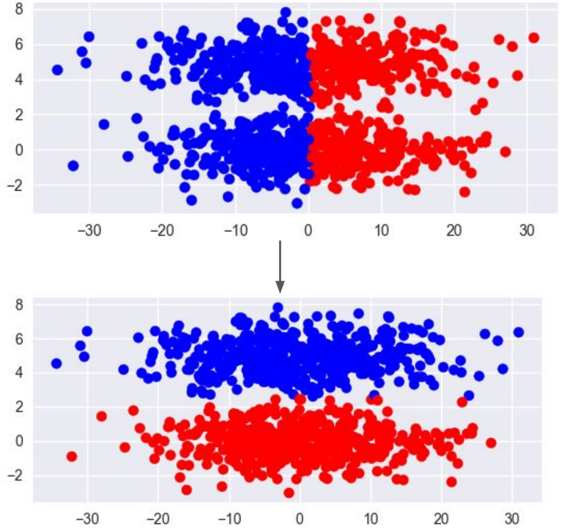
Clustering methods use distance-based measures

Algorithms become sensitive to large values

Min-Max scaler to map values to [0,1]

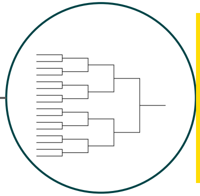
Before normalising

After normalising

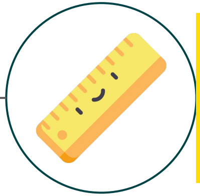


Other models, settings, and variables were considered

Alternative models



Hierarchical clustering: does not naturally capture spherical shape of groups



Gower's distance: does not consider relative riskiness

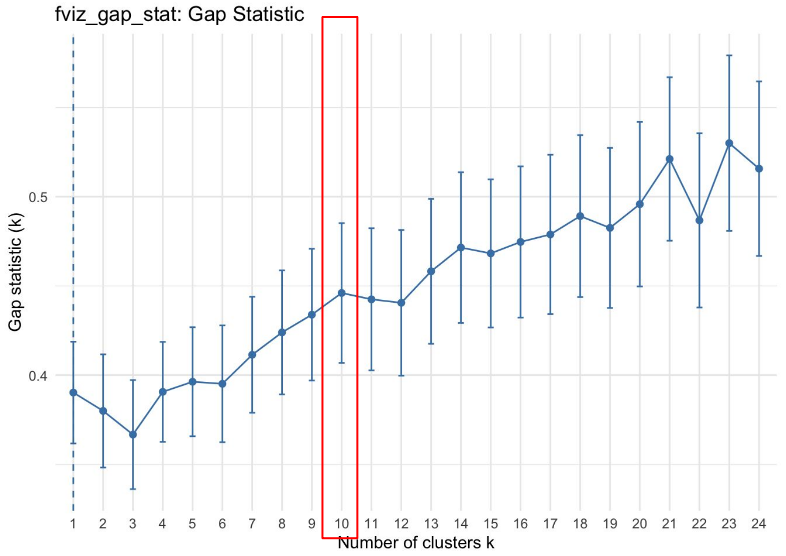
Other models, settings, and variables were considered

Alternative feature engineering

- **Split data** into fire and non-fire claims
- Include **other variables**, e.g. materials
- 3 Aggregate by **ANZSIC3** code

In **K-means clustering**, choosing the right **number of clusters** is important

Gap plot shows how gap between clusters increases with k



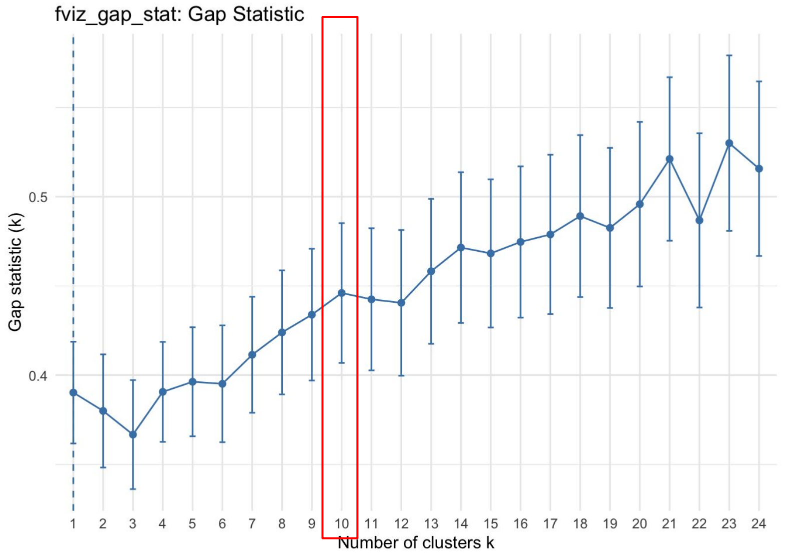
Elbows at 10, 14, 18, 21, 23

Too many clusters lead to overfitting, too many categorical variables

10 clusters chosen

Choosing the right **number of clusters** is important

Gap plot shows how gap between clusters increases with k



Elbows at 10, 14, 18, 21, 23

Too many clusters lead to overfitting, too many categorical variables

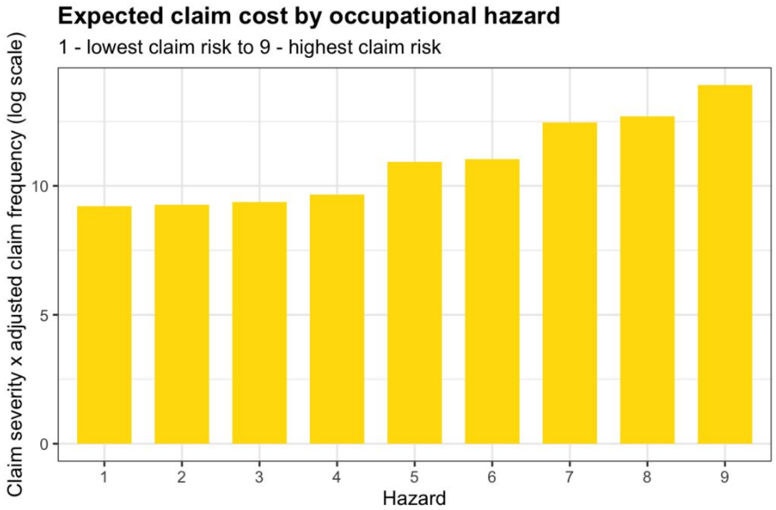
10 clusters chosen

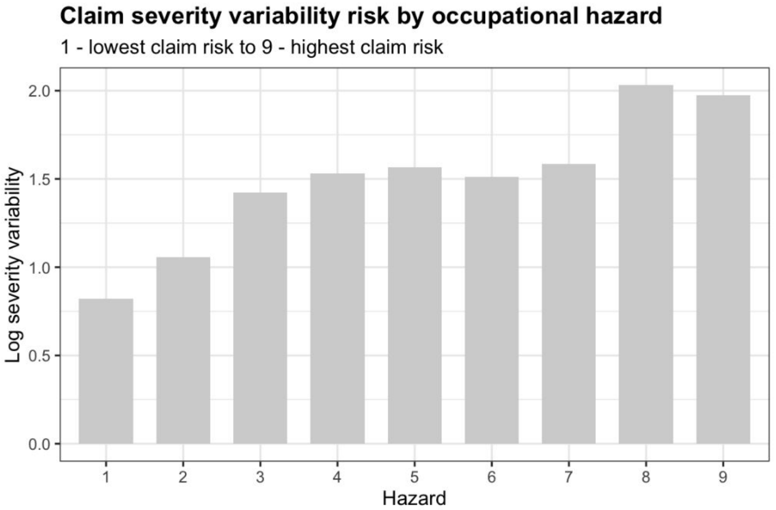
Combine groups 9 and 10

9 distinct groups

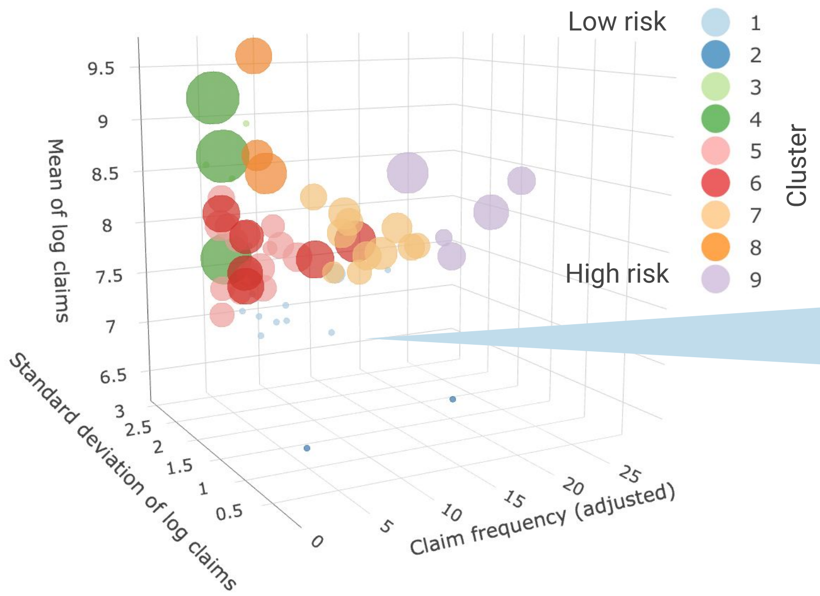
Clusters were **ranked** based on **expected claim size**

Expected claim size ≈ Claim frequency × Claim severity





Education and media businesses are prevalent in the **low-risk** group

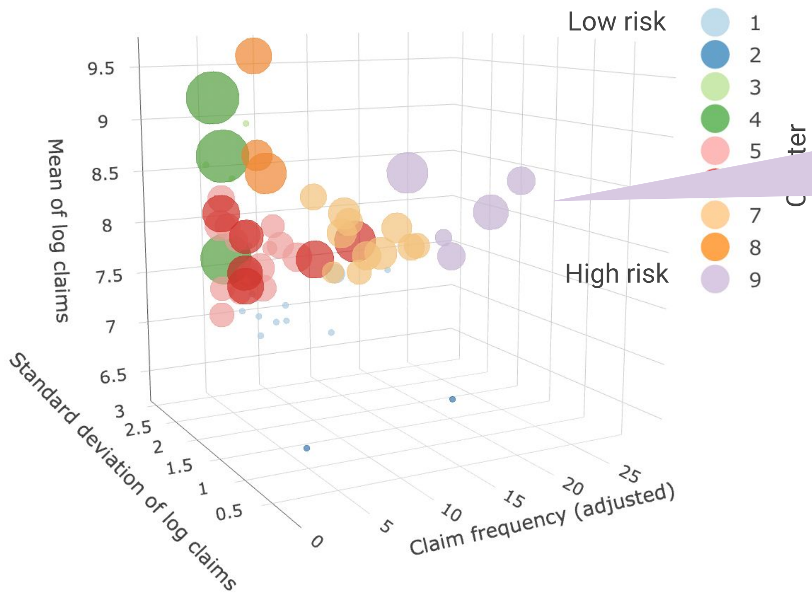


Size of bubble indicates proportion of fire claims

Group 1 includes:

- Education (adult, tertiary)
- Media and arts (broadcasting, motion picture, sound recording, performing arts)
- Construction
- Telecommunications

Retail and hospitality makes up the **riskiest** occupations

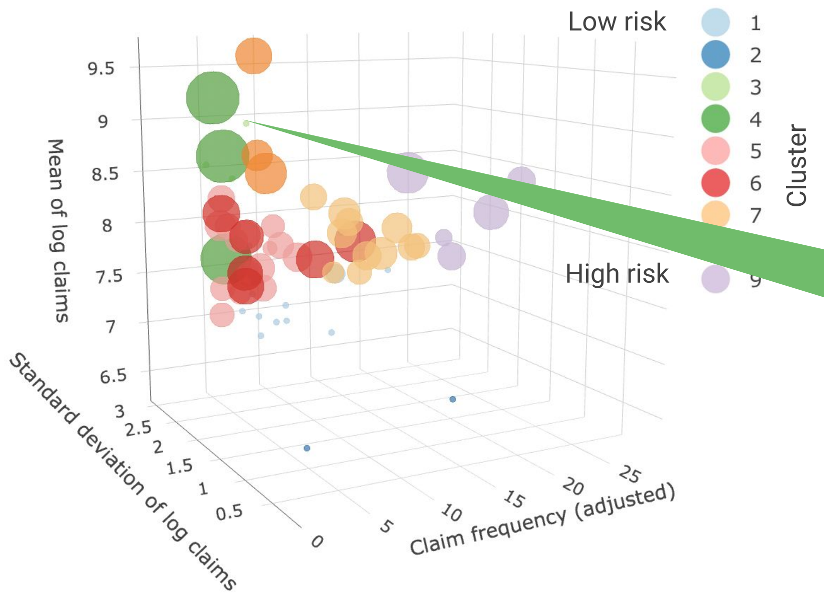


Size of bubble indicates proportion of fire claims

Group 9 includes:

- Fuel Retailing
- Food Retailing
- Food and Beverage Services
- Heritage Activities
- Library and Information Services

Group 4 suffers from large claims, but benefits from a low claim frequency

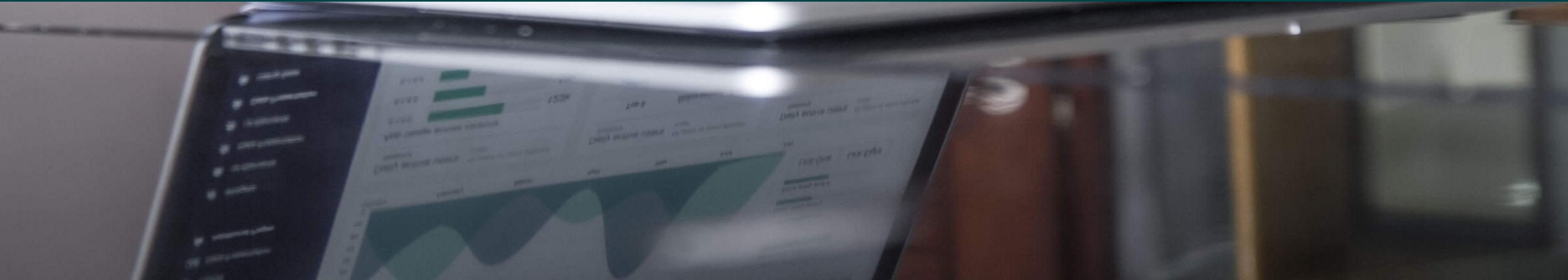


Size of bubble indicates proportion of fire claims

Group 4 includes:

- Agriculture
- Transport support services
- Internet service providers

Claims Cost Models



The data was **split into a training and testing set**

Objective: To accurately model working claims cost for building insurance

The data was randomly split into...

Training set	**Testing set**
80% of data	20% of data
To aid in model tuning and fitting processes	To evaluate **performance** across models

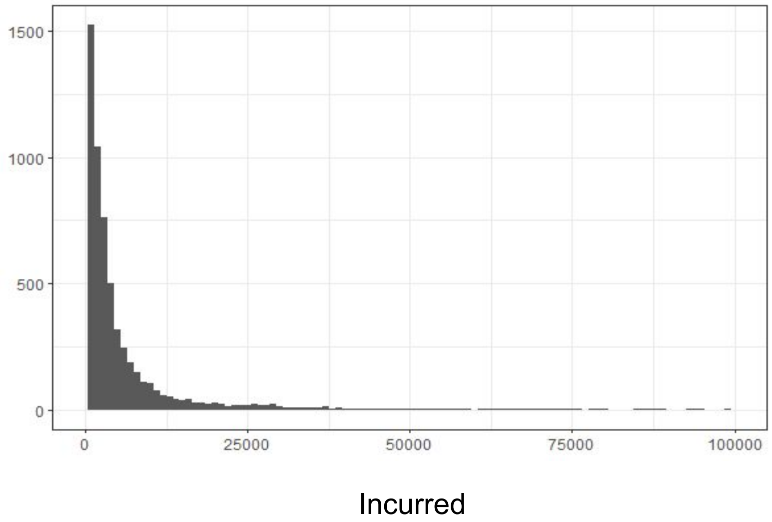
Root mean square error (RMSE)

- Learning objective (for training models)
- Performance metric (for model selection)

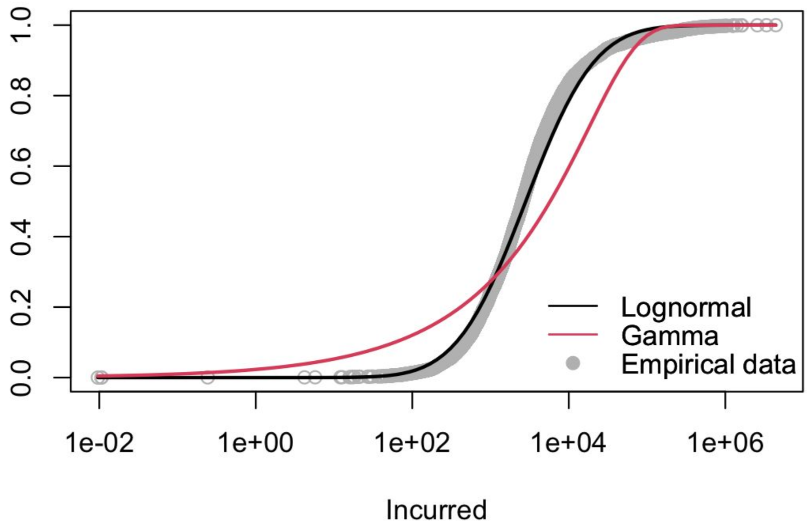
$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y_i} - y_i)^2}{n}}$$

We **fit a normal distribution** to the **log transformed** data to optimise the GLM

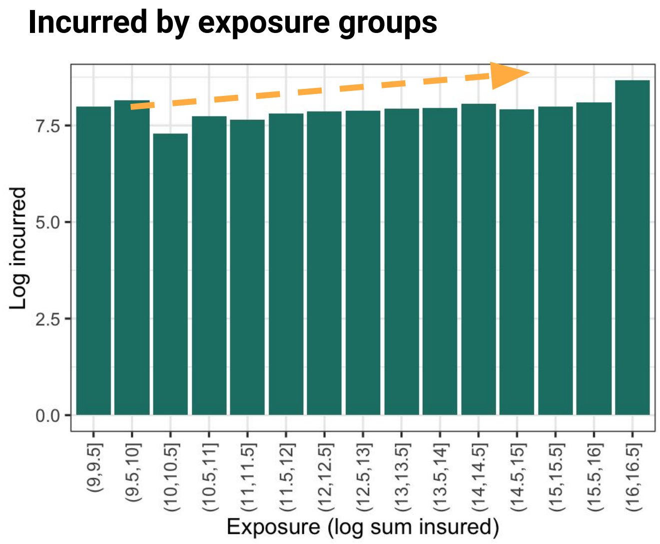
Distribution of total incurred

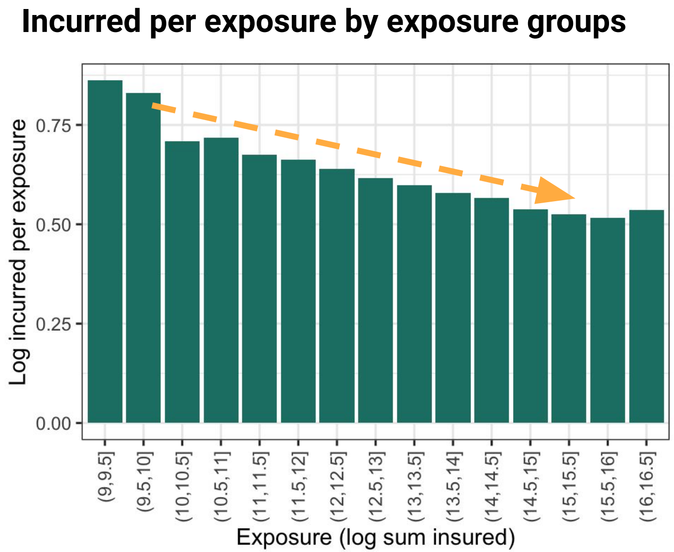


Empirical and fitted parametric CDFs (without predictors)



Sum insured was treated as a predictor





Relationship between incurred and sum insured may not be strictly pro-rata

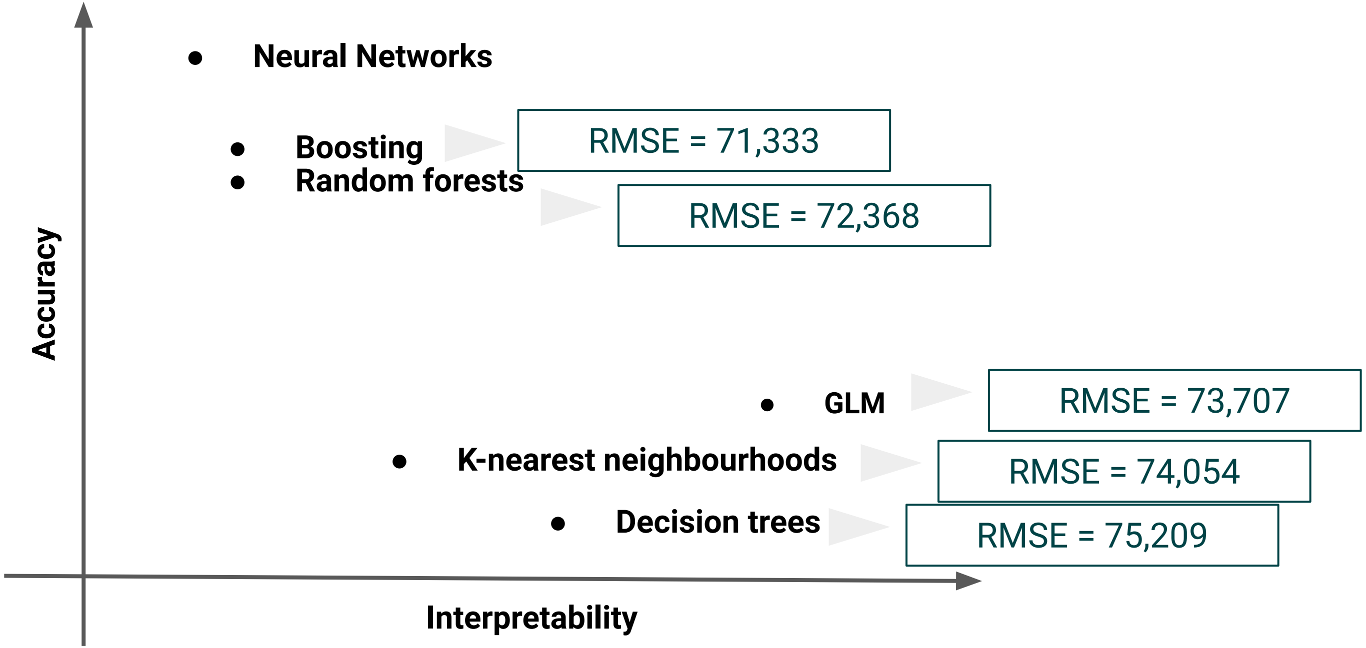
Other **model considerations** were assessed and included to the model

Material types, fire protection, occupations, postcodes	In line with EDA results
Superimposed inflation	yearIncurred as a categorical predictor
Fire/non-fire vs aggregate model	Balanced the lack of data and the significantly different distributions

The models need to be both **accurate and interpretable** to meet business objectives

Business objectives	
Accuracy	**High** - Accurate prediction for claim costs is the basis for competitive and sustainable pricing schemes.
Interpretability	**High** - It is crucial to understand the risk drivers in order to inform product design and risk management, and communicate the insights to other teams and/or management.
Efficiency	**Low** - The dataset is relatively small (thousands of observations x 10-20 features). The models need not to be re-trained frequently.

But it is well known that a **compromise exists** between accuracy and interpretability



Data preparation → EDA → Occupation grouping → **Claims cost models** → Limitations & next steps

To balance these objectives, our team proposes an innovative model: **XGBoosted GLM!**

Baseline GLM as the backbone	**XGBoosting layer**
Captures the overall pattern	Learns remaining model structure in the GLM residuals
Interpretative power from GLM ensures compliance	Prediction accuracy from XGBoost

XGBoosted GLM maintains XGBoost's **accuracy while retaining interpretability** of GLM

Error comparison of GLM, XGBoost, and XGBoosted GLM

RMSE	GLM	XGBoost	XGBoosted GLM	XGBoosted GLM: Improvement over GLM baseline
Fire claims	525,465	505,244	508,709	**3.2%**
Non-fire claims	22,841	22,482	22,456	**1.7%**
Aggregate	73,707	71,333	71,660	**2.8%**

GLM identified **different significant predictors** for fire and non-fire claims

Fire

Variable	Estimated effect	F-statistics	Significance
buildingSI	**1.17e-07**	**8.008**	★★☆
YearIncurred	Multiple factors	1.126	
yearsInsured	−7.38e-03	0.424	
regionRisk	**Multiple factors**	**1.650**	★☆☆
locality	Multiple factors	1.123	
roofType_resist	−6.25e-01	1.429	
wallType_resist	−1.25	1.017	
floorType_resist	4.53e-01	0.307	
fire_detection	**6.50e-01**	**5.695**	★☆☆
fire_extinguishing	3.23e-01	1.327	
hazard	**Multiple factors**	**1.988**	☆☆☆

Significance: ★★★ <0.001; ★★☆ <0.01; ★☆☆ <0.05; ☆☆☆ <0.10

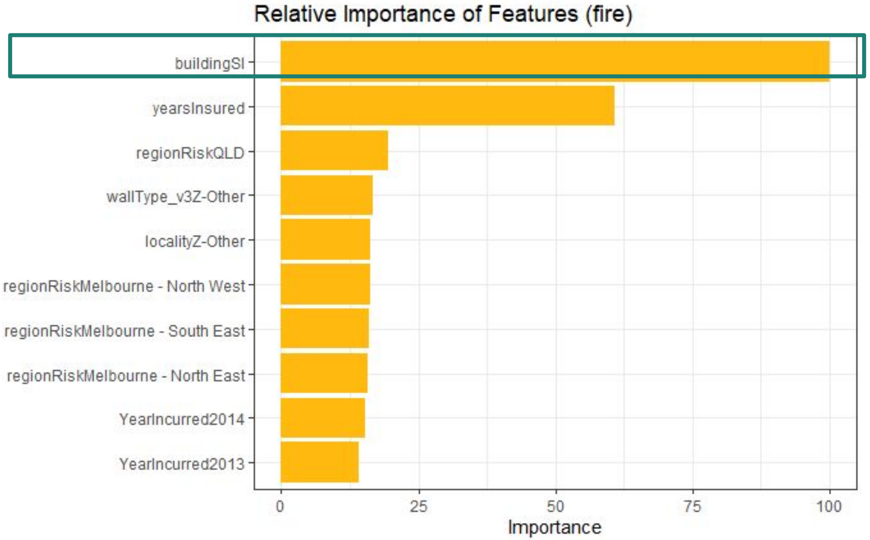
Non-fire

Variable	Estimated effect	F-statistics	Significance
buildingSI	**2.28e-08**	**9.433**	★★☆
YearIncurred	**Multiple factors**	**12.950**	★★★
yearsInsured	**−8.19e-03**	**6.382**	★☆☆
regionRisk	**Multiple factors**	**2.245**	★★★
locality	Multiple factors	1.413	
roofType_resist	−7.76e-02	0.438	
wallType_resist	4.94e-01	0.107	
floorType_resist	**−6.03e-01**	**6.928**	★★★
fire_detection	8.93e-02	0.363	
fire_extinguishing	−5.75e-02	0.868	
hazard	**Multiple factors**	**3.813**	★★★
peril	**Multiple factors**	**18.985**	★★★

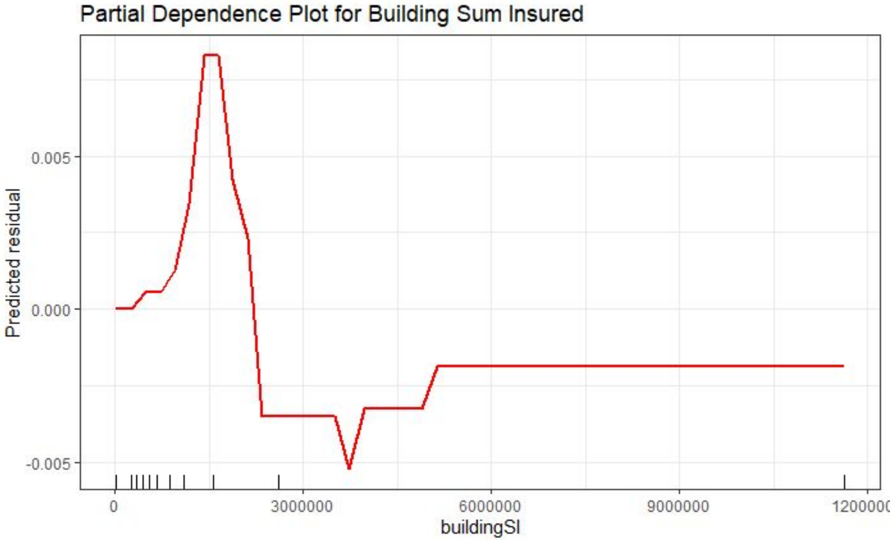
The split models together have an **RMSE of 73,726**, a vast improvement over the previous two models!

XGBoost effectively picked up the **remaining patterns**: fire example

Variable importance

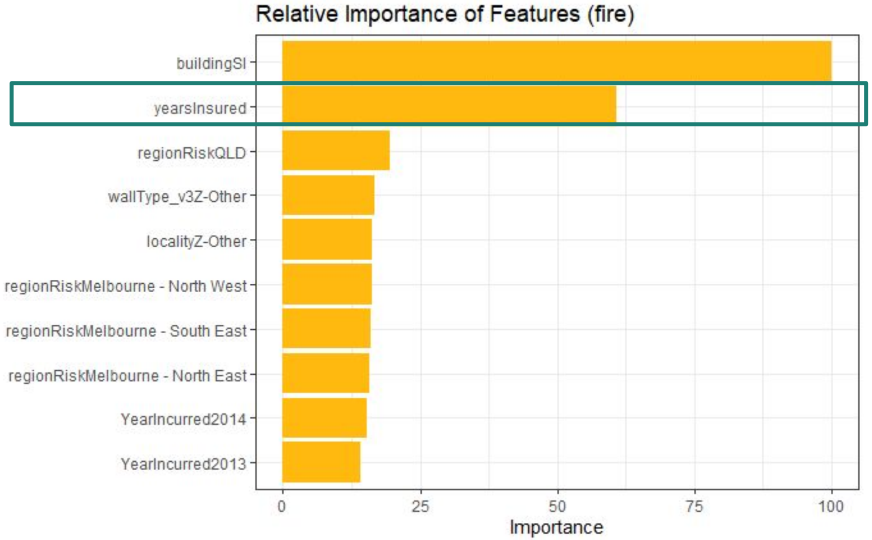


Partial dependence

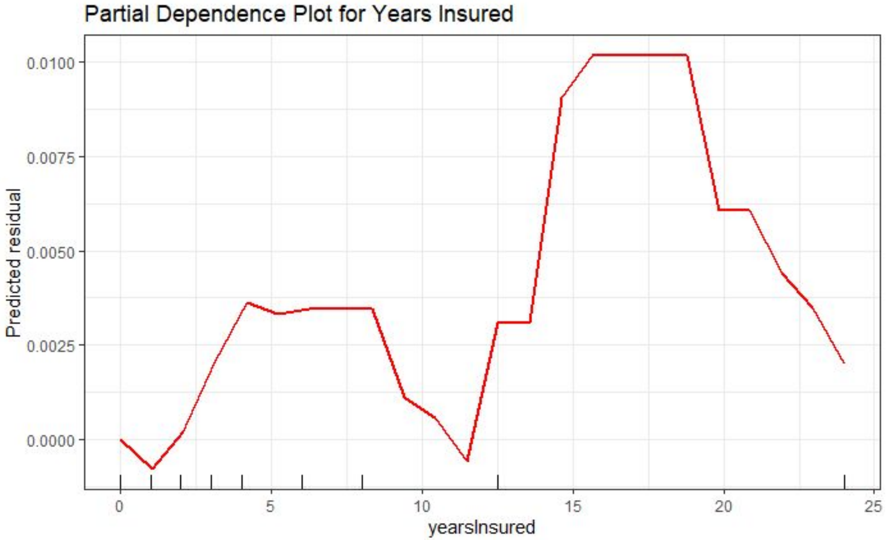


XGBoost effectively picked up the **remaining patterns**: fire example

Variable importance



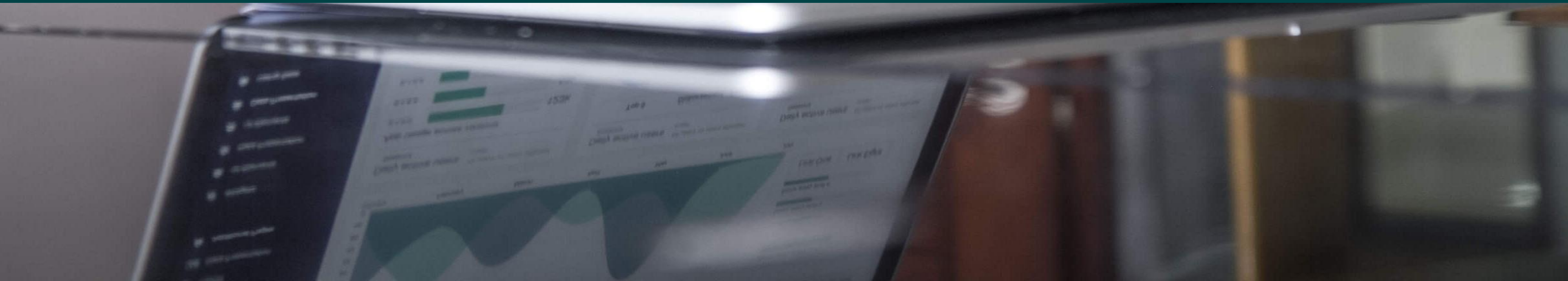
Partial dependence



We make three recommendations to make our model **commercially viable**

Converting models into rating tables	Further experimentation with the hybrid model	Integration with peril classifier (either model or underwriter)
Discriminate customers based on risk drivers identified by GLM (e.g., buildingSI, yearIncurred, occupation)	Consider other parametric models for baseline and/or deep learning techniques for learning residuals	Instead of assuming known perils, we can integrate our model with peril classifier that estimates Pr(fire claim)

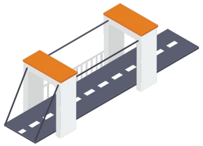
Limitations & Next Steps



Our results inherited any limitations in the data sources

Limitations

1. Relatively small size of data may constrain the performance of ML models
2. Lack of policy information/exposure data



Next Steps

1. Augment with industry and external datasets (e.g. population density)
2. Explore additional models (e.g. credibility models)
3. Linking fire and non-fire models trained on separate datasets