Dynamic Accident Risk Model for Citi Bike Trips in NYC

1 Goal

We build a **dynamic risk model** that outputs the probability (in %) that a Citi Bike trip results in an accident, conditional on the rider starting at a specific station and the current time. This estimate can be used to derive a **dynamic per-trip health insurance price**.

2 Notation

- S: set of all stations.
- $s \in S$: a station.
- $s_{\text{cur}} \in S$: current (start) station.
- $s_{\text{dest}} \in S$: destination station.
- r: a (random) trip.
- $r_{s_1 \to s_2}$: trip starting at s_1 and ending at s_2.
- t_r: time window associated with trip r (e.g. start time or start-to-end interval).
- We write $P(\cdot)$ for probabilities of events. For continuous variables, we write $f(\cdot)$ for probability density functions (pdfs).

3 Model decomposition

For a fixed start station s_{cur}:

$$P(\text{accident} \mid s_{\text{cur}}) = \sum_{s_{\text{dest}} \in S \setminus \{s_{\text{cur}}\}} P(\text{accident}, r_{s_{\text{cur}} \to s_{\text{dest}}} \mid s_{\text{cur}}) \tag{1}$$

$$= \sum_{s_{\text{dest}} \in S \setminus \{s_{\text{cur}}\}} P(r_{s_{\text{cur}} \to s_{\text{dest}}} \mid s_{\text{cur}}) \, P(\text{accident} \mid r_{s_{\text{cur}} \to s_{\text{dest}}}). \tag{2}$$

The second line follows from the **law of total probability** and the product rule.

This yields two subproblems:

1. Destination choice:
$$P(r_{s_{\text{cur}} \to s_{\text{dest}}} \mid s_{\text{cur}}).$$
2. Conditional trip risk:
$$P(\text{accident} \mid r_{s_{\text{cur}} \to s_{\text{dest}}}).$$

4 Destination choice

Reference. Notebook `03_citibike_FE_Modelling` (trained for a fixed s_{cur}).

$$P(r_{s_{\text{cur}} \to s_{\text{dest}}} \mid s_{\text{cur}}).$$

We model the destination distribution with a probabilistic classifier (multinomial logistic regression), producing predicted probabilities for each candidate destination.

Issue: class explosion. The number of possible destination stations is large (on the order of ~ 2000), which makes a direct multi-class formulation noisy and data-inefficient.

Mitigation. We restrict the target space to the k most frequent destination stations (calculated for each s_{cur} individually) and collapse all remaining stations into a single "Other" class:

$$\mathcal{C} = \{s^{(1)}, \dots, s^{(k)}, \text{Other}\}.$$

5 Conditional trip risk

We approximate the conditional accident probability as a ratio of expected accident count to expected rider count on the route during the relevant time window:

$$P(\text{accident} \mid r_{s_{\text{cur}} \to s_{\text{dest}}}) \approx \frac{\mathbb{E}[N_{\text{acc}}(r_{s_{\text{cur}} \to s_{\text{dest}}})]}{\mathbb{E}[N_{\text{riders}}(r_{s_{\text{cur}} \to s_{\text{dest}}})]}. \tag{3}$$

5.1 Expected accident count on a route

We decompose the expected accident count into (notation: $r = r_{s_{\text{cur}} \to s_{\text{dest}}}$):

$$\mathbb{E}[N_{\text{acc}}(r_{s_{\text{cur}} \to s_{\text{dest}}})] = N_{\text{acc,day}}(d(t_r)) \cdot \int_{\mathcal{R}(x_r, t_r)} f_{\text{acc}}(x, \tau)\, dx\, d\tau, \tag{4}$$

where $d(t_r)$ is the calendar day of the trip window, $x = (\text{lat}, \text{lng})$ is location, τ is time-of-day, and $\mathcal{R}(x_r, t_r)$ denotes the spatio-temporal "tube" of the trip.

5.1.1 Daily accident volume

Reference. `MV_Collision` notebook.

$$N_{\text{acc,day}}(d).$$

We predict the total number of accidents per day using regression models (Linear Regression, XGBoost, Random Forest) and a boosted hybrid approach (linear model + XGBoost on residuals).

5.1.2 Spatio-temporal accident intensity

$$f_{\text{acc}}(x, \tau).$$

We model accident intensity over location and time-of-day via a **mixture of Gaussians (MoG)** fitted to all accidents with injured cyclists aggregated by $(\text{lat}, \text{lng}, \tau)$.

Model assumption. The accident distribution in New York only depends on time of day.

5.2 Expected rider count on a route

Analogously:

$$\mathbb{E}[N_{\text{riders}}(r_{s_{\text{cur}}\to s_{\text{dest}}})] = \bar{N}_{\text{riders,NYC}}(t_r) \cdot \int_{\mathcal{R}(x_r)} f_{\text{riders}}(x)\, dx. \tag{5}$$

Model assumption. We assume that the traffic distribution is stationary for easier modelling.

5.2.1 Citywide rider volume

We approximate citywide rider volume by the number of active Citi Bike rides at time of the trip start, scaled by a constant factor (there are bicycle riders which aren't riding Citi Bikes). Since the average ride duration is short ($T \approx 12$ minutes), the scaled active rider count is a reasonable proxy:

$$\bar{N}_{\text{riders}}(t_r) \approx N_{\text{citibike\_riders}}(t_{\text{start}}) \cdot c_{\text{scaling}}.$$

We estimate the scaling constant c as

$$c = \frac{\text{Total bicycle rides in New York City in 2025}}{\text{Total Citi Bike trips in New York City in 2025}},$$

using Citi Bike data and NYC bicycle statistics (`https://www.nyc.gov/html/dot/html/bicyclists/bikestats.shtml`).

5.2.2 Spatio-temporal rider intensity

For each trip we map start and end coordinates on the map and fit a **mixture of Gaussians (MoG)** to obtain a normalized rider density:

$$f_{\text{riders}}(x).$$

5.3 Approximating Conditional trip risk

$$P(\text{accident} \mid r_{s_{\text{cur}}\to s_{\text{dest}}}) \approx \frac{\mathbb{E}[N_{\text{acc}}(r_{s_{\text{cur}}\to s_{\text{dest}}})]}{\mathbb{E}[N_{\text{riders}}(r_{s_{\text{cur}}\to s_{\text{dest}}})]}. \tag{6}$$

$$\approx \frac{N_{\text{acc,day}}(d(t_r)) \cdot \int_{\mathcal{R}(x_r,t_r)} f_{\text{acc}}(x,\tau)\, dx\, d\tau}{N_{\text{riders,NYC}}(t_r) \cdot \int_{\mathcal{R}(x_r)} f_{\text{riders}}(x)\, dx} \tag{7}$$

$$= \frac{N_{\text{acc,day}}(d(t_r))}{N_{\text{riders,NYC}}(t_r)} \cdot \frac{\int_{\mathcal{R}(x_r,t_r)} f_{\text{acc}}(x,\tau)\, dx\, d\tau}{\int_{\mathcal{R}(x_r)} f_{\text{riders}}(x)\, dx} \tag{8}$$

Assume that the accident density function will not be time dependent for this duration of the trip (stop integrating over it, multiply with the constant (trip duration) and a representative value).

$$\approx \frac{N_{\text{acc,day}}(d(t_r))}{N_{\text{riders,NYC}}(t_r)} \cdot \frac{\Delta\tau \int_{\mathcal{R}(x_r)} f_{\text{acc}}(x,\bar{\tau}_r)\, dx}{\int_{\mathcal{R}(x_r)} f_{\text{riders}}(x)\, dx} \tag{9}$$

Now split $\mathcal{R}(x_r)$ into n regions $\mathcal{R}_1, \ldots, \mathcal{R}_n$ of equal size.

$$\approx \frac{N_{\text{acc,day}}(d(t_r))}{N_{\text{riders,NYC}}(t_r)} \cdot \Delta\tau \cdot \frac{\sum_{i=1}^{n} \int_{\mathcal{R}_i} f_{\text{acc}}(x,\bar{\tau}_r)\, dx}{\sum_{i=1}^{n} \int_{\mathcal{R}_i} f_{\text{riders}}(x)\, dx} \tag{10}$$

Assume that f_{riders} and f_{acc} are constant on the regions.

$$\approx \frac{N_{\text{acc,day}}(d(t_r))}{N_{\text{riders,NYC}}(t_r)} \cdot \Delta\tau \cdot \frac{\sum_{i=1}^{n} \Delta\mathcal{R}_i \, f_{\text{acc}}(\bar{x}_i, \bar{\tau}_r)}{\sum_{i=1}^{n} \Delta\mathcal{R}_i \, f_{\text{riders}}(\bar{x}_i)} \tag{11}$$

By definition, all $\Delta\mathcal{R}_i$ are of equal size.

$$\approx \frac{N_{\text{acc,day}}(d(t_r))}{N_{\text{riders,NYC}}(t_r)} \cdot \Delta\tau \cdot \frac{\sum_{i=1}^{n} f_{\text{acc}}(\bar{x}_i, \bar{\tau}_r)}{\sum_{i=1}^{n} f_{\text{riders}}(\bar{x}_i)} \tag{12}$$

6 Final Model

The final dynamic risk estimate is:

$$\begin{aligned}
P(\text{accident} \mid s_{\text{cur}}) &= \sum_{s_{\text{dest}} \in S \setminus \{s_{\text{cur}}\}} P(r_{s_{\text{cur}} \to s_{\text{dest}}} \mid s_{\text{cur}}) \, P(\text{accident} \mid r_{s_{\text{cur}} \to s_{\text{dest}}}) && (13)\\
&= \sum_{s_{\text{dest}} \in S \setminus \{s_{\text{cur}}\}} P(r_{s_{\text{cur}} \to s_{\text{dest}}} \mid s_{\text{cur}}) \cdot \frac{N_{\text{acc,day}}(d(t_r))}{N_{\text{riders,NYC}}(t_r)} \cdot \Delta\tau \cdot \frac{\sum_{i=1}^{n} f_{\text{acc}}(\bar{x}_i, \bar{\tau}_r)}{\sum_{i=1}^{n} f_{\text{riders}}(\bar{x}_i)} && (14)\\
&= \Delta\tau \cdot \frac{N_{\text{acc,day}}(d(t_r))}{N_{\text{citibike_riders}}(t_{\text{start}}) \cdot c_{\text{scaling}}} \cdot \sum_{s_{\text{dest}} \in S \setminus \{s_{\text{cur}}\}} P(r_{s_{\text{cur}} \to s_{\text{dest}}} \mid s_{\text{cur}}) \cdot \frac{\sum_{i=1}^{n} f_{\text{acc}}(\bar{x}_i, \bar{\tau}_r)}{\sum_{i=1}^{n} f_{\text{riders}}(\bar{x}_i)}. && (15)
\end{aligned}$$

Now we have everything needed to compute this:

- $\Delta\tau$: average trip duration (02_Citi_Bike_EDA)
- $N_{\text{acc,day}}(d(t_r))$: predicted daily number of accidents from a linear regression model (MV_Collision)
- $N_{\text{citibike\_riders}}(t_{\text{start}})$: current number of active rides (citi bike live data feed)
- $P(r_{s_{\text{cur}} \to s_{\text{dest}}} \mid s_{\text{cur}})$: probability that a new trip ends at destination s_{dest}, modelled by logistic regression (`03_Citibike_FE_Modelling`)
- $\bar{x}_i$: take them equally spaced from a straight line between s_{cur} and s_{dest}
- $\bar{\tau}_r$: calculate trip start + 1/2 average trip duration
- f_{acc}: MoG on bike accident data with features lat,long, time of day (to do)
- f_{riders}: MoG on citibike trips, from each data point use start and end coordinates (to do)

7 Pricing Health Insurance

We can directly calculate the insurer's expected cost for a ride starting at s_{cur}:

$$\mathbb{E}[\text{Cost} \mid s_{\text{cur}}] = \mathbb{P}(\text{acc} \mid s_{\text{cur}}) \, \mathbb{E}[\text{Insurance Payout} \mid \text{acc}] \,. \tag{16}$$

Adding a premium and we already have the dynamic price for the customer:

$$\text{Price} = (1+\lambda) \, \mathbb{E}[\text{Cost} \mid s_{\text{cur}}] \,. \tag{17}$$