Institute and Faculty of Actuaries

C5 First Steps in Machine Learning for Individual Claims Reserving

Bob Gore, Grant Thornton
Jon Read, NFU Mutual

12 September 2019

Agenda

- A brief overview of machine learning
- The Techniques used for Large Claims Reserving
- The motivation and benefit for starting the project
- The data set and data mining techniques
- The modelling approach
- Results
- Conclusions from the Project

Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

What exactly *is* ‘machine learning’?

12 September 2019

What is machine learning?

- Machine learning is a data analytics technique spanning a range of algorithms covering computational and statistical methods to “learn” directly from data without relying on any predefined or assumed relationships

Main differences from statistics

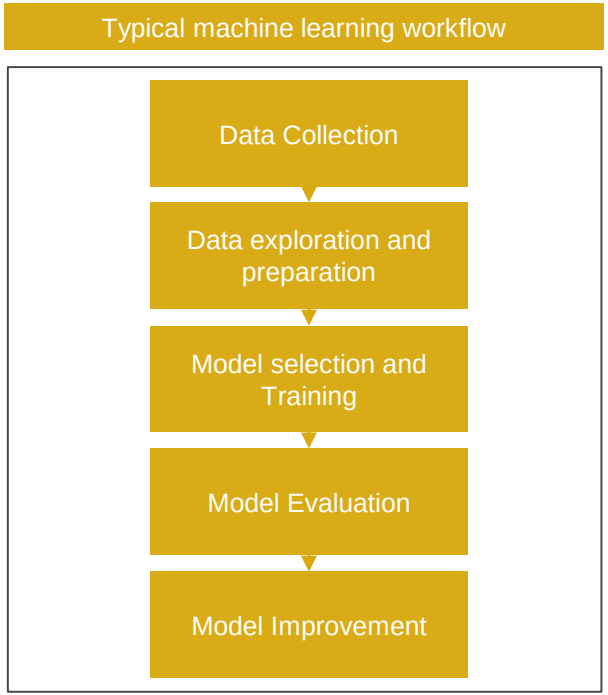
- Greater emphasis on optimization, performance and prediction over basic inference
- Requires no prior assumptions about the underlying relationships between the variables included in the analysed data
- So less emphasis on data collection, statistical properties of any derived estimators, and the underlying distribution
- Algorithms are often applied to high dimensional datasets.

Main benefits of machine learning

- Algorithms learn and can develop from examples, rather than them being programmed for a specific outcome, given a particular input
- Supervised learning is where algorithms are trained via data on examples of the correct output for a certain problem to establish prediction or classification.

Institute and Faculty of Actuaries

Machine Learning: the Basic Workflow



1. **Data Collection -** Gathering the data to be used for training and developing the ML model. This data contains the correct output for every example provided.

2. **Data exploration -** The quality of any machine learning model is heavily dependent on the quality of the data it uses. It is important in this phase to clean the data and make available data features which are likely to beneficial to the model whilst at the same time eliminating unnecessary or incomplete data.

3. **Model selection and training -** Here a suitable ML algorithm is selected for the particular problem to be solved. The nature of the problem influences model selection. Model is trained on a training set of the overall data to establish the form of the model.

4. **Model evaluation** – It is important to evaluate the predictive power of a model from its learning experience using a separate test set of data held back from the original data set. The model is evaluated on agreed quality measures relevant to the problem.

5. **Model improvement** – In order to develop the model further different strategies can be employed including selection of different types of model. Additional data may be sought or may be mined from existing data to make it more accessible to the model e.g. unstructured text data.

Institute and Faculty of Actuaries

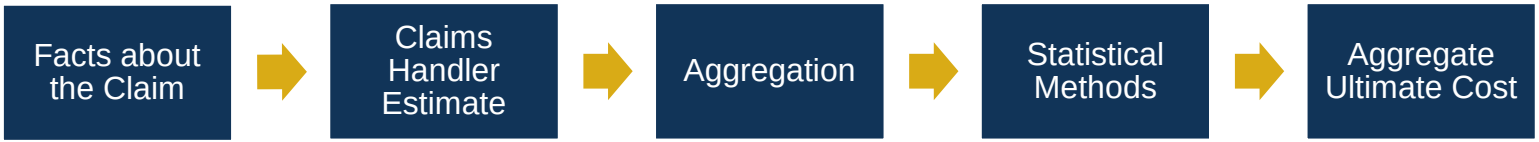
Institute and Faculty of Actuaries

The Techniques used for Large Claims Reserving

12 September 2019

Large Injury Claims Reserving

- Chain Ladder Methods remain the Industry Standard, with known limitations:
  - Reliance on estimates set by claims handlers
  - Claims are aggregated and assumed to be homogenous
  - Historical patterns of development assumed appropriate for modelling the future
  - Results calculated at an aggregate level.



Institute and Faculty of Actuaries

The Motivation and Benefit for Starting the Project

12 September 2019

Improving Large Claims Reserving

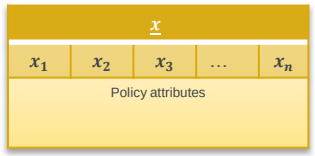
- With enough data, could we build a model to address these limitations?
  - Remove judgements made by claims handlers
  - Remove assumption that claims are homogenous
  - Reduce reliance on appropriateness of historical patterns
  - Calculate results at individual claim level.
- Challenge (as always!) is having enough good quality data.

Institute and Faculty of Actuaries

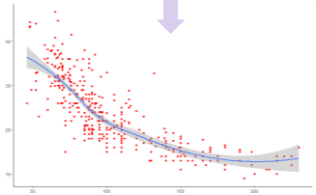
Project objective

- Build a model to estimate ultimate cost of individual RBNS large injury claims
- Make use of all data recorded by claims handlers
- Test value of qualitative / unstructured / incomplete data, as well as quantitative data
- As a first step, ignore any time dependency – i.e. latest data regardless of stage of development
- Prioritise accuracy over interpretability
- Deliver a model which could be easily updated and refreshed when new data emerges.

High level process flow



Pre-processed Claims Data



Modelling & Insights (via R code)

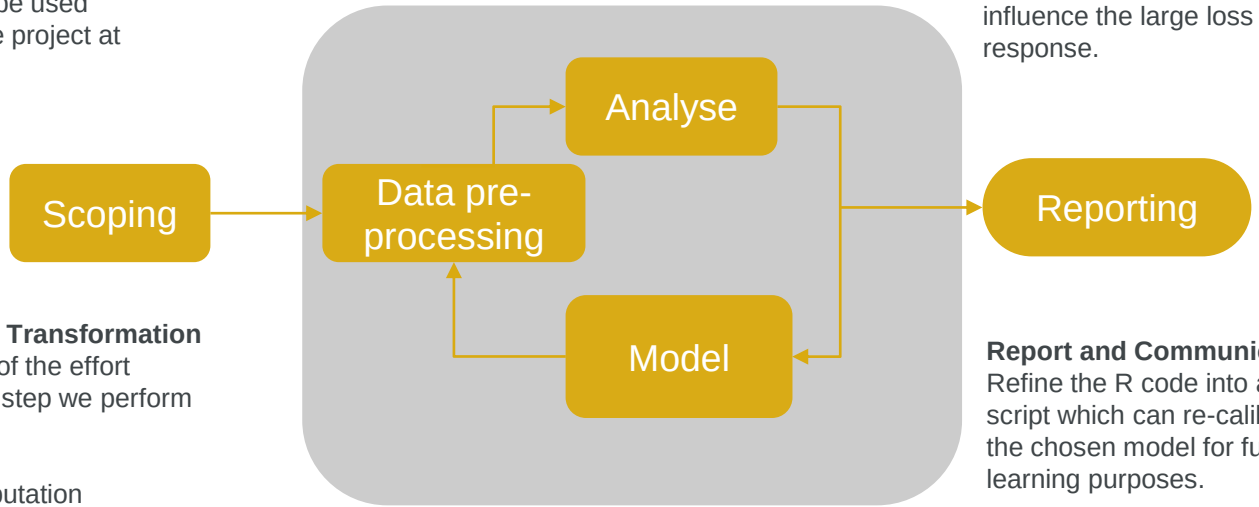
Model estimates of ultimate settlement (Production flow)

Final Machine Learning Model (Training flow)

Institute and Faculty of Actuaries

Model building approach

Define the data and aims
Clarify the data to be used and the aims of the project at the outset.



Analyse and Refine
Perform subset selection methods in order to identify the most statistically significant variables which influence the large loss response.

Data Cleaning and Transformation
This is where most of the effort resides. During this step we perform tasks such as:

- Missing value imputation
- Re-structuring the existing dataset
- Creating new variables based on existing variables.

Model Training and Testing
In order to select the best model, amongst all considerations, it is necessary to tune model parameters on a training set using cross-validation and subsequently test model accuracy on a separate randomly sampled test set, according to the relevant error statistic

Report and Communicate
Refine the R code into a script which can re-calibrate the chosen model for future learning purposes.

Report on findings

Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

The data set and data mining techniques

12 September 2019

Quote

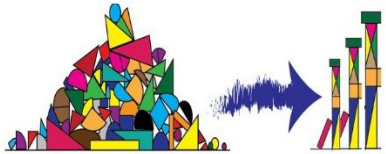
"Data scientists spend 60% of their time on cleaning and organising data. Collecting data sets comes second at 19% of their time meaning data scientists spend around 80% of their time in preparing and managing data for analysis"

Institute and Faculty of Actuaries

Key tasks

- What pre-processing tasks must we perform before modelling?

Variable selection



Feature engineering

Handling of categorical variables

Missing value imputation

Data type conversion (e.g. ‘Strings’ to ‘Dates’)

Days from 1st LL Indicator to HOC - </>30 days

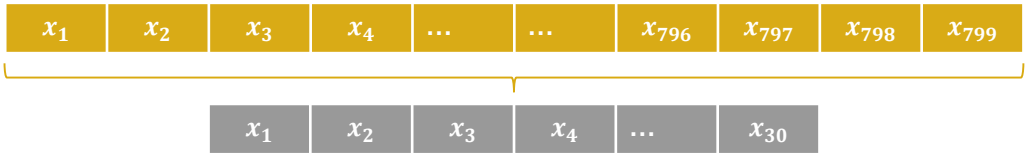
LL_to_HOC

Variable name abbreviations

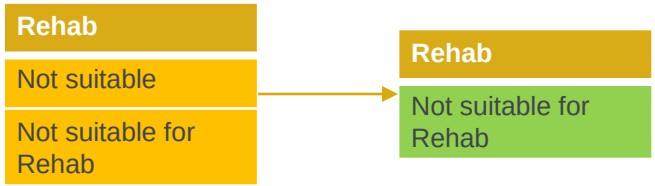
Institute and Faculty of Actuaries

Pre-processing: overview of key steps

1. Variable Selection: reduction of the initial set of c.800 variables to the most important set of variables



2. Pre-processing: this involves tasks such as the restructuring of categorical variables and missing value imputation



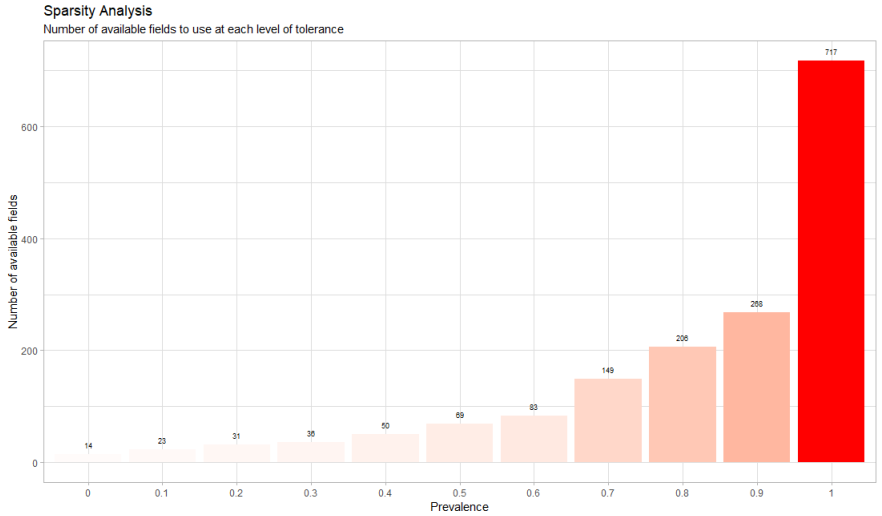
3. Feature Engineering: This involves using existing variables to create more useful features for the learning process – an example is shown to the right-hand side.

…	…	Loss Date	Intimation Date	Time to Report (Days)
…	…	22/11/2010	23/11/2010	1
…	…	16/06/2012	22/06/2012	6

Institute and Faculty of Actuaries

Variable Selection

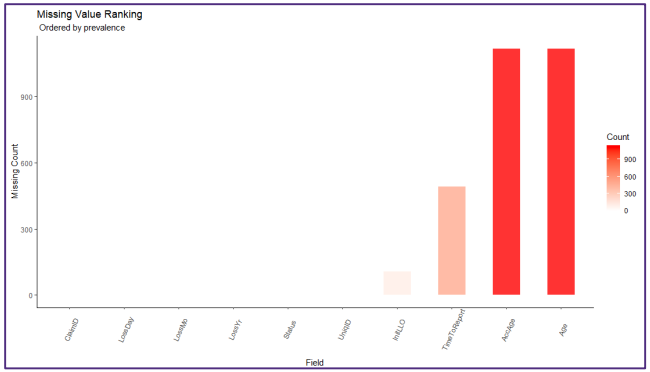
- Common feature is missing or incomplete data
- Need to ensure that the algorithm is not adversely affected by missing value imputation.

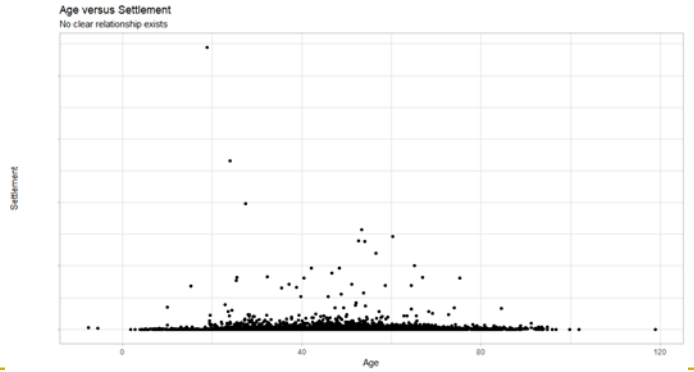


Institute and Faculty of Actuaries

Treatment of Missing Values

- Imputation or replacement of missing values is necessary in order to train most machine learning models effectively
- Variety of methods available – we used straightforward approach
- Missing values prevalence had little predictive impact here





Institute and Faculty of Actuaries

Treatment of Categorical Variables

- Machine learning algorithms can only learn from categorical data effectively if each categorical variable consists of a limited number of levels.
- Most common categories were used and rest grouped into ‘Other’ based on a judgment of data quality

There is no value added in separating these levels and the algorithm should not interpret these as ‘separate’ – they are the same for modelling purposes.

Rehab	Count
Not suitable	789
Not suitable for Rehab	987
Yes – by HO	234
Yes	210
Potential	123
…	…

Institute and Faculty of Actuaries

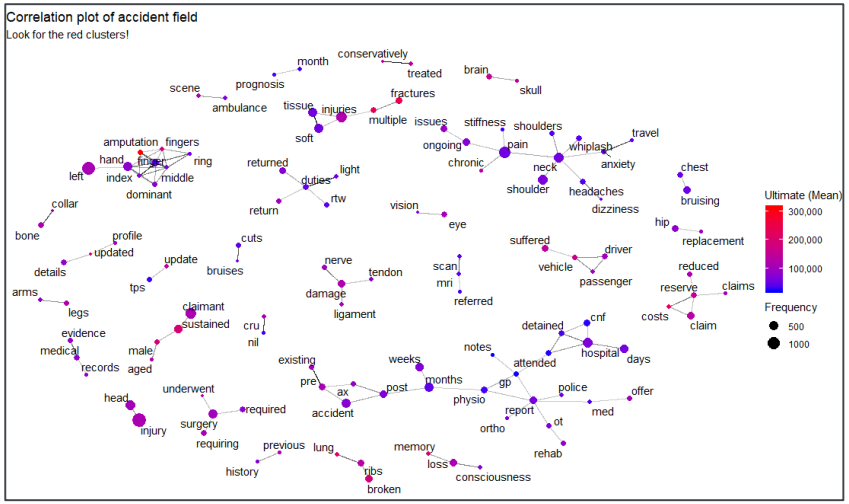
Exploratory Data Analysis

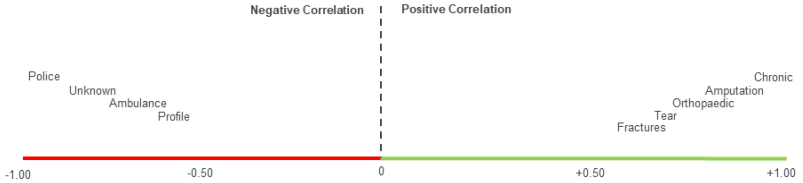
- Exploratory analysis of the effect of different field values on the ultimate position of each claim
- Provided an understanding of which field values had the greatest influence on eventual settlement

Institute and Faculty of Actuaries

Text Mining

- One of the project objectives was to investigate the ability to gain insight from unstructured text data
- Based on our analysis methodology, we assigned numerical severity ratings (from -1 to 1) to words which were sufficiently correlated with the target settlement fee across each claim.





Institute and Faculty of Actuaries

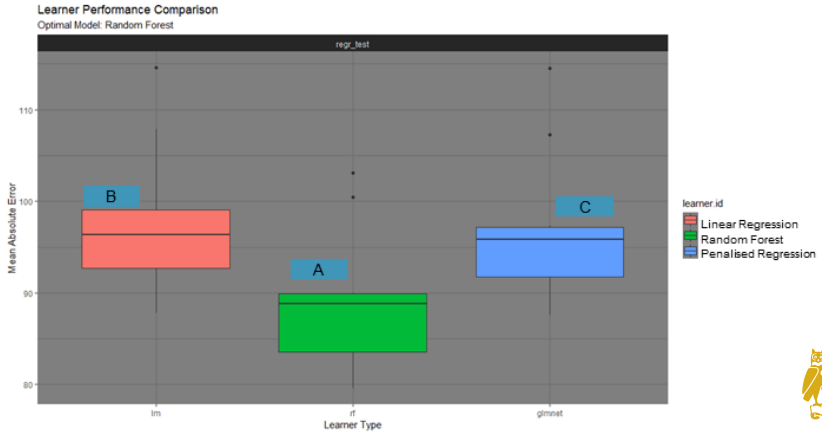
Institute and Faculty of Actuaries

The modelling approach

12 September 2019

Modelling approach

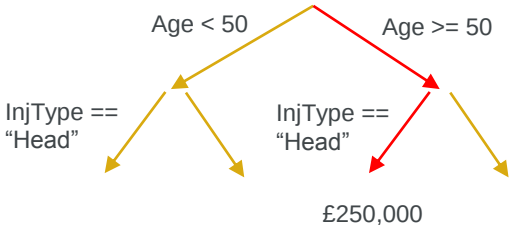
- The objective was to develop a model that could:
  - Take into consideration all the information in claims data, even if it was unstructured
  - Use this data with a machine learning model to provide a more accurate and precise ‘Best Estimate’
  - Create a process to learn from experience and improve as the data develops over time.
- Certain models are more interpretable than others however it can come at the expense of predictive accuracy. Since the objective was to optimise predictive accuracy less emphasis was placed on interpretability
- Trained limited number of machine learning methods : Linear Regression , Elastic Net Regression (Lasso / Ridge), and Random Forest
- Tested all algorithms through the machine learning packages in R.



Institute and Faculty of Actuaries

Random Forest

- The ‘Random Forest’ is actually a collection of decision trees.
- A simple explanation of the model is that it takes many different estimates (from each tree) and aggregates (averages) each estimate into one single estimate.
- Each branch corresponds to a different ‘split’ on a given variable and at the end of all the branches lies a given number of points (say 2 in the simple case below) where the data has been sufficiently reduced.
- The two observations pertaining to the decision path highlighted in red give a prediction of £250,000 – an average of £200,000 and £300,000.

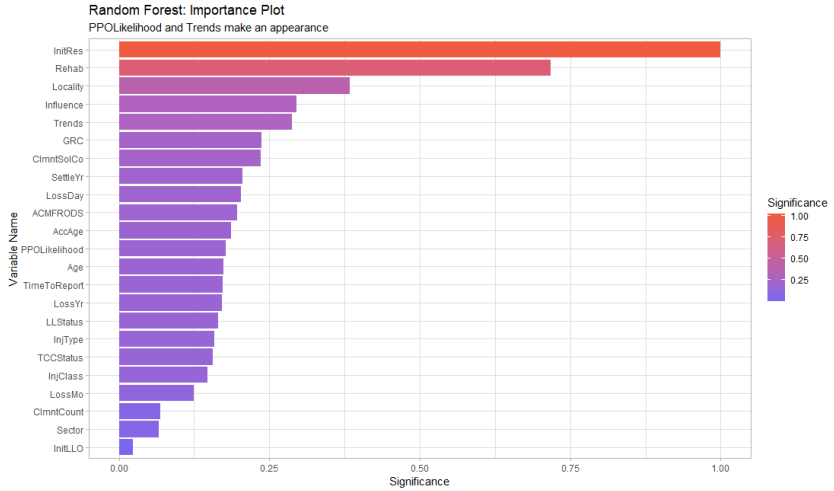


Age	Injury Type	Settlement
28	Spinal	750,000
56	Head	200,000
60	Head	300,000
23	Head	400,000

Institute and Faculty of Actuaries

Random Forest: Importance Plot

- The plot illustrates the relative ‘importance’ rating of each of the variables used by the model
- The importance of a variable is roughly equivalent to how often it is chosen as a primary splitting variable
- This plot provides some interpretability of the random forest and can also provide users with information on which variables may be biased.



Institute and Faculty of Actuaries

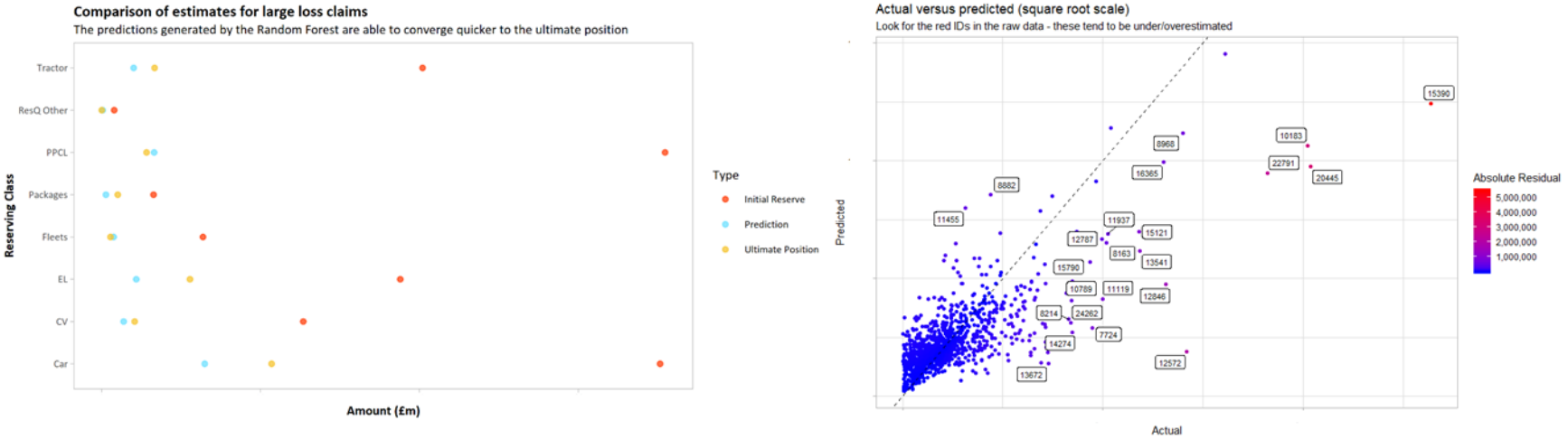
Institute and Faculty of Actuaries

Results

12 September 2019

Results

- Model was tested on ”unseen” out of sample data
- The model was overall, on average, relatively close to the true values on holdout data
- Tendency for the model to underestimate larger claims.



Institute and Faculty of Actuaries

Conclusions from the Project

12 September 2019

Conclusions

- **Benefits:**
  - In relatively small time / cost, machine learning techniques can provide a reasonable alternative approach
  - Improved understanding of NFU Mutual's data quality and availability
  - Opportunities in data visualisation and power of statistical software packages
- **Things to consider further:**
  - How will the model predict and react to emerging data?
  - Quantity and quality of data is key
  - Uncertainties created by change in Ogden discount rate remain
  - Communication of underlying assumptions becomes more challenging
  - Random Forest method makes it difficult to interpret drivers underlying results

Institute and Faculty of Actuaries

Conclusions – Where Next?

- Do current industry methods leverage the value of (increasingly robust) data available?
- Role of the Reserving Actuary is a long way from being automated!
- However, Actuaries must stay ahead of the fast moving progress in machine learning
- Over to the Machine Learning in Reserving Working Party!

Questions

Comments

Institute and Faculty of Actuaries