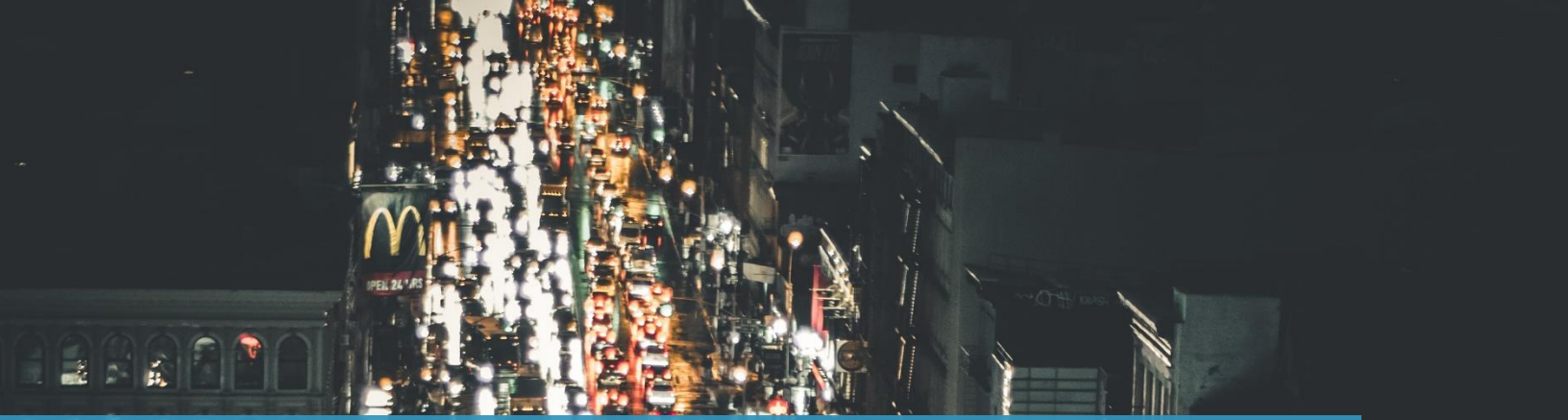


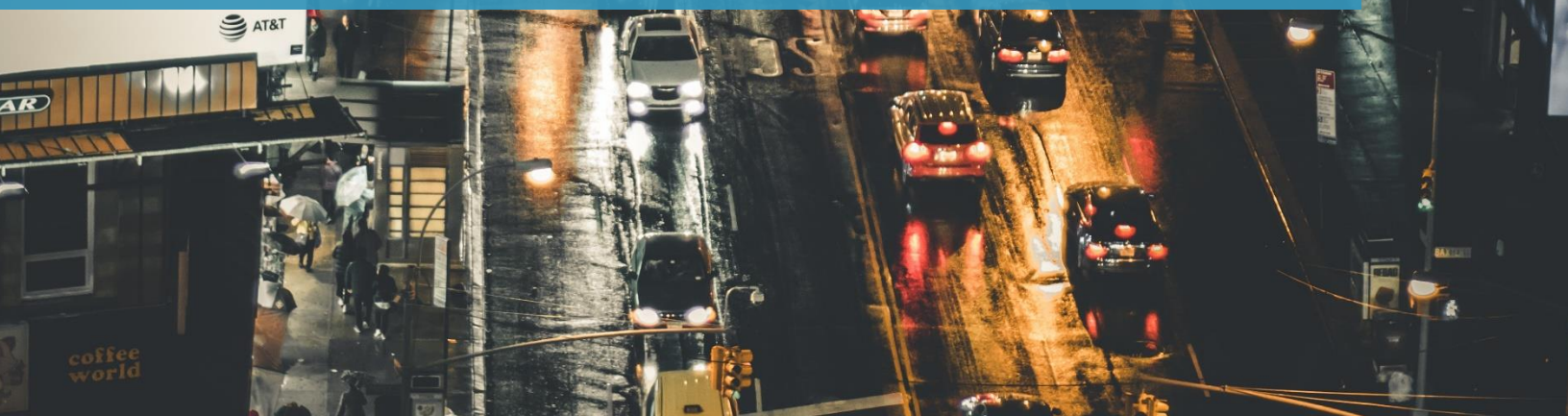
CONSUMER VEHICLE TOOLKIT

Helping you shop and drive smarter

Cover Note & Technical Specs

Michaël Bordeleau-Tassile

Montréal, Québec, Canada mbordeleau@gmail.com



Preface

ABOUT ME

I am an actuary (FCAS 2011) from Montréal, Québec, Canada. For more than a decade, I have been working in Home & Auto Insurance, using proprietary software for both data preparation (VBA/SAS) and modeling (Emblem). Last year, I embarked on a journey to learn the open-source R programming language and use machine learning algorithms.

HACKTUARY 2022

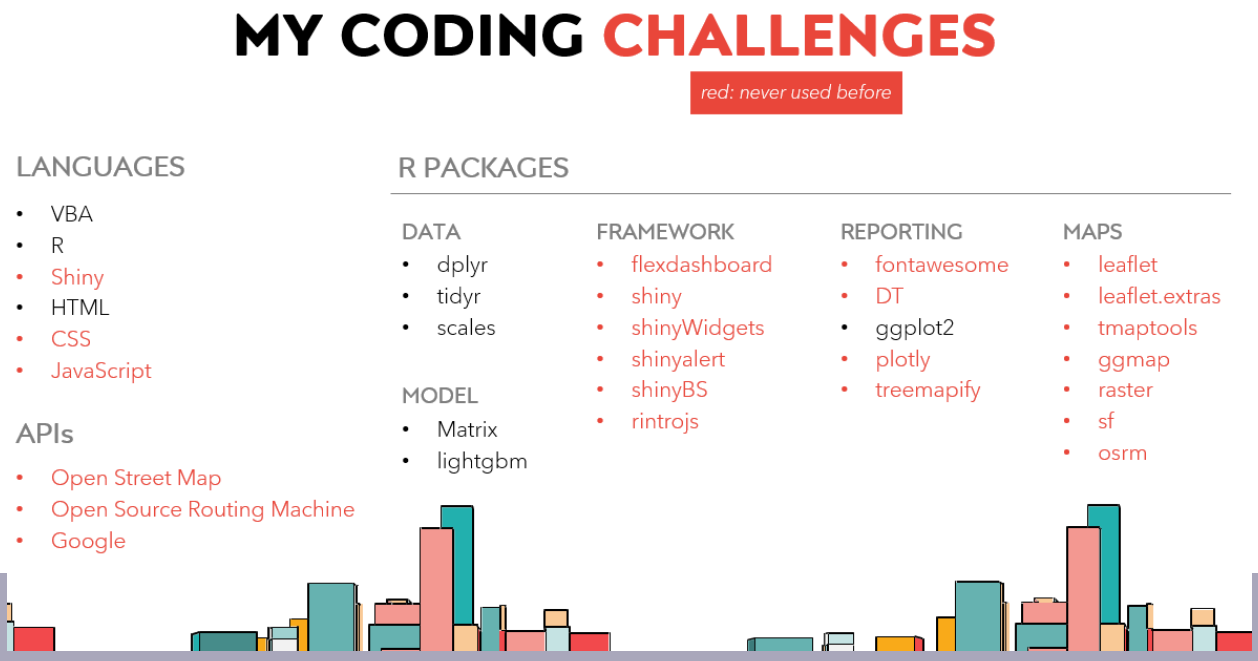
> *"a contest designed to showcase the actuarial skill set in developing novel risk engineering solutions, using publicly available data and featuring interactive elements"*

This challenge resonated with me, especially with my recent personal learning journey. It was the natural progression after learning data preparation and modeling in an open-source language. Developing a meaningful and appreciable interactive interface to deliver results, is no small feat.

In actuarial work, we commonly say 80% of the time is spent on data preparation and 20% on modeling. Having little to no experience on application development, I humbly underestimated the work needed to render applications with interactive features. I believe the former ratio could be modified to 40% data-prep, 10% modeling and 40% application development.

Moreover, the data preparation phase was additionally unusually long, as I had to extensively search the internet for valuable open data.

Finally, the challenge had me work through a lot of open R packages and APIs which I had no prior experience with.



Toolkit

OBJECTIVE

The central principle is to provide various interactive insights throughout the relationship cycle of a user and its vehicle.

First, when shopping for a vehicle, we provide tailored guidance on which make/model would be a better fit and give a relative indication of the insurance cost by vehicle.

Secondly, we provide insights on where to drive based on collision risk and where to park based on theft risk. As it covers Toronto and Montréal, that section is relevant for a total 10M population[1,2].

Thirdly, we provide numerous insights on various temporal and road contexts and how they influence accident propensity and severity.

As a reminder, for a better viewing experience, please use a screen resolution of 1920 x 1080, in a maximized browser window, with a zoom level no greater than 100%. Ideally, introduction page should show no scrollbar.

Over the next pages, we will discuss the data sources used, methodologies and provide examples of specific challenges we encountered during the development.

[1] https://en.wikipedia.org/wiki/Greater_Toronto_Area

[2] https://en.wikipedia.org/wiki/Greater_Montreal

Shop

OBJECTIVE

Help the user shop vehicles based on its profile and desired vehicle specifications. The application provides tailored suggestions, along with insurance price indications, environmental ratings, and body color popularity.

DATA SOURCES

1) Vehicles in Circulation: 2020 (Government of Québec fr-only)
2) Car Models by Manufacturer, Category, and Year (alternative) (Open-Source)
3) Fuel Consumption ratings (Government of Canada)
4) NHTSA Safercar (U.S. Department of Transportation)
5) How Cars Measure Up (Insurance Bureau of Canada)
6) How Age and Gender Affect Car Insurance Rates (Forbes)

METHODOLOGY

We started with the most important dataset (#1) which contains exhaustive details on the 6.8M registered vehicles in Quebec for 2020. Details include vehicle characteristics (make, model, year, weight, color, …) and owner characteristics (gender, age, city, …).

We used dataset #2 to get the body profile information (pickup, sedan, SUV, etc.).

From this we built and trained a multi-class gradient boosting machine using LightGBM.
Objective was to find the most suitable Make and Model for any given individual. The logic is that the mature consumer vehicle market can provide insight on what vehicle is more appropriate based on gender, age, surroundings, etc.

Ex: A 25-year-old male interested in a heavy sedan will have *Subaru WRX* as a suggestion.
All things being equal, a 70-year-old male will have *Subaru Legacy* as a suggestion.

A second multi-class model was built and trained on the Body Color.

Ex: A 25-year-old male will have *Black* as the color of choice.
All things being equal, a 70-year-old male will have *Gray* as the color of choice.

Dataset #3 is used to return (and sort by) the Fuel Consumption, CO2, and Smog Ratings.
Dataset #4 is used to return (and sort by) the Crash Safety Ratings.

Dataset #5 is an insurance relative index (with 100 being the average) of the claim cost per vehicle, per coverage.

We built the Theft risk indication by isolating and clustering the Comprehensive coverage.

Finally, we used dataset #6 to interactively display the importance of the Age & Gender on the insurance cost.

CHALLENGES

No Unique Key

The biggest challenge was the absence of a unique single key to appropriately merge information by make and model. We had to perform extensively tedious validations and make a couple dozens of adjustments to allow for a proper merge between all 5 vehicle data sources.

Color Coded Value

Dataset #5, “How Cars Measure Up” from the Insurance Bureau of Canada is in an Excel spreadsheet and some of its value is only color coded. Extracting and converting the color information into numerical values was not as trivial as expected. VBA macros had to be developed to perform the task.

Interpolation

Dataset #6, “How Age and Gender Affect Car Insurance Rates” from Forbes, has only a few data points. Excel is very powerful at charting data and creating smoothed curves. We were satisfied with the curve and wanted to retrieve all the values along the curve so that it can be displayed interactively. Unfortunately, retrieving values along a smoothed curve is not something that Excel offers. Again, VBA macros had to be developed to perform the task.

Map

OBJECTIVE

Help the user map its route and assess both itinerary and parking risk. The application calculates risk exposure using historical collisions and theft events.

DATA SOURCES

1) Toronto Boundary (Toronto Open Data)
2) Montréal Boundary (Government of Québec fr-only)
3) Traffic Collisions (Toronto Police Service)
4) Theft from Motor Vehicle (Toronto Police Service)
5) Auto Theft (Toronto Police Service)
6) Road Collisions (alternative) (Montreal Open Data)

METHODOLOGY

We used dataset #1 and #2 to facilitate the mapping of both cities' boundaries using powerful Leaflet (open-source library for web mapping).

Collisions datasets, #3 and #6, were pre-processed. All individual collision datapoints (730,000) were summarized into various heatmap rasters, varying by types of collisions (all vs only-injuries).

Mapping functionalities (search, drag & drop) were enhanced by using OSM geocoding engine. (OpenStreetMap). Itinerary routing is provided by OSRM (the Open-Source Routing Machine) and its open-limited-use server. We then calculate risk exposures by computing overlaps between both heatmap raster and itinerary route.

Thefts datasets, #4 and #5, were pre-processed, and all individual theft events (109,000) were summarized into heatmap rasters. We calculate a 500m radius circle around the itinerary endpoint (parking spot) and compute the overlap between theft heatmap rasters and that circle. We return a qualification of the parking risk based on the overlap computation.

CHALLENGES

Interactivity

Whenever there is interactivity, there is a heightened potential for unintended usage which can make the application crash. The map section is the most interactive segment, with search and drag & drop functionalities, multiple calls to third-party open-source services and plenty of under-the-hood risk computation.

Identifying bugs, developing routines to catch and manage errors was very time consuming.

Ultimate UX

We focused on developing a 100% open-source application. However, to provide the ultimate best user experience, we additionally integrated support for Google Places which allows for Auto-Completed search boxes, and more versatile search terms (Places Names, Addresses, Postal Codes, …).

To enable this feature, the application administrator should input the Google API key in the .Rmd file before server deployment. It is currently enabled on our online application.

Memory Management

While loading all heatmap layers at a high resolution on a regular desktop worked just perfect, the deployment server (shinyapps.io) has strict memory availability for its free tier.

Significant effort was put into optimizing memory usage and reducing map resolution so that the application could run smoothly without triggering a server shutdown. Heatmap resolution would be a nicer provided we deploy the application on a server with an annual fee of $500.

Risk Factors

OBJECTIVE

Help the user understand driving risk versus when and where it drives.

Answers to many questions:
Which month is riskier? Which day is riskier? Which hour of the day is riskier? What type of road? Etc.

DATA SOURCES

1) Accident Reports (Government of Québec fr-only)

METHODOLOGY

Dataset #1, details 1.4M unique accident reports filled out by police officers over the last 10 years.

Data was pre-processed with cleaning and grouping. 14 unique graphs are produced.

- 3 smoothed curves to illustrate the accident propensity and severity risk across different temporal contexts (seasons, days, hours).
- 11 treemap charts to illustrate the accident severity across various road contexts (surface, weather, configuration, etc.).

CHALLENGES

Graphs

Creating graphs from code lines is powerful for replicability.
However, it can be quite laborious to get from an idea to an appreciable result.