

Analysis of the “automobile-loss-prediction” dataset

Illinois State University - ACC 471 - Final Report

Jared Musil & Jake McNair

2017-11-29

Contents

| | | |
|----------|-------------------------------|-----------|
| 1 | Introduction | 3 |
| 2 | Problem Description | 4 |
| 3 | Data | 5 |
| 4 | Methods Used | 9 |
| 5 | Results | 11 |
| 5.1 | Regression Tree | 11 |
| 5.2 | Classification Tree | 12 |
| 6 | Reccomentations | 13 |
| 7 | Future Analysis | 14 |
| 8 | Conculsion | 15 |

Chapter 1

Introduction

The ability to utilize analytics to predict automobile loss is an area of active research and application throughout the insurance and fin-tech industries. All of the “big four” US domiciled auto insurers being State Farm, Geico, Allstate, and Progressive are actively engaging in research to operationalize analytical models to increase operational efficiency. [citation needed...]. This dataset is representative of claims data common to all of these auto insurance providers, and the industry at large.

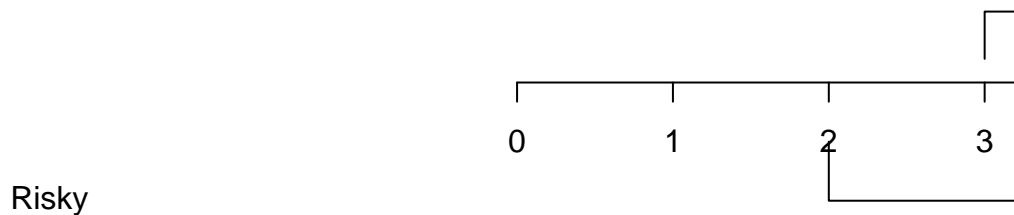
From a consumer standpoint, this has the potential to reduce average claim times, reduce premium costs, and improve claims decisions (total loss, not total loss).

Throughout this report, the columns of our dataset will be referred to as factors, and the rows of our dataset will be referred to as records. This is because it follows the terminology used by the R statistical programming language, which was the analytical tool used in this report. This was chosen to allow for reproducible 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 at the following URL: <https://github.com/jaredmusil/acc471-final-report>

Chapter 2

Problem Description

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process “symboling”. A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.



The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...), and represents the average loss per car per year.

Chapter 3

Data

Before doing any analysis, the factors within the dataset were first checked for missing or invalid data. The individual factors can be described as follows: 15 continuous, 10 nominal, and 1 integer.

Seven of the factors contained missing or improperly coded data. In this dataset in particular, all missing data has been coded with the value of ?. In all cases below, the records containing the missing data have been removed.

| Index | Factor | Number of records missing a value |
|-------|-------------------|-----------------------------------|
| 2 | normalized-losses | 41 |
| 6 | num-of-doors | 2 |
| 19 | bore | 4 |
| 20 | stroke | 4 |
| 22 | horsepower | 2 |
| 23 | peak-rpm | 2 |
| 26 | price | 4 |

Of the original 205 records, X were removed because they contained missing data for the **normalized-losses** factor, which was coded as a ?. This resulted in a dataset of 164 records of clean data. No other factors needed cleaning up, as the data was properly coded for each record.

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 **symboling**.

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

| | | | | | |
|----|---------------|-----------|------------|------------|----------|
| ## | X1 | X2 | X3 | X4 | X5 |
| ## | Min. :-2.0000 | Min. : 65 | toyota :31 | diesel: 15 | std :136 |

Table 3.2: Data Dictionary - Initial

| N | Description | Values |
|----|-------------------|---|
| 1 | symboling | -3, -2, -1, 0, 1, 2, 3 |
| 2 | normalized-losses | continuous from [65 to 256] |
| 3 | make | alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, merced |
| 4 | fuel-type | diesel, gas |
| 5 | aspiration | std, turbo |
| 6 | num-of-doors | four, two |
| 7 | body-style | hardtop, wagon, sedan, hatchback, convertible |
| 8 | drive-wheels | 4wd, fwd, rwd. |
| 9 | engine-location | front, rear |
| 10 | wheel-base | continuous from [86.6 to 120.9] |
| 11 | length | continuous from [141.1 to 208.1] |
| 12 | width | continuous from [60.3 to 72.3] |
| 13 | height | continuous from [47.8 to 59.8] |
| 14 | curb-weight: | continuous from [1488 to 4066] |
| 15 | engine-type | dohc, dohcv, l, ohc, ohcf, ohcv, rotor |
| 16 | num-of-cylinders | eight, five, four, six, three, twelve, two |
| 17 | engine-size | continuous from [61 to 326] |
| 18 | fuel-system | 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi |
| 19 | bore | continuous from [2.54 to 3.94] |
| 20 | stroke | continuous from [2.07 to 4.17] |
| 21 | compression-ratio | continuous from [7 to 23] |
| 22 | horsepower | continuous from [48 to 288] |
| 23 | peak-rpm | continuous from [4,150 to 6,600] |
| 24 | city-mpg | continuous from [13 to 49] |
| 25 | highway-mpg | continuous from [16 to 54] |
| 26 | price | continuous from [5,118 to 45,400] |

```

## 1st Qu.: 0.0000    1st Qu.: 94    nissan :18    gas    :149    turbo: 28
## Median : 1.0000    Median :115    mazda  :15
## Mean   : 0.7927    Mean   :122    honda  :13
## 3rd Qu.: 2.0000    3rd Qu.:150    subaru :12
## Max.   : 3.0000    Max.   :256    volvo  :11
##                                     (Other):64
##      X6              X7              X8              X9              X10
## ?    : 1    convertible: 2    4wd: 8    front:164    Min.    : 86.60
## four:95    hardtop    : 5    fwd:106                1st Qu.: 94.50
## two :68    hatchback  :60    rwd: 50                Median : 96.55
##                sedan      :80                Mean   : 98.16
##                wagon      :17                3rd Qu.:100.40
##                                     Max.    :115.60
##
##      X11              X12              X13              X14              X15
## Min.    :141.1    Min.    :60.3    Min.    :49.40    Min.    :1488    dohc : 8
## 1st Qu.:165.7    1st Qu.:64.0    1st Qu.:52.00    1st Qu.:2091    l     : 8
## Median :172.0    Median :65.4    Median :54.10    Median :2368    ohc  :124
## Mean   :172.2    Mean   :65.6    Mean   :53.77    Mean   :2458    ohcf : 12
## 3rd Qu.:177.8    3rd Qu.:66.5    3rd Qu.:55.50    3rd Qu.:2786    ohcv : 8
## Max.    :202.6    Max.    :71.7    Max.    :59.80    Max.    :4066    rotor: 4
##
##      X16              X17              X18              X19              X20
## eight: 1    Min.    : 61.0    1bbl:11    3.62    :20    3.03    :14
## five : 7    1st Qu.: 97.0    2bbl:63    3.15    :15    3.15    :14
## four :137    Median :109.0    4bbl: 3    3.19    :15    3.4     :13
## six  : 14    Mean   :118.0    idi :15    2.97    :12    3.23    :12
## three: 1    3rd Qu.:131.8    mfi : 1    3.03    :10    2.64    :11
## two  : 4    Max.    :258.0    mpfi:66    2.91    : 7    3.29    : 9
##                spdi: 5    (Other):85    (Other):91
##      X21              X22              X23              X24
## Min.    : 7.00    Min.    : 48.00    Min.    :4150    Min.    :15.00
## 1st Qu.: 8.70    1st Qu.: 69.00    1st Qu.:4800    1st Qu.:22.00
## Median : 9.00    Median : 91.00    Median :5200    Median :26.00
## Mean   :10.13    Mean   : 96.21    Mean   :5138    Mean   :26.27
## 3rd Qu.: 9.40    3rd Qu.:114.00    3rd Qu.:5500    3rd Qu.:31.00
## Max.    :23.00    Max.    :200.00    Max.    :6600    Max.    :49.00
##
##      X25              X26
## Min.    :18.00    Min.    : 5118
## 1st Qu.:28.00    1st Qu.: 7446
## Median :32.00    Median : 9268
## Mean   :31.85    Mean   :11467
## 3rd Qu.:37.00    3rd Qu.:14559
## Max.    :54.00    Max.    :35056

```

##

Chapter 4

Methods Used

A number of analytical methods are available for use such as decision trees, classification trees, regression, multiple-regression. Not all of these techniques makes sense for our purposes as they are used to predict different types of information.

We utilized X methods in our analysis, while settling on regression trees for our final recommendation.

The main goal of our analysis is to predict how risky a particular car is, and therefore Regression trees make the most sense.

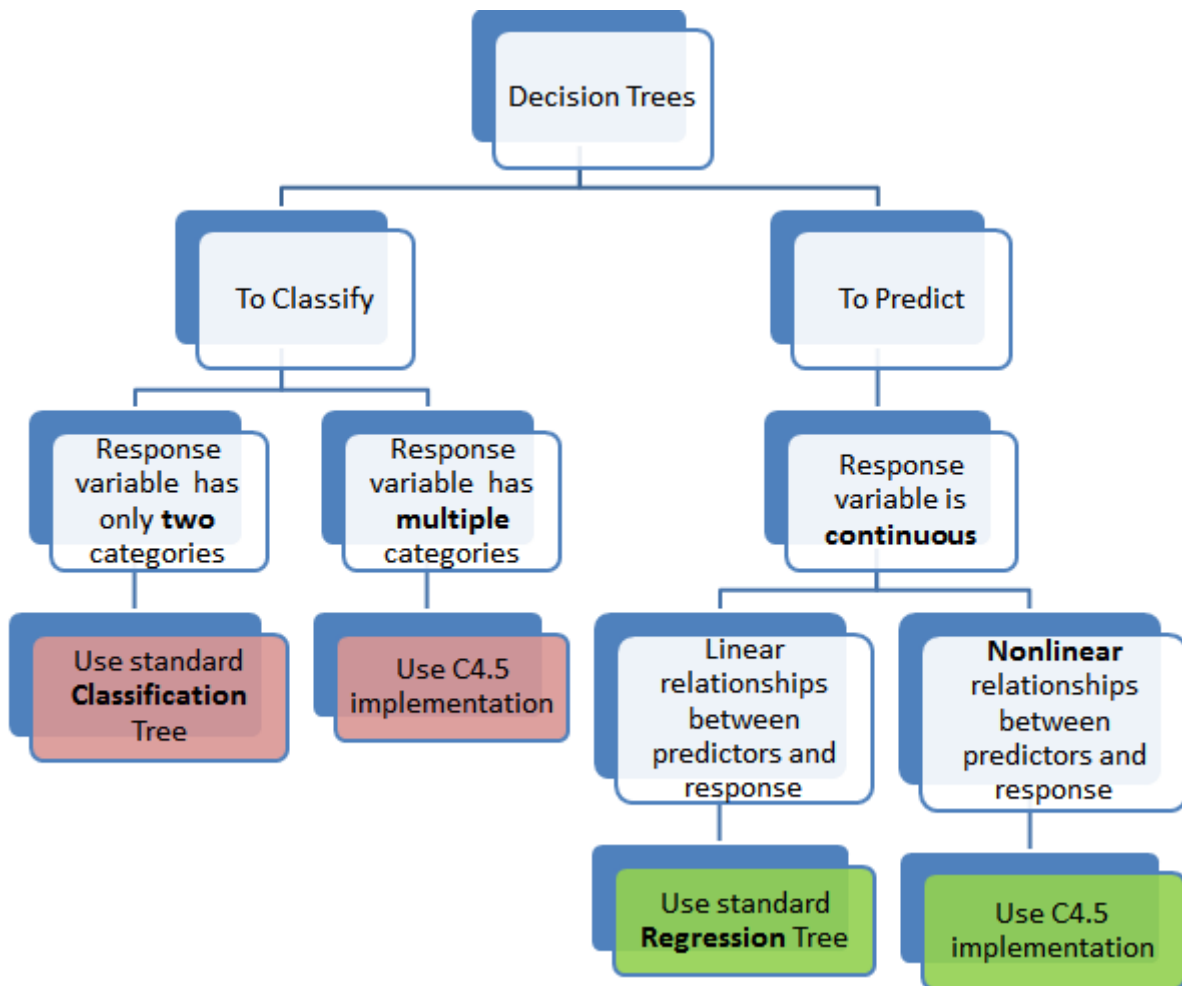


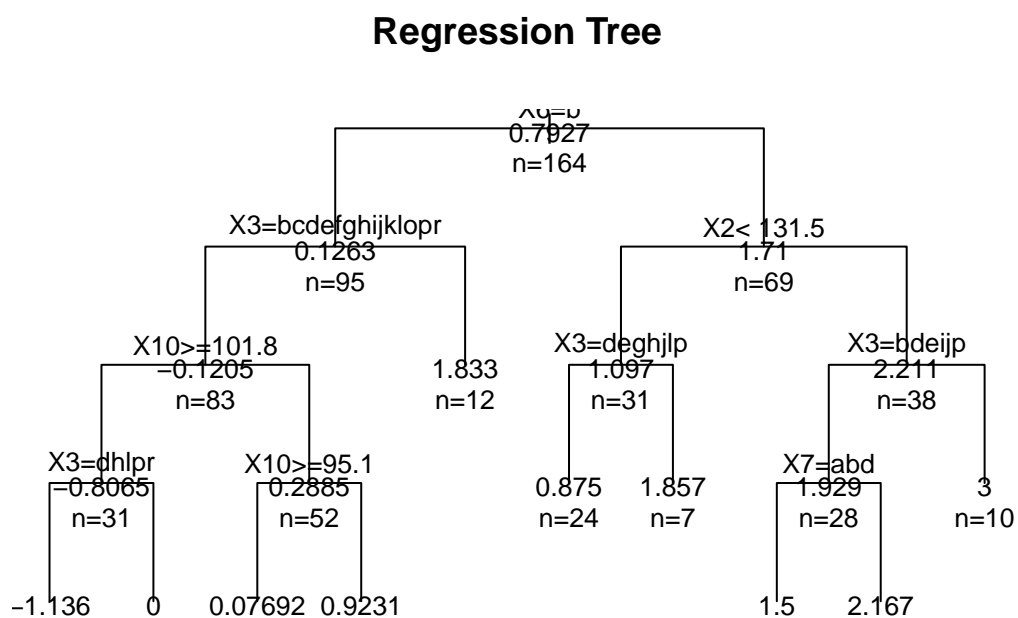
Figure 4.1: Source: <http://www.simafore.com/blog/bid/62482/2-main-differences-between-classification-and-regression-trees>

Chapter 5

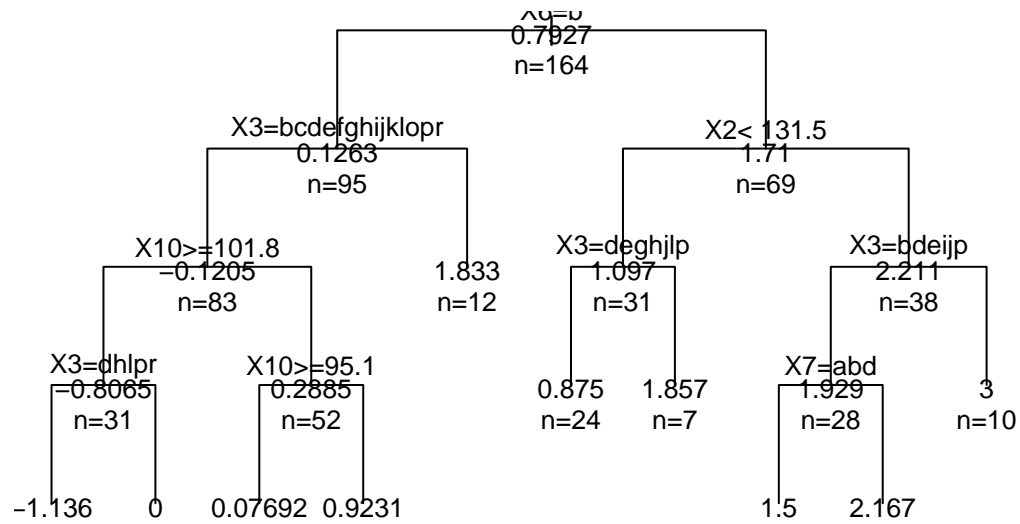
Results

...

5.1 Regression Tree



Pruned Regression Tree



Lift Chart

...

Decile Chart

...

5.2 Classification Tree

Lift Chart

...

Decile Chart

...

Chapter 6

Reccomentations

...

Chapter 7

Future Analysis

As with any data analysis, the quality of the input data will determine the quality of the resulting models. In this case we started with 26 factors. A good way to increase the quality of the model would be to provide it with more factors and potentially more levels within the factors.

All of this data also is only related to the automobile itself, and does not account for the individual driving it. While some behavioral and demographic factors protected by federal law from being used for analysis like race and religion(CITE), Others such as gender are allowed. Including these behavioral factors as inputs into the model would be an opportunity to strengthen the existing model. Technology and in particular the increase of telematics within vehicles and internet of things (IoT) connected devices, will increase the ubiquity and variety of this datastream. With the advances in autonomous vehicles, behavioral factors may impact results less, but is something to monitor for the future of auto risk classification.

Chapter 8

Conculsion

Given the results of this analysis, we

Bibliography