**1. Overview**

Road safety is a function of three components: driver, vehicle and road, on top of which the context of the trip plays an important role.

**2. Timeline**

- Defined our long-term vision
- Explored and visualized GM data
- Create pipeline of first project
- Learn from first iteration and expand analysis
- Data processing second project
- Build entire second project pipeline
- Presentations in Austin and Detroit

**3. First product goal**

To explain crashes, our first goal is to understand how a given vehicle or behavior feature impacts safety. Therefore, we built a reproducible framework that quantifies the real effect of any feature on safety, which relies on three main data sources:

<table>
<thead>
<tr>
<th>Data type</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle data</td>
<td>5.2M</td>
</tr>
<tr>
<td>Trip characteristics data</td>
<td>8.3M</td>
</tr>
<tr>
<td>Severe crashes data</td>
<td>37K</td>
</tr>
</tbody>
</table>

We have identified and rigorously removed three biases: the vehicle characteristics, the driving time and the driving behavior.

**4. Driver score method comparison**

To tackle the driving behavior bias, we built a driver score, derived from predicted crash likelihood. Below are the testing AUC for combinations of resampling and modeling techniques:

<table>
<thead>
<tr>
<th>Resampling techniques</th>
<th>Modeling</th>
<th>Over-sampling</th>
<th>SMOTE1</th>
<th>SMOTEEN2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.727</td>
<td>0.780</td>
<td>0.722</td>
<td></td>
</tr>
<tr>
<td>FFNN (Neural Network)</td>
<td>0.752</td>
<td>0.722</td>
<td>0.723</td>
<td></td>
</tr>
<tr>
<td>CART</td>
<td>0.699</td>
<td>0.689</td>
<td>0.667</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.782</td>
<td>0.715</td>
<td>0.788</td>
<td></td>
</tr>
</tbody>
</table>

1 = Synthetic minority over-sampling technique
2 = Combine over- and under-sampling using SMOTE and edited nearest neighbors

**5. Bias removal techniques**

In order to analyze the unbiased effect of a given feature on safety, we create groups of similar drivers and compare treated and untreated drivers in each group, using the following techniques:

- **K-Means clustering** to gather similar drivers together,
- **Causal inference estimates** to simulate a randomized experiment,
- A custom mixed-integer optimization model:

\[
\min_{\text{driver assignment to groups}} \sum_{\text{groups}} \left( \text{proportion discrepancies between treated and untreated drivers} \right)
\]

**6. Results & impact**

We applied our framework to uncover the real effect of the forward collision alert safety option. Customers, GM and insurance companies, which can better adapt safety option pricing, can benefit from such findings.

**7. Second project goal**

To deepen our understanding of crashes, we built a road risk score. Since GM does not store road data, we web-scraped open data about the Detroit area:

<table>
<thead>
<tr>
<th>Data source</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Road features</td>
<td>20K road segments</td>
</tr>
<tr>
<td>Road geometry</td>
<td>260K road segments</td>
</tr>
<tr>
<td>Intersections</td>
<td>18K intersections</td>
</tr>
</tbody>
</table>

We split roads into segments, extracted features about them and matched crashes to those segments.

**8. Crash locations matching**

- Step 1: Raw features only
- Step 2: Feature engineering (binning) and spatial data imputation
- Step 3: Add interactions between features
- Step 4: Process intersection data
- Step 5: Feature engineering: road angle
- Step 6: Process GM telemetry data
- Step 7: Resample the training data

**9. Road risk score impact**

Our road risk score can be impactful in many ways:

- Enhance safety awareness and strengthen GM’s position as a safety leader using interpretable coefficients
- Help city-planners decide which road to renovate and how
- Partner with insurance companies to reduce number of crashes by incentivizing their customers to take safer routes

**10. Predictive power comparison & evolution**

We investigated the driver, vehicle and road’s impact on safety and potential use cases.

For the first project, we recommend to apply our framework to find the real effect of weather and the effect of loud radio volume on safety.

GM should also start engaging insurance companies with both projects to reduce crashes and monetize its data. In particular, our first product can help insurance companies adapt safety option pricing, while our road risk score product can be turned into a white-label app and used to incentivize customers to take safer routes.

To achieve the zero crashes goal, GM should build a personalized road risk score that would lead to the safest routing algorithm. first, find features that are relevant to safety using our first product’s framework, then incorporate them into our second product’s risk score model that would be unique to a driver in a given vehicle.

**Conclusion and recommended next steps**

The plot of Detroit’s road risk score shows consistency from a segment to the next one. Moreover, it is often possible to find a safer route without making an important detour.