Nationwide: Telematics Assessment Exercises

Jason Barkeloo, PhD

Table of Contents

Code location for further fleshed out examples

- All code for these exercises can be found via these links as ipython/jupyter notebooks located on my github in addition to attachments sent with the presentation
    - Part 1: github: BarkelooNationwideAssessmentPart1.ipynp
    - Part 2: github: BarkelooNationwideAssessmentPart2.ipynp

Tasks to be Completed

Analysis Task:

- 1: Data Cleaning
- 2: Setting of hard braking and acceleration thresholds based on the data
- 3: Trip-by-trip Analysis and Summary

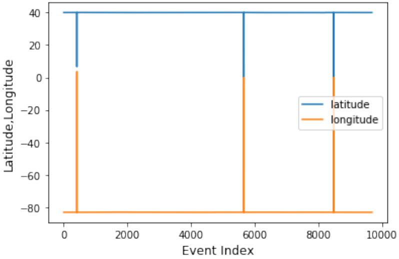
Data Set Overview:

- 9687 rows of 4 variables including:
  - trip_id: a trip number identifier
  - local_dtm: a datetime timestamp of the event entry
  - latitude: latitudinal coordinate
  - longitude: longitudinal coordinate

Datasets are loaded into pandas dataframes for further analysis

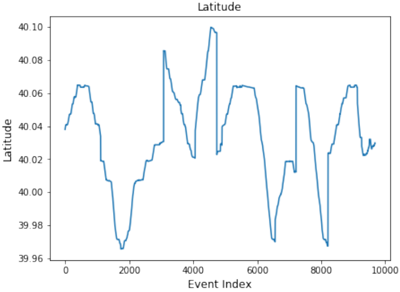
Data Cleaning, Gross Features

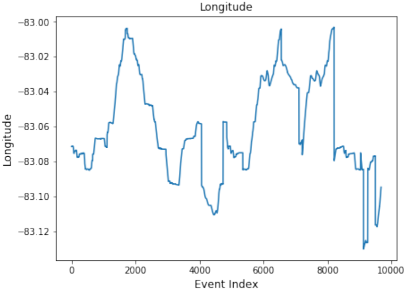
- 3 large unphysical features occur in the dataset (teleportation across the globe for 2-4 seconds)
- These events are pruned by requiring the latitude and longitude be within 2° of the median for the data set
- This includes an area on the order of the state of Ohio
  - Assumption: the sensors are used for checking daily driving habits and not long, rare, road trips
  - No other points are removed under this cut, only these large outliers. If this assumption is false (i.e., long-haul truck drivers use these), then this would need to be adapted



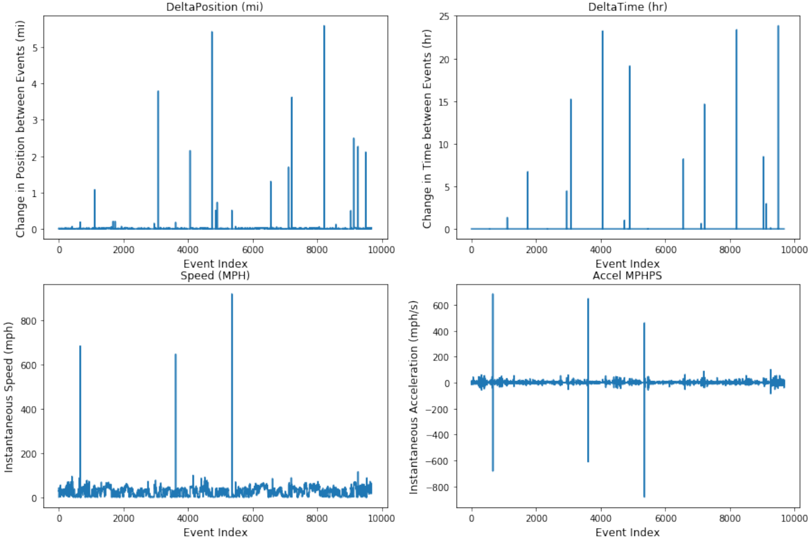
Result of Gross Cleaning

- The median cut above leaves the longitude and latitude plots in a reasonable state
- Some very fast jumps are still seen which are coincident, typically, with a change in trip_id (GPS drift while off)
- Can calculate distance between any two points using the geodesic distance making use of geopy package
- From this data and corresponding timestamps in local_dtm plots of the speed $s = \frac{\Delta\text{Position}}{\Delta\text{Time}}$ and acceleration $a = \frac{\Delta\text{Speed}}{\Delta\text{Time}}$ can be made





Further Cleaning - ΔPosition, ΔTime, Speed, Acceleration

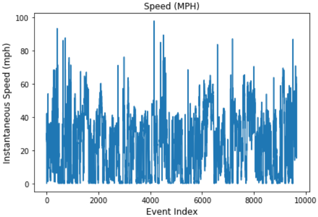


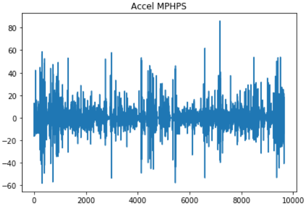
More Features to be Cleaned

- From ΔPosition, ΔTime we see the large number of drifts which account for the GPS drift from trip differences
- 15 events: These jumps will not be an issue when analyzing trip-by-trip as the change in position starts from the first point of the trip
- Speed and Acceleration plots show an additional 3 further unphysical events. These are resultant from small GPS errors for a few seconds and need to be cleaned
- Another issue comes when ΔTime between events is 0 i.e., if the frequency drops below 1Hz and the readings are taken within a second.
  - 24 events: A 0th order approach is taken to these points and only the first point is kept. An alternative would be averaging the latitude/longitude for the points. This would be a change within the same second and as such would not have much of an effect that is not then averaged out in the acceleration

Gross Feature Cleaning - Speed and Acceleration Plots

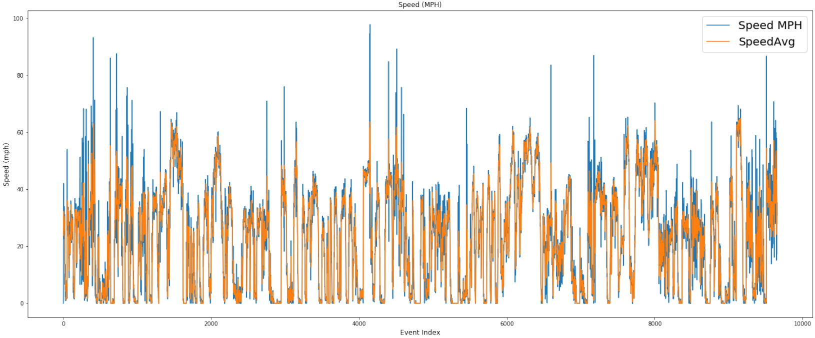
- The clear erroneous events in the speed and acceleration curves are cleaned by looking at large speed values ($>$ 100mph) using coincidence points that also correspond to accelerations that are not possible by the majority of cars ($>$ 30mph/s)
  - After these cleaning steps have occured most of the obvious points have been removed
  - Remaining oscillations are closer to the scale of the data
  - To smooth out itinerant spikes and general noise, a rolling average using a 3 event window is used on speed and acceleration





Speed, After Cleaning

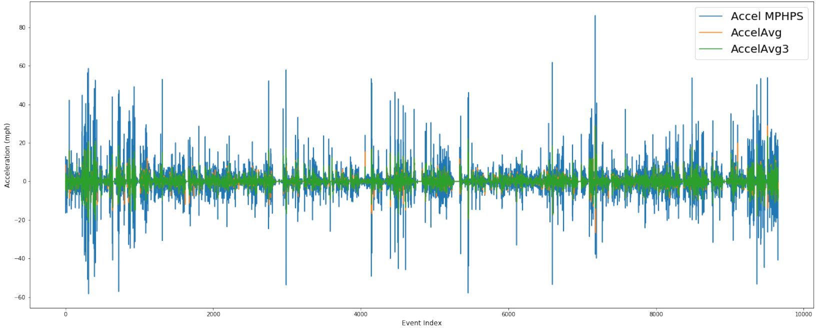
- Window size 3 average helps filter noise, still keeps large fast features



Acceleration, After Cleaning

- Accel: Directly calculated from change in speed values
- AccelAvg: Calculated using the change in the rolling average of speed values
- AccelAvg3: Calculated using the rolling average of acceleration values

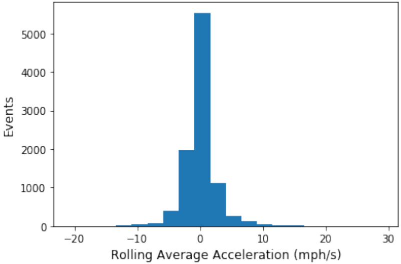
AccelAvg3 is the least spiking and as such will be used as the acceleration value going forward for threshold setting



Task 2: Setting Hard Event Thresholds

Hard Braking/Acceleration Events

- Assume average acceleration is roughly normal (mean = 0.07, std= 2.71) to consider positive and negative accelerations half-normal distributions → $\sigma = \bar{a}\sqrt{\pi/2}$
  - Positive Acceleration- mean: 1.71 mph/s std: 2.14 mph/s
  - Negative Acceleration- mean: -1.62 mph/s std: -2.03 mph/s
- Thresholds set for both distributions at 2 standard deviations away from the mean
  - Hard Acceleration: >5.99 mph/s
  - Hard Braking: <-5.68 mph/s



Hard Event and Idle Time Definition

Hard Events

- Number of peaks beyond the threshold using the rolling average acceleration
- Using rolling average and looking for local peaks in the acceleration landscape limits multicounting of the same ‘Event’

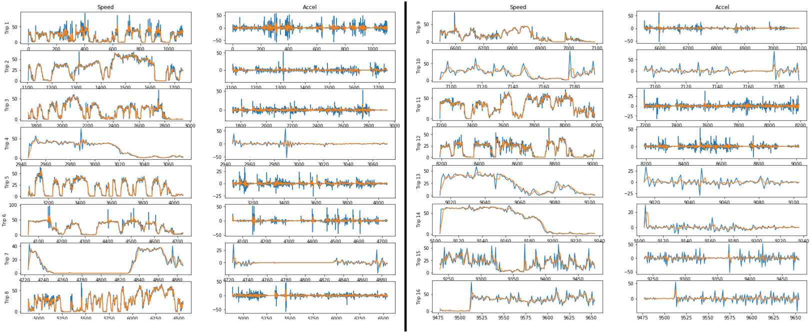
Idle Time Definition

- Total time spent with rolling average speed <1mph

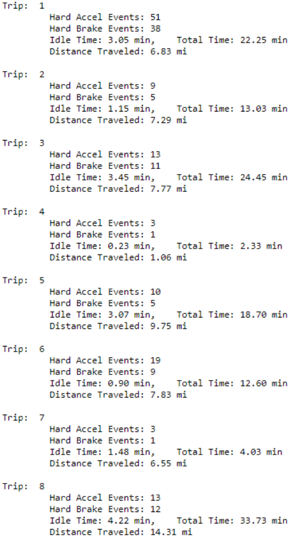
Task 3: Trip-by-Trip Summaries

Trip-by-Trip Speed and Acceleration Plots

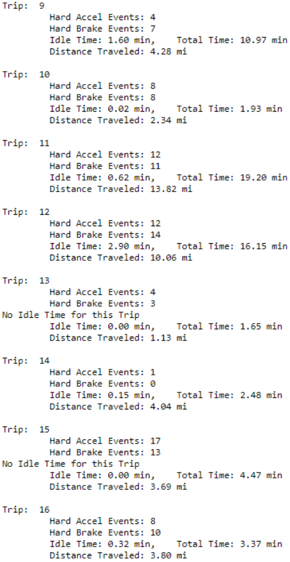
- Blue are raw values and orange are rolling averages



Trip Summaries



```
Trip:  1
         Hard Accel Events: 51
         Hard Brake Events: 38
         Idle Time: 3.05 min,    Total Time: 22.25 min
         Distance Traveled: 6.83 mi

Trip:  2
         Hard Accel Events: 9
         Hard Brake Events: 5
         Idle Time: 1.15 min,    Total Time: 13.03 min
         Distance Traveled: 7.29 mi

Trip:  3
         Hard Accel Events: 13
         Hard Brake Events: 11
         Idle Time: 3.45 min,    Total Time: 24.45 min
         Distance Traveled: 7.77 mi

Trip:  4
         Hard Accel Events: 3
         Hard Brake Events: 1
         Idle Time: 0.23 min,    Total Time: 2.33 min
         Distance Traveled: 1.06 mi

Trip:  5
         Hard Accel Events: 10
         Hard Brake Events: 5
         Idle Time: 3.07 min,    Total Time: 18.70 min
         Distance Traveled: 9.75 mi

Trip:  6
         Hard Accel Events: 19
         Hard Brake Events: 9
         Idle Time: 0.90 min,    Total Time: 12.60 min
         Distance Traveled: 7.83 mi

Trip:  7
         Hard Accel Events: 3
         Hard Brake Events: 1
         Idle Time: 1.48 min,    Total Time: 4.03 min
         Distance Traveled: 6.55 mi

Trip:  8
         Hard Accel Events: 13
         Hard Brake Events: 12
         Idle Time: 4.22 min,    Total Time: 33.73 min
         Distance Traveled: 14.31 mi
```



```
Trip:  9
         Hard Accel Events: 4
         Hard Brake Events: 7
         Idle Time: 1.60 min,    Total Time: 10.97 min
         Distance Traveled: 4.28 mi

Trip:  10
         Hard Accel Events: 8
         Hard Brake Events: 8
         Idle Time: 0.02 min,    Total Time: 1.93 min
         Distance Traveled: 2.34 mi

Trip:  11
         Hard Accel Events: 12
         Hard Brake Events: 11
         Idle Time: 0.62 min,    Total Time: 19.20 min
         Distance Traveled: 13.82 mi

Trip:  12
         Hard Accel Events: 12
         Hard Brake Events: 14
         Idle Time: 2.90 min,    Total Time: 16.15 min
         Distance Traveled: 10.06 mi

Trip:  13
         Hard Accel Events: 4
         Hard Brake Events: 3
No Idle Time for this Trip
         Idle Time: 0.00 min,    Total Time: 1.65 min
         Distance Traveled: 1.13 mi

Trip:  14
         Hard Accel Events: 1
         Hard Brake Events: 0
         Idle Time: 0.15 min,    Total Time: 2.48 min
         Distance Traveled: 4.04 mi

Trip:  15
         Hard Accel Events: 17
         Hard Brake Events: 13
No Idle Time for this Trip
         Idle Time: 0.00 min,    Total Time: 4.47 min
         Distance Traveled: 3.69 mi

Trip:  16
         Hard Accel Events: 8
         Hard Brake Events: 10
         Idle Time: 0.32 min,    Total Time: 3.37 min
         Distance Traveled: 3.80 mi
```

Part 2: Modeling - Simulated Dataset Overview

Summary of 30,000 vehicles 1Hz telematics datasets

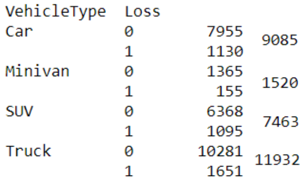
- Vehicle - Effectively an index on the data
- Days - Number of days data was collected (365 for all)
- Distance - Total number of miles vehicle was driven during data collection
- HardBrakes - Number of hard braking events detected
- HardAccelerations - Number of hard acceleration events detected
- NightTime_Pct - Percentage of total miles driven at night
- VehicleType - str description of type of vehicle
- Loss - Indicator if vehicle has been in a collision

Want to build a model that will optimize recognition of Loss events using these values

Task 4: Statistical Significance of Loss Between Vehicle Types

- Assuming the Loss populations are sampled from a binomial distribution with probablity LossPerType/TotalPerType, then a z-test can be conducted to determine if the null hypothesis (distributions are sampled from the same distribution) can be rejected
- For a significance $\alpha = 0.05$ a z-value greater than the critical value of $z_c = 1.64$ implies rejection of the null hypothesis
- For repeated tests the Look-Elsewhere effect/multi-comparison problem should also be taken into consideration, changing the critical value $z_c = 2.64$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}}$$



```
VehicleType  Loss
Car          0          7955
             1          1130   9085
Minivan      0          1365
             1           155   1520
SUV          0          6368
             1          1095   7463
Truck        0         10281
             1          1651  11932
```

Statistical Significance Between Vehicle Types

The conclusions to be drawn depend on how liberal the definition of statistical significance being used is

The use of $p<0.05$ is somewhat arbitrary but is what will be used here as it is a standard choice of convention

- z value for Car and Minivan: 2.48
- z value for Car and SUV: 4.19
- z value for Car and Truck: 2.96
- z value for Minivan and SUV: 4.59
- z value for Minivan and Truck: 3.92
- z value for SUV and Truck: 1.62

The null hypothesis cannot be rejected for the combination of Cars and Minivans and the combination of SUVs and Trucks

The implication then is that there are 2 distributions being sampled for these simulated events. This matches intuition as Trucks/SUVs exist in a cargo-loading domain while Cars/Minivans are what parents may gravitate toward

Task 5: Are Hard Brakes and Accelerations equally important in predicting risk?

Basic stats about the HardBrakes and HardAccelerations per Loss event and comparing to NoLoss events can give us insight on the separation power of these variables

Loss Events:

- HardBrakes - mean: 170.24, median: 98
- HardAccelerations - mean: 138.25, median: 68

NoLoss Events:

- HardBrakes - mean: 167.44, median: 98
- HardAccelerations - mean: 104.53, median: 56

Looking at the median values of these Loss events leads to the conclusion that Loss events have more HardAccelerations but a similar number of HardBrakes to NoLoss events

This matches my intuition that a larger number of HardAccelerations is an indicator of more aggressive driving

Model Building

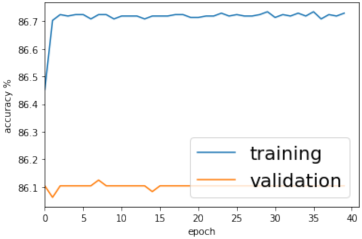
Primarily employing densely connected feed forward neural networks for event classification was chosen as it is the machine learning model I have most experience with for binary classification

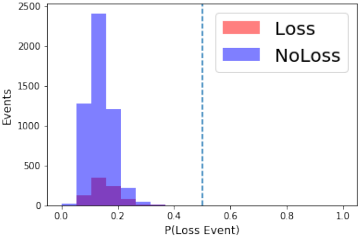
- 1 input layer with all potentially useful features (Distance, HardBrakes, HardAccelerations, NightTime_Pct, VehicleType)
- 2 hidden layers with 20 nodes each
- Each hidden layer has 20% dropout to avoid overfitting
- 1 output layer
- Activation Function: ReLU on input nodes and hidden layers with Sigmoid on the output layer
  - ReLU is efficient and reduces vanishing gradient problems as the gradient is linear throughout the positive regime
  - Sigmoid gives an output that we can interpret as a probability distribution
- Optimization Function: Adam (Adaptive moment estimation)
- Loss Function: Binary Cross-entropy

A training (64%)/testing(16%)/validation(16%) random set split was done to help ensure unbiased results

Naive Approach Neural Network

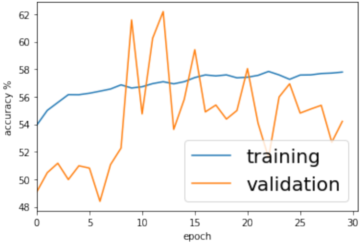
- Naively we could train a neural network on the data classes as given
- With enough separation power i.e., variables distinct enough in each class, this can be used for event classification
- This is not the case for this dataset, only a few variables inputs with a lot of distribution overlap
- This would then be expected to fail with a total accuracy that trends toward the class representation of the majority class, which is seen here

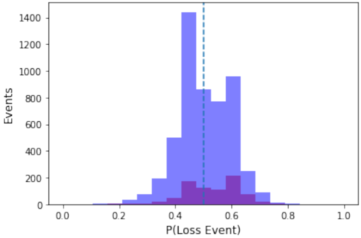




Neural Network with ADASYN Upsampling

- In order to ensure equal class representation the minority class is upscaled using synthetic data created using Adaptive Synthetic (ADASYN) over-sampling
- ADASYN generates synthetic data using weighted distributions for different minority class examples, generating more data on harder to learn events
- ADASYN shifts the decision boundary toward harder to learn events

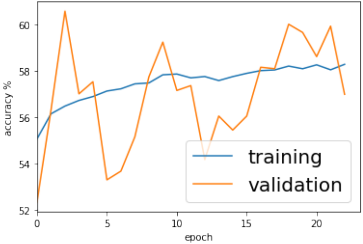


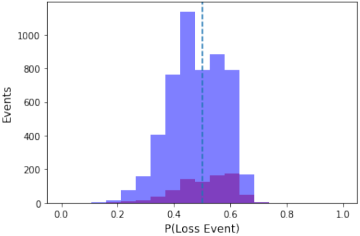


Loss Events P(Loss Event): mean: 0.535, std: 0.094
NoLoss Events P(LossEvent): mean: 0.503, std: 0.092
Loss Event Accuracy: 61.2%

Neural Network with SMOTE Upsampling

- Another network was created and trained using Synthetic Minority Oversampling Technique (SMOTE) over-sampling with similar results
- SMOTE generates synthetic data that is similar to, but not exactly like the minority class, using a nearest-neighbors approach and fills in space between neighbors





Loss Events P(Loss Event): mean: 0.510, std: 0.096
NoLoss Events P(LossEvent): mean: 0.480, std: 0.097
Loss Event Accuracy: 55.7%

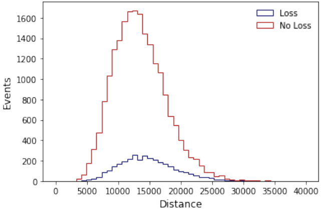
Model Comments

- Neural networks have been created and trained on a limited set of input variable with success in determination of Loss events
- The addition of further independent input variables would help the separation of the neural network greatly
- A bifurcation of the distributions is starting to occur with the ADASYN network, more input variables and events is likely to cause a major splitting of the distribution into likely Loss events and likely NoLoss events
- Boosted decision tree (BDT) models were also employed in the Jupyter notebook to slightly different ends

Additional Research Questions: 1

Can Multiple Loss Models be Useful Based on Driver Class i.e., Rural Vs. Urban Drivers?

- Rural and Urban drivers face different landscapes of challenges on their daily travels
- Requirement: GPS definition of urban environments
- Expect longer distance/trip for rural drivers while urban drivers have more stop-and-go traffic
- Larger Distances and a larger amount of HardAccelerations are both positvely correlated with loss this seems to be an interesting intersection of these correlations



Additional Research Questions: 2

Can Adherance to Speed Limit be Calculated and Used in Analysis of Loss?

- GPS Coordinates can be traced back to roads that have known speed limits (i.e., Apple Maps shows expected speed limits while navigating)
- An advanced analysis on a percentage of time above some threshold around the speed limit could point towards more high-risk behavior
- Including this additional variable in loss models could prove beneficial for risk modeling

Additional Research Questions: 3

Does Average Acceleration During Bearing Change Provide Benefit in Modeling Risk?

- Events with large accelerations have been used throughout my analysis within this assessment: HardBreaks and HardAccelerations
- Hard Turning i.e. a large acceleration as measured by an accelerometer when GPS bearing changes by $75 - 115^{\circ}$
- HardTurns could prove to be another advantageous metric for evaluating the driving habits of an individual and could be corrected in a similar way as HardBrakes/Accelerations

Additional Dataset Attributes

Additional attributes to add to the dataset for model improvement

- Driver Age - Age is historically one of the major factors impacting insurance rates (i.e., Age$\geq$ 25 leads to a lower rate) if this is true it should be beneficial as a discriminating variable
- Driver Sex - Another historical risk factor for risk that would be both interesting to analze and I have always been curious about
- Driving Record: Number of at-fault incidents/Violations - A driver with 0 to 1 at fault incident, especially with a long driving record (correlation to Age), is surely less likely to cause additional Loss Events
- Driver Home Location: Address/Zip Code - Leads towards a start of my urban/rural question on a smaller scale but will give first order estimations of the day-to-day driving experience. This also can lead towards accounting for weather induced Loss
- Driver Credit Score - As a stand-in for financial responsibility and fitness credit scores could have a small effect on the model as a surrogate for overall responsibility

Estimate Sample Size Needed for Additional Research

All of these estimates are done assuming access to the additional discriminating variables about the driver I have requested

- Question 1 (Rural/Urban Loss): Equal sized data sets (50k each, Rural/Urban Drivers) with the additional variables should be sufficient for results to measure the ability to calculate differential loss probability between the two classes
- Question 2 (Speed Limit Adherance): Further analysis could be done on every existing dataset to include checks to the posted speed limits and amount of time spend significantly above those speed limits. The nontrivial part is integration of GPS data to speed limit data which can be done after sensors have been collected

Estimate Sample Size Needed for Additional Research

- Question 3 (HardTurn Accelerations): A dataset of 100-200k trips would be a sufficient minimum to start setting a threshold for the acceleration occuring during turns. The GPS sensors would need to have an internal accelerometer to get more precise instantaneous information during the turn than could be reasonably expected from GPS coordinates alone. A dataset of this many trips would include perhaps millions turns even though a small fraction of those would be HardTurn accelerations

While more data is always better, datasets slightly larger or on the order of what was given for this example with additional discriminating variables assumed to be correlated with Loss prediction will allow a machine learning algorithm much more room to improve and isolate the phase space of high-risk drivers