Statistical Machine Learning for Risk and Actuarial Applications
Assignment: Fatal Road Crashes in Victoria
T2 2023

Due time: Friday 28th July 2023 11.55 am (sharp)

1 Skills Developed

This assignment provides you with an opportunity to apply techniques you have learnt in the course to a business task involving data. In addition, your skills in understanding and applying advanced research works (including the text and any additional reference material you consider) will be developed via this assignment. Communication of the results of your investigations and analysis is also an important skill developed.

2 Task

You are an analyst at a data analytics consulting firm. Your manager has tasked you with providing a report for VicRoads, the agency of the state of Victoria responsible for road management and road safety. One of VicRoads' functions is supporting the development of strategies to improve the safety of the road system.

For this particular project, VicRoads wants to understand what drives fatal accidents so that they can design appropriate policies to reduce fatalities on the roads in Victoria. They also want to take a proactive approach to prevention and want to produce a predictive model to identify drivers who are more likely to suffer a fatal accident so that they can target an information campaign.

VicRoad has provided your consulting firm with the dataset `VicRoadFatalData.csv` of road crashes with information about accidents including the road and weather conditions, the drivers and vehicles involved in the crash, and whether the accident was fatal or not. To inform these strategies your manager has asked you to do the following tasks and write a short report to comment and summarise on your results:

1. Perform a brief exploratory data analysis of the characteristics of the accidents in the dataset.
2. Build a model to analyse the relationship between fatal accidents and the different variables available in the dataset. In particular, but not limited to, the study of differences in fatality outcomes based on the road and weather conditions at the time of the accident, and the characteristics of the drivers and vehicles involved.
3. Build a predictive model based on the demographics of the drivers and their vehicles to predict which drivers are more likely to be engaged in a fatal accident. In particular, out of 10,000 drivers in the `Drivers_Eval.csv` evaluation dataset you must identify 2,500 drivers who are the most likely to be involved in a fatal accident and who could be the target of a prevention campaign.

See more details on the tasks below.

3 Additional information and mark allocation

3.1 Data

The dataset consists of real road crashes in Victoria obtained from the CrashStats datasets provided by VicRoads. The CrashStats data allows us to analyse serious vehicle crashes based on time, location, conditions, crash type, road user type, object hit etc.

The training dataset `VicRoadFatalData.csv` has been prepared by linking and cleaning several of the files available in CrashStats. This dataset includes data for 200,000 drivers involved in road crashes in Victoria between 2006 and 2020. For each of the 200,000 instances we have information on 27 variables describing the driver, the driver's vehicle, and other accident-related information (e.g., time, road conditions, etc.). These are described below:

Driver data

FIELD NAME	FIELD DEFINITION	FIELD DOMAIN
`DRIVER_ID`	Unique identifier for each driver	Text
`SEX`	Sex or gender of the driver	Male (`M`), Female (`F`), Unknown (`U`)
`AGE`	Age of the driver at the time of the accident	Integer
`Age Group`	Age group that the driver falls into	16–17, 18–21, etc.
`LICENCE_STATE`	State where the driver's license is registered	Victoria, Other
`HELMET_BELT_WORN`	Whether a helmet or seatbelt was worn by the driver	Seatbelt worn, Seatbelt not worn, Other

Vehicle data

FIELD NAME	FIELD DEFINITION	FIELD DOMAIN
`VEHICLE_ID`	Unique identifier for each vehicle	Text
`VEHICLE_YEAR_MANUF`	The year the vehicle was manufactured	Year
`VEHICLE_BODY_STYLE`	The body style of the vehicle	Sedan, Coupe, etc.
`VEHICLE_MAKE`	The make of the vehicle	Toyota, Ford, etc.
`VEHICLE_TYPE`	The type or category of vehicle	Car, Taxi, etc.
`FUEL_TYPE`	The type of fuel the vehicle uses	Petrol, Diesel, etc.
`VEHICLE_COLOUR`	The colour of the vehicle	Various colors
`OWNER_POSTCODE`	The postcode of the vehicle's owner	Postcode
`TOTAL_NO_OCCUPANTS`	The total number of occupants in the vehicle at the time of the accident	Integer

Other Accident data

FIELD NAME	FIELD DEFINITION	FIELD DOMAIN
`ACCIDENT_NO`	Unique identifier for each accident	Text
`ACCIDENTDATE`	The date of the accident	Date
`ACCIDENTTIME`	The time of the accident	Time
`DAY_OF_WEEK`	The day of the week when the accident occurred	Monday, Tuesday, etc.
`ACCIDENT_TYPE`	The type of accident	Various types
`LIGHT_CONDITION`	The light condition at the time of the accident	Day, Dark Street lights on, etc.

FIELD NAME	FIELD DEFINITION	FIELD DOMAIN
`ROAD_GEOMETRY`	The layout of the road where the accident occurred	Various types
`SPEED_ZONE`	The speed limit in the area where the accident occurred	Various speed limits
`SURFACE_COND`	The condition of the road's surface at the time of the accident	Wet, Dry, Other.
`ATMOSPH_COND`	The atmospheric condition at the time of the accident	Clear, Raining, Fog, Other
`ROAD_SURFACE_TYPE`	The type of road surface where the accident occurred	Paved, Unpaved, Gravel

Output variable (desired target)

FIELD NAME	FIELD DEFINITION	FIELD DOMAIN
fatal	Whether the accident was fatal or not	Yes (`TRUE`), No (`FALSE`)

You can see further information about the variables in the metadata file provided by VicRoads.

These data has been preprocessed and does not have any missing values. However, these data have several characteristics which could make the design of the some of the models challenging:

- *Imbalanced data*: only about 2% of all accidents in the training data set were fatal. This may be especially relevant for the prediction task. You may wish to have a look at the paper by He and Garcia (2009) for strategies for dealing with imbalanced data in classification problems.
- *Categorical variables with many levels*: Some of the variables in the dataset may have many levels which could be problematic if you try to use them in some model. For example, the variable `OWNER_POSTCODE` has 1500+ possible postcodes. You can choose to ignore some of these variables. Alternatively, you can reduce the number of levels by doing groupings of similar levels.

3.2 Analysis, Modelling and Discussion [85 Marks]

Mark allocation for the assignment can be found in the rubric attached, and also refer to the information below for more details on the tasks.

Note that, as in any consulting exercise, there are many alternative valid approaches that can be used. You can choose how to perform the tasks as long as they are justifiable and justified. What is important is the rationale for your chosen methods, and the associated insights and recommendations.

You may also wish to engage in extra research beyond what is covered in the course – please feel free to do so. Although the marks for each component of the assignment are capped, innovations will be encouraged and will potentially offset issues if present. Note however that it is possible to attain full marks without significant innovation.

3.2.1 Exploratory data analysis of the accidents, drivers and vehicles [10 Marks]

For this part you should do exploratory data analysis (EDA) of the dataset to summarise the conditions at the time of the accidents, the profile of the drivers, and characteristics of the vehicles. Your presentation of the different EDA summary metrics should be accompanied by a discussion of the insights you get from the summary metrics.

The StoryWall formative activity for Week 5 also involves EDA of this dataset. You can use the answers and discussions from that activity as an starting point for this task.

3.2.2 Modelling the relationship between fatal accidents and the different variables [35 Marks]

For this part you should build a generalised linear model to understand the relationship between fatal accidents and the different variables available in the `VicRoadFatalData.csv` dataset.

You should provide in the main report the results a detailed analysis using your selected model, along with justification of why the particular model was chosen. Your analysis should also be accompanied by a discussion of the insights you get from the model. Provide details about the construction of the model in the technical appendix

3.2.3 Predictive modelling of drivers and fatal accidents [40 Marks]

Due to its constant interaction with drivers, VicRoads can access data on the demographics of the drivers and their vehicles before they are engaged in a fatal accident.

For this part you should develop a predictive model based *only* on the demographics of the drivers and their vehicles to predict which drivers are more likely to be engaged in a fatal accident.[1]

Using the driver and vehicle data you have (see section on data for details), develop predictive models using various methods such us (but not limited to): logistic regression, k-nearest neighbours, logistic regression with lasso and ridge, classification trees and their extensions. Provide the results and analysis associated with each of these methods in the technical appendix; this should include discussion on the choice of the tuning parameter(s). A very brief summary of each approach should also be included. f Note that you should also provide in the main report the results and a detailed analysis using your selected predictive model, along with justification of why the particular model was chosen.

The evaluation dataset (`Drivers_Eval.csv`) comprises 10,000 current drivers who could be targeted for the prevention campaign. From this 10,000 drivers you have to select 2,500 who you believe are the most likely to be involved in a fatal accident and who could be the target of a prevention campaign (see the submission section for details). This evaluation dataset has the same format as the `VicRoadFatalData.csv` training dataset but doesn't include the column `fatal` nor the accident-related variables. We know the column `fatal` for the dataset and we will release it after the due date of the assignment.

Your should also provide in the main report a non-technical description of the characteristic of the drivers you are suggesting to target for the prevention campaign.

3.2.3.1 Marks for predictive performance The quality of your predictions on the evaluation data will have a (minor) impact on your mark. More specifically, 10 of the 40 marks of the prediction task will be associated to the accuracy of your predictions. That is, your accuracy score a is the fraction of the $n = 2,500$ drivers you selected which were in fact in fatal accidents.

The marks you will get for the accuracy criterion will be given by the following:

$$\text{Marks} = \begin{cases} \frac{4}{40} \times n \times a & \text{if } a \times n \leq 40 \\ 4 + \frac{6}{C-50}(a \times n - 40) & \text{if } a \times n \geq 40 \end{cases}$$

where C is maximum number of drivers successfully identified as being in a fatal accident by a student in the class.

The number of correct drivers in fatal accidents you would identify if your prediction algorithm were to just randomly sample the 10,000 clients is 40 (on average). Therefore, if your number of correct predictions is below 40, your mark will between 0 and 4. If your number of correct predictions is above 40, then your mark is scaled from 4 up to 10.

[1] Please note that, with the data at our disposal, we are not able to predict which drivers will be involved in accidents. Instead, we are capable of predicting the conditional probability that, given a driver is involved in an accident, the accident results in fatalities.

3.3 Presentation Format and Communication [15 Marks]

Communication of quantitative results in a concise and easy-to-read manner is a skill that is vital in practice. As such, marks will be given for the presentation of your results. In order to maximise your presentation marks you may wish to consider issues such as: table size/readability, figure axis/formatting, ease of reading, grammar/spelling, and report structure. You may also wish to consider the use of executive summaries and appendixes, where appropriate. Provide sufficient details in the main body of the reader so that they can judge what you are doing, using appendices for non-essential but useful results as necessary.

In writing the report, you can assume that the client and your manager are both familiar with the basics of statistical learning. Note that sufficient detail must be provided (in either the report body and/or appendices) so that the reviewer can follow all the steps and derivations required in your work.

Note that a **maximum page limit of 6 pages** (including tables and graphs but excluding references) is applicable to the main body of the report.[2] You should also consider the rubric for the presentation component. There is no limit to the size of the appendix. Furthermore your answer should satisfy the following formatting requirements: (i) font: Times, 12 pt or equivalent size and (ii) margins: all four of at least 2 cm.

3.4 Software

You may choose which software language to use (e.g., `R`, `Python`, or other), however, nearly every function you will be required to use for this task is available in `R`. Note also that code enabling you to perform most of the modelling can be found in the learning activities of the course. If you use any simplifying assumptions in your modelling, they must be clearly identified and justified.

3.5 Assignment submission procedure

3.5.1 Turnitin submission

Your assignment report must be uploaded as a **unique document**. As long as the due date is still future, you can resubmit your work; the previous version of your assignment will be replaced by the new version.

Assignments must be submitted via the Turnitin submission box that is available on the course Moodle website. Turnitin reports on any similarities between their own cohort's assignments, and also with regard to other sources (such as the internet or all assignments submitted all around the world via Turnitin). More information is available at the UNSW Turnitin page. Please read this page, as we will assume that you are familiar with its content.

Please **also attach any programming code and/or sample spreadsheet output** used in your analysis as a separate file in the dedicated "code_sample" Moodle assignment box on the course webpage. These will be referred to by the marker only if needed, and in particular the **main assignment (with appendix) should be self contained**.

In the dedicated "selected_drivers" Moodle assignment box, please also attach a **csv** file containing the **list of 2,500 drivers which you recommend contacting**. Your csv file must contain 2,500 unique `DRIVER_ID` numbers identifying the drivers you recommend to contact. See the file `selected_drivers.csv` for a sample of how your recommendation file should look.

3.5.2 Late submission

Please note that it is School policy that late submission of assignments will incur in a penalty.

[2] Please kindly note that this is a maximum - you should feel free to use less pages if it is sufficient!

When an assessment item had to be submitted by a pre-specified submission date and time and was submitted late, the School of Risk and Actuarial Studies will apply the following policy. Late submission will incur a penalty of 5% per day or part thereof (including weekends) from the due date and time. An assessment will not be accepted after 5 days (120 hours) of the original deadline unless special consideration has been approved. An assignment is considered late if the requested format, such as hard copy or electronic copy, has not been submitted on time or where the 'wrong' assignment has been submitted. **Students who are late are still required to upload documents to the appropriate submission boxes. Additionally, they must also send their assignments to the LIC (Lecturer in Charge) via email.** The date and time of reception of the e-mail determines the submission time for the purposes of calculating the penalty.

You need to check your document once it is submitted (check it on-screen). **We will not mark assignments that cannot be read on screen.**

Students are reminded of the risk that technical issues may delay or even prevent their submission (such as internet connection and/or computer breakdowns). Students should then consider either submitting their assignment from the university computer rooms or **allow enough time (at least 24 hours is recommended) between their submission and the due time. No paper copy will be either accepted or graded.**

3.5.3 Plagiarism awareness

Students are reminded that the work they submit must be their own. While we have no problem with students working together on the assignment problems, the material students submit for assessment must be their own.

Students should make sure they understand what plagiarism is — cases of plagiarism have a very high probability of being discovered. For issues of collective work, having different persons marking the assignment does not decrease this probability.

3.5.4 Generative AI policy

You are allowed to use Generative AI to help you with editing, planning, idea generation, or coding. However, please include an Appendix in the report titled **"Generative AI usage"** explaining what you used AI for and, if applicable, outlining what prompts you used. If **you did not use Generative AI write in this Appendix that generative AI was not used**.

References

He, Haibo, and Edwardo A. Garcia. 2009. "Learning from imbalanced data." *IEEE Transactions on Knowledge and Data Engineering* 21 (9): 1263–84. https://doi.org/10.1109/TKDE.2008.239.