Road Roughness Condition Analysis with Spark

Giulio Leonardi

2024

1 Introduction

Dataset Presentation

This project is a data-driven analysis of the *Passive Vehicular Sensors* dataset, situated within the field of Vehicle Perception.

The dataset comprises measurements collected during multiple driving sessions using various sensors installed in the vehicle. Each driving session is represented as a long multivariate time series. A total of 9 driving sessions were conducted, distinguished by being performed by 3 different drivers across 3 different scenarios. Each session is stored in a separate file named PVS[n], where n is the session number.

The features of the time series are derived from data captured by the following sensors:

- **GPS**: A single unit mounted on the dashboard. It records location and speed, with a sampling rate of 1 Hz.
- **Accelerometers (Acc)**: two on the dashboard (Left and Right), two above the front suspensions, and two below the front suspensions. They measure acceleration (m/s^2) along three axes, with a sampling rate of 100 Hz.
- **Magnetometers (Mag)**: two on the dashboard (Left and Right), two above the front suspensions, and two below the front suspensions. They record the ambient geomagnetic field (μT) along three axes, with a sampling rate of 100 Hz.
- **Gyroscopes (Gyro)**: two on the dashboard (Left and Right), two above the front suspensions, and two below the front suspensions. They measure the rotation rate (deg/s) along three axes, with a sampling rate of 100 Hz.
- **Thermometers (Temp)**: two on the dashboard (Left and Right), two above the front suspensions, and two below the front suspensions. They measure temperature (°C), with a sampling rate of 100 Hz.

Additionally, each timestep in every time series is assigned several class labels: *Road Surface Type*, *Road Surface Condition*, *Road Roughness Condition*, and *Speed Bump Types*. In this study, the only label used *is Road Roughness Condition*, which will be referred to simply as "Road Condition" from this point onward. This label can take one of three values: *Bad*, *Regular*, or *Good*.

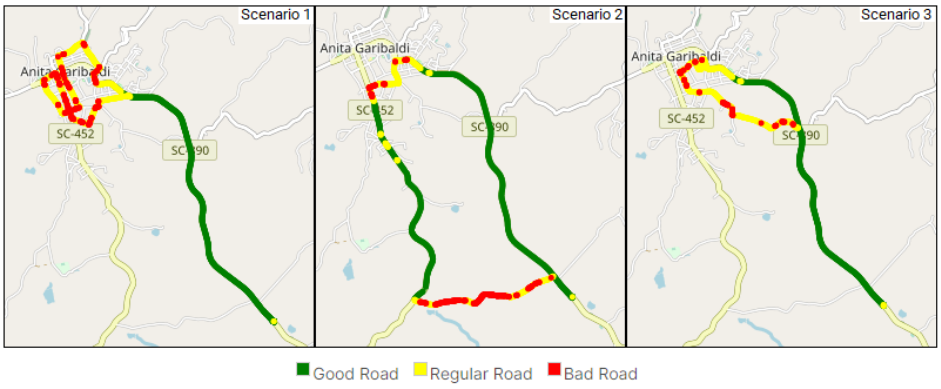


Figure 1: The three different scenarios with the distribution of the Road Roughness Condition labels

-	PVS1	PVS2	PVS3	PVS4	PVS5	PVS6	PVS7	PVS8	PVS9
Vehicle	Saveiro	Saveiro	Saveiro	Bravo	Bravo	Bravo	Palio	Palio	Palio
Driver	1	1	1	2	2	2	3	3	3
Scenario	1	2	3	1	2	3	1	2	3
Distance	13.81km	11.62km	10.72km	13.81km1	11.62km	10.72km	13.81km	11.62km	10.72km
N. Rows	144036	124684	105816	132492	133877	96279	128548	123618	91555

Table 1: Dataset Composition

In table 1, the composition of the dataset across its 9 subsets is shown. Overall, it contains a total of 1,080,905 records. Other dataset characteristics will be presented and analyzed in Section 2.

Research Questions

The following research questions guide this project:

1. Can meaningful features of the road be identified through an unsupervised analysis of the measurements?
2. Is it possible to develop a prediction model for road conditions that is independent of the driver, vehicle, and location?
3. Which types of features are most important among the following: Speed, Accelerometers, Magnetometers, Gyroscopes, and Thermometers?

2 Data Exploration

This section presents the feature analysis phase of the dataset, aiming to identify the most promising features for the analysis and to explore the relationships between them.

Note: To standardize the naming of the features, I will refer to them using the following format: [sensor abbreviation]_[spatial axis, if applicable]_[vehicle part]_[vehicle side]. For example, acc_x_dashboard_L represents the acceleration along the x-axis recorded by the sensor located on the left side of the vehicle's dashboard. This convention does not apply to features recorded by the GPS, which are simply named speed, longitude, and latitude.

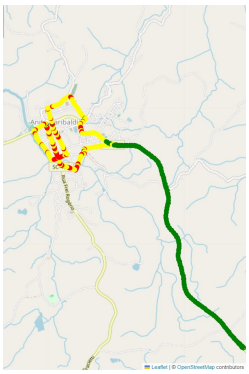
The dataset contains a total of 59 features. Analyzing every single feature would be both impractical and useless. Therefore, I decided to select a subset of features for analysis, choosing at least one feature from each sensor group. At this stage, I considered only the sensors positioned on the dashboard, as this intuitively minimizes measurement differences between the left and right sides (so I selected only the left side features).

The analyzed features were: speed, acc_x_dashboard_L, acc_y_dashboard_L, acc_z_dashboard_L, gyro_x_dashboard_L, gyro_y_dashboard_L, temp_dashboard_L, mag_x_dashboard_L, mag_y_dashboard_L, mag_z_dashboard_L.

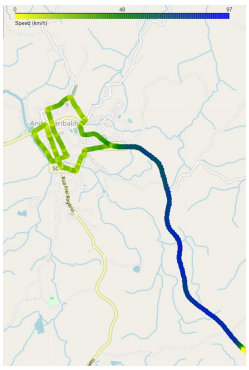
To analyze and create plots of the features, an approximation was applied to the time series using the PAA (Piecewise Aggregate Approximation) method. In this method, a time series is divided into n segments of equal size, and the mean of each segment is calculated. This mean value represents the corresponding group of observations in the approximated time series. In this case, each time series was divided into segments of 200 observations (a 200:1 approximation ratio). For each feature and each group of observations, in addition to the mean, the standard deviation was also calculated, as it was considered important for the analysis.

Exploration 1: Feature Distribution on Scenario

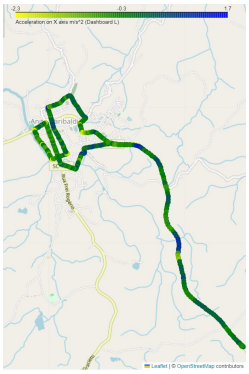
The first part of data understanding involved analyzing the distribution of the features in Scenario 1, this process provides three different perspectives: it allows for analyzing the distribution of each feature along the road, helping to visually identify patterns in the individual features; it enables comparison with the Road Condition Label, to assess whether the feature might be important for the classification task; and it allows for comparing features with each other, making it possible to notice any correlations.



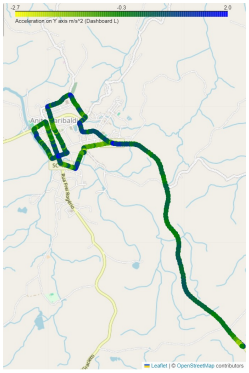
(a) Road Condition



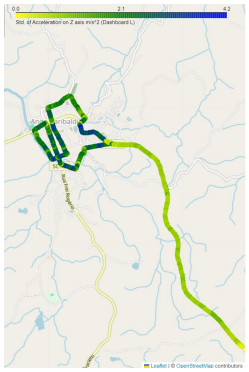
(b) Speed



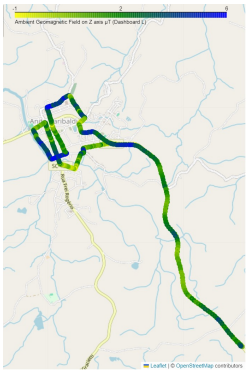
(c) acc_x_dashboard_L



(d) acc_y_dashboard_L



(e) std_acc_z



(f) mag_z_dashboard_L

Figure 2: The distribution of the features and the label on the Scenario 1, for the Driver 1.

As we can see in figure 2, among the analyzed features, the most promising ones for classification are speed and std_acc_z, which is not surprising: roads with worse conditions appear to be those where the vehicle tends to go slower; on the other hand, the standard deviation of the acceleration along the z-axis appears bigger while driving on a bad surface, probably due to vehicle instability. The other three features shown, although are visually less significant, may hide more complex patterns or could potentially be more useful for other tasks rather than Road Condition Classification.

Additionally, by analyzing the graphs of other features not shown in figure 2, I noticed that some features seem to be highly correlated with each other. This topic will be explored further in the feature selection phase presented in Section 3.

Exploration 2: Feature Distribution by Driver

In this part of the analysis, for the features speed, std_acc_z_dashboard_L, acc_x_dashboard_L, and acc_y_dashboard_L, the time series of the three different drivers in Scenario 1 will be plotted and compared. This allows us to observe patterns in the time series that might be related to road characteristics, as well as to examine the differences in the distribution of these features across the different drivers, thus visually evaluating the impact of the driver and/or vehicle on the measurements.

The drivers traveled the same scenario with different times, so to compare the time series, it was decided to show them in relation to distance rather than timesteps. The GPS had a sampling rate of 1 Hz, while the original time series had a time step of 0.01 seconds. I compressed the original time series with a 200:1 scale, meaning that each current time step is spaced 2 seconds apart (0.01 x 200). Therefore, I can calculate the distance traveled by multiplying the speed by 2, and I compute the cumulative distance by summing the previous values as well.

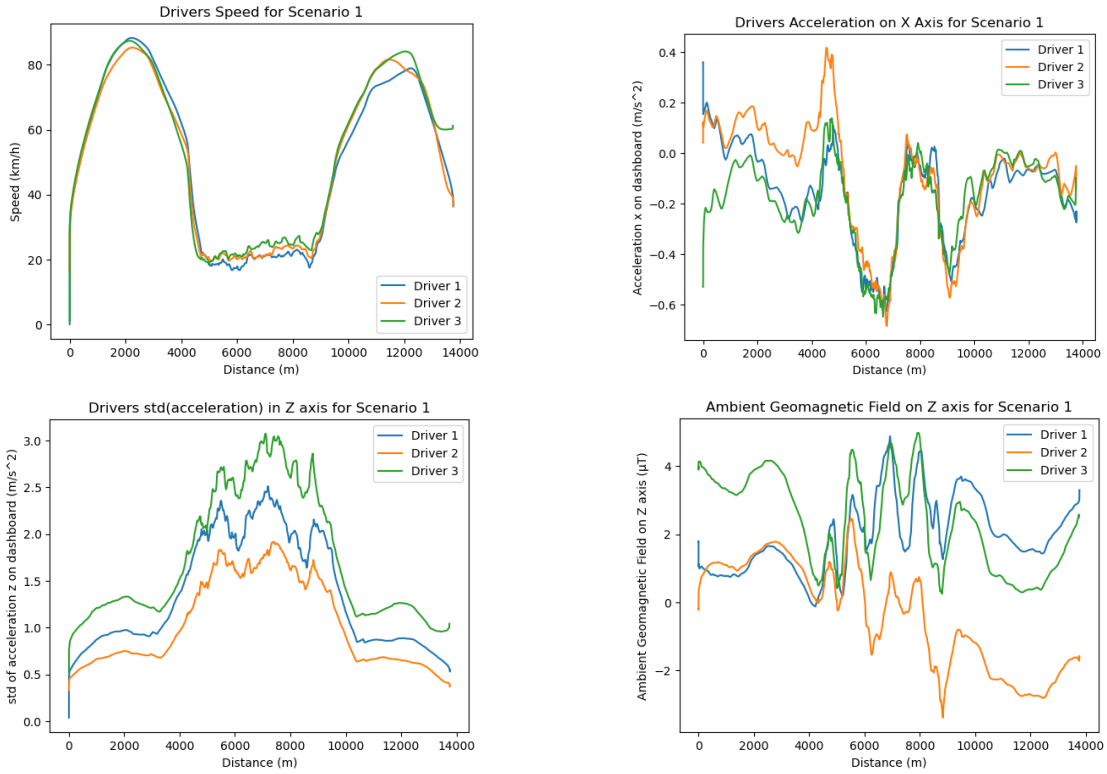


Figure 3: Univariate time series of the features for the different drivers on the Scenario 1.

Regarding speed, the three time series are very similar to each other. This suggests that, in this dataset, speed may provide more information about road characteristics, such as its condition or whether it belongs to an urban area, rather than about the behavior of individual drivers. The interpretation of the graph is therefore quite straightforward: all three drivers start from a standstill on a road in good condition (and outside the urban area); here, the speed increases to a peak and then decreases as they approach the urban area (which in this case also involves poor road conditions). In this segment, speed fluctuates more but remains between 15 km/h and 30 km/h. Finally, the vehicle returns to the initial road in the opposite direction, so we observe a bell-shaped curve that mirrors the initial one.

For std_acc_z_dashboard_L, the shape of the plot also has a straightforward interpretation. In the first segment, the vehicle is much more stable, while in the segment with poor road conditions, the vehicle is likely much less stable, leading to an increase in the standard deviation of the acceleration along the Z-axis. However, in this case, the three drivers' time series have noticeably different amplitudes; this indicates that while the shape is likely derived from road characteristics, the amplitude depends on the driver and/or the vehicle. This could simply be explained by the greater stability of some vehicles compared to others.

Regarding acc_x_dashboard_L, which represents the vehicle's longitudinal acceleration, it appears that the most significant differences between the three drivers occur in the first part of the circuit. Beyond the 5 km mark, all three drivers seem to follow a specific pattern in their accelerations and decelerations.

Finally, the most challenging graph to interpret is mag_z_dashboard_L. Here too, a pattern in the shape can be observed, but given the nature of this feature, it remains difficult to provide a meaningful interpretation. The ambient geomagnetic field can also vary depending on road conditions, such as when the road surface is wet or due to the presence of temporary debris, therefore, it is better to avoid making overly speculative analysis.

3 Data Preparation

Structural Feature Extraction

As previously explained, in this dataset, each driving process represents a single time series, where each timestep is also associated with a Road Condition Label. To address the research questions about Road Condition Classification, it was necessary to model the granularity of individual road segments rather than the entire driving process. For

this reason, the approach involved extracting multiple time series of equal length from each driving process.

To extract the time series, a sliding window was applied to each driving process with a window size of 200. Each time series thus represents a road segment covered in 2 seconds. The choice of the sliding window size is entirely arbitrary, and future developments could investigate the impact of varying this hyperparameter. To assign a class to each time series, the mode of the labels from the observations within the window was used.

Furthermore, in this study, the structural representation of time series was chosen as the focus. Instead of using the raw shape of the time series, structural statistics were employed as predictors. Future work could explore answering the same research questions using a shape-based representation.

For each time series, the mean and standard deviation of its features (excluding timestep, latitude, and longitude) were calculated. The final result is a tabular dataset with 110 independent variables and a single dependent variable (Road Condition Label).

Features Selection

As observed during the Data Exploration phase, measurements from various sensors can be highly correlated. This could be problematic for several reasons:

- many classification algorithms are negatively affected by multicollinearity among independent variables;
- even for algorithms that are not sensitive to multicollinearity, a smaller number of features typically reduces computational time;
- interpreting feature importance becomes excessively complex with a large number of features.

For these reasons, feature selection was performed based on linear correlation. After computing the correlation matrix, 310 pairs of features with an absolute correlation value greater than 0.75 were identified. Consequently, from each group of highly correlated features, only one was arbitrarily selected. Using this method, 81 variables were removed from the initial set of 110. Therefore, the 29 independent variables to be used for subsequent tasks are: avg_speed, avg_acc_x_dashboard_L, avg_acc_y_dashboard_L, avg_acc_z_dashboard_L, avg_acc_z_above_suspension_L, avg_acc_z_below_suspension_L, avg_acc_z_dashboard_R, avg_acc_z_above_suspension_R, avg_acc_z_below_suspension_R, avg_gyro_x_dashboard_L, avg_gyro_y_dashboard_L, avg_gyro_z_dashboard_L, avg_mag_z_dashboard_L, avg_temp_dashboard_L, avg_temp_above_suspension_L, std_acc_z_dashboard_L, std_gyro_z_dashboard_L, std_mag_x_dashboard_L, std_mag_y_dashboard_L, std_mag_z_dashboard_L, std_mag_x_above_suspension_L, std_mag_y_above_suspension_L, std_mag_z_above_suspension_L, std_mag_x_above_suspension_R, std_mag_y_above_suspension_R, std_mag_z_above_suspension_R, std_temp_dashboard_L, std_temp_above_suspension_L, std_temp_below_suspension_L.

4 Anomaly Detection

In this section, the driving processes recorded in Scenario 1 were analyzed to identify anomalies. For this task, the raw representation of the data (rather than structural features) was used, as it is generally more effective to detect anomalies directly from sensor observations. However, to avoid the curse of dimensionality, only the variables retained after the feature selection phase were considered. Additionally, the speed feature was excluded to prevent it from having an excessive impact and revealing only trivial patterns, thereby leveraging other variables to detect more meaningful anomalies.

The algorithm used for anomaly detection was the Isolation Forest, with a contamination parameter set to 0.005. This means that each observation was assigned an outlier score, but only 0.5% of the data was labeled as outliers by the algorithm.

The algorithm was trained separately on the data from PVS1, PVS4, and PVS7, resulting in three distinct models. These models were then used for inference on the respective training sets.

Driving Processes Point of View

In this experiment, the outlier scores of the three driving processes were kept separate to compare the distribution of outlier scores across the three time series. This approach allows us to visually analyze how many outliers appear unique to a specific driving process, potentially indicating driving behavior anomalies or dynamic situations (e.g., a pedestrian crossing unexpectedly). Conversely, if the outliers are similarly distributed along the route, they are more likely to relate to static road features, which are the primary focus of this study.

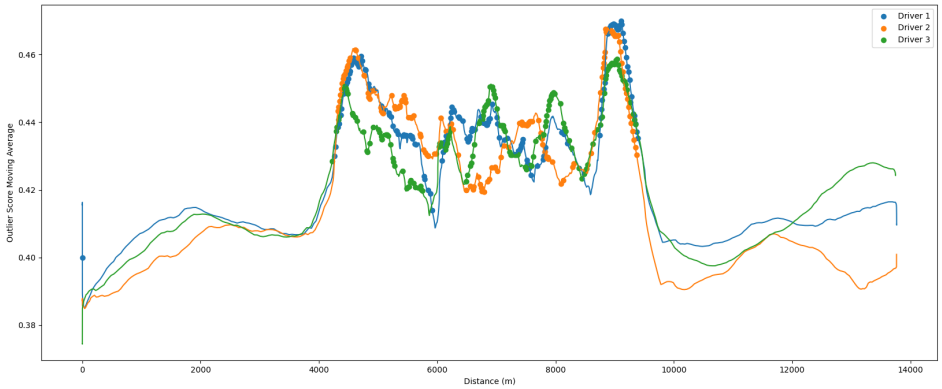


Figure 4: Distribution of the outlier scores assigned by the Isolation Forest models for the three drivers in Scenario 1, with points representing the observations labeled as outliers.

As we can see from Figure 4, almost all the outliers are concentrated in the central part of the route, which corresponds to the urban section. Additionally, we observe that there is a clear common pattern in the outlier score distribution across the three time series. It seems possible, therefore, to identify anomalies that are not related to the driving behavior or vehicle, but are inherent to the road itself. However, it is important to note that to the time series in the graph is applied a denoising process using a moving average (window of 32 observations), so the actual distribution of the outlier scores is much noisier than what is shown.

Road Point of View

Since the previous subsection highlighted a potentially interesting pattern, the aim of this experiment is to identify road anomalies by combining the outlier scores and outlier labels assigned to the three different driving processes.

To achieve this, the observations classified as outliers for each time series were selected, and only anomalies occurring at the same location (with some tolerance) in all three driving processes were kept. Therefore, for each anomaly in one time series, if a very close anomaly (in terms of coordinates) is also present in the other two time series, it is selected. This should allow for modeling static anomalies of the road segment, rather than those specific to an individual driving process. Finally, the outlier score assigned to each anomaly was calculated as the average of the three outlier scores from the three driving processes.

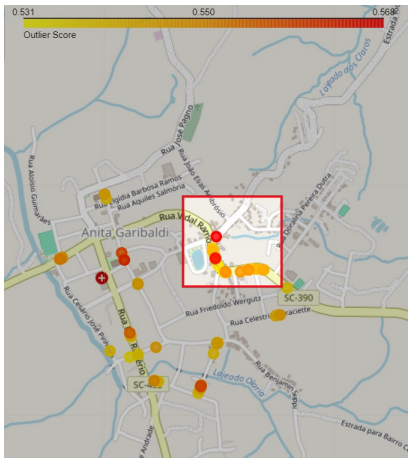


Figure 5: The detected road anomalies for Scenario 1, with a dense cluster of anomalies highlighted in red.

Thanks to this process, 92 outliers were detected, all in the urban area of the scenario, scattered across different zones. Trying to explain what these anomalies might represent is quite complex, so I performed a simple visual analysis, focusing on the section of the road highlighted in Figure 5. In this segment, there are several anomalies, and to try to understand their meaning, I manually analyzed the video recordings made during the driving processes, correlating the correct road sections using the timestep. The anomaly at north in the highlighted area in Figure 5, which has a very high outlier score, appears to correspond to a cobblestone bump (Figure 6a). In general, this entire

section of the road is quite uneven and full of potholes, such as the one shown in Figure 6b, so it seems reasonable that there is a cluster of anomalies in this area.

(a) Cobblestone Speed Bump

(b) Road Pothole

Figure 6: Two possible anomalies detected in the road segment highlighted in Figure 5

Outlier Score Explainability

Finally, a simple approach was tested to try to explain which features were most important for the Isolation Forest models in assigning the outlier score.

To do this, three regression decision trees (one for each Isolation Forest model) were trained, using the outlier score as the target variable. Of course, no hyperparameter tuning was performed, as overfitting is not a concern in this approach (in fact, it is the goal).

To evaluate how effective these decision trees are in replicating the black-box models, the R2 score was calculated on the training set. The results were: 0.46 for PVS1, 0.51 for PVS4, and 0.42 for PVS7. We can therefore confirm that the decision trees have a significant effectiveness in modeling the outlier score.

At this point, the feature importances were extracted from the three decision trees, and for each feature, the average importance across the three models was calculated. The results are shown below.

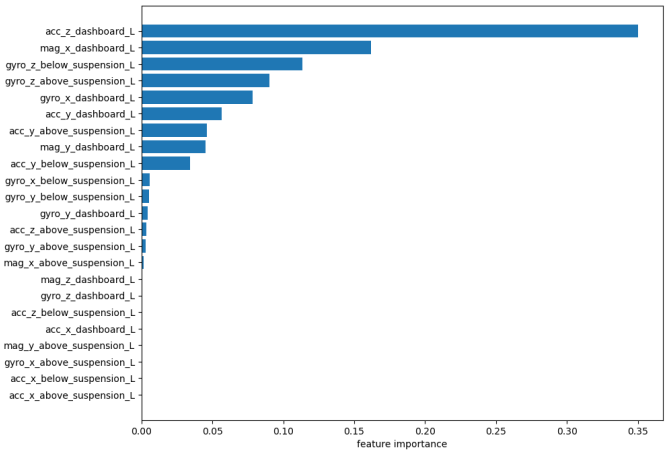


Figure 7: Feature importances for the outlier score decision tree regressor

An interesting observation is that the most important feature in determining the outlier score is the acceleration on the Z-axis. This seems to support the observations from the previous subsection, where the anomalies could represent road potholes or bumps, as the car would bounce and thus experience extreme acceleration values on the Z-axis.

5 Road Condition Classification

In this section, the structural feature representation of the dataset was used to attempt a classification of the Road Condition class (bad/regular/good). Three different algorithms were employed: logistic regression, serving as a baseline to understand the performance achievable with a simple linear model; and subsequently, the Random Forest and Light Gradient Boosting Machine algorithms.

For each of these algorithms, a hyperparameter tuning phase was performed using 5-fold cross-validation before evaluation on the test set. This phase will not be further detailed in this report.

Experimental Setup

Research Question 2, presented in Section 1, clearly emphasizes the need for a road condition prediction model that is independent of the driver, vehicle, and location. Therefore, it is crucial to construct the training set and test set while adhering to these constraints.

To ensure independence from the driver and vehicle, the training data was composed of driving processes from drivers 1 and 3. Conversely, the test set included only measurements from driver 2. This driver was selected because, as observed during the Data Exploration phase, particularly in Figure 3, driver 2 appears to be the most distinct among the three.

Achieving independence from location is more challenging, as shown in Figure 2a, where all three scenarios share several overlapping road segments. To address this, only the unique section of Scenario 2, which does not overlap with other scenarios, was included in the test set. Specifically, observations from PVS5 (Driver 2, Scenario 2) were selected, ensuring that none were in the same location (within a tolerance) as those from PVS4 and PVS6.

Figure 8: In blue, the observations selected for the test; in red, the observations recorded in Scenarios 1 and 2.

As shown in Figure 8, there are no longer overlapping observations between the test set and Scenarios 1 and 3. In conclusion, the training set consists of data from PVS1, PVS3, PVS7, and PVS9, comprising a total of 469955 observations. Meanwhile, the independent subset derived from PVS5 serves as the test set, whose composition is detailed below.

Class	n. record	% record
bad	45761	54.44
regular	15053	17.91
good	23244	27.65

Table 2: Test Set Composition

Classification Results

The models selected during hyperparameters tuning are:

- **Logistic Regression** (maxIter=100, regParam=0.01), with a validation accuracy of 0.8686.
- **Random Forest** (numTrees=200), with a validation accuracy of 0.8658.
- **LGBM** (learningRate=1, featureFraction=1, numTrees=300), with a validation accuracy of 0.9779.

Below the test set evaluation results.

Model	Accuracy	F1 score	F1 "Bad"	F1 "Regular"	F1 "Good"
Logistic Regression	0.4209	0.3210	0.0006	0.3510	**0.9274**
Random Forest	0.4120	0.2930	0	0.3557	0.8292
LGBM	**0.7511**	**0.7648**	**0.7594**	**0.5543**	0.9118

Table 3: Evaluation metrics on the test set

The first observation regarding the results is that the performance on the test set is significantly lower than that achieved during validation. This highlights the simplicity of solving the task when the training data includes observations belonging to the same driver or the same road segment. On the other hand, developing a prediction model independent of the driver, vehicle, and location is a task with a much higher level of complexity.

Regarding test set performance, the logistic regression and random forest models appear to only succeed in identifying the Good class, while they struggle significantly with identifying Bad road segments. This suggests that the Good class is well-defined and easier to distinguish, whereas the models face great difficulty differentiating between the Bad and Regular classes.

In contrast, the LGBM model performs significantly better than the others. It demonstrates an ability to recognize the Bad and Regular classes to a certain extent. This allows it to achieve 75% accuracy and a 76% F1 score. We can therefore conclude that the model has meaningful predictive capability for this task.

Features Importance

Since the only model that proved to be significant was the LGBM, the features importance was extracted and analyzed exclusively for this model.

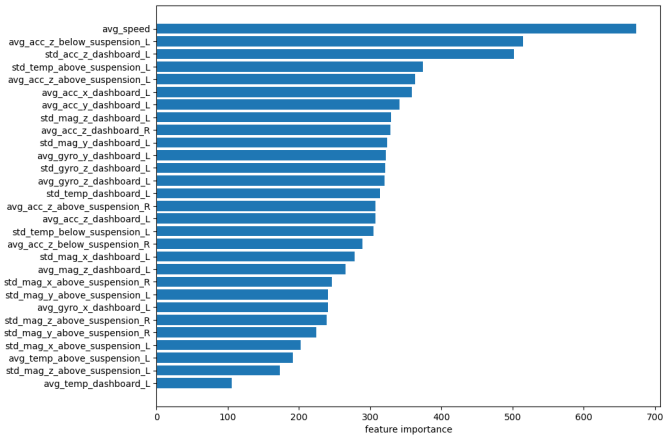


Figure 9: Feature importances for the LGBM classification model

From Figure 9, we can make several observations:

- The most important feature, as inferred during the Exploration phase, is speed.
- In second and third place are avg_acc_z_below_suspension_L and std_acc_z_dashboard_L. This confirms the previously presented intuition that acceleration along the z-axis is highly significant in characterizing the road segment, modeling vehicle instability and sudden bumps.
- Beyond speed, the most interesting measurements are likely those from the accelerometers.

6 Conclusions

In conclusion, I attempt to answer the research questions presented in the introduction:

1. *Can meaningful features of the road be identified through an unsupervised analysis of the measurements?* From the analysis in Section 4, using an anomaly detection approach, 92 anomalies were identified in Scenario 1. These anomalies were consistent across all three drivers and appear to represent potholes, bumps, or deteriorated road segments.

2. *Is it possible to develop a prediction model for road conditions that is independent of the driver, vehicle, and location?*
   Although Section 4 revealed that training a model under these constraints is highly challenging, the LGBM model achieved good results, demonstrating significant predictive capabilities for road conditions.

3. *Which types of features are most important among the following: Speed, Accelerometers, Magnetometers, Gyroscopes, and Thermometers?*
   Speed appears to be the most important feature, as anticipated during the Exploration phase and later confirmed through the analysis of feature importance in the classification model. Other significant features pertain to vehicle acceleration, particularly acceleration along the z-axis, as noted during the Exploration phase and further observed in both the Outlier Score Explainability section and the feature importance analysis for classification. The remaining features appear to be secondary.