Car Insurance Claims Classification Report

Hamed Vaheb
hamed.vaheb.001@student.uni.lu

January 13, 2023

Abstract

This report describes the project of analysis and prediction of the car insurance claims dataset. The dataset includes historical data of the policyholders (e.g., age, gender, vehicle details , etc.), among which the variable of interest is the outcome of insurance, indicating whether a customer's is approved or not. The `aiinsurance` R package [1] is developed to make this work reproducible, accessible, and equipped with advanced features, e.g., a pipeline that performs the main steps of this work in a single command, and an interactive app to display the performance of the models. Using package's functions, two machine learning models, i.e., logistic regression and random forest are implemented to predict the outcomes based on solely the historical data. The results indicate promising prediction power with various evaluation metrics (e.g., accuracy, precision, recall, etc.) of over 80 percent. In addition to prediction, informative insights from the dataset and models have been drawn, e.g., by dint of the random forest model, the variables of the dataset that perform effective role in the prediction become evident.

1 Introduction

Rapid advances in artificial intelligence (AI) and machine learning are creating products and services with the potential not only to change the environment in which actuaries operate but also to provide new opportunities within actuarial science [2].

The use of statistical learning models has been a common practice in actuarial science since the 1980s. It was not long after since the field adopted classical models, such as linear models and generalized linear models (GLMs). While actuaries use GLMs frequently in practice, it was in the past few years that the use of AI and machine learning, and hence more modern models garnered significant attention in the field [3].

The goal of this work is to predict status of policyholders' claims. The status lies at the "outcome" column of the car insruance dataset, indicating whether a customer has claimed his loan or not.

A classical linear model and a modern nonlinear model is used and compared. The former model is logistic regression, which is an example of GLMs, accommodated to classification setting, i.e., for predicting discrete classes, in our case, status of claims. The latter model is a nonlinear tree-based model.

The remainder of this work is organized as the following: In section 2, main concepts included in the work are explained. In 2.1, the concept of car insurance is introduced followed by an elaboration on claims and how they interact among an insurer and a policyholder. In section 2.2, the machine learning models, logistic regression 2.2.1, and random forest 2.2.2 are briefly explained. As a contribution of this work, the `aiinsurance` package [1], introduced in 3.1 is developed, which automates many of the classification tasks, uses advanced features, e.g., runs a pipeline that performs the main steps of this work in a single command, and it provides an interactive shiny app to display the performance of the models.

2 Preliminary Concepts

2.1 Car Insurance Claims

Car insurance is a type of insurance policy that financially protects drivers in the event of an accident or theft. There are several types of car insurance coverage, including liability insurance, collision insurance, comprehensive insurance, and personal injury protection.

An insurance claim is a formal request by a policyholder to an insurance company for coverage or compensation for a covered loss or policy event, in our case, a car accident. The insurance company either accepts or rejects the claim. If it is approved, the insurance company will issue payment to the insured or an approved interested party on behalf of the insured.

Predicting the outcome of claims be utilized to better understand the customer strata and incorporate the findings throughout the insurance policy enrollment (including the underwriting and approval or rejection stages), triage claims and automate where possible, gradually obviating the need for human interaction, and optimize the entire insurance policy enrollment process flow. [4]

2.2 Models

2.2.1 Logistic Regression

2.2.2 Random Forest

Recent models learn nonlinear transformations and interactions between variables from the data without manually specifying them. This is performed implicitly with tree-based models, e.g., random forest.

3 Implementation and Results

3.1 `aiinsurance` Package

3.2 Dataset Analysis

3.3 Classification Models

4 Conclusion

5 Headings: first level

LaTeX command can be used to reference other section. See Section 5. However, you can also use **bookdown** extensions mechanism for this.

5.1 Headings: second level

You can use equation in blocks

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j|y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}$$

But also inline i.e $z = x + y$

5.1.1 Headings: third level

Another paragraph.

6 Examples of citations, figures, tables, references

You can insert references. Here is some text (**kour2014real?**; **kour2014fast?**) and see (**hadash2018estimate?**).

The documentation for `natbib` may be found at

You can use custom blocks with LaTeX support from **rmarkdown** to create environment.

http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf%7D

Of note is the command `\citet`, which produces citations appropriate for use in inline text.

You can insert LaTeX environment directly too.

```
\citet{hasselmo} investigated\dots
```

produces

> Hasselmo, et al. (1995) investigated...

https://www.ctan.org/pkg/booktabs

6.1 Figures

You can insert figure using LaTeX directly.

See Figure 1. Here is how you add footnotes. [^Sample of the first footnote.]

But you can also do that using R.

```
plot(mtcars$mpg)
```

You can use **bookdown** to allow references for Tables and Figures.

6.2 Tables

Below we can see how to use tables.

See awesome Table~1 which is written directly in LaTeX in source Rmd file.

You can also use R code for that.

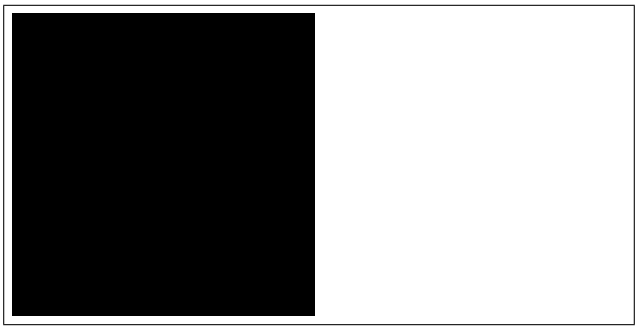

Figure 1: Sample figure caption.

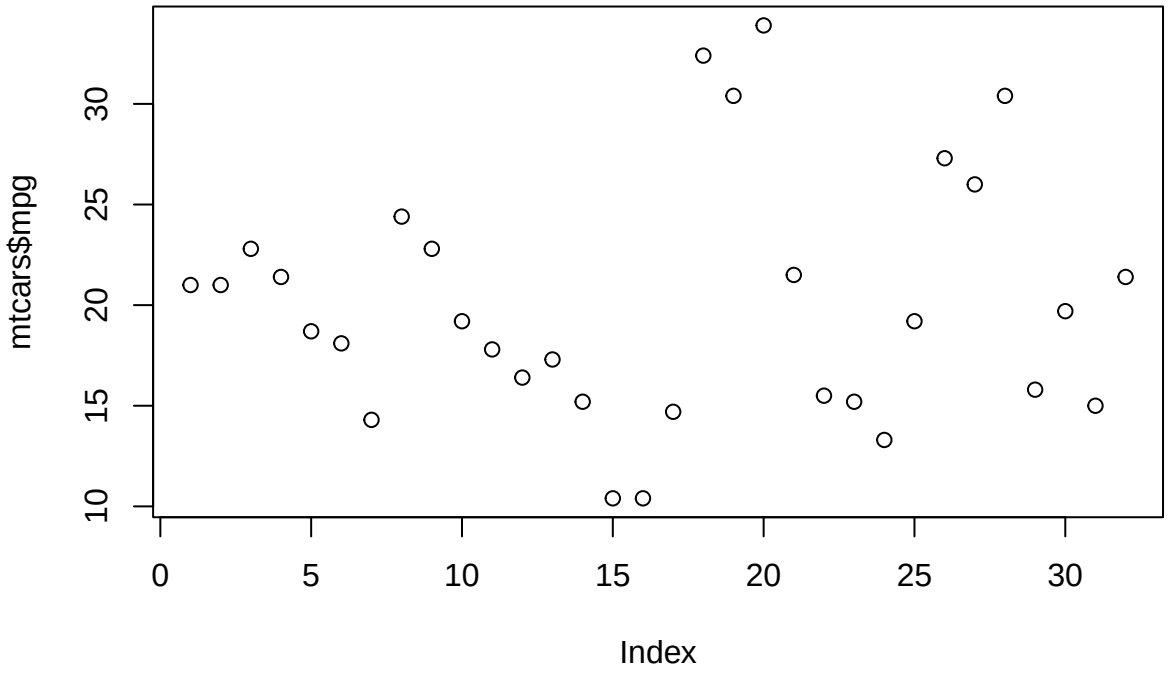

Figure 2: Another sample figure

Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

```
knitr::kable(head(mtcars), caption = "Head of mtcars table")
```

Table 2: Head of mtcars table

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

6.3 Lists

- Item 1
- Item 2
- Item 3

References

[1] Hamed Vaheb. AI Insurance Package. Available at http://github.com/berserkhmdvhb/aiinsurance.

[2] Ronald Richman. Ai in actuarial science – a review of recent advances – part 1. *Annals of Actuarial Science*, 15(2):207–229, 2021.

[3] Christopher Blier-Wong, Hélène Cossette, Luc Lamontagne, and Etienne Marceau. Machine learning in p&c insurance: A review for pricing and reserving. *Risks*, 9(1), 2021.

[4] Mohamed Hanafy and Ruixing Ming. Classification of the insureds using integrated machine learning algorithms: A comparative study. *Applied artificial intelligence*, 36(1), 2022.