|  |
| --- |
|  |
| Feature Extraction |

Revision Control

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Version No | Date | Change / Activity | By | Approver | Circulation |
| PA1 | 24 -Jan-18 | Initial Draft | Saurabh | Rajeev | Internal |
| A |  |  |  |  |  |

Reference Documents

|  |  |  |  |
| --- | --- | --- | --- |
| S No | Document Name | Version / Date | Owner |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Copyright: SENSORISE 2018

SENSORISE Confidential and Proprietary Information

All Rights Reserved

The contents of this document are restricted for circulation; please use pursuant to company instructions and the terms of the Non-Disclosure Agreement.

Table of Contents

[1 Introduction 3](#_Toc534036884)

[2 Feature Extraction 4](#_Toc534036885)

[2.1 High Level Description 5](#_Toc534036886)

[2.2 Segment Features 6  
2.3 IMEI Features 7](#_Toc534036887)

# Introduction

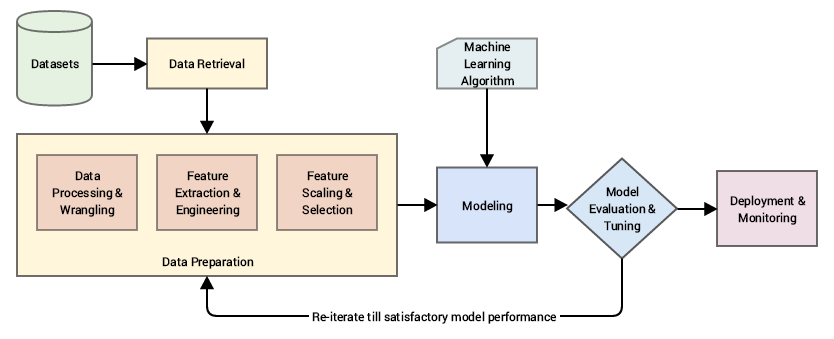
This document explains the features which we will be extracting for our driving behavior to de done correctly and aptly.

* 1. Why do we need Features?

In the context of machine learning, a feature can be described as a characteristic, or a set of characteristics, that explains the occurrence of a phenomenon. When these characteristics are converted into some measurable form, they are called features

* 1. What is Feature Extraction?

It is a set of designed features which would help us in extracting necessary actionable insights aiding us in drawing necessary inferences and model building for machine learning



The model shows us how feature extraction will help us in delivering a full scalable model of predicting driver behavior, with respect to model development this particular process plays a very important part.

1. Feature Extraction

2.1 High Level Description

* In driver analytics behavior we need actionable insights which could guide us in model building
* The insight from data will be enabled with features which would be particular set of characteristics of data like speed, distance, path deviation which would convey information for driving analytics behavior
* If for model building particular set of features are missing or needs to be extracted, feature extraction enables us to extract the essential parameters again and feed it to the model (loop back)
* Information to data is extracted from the IMEI segmented files which are generated after data cleaning of port files are done
* Since each cluster of IMEI is labelled particularly with data points from “Start” to “Stop” we need to extract features from these files to build model
* All important features are primarily extracted from few important fields of IMEI data:

1. Latitude and longitude for distance and speed calculation
2. Time stamp for calculating time difference and duration between each point
3. Geospatial information also plays an important role in driver behavior which can be extracted from coordinate information

2.2 Feature Table

1. **Segment Features**

|  |  |  |
| --- | --- | --- |
| Features | Parameters | Description |
| Distance | * Total distance * Distance in hour of day and day of week | In each segment how much a vehicle has travelled and how much has the vehicle moved in general in each stretch of an hour and particular day of week |
| Displacement | * Nature of path * Average Displacement | What has been the nature of path travelled by vehicle during the stretch i.e circular or rectilinear  Average displacement on particular path in each hour |
| Speed | * Max Speed * Average Speed * Speed deviation * Speed Peaks * Max minimum and mean of speed peak | Speed of vehicle in each stretch and average speed in each stretch in an hour  Deviation of speed and number of peaks in speed, will give information of driver behavior |
| Time | * Active hour * Total driving time * Maximum and minimum stretch of journey | For much total time has the vehicle driven  When does the vehicle travel most in hour of day and how much at a stretch does it drives? |
| Nature of travel | * Geographical location * Path travelled | First last and middle geographical coordinates in travel sprint or stretch |
| Acceleration | * Acceleration and Retardation  1. Maximum and Minimum acceleration 2. Standard deviation of acceleration 3. Mean of acceleration 4. Weight, maximum peak, mean and standard deviation of max peak | The value suggests us how much vehicle has actually speeded up or down during the course of its sprint getting key insights about driver behavior |

1. **IMEI and IMEI Trip Features**

|  |  |
| --- | --- |
| Features | Description |
| Are there any Day of Week or Time of Day Patterns | DOW patterns assist us in knowing, whether how much a vehicle has travelled on particular days. Is there any pattern on particular days of week, which would assist us in getting insight of driver behavior?  TOD pattern will help us in knowing during which Time of Day vehicle has moved frequently. It would show driver active hours and its efficiency time of driver, crucial parameter for model building. |
| Segments in data | Exactly how many stretches has the vehicle made during the course of its journey.  Root cause analysis could also be done by further securitizing the parameters and other factors |
| Frequency of travel | Exactly how many segments or how many sprints of travels has the vehicle made. |
| Time of travel | What has been the total time of travel in all stretches and what’s the average time of all stretches. |
| Path and travel geography | Has there been any concurrency in path? Has the same path been travelled over and over or has the path been accessed only once?  Where and which geo location has been a constant commute and what have been the stretches in same geographical location |
|  | |

1. Program flow for feature Extraction
   1. High Level Description

* Feature extraction needs to be done at segment level, as well as IMEI level
* Given, each segment of data we need to calculate different parameters for features to be extracted for analysis and getting insights on driver behavior
* Different parameters extracted at IMEI level, are related to geographical location, patterns in vehicle movement overall
* While at each segment these values are related to driver behavior i.e nature of driving
  1. Program Flow:
* With each segment of data:

1. Find distance travelled during each week of day and hour of day
2. Find total distance with respect to each point and displacement between start and stop point
3. Calculate time between and each successive points and speed at each segment between each point
4. Find maximum, mean and standard deviation of peaks in the speed columns of data
5. Calculate acceleration and deacceleration successively with respect to each point in continuous segment of data
6. Find various parameters related to acceleration and retardation like maximum, minimum, mean and standard deviation
7. Find distance of first, middle and last point in the segment with respect to reference point to figure out the path travel in geo coordinate system
8. **Check Points**
   1. High Level Description

* Since we have a geographical large area of many vehicles together there could be a possibility of vehicle moving in different distance intervals i.e short distances to large distances
* For finding route of a vehicle a proper path needs to be found out where vehicle has at least moved more than some distance(say 1 K.M)
* Based on this criterion we will select only those data points which have valid start or end points select start or end accordingly
  1. Input
* Input to the program is Data frame having latitude and longitude of start and end points
* Reference distance (dist\_ref) for selecting points based on our distance criterion
  1. Output

Output of the program is data frame column labelled which have label having 3 labels “S”, “E” or “SE”

* 1. Program Flow

1. From iteratively second point find distance between start point and end point of first data point
2. If the distance is more than reference distance specified:

* Label the second point as “S”
* If first point is marked as “S”:
* Label the point as “SE”
* Else:
* Label the first point as “E”

1. All those points which are labelled as “S” or “SE” or ”E” are having problem in

Start , end or both start and end points and shouldn’t be selected for our analysis

1. **Same Geo Points**
   1. High Level Description

* K-Means clustering for geographically clustering same data points having same start point is creating problem as boundary points are getting associated with different clusters on different iterations.
* Since starts points and end points are closely associated, we only select those data points which lies away from end points and cluster them together
* For this purpose, we use point distance approach to calculate distances between points to get closely associated points grouped together, so that clustering could be done properly
* Here a clustering method is developed where distances are calculated based on randomly selected data point and all those points which are nearby points are clustered together based on specified distance(cluster formed)
* The center of the cluster(centroid) is updated based on the grouped points which is nothing than mean of all the clustered points
* The points are again clustered based on maximum of specified distance or maximum of twice mean of inter-cluster distances
* The procedure of same point geographical grouping is repeated until there are no points left which are not in range of centroid
  1. Input
  + Input to the program is dataframe(X) having latitude and longitude
  + Reference distance(ref\_dist) which is user defined distance calculated for getting cluster points
  1. Output
* Output of the program is same data frame having a column of cluster associating cluster number of all those points
* Another data frame centroid which contains centroid of all the clusters formed
  1. Program Flow:

1. Find distance of first point with respect to all points in the geographical region spread over i.e use Euclidean distance after proper scaling
2. Find those points which are less than the reference distance 1 and group them together to find the Centroid 1 which is mean value of those coordinates
3. Now find mean of inter cluster distances i.e distance of remaining points from the mean
4. Now of all those remaining points find all the respective distances with centroid of cluster
5. Select only those points from the remaining points which have distances having the maximum value of twice of mean centroid or reference distance(specified) and call it reference distance 2
6. Iteratively repeat this process i,e steps (2-5) till the minimum distance found is not less than some maximum of reference distance 1 or reference distance 2
7. Repeat all these steps (1-6) until no point are remaining which is to be clustered and we have associated points grouped together based on distances