Insurance Pricing With GLMs

Alan Chalk

2025-01-22

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

1 Introduction

Dear Reader,

2 Data preparation

Introduction

This chapter will provide code examples for various aspects of data preparation that should be addressed before modelling starts. It is very tempting once a dataset is available, to jump right in to the modelling stage - because that is interesting and data preparation is boring. Or because you are being chased by management for a new and improved rating plan which they in turn have promised their boss to have ready by yesterday. In this regard they will often quote the 80-20 rule to you. Meaning that you can get 80\% of the benefit of a data cleansing exercise by just 20\% of the effort. (Or they might use the phrase "perfect is the enemy of good"). This magical rule, means that if you need a month to do data cleansing, your boss can reasonably expect it of you in just under a week, and that whatever has been left undone will not matter to the final result. There is no simple answer to such time pressures. But realistic planning at the start of a project can help decide what should and should not be attempted, given the time available.

A typical outcome of poor data preparation is getting to the end of a few days of modelling and then realizing that you included a feature which should not be included. Or you treated a feature as numeric which should been treated as a character / factor. An example of the latter is the number of doors of a vehicle (in auto insurance). Treating this as a numeric feature will often lead to the assumption that the risk of accident decreases (or increases) linearly as the number of doors increases where this might not be true. The net result of such errors can vary from some minor changes need to be made post-hoc, to having to redo the whole exercise.

This chapter covers naming convention for our features both from the point of view of understanding what they are, and from a computer programming perspective. Character and numeric features will be discussed. And we will also look at feature screening, missing values and general sense-checking.

3 Summary