Actuaries Institute.

Whole person. Whole system.

IDSS 2023

12 – 14 November
Hobart

Leveraging unstructured text data to improve a statistical lifetime cost of claim model

Michael McLean, Nikolay Nikolaev

This presentation has been prepared for the Actuaries Institute 2023 Injury and Disability Schemes Seminar. The Institute Council wishes it to be understood that opinions put forward herein are not necessarily those of the Institute and the Council is not responsible for those opinions.

Table of Contents

Introduction

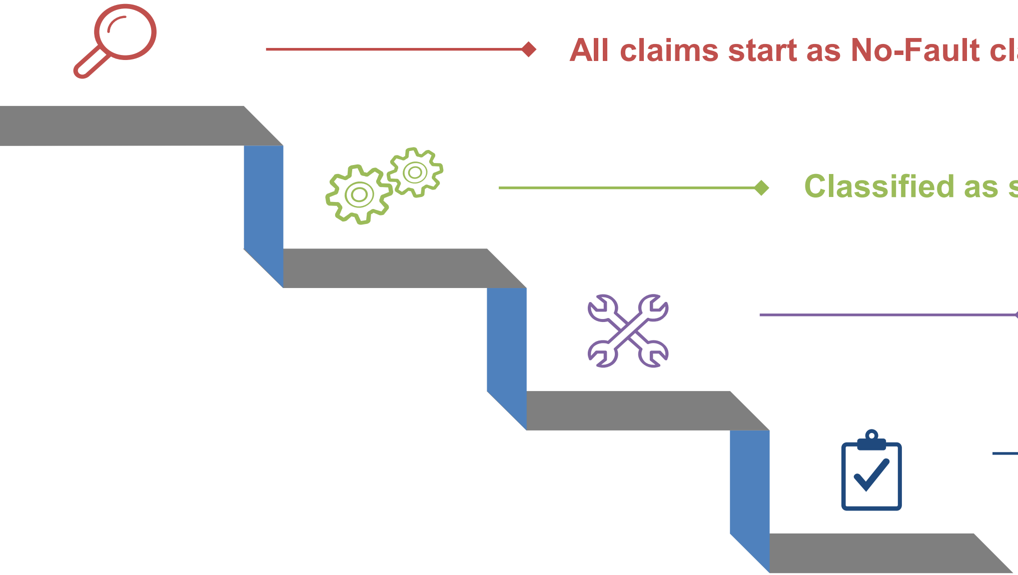
Lifetime Cost of Claim (LCC) model for accident compensation claims

- Case reserves
- Claim management
- Strategic intervention

Our research

- Can unstructured text add value when modelling LCC?
- Test on one component of the LCC model we built for a large Scheme

Lifetime Cost of Claim modelling



All claims start as No-Fault claims

Classified as severe or non-severe

No-Fault claims can become a Common Law claim

Common Law claims

- Model probability of a Common Law lodgement
- Size given lodgement

Predicting the probability of CL lodgement

Data

- Historical claims information from 1995 to 2023Q1
- Observed outcomes of common law lodgements as at 2023Q1

Model

- Snapshot of claims at 9 months from accident date
- GBM to model the relationship between claim info and common law lodgement

Splitting

- Accident years 2009-17 used for Training (70%) and Validation (30%)
- Accident year 2018 used as Out-Of-Time test set

Data

Claims header file
All info relating to claimant, injury

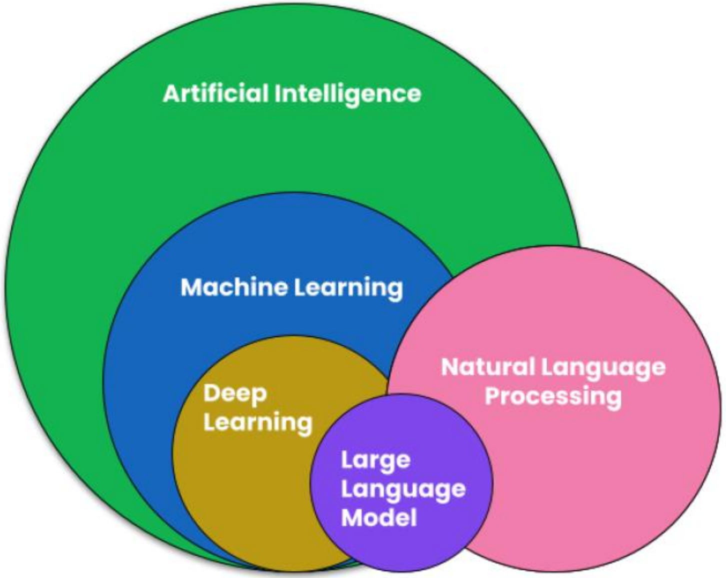
Payments
Transactions payment data

Road safety data
Info on crash, vehicles, drivers involved, drugs, alcohol

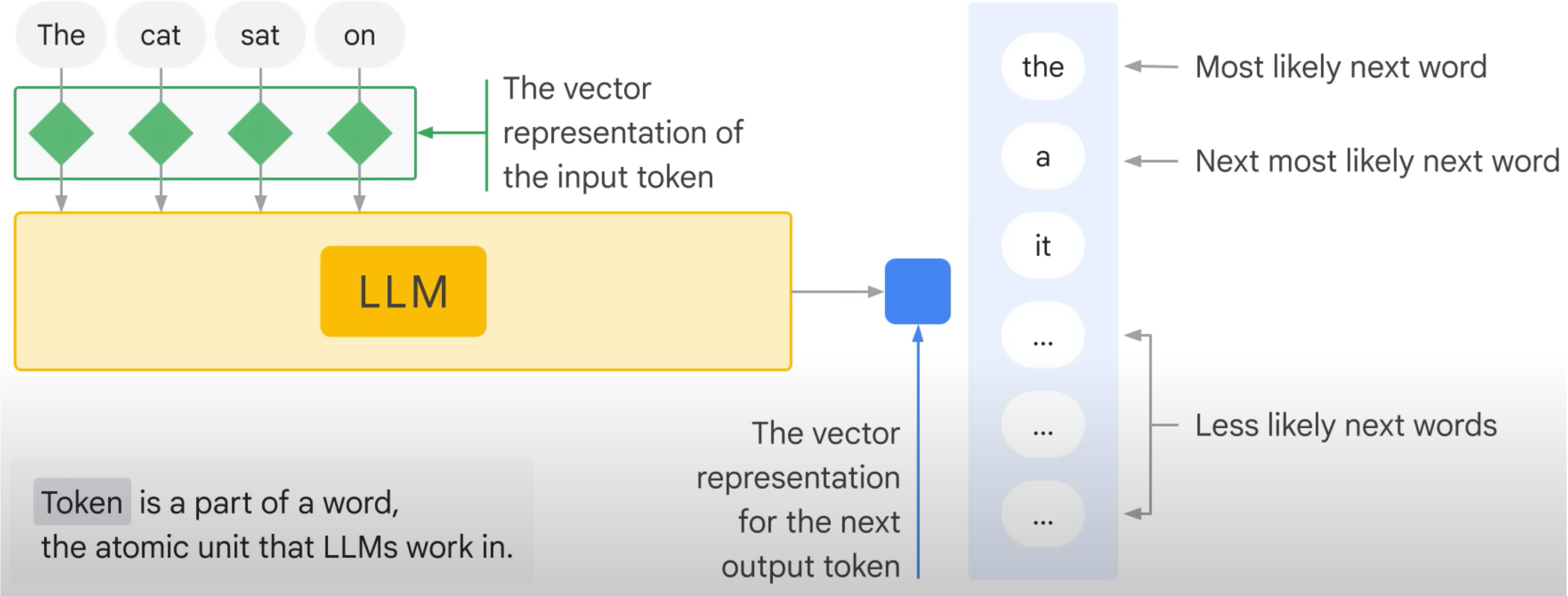
Free form text data
Five distinct types (e.g. case notes, external documents, phone calls)

Census data
SEIFA index, vehicle density, remoteness

Large Language Models



Large Language Models



Sentence embeddings

Request letter, cheque and authority received from Solicitors

Requesting an x-ray and a MRI

Appointment item for ultrasound

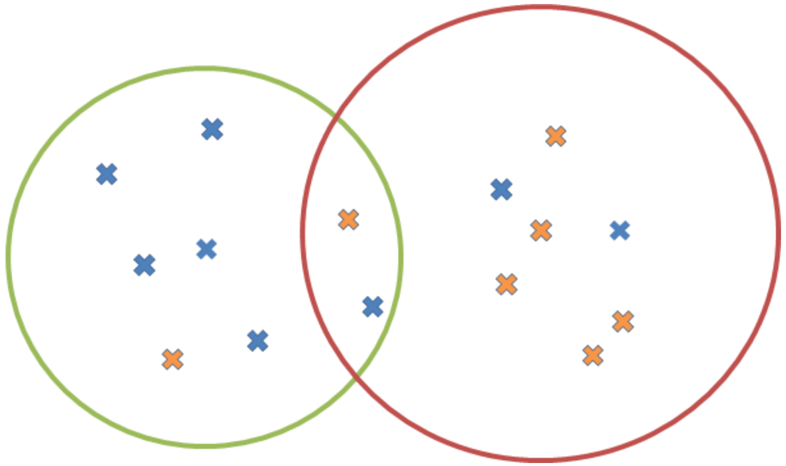
Sentence Transformer (using LLM)

[0.5, -0.3, 0.2, 0.7,…]

[-0.3, 0.5, 0.3, -0.8,…]

[-0.2, 0.2, 0.7, -0.4,…]

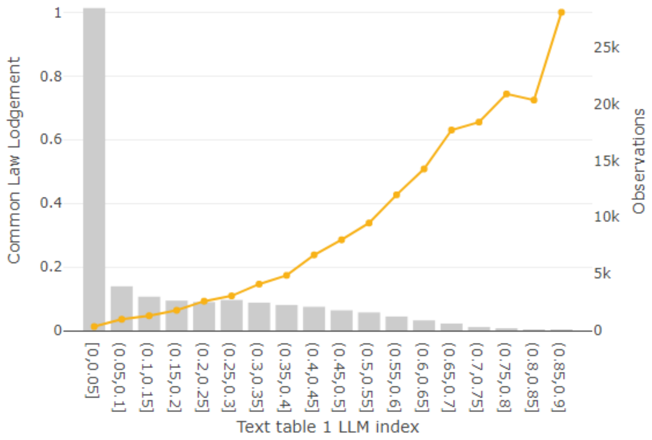
Use LLM for finding similar claims

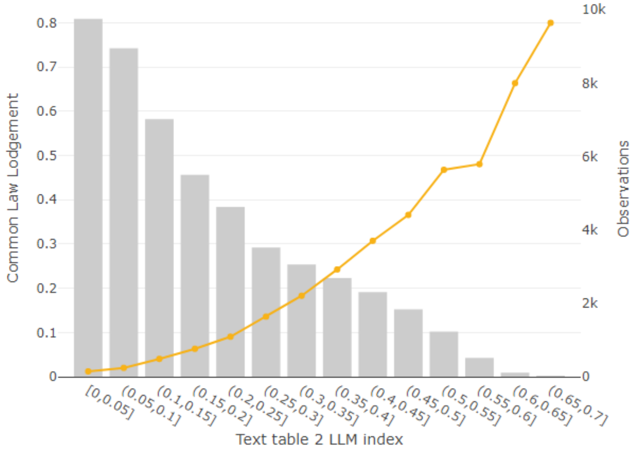


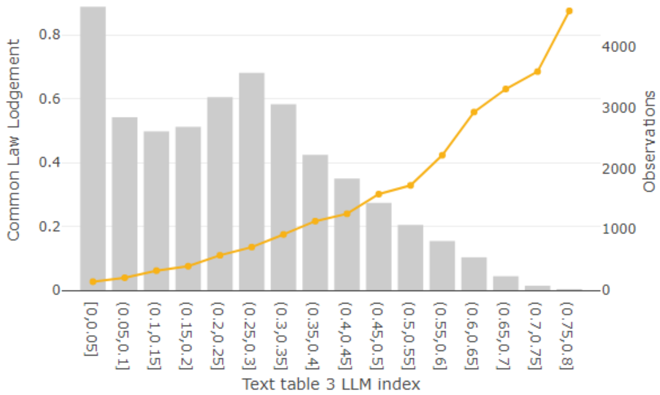
Intuition: Can historical claims with similar claim descriptions help predict the probability of a common law claim?

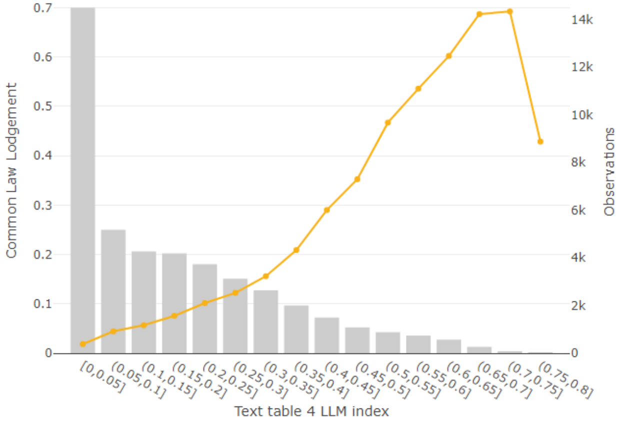
- Embed claim texts into numerical vectors that capture the meaning using open source LLMs (mpnet, gte)
- Average embeddings per claim and text type to represent claim as a whole
- For each claim
  - Find other claims that have similar text descriptions
  - Derive a score based on how many neighbours lodged a CL
- Use the derived score as a predictor in the modelling

LLM based text score









Text-based neighbours strongly predict CL probability

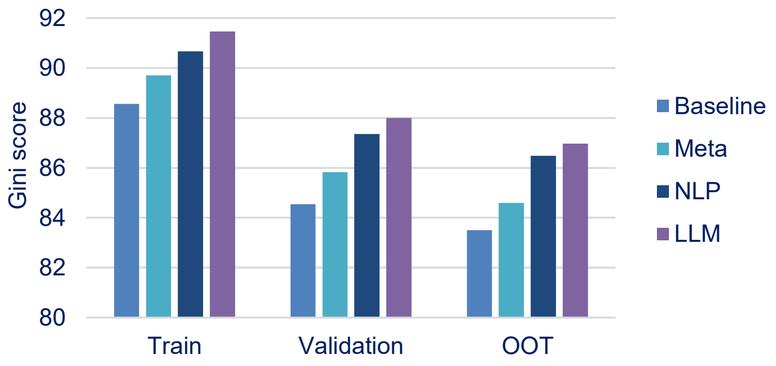
Experimental setup

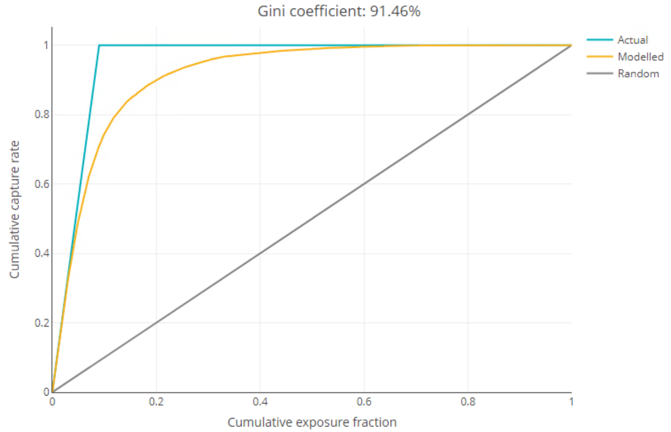
	Features			
Models	*Structured data*	*Text meta-data*	*NLP*	*Text data embeddings*
#1: Baseline				
#2: Meta				
#3: NLP				
#4: LLM				

Evaluation: Gini coefficient

Results

- Model performance improves with each new set of features
- Results demonstrate the strong predictive power of claim text information
- Using LLM features results in the strongest model

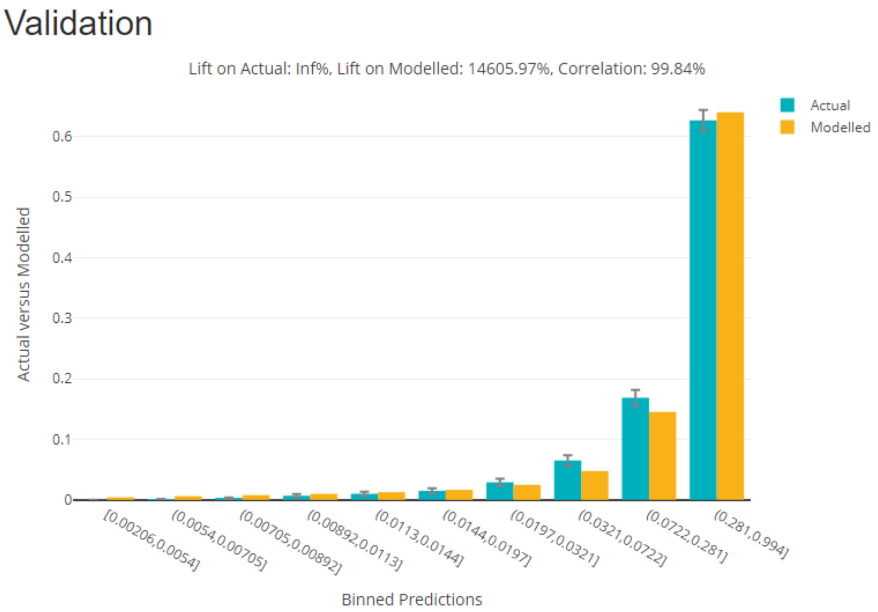




% Common Law Lodgements Identified					
Validation	% of claims	Baseline	Meta	NLP	LLM
	2%	20%	20%	20%	20%
	5%	42%	43%	45%	45%
	10%	64%	64%	68%	68%
	25%	87%	89%	89%	91%

Risk differentiation

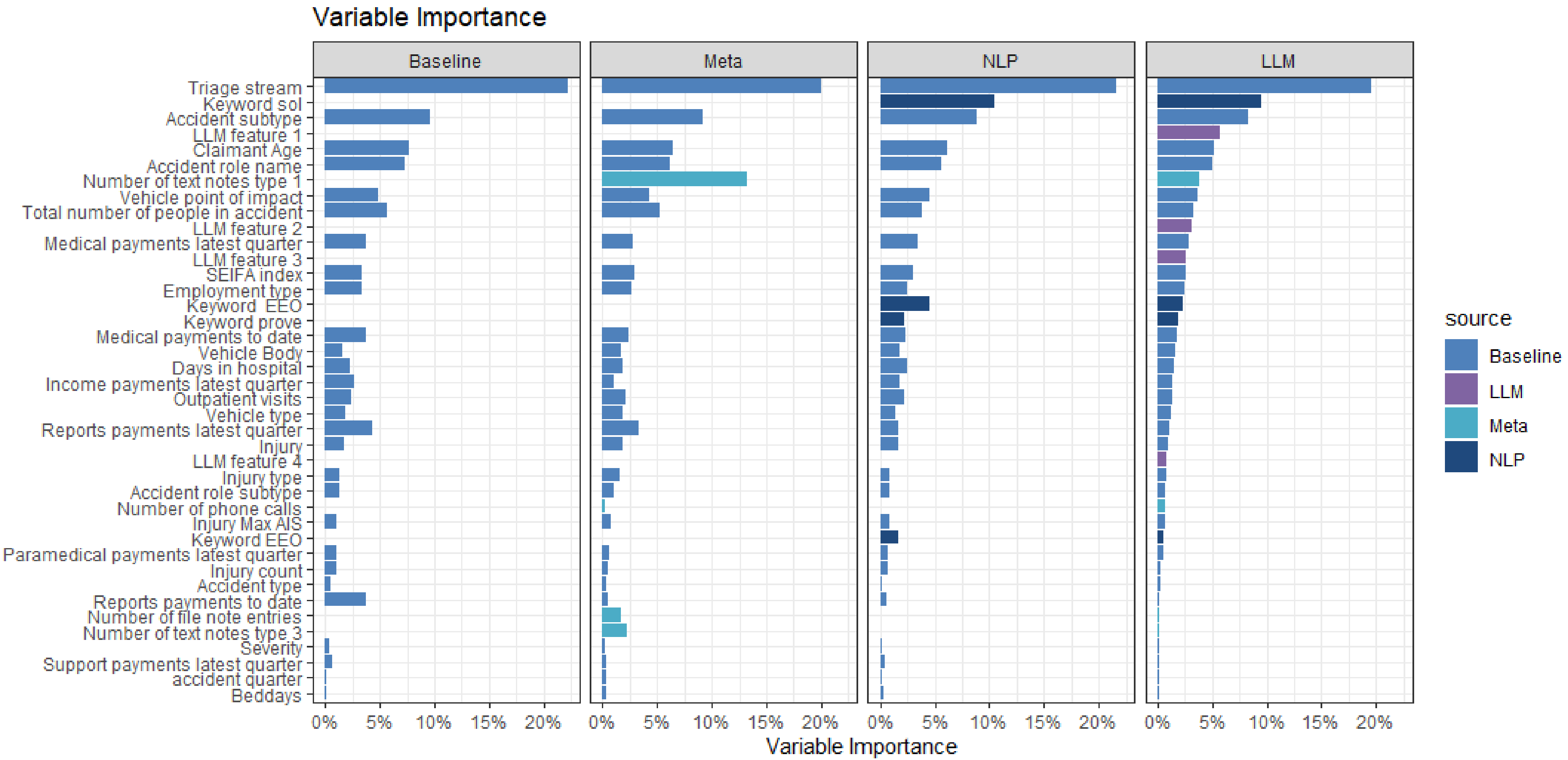
- LLM based model achieves the best results and validates well on unseen data
- Well calibrated predictions for the probability of a claim to become a common law claim
- Strong risk differentiation achieving a high model lift



Feature Importance

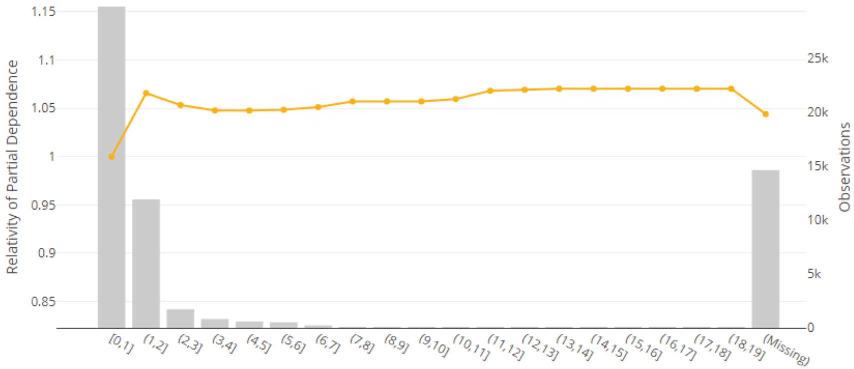
IDSS 2023

Actuaries Institute.

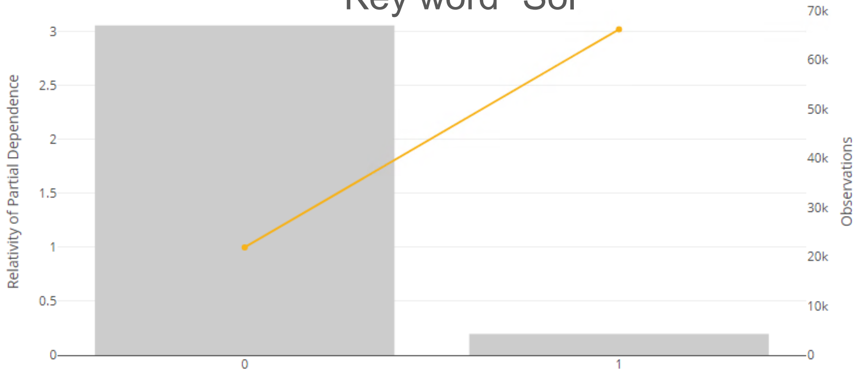


Partial Dependence

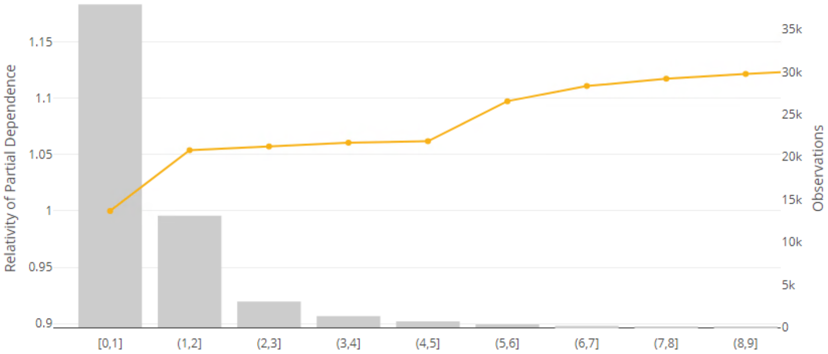
LLM Feature 1



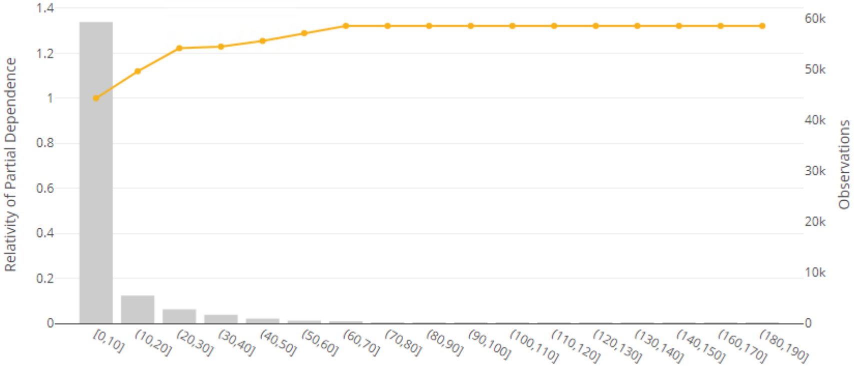
Key word “Sol”



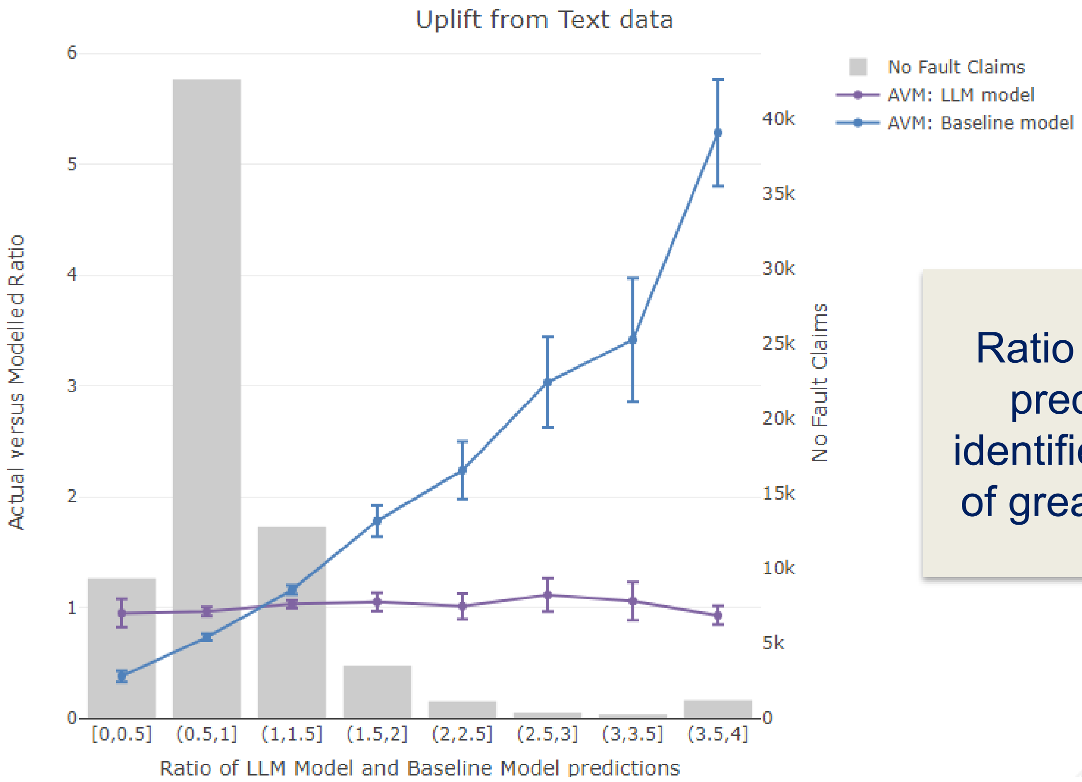
LLM Feature 2



Count of external document notes



Model Comparison

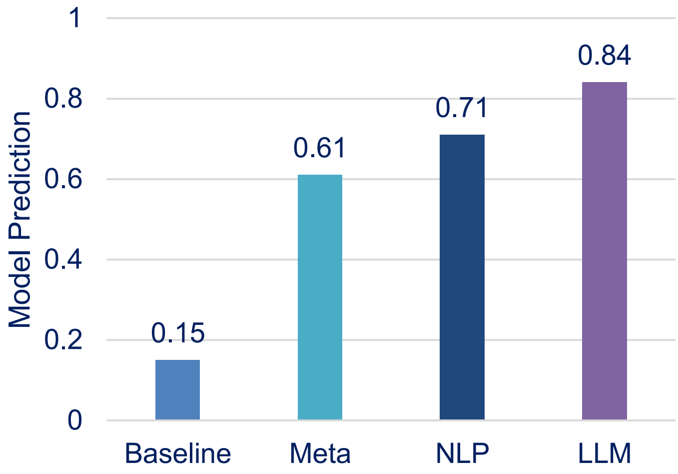


Ratio of model predictions identifies regions of greatest uplift

Example – Claim 1

Outcome: Common Law Lodgement

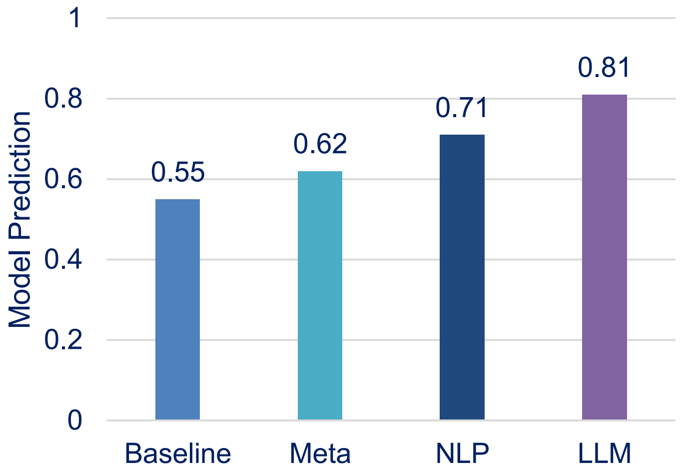
Injury detail	Fractures – Limb
Role	Passenger/Pillion
Total sum to date	$24k
Days in hospital	9
Age	19
Text records counts	0-20
LLM 10-NN scores	0.7
Common keywords	TAXI, Support, Form, General, approval



Example – Claim 2

Outcome: Common Law Lodgement

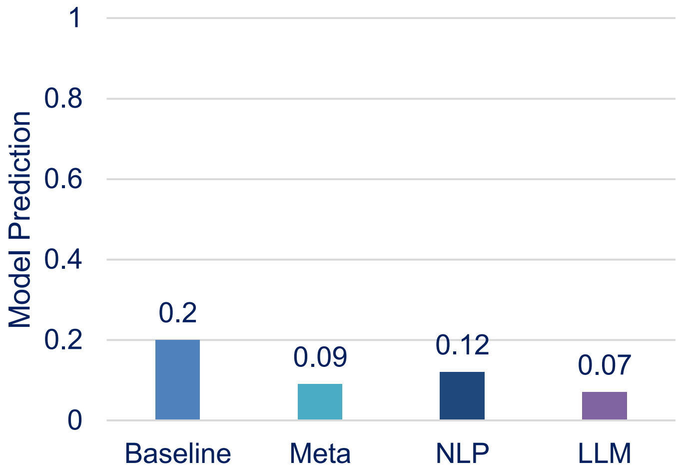
Injury detail	Brain Injury (Mild) / Head Injury (Ill defined)
Role	Bicyclist
Total sum to date	$109k
Days in hospital	23
Age	49
Text records counts	70+
LLM 10-NN scores	0.9
Common keywords	Prove, Uploaded, Received, Correspondence, Benefits, Care, Dr, Report, LOE, Form, Rehabilitation, Support, Treating, Certificate, Services, RTW, letter, Practitioner, employer, Income



Example – Claim 3

Outcome: No Common Law Lodgement

Injury detail	Fractures – Limb
Role	Bicyclist
Total sum to date	$83k
Days in hospital	11
Age	51
Text records counts	0-7
LLM 10-NN scores	0.1
NLP	Prove
Common keywords	Police, Report, Incident, Confidential



Future work

Utilize more unstructured text data – documents, e-forms, medical reports etc.

Fine-tuning the large language model or method of aggregation of the embeddings

Use commercial private versions of more powerful models (e.g. chat GPT) instead of the smaller open source LLMs

Apply to other components of lifetime cost of claims model (e.g. cost of No-Fault claims)

Conclusions

- Unstructured text data significantly improves compensation claims predictive model performance
- Schemes, insurers and claims service providers have a valuable asset which can be utilized at scale with potential significant improvements in claim management and reserving
- Large Language Models are a powerful tool for extracting signal out of unstructured data
- LLM field is emerging and improving rapidly - expect better results in the future with advances in technology

Actuaries Institute.

Thank you

IDSS 2023

12 – 14 November
Hobart