Spain Crash Data Analysis

Seyed Saber Naseralavi Seyed Ehsan Jafarinasab

This version: August 21, 2023 (First version: August 16, 2023)

Abstract: In this short text, the general ideas of some potential research that can be done with Spanish accident data are mentioned and then an example of them is implemented. Also, some points are mentioned for future directions.

1 Introduction

I am writing this text for my dear friend and colleague Mr. Jafari Nasab for two purposes:

1. This document was written through a Quarto project in R Studio. This document itself will be the apparent basis of our future research. There are things to learn that you will learn by carefully going through the details of the documents in the project folder. Of course, we will discuss it at the right time.
2. Ideas for future research are suggested and an example is implemented.

If the terms "machine learning" and "crash severity prediction" or similar items are searched on Mendeley's website (link), we will find that firstly, a large amount of articles have been written in these fields, and secondly, this amount of articles has increased in recent years. I have downloaded and reviewed these articles more or less in recent months and created an Excel file for reference management. Through the investigations, I realized that the articles often focus on a part of the whole data, which I show below with my own division:

- Location
  - Freeway
  - Highway-rail crossing
  - Intersection
  - Workzone
- Factor
  - Age

  - Time-of-day
  - Alcohol
  - Seat belt
  - At-fault
  - Gender
  - Visibility condition
  - Weather conditions

- Type
  - Car-truck
  - Cross-median
  - Fixed object
  - Hazardous material
  - Head on
  - Rear-end
  - Rollover
  - Hit and run
  - Multi vehicles
  - Single vehicle
  - Red light running crashes
  - Roadway departure

- User
  - Cyclists
  - Pedestrian
  - Motorcycle
  - Large truck
  - Teen drivers
  - Children and adolescents involved

IN Section 2, these items are explained in more detail. I believe that in order to be able to publish a high-quality research paper with Spanish accident data, each of these divisions can be the basis of our research. Therefore, to begin with, we should focus our attention on only one part and while carefully studying the research literature and related articles, extract the desired data from the overall raw data. Then we will gradually advance the research work by analyzing the extracted data and building a machine learning model and interpreting it as much as possible. Obviously, as a research work starts, dozens of small and big ideas are brainstormed. Maybe some of those ideas will be implemented in the research and finally at some point of time the article will be finalized with the agreement of the authors and the research will end and the final output will be prepared for submission to the target journal.

The text is organized in such a way that after the introduction, the research literature is reviewed in Section 2. In Section 3, the method is described. Then the results are presented

in Section 4. In Section 5, the results are discussed, and finally, in Section 6, conclusions are made.

2 Literature Reveiw

As mentioned in the Section 1, in this section, the research literature is presented very briefly and only by referring to the research in each division:

- Location
  - Freeway (Li et al., 2020; Li et al., 2018; Wen et al., 2023)
  - Highway-rail crossing (Keramati et al., 2020; Kutela et al., 2023)
  - Intersection (Lin and Fan, 2021; Russo et al., 2023; Sharafeldin et al., 2022; Zhu, 2022)
  - Workzone (Dimitrijevic et al., 2023; Ghasemzadeh and Ahmed, 2019; Yu et al., 2020)
- Factor
  - Age (Lee et al., 2023; Mafi et al., 2018)
  - Time-of-day (Behnood and Mannering, 2019; Song et al., 2021)
  - Alcohol (Lasota et al., 2020; Liu and Fan, 2020; Shyhalla, 2014; Song et al., 2021; Wu and Zhang, 2018)
  - Seat belt (Abay et al., 2013; Kim et al., 2021)
  - At-fault (Rezapour et al., 2020)
  - Gender (Amarasingha and Dissanayake, 2014; Billah et al., 2022; Fu et al., 2021; Lee et al., 2023; Mafi et al., 2018)
  - Visibility condition (Harris et al., 2023; Li et al., 2018)
  - Weather conditions (Ghasemzadeh and Ahmed, 2019; Naik et al., 2016; Sawtelle et al., 2023; Yazdani and Nassiri, 2021; Zeng et al., 2020; Zhai et al., 2019)
- Type
  - Car-truck (Song et al., 2023; Zhou et al., 2020)
  - Cross-median (Das et al., 2018; Hu and Donnell, 2011; Lu et al., 2010)
  - Fixed object (Holdridge et al., 2005; Yan et al., 2022)
  - Hazardous material (Ahmed et al., 2020; Iranitalab et al., 2018; Shen and Wei, 2021; Sun et al., 2022; Xing et al., 2020)
  - Head on (Kardar and Davoodi, 2020; Liu and Fan, 2020)
  - Rear-end (Champahom et al., 2020; Mohamed et al., 2017; Prajongkha et al., 2023; Shao et al., 2020; Yu et al., 2020)
  - Rollover (Bullard et al., 2023; Hu and Donnell, 2011; Khan and Vachal, 2020; Rezapour and Ksaibati, 2022)

  - Hit and run (Jiang et al., 2021; Sivasankaran and Balasubramanian, 2022; Zhou et al., 2018; Zhu and Wan, 2021)
  - Multi vehicles (Ma et al., 2023; Wen et al., 2023; Yazdani and Nassiri, 2021)
  - Single vehicle (Ma et al., 2023; Naik et al., 2016; Roque et al., 2021; Sivasankaran et al., 2021)
  - Red light running crashes (Shaaban et al., 2021; Zhang et al., 2021)
  - Roadway departure (Alhasan et al., 2018; Peng et al., 2012; Yu et al., 2021)
- User
  - Cyclists (Boufous et al., 2012; Eriksson et al., 2022)
  - Pedestrian (Kim et al., 2017; Sivasankaran and Balasubramanian, 2022; Zafri et al., 2020; Zhai et al., 2019)
  - Motorcycle (Farid and Ksaibati, 2021; Kitali et al., 2022; Prajongkha et al., 2023; Rezapour et al., 2020; Salum et al., 2019; Wahab and Jiang, 2020)
  - Large truck (Azimi et al., 2022, 2020; Behnood and Al-Bdairi, 2020; Behnood and Mannering, 2019; Hosseinzadeh et al., 2021; Li et al., 2020; Okafor et al., 2022; Wu et al., 2023; Zhu and Srinivasan, 2011)
  - Teen drivers (Duddu et al., 2019; Hossain et al., 2023; Villavicencio et al., 2022)
  - Children and adolescents (Rezapour and Ksaibati, 2021; Theofilatos et al., 2021)

It is obviously in some articles, some combination conditions of the above division are used. The review of the above references was done very superficially. Of course, when we decide to work on a specific topic, we should try to extract all related articles in more detail. Then plan to study and categorize the effective factors affecting that particular issue so as to extract the consistency and inconsistency of the research results.

3 Method

The aim of the research here is to present a machine learning model to predict the severity of accidents with Spanish accident data. We consider the response variable as the severity of accidents in two levels of injury or non-injury accident. Obviously, a fatal accident is included in the group of injury accidents. The research method is as follows:

1. First, the accident data of Spain is loaded
2. Through exploratory analysis, the data is checked and some graphs are drawn
3. Some variables are selected as predictive variables. In this case, only previous experiences are used and special feature engineering is not used
4. A decision tree machine learning model is built
5. With cross-validation, the built model is checked and then finalized 6
6. The final model is evaluated

4 Results

4.1 Load Data

From the aspect of reproducible research, it is better to always describe the research process from the beginning, that is, from the place where the data is read from the primary file. But for convenience here, the R file that has already been created as raw data is copied to the project folder, and to start each research, the data is loaded with the following code.

```
load("RawData.RData")
```

Often, we put the `#| cache: TRUE` for each chunk of code to save the time of subsequent executions.

4.2 library

To start each analysis, we first bring all the required libraries. For repeatable research, this is an important point that must be taken into account. In some cases, I saw that it is not paid attention to, and that library is brought to the place where a special library is needed.

```
#library
library(vtable)
library(caret)
library(janitor)
library(tidyverse)
library(scales)
library(lubridate)
library(RSocrata)
library(tidymodels)
library(themis)
library(baguette)
```

4.3 EDA

In the code below, the names of the variables are modified.

```
class(MGE_drv_acc_veh)
crash_raw <- as_tibble(MGE_drv_acc_veh)
class(crash_raw)
names(crash_raw)
```

```
#names modification
#names(df)
crash_raw <- crash_raw |>
  clean_names("upper_camel", abbreviations = c("ID", "KM"))
names(crash_raw)
```

The code below is not implemented here, but the code chunk is important and useful for quickly checking and reducing variables.

```
(nzv <- nearZeroVar(crash_raw, saveMetrics= TRUE))
dim(crash_raw)
nzv <- nearZeroVar(crash_raw)
names(crash_raw[nzv])
crash_raw <- crash_raw[, -nzv]
dim(crash_raw)
```

Here the research data is prepared and the response variable and predictor variables are selected.

```
crash <- crash_raw %>%
  arrange(desc(AccidentDate)) %>%
  transmute(injuries = if_else(TotalInjMore24H30D > 0, "injuries", "none"),
            AccidentDate,
            Age,
            Sex,
            BeltUse,
            Month,
            Weekdays,
            Hour,
            RoadType,
            TotalVehicles,
            Speed,
            SpeedLimit,
            WeatherCondition,
            LightningCondition,
            SurfCondition,
            AccTypeCollision) %>%
  na.omit()
```

The for loop in the code chunk below is very useful for identifying variables. With a better understanding of the values and distribution of variables, subsequent decisions for each variable, including selection, regrouping, or modification, will be easier.

```r
#df <- crash_raw
df <- crash
df[df == 998 | df == 999] <- NA

for (col in names(df)) {
  uniq_val <- unique(df[[col]])
  n_uniq <- length(uniq_val)
  n_miss <- sum(is.na(df[[col]]))
  if (n_uniq < 100) {
    print(paste("Column:", col, "- Number of unique values:", n_uniq))
    print(paste("Column:", col, "- Number of missing values:", n_miss))
    tbl <- table(df[[col]])
    print(paste("Column:", col, "- Ordered Frequency Table:"))
    print(tbl[order(tbl, decreasing = TRUE)])
  }
}
```

```
[1] "Column: injuries - Number of unique values: 2"
[1] "Column: injuries - Number of missing values: 0"
[1] "Column: injuries - Ordered Frequency Table:"

    none injuries
  131353    11351
[1] "Column: Age - Number of unique values: 99"
[1] "Column: Age - Number of missing values: 961"
[1] "Column: Age - Ordered Frequency Table:"

  42   41   45   43   44   40   46   39   48   47   38   27   36   30   50   28
3617 3483 3479 3438 3430 3417 3405 3327 3258 3192 3144 3112 3078 3061 3046 3038
  29   37   49   26   23   25   31   32   34   33   51   24   35   22   52   21
3037 3031 3015 3009 2999 2969 2963 2953 2898 2879 2850 2821 2814 2777 2776 2676
  54   53   20   55   56   19   57   58   59   60   61   62   18   63   64   65
2676 2623 2587 2389 2209 2178 2122 2033 1828 1725 1566 1544 1301 1291 1147 1018
  66   67   68   69   71   70   17   72   73   74   75   76   16   77   78   79
 896  849  845  755  720  703  666  644  615  574  543  521  502  430  402  334
  80   81   15   83   82   85   84   14   86   87   13   88   12   89   90   11
 315  264  256  248  244  182  170  131  129  108   86   81   64   54   39   21
  91   10    8   92    5   94    7    9   93    6    4   95   96   97   98   99
  20   18   15   12   10    9    7    7    7    5    2    2    2    2    2    1
 100  121
   1    1
[1] "Column: Sex - Number of unique values: 3"
```

```
[1] "Column: Sex - Number of missing values: 372"
[1] "Column: Sex - Ordered Frequency Table:"

     1      2
104878  37454
[1] "Column: BeltUse - Number of unique values: 4"
[1] "Column: BeltUse - Number of missing values: 30478"
[1] "Column: BeltUse - Ordered Frequency Table:"

    1     3     2
85817 15977 10432
[1] "Column: Month - Number of unique values: 12"
[1] "Column: Month - Number of missing values: 0"
[1] "Column: Month - Ordered Frequency Table:"

   10     7     9    11     6    12     8     5     2     1     3     4
13984 13538 12992 12815 12540 12415 12132 11225 10929 10771 10747  8616
[1] "Column: Weekdays - Number of unique values: 7"
[1] "Column: Weekdays - Number of missing values: 0"
[1] "Column: Weekdays - Ordered Frequency Table:"

   Friday  Thursday Wednesday    Monday   Tuesday  Saturday    Sunday
    23788     21546     21120     21045     20968     18769     15468
[1] "Column: Hour - Number of unique values: 24"
[1] "Column: Hour - Number of missing values: 0"
[1] "Column: Hour - Ordered Frequency Table:"

   14    13    18    12    19    15    17    11    16    20     8     9    10
11501 10628 10115  9380  9263  9061  8845  8408  7852  7731  7703  7476  7142
    7    21    22     6    23     0     5     1     2     4     3
 6241  5667  4217  2612  2532  1780  1311  1128   732   705   674
[1] "Column: RoadType - Number of unique values: 14"
[1] "Column: RoadType - Number of missing values: 0"
[1] "Column: RoadType - Ordered Frequency Table:"

    9     6     3     5     1     2    14     8    10     7     4    11    13
45045 44694 24635 12107  6226  3318  2740  1255  1204   728   383   167   106
   12
   96
[1] "Column: TotalVehicles - Number of unique values: 18"
[1] "Column: TotalVehicles - Number of missing values: 0"
[1] "Column: TotalVehicles - Ordered Frequency Table:"
```

```
    2     1     3     4     5     6     7     8    11     9    97    10    13
85605 29790 18506  5632  1840   629   235   141    68    67    63    34    26
   19    22    14    12    18
   19    19    15    12     3
[1] "Column: Speed - Number of unique values: 3"
[1] "Column: Speed - Number of missing values: 8275"
[1] "Column: Speed - Ordered Frequency Table:"

     2      1
115041  19388
[1] "Column: SpeedLimit - Number of unique values: 28"
[1] "Column: SpeedLimit - Number of missing values: 8666"
[1] "Column: SpeedLimit - Ordered Frequency Table:"

   40   100    30    50    80    60    70    90    20   120    10    45    15
26277 25547 19000 18605 17249 11954  6695  3945  2576  1614   428    35    34
   25   110     5    55    81   108    36    48    59    68    69    24    28
   32    22     3     3     3     3     2     2     2     2     2     1     1
   58
    1
[1] "Column: WeatherCondition - Number of unique values: 8"
[1] "Column: WeatherCondition - Number of missing values: 593"
[1] "Column: WeatherCondition - Ordered Frequency Table:"

     1      2      3      4      7      6      5
120150   9262   9188   2243    965    168    135
[1] "Column: LightningCondition - Number of unique values: 7"
[1] "Column: LightningCondition - Number of missing values: 590"
[1] "Column: LightningCondition - Ordered Frequency Table:"

     1      4      6      5      2      3
102977  14841   9764   6642   5115   2775
[1] "Column: SurfCondition - Number of unique values: 10"
[1] "Column: SurfCondition - Number of missing values: 6"
[1] "Column: SurfCondition - Ordered Frequency Table:"

     1      3      8      2      9      7      4      5      6
122872  16286   1294    681    542    384    279    198    162
[1] "Column: AccTypeCollision - Number of unique values: 13"
[1] "Column: AccTypeCollision - Number of missing values: 0"
[1] "Column: AccTypeCollision - Ordered Frequency Table:"

    4     2     3    13     5    10     6     1     7     9    12    11     8
```

```
35549 29423 16470 15225  8134  7563  7483  6737  6706  6346  1924   708   436
```

```
df <- df |> na.omit()
crash <- df
```

4.4 Plot

Figure 1 shows the count of traffic crashes from 2019-2022 by injury and no injury crashes. The top line represents crashes with injuries, and the bottom line represents crashes without injuries.

```
crash %>%
  mutate(AccidentDate = floor_date(AccidentDate, unit = "week")) %>%
  count(AccidentDate, injuries) %>%
  filter(AccidentDate != last(AccidentDate),
         AccidentDate != first(AccidentDate)) %>%
  ggplot(aes(AccidentDate, n, color = injuries)) +
  geom_line(size = 1.5, alpha = 0.7) +
  scale_y_continuous(limits = (c(0, NA))) +
  labs(x = NULL, y = "Traffic crashes per week", color = "Injuries?")
```

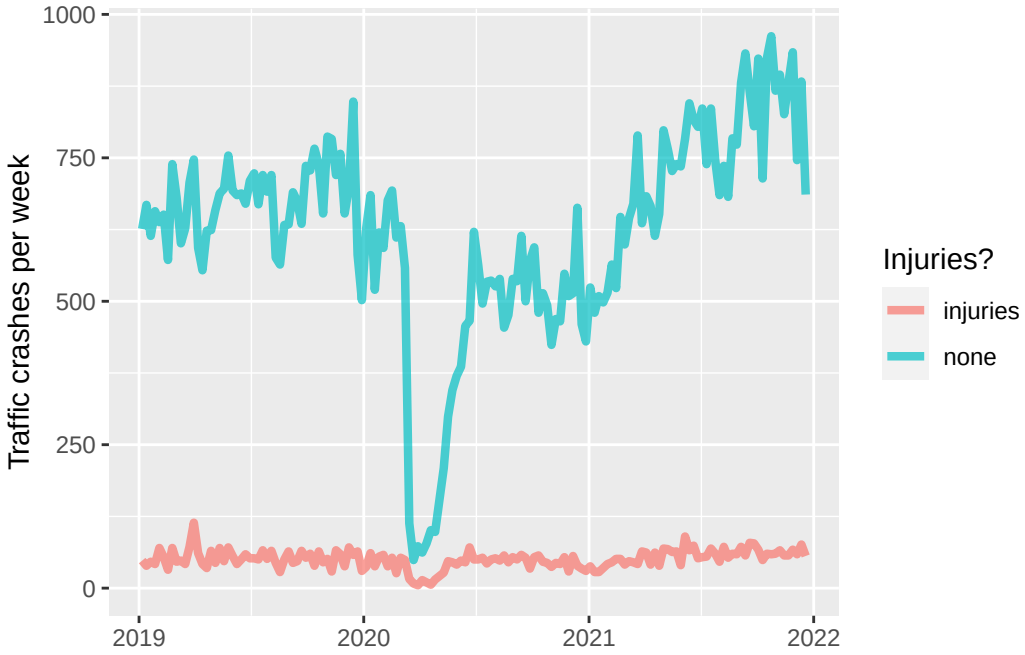

Figure 1: Comparison of injury and no injury crashes between 2019-2022

The sharp drop in the number of accidents in the graph is probably related to the Covid-19 pandemic. Therefore, it may be worth as a research topic, accidents and their influencing factors in this time period, compared with other normal periods.

```
crash %>%
  mutate(AccidentDate = wday(AccidentDate, label = TRUE)) %>%
  count(AccidentDate, injuries) %>%
  group_by(injuries) %>%
  mutate(percent = n / sum(n)) %>%
  ungroup() %>%
  ggplot(aes(n, AccidentDate, fill = injuries)) +
  geom_col(position = "dodge", alpha = 0.8) +
  labs(x = "crashes", y = NULL, fill = "Injuries?")
```

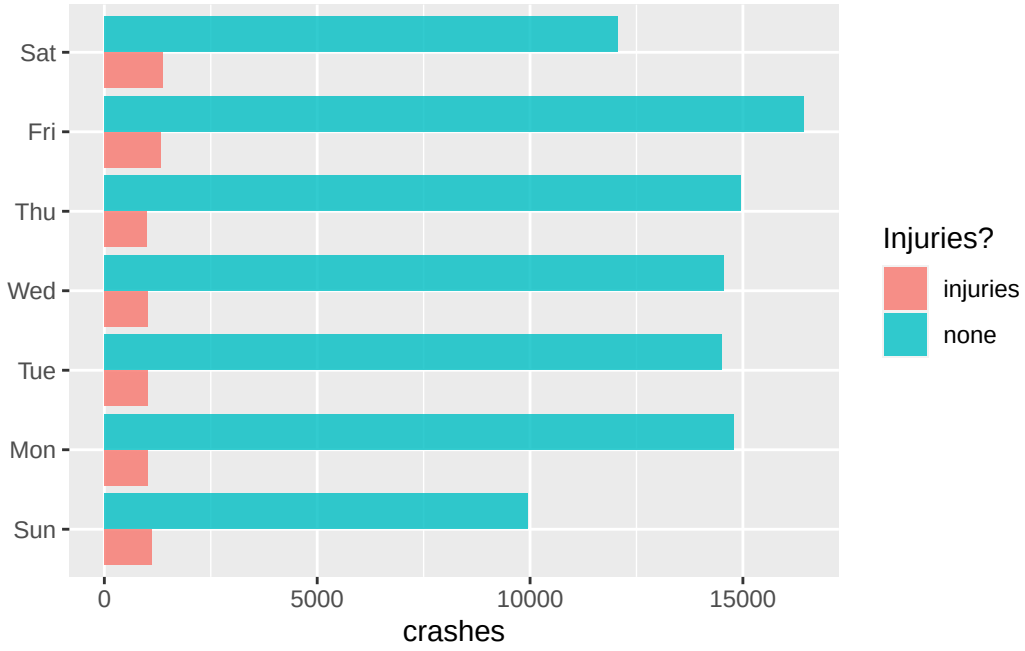

Figure 2: Traffic accidents on weekdays by injury and non-injury

4.5 Build a model

Tidymodel meta-package tools make the modding process easier, and that's why I recommend using them.

```
set.seed(1212)
crash_split <- initial_split(crash, strata = injuries)
crash_train <- training(crash_split)
crash_test <- testing(crash_split)

set.seed(123)
crash_folds <- vfold_cv(crash_train, strata = injuries)
crash_folds
```

```
#  10-fold cross-validation using stratification
# A tibble: 10 x 2
   splits               id
   <list>               <chr>
 1 <split [70961/7885]> Fold01
 2 <split [70961/7885]> Fold02
```

```
 3 <split [70961/7885]> Fold03
 4 <split [70961/7885]> Fold04
 5 <split [70961/7885]> Fold05
 6 <split [70961/7885]> Fold06
 7 <split [70962/7884]> Fold07
 8 <split [70962/7884]> Fold08
 9 <split [70962/7884]> Fold09
10 <split [70962/7884]> Fold10
```

```
names(crash)
```

```
 [1] "injuries"           "AccidentDate"       "Age"
 [4] "Sex"                "BeltUse"            "Month"
 [7] "Weekdays"           "Hour"               "RoadType"
[10] "TotalVehicles"      "Speed"              "SpeedLimit"
[13] "WeatherCondition"   "LightningCondition" "SurfCondition"
[16] "AccTypeCollision"
```

```
crash_rec <- recipe(injuries ~ ., data = crash_train) %>%
  step_downsample(injuries)

bag_spec <- bag_tree(min_n = 10) %>%
  set_engine("rpart", times = 25) %>%
  set_mode("classification")
crash_wf <- workflow() %>%
  add_recipe(crash_rec) %>%
  add_model(bag_spec)

crash_wf
```

```
== Workflow ====================================================================
Preprocessor: Recipe
Model: bag_tree()

-- Preprocessor ----------------------------------------------------------------
1 Recipe Step

* step_downsample()

-- Model -----------------------------------------------------------------------
```

```
Bagged Decision Tree Model Specification (classification)

Main Arguments:
  cost_complexity = 0
  min_n = 10

Engine-Specific Arguments:
  times = 25

Computational engine: rpart
```

```
doParallel::registerDoParallel()
crash_res <- fit_resamples(crash_wf,
                           crash_folds,
                           control = control_resamples(save_pred = TRUE))
```

```
collect_metrics(crash_res)
```

```
# A tibble: 2 x 6
  .metric  .estimator  mean     n std_err .config
  <chr>    <chr>      <dbl> <int>   <dbl> <chr>
1 accuracy binary     0.681    10 0.00168 Preprocessor1_Model1
2 roc_auc  binary     0.757    10 0.00233 Preprocessor1_Model1
```

```
crash_fit <- last_fit(crash_wf, crash_split)
collect_metrics(crash_fit)
```

```
# A tibble: 2 x 4
  .metric  .estimator .estimate .config
  <chr>    <chr>          <dbl> <chr>
1 accuracy binary         0.673 Preprocessor1_Model1
2 roc_auc  binary         0.758 Preprocessor1_Model1
```

Figure 3 shows a variable importance plot. The variable `AccidentDate` has the highest variable importance.

```
crash_imp <- crash_fit$.workflow[[1]] %>%
  pull_workflow_fit()
crash_imp$fit$imp %>%
```

```
  slice_max(value, n = 10) %>%
  ggplot(aes(value, fct_reorder(term, value))) +
  geom_col(alpha = 0.8, fill = "midnightblue") +
  labs(x = "Variable importance score", y = NULL)
```

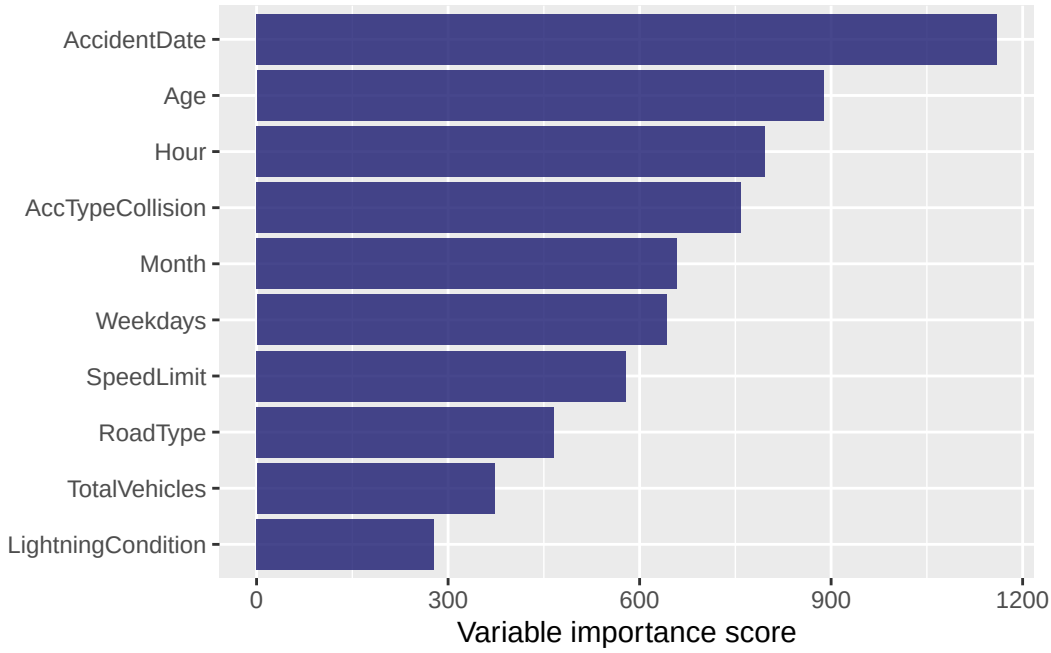

Figure 3: The importance of predictor variables to describe the severity of accidents

Figure 4 shows an ROC curve over a graph of 1-specificity vs. sensitivity.

```
collect_predictions(crash_fit) %>%
  roc_curve(injuries, .pred_injuries) %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity)) +
  geom_line(size = 1.5, color = "midnightblue") +
  geom_abline(lty = 2, alpha = 0.5,     color = "gray50",     size = 1.2) +
  coord_equal()
```

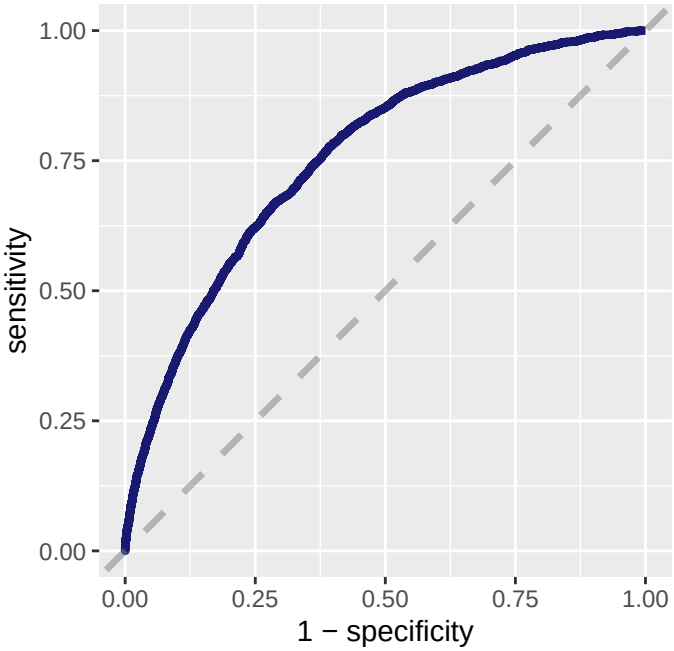

Figure 4: ROC curve

I think the results for the initial modeling are good and promising.

5 Discussion

The analysis and modeling done here is just a small example, but it is the basis of all research work. It is necessary to mention a few points here:

1. Everything goes back to the definition of the problem: modeling should be done on a part of the data according to the definition of the problem (for example, according to the division mentioned in the Section 2).
2. Modeling can be further segmented for an extracted data and the differences in its different parts can be determined. For example, in the beginning, the aim is to investigate the importance of variables in the severity of rear-end accidents in Spain; Then, the goal is to compare the modeling results by gender in such accidents.
3. Fortunately, accident data is now available in different parts of the world. Actual accident data at any particular location is valuable for research. You yourself suggested earlier that a comparative study could be the basis of one of our researches, and I agree with this. Of course, it is a good idea to use data from Iran for comparison.

6 Conclosion

A few points are mentioned as a summary:

1. With the passage of time, it becomes easier and harder to publish articles in high-quality journals! Easier because there are so many research assistant tools available. More difficult because the expectations of journals and audiences have risen: about 5 years ago, only one machine learning modeling was enough to accept an article, but now the model often has to be interpreted as much as possible; In this regard, I recommend reading the book Interpretable Machine Learning.
2. In order to publish a group of articles in any research field, one must first spend a lot of time learning the concepts of that research field. Of course, there is no end to learning. It is also necessary to spend time on training to conduct reproducible research. We moved enough on the runway and reached the right speed for flight, and now it's time to fly. From now, learn everything else in a **specific project** and with a **specific goal**. This saves a lot of time. I hope that if you have the opportunity and spend about 10 hours a day, then within several months you will write at least 5 high-quality articles that are ready for submission.
3. We agreed to publish 10 joint articles. According to the reasons given in the following cases, I suggest that we divide the work: you manage 5 articles with accident data and I manage the other 5 articles. Of course, we present our opinion for each joint article, but you and I do mental, intellectual and practical outsourcing for the article under the management of the other party. Now the reasons:
    - You only need an article related to accident data for your doctoral thesis, and on the other hand, I and other students are researching on various topics and non-accident data.
    - To defend your doctoral thesis, you need articles in which the first and second names are you and your supervisor, and I am the third person in your articles as a consultant professor. On the other hand, I need articles to be the first author for my career promotion. Therefore, for a win-win game for both, it is better to research together, but each of us pursue our own goals.
4. Important point: spend one time of the day only for critical literature reading and creative thinking: always defining a good problem greatly facilitates the path of publishing an article.

Please provide your comments. Remember, I am also eager to learn from you in the fields of publishing articles. Wishing you much success

References

Abay, K.A., Paleti, R., Bhat, C.R., 2013. The joint analysis of injury severity of drivers in two-vehicle crashes accommodating seat belt use endogeneity. Transportation Research Part B: Methodological 50. doi:10.1016/j.trb.2013.01.007

Ahmed, I.U., Gaweesh, S.M., Ahmed, M.M., 2020. Exploration of hazardous material truck crashes onwyoming's interstate roads using a novel hamiltonian monte carlo markov chain bayesian inference. Transportation Research Record 2674. doi:10.1177/0361198120931103

Alhasan, A., Nlenanya, I., Smadi, O., MacKenzie, C.A., 2018. Impact of pavement surface condition on roadway departure crash risk in iowa. Infrastructures 3. doi:10.3390/infrastructures3020014

Amarasingha, N., Dissanayake, S., 2014. Gender differences of young drivers on injury severity outcome of highway crashes. Journal of Safety Research 49. doi:10.1016/j.jsr.2014.03.004

Azimi, G., Rahimi, A., Asgari, H., Jin, X., 2022. Injury severity analysis for large truck-involved crashes: Accounting for heterogeneity, in: Transportation Research Record. doi:10.1177/03611981221091562

Azimi, G., Rahimi, A., Asgari, H., Jin, X., 2020. Severity analysis for large truck rollover crashes using a random parameter ordered logit model. Accident Analysis and Prevention 135. doi:10.1016/j.aap.2019.105355

Behnood, A., Al-Bdairi, N.S.S., 2020. Determinant of injury severities in large truck crashes: A weekly instability analysis. Safety Science 131. doi:10.1016/j.ssci.2020.104911

Behnood, A., Mannering, F., 2019. Time-of-day variations and temporal instability of factors affecting injury severities in large-truck crashes. Analytic Methods in Accident Research 23. doi:10.1016/j.amar.2019.100102

Billah, K., Sharif, H.O., Dessouky, S., 2022. How gender affects motor vehicle crashes: A case study from san antonio, texas. Sustainability (Switzerland) 14. doi:10.3390/su14127023

Boufous, S., Rome, L.D., Senserrick, T., Ivers, R., 2012. Risk factors for severe injury in cyclists involved in traffic crashes in victoria, australia. Accident Analysis and Prevention 49. doi:10.1016/j.aap.2012.03.011

Bullard, C., Jones, S., Adanu, E.K., Liu, J., 2023. Crash severity analysis of single-vehicle rollover crashes in namibia: A mixed logit approach. IATSS Research 47. doi:10.1016/j.iatssr.2023.07.002

Champahom, T., Jomnonkwao, S., Watthanaklang, D., Karoonsoontawong, A., Chatpattananan, V., Ratanavaraha, V., 2020. Applying hierarchical logistic models to compare urban and rural roadway modeling of severity of rear-end vehicular crashes. Accident Analysis and Prevention 141. doi:10.1016/j.aap.2020.105537

Das, S., Dutta, A., Jalayer, M., Bibeka, A., Wu, L., 2018. Factors influencing the patterns of wrong-way driving crashes on freeway exit ramps and median crossovers: Exploration using "eclat" association rules to promote safety. International Journal of Transportation Science and Technology 7. doi:10.1016/j.ijtst.2018.02.001

Dimitrijevic, B., Asadi, R., Spasovic, L., 2023. Application of hybrid support vector machine models in analysis of work zone crash injury severity. Transportation Research Interdisci-

plinary Perspectives 19. doi:10.1016/j.trip.2023.100801

Duddu, V.R., Kukkapalli, V.M., Pulugurtha, S.S., 2019. Crash risk factors associated with injury severity of teen drivers. IATSS Research 43. doi:10.1016/j.iatssr.2018.08.003

Eriksson, J., Niska, A., Forsman, Å., 2022. Injured cyclists with focus on single-bicycle crashes and differences in injury severity in sweden. Accident Analysis and Prevention 165. doi:10.1016/j.aap.2021.106510

Farid, A., Ksaibati, K., 2021. Modeling severities of motorcycle crashes using random parameters. Journal of Traffic and Transportation Engineering (English Edition) 8. doi:10.1016/j.jtte.2020.01.001

Fu, W., Lee, J., Huang, H., 2021. How has the injury severity by gender changed after using female dummy in vehicle testing? Evidence from florida's crash data. Journal of Transport and Health 21. doi:10.1016/j.jth.2021.101073

Ghasemzadeh, A., Ahmed, M.M., 2019. Exploring factors contributing to injury severity at work zones considering adverse weather conditions. IATSS Research 43. doi:10.1016/j.iatssr.2018.11.002

Harris, L., Ahmad, N., Khattak, A., Chakraborty, S., 2023. Exploring the effect of visibility factors on vehicle–pedestrian crash injury severity. Transportation Research Record: Journal of the Transportation Research Board. doi:10.1177/03611981231164070

Holdridge, J.M., Shankar, V.N., Ulfarsson, G.F., 2005. The crash severity impacts of fixed roadside objects. Journal of Safety Research 36. doi:10.1016/j.jsr.2004.12.005

Hossain, M.M., Zhou, H., Sun, X., 2023. A clustering regression approach to explore the heterogeneous effects of risk factors associated with teen driver crash severity. Transportation Research Record: Journal of the Transportation Research Board 2677. doi:10.1177/03611981221150927

Hosseinzadeh, A., Moeinaddini, A., Ghasemzadeh, A., 2021. Investigating factors affecting severity of large truck-involved crashes: Comparison of the SVM and random parameter logit model. Journal of Safety Research 77. doi:10.1016/j.jsr.2021.02.012

Hu, W., Donnell, E.T., 2011. Severity models of cross-median and rollover crashes on rural divided highways in pennsylvania. Journal of Safety Research 42. doi:10.1016/j.jsr.2011.07.004

Iranitalab, A., Kang, Y., Khattak, A., 2018. Modeling the probability of hazardous materials release in crashes at highway–rail grade crossings. Transportation Research Record 2672. doi:10.1177/0361198118780885

Jiang, X., Han, M., Guo, R., Zhang, G., Fan, Y., Li, X., Bai, W., Wei, M., Liang, Q., 2021. Examining the underlying exposures of hit-and-run and non-hit-and-run crashes. Journal of Transport and Health 20. doi:10.1016/j.jth.2020.100995

Kardar, A., Davoodi, S.R., 2020. A generalized ordered probit model for analyzing driver injury severity of head-on crashes on two-lane rural highways in malaysia. Journal of Transportation Safety and Security 12. doi:10.1080/19439962.2019.1571550

Keramati, A., Lu, P., Iranitalab, A., Pan, D., Huang, Y., 2020. A crash severity analysis at highway-rail grade crossings: The random survival forest method. Accident Analysis and Prevention 144. doi:10.1016/j.aap.2020.105683

Khan, I.U., Vachal, K., 2020. Factors affecting injury severity of single-vehicle rollover crashes in the united states. Traffic Injury Prevention 21. doi:10.1080/15389588.2019.1696962

Kim, J.M., Kim, S.C., Lee, K.H., Kim, H.J., Kim, H., Lee, S.W., Na, D.S., Park, J.S., 2021. Preventive effects of seat belts on traumatic brain injury in motor vehicle collisions classified by crash severities and collision directions. European Journal of Trauma and Emergency Surgery 47. doi:10.1007/s00068-019-01095-4

Kim, M., Kho, S.Y., Kim, D.K., 2017. Hierarchical ordered model for injury severity of pedestrian crashes in south korea. Journal of Safety Research 61. doi:10.1016/j.jsr.2017.02.011

Kitali, A.E., Kidando, E., Alluri, P., Sando, T., Salum, J.H., 2022. Modeling severity of motorcycle crashes with dirichlet process priors. Journal of Transportation Safety and Security. doi:10.1080/19439962.2020.1738613

Kutela, B., Kitali, A.E., Kidando, E., Mbuya, C., Langa, N., 2023. Exploring the need to model severity of single- and multi-occupant vehicles crashes separately: A case of crashes at highway-rail grade crossings. International Journal of Transportation Science and Technology. doi:10.1016/j.ijtst.2022.11.002

Lasota, D., Goniewicz, M., Kosson, D., Ochal, A., Krajewski, P., Tarka, S., Goniewicz, K., Mirowska-Guzel, D., 2020. Effects of ethyl alcohol on injuries severity according to injury severity scales in pedestrian fatal injury in traffic crashes. International Journal of Injury Control and Safety Promotion 27. doi:10.1080/17457300.2019.1665551

Lee, D., Guldmann, J.M., Rabenau, B. von, 2023. Impact of driver's age and gender, built environment, and road conditions on crash severity: A logit modeling approach. International Journal of Environmental Research and Public Health 20. doi:10.3390/ijerph20032338

Li, J., Liu, J., Liu, P., Qi, Y., 2020. Analysis of factors contributing to the severity of large truck crashes. Entropy 22. doi:10.3390/e22111191

Li, N., Park, B.B., Lambert, J.H., 2018. Effect of guardrail on reducing fatal and severe injuries on freeways: Real-world crash data analysis and performance assessment. Journal of Transportation Safety and Security 10. doi:10.1080/19439962.2017.1297970

Lin, Z., Fan, W., 2021. Cyclist injury severity analysis with mixed-logit models at intersections and nonintersection locations. Journal of Transportation Safety and Security 13. doi:10.1080/19439962.2019.1628140

Liu, P., Fan, W., 2020. Modeling head-on crash severity with drivers under the influence of alcohol or drugs (DUI) and non-DUI. Traffic Injury Prevention 21. doi:10.1080/15389588.2019.1696964

Lu, G., Chitturi, M.V., Ooms, A.W., Noyce, D.A., 2010. Ordinal discrete choice analyses of wisconsin cross-median crash severities. Transportation Research Record. doi:10.3141/2148-06

Ma, J., Ren, G., Li, H., Wang, S., Yu, J., 2023. Characterizing the differences of injury severity between single-vehicle and multi-vehicle crashes in china. Journal of Transportation Safety and Security 15. doi:10.1080/19439962.2022.2056931

Mafi, S., AbdelRazig, Y., Doczy, R., 2018. Machine learning methods to analyze injury severity of drivers from different age and gender groups. Transportation Research Record 2672. doi:10.1177/0361198118794292

Mohamed, S.A., Mohamed, K., Al-Harthi, H.A., 2017. Investigating factors affecting the occurrence and severity of rear-end crashes. Transportation Research Procedia 25. doi:10.1016/j.trpro.2017.05.403

Naik, B., Tung, L.W., Zhao, S., Khattak, A.J., 2016. Weather impacts on single-vehicle truck crash injury severity. Journal of Safety Research 58. doi:10.1016/j.jsr.2016.06.005

Okafor, S., Adanu, E.K., Jones, S., 2022. Severity analysis of crashes involving in-state and out-of-state large truck drivers in alabama: A random parameter multinomial logit model with heterogeneity in means and variances. Heliyon 8. doi:10.1016/j.heliyon.2022.e11989

Peng, Y., Geedipally, S., Lord, D., 2012. Effect of roadside features on single-vehicle roadway departure crashes on rural two-lane roads. Transportation Research Record. doi:10.3141/2309-03

Prajongkha, P., Kanitpong, K., Jensupakarn, A., 2023. Factors contributing to the severity of motorcycle rear-end crashes in thailand. Traffic Injury Prevention 24. doi:10.1080/15389588.2022.2127320

Rezapour, M., Ksaibati, K., 2022. Contributory factors to the severity of single-vehicle rollover crashes on a mountainous area, generalized additive model. International Journal of Injury Control and Safety Promotion 29. doi:10.1080/17457300.2021.2011927

Rezapour, M., Ksaibati, K., 2021. Modeling crashes involving children, finite mixture cumulative link mixed model. International Journal of Injury Control and Safety Promotion 28. doi:10.1080/17457300.2021.1964088

Rezapour, M., Nazneen, S., Ksaibati, K., 2020. Application of deep learning techniques in predicting motorcycle crash severity. Engineering Reports 2. doi:10.1002/eng2.12175

Roque, C., Jalayer, M., Hasan, A.S., 2021. Investigation of injury severities in single-vehicle crashes in north carolina using mixed logit models. Journal of Safety Research 77. doi:10.1016/j.jsr.2021.02.013

Russo, B.J., Yu, F., Smaglik, E.J., 2023. Examination of factors associated with fault status and injury severity in intersection-related rear-end crashes: Application of binary and bivariate ordered probit models. Safety Science 164. doi:10.1016/j.ssci.2023.106187

Salum, J.H., Kitali, A.E., Bwire, H., Sando, T., Alluri, P., 2019. Severity of motorcycle crashes in dar es salaam, tanzania. Traffic Injury Prevention 20. doi:10.1080/15389588.2018.1544706

Sawtelle, A., Shirazi, M., Garder, P.E., Rubin, J., 2023. Driver, roadway, and weather factors on severity of lane departure crashes in maine. Journal of Safety Research 84. doi:10.1016/j.jsr.2022.11.006

Shaaban, K., Gharraie, I., Sacchi, E., Kim, I., 2021. Severity analysis of red-light-running-related crashes using structural equation modeling. Journal of Transportation Safety and Security 13. doi:10.1080/19439962.2019.1629137

Shao, X., Ma, X., Chen, F., Song, M., Pan, X., You, K., 2020. A random parameters ordered probit analysis of injury severity in truck involved rear-end collisions. International Journal of Environmental Research and Public Health 17. doi:10.3390/ijerph17020395

Sharafeldin, M., Farid, A., Ksaibati, K., 2022. Injury severity analysis of rear-end crashes at signalized intersections. Sustainability (Switzerland) 14. doi:10.3390/su142113858

Shen, X., Wei, S., 2021. Severity analysis of road transport accidents of hazardous materials with machine learning. Traffic Injury Prevention 22. doi:10.1080/15389588.2021.1900569

Shyhalla, K., 2014. Alcohol involvement and other risky driver behaviors: Effects on crash initiation and crash severity. Traffic Injury Prevention 15. doi:10.1080/15389588.2013.822491

Sivasankaran, S.K., Balasubramanian, V., 2022. Investigation of factors contributing to pedes-

trian hit-and-run crashes in india. Journal of Transportation Safety and Security 14. doi:10.1080/19439962.2020.1781313

Sivasankaran, S.K., Rangam, H., Balasubramanian, V., 2021. Investigation of factors contributing to injury severity in single vehicle motorcycle crashes in india. International Journal of Injury Control and Safety Promotion 28. doi:10.1080/17457300.2021.1908367

Song, D., Yang, X., Yang, Y., Cui, P., Zhu, G., 2023. Bivariate joint analysis of injury severity of drivers in truck-car crashes accommodating multilayer unobserved heterogeneity. Accident Analysis and Prevention 190. doi:10.1016/j.aap.2023.107175

Song, L., Fan, W. (David), Li, Y., 2021. Time-of-day variations and the temporal instability of multi-vehicle crash injury severities under the influence of alcohol or drugs after the great recession. Analytic Methods in Accident Research 32. doi:10.1016/j.amar.2021.100183

Sun, M., Zhou, R., Jiao, C., Sun, X., 2022. Severity analysis of hazardous material road transportation crashes with a bayesian network using highway safety information system data. International Journal of Environmental Research and Public Health 19. doi:10.3390/ijerph19074002

Theofilatos, A., Antoniou, C., Yannis, G., 2021. Exploring injury severity of children and adolescents involved in traffic crashes in greece. Journal of Traffic and Transportation Engineering (English Edition) 8. doi:10.1016/j.jtte.2020.07.005

Villavicencio, L., Svancara, A.M., Kelley-Baker, T., Tefft, B.C., 2022. Passenger presence and the relative risk of teen driver death. Journal of Adolescent Health 70. doi:10.1016/j.jadohealth.2021.10.038

Wahab, L., Jiang, H., 2020. Severity prediction of motorcycle crashes with machine learning methods. International Journal of Crashworthiness 25. doi:10.1080/13588265.2019.1616885

Wen, H., Ma, Z., Chen, Z., Luo, C., 2023. Analyzing the impact of curve and slope on multi-vehicle truck crash severity on mountainous freeways. Accident Analysis and Prevention 181. doi:10.1016/j.aap.2022.106951

Wu, J., Rasouli, S., Zhao, J., Qian, Y., Cheng, L., 2023. Large truck fatal crash severity segmentation and analysis incorporating all parties involved: A bayesian network approach. Travel Behaviour and Society 30. doi:10.1016/j.tbs.2022.09.003

Wu, Q., Zhang, G., 2018. Formulating alcohol-influenced driver's injury severities in intersection-related crashes. Transport 33. doi:10.3846/16484142.2016.1144221

Xing, Y., Chen, S., Zhu, S., Zhang, Y., Lu, J., 2020. Exploring risk factors contributing to the severity of hazardous material transportation accidents in china. International Journal of Environmental Research and Public Health 17. doi:10.3390/ijerph17041344

Yan, X., He, J., Wu, G., Zhang, C., Wang, C., Ye, Y., 2022. Differences of overturned and hit-fixed-object crashes on rural roads accompanied by speeding driving: Accommodating potential temporal shifts. Analytic Methods in Accident Research 35. doi:10.1016/j.amar.2022.100220

Yazdani, M., Nassiri, H., 2021. The effect of weather on the severity of multi-vehicle crashes: A case study of iran. Proceedings of the Institution of Civil Engineers: Transport 174. doi:10.1680/jtran.18.00080

Yu, M., Ma, C., Shen, J., 2021. Temporal stability of driver injury severity in single-vehicle roadway departure crashes: A random thresholds random parameters hi-

erarchical ordered probit approach. Analytic Methods in Accident Research 29. doi:10.1016/j.amar.2020.100144

Yu, M., Zheng, C., Ma, C., 2020. Analysis of injury severity of rear-end crashes in work zones: A random parameters approach with heterogeneity in means and variances. Analytic Methods in Accident Research 27. doi:10.1016/j.amar.2020.100126

Zafri, N.M., Prithul, A.A., Baral, I., Rahman, M., 2020. Exploring the factors influencing pedestrian-vehicle crash severity in dhaka, bangladesh. International Journal of Injury Control and Safety Promotion 27. doi:10.1080/17457300.2020.1774618

Zeng, Q., Hao, W., Lee, J., Chen, F., 2020. Investigating the impacts of real-time weather conditions on freeway crash severity: A bayesian spatial analysis. International Journal of Environmental Research and Public Health 17. doi:10.3390/ijerph17082768

Zhai, X., Huang, H., Sze, N.N., Song, Z., Hon, K.K., 2019. Diagnostic analysis of the effects of weather condition on pedestrian crash severity. Accident Analysis and Prevention 122. doi:10.1016/j.aap.2018.10.017

Zhang, G., Tan, Y., Zhong, Q., Hu, R., 2021. Analysis of traffic crashes caused by motorcyclists running red lights in guangdong province of china. International Journal of Environmental Research and Public Health 18. doi:10.3390/ijerph18020553

Zhou, B., Li, Z., Zhang, S., 2018. Comparison of factors affecting crash severities in hit-and-run and non-hit-and-run crashes. Journal of Advanced Transportation 2018. doi:10.1155/2018/8537131

Zhou, B., Wang, X., Zhang, S., Li, Z., Sun, S., Shu, K., Sun, Q., 2020. Comparing factors affecting injury severity of passenger car and truck drivers. IEEE Access 8. doi:10.1109/ACCESS.2020.3018183

Zhu, S., 2022. Analyse vehicle–pedestrian crash severity at intersection with data mining techniques. International Journal of Crashworthiness 27. doi:10.1080/13588265.2021.1929002

Zhu, S., Wan, J., 2021. Cost-sensitive learning for semi-supervised hit-and-run analysis. Accident Analysis and Prevention 158. doi:10.1016/j.aap.2021.106199

Zhu, X., Srinivasan, S., 2011. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. Accident Analysis and Prevention 43. doi:10.1016/j.aap.2010.07.007

Online appendix

6.1 Attach R session info in appendix

Since R and R packages are constantly evolving you might want to add the R session info that contains information on the R version as well as the packages that are loaded.

```
R version 4.2.2 (2022-10-31 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19045)

Matrix products: default

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] rpart_4.1.19       baguette_1.0.1     themis_1.0.1       yardstick_1.2.0
 [5] workflowsets_1.0.1 workflows_1.1.3    tune_1.1.1         rsample_1.1.1
 [9] recipes_1.0.7      parsnip_1.1.0      modeldata_1.2.0    infer_1.0.4
[13] dials_1.2.0        broom_1.0.5        tidymodels_1.1.0   RSocrata_1.7.12-4
[17] scales_1.2.1       lubridate_1.9.2    forcats_1.0.0      stringr_1.5.0
[21] dplyr_1.1.2        purrr_1.0.2        readr_2.1.4        tidyr_1.3.0
[25] tibble_3.2.1       tidyverse_2.0.0    janitor_2.2.0      caret_6.0-94
[29] lattice_0.21-8     ggplot2_3.4.2      vtable_1.4.4       kableExtra_1.3.4

loaded via a namespace (and not attached):
 [1] Cubist_0.4.2.1       colorspace_2.1-0     ellipsis_0.3.2
 [4] class_7.3-20         snakecase_0.11.0     rstudioapi_0.15.0
 [7] farver_2.1.1         listenv_0.9.0        furrr_0.3.1
[10] earth_5.3.2          mvtnorm_1.2-2        prodlim_2023.03.31
[13] fansi_1.0.4          xml2_1.3.5           codetools_0.2-18
[16] splines_4.2.2        doParallel_1.0.17    libcoin_1.0-9
[19] knitr_1.43           Formula_1.2-5        jsonlite_1.8.7
[22] pROC_1.18.4          compiler_4.2.2       httr_1.4.6
[25] backports_1.4.1      Matrix_1.6-0         fastmap_1.1.1
[28] cli_3.6.1            htmltools_0.5.6      tools_4.2.2
[31] partykit_1.2-20      gtable_0.3.3         glue_1.6.2
[34] reshape2_1.4.4       Rcpp_1.0.11          DiceDesign_1.9
[37] vctrs_0.6.3          svglite_2.1.1        nlme_3.1-160
[40] iterators_1.0.14     inum_1.0-5           timeDate_4022.108
[43] gower_1.0.1          xfun_0.40            globals_0.16.2
```

```
[46] rvest_1.0.3          timechange_0.2.0     mime_0.12
[49] lifecycle_1.0.3      future_1.33.0        MASS_7.3-58.1
[52] ipred_0.9-14         hms_1.1.3            parallel_4.2.2
[55] yaml_2.3.7           C50_0.1.8            TeachingDemos_2.12
[58] stringi_1.7.12       plotrix_3.8-2        foreach_1.5.2
[61] butcher_0.3.2        lhs_1.1.6            hardhat_1.3.0
[64] lava_1.7.2.1         rlang_1.1.1          pkgconfig_2.0.3
[67] systemfonts_1.0.4    evaluate_0.21        labeling_0.4.2
[70] tidyselect_1.2.0     parallelly_1.36.0    plyr_1.8.8
[73] magrittr_2.0.3       R6_2.5.1             generics_0.1.3
[76] pillar_1.9.0         withr_2.5.0          survival_3.4-0
[79] nnet_7.3-18          future.apply_1.11.0  ROSE_0.0-4
[82] utf8_1.2.3           tzdb_0.4.0           rmarkdown_2.24
[85] grid_4.2.2           data.table_1.14.8    plotmo_3.6.2
[88] ModelMetrics_1.2.2.2 digest_0.6.33        webshot_0.5.5
[91] stats4_4.2.2         munsell_0.5.0        GPfit_1.0-8
[94] viridisLite_0.4.2
```

6.2 All the code in the paper

To simply attach all the code you used in the PDF file in the appendix see the R chunk in the underlying `.qmd` file:

```r
knitr::opts_chunk$set(cache = FALSE)
# Use cache = TRUE if you want to speed up compilation

knitr::opts_knit$set(output.format = "html")  # Set to "html" for HTML output

# A function to allow for showing some of the inline code
rinline <- function(code){
  html <- '<code  class="r">``` `r CODE` ```</code>'
  sub("CODE", code, html)
}
load("RawData.RData")
#library
library(vtable)
library(caret)
library(janitor)
library(tidyverse)
library(scales)
library(lubridate)
```

```r
library(RSocrata)
library(tidymodels)
library(themis)
library(baguette)
class(MGE_drv_acc_veh)
crash_raw <- as_tibble(MGE_drv_acc_veh)
class(crash_raw)
names(crash_raw)
#names modification
#names(df)
crash_raw <- crash_raw |>
  clean_names("upper_camel", abbreviations = c("ID", "KM"))
names(crash_raw)
(nzv <- nearZeroVar(crash_raw, saveMetrics= TRUE))
dim(crash_raw)
nzv <- nearZeroVar(crash_raw)
names(crash_raw[nzv])
crash_raw <- crash_raw[, -nzv]
dim(crash_raw)
crash <- crash_raw %>%
  arrange(desc(AccidentDate)) %>%
  transmute(injuries = if_else(TotalInjMore24H30D > 0, "injuries", "none"),
            AccidentDate,
            Age,
            Sex,
            BeltUse,
            Month,
            Weekdays,
            Hour,
            RoadType,
            TotalVehicles,
            Speed,
            SpeedLimit,
            WeatherCondition,
            LightningCondition,
            SurfCondition,
            AccTypeCollision) %>%
  na.omit()


#df <- crash_raw
```

```r
df <- crash
df[df == 998 | df == 999] <- NA

for (col in names(df)) {
  uniq_val <- unique(df[[col]])
  n_uniq <- length(uniq_val)
  n_miss <- sum(is.na(df[[col]]))
  if (n_uniq < 100) {
    print(paste("Column:", col, "- Number of unique values:", n_uniq))
    print(paste("Column:", col, "- Number of missing values:", n_miss))
    tbl <- table(df[[col]])
    print(paste("Column:", col, "- Ordered Frequency Table:"))
    print(tbl[order(tbl, decreasing = TRUE)])
  }
}

df <- df |> na.omit()
crash <- df

crash %>%
  mutate(AccidentDate = floor_date(AccidentDate, unit = "week")) %>%
  count(AccidentDate, injuries) %>%
  filter(AccidentDate != last(AccidentDate),
         AccidentDate != first(AccidentDate)) %>%
  ggplot(aes(AccidentDate, n, color = injuries)) +
  geom_line(size = 1.5, alpha = 0.7) +
  scale_y_continuous(limits = (c(0, NA))) +
  labs(x = NULL, y = "Traffic crashes per week", color = "Injuries?")
crash %>%
  mutate(AccidentDate = wday(AccidentDate, label = TRUE)) %>%
  count(AccidentDate, injuries) %>%
  group_by(injuries) %>%
  mutate(percent = n / sum(n)) %>%
  ungroup() %>%
  ggplot(aes(n, AccidentDate, fill = injuries)) +
  geom_col(position = "dodge", alpha = 0.8) +
  labs(x = "crashes", y = NULL, fill = "Injuries?")
set.seed(1212)
crash_split <- initial_split(crash, strata = injuries)
crash_train <- training(crash_split)
crash_test <- testing(crash_split)
```

```r
set.seed(123)
crash_folds <- vfold_cv(crash_train, strata = injuries)
crash_folds
names(crash)

crash_rec <- recipe(injuries ~ ., data = crash_train) %>%
  step_downsample(injuries)

bag_spec <- bag_tree(min_n = 10) %>%
  set_engine("rpart", times = 25) %>%
  set_mode("classification")
crash_wf <- workflow() %>%
  add_recipe(crash_rec) %>%
  add_model(bag_spec)

crash_wf
doParallel::registerDoParallel()
crash_res <- fit_resamples(crash_wf,
                           crash_folds,
                           control = control_resamples(save_pred = TRUE))
collect_metrics(crash_res)
crash_fit <- last_fit(crash_wf, crash_split)
collect_metrics(crash_fit)
crash_imp <- crash_fit$.workflow[[1]] %>%
  pull_workflow_fit()
crash_imp$fit$imp %>%
  slice_max(value, n = 10) %>%
  ggplot(aes(value, fct_reorder(term, value))) +
  geom_col(alpha = 0.8, fill = "midnightblue") +
  labs(x = "Variable importance score", y = NULL)
collect_predictions(crash_fit) %>%
  roc_curve(injuries, .pred_injuries) %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity)) +
  geom_line(size = 1.5, color = "midnightblue") +
  geom_abline(lty = 2, alpha = 0.5,     color = "gray50",     size = 1.2) +
  coord_equal()
print(sessionInfo(), local = FALSE)
```