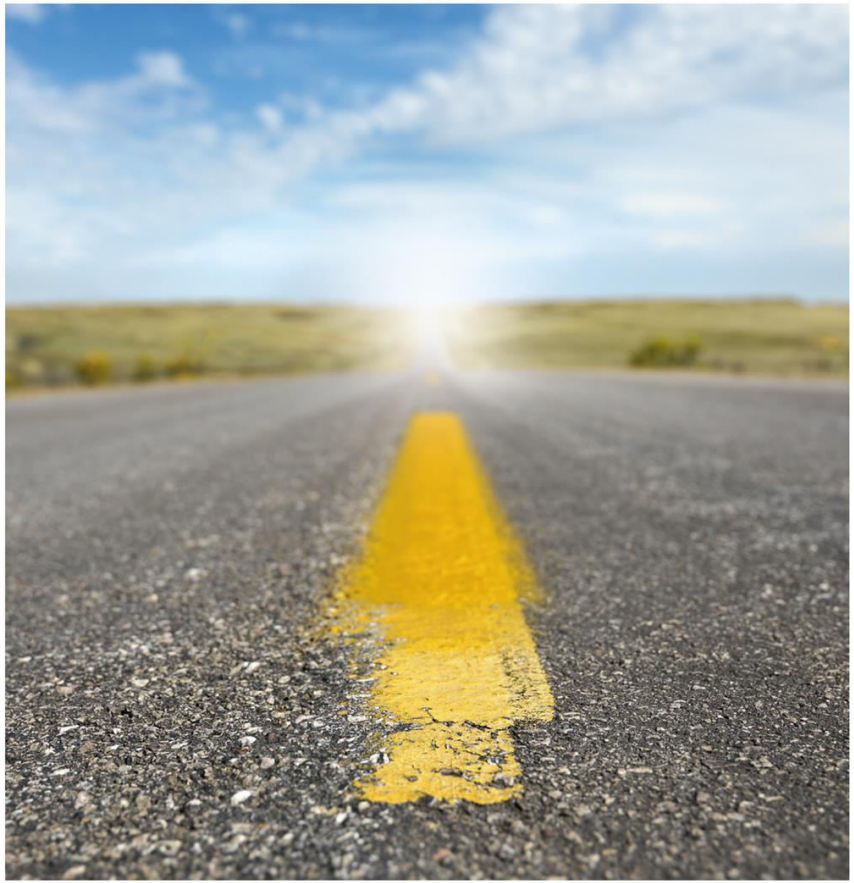
Injury Schemes Seminar

Road to Recovery

Actuaries Institute

8-10 November 2015 • Hilton • Adelaide



Injury Schemes Seminar
Road to Recovery
8-10 November 2015 • Hilton • Adelaide

Actuaries Institute

Analytics-assisted triage of workers' compensation claims

Ivan Lebedev, Inna Kolyshkina, Marcus Brownlow, and Colin Khoo

This presentation has been prepared for the Actuaries Institute 2015 Injury 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.

Overview

- Motivation
- Business context
- Methodology
- Toolbox review
- Results
- Key learnings

Motivation

- Share our learnings
- Discuss the difficulties
- Encourage the audience to try using advanced data analytics techniques for their own problems

BUSINESS CONTEXT

Business context

- Pilot project for ReturnToWorkSA to develop data analytics capability.
- Workers' compensation claims:
  - Entitlement to income support, medical costs, etc.
  - Entitlement continues until recovery (under the new Act this is limited to 2+1 years)
  - Claim duration is the major cost driver
  - Ability to identify high risk claims is important for targeting case management effort

Aims of the study

- Predict whether or not a claim will stay for more than 1 year on benefits from the information known at 13 weeks from lodgement
- Identify the most important predictors of claim duration and derive interpretable business rules

Modelling dataset

- Worker information (gender, occupation, income estimate, age at the time of injury, etc)
- Employer and Industry data (remuneration, industry classification, etc)
- Injury information (injury date, nature of injury, body part, etc)
- Previous claims history of the same claimant (for example, the number and cost of prior claims)
- Altogether, there were ~ 200 variables

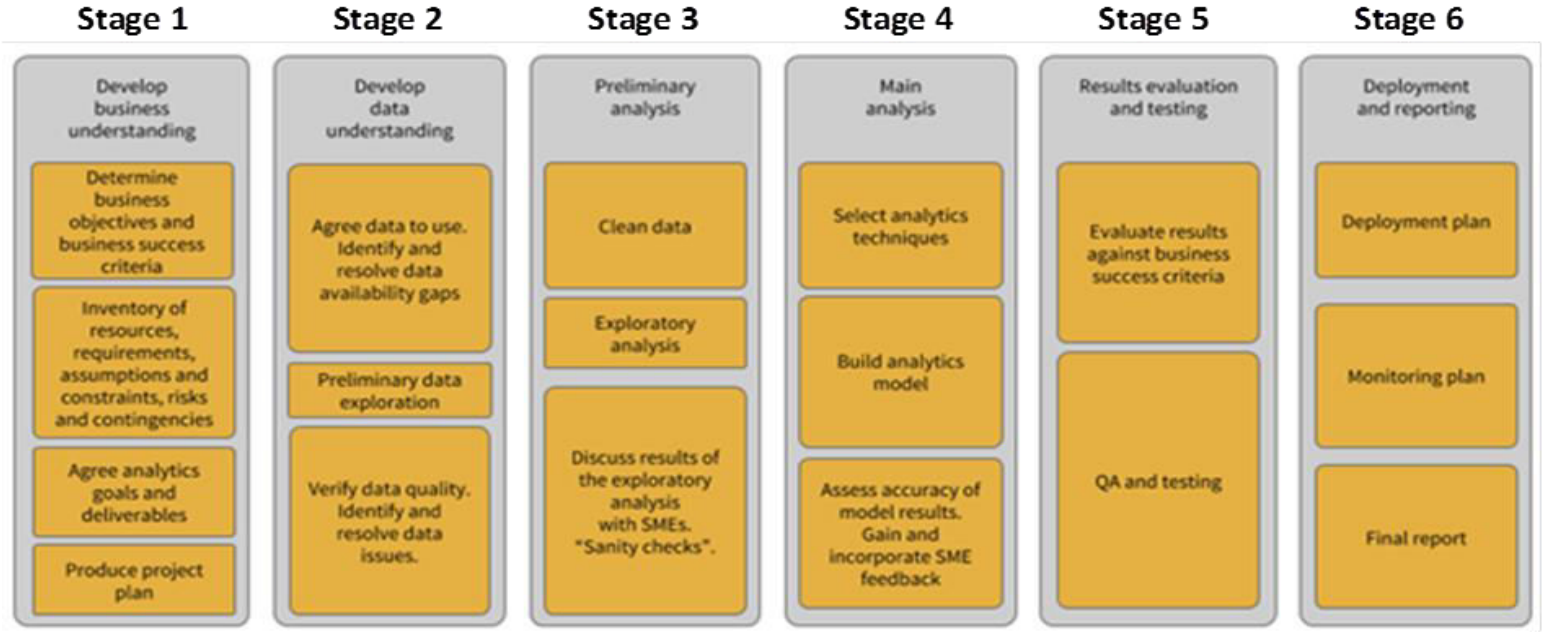
METHODOLOGY

Why advanced data analytics?

- To assess the business potential for ReturnToWorkSA

- Traditional methods (e.g. GLMs) would be inefficient for this task because of
  - Complex data – outliers, missing values, some highly categorical fields
  - A large pool of potential predictors
  - The aim to derive interpretable business rules

Main steps of the data analytics process



Main steps

- Data review, audit and pre-processing

- Predictive modelling Stage 1. Evaluate and refine predictive potential of the data.
  - Evaluate predictive value of the data
  - Enrich data. Identify factors that can be added to improve predictive value of the data
  - Select the most important predictors out of many available variables

  Tools: Random Forests, GBM, LASSO

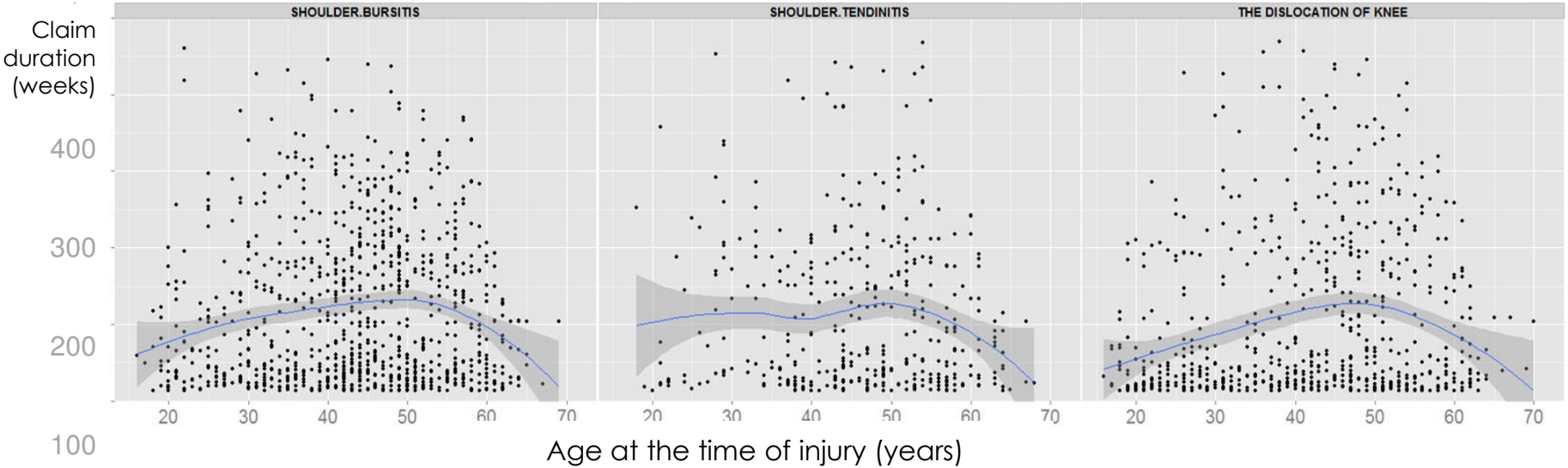
- Predictive modelling Stage 2. Build the final model

  Tools: Decision Trees + Association Rules

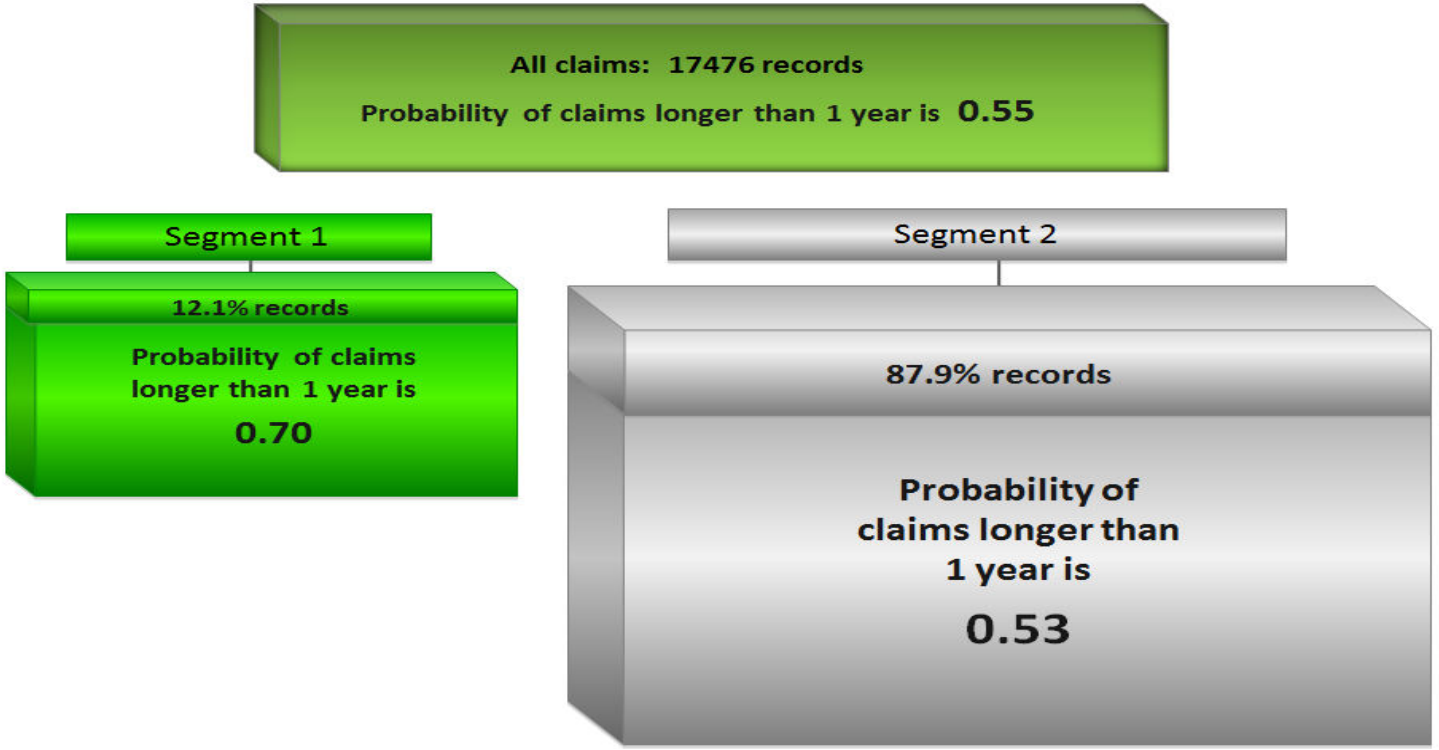
PREDICTIVE MODELLING STAGE 1

Extent of random variation in the data.

Claim duration vs. injury age for selected injury nature and location



Segmentation of claims by probability of duration longer than 1 year (based on the initially provided data)



Data enrichment

- Our approach:
  - external research
  - RTWSA subject matter expert advice
- New predictors added:
  - lag between injury occurrence and claim lodgment
  - information on the treatment received (for example, type of providers visited, number of visits, provider specialty)
  - information on potent opioid prescription
  - details of the claim that occurred immediately prior to the current claim:
    - injury nature and location of that claim
    - whether it was related to the same or similar injury as the current claim.

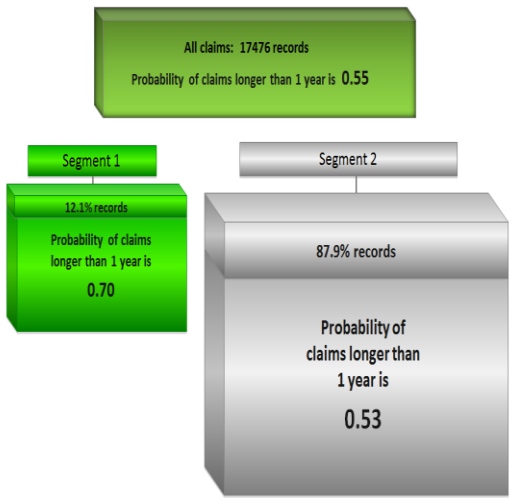
PREDICTIVE MODELLING STAGE 2

Key results

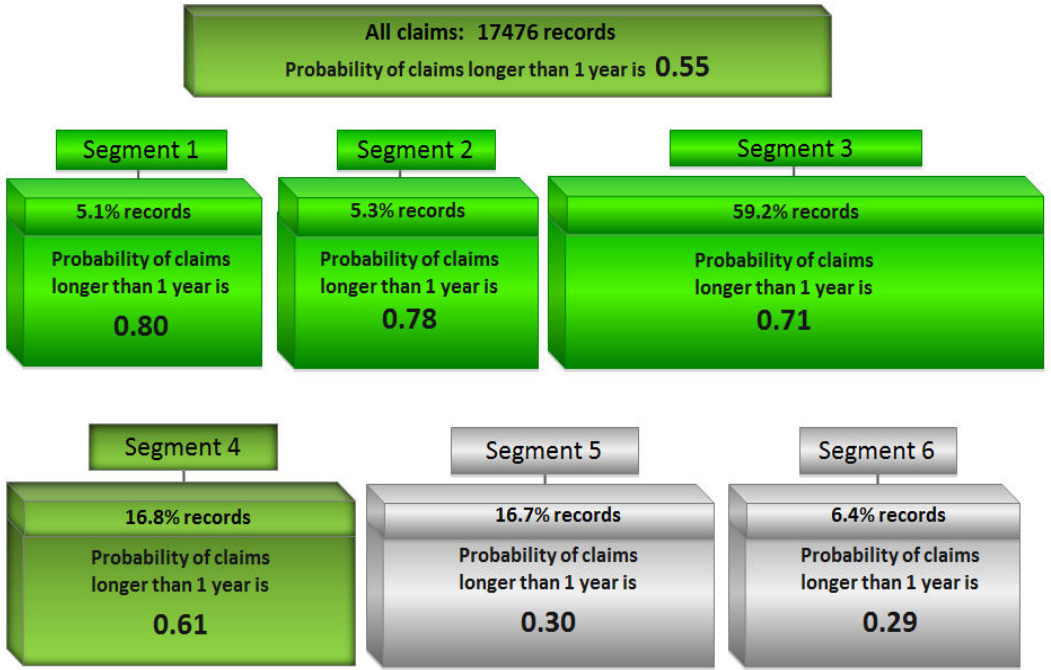
- Derived business rules to identify claims that have high risk of long duration. The rules are based on the information available at the first 13 weeks of the claim.
- Some important predictors of duration are:
  - Injury nature and location
  - Worker age, occupation, industry
  - Lag between injury and its report.
  - Claimant's previous history with RTWSA
  - Service provider (doctors, physiotherapists etc.) identity and specialty

Enriching the data significantly improved predictive outcome

Segmentation of claims by probability of duration longer than 1 year
(based on the initially provided data)



Segmentation of claims by probability of duration longer than 1 year



Business rules are intuitive and easy to use

- The segments are described by business rules that are easily interpretable by the business.
- The business rules can be used to identify new claims that have high probability of long duration as early as possible.
- From the technical point of view, the business rules can be used to score claims efficiently as they are easily expressed as SQL code.

For example, rules for claims that have 80% probability of lasting 1 year or longer:

NATURE AND LOCATION OF INJURY = LOWER BACK DISC DISPLACEMENT OR NECK DISC DISPLACEMENT OR ...

AND

MOST FREQUENTLY VISITED PROVIDER SPECIALTY = DIAGNOSTIC RADIOLOGY OR GENERAL SURGERY OR PSYCHOLOGY OR...

AND ...

TOOLBOX OVERVIEW

Random Forests

- Key idea
  - Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
- Strengths: delivers maximum predictive value from the data
- Weaknesses: hard to interpret
- Ease of use? no

Stochastic Gradient Boosting

- Key idea
  - The method produces a prediction model in the form of an ensemble of decision trees. It builds the model in a stage-wise fashion and generalizes them by allowing optimization of a loss function
- Strengths: delivers maximum predictive value from the data
- Weaknesses: hard to interpret
- Ease of use? no

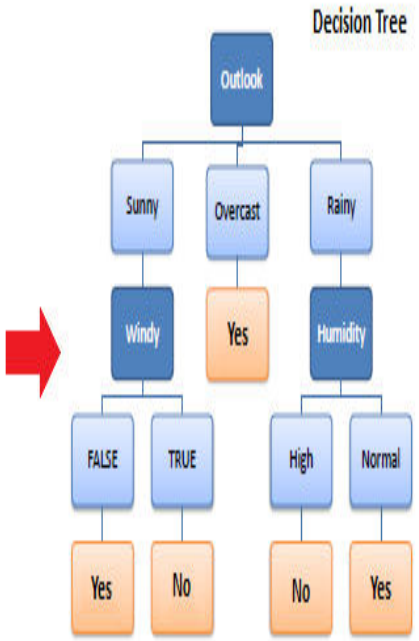
LASSO

- Key idea
  - LASSO (Least Absolute Shrinkage and Selection Operator) is a regression method that involves penalizing the absolute size of the coefficients. The larger the penalty applied, the further estimates are shrunk towards zero. This is convenient when dealing with many potential predictors or with highly correlated predictors.
- Strengths: quick, delivers good predictive value from the data, output looks similar to the traditional regression output
- Weaknesses: results can be somewhat affected by strong multicollinearity
- Ease of use is moderate

Decision Trees

- Key idea
  - Decision tree breaks down a dataset into incrementally smaller and smaller subsets based on minimisation of a statistical criterion (e.g. Gini or entropy)
- Strengths: interpretable, quick
- Weaknesses: over-fitting, coarseness, poor representation of linear structure
- Ease of use - yes

Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



Further information on the tools/methods

There are many books and articles that describe the methods we applied.

The quickest to access is the classical textbook "Elements of Statistical Learning" by the Stanford University professors Tibshirani, Friedman and Hastie who were involved in creation of most of the methods listed in the talk. The book is available online:

http://www-stat.stanford.edu/~tibs/ElemStatLearn/