Analysis of the "automobile-lossess-prediction" dataset

Illinois State University - ACC 471 - Final Report

Jared Musil & Jake McNair

2017-11-29

Table of Contents

# Prerequisites

## Foreword

This report was written using the R package Bookdown. This was done as it allows for reproducable research of our data, methods, and results. Where appropriate, the code has been included inline with the results. All other methods are contained within the Appendix.

It is also available as a website reading on mobile devices, and and epub for reading offline.

## Markdown Test

is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation .

For now, you have to install the development versions of **bookdown** from Github:

devtools::install\_github("rstudio/bookdown")

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading #.

To compile this example to PDF, you need to install XeLaTeX.

# Introduction

Throughout this report, the columns of our dataset will be refered to as factors, and the rows of our dataset will be refered to as reccords. This is

The technology used to generate this report was the R statistical programming language. This was chosen to allow for reproducable research and full transparency of the methods used to arrive at our conclusions. The code itself has been omitted from the report for brevity, but is available for review and reuse in this reports gitlab page [link][gitlab].

[gitlab](https://github.com/jaredmusil/acc471-final-report)

# Problem Description

The ability to utilize analytics to predict automobeile lossess is a area of active research and application throughout the insurance and fin-tech industries. All of the "big four" US domiciled auto insurrers being State Farm, Geico, Allstate, and Progressive are actively engaging in research to operationalize analytical models to increase operational efficency. [citation needed...]. This dataset is representitive of claims data common to all of these auto insurance providers, and the industry at large.

From a consumer standpoint, this has the potential to reduce average claim times, reduce premium costs, and improve claims decisions (total loss, not total loss).

# Data

Before doing any analysis, the feactors withing data set were first checked for missing or invalid data. Of the original 205 reccords, 41 were removed because they contained missing data for the normalized-lossess factor, which was coded as a ?. This resulted in a dataset of 164 reccords of clean data. No other factors needed cleaning up, as the data was properly coded for each reccord.

library(readxl)  
data\_dict <- readxl::read\_xlsx("automobile-losses-data-dictionary.xlsx")  
knitr::kable(head(data\_dict, 20), caption = 'Data Dictionary - Initial', booktabs = TRUE)

Table 1 Data Dictionary - Initial

|  |  |  |  |
| --- | --- | --- | --- |
| # | Description | Values | Keep |
| 1 | symboling | -3, -2, -1, 0, 1, 2, 3 | No |
| 2 | normalized-losses | continuous from [65 to 256] | Yes |
| 3 | make | alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo | Yes |
| 4 | fuel-type | diesel, gas | Yes |
| 5 | aspiration | std, turbo | Yes |
| 6 | num-of-doors | four, two | Yes |
| 7 | body-style | hardtop, wagon, sedan, hatchback, convertible | Yes |
| 8 | drive-wheels | 4wd, fwd, rwd. | Yes |
| 9 | engine-location | front, rear | Yes |
| 10 | wheel-base | continuous from [86.6 to 120.9] | Yes |
| 11 | length | continuous from [141.1 to 208.1] | Yes |
| 12 | width | continuous from [60.3 to 72.3] | Yes |
| 13 | height | continuous from [47.8 to 59.8] | Yes |
| 14 | curb-weight: | continuous from [1488 to 4066] | Yes |
| 15 | engine-type | dohc, dohcv, l, ohc, ohcf, ohcv, rotor | Yes |
| 16 | num-of-cylinders | eight, five, four, six, three, twelve, two | Yes |
| 17 | engine-size | continuous from [61 to 326] | Yes |
| 18 | fuel-system | 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi | Yes |
| 19 | bore | continuous from [2.54 to 3.94] | Yes |
| 20 | stroke | continuous from [2.07 to 4.17] | Yes |

Of these factors, 10 of the initial 26 were removed, resulting in the 16 factors that will be used in analysis. These factors are noted in green in Keep column of the above table.

The objective factor in the dataset is determined to be ``.

Next, the data was partitioned into three groups named *training*, *test*, and *validation*. This was

data <- readxl::read\_xlsx("automobile-losses.xlsx")  
knitr::kable(head(data, 20), caption = 'Dataset', booktabs = TRUE)

Table 2 Dataset

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
| 3 | ? | alfa-romero | gas | std | two | convertible | rwd | front | 88.6 | 168.8 | 64.1 | 48.8 | 2548 | dohc | four | 130 | mpfi | 3.47 | 2.68 | 9.0 | 111 | 5000 | 21 | 27 | 13495 |
| 3 | ? | alfa-romero | gas | std | two | convertible | rwd | front | 88.6 | 168.8 | 64.1 | 48.8 | 2548 | dohc | four | 130 | mpfi | 3.47 | 2.68 | 9.0 | 111 | 5000 | 21 | 27 | 16500 |
| 1 | ? | alfa-romero | gas | std | two | hatchback | rwd | front | 94.5 | 171.2 | 65.5 | 52.4 | 2823 | ohcv | six | 152 | mpfi | 2.68 | 3.47 | 9.0 | 154 | 5000 | 19 | 26 | 16500 |
| 2 | 164 | audi | gas | std | four | sedan | fwd | front | 99.8 | 176.6 | 66.2 | 54.3 | 2337 | ohc | four | 109 | mpfi | 3.19 | 3.4 | 10.0 | 102 | 5500 | 24 | 30 | 13950 |
| 2 | 164 | audi | gas | std | four | sedan | 4wd | front | 99.4 | 176.6 | 66.4 | 54.3 | 2824 | ohc | five | 136 | mpfi | 3.19 | 3.4 | 8.0 | 115 | 5500 | 18 | 22 | 17450 |
| 2 | ? | audi | gas | std | two | sedan | fwd | front | 99.8 | 177.3 | 66.3 | 53.1 | 2507 | ohc | five | 136 | mpfi | 3.19 | 3.4 | 8.5 | 110 | 5500 | 19 | 25 | 15250 |
| 1 | 158 | audi | gas | std | four | sedan | fwd | front | 105.8 | 192.7 | 71.4 | 55.7 | 2844 | ohc | five | 136 | mpfi | 3.19 | 3.4 | 8.5 | 110 | 5500 | 19 | 25 | 17710 |
| 1 | ? | audi | gas | std | four | wagon | fwd | front | 105.8 | 192.7 | 71.4 | 55.7 | 2954 | ohc | five | 136 | mpfi | 3.19 | 3.4 | 8.5 | 110 | 5500 | 19 | 25 | 18920 |
| 1 | 158 | audi | gas | turbo | four | sedan | fwd | front | 105.8 | 192.7 | 71.4 | 55.9 | 3086 | ohc | five | 131 | mpfi | 3.13 | 3.4 | 8.3 | 140 | 5500 | 17 | 20 | 23875 |
| 0 | ? | audi | gas | turbo | two | hatchback | 4wd | front | 99.5 | 178.2 | 67.9 | 52.0 | 3053 | ohc | five | 131 | mpfi | 3.13 | 3.4 | 7.0 | 160 | 5500 | 16 | 22 | ? |
| 2 | 192 | bmw | gas | std | two | sedan | rwd | front | 101.2 | 176.8 | 64.8 | 54.3 | 2395 | ohc | four | 108 | mpfi | 3.5 | 2.8 | 8.8 | 101 | 5800 | 23 | 29 | 16430 |
| 0 | 192 | bmw | gas | std | four | sedan | rwd | front | 101.2 | 176.8 | 64.8 | 54.3 | 2395 | ohc | four | 108 | mpfi | 3.5 | 2.8 | 8.8 | 101 | 5800 | 23 | 29 | 16925 |
| 0 | 188 | bmw | gas | std | two | sedan | rwd | front | 101.2 | 176.8 | 64.8 | 54.3 | 2710 | ohc | six | 164 | mpfi | 3.31 | 3.19 | 9.0 | 121 | 4250 | 21 | 28 | 20970 |
| 0 | 188 | bmw | gas | std | four | sedan | rwd | front | 101.2 | 176.8 | 64.8 | 54.3 | 2765 | ohc | six | 164 | mpfi | 3.31 | 3.19 | 9.0 | 121 | 4250 | 21 | 28 | 21105 |
| 1 | ? | bmw | gas | std | four | sedan | rwd | front | 103.5 | 189.0 | 66.9 | 55.7 | 3055 | ohc | six | 164 | mpfi | 3.31 | 3.19 | 9.0 | 121 | 4250 | 20 | 25 | 24565 |
| 0 | ? | bmw | gas | std | four | sedan | rwd | front | 103.5 | 189.0 | 66.9 | 55.7 | 3230 | ohc | six | 209 | mpfi | 3.62 | 3.39 | 8.0 | 182 | 5400 | 16 | 22 | 30760 |
| 0 | ? | bmw | gas | std | two | sedan | rwd | front | 103.5 | 193.8 | 67.9 | 53.7 | 3380 | ohc | six | 209 | mpfi | 3.62 | 3.39 | 8.0 | 182 | 5400 | 16 | 22 | 41315 |
| 0 | ? | bmw | gas | std | four | sedan | rwd | front | 110.0 | 197.0 | 70.9 | 56.3 | 3505 | ohc | six | 209 | mpfi | 3.62 | 3.39 | 8.0 | 182 | 5400 | 15 | 20 | 36880 |
| 2 | 121 | chevrolet | gas | std | two | hatchback | fwd | front | 88.4 | 141.1 | 60.3 | 53.2 | 1488 | l | three | 61 | 2bbl | 2.91 | 3.03 | 9.5 | 48 | 5100 | 47 | 53 | 5151 |
| 1 | 98 | chevrolet | gas | std | two | hatchback | fwd | front | 94.5 | 155.9 | 63.6 | 52.0 | 1874 | ohc | four | 90 | 2bbl | 3.03 | 3.11 | 9.6 | 70 | 5400 | 38 | 43 | 6295 |

data <- readxl::read\_xlsx("automobile-losses.xlsx")  
summary(data)

## 1 2 3   
## Min. :-2.0000 Length:205 Length:205   
## 1st Qu.: 0.0000 Class :character Class :character   
## Median : 1.0000 Mode :character Mode :character   
## Mean : 0.8341   
## 3rd Qu.: 2.0000   
## Max. : 3.0000   
## 4 5 6   
## Length:205 Length:205 Length:205   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## 7 8 9 10   
## Length:205 Length:205 Length:205 Min. : 86.60   
## Class :character Class :character Class :character 1st Qu.: 94.50   
## Mode :character Mode :character Mode :character Median : 97.00   
## Mean : 98.76   
## 3rd Qu.:102.40   
## Max. :120.90   
## 11 12 13 14   
## Min. :141.1 Min. :60.30 Min. :47.80 Min. :1488   
## 1st Qu.:166.3 1st Qu.:64.10 1st Qu.:52.00 1st Qu.:2145   
## Median :173.2 Median :65.50 Median :54.10 Median :2414   
## Mean :174.0 Mean :65.91 Mean :53.72 Mean :2556   
## 3rd Qu.:183.1 3rd Qu.:66.90 3rd Qu.:55.50 3rd Qu.:2935   
## Max. :208.1 Max. :72.30 Max. :59.80 Max. :4066   
## 15 16 17 18   
## Length:205 Length:205 Min. : 61.0 Length:205   
## Class :character Class :character 1st Qu.: 97.0 Class :character   
## Mode :character Mode :character Median :120.0 Mode :character   
## Mean :126.9   
## 3rd Qu.:141.0   
## Max. :326.0   
## 19 20 21 22   
## Length:205 Length:205 Min. : 7.00 Length:205   
## Class :character Class :character 1st Qu.: 8.60 Class :character   
## Mode :character Mode :character Median : 9.00 Mode :character   
## Mean :10.14   
## 3rd Qu.: 9.40   
## Max. :23.00   
## 23 24 25 26   
## Length:205 Min. :13.00 Min. :16.00 Length:205   
## Class :character 1st Qu.:19.00 1st Qu.:25.00 Class :character   
## Mode :character Median :24.00 Median :30.00 Mode :character   
## Mean :25.22 Mean :30.75   
## 3rd Qu.:30.00 3rd Qu.:34.00   
## Max. :49.00 Max. :54.00

# Methods Used

We utilized X methods in our analysis, while settlling on regression trees for our final reccomendations.

# Results

...

## Example one

## Example two

# Reccomentations

...

# Future Analysis

...

# Conculsion

...