June 16th, 2019

Safety: Driver Behaviour Analysis Using Telematics Data

Shubham Gupta

Executive Summary

End-to-end overview of analytical framework

Problem Statement

- Goal: Given the telematics data for each trip and the label if the trip is tagged as dangerous driving, derive a model that can detect dangerous driving trips
- Solution approach: Build model that accurately predicts the dangerous driving trip
- Produce empirical, model-derived insights into the factors that drive dangerous behaviour (Harsh Brake, Harsh Acceleration, Sharp Turns, Cornering

Data Preparation and Feature Engineering

- Extensive feature engineering to distill time-series data into a set of representative characteristics of the trip
- Features created at trip level (average speed, average acceleration, trip duration, count of stops, duration of stop)
- Detecting driving events (speeding stopping, sudden braking, fast u-turn) using sliding window of 10s with overlap of 5s
- Data cleaning included minor data imputation and outlier removal

Modeling

- Ensembles of two models, first model based on tree approach build on metadata, no. of driving events and second model based on sequence approach build on sequence/occurrence of danger driving event
- **LightGBM boosted tree model –** based on trip details and count of events
- **LSTM Sequence model** – based on sequence of events (accelerate, brake and then again harsh accelerate
- Training/validation splits to 80/20 ratio and performance tested on 5-CV folds

Performance and Insights

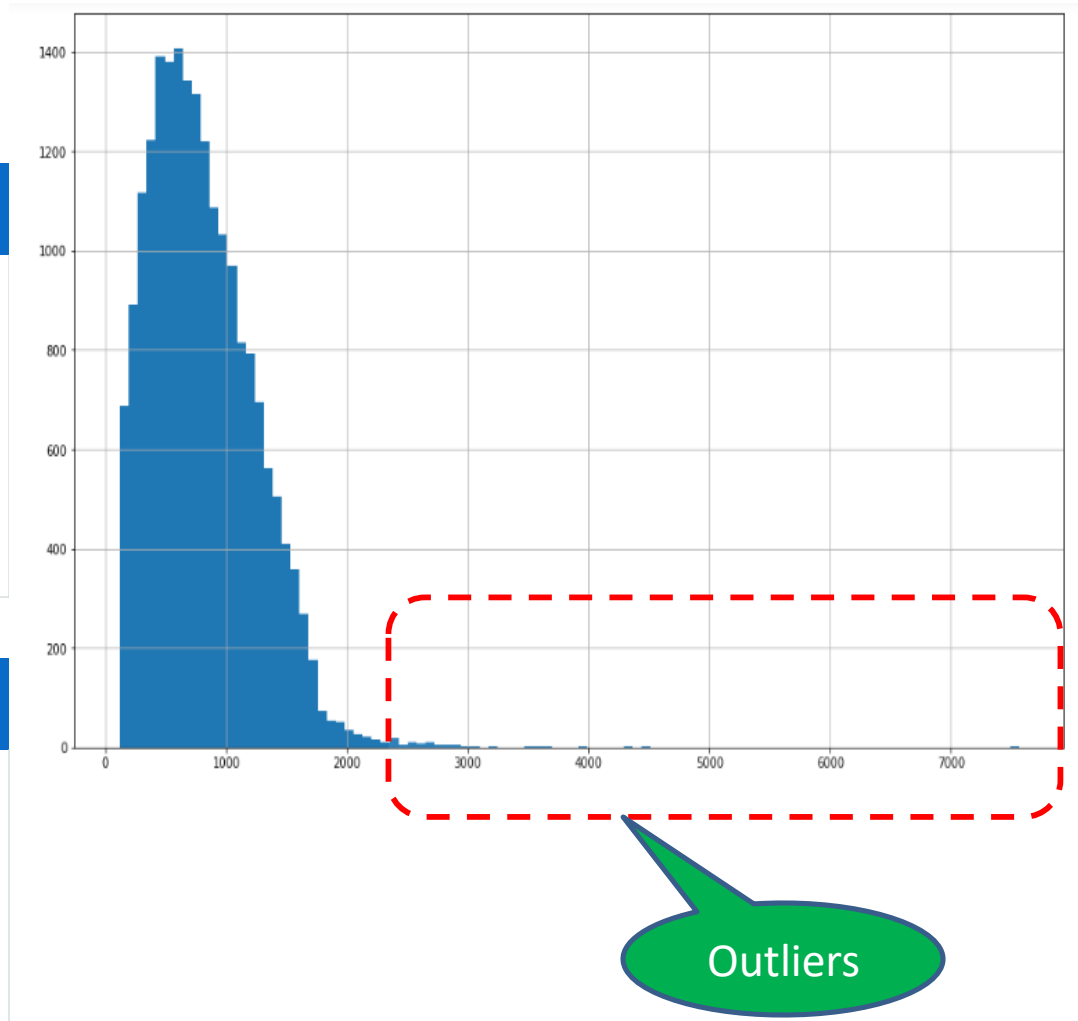
- Best AUC on holdout validation data (20%): **0.761**
- Model-driven insights derived using driving events detection and SHAP (Shapley Additive Explanations)
- Primary factor that influences the dangerous trip – Occurrence of events
- Additional key factors: Maximum Speed, Stop duration, fraction of harsh braking

Data Understanding

Telematics Data - four-wheel driver's smartphone

Snapshot of Data

- Total number of trip – 20K
- Thousands of telematics data points per trip have acceleration (3-axis), gyro (3-axis), Speed, Bearing and Accuracy information available at every 1s interval
- Target Label per trip (1-indicate dangerous driving) – Slightly unbalanced 25% positive points
- Trip duration varies from 2min to 2.5 hrs; on the average of 15 mins* (figure on right)



Cleaning of Data

- Removed outlier data points – Trip duration greater than 2.5 hrs
- Data Missing during the trip (removed if % of missing data is more than 90%)
- Received signal data is weak – indicated by speed value of -1 and higher accuracy value trip (removed if % of weak signal data is more than 90%)
- Applied special treatment of data missing and weak signal at feature engineering process

*After Data Cleaning; Driver forgot to end the trip

Feature Engineering

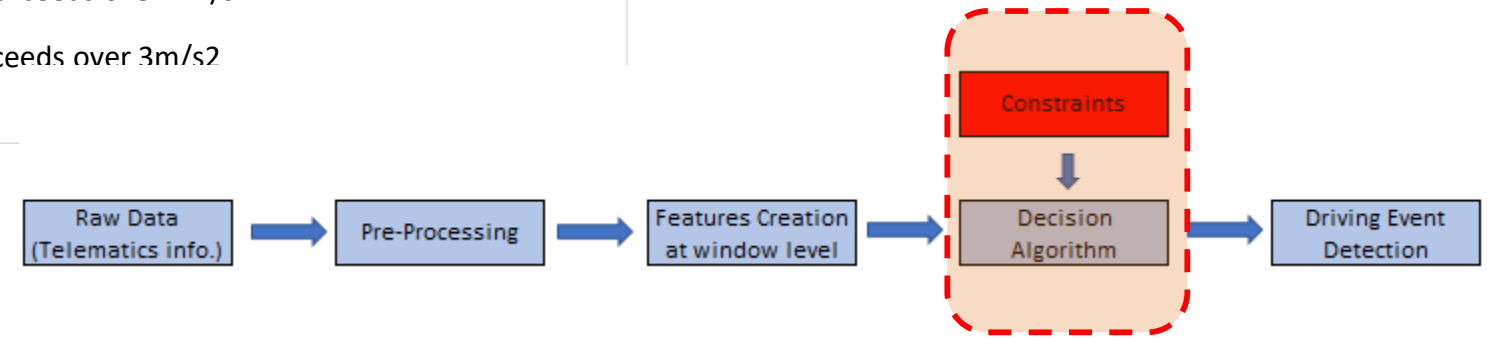
Characteristics of the Trip

Overall Trip Level Features

- Statistical Features (Min, Max, Std, Mean , Quantile) – Accuracy , Speed, Gyro, Acceleration, Bearing, Acceleration Calculated using speed
- Trip Duration, Missing data%, Average Speed, Number of stops, longest streak of stop
- Detection of Events – Turns, Braking, Accelerator, Long Stop Detection, Quick Acceleration, Speeding, Harsh Braking, Cornering
- Calibration of data – While removing impact of irrelevant events (missing data and weak signals) on interested events interested events (brake, stop , turn)
- Fraction of braking events exceeds over 2m/s2
- Fraction of acceleration exceeds over 3m/s2

Interaction between the Driving Events

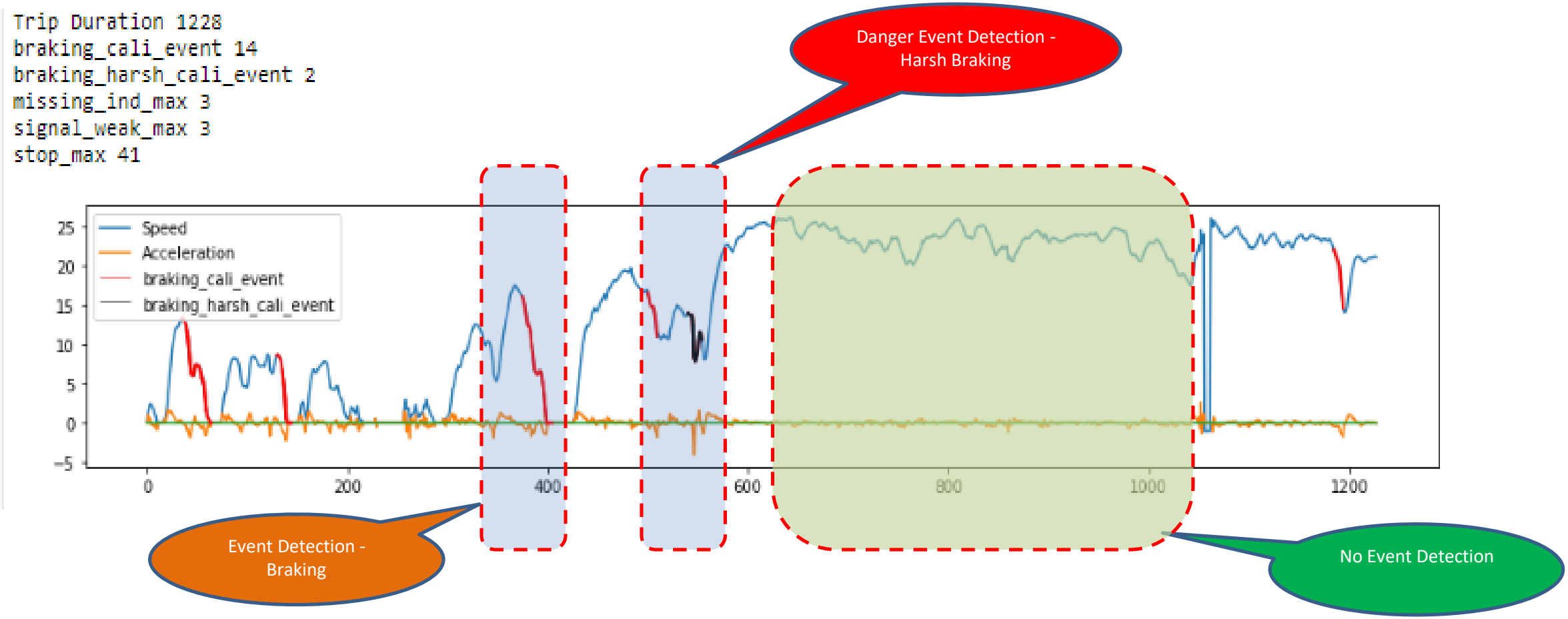
- Sometime only detection of events would not fully characteristics the trip. There are lot of cross events possibility and impossible to detect using Decision Algorithm. Hence using sequence model (LSTM model) to capture such scenario
- Combination of events - Turn with high speed (combination of turn and speed event)
- Sequence/occurrence of events add significant value of the driver behaviour
- Example: Turn, accelerator and then sudden break – Danger Behaviour
- Stop may be in traffic and then sudden acceleration – Danger Behaviour



For Trip Id 130

Feature Engineering – Driving Events Detection

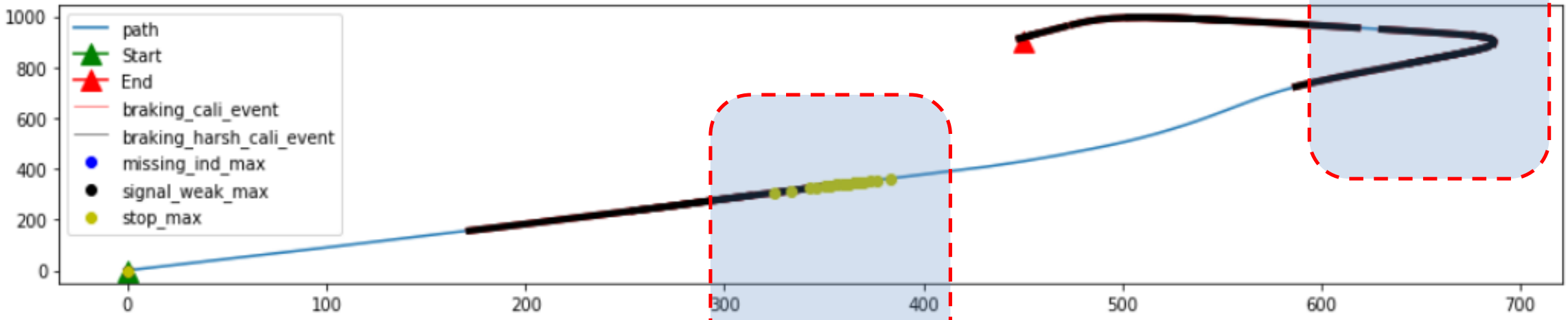
```
Trip Duration 1228
braking_cali_event 14
braking_harsh_cali_event 2
missing_ind_max 3
signal_weak_max 3
stop_max 41
```



For Trip Id 13 Refer (06_Visualization Script)

Feature Engineering – Interaction/Sequence of Events

```
Trip Duration 150
braking_cali_event 11
braking_harsh_cali_event 11
missing_ind_max 0
signal_weak_max 0
stop_max 9
```



Danger Event Detection – Harsh Braking and turning

Danger Event Detection – Harsh Braking and then significant stop

For Trip Id 130 (Refer 06_Visualization Script)

Modeling

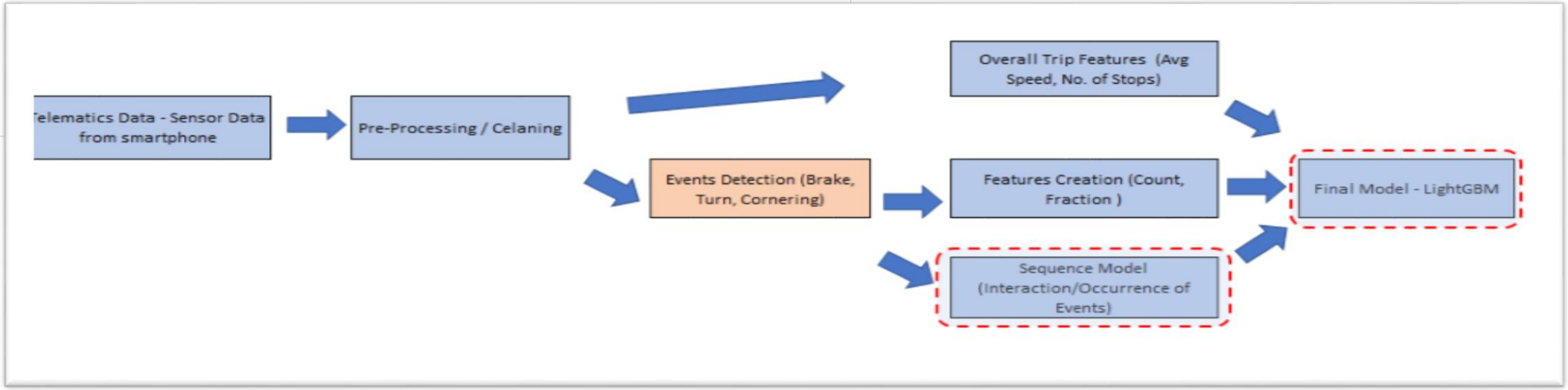
Modeling and validation techniques

Primary Modeling Technique: Gradient Boosted Trees (LightGBM)

- A state-of-the-art implementation of gradient-boosted decision trees with several attractive features:
- Training/Validation – 80/20 ratio at trip level
- Used 5-CV folds to insure model stability and performance
- Used Bayesian approach for Hyper parameters tuning
- Ensemble the sequence model score in LightGBM model

Sequence Model - LSTM

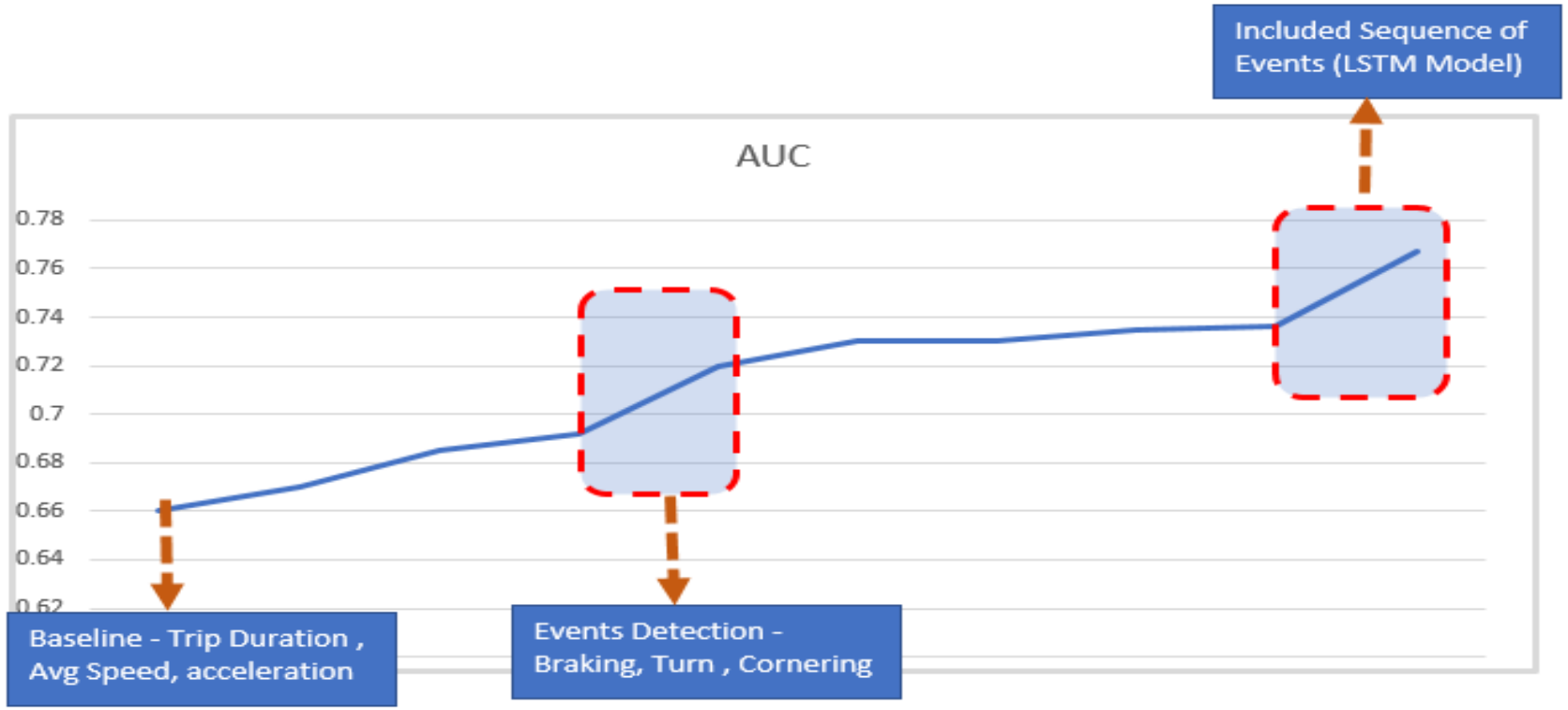
- Used LSTM model to captured interaction/sequence between events
- Considered 6 events - Braking, Acceleration, Accuracy, Stop, Missing and Signal weak events for sequence model
- Built deep model of 3 layers; with [8,5,1] neurons configuration



Performance

Prediction accuracy

- Best Model Accuracy → 0.761
- Total Features included in LightGBM model – 47
- Total Events Included in LSTM Model – 6 Features and maximum 200 timestamp data points
- Starting with the simple baseline model , using events detection features , sequence of events score and ending up with a complex ensemble model of LightGBM and LSTM

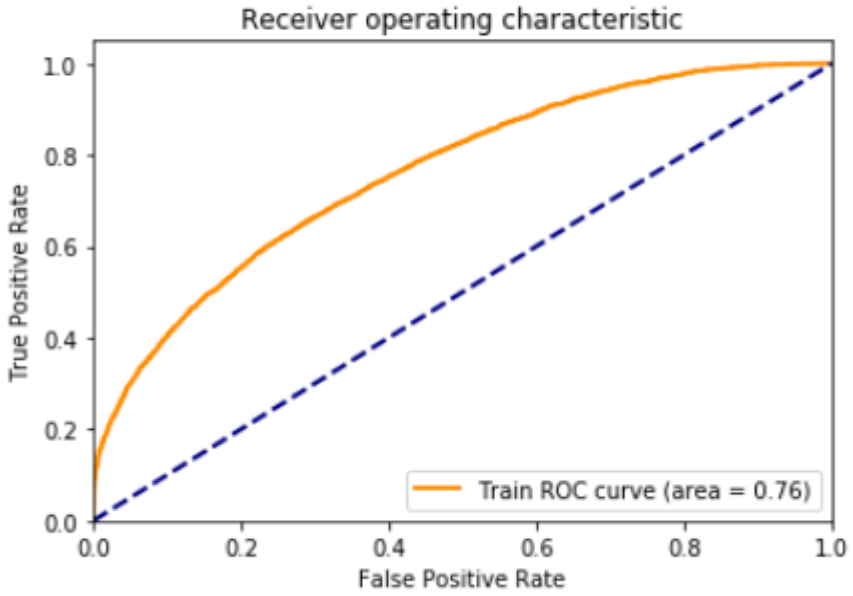


Performance

Prediction accuracy

- Best Model Accuracy → 0.761
- Total Features included in LightGBM model – 47
- Total Events Included in LSTM Model – 6 Features and maximum 200 timestamp data points

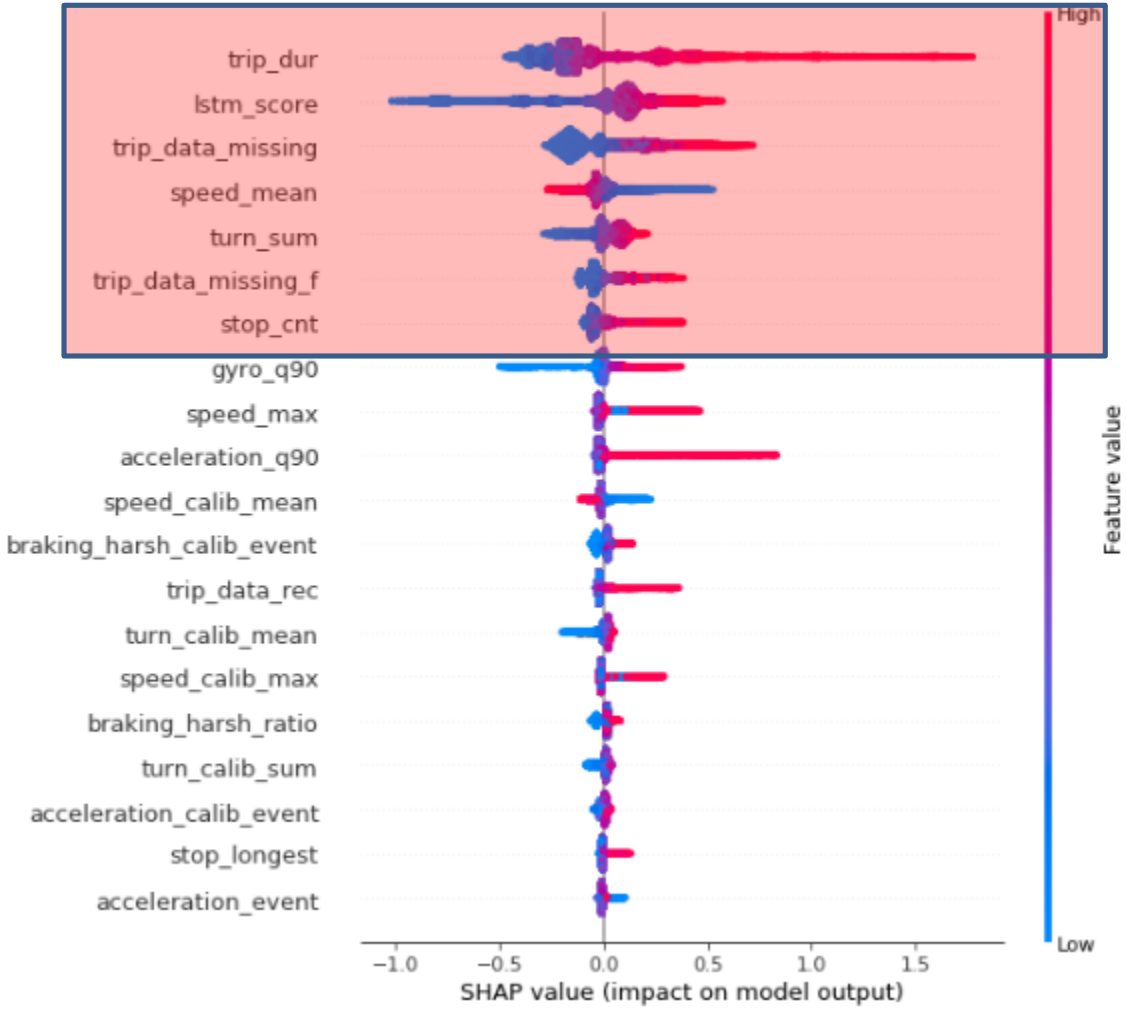
```
Train ROC -->  0.7611096488616138
```



Insights

SHAP overview, most important predictor

Primary model interpretation technique: SHAP (Shapley Additive Explanations)



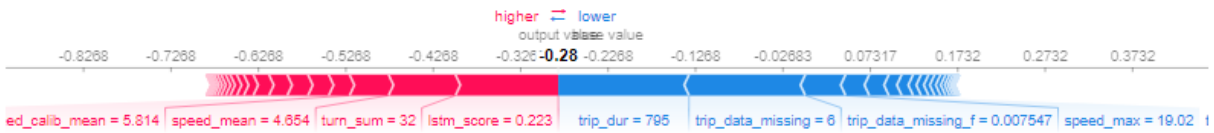
Top Factors that characteristics danger trip

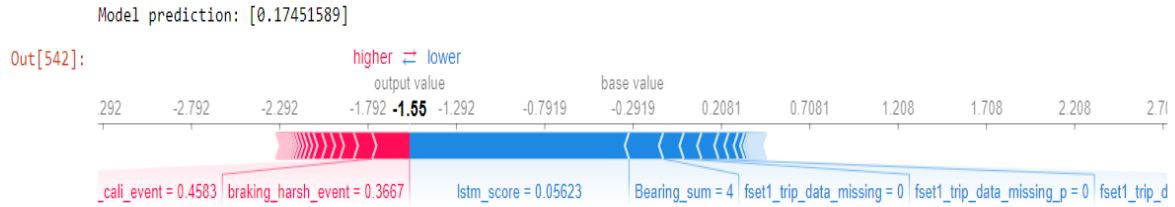
- trip-dur: higher the duration, more data points, more danger driving event and higher chance of dangerous driving
- lstm_score: lower the value indicate less chance of dangerous driving events
- trip_data_missing: Indicates some abnormal events, example accident
- turn_sum: lower the number of turns, lower the possibilities of dangerous events
- stop_cnt: higher the value; higher the chance of abnormal events

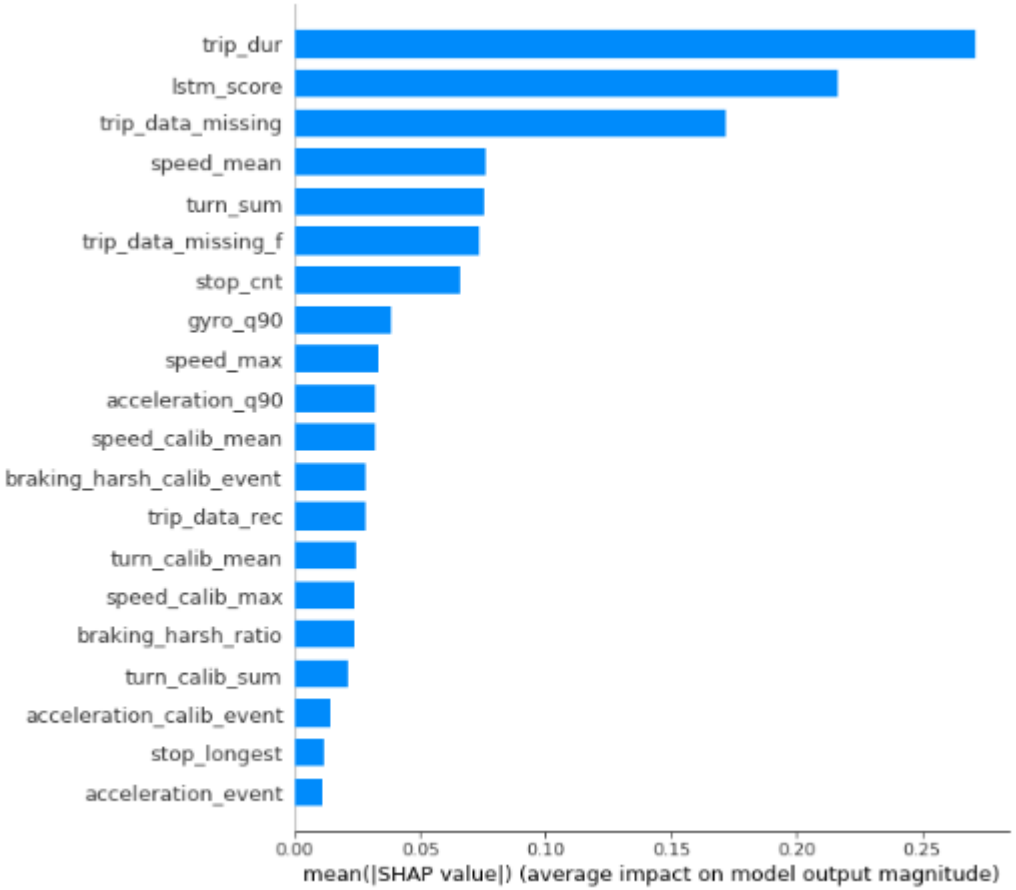
https://github.com/slundberg/shap

Insights

We can visualization Impact of factors at trip level and overall score of features







https://github.com/slundberg/shap

Conclusion

Future work

Calibration of Sensor Data

- Aligned acceleration and gyro direction with vehicle direction to accurate prediction of events (braking, turn)
- We can leverage acceleration, gyro 3-axis data if we aligned vehicle data with sensor direction

Additional Data

- Map data
  - Use for location identification which in turn helps to find the event of interest (Stop events)
  - Example : Detect whether stop events are normal event i.e. area where vehicle stop
  - Determine the quality of area (road quality; urban/rural area)
- Driver Information
  - If we get driver information also, we can build model at driver level and calculate driver driving score
  - Identify area where driver needs to improve

Alternative Modeling Approaches

- Use deep learning model – sequence model to capture better interaction/occurrence between the events (turn, acceleration)