Vrije Universiteit Amsterdam

Bachelor Thesis

Eco-driving Analysis and Driving Style Feedback for Passenger Cars Using OBD-II Data and Machine Learning

---

Author: Yudong Fan 2662762

1st supervisor:	Kees Verstoep
daily supervisor:	Kees Verstoep
2nd reader:	Dr. Thilo Kielmann

A thesis submitted in fulfillment of the requirements for the VU Bachelor of Science degree in Computer Science

July 5, 2022

Abstract

Eco-driving is a driving style which has a great potential in reducing fuel consumption at a relatively low cost. In order to promote an Eco-driving style, an Eco-driving assistant is designed to analyze driving performances and provide Eco-driving feedback accordingly. The system makes use of the data collected by a vehicle onboard diagnostic system (OBD-II) and machine learning techniques to perform a comprehensive analysis on various driving events and the general driving style. As a result, the system can evaluate drivers' general driving styles and provide comprehensive textual feedback on specific driving behaviors to help drivers to enhance their Eco-driving practice. By following the feedback, drivers are expected to drive in a more Eco-friendly manner and ultimately reduce the fuel consumption for their trip.

Contents

1

Introduction

Nowadays, environmental crises such as energy shortage and global warming grab more attention worldwide as human society keeps growing and evolving fast. Sustainability becomes more and more important and is relevant to everyone. One of the most important factors which gives rise to those environmental crises is transportation. According to a report published by IPCC (Intergovernmental Panel on Climate Change) [1], in 2010, transportation made up about 14% of 2010 global greenhouse gas emissions. Moreover, global greenhouse gas emission has increased significantly since 1900. For example, the global CO_2 emissions in 2014 have increased by about 150% since 1970. Among various emission sources during that period, fossil fuel combustion and industrial processes contribute about 80% of the total greenhouse gas emissions increased. Scientists claim that if the current energy consumption keeps growing without any limitation, the reserves for fossil fuels on earth will be emptied within 21 century [2]. Consequently, it is crucial to take actions to reduce fuel consumption and pollution from transportation in order to cope with the larger sustainability and environmental issues.

1.1 Context

Eco-driving is a driving style which can reduce fuel consumption for vehicles by asking drivers to drive in a certain way. In general, it involves actions such as accelerating and decelerating moderately, keeping the speed stable, avoiding excessive idling, etc. [3]. Eco-driving is an attractive method to promote not only because it reduces fuel consumption and produces tangible safety benefits, but also because it has a relatively low or even no cost in comparison to other methods which are designed to achieve the same goal. However, Eco-driving is often overlooked as many might doubt its actual effect [3].

In fact, researches have shown that, in terms of saving energy, Eco-driving can perform surprisingly well. According to Ford [4], roughly 25% improvements in fuel economy during short-term driving can be achieved with the correct Eco-driving style. Furthermore,

when driving in normal and sustained situations, drivers can save 5% of fuel after initial training and 10% of fuel when there is continuous feedback on the driving style [3]. In short, Eco-driving has great potential in terms of saving energy especially when there are appropriate assistants. This thesis focuses on analyzing driving data from a vehicle onboard diagnostic system (OBD-II), and ultimately design a system which is capable of providing real-time, intuitive, flexible, and comprehensive Eco-driving feedback and suggestions.

1.2 Problem Statement

The specific problem addressed in this thesis is how to design an Eco-driving assistant in order to reduce fuel consumption. The essence of this target can be broken down into the question of how to analyze links between driving data and Eco-driving in a scientific manner. In other words, how to map driving data collected by machines with driving behaviors performed by drivers, analyze correlations between driving data and fuel consumption, and eventually transform data into human readable and feasible Eco-driving feedback. As discussed in the previous section, Eco-driving has great potential in reducing fuel consumption in case adequate support is provided. Therefore, answering this problem is valuable as it has practical application and benefits. In addition, regardless of the improvement in the car industry and material science, Eco-driving retains its value as long as energy has limitations. Even in the case when fully self-driving cars become reality in the future, Eco-driving assistants can still be deployed. The only difference is that the Eco-driving principles will be enforced by machines instead of humans in favor of better fuel efficiency.

1.3 Research Questions

In order to generate comprehensive feedback for Eco-driving, the system needs to be able to perform accurate, data-driven, and timely driving style analysis. The first step is to justify the concept of Eco-driving. Namely, distinguishable metrics for Eco-driving need to be defined in a countable way because Eco-driving itself is merely an abstract doctrine and thus is very hard to measure and calculate without specific metrics. The first question then is about how to define Eco-driving in a measurable manner?

Next, in order to discover connections between the actual driving events and the driving data, the second question is how to interpret and classify Eco-driving events with driving data? More specifically, what features in the data are influential in terms of changing fuel consumption, how do they relate to actual driving behaviors, and how to classify those driving behaviors into different driving styles?

Finally, in order to provide comprehensive driving style feedback and advice, general driving style detection and clustering are needed. The question is how to differentiate

different driving styles? Essentially, the challenge is about how to determine the driving style given a complex feature space? After that, comprehensive driving style analysis is possible, and the last question is how to provide Eco-driving feedback accordingly for both real-time continuous feedback during a trip and summarized accumulated feedback after a trip?

1.4 Research Methodology

To answer the research questions, a design system methodology proposed by K. Peffers et al. [5] is adopted. In principle, the methodology incorporates the following activities:

- Activity 1: Problem identification and motivation
- Activity 2: Define the objectives for a solution
- Activity 3: Design and development
- Activity 4: Demonstration
- Activity 5: Evaluation
- Activity 6: Communication

Although it is a general guideline to the structure of the thesis, it offers important insights from the perspective of design science research methodology including practice rules and principles for design system research. So far, activity 1 and 2 are addressed in previous sections.

1.5 Thesis Contributions

This thesis intends to design and prototype an Eco-driving assistant which is capable of analyzing driving behaviors, classifying driving styles, and providing Eco-driving feedback. It focuses mostly on driving data analysis. Specifically, it combines prior studies and frames an Eco-driving model which can be widely adopted as a baseline in order to perform general driving style analysis with limited driving data. It uncovers hidden correlations between driving data and fuel consumption. Furthermore, it attempts to explain these correlations and define more factors which have significant meaning in both analyzing and explaining Eco-driving based on them. Lastly, it promotes a practical process to generate real-time and accumulated Eco-driving feedback. In general, it offers both intuition and practical application in the field of Eco-driving analysis with OBD-II data.

1.6 Plagiarism Declaration

I confirm that this thesis work is my own work, is not copied from any other source(person, Internet, or machine), and has not been submitted elsewhere for assessment.

1.7 Thesis Structure

Chapter 2 introduces essential background knowledge about the concept of OBD-II data, the limitations in the experiment, and the datasets used in this thesis. Chapter 3 introduces the workflow of the design in detail. It also explains principles and components of the system. Chapter 4 performs a case study using the Eco-driving system. Chapter 5 evaluates performance of the system. The evaluation includes reporting of the results, explanation of the experimental setup, limitations of the system, and a summary on evaluation of the system. Chapter 6 describes prior studies and efforts which are related to this thesis. Finally, Chapter 7 wraps up the thesis with answers to the research questions, limitations to the current work, and future work directions.

2

Background

Beyond the concept of Eco-driving, OBD-II data is another essential prerequisite to this thesis. OBD is the acronym for On-Board Diagnostics, and OBD-II represents the second generation of OBD. In short, OBD-II is a built-in self-diagnostic system which is capable of accessing, monitoring, transmitting, and reporting vehicle operational status such as engine speed, engine load, throttle position, etc., through various vehicle subsystems [6]. Furthermore, OBD-II works on top of a message-based protocol called CAN (Controller Area Network). CAN is a typical bus-based communications standard, so it is also often referred to as CAN-Bus. It allows vehicles' electronic units to communicate with each other in order to monitor and transfer data (errors, value adjustments, etc.) [7]. OBD-II was developed in 1992, and was mandatory to be installed in all cars and small trucks in 1996 in the United States [8]. Later in 2003, Europe has also adopted OBD-II on both petrol and diesel vehicles. As of today, almost all newer cars around the world support OBD-II and run on CAN (ISO 15765) [9]. As mentioned in previous sections, all driving data used in this thesis were collected using OBD-II scanner. This type of data in general is called OBD-II data. More specifically, a special set of codes called the OBD-II parameter IDs (OBD-II PIDs) are used to request data from vehicles. A standard OBD-II PIDs table can be found in [10]. In short, OBD-II data has significant practical usage as it provides a universal standard to driving data which is exceptionally helpful in analyzing driving styles.

The original datasets used for the experiments in this thesis are acquired through a public database on Kaggle [11]. After initial data wrangling and engineering, the actual dataset consists of two subsets. The first contains information about 3 gasoline passenger cars and 3 drivers for their daily routes, and the second contains information about 1 gasoline passenger car and 19 drivers for the same route. Specific features in the driving data are shown in Table 2.1 (Details of the features can be found in the standard OBD-II PIDs table [10]). Among all features, fuel rate and fuel consumption are calculated using the formula (1) and (2) respectively given by Meseguer et al.[12].

(1) *Fuel rate [l/h] = (MAF · 3600)/* AFR_A *· FD*

Where MAF refers to Mass Air Flow (g/s), AFR_A refers to the Air-to-Fuel Ratio which is 14.7 for gasoline , and FD refers to Fuel Density (g/l) which is 820 for gasoline [12].

(2) *Fuel Consumption [l/100km] =* $\frac{FuelRate[l/h]}{Speed[km/h]}$ *· 100*

It is important to note that the experiment in this thesis has certain limitations. Namely, the experiment only takes into account the OBD-II data provided by the datasets. External factors such as traffic, road type, weather, altitude, etc. are omitted because the purpose of this thesis is to focus on analyzing driver behaviors and internal driving data collected by OBD-II. Thus, the experiment made an assumption that part of the external conditions are implicitly reflected in the collected driving data. The result therefore is constrained and might deviate from real-life driving circumstances given different external factors. However, it still provides insightful conclusions from the perspective of internal Eco-driving analysis.

In addition, all the machine learning algorithms used in this thesis are adopted from Scikit-learn [13] which is an effective machine learning toolkit in Python. Since the main subject of this thesis is not machine learning, theories of the machine learning concepts will not be introduced in details, more exhaustive information about machine learning concepts mentioned in this thesis can be found in the Scikit-learn user guide [14].

Index	Features	Unit
0	Barometric pressure	kPa
1	Engine coolant temperature	$^{\circ}C$
2	Fuel level	%
3	Engine load	%
4	Ambient air temperature	$^{\circ}C$
5	Engine speed	rpm
6	Intake manifold pressure	kPa
7	Mass air flow (MAF)	g/s
8	Air intake temperature	$^{\circ}C$
9	Speed	km/h
10	Short term fuel trim bank 1	%
11	Engine run-time	sec
12	Throttle position	%
13	Ignition timing advance	degree
14	Equivalence ratio	ratio
15	Fuel rate	l/h
16	Fuel consumption	l/100km

Table 2.1: Features in the dataset.

3

Design of the Eco-driving assistant: an analytical Eco-driving system

3.1 The driving performance score metric

In pursuit of a general Eco-driving model, the metric of Eco-driving needs to be defined first. In fact, it is challenging to find a mathematical formula which is able to tell precisely the boundary of proper Eco-driving style because external factors such as the driving environment, the vehicle's physical condition, etc. are not fixed and not always accessible. Even when there occasionally exist circumstances where all relevant data are accessible and the mathematical formula is totally correct, theoretically optimal solutions are not always feasible for humans. For example, certain Eco-driving attempts like shifting a gear may not be applicable due to the concern for safety or traffic in some cases. Also, drivers may simply choose to not follow the Eco-driving practice for a variety of reasons. For instance, a driver might not willing to lower his speed in order to save fuel when he is in hurry to catch up an appointment. In short, circumstances may force non-optimal driving choices, so a better way to assess driving style and provide feasible Eco-driving advice is to enforce a relative and flexible metric instead of a fixed formula. Therefore, a dynamic metric for Eco-driving is adopted following the approach given by Khedkar et al. [15] (Fig. 3.1). In general, this metric assesses Eco-driving performances for a certain driving instance and returns an Eco-driving performance score ranging from 0 to 100 based on the best Eco-driving performance recorded for the same car and the same trip before. It not only provides a possible solution to quantify and then compare Eco-driving performances, but also ensures that the result of the system is realistic because the result is derived from real human performances.

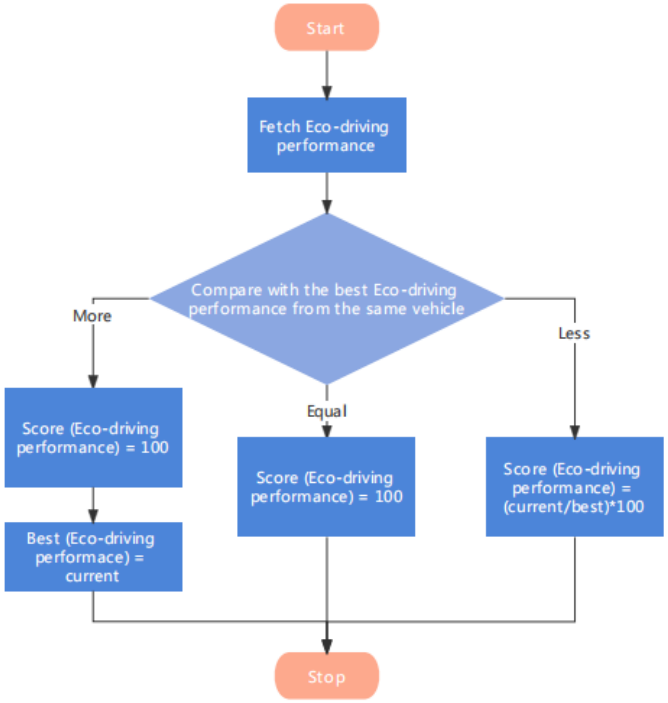

Figure 3.1: Eco-driving performance score metric. [15]

3.2 Uncover hidden correlations to fuel consumption

Next, given the dynamic Eco-driving performance score metric (Fig. 3.1), specification of Eco-driving performances needs to be made. In other words, the parameters passed into the performance metric need to be determined. In principle, not all features in the driving data are related to fuel consumption, and the performance score metric should only take into account driving data which are influential in terms of changing fuel consumption. Thus, the task in this step is to find those strongly correlated features in the driving data. Notably, there are some theoretically obvious correlations such as the features which are directly involved in the fuel consumption formula (see (1) and (2)). Specifically, these features are Speed, MAF, and Fuel rate. However, there might exist obscure correlations between other driving data and fuel consumption.

In order to achieve this goal, supervised machine learning is introduced as a methodology to uncover hidden correlations between fuel consumption and other driving data. That is, a regression machine learning task is formed in which the fuel consumption is the target value given all other features as the training data. The purposes of this machine learning task are not only to train good predictive models which are capable of predicting fuel consumption accurately, but also to backtrack influential features besides those theoretically correlated data mentioned above. Initially, auto-sklearn [16] was used to establish a

baseline for the machine learning task. In general, auto-sklearn is an automated machine learning toolkit which is capable of automatic machine learning algorithm selection and hyperparameter tuning. It takes advantage of Bayesian optimization in order to select best performing models and model configurations. Moreover, it includes a meta-learning step to warm-start the Bayesian optimization procedure in favor of better efficiency and an ensemble construction in pursuit of better accuracy [16]. For a regression task, auto-sklearn will search and select algorithms in its regression algorithm bank (specification for algorithms can be found in Table 4.1) and tune the hyperparameters for those algorithms automatically. Consequently, it is able to provide a list of regression algorithms with their best performing configurations which suits the machine learning task the most. Although the auto-sklearn processes are mostly black-box operations, the result can still be heuristic and helpful in terms of understanding the trait of the dataset and selecting algorithms with good adaptability to the dataset.

RMSE (Root Mean Squared Error) and MAPE (Mean Absolute Percentage Error) were used as the metrics to evaluate the result of regression algorithms. On the one hand, RMSE (see (3)) is a measure of errors between paired observations expressing the same phenomenon. It is especially good in terms of interpretation of loss because it produces output values in the same unit as the required output variable [17]. On the other hand, MAPE (see (4)) measures the accuracy of a model as a percentage. It is easier to understand because variable's units are scaled to percentage units, and it avoids the problem of positive and negative errors canceling each other because it calculates the absolute values of errors [18]. After that, manual observation and adjustment based on the auto-sklearn result can be made. Finally, models with relatively high score in predicting fuel consumption will be selected for the next step.

(3) *RMSE* = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(A_i - F_i\right)^2}$

(4) *MAPE* = $\frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right|$

where A_i is the actual value, F_i is the forecast value, n is the number of fitted points.

Once the good models were selected, permutation feature importance [14] technique was applied to target important features in terms of predicting fuel consumption. Permutation feature importance is the decrease in the model score when a single feature value is randomly shuffled. This procedure breaks the link between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature [19]. Particularly, the predictive power of the applied model needs to be evaluated prior to computing importance because permutation feature importance does not reflect the intrinsic predictive value of a feature by itself but how important a feature is for a certain model. Therefore, the previous step only passes models with good scores in predicting fuel consumption because the permutation feature importance technique only

works well when the applied models are robust in predicting the target value. Then, features in the dataset which have strong correlation to fuel consumption are discovered. So far, valuable features in the driving data are selected. They can be used to further develop other important features. For example, fuel efficiency can be derived given the fuel consumption and distance travelled. As a result, a complex Eco-driving feature space is generated.

3.3 Instantaneous driving style analysis

After all the influential features in terms of fuel consumption are gathered, some of them are passed into the Eco-driving performance score metric for relative scores. Specifically, relative fuel efficiency score, relative throttle score, and relative Eco-driving score are generated by the metric and added to the complex Eco-driving feature space. However, some features such as speed-RPM ratio and vehicle speed are either self-explanatory or do not suit the performance metric, therefore not all features are necessary to go through the performance score metric and be transformed into scores before they can be put into comparison and evaluation. Specification for the complex Eco-driving feature space can be found in Table 3.1 [20][21]. Eventually, combining scores generated by the Eco-driving performance metric and self-explanatory driving data results in a multi-dimensional reflection in terms of driving style for every driving index in the driving data. After that, it is possible to do a white-box analysis and give multi-dimensional instantaneous driving style feedback accordingly based on the driving classification metrics (Table 3.2, 3.3, 3.4, 3.5, 3.6) . Among these metrics, Table 3.2, 3.3, 3.4, and 3.5 added feedback guidelines for trip analysis on top of the original metrics provided by Massoud et al. [20], and Table 3.6 is constructed based on the safe-driving suggestion given by Chen et al. [21].

Feature name	Formula	Definition
Fuel efficiency	$\frac{KilometersTravelled}{FuelConsumptionSoFar}$	A value represents the efficiency of fuel consumption for a certain distance, the higher the better.
Relative fuel efficiency score (0-100)	Pass the fuel efficiency to the performance metric (Figure 3.1), the initial scores are set to be 50.	A score ranging from 0-100 which scores the relative fuel efficiency, the higher the better.

Throttle efficiency	$100 - CurrentThrottlePosition$	A value represents the throttle efficiency for a certain driving index, the higher the better. It is calculated in this way because the less open the throttle is, the less fuel consumption the engine needs.
Relative throttle score (0-100)	Pass the throttle efficiency to the performance metric (Figure 3.1), the initial scores are set to be 50.	A score ranging from 0-100 which scores the relative throttle efficiency, the higher the better.
Eco-driving score (0-100)	$(0.75 * RelativeFuelEfficiencyScore) + (0.25 * ThrottleEfficiency)$, the 0.75 and 0.25 in the formula consist of balancing the trade-off between the impact of driving pattern and other factors on fuel economy (e.g., weather condition) [22].	An indicator of the trade-off between the fuel efficiency and throttle efficiency ranging from 0 to 100. The higher the score, the better the Eco-driving performance.
Relative Eco-driving score (0-100)	Pass the Eco-driving score to the performance metric (Figure 3.1), the initial scores are set to be 50.	A score ranging from 0-100 which scores the relative Eco-driving score, the higher the better.
Modified speed (km/h)	$\frac{Speed}{MaxSpeed}$, by default the max speed is 220 km/h.	A variation of the vehicle speed. It is used to generate a rational Speed-RPM ratio.
Modified RPM	$\frac{RPM}{MaxRPM}$, by default the max RPM is 8000.	A variation of the engine RPM. It is used to generate a rational Speed-RPM ratio.
Speed-RPM ratio	$\frac{ModifiedSpeed}{ModifiedRPM}$	An indicator of the current gear load. A value between 0.9 to 1.3 is considered good in terms of gear performance [21]

Speed change rate	$\frac{Speed_{t2}-Speed_{t1}}{t2-t1}$, where t1 and t2 represent engine run time for two continuous driving indices. The driving indices is pegged with the OBD-II sampling frequency which is 6 seconds in this experiment. This peg holds for all formula involving continuous driving indices.	A value indicate the change of vehicle speed during a certain time.
RPM change rate	$\frac{RPMt2-RPMt1}{t2-t1}$, where t1 and t2 represent engine run time for two continuous driving indices.	A value indicate the change of engine RPM during a certain time
Acceleration (m/s^2)	$\frac{(Speed_{t2}-Speed_{t1})\div 3.6}{t2-t1}$, where t1 and t2 represent engine run time for two continuous driving indices. The number 3.6 in the numerator is used to convert unit from km/h to m/s^2.	Similar to the speed change rate. The only difference is that the unit for acceleration is m/s^2.
Car jerk (m/s^3)	$\frac{Acceleration_{t2}-Acceleration_{t1}}{t2-t1}$, where t1 and t2 represent engine run time for two continuous driving indices.	A variation of the acceleration. It indicates a driver's acceleration profile such as forced acceleration that requires more fuel consumption

Table 3.1: Eco-driving feature space [20][21].

Engine RPM (%)	Class	Feedback (instant)	Feedback (trip)
< 2000	Calm driver	Your engine RPM is good	Your average RPM performance is good
2000-2999	Moderate driver	Stay less than 3000 RPM to save fuel	Your average RPM performance has space for improvements, try staying less than 3000 RPM to save fuel
3000-	Aggressive driver	Slow down or shift up a gear to save fuel	Your average RPM performance has space for improvements, try slowing down your speed or shift up a gear to save fuel

Table 3.2: Driving classification with RPM [20].

Car jerk (m/sec)	Class	Feedback (instant)	Feedback (trip)
-4 <= jerk <= 3	Moderate driver	Your acceleration and deceleration are good	Your average acceleration and deceleration are good
jerk > 3	Aggressive driver	Avoid forced acceleration	Your average acceleration is aggressive, try avoiding forced acceleration
jerk < -4	Aggressive driver	Avoid sharp deceleration	Your average deceleration is aggressive, try avoiding sharp deceleration

Table 3.3: Driving classification with car jerk [20].

Also, driving styles can be clustered using unsupervised machine learning algorithms. That is, based on the Eco-driving feature space, all driving indices in the driving data are clustered into 3 types of driving style. Namely, saving drivers, normal drivers, and careless drivers based on the average fuel consumption for each cluster. In this thesis, K-means clustering [14] has been implemented for the clustering task. In general, it is easy to implement and can be applied to large datasets. Moreover, the biggest advantage which makes K-means stand out for this clustering task is its high efficiency in computing time [23]. Because real-time clustering requires continuous and rapid calculations, the efficiency of the algorithm is exceptionally important to make the system run smoothly without lagging. However, the process of clustering is more of a black-box operation especially when a complex feature space is involved, so it is harder to explain in comparison to the driving classification metrics. Nevertheless, even though the clustering process is poor in terms of interpretability, the resulting clusters are more comprehensive because the algorithm is able to calculate and make sense of all relevant features at the same time. As a result, every driving index will be assigned to a driving style class based on the driving data. According to the clustering result, feedback for general driving style will be provided using the classification metric shown in Table 3.7.

At this point, the system developed two pipelines to evaluate divers' instantaneous driving style at a certain driving instance based on the features in the corresponding driving data. One of them is a white-box analysis which is based on explicit classification metrics and capable of providing multi-dimensional feedback on driving behaviors. Another one is more of a black-box analysis which incorporates unsupervised machine learning to provide a general driving style evaluation.

Throttle position (%)	Class	Feedback (instant)	Feedback (trip)
0-39	Calm driver	Your throttle performance is good	Your average throttle performance is good
40-59	Moderate driver	Press the accelerator pedal gently for saving fuel	Your average throttle performance has place for improvements, try pressing the accelerator pedal gently for saving fuel
60-100	Aggressive driver	Release the accelerator pedal gradually, too much fuel is supplied to the engine	Your average throttle performance is aggressive, too much fuel is supplied to the engine. Try reducing the level of accelerator pedal pressed for saving fuel

Table 3.4: Driving classification with throttle position [20].

3.4 Accumulated driving style analysis

Accumulated driving style analysis is carried out using the mean values of the driving data for the studied trip. For instance, the mean value of fuel efficiency scores gathered in one trip will then becomes the fuel efficiency score for that trip. Following this, similar features for trips instead of driving indices can be calculated, and similar analysis using the two established pipelines can be adopted with only slight modification on the feedback text (Table 3.2, 3.3, 3.4, 3.5, 3.6). Besides, another classification metric for trips using the trip Eco-driving scores can be applied. Following the Eco-driving performance score metric (Fig. 3.1), relative trip fuel efficiency score and throttle efficiency score can be generated using the mean values and max values of fuel efficiency and throttle efficiency for that trip. Then, Eco-driving scores for trips which measures the general Eco-driving performance for an entire trip is calculated, and corresponding feedback can be given based on the driving classification with relative trip Eco-driving score (Table 3.8) [20]. The reason why this metric with Eco-driving score is only appropriate for accumulated driving style analysis instead of instantaneous driving style analysis is because the Eco-driving score for a driving index is a higher level assessment of that driving index based on the fuel efficiency and throttle efficiency, and it is heavily dependent on external factors of that driving index which this experiment does not have any control of. Thus, numerous instantaneous Eco-driving scores cannot reflect the accurate driving style assessment because they can easily get biased due to the abnormal instantaneous fuel or throttle efficiency caused by unknown external factors. However, after calculating means of the Eco-driving

Speed (km/h)	Class	Feedback (instant)	Feedback (trip)
Current speed < Max speed	Moderate driver	Your speed is safe	Your average speed is good
Current speed = Max speed	Moderate driver	Be careful, reaching the legal speed limit	Your average speed is close to the speed limit, be careful
Else	Aggressive driver	Over-speeding, slow down for safety and fuel saving	Your average speed is exceeding the speed limit, slow down for safety and fuel saving

Table 3.5: Driving classification with speed [20].

Speed-RPM ratio	Class	Feedback (instant)	Feedback (trip)
1.3 > Speed-RPM ratio > 0.9	Moderate driver	Keep the current gear for Eco-driving	Your average gear shifting is good
Else	Aggressive driver	Abnormal Speed-RPM ratio detected, try shifting a gear to save fuel	Your average gear shifting has space for improvements, try shifting a gear when abnormal Speed-RPM ratios are detected

Table 3.6: Driving classification with Speed-RPM ratio [21].

scores, the influence of external factors is weakened because uncertainties from those outliers are equally spread in the trip. Therefore, the classification with Eco-driving score can be adapted to accumulated driving style analysis. Although it is a similar evaluation on general driving style with the clustering method, it has a greater interpretability and therefore can be a good support or alternative to the clustering method. Last, an utility function which is designed in order to give a more intuitive and perhaps more impressive evaluation of drivers' driving style is provided. This utility function fetches the current fuel price and calculates the cost for the analyzed trip. Also, the function stores the best cost for a certain trip and compares it with the analyzed trip. Consequently, the function will tell the driver whether he/she saved or spent more money on a certain trip.

In summary, the design of the Eco-driving assistant incorporates two general pipelines for driving style analysis. Namely, a white-box analysis using explicit driving classification metrics and a black-box analysis using K-means clustering to cluster driving styles based on the driving data. Notably, the white-box analysis for accumulated driving style analysis incorporates two more evaluation metrics. That is, the driving style classification with trip Eco-driving score analysis and a trip cost analysis to provide a more comprehensive accumulated driving style analysis. Consequently, the system is able to provide both

Clustering result	Class	Feedback (instant)	Feedback (trip)
Saving driver	Saving driver	Your current driving style is Eco-friendly	Your average driving style based on comprehensive analysis is Eco-friendly
Normal driver	Normal driver	Your current driving style is normal, try following other advice to make progress	Your average driving style based on comprehensive analysis is normal, try following other advice to make progress
Careless driver	Careless driver	Your current driving style has space for improvements, try following other advice to make progress	Your average driving style based on comprehensive analysis has space for improvements, try following other advice to make progress

Table 3.7: Driving classification with driving style clustering.

real-time instantaneous driving style feedback during driving and accumulated driving style analysis for a trip after driving.

Relative trip Eco-driving score	Class	Feedback (trip)
60-100	Saving driver	Your average driving style based on fuel efficiency and throttle efficiency is Eco-friendly
40-59	Normal driver	Your average driving style based on fuel efficiency and throttle efficiency is normal, try following other advice to make progress
0-39	Careless driver	Your average driving style based on fuel efficiency and throttle efficiency has space for improvements, try following other advice to make progress

Table 3.8: Driving classification with trip Eco-driving score [20].

4

Case study: driving style analysis for 19 drivers

4.1 Train regression models and extract important features

In this chapter, a case study concerning driving style analysis for 19 drivers is carried out. As introduced in Chapter 3, before the system can appropriately start processing driving performances, the hidden correlations to fuel consumption need to be uncovered first. Besides normal approaches such as drawing and observing correlation maps for features, another approach taken is to apply the permutation feature importance technique. The first step is to train models which are capable of predicting fuel consumption accurately. For the regression task, datasets of 3 drivers with 3 cars and 19 drivers with 1 car are integrated to make a greater diversity of the dataset because the result is expected to be general and therefore can be adopted widely for different car makes and models. In total, 17105 available driving indices are selected. Among those indices, 10000 of them are set to be the training and validation set, and 7105 of them are set to be the test set. Features involved are the original 16 features (Table 2.1) with fuel consumption being the target value of the regression task. After applying the auto-sklearn, the results for 1 hour training are shown in Table 4.1.

Models with their RMSE less than 2 or MAPE less than 0.05 are considered good and selected. Then, permutation feature importance is applied on them to target influential features in terms of predicting fuel consumption. The idea is to extract those more important features and in the meantime remove irrelevant features to get rid of the undesired noise. After applying the permutation feature importance, Speed, MAF, Engine speed, Throttle position, Intake manifold pressure, Engine load, Fuel level, and Fuel rate appear to be the most important features.

Then, a critical observation about the generated features is done in order to check the validity of the generated features. Specifically, speed, MAF, and fuel rate can be de-

Index	Regression algorithms	RMSE	MAPE
0	MLP (multi layer perceptron)	1.762	0.140
1	Extra tree	1.810	0.018
2	Adaboost	1.934	0.054
3	Decision tree	2.151	0.043
4	Random forest	2.316	0.028
5	linear SVR (support vector regressor)	2.634	0.169
6	K-nearest neighbors	3.140	0.065
7	ARD (automatic relevance determination) regression	4.117	0.271
8	SGD (stochastic gradient descent)	4.227	0.175
9	SVR (support vector regressor)	4.324	0.361
10	Gaussian process	12.394	1.173

Table 4.1: Auto-sklearn result 16 features.

scribed as the primary features because they are directly involved in the calculation of fuel consumption (see (1) and (2)). Others can be described as the secondary features because they do not contribute to the fuel consumption directly. However, most of them either have great influence on the primary features or correlate to the fuel consumption in a special way. Engine speed is a hidden variable to the fuel consumption formula. It is often referred to as engine RPM (revolutions per minute) as it indicates how many times the engine's crankshaft rotates in a minute. The larger the engine RPM, the more times each piston goes up and down in a vehicle's cylinder [24]. Therefore, higher engine RPM implies both higher engine power and fuel consumption to the vehicle. Throttle position is directly controlled by the driver when the accelerator is pressed. It indicates the air intake of the engine. Especially, engine RPM is heavily dependent on the throttle position [24]. In short, the more open the throttle, the higher the engine RPM, and hence the higher fuel consumption. Intake manifold pressure indicates the intake manifold vacuum that exists in the intake manifold after the throttle. It can be used to estimate the MAF of an internal combustion vehicle [12]. Essentially, the larger the intake manifold pressure, the larger the MAF. Engine load is an indicator of the capacity of the engine to produce power. It is mostly determined by throttle position, current airflow, standard temperature and pressure, and ambient air temperature [25]. In general, a larger value of engine load represents a larger part of the engine being occupied for producing power, and therefore the fuel consumption is also higher.

Last, fuel level ended up being the only exception which theoretically does not imply any sort of correlation to the fuel consumption or other important features. It is merely an indicator of the current fuel left in the fuel tank. After taking a more thorough look into the dataset, it appears to be more of an independent feature because the correlation between the fuel level and fuel consumption is mostly random and unstable. However, it is coincidentally correlated with the fuel consumption in numerous trips. Specifically,

5 out of 22 studied trips show relatively strong correlations between fuel level and fuel consumption, 3 of them have incomplete fuel level data, and others only entail relatively weak correlation between fuel level and fuel consumption. Also, the correlation values fluctuate severely within a large range as the largest correlation value is about 213 times larger than the smallest correlation value. Given the limited size of the dataset, this is likely the reason why the permutation feature importance is misled in some occasions. Given these observations, an assumption is made that the fuel level is a random noise to this experiment which just appears by chance. Therefore, it is excluded from the selected features. Consequently, the most influential features are Speed, MAF, Engine speed, Throttle position, Intake manifold pressure, Engine load, and Fuel rate.

To further support this conclusion, one feasible method is to check whether more robust models can be built based on those selected features while keeping other parts of the experiment unchanged. In order to do that, another round of 1 hour training with the 7 selected features using the auto-sklearn with the same configuration was carried out (Table 4.2). As a result, although the RMSE of a few models slightly fluctuated around the old values, all models ended up with a better MAPE. Some of the improvements are quite obvious. Especially for the MLP, linear SVR, and Gaussian process models, the improvements in both RMSE and MAPE are significant in comparison to the old values. This observation is able to prove that the extracted features indeed have strong correlations to the target value.

Index	Regression algorithms	RMSE	MAPE
8	MLP (multi layer perceptron)	0.963	0.074
6	linear SVR (support vector regressor)	1.585	0.056
3	Extra tree	1.824	0.017
0	Adaboost	1.851	0.050
9	Random forest	2.040	0.025
2	Decision tree	2.135	0.039
5	K-nearest neighbors	3.151	0.056
10	SGD (stochastic gradient descent)	3.702	0.156
1	ARD (automatic relevance determination) regression	4.159	0.268
7	SVR (support vector regressor)	5.141	0.091
4	Gaussian process	6.771	0.845

Table 4.2: Auto-sklearn result 7 features.

Another method worth trying out in order to build more robust models is to do a more tendentious feature engineering. Namely, deriving more features based on the selected features to further increase the model performances. Given all features selected are numerical, the strategy is to add numerical derivations of the selected features such as 2-way cross products, 3-way cross products, squares, etc.. For example, suppose there are feature A, B, and C. Then A*B, B*C, C*A are the 2-way cross products and so on.

This strategy is proven to be efficient and effective in terms of encoding nonlinearity in the feature space [26]. In other words, it gives linear models the ability to also fit nonlinear aspects of the data. Hence, it can make weak models stronger. Following this strategy, a feature space consisting of 43 features is generated based on the 7 selected features. The added features are square roots, squares, cubes, and 2-way cross products of the original 7 features. On top of this new feature space, models are further tuned manually using RandomizedSearchCV [14]. In short, it is a technique that allows randomized search on a predetermined hyperparameter space. In comparison to exhausted search on the hyperparameter space, it only tries a fixed number of parameter settings that are sampled from the specified hyperparameter space. Since the number of parameter settings can be set manually, the randomized search can be customized to balance the need between search power and time efficiency. As a result, MLP stands out to be the most robust model for this specific regression task with an RMSE of 0.299 and MAPE of 0.002. In other words, the result of the MLP model entails that the average difference between the predicted value and the actual value is approximately 0.2% . This result is exceptionally good in comparison to other models.

Notably, after applying the permutation feature importance on this MLP model, the most important features end up containing mostly the primary features and their derivations. In detail, the 3 most important features are the square root of fuel rate, the square root of MAF, and the square root of speed. Besides, the only independent secondary feature existing on the list is the engine RPM. This indicates that the primary features are dominant in terms of feature importance especially when the nonlinearity of the primary features are encoded, and the engine RPM is also weighted heavily in terms of predicting fuel consumption. This makes sense according to the physical relation between the engine speed and the fuel consumption discussed above. Other secondary features appear only in the cross products with the primary features. Among these cross products, the most important combinations are speed * throttle position, speed * engine RPM, speed * engine load, and speed * intake manifold pressure. Theoretically, these combinations do not have exclusive interpretations themselves. However, the RMSE and MAPE of the MLP model have dropped approximately 66% and 287% respectively after removing those combinations consisting of secondary features. Thus, those secondary features involved in the combinations are valuable as they can be associated with the primary features to encode more hidden nonlinear aspects of the data and contribute to a better predictive model. In conclusion, although it is still unclear how much of the improvements are from the feature choice and how much of the improvements are from the feature crossing, this observation still provides useful insight that the selected features are influential in terms of predicting fuel consumption.

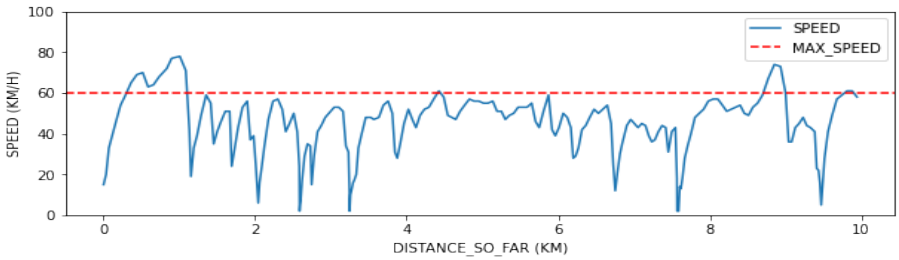

Figure 4.1: Speed classification for s4.

4.2 Instantaneous driving style analysis for a random driver s4

Next, in addition to the selected 7 features, generated features and scores(see Table 3.2) are derived accordingly and added to the feature space. The system can now start instantaneous driving performance analysis and provide driving style feedback following the guide of the Eco-driving performance score metric and the driving classification metrics. The dataset used for the analysis is only the dataset of 1 car with 19 drivers for the same route because the driving performance metric can only compare driving performances for driving indices recorded for the same trip. The studied trip is 10 kilometers long and all driving data are assumed to be collected under the same external conditions. In this experiment, a random driver's driving data (s4) is selected to be the exhibition of the analysis process, the rest driving data are prerecorded to the system to prepare for the analysis.

On the one hand, the white-box analysis keeps running while new driving data for the next driving index is fetched. Figure 4.1 - 4.5 visualize the white-box analysis for s4. On the other hand, the black-box analysis using K-means clustering also played out at the same time. Since K-means clustering is a distance-based clustering algorithm, MinMaxScaler [14] was applied to normalize the data for a better clustering result. The instantaneous driving style clustering result is shown in Figure 4.6, and a snippet of the system log for s4 at driving index 42 looks like:

```
- Stay less than 3000 RPM to save fuel.
- Your acceleration and deceleration are good.
- Your throttle performance is good.
- Your speed is safe.
- Abnormal Speed-RPM ratio detected, try shifting a gear to save fuel.
- Your current driving style is normal,
  try following other advice to make progress.
```

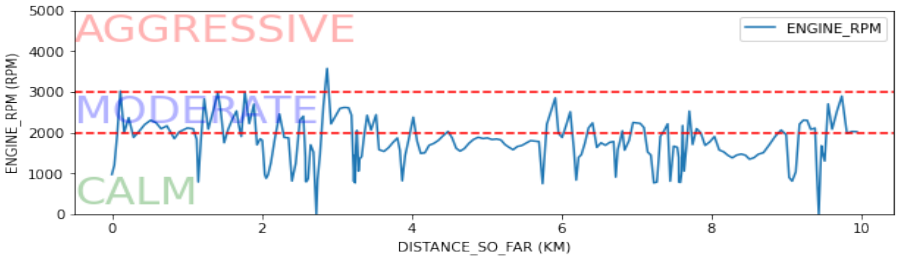

Figure 4.2: RPM classification for s4.

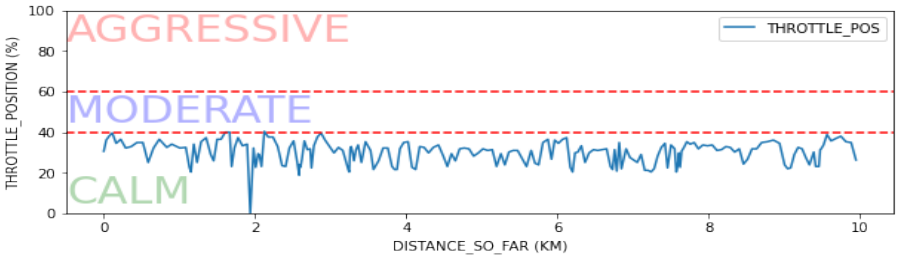

Figure 4.3: Throttle classification for s4.

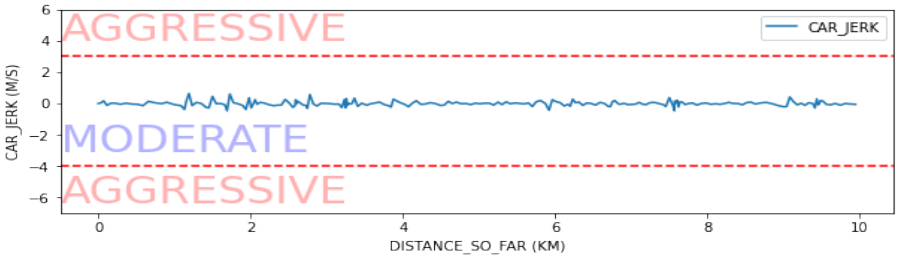

Figure 4.4: Car jerk classification for s4.

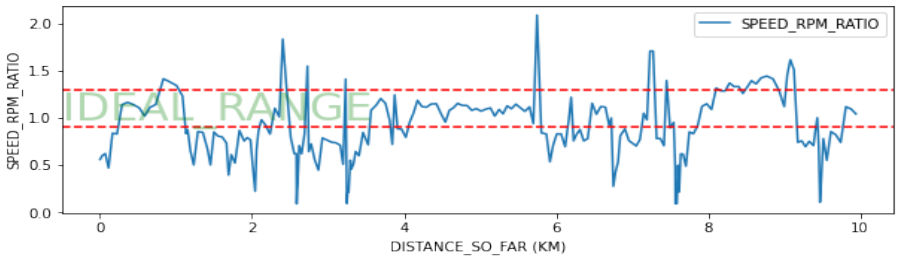

Figure 4.5: Speed-RPM classification for s4.

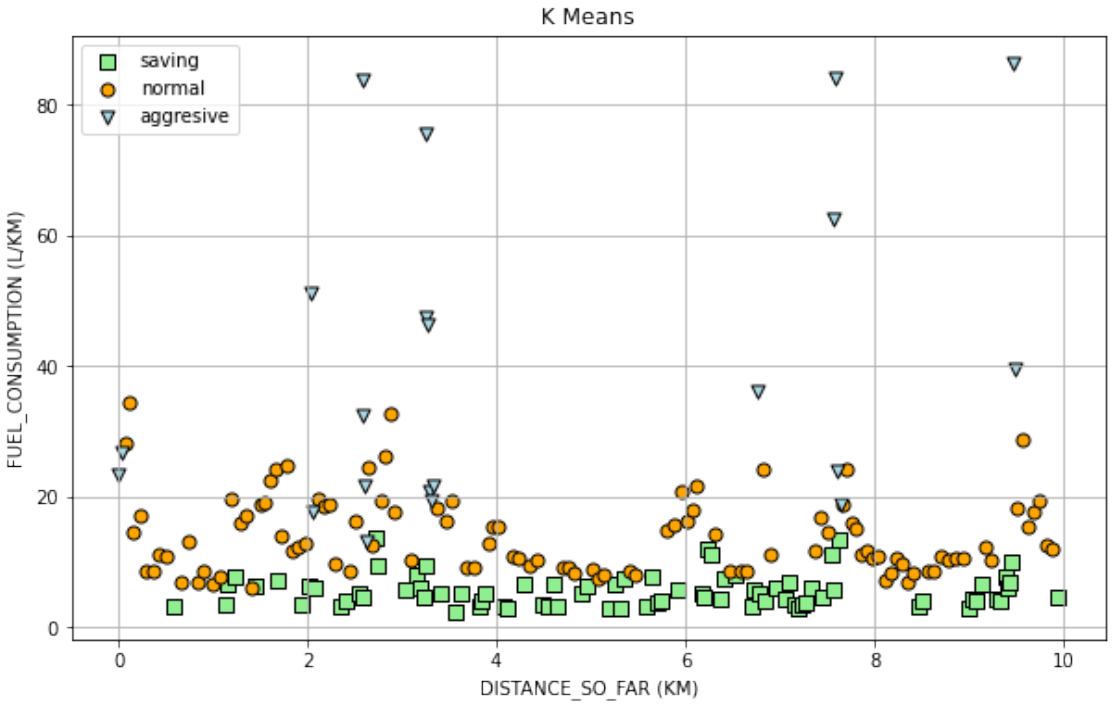

Figure 4.6: Instantaneous driving style K-means clustering for s4.

4.3 Accumulated driving style analysis for a random driver s4

As discussed in Chapter 3, the accumulated driving style analysis is carried out using the mean values of the driving data for the studied trip. The accumulated white-box analysis for s4 consists of 3 parts. Namely, the same procedure as the instantaneous white-box analysis using the driving classification metrics (Table 3.2, 3.3, 3.4, 3.5, 3.6), the general driving style clustering based on the trip Eco-driving score using the trip Eco-driving classification metric (Table 3.8), and the trip cost analysis. Meanwhile, the accumulated black-box analysis for s4 using K-means clustering is also done. As a result, the accumulated driving style clustering for all drivers based on the trip Eco-driving score is shown in Figure 4.7, the accumulated driving style clustering for all drivers based on the K-means clustering is shown in Figure 4.8, and the overall system log for s4's trip analysis looks like:

```
- Your average RPM performance is good.
- Your average acceleration and deceleration are good.
- Your average throttle performance is good.
- Your average speed is good.
- Your average gear shifting has space for improvements,
  try shifting a gear when abnormal Speed-RPM ratios are detected.
- Your average driving style based on fuel efficiency and throttle
  efficiency has space for improvements, try following other advice
```

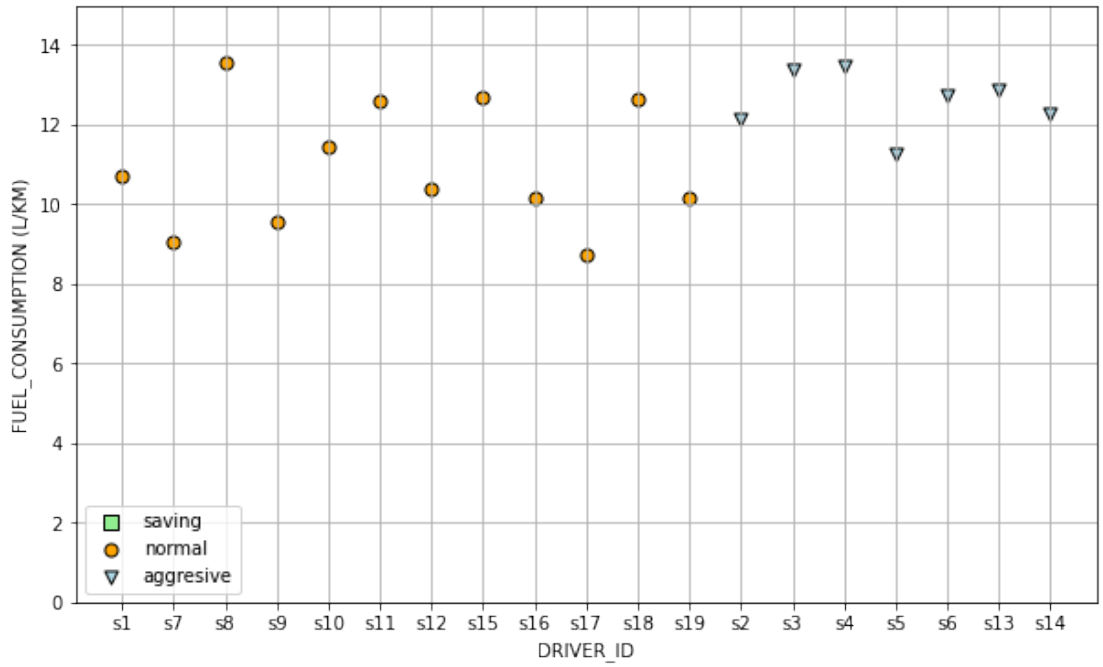

Figure 4.7: Accumulated driving style clustering based on the trip Eco-driving score for s4.

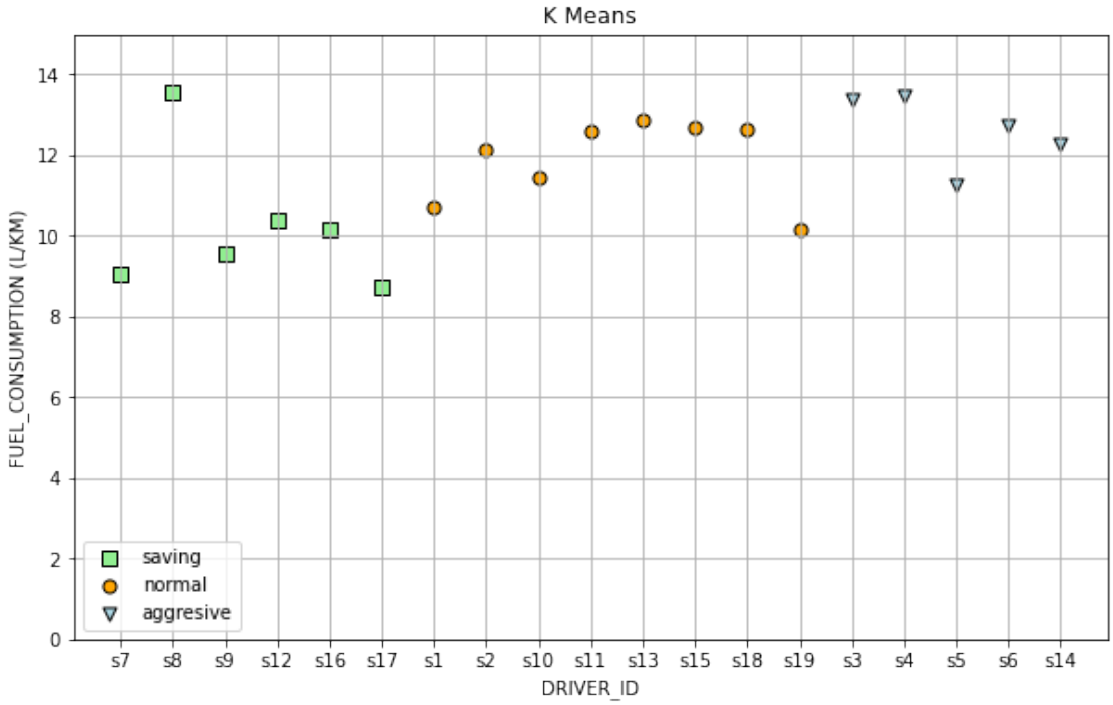

Figure 4.8: Accumulated driving style K-means clustering for s4

```
  to make progress.
- Your average driving style based on comprehensive analysis has
  space for improvements, try following other advice to make progress.
- Your trip cost is 2.39 Euros. The best trip cost is 2.07 Euros.
  Your trip spent 0.32 more Euros than the best saver.
```

5

Evaluation

After conducting the case study, the general structure and some example outputs of the system are demonstrated. Essentially, the system performs a white-box analysis and a black-box analysis to assess the driving performances on top of a comprehensive feature space which is generated based on the selected 7 most important features in terms of predicting fuel consumption. The system is able to provide both instantaneous and accumulated driving style analysis and feedback. The next problem is about how to evaluate the effectiveness of the system.

5.1 Experiment setup

This chapter evaluates the performance of the Eco-driving system in context of the case study discussed in chapter 4. Specifications for the reproducibility of the experiment can be found in the Appendix A.

5.2 The impact of the instantaneous feedback

After applying the white-box analysis on s4's driving data, the relation between the vehicle speed classification and the Eco-driving performance is shown in Figure 5.1. It is clear that the vehicle speed classification alone cannot reflect the Eco-driving performance because there is no recognizable linear correlation between the fuel consumption and the speed classification. In other words, high vehicle speed does not necessarily imply high fuel consumption. Thus, the feedback based on the vehicle speed alone is more due to the concern of safety instead of Eco-driving.

The relation between the engine RPM classification and the Eco-driving performance is shown in Figure 5.2. In comparison to the vehicle speed, the pattern between the engine speed and the fuel consumption is more obvious. However, although the majority of calm and moderate engine RPM in general tend to imply a lower fuel consumption, there

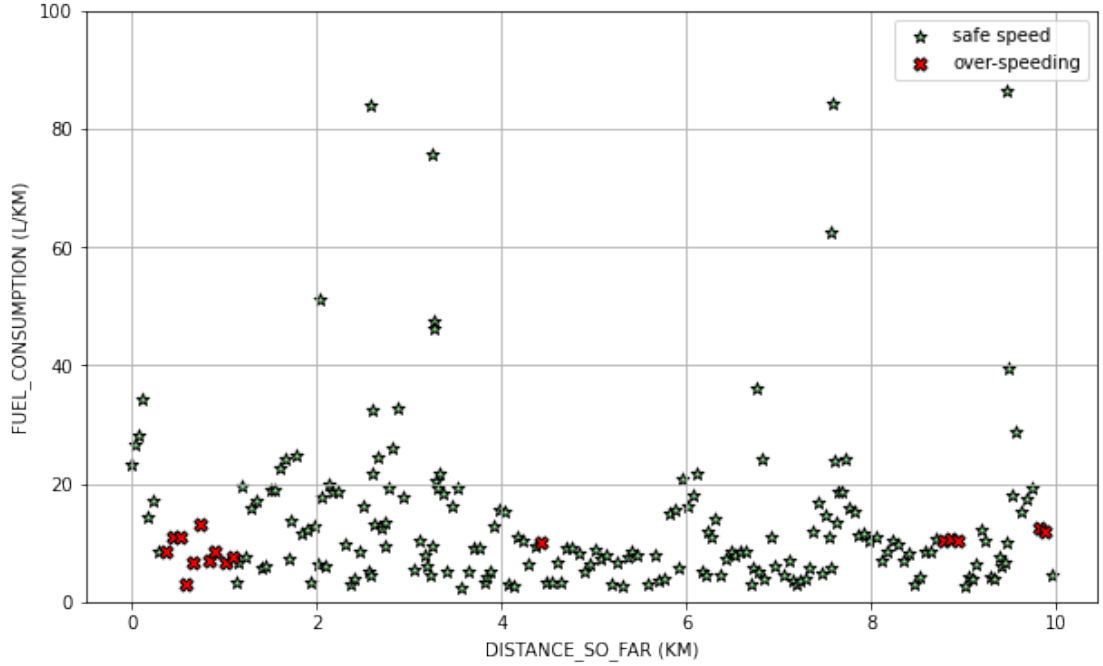

Figure 5.1: Speed analysis for s4

are numerous outliers which also indicate the possibility of exceptions. Therefore, even though the individual engine RPM classification is more instructive in terms of improving Eco-driving performances, there is a chance that it is biased when no other perspective is taken into account.

The relation between the speed-RPM ratio classification and the Eco-driving performance is shown in Figure 5.3. The result is more intuitive as the abnormal speed-RPM ratio tends to imply a higher fuel consumption in general. It provides an useful insight that the feedback based on the speed-RPM ratio classification have the potential to correctly reflect the Eco-driving performances, and following the feedback to maintain an ideal range of speed-RPM ratio can actually keep the fuel consumption at a relatively low level.

For the throttle position, given the throttle position classification for s4 is mostly calm (as shown in Fig. 4.3), throttle position classification for another driver with a more fluctuating throttle position (s14) is used to better demonstrate the impact of the throttle position classification. As shown in Figure 5.4, most of the calm throttle position indeed results in relatively lower fuel consumption. However, the difference between the moderate and aggressive throttle position is not obvious, and there are also numerous outliers that appear in the classification. In conclusion, the throttle position classification is generally instructive, but it can also get biased when looking at it individually.

The car jerk classifications for all drivers throughout the entire experiment end up having non aggressive index at all. As discussed in Table 3.1, car jerk is a variation of the

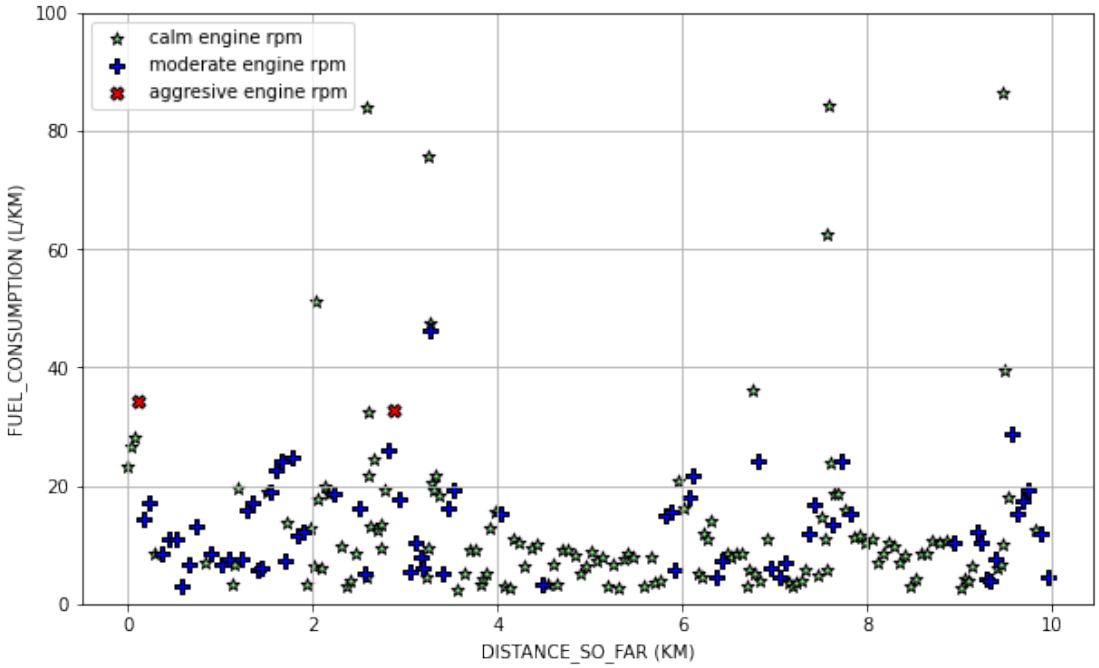

Figure 5.2: Engine RPM analysis for s4

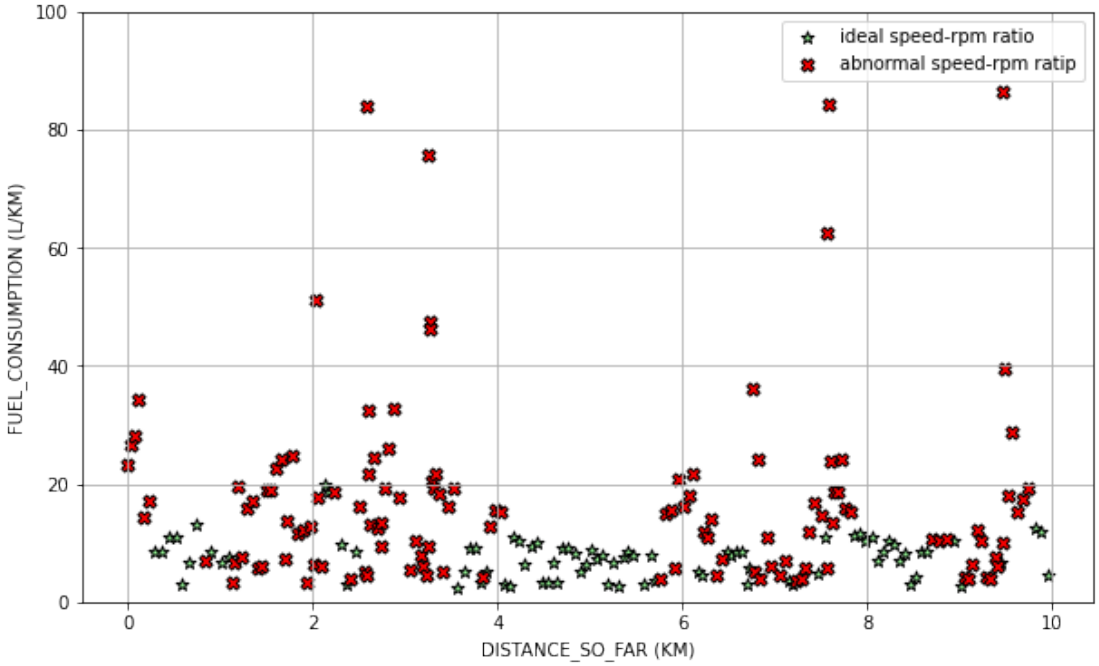

Figure 5.3: Speed-RPM ratio analysis for s4

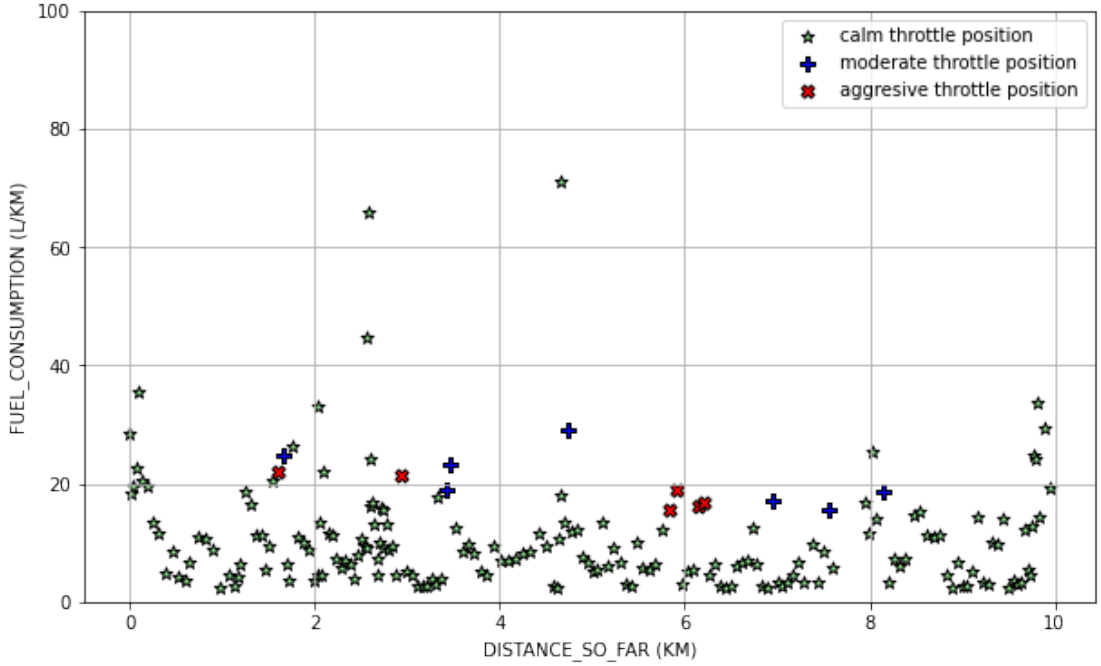

Figure 5.4: Throttle position analysis for s14

acceleration. Generally, being aggressive in the car jerk classification implies a radical acceleration or deceleration profile. Unfortunately, evaluating the impact of the car jerk classification is then inapplicable in this experiment. However, this observation is still able to provide an useful insight about the car jerk classification. Namely, the metric for the car jerk classification might be too harsh or unrealistic for the majority of driving indices. Adjustments on the threshold values might improve the metric to suit more driving indices. Nevertheless, it is not concerned in this thesis.

Besides, the black-box driving style clustering is a more comprehensive evaluation. As shown in Figure 4.6, 3 different driving styles are clearly distinguishable. It is very intuitive and reasonable as various aspects are considered simultaneously. However, it is hard to explain the process and provide specific feedback based on the black-box analysis.

In summary, all white-box classification can more or less reflect the Eco-driving performance which means following the feedback accordingly indeed have potential to reduce the fuel consumption. However, in comparison to the comprehensive black-box analysis (Fig. 4.6), the effect of the classifications are weaker and more prone to errors when looking at each classification individually. It is possible that the feedback based on any single classification rule turns out to be in contrast to the real situation. Thus, the white-box analysis and the black-box analysis should be considered together to have a more thorough and correct understanding of the driving style. The feedback based on the white-box classification rules are better to be seen as a supportive explanation to the

black-box analysis instead of independent modular feedback.

5.3 The difference between the white-box and black-box analysis in the trip Eco-driving style clustering

On the one hand, the white-box analysis is an evaluation based on a set of fixed and explicit universal metrics, so it is stable, reproducible, and interpretable. However, there is no guarantee that the metric is always correct and suitable for a variety of driving conditions. On the other hand, the black-box analysis is an evaluation based on dynamic driving data and prerecorded histories. It is more of a relative measure of driving styles. Depending on the amount of prerecorded data, the black-box analysis has the potential to produce more accurate and realistic driving style prediction than the white-box analysis. Nevertheless, the black-box analysis is poor in terms of interpretability, and it requires a fair amount of pre-recorded data to function properly which can be considered as a drawback to the applicability of the system.

As shown in Figure 4.7 and 4.8, the results of the trip driving style clustering for the white-box and black-box analysis are very different. In addition, Fowlkes–Mallows index [27][14] is introduced to evaluate the similarity between the two clustering results. In short, Fowlkes-Mallows index calculates the geometric mean between precision and recall. It returns a score which indicates the similarity between two clusters. The score is ranged from 0 to 1, with higher values being better. As a result, the Fowlkes-Mallows score for the two clusters is 0.604. It shows that there is a great divergence between the white-box and black-box analysis result.

Notably, the white-box analysis looks more rigid than the black-box analysis because no driver is clustered as the saving driver even for the driver with the least average fuel consumption. It is possible that there is actually no saving driver at all given only 19 drivers are involved in this experiment. However, it is also possible that the poor Eco-driving scores are due to the unknown external factors. For example, the traffic, the weather, the car model, etc.. Since the white-box metric only takes into account the fuel efficiency and the throttle efficiency, it is possible that the result of the white-box analysis is prejudiced, and therefore it might deviate from the real situation. In contrast, the black-box analysis looks more natural as all driving styles have drivers being assigned to and the relation between the driving style and the average fuel consumption is intuitive. Namely, a driver with a high average fuel consumption is more likely to be ranked as an aggressive driver and vice versa. Nevertheless, there is also a chance that all samples in the experiment have relatively poor Eco-driving performances, so even the most saving driver in the experiment actually drives non Eco-friendly. Then, the result of the clustering can also be biased given the samples are biased at first.

In summary, both white-box and black-box analysis for the trip Eco-driving style analysis have their unique advantages and disadvantages. Neither of them can totally disapprove

of the other one. They should be considered from two dimensions and be used to support or contrast each other to make a more comprehensive conclusion. In order to determine which one actually reflects the real situation when two analysis are in contrast to each other, more details need to be checked accordingly. All in all, the white-box analysis is always a good universal baseline to check, and when the amount of prerecorded driving data is abundant, the result of the black-box analysis can be a more valuable reference.

5.4 Negative results for the experiment

As shown in Figure 4.8, the driving style assigned to driver s8 seems to be an outlier in the trip driving style clustering. For some unknown reasons s8 is tagged as a saving driver even with the largest average fuel consumption. It is hard to track the cause for this exception given the complexity of the feature space and the nature of the black-box analysis methodology. Although there is only 1 significant error appearing in the experiment, it indicates that the result of the comprehensive black-box analysis under the current experiment setting can also get biased.

In both instantaneous and accumulated analysis, occasionally there exists identical feedback for drivers with two different driving styles. For example, a driving index which is clustered as saving might have the exact same feedback as another driving index which is clustered as aggressive. This problem occurs for two possible reasons. One is that because the white-box feedback rules are coarsely defined, minor differences in driving behaviors are overlooked in some cases. Another one is that the white-box feedback may accidentally mismatch the corresponding black-box analysis either because the black-box analysis involves more features than the white-box analysis does or simply because the black-box algorithm works differently in certain cases. After all, the black-box analysis by nature is hard to justify, and this could be a major dilemma to further refine the system.

5.5 Limitations to validity

The Eco-driving performance score metric (Fig. 3.1) plays a very important role in this thesis. However, the most obvious drawback of this approach is the requirement of a large amount of prerecorded driving data. One possible solution to this problem is to preset default scores for different driving conditions, so it is possible to warm-start the approach even when there is no prerecorded driving data. However, determining the default scores can be intricate, and it requires more external data such as the road type to be provided. Also, the frequency for windowing the trip and resetting the evaluation procedure matters because a trip can last long and may include different driving environments. For example, a trip may start on a city road, go through a segment of highway, and end on a city road again. In this case, the best Eco-driving score is likely to be generated during the highway driving. However, comparing a score generated on highways to scores generated on

city roads makes less sense in terms of relative driving style analysis. Ideally, windows for all different driving environments need to be specified clearly, so the relative driving performance score is always generated in comparison to the best driving performance recorded under the same circumstances. Consequently, the more thorough the specification, the more accurate the evaluation. Nevertheless, as mentioned in Chapter 2, this thesis focuses only on the OBD-II data and part of the external conditions are assumed to be implicitly reflected in the collected driving data. Hence, although the studied trip in this thesis is only 10 kilometers long, the implementation of this approach for this thesis may inevitably contain certain bias and limitations. Moreover, it is important to preprocess the data with extra care because this approach relies predominantly on the largest value, it is vulnerable to outliers with extreme large values. Also, regardless of the amount of prerecorded data, it is always possible that the best trip recorded for the same car and route is still poor in terms of Eco-driving. In that case, the returned scores will then be skewed as well. Nonetheless, this situation is less likely to happen when the amount of prerecorded driving data gets larger and more diverse.

Moreover, the k-means clustering in the black-box analysis requires the number of clusters to be determined prior to the clustering. This means that the predetermined 3 driving styles are independent of the actual driving data. This can be inappropriate as the data might imply more detailed or more coarse driving style clustering. Either case, the accuracy of the black-box analysis is limited.

Finally and most importantly, the dataset used in this experiment is small. There are only 19 drivers and approximately 3500 driving indices involved which is a very restrained size for a data-driven system like this. Ideally, a larger dataset is desired in order to build a more robust system. Nonetheless, there also exists a possibility that drivers might be able to evaluate his/her progress towards Eco-driving using his/her driving data overtime. That is, for drivers who tend to maintain a regular routine on a certain route, driving style analysis against the driving history for that route can be done to learn the difference in the driver's driving performances overtime. In this case, the Eco-driving performance score metric is more self-sustaining, and the driving style analysis works more effectively overtime even when there are only limited prerecorded driving data in the system. All in all, the more data there is, the more accurate the Eco-driving performance score metric and the driving style analysis will be.

5.6 Summary

In conclusion, the Eco-driving system is capable of generating both instantaneous and accumulated driving style feedback. The design integrates a white-box and a black-box analysis in order to satisfy the need for both the robustness and the interpretability of the system. Specifically, the black-box analysis on the one hand is a comprehensive evaluation which takes into account a complex feature space and can generate a relative

data-based estimate of the driving style. On the other hand, the white-box analysis works correspondingly as a supportive explanation to the black-box analysis. There are certain errors and limitations in the system. However, it prototypes a potential Eco-driving assistant which can be adopted widely to improve the Eco-driving practice and eventually save more fuel.

6

Related Work

A general methodology to predict the ideal fuel consumption for vehicles using a series of polynomial fuel consumption models are provided by Saerens et al. [28]. They offer a possible global approach to assess driving style for drivers. Namely, the polynomial fuel consumption models can take in certain engine data such as engine speed and engine load to generate an estimation of the minimum fuel consumption accordingly. The estimation can then be used against the driver's actual fuel consumption to evaluate drivers' driving style. Notably, even though different models have unique traits as they focus on different aspects of the engine data, the general applicability of this approach is good because all data required for the fuel consumption models are accessible via OBD-II. However, not all models are robust in terms of predictive power. For example, the load-based and speed-based models are coarser than the power-based and torque-based model. Unfortunately, this approach does not suit the experiment in this thesis because the dataset for the thesis does not contain engine power and engine torque which are the required parameters for those more accurate models. In addition, two load-based and speed-based models are tried out, but the results turn out to be problematic because the minimum fuel consumption estimations are mostly higher than the driver's actual fuel consumption. In conclusion, the polynomial fuel consumption models have great potential in terms of predicting minimum or ideal fuel consumption based on only a few easily accessible OBD-II data. Although it does not suit the current dataset, it provides useful insights about engine data that are valuable in terms of influencing fuel consumption.

An assessment of driving behavior on instantaneous fuel consumption is carried out by Meseguer et al. [12]. In short, it incorporates an Android application to automatically generate a coefficient for driving style. The coefficient is ranged from 0 to 1, and driving styles can be classified to quiet, normal, and aggressive according to the coefficient. In order to evaluate the application, it introduces a series of formulas to calculate vehicles' instantaneous fuel consumption and CO_2 emission using OBD-II data and other constants. Even though it does not contain theoretical details about the driving style application, it offers a

general idea about how to quantify Eco-driving performances and classify driving styles. Also, given the formula it provides, crucial relationships between certain OBD-II data and fuel consumption is revealed. In comparison to this thesis, it also evaluates the driving style classification on CO_2 emission which is also a good angle to study when considering the idea of Eco-driving. In summary, it offers heuristic guidelines on how to evaluate driving style and claims that more aggressive driving style is indeed more likely to result in higher fuel consumption and CO_2 emission.

In terms of profiling Eco-driving behaviors and providing feedback accordingly, the work from Massoud et al. [20] established a set of well-defined rules to specifically react to various driving indices on different aspects. Moreover, explicit feedback regarding different driving data is given. In fact, the design of the white-box analysis in this thesis is primarily inspired by those rules. Due to the multi-dimensional consideration of the driving indices, modular driving style analysis is possible, and the feedback becomes more comprehensive and explicit. Besides, the concept of Eco-driving score which is an indicator of the general Eco-driving performance is defined. The idea behind the Eco-driving score is to measure the trade-off between the fuel efficiency and throttle efficiency, and it turns out to be very intuitive and robust in terms of representing the general driving style. Nevertheless, according to the evaluation of the experiment in this thesis, fixed universal Eco-driving metrics might be rigid and result in unexpected errors. Thus, this thesis combines a black-box analysis in order to make the analysis result more realistic and flexible.

Multiple studies on integrating machine learning techniques with the Eco-driving analysis are available. Chen et al. [21] conduct a driving behavior analysis using OBD-II data and Adaboost algorithm. The key concept of their work is to use an Adaboost classifier to label driving indices as normal or abnormal based on a set of normal driving rules. In comparison to this thesis, it focuses predominantly on the general driving condition instead of fuel consumption. Thus, it does not reflect much about Eco-driving practice. However, although the proposed classification is relatively coarse as only 2 classes are specified, they sufficiently motivate features that are relevant in predicting the normal driving condition. In detail, it discusses the relation between vehicle speed and engine speed thoroughly, and proposes an ideal range for the relative ratio of the vehicle speed and engine speed in the same gear. As a result, this observation is exceptionally useful in terms of evaluating gear efficiency in the field of Eco-driving analysis. Yen et al. [29] combines OBD-II data and recurrent neural networks to predict vehicles' fuel consumption on different road types. Consequently, it breaks down and explains the relations between the influential features in the driving data to the fuel consumption and designs a prototype of Eco-driving assistant based on a deep learning model which has an accuracy over 96% in terms of predicting fuel consumption. In comparison to this thesis, the Eco-driving assistant of their work focuses only on real-time reflection of various driving events in correspondence to the fuel consumption, whereas the Eco-driving assistant in this thesis provides both real-time and accumulated driving style clustering and textual Eco-driving

feedback. Last, a driver clustering system using hierarchical clustering algorithms and OBD-II data is designed by Zardosht et al [30]. In conclusion, two clusters of drivers with different driving styles are introduced based on their hierarchical clustering analysis. The main difference between the clustering from their work and this thesis is the way to determine the number of existing clusters. Specifically, they evaluate the distances between data points and decide the number of clusters based on the differences between major clusters of points, whereas this thesis has predetermined the number of clusters prior to the clustering process. It is arguable which approach makes more sense in general. As a matter of fact, both of them have unique advantages in different use cases. Hence, the ideal choice for clustering approach may vary based on the use case scenario.

7

Conclusion

In conclusion, this thesis designs and prototypes an Eco-driving assistant which is capable of analyzing driving performances and providing both instantaneous and accumulated Eco-driving feedback using OBD-II data and machine learning techniques. Initially, 7 influential features to the fuel consumption are extracted, and a complex Eco-driving feature space (Table 3.1) is derived on top of those influential features. Then, a set of analysis consisting of a white-box analysis and a black-box analysis are carried out simultaneously based on the Eco-driving feature space. On the one hand, the white-box analysis evaluates various crucial driving performances and provides respective feedback based on a set of explicit classification metrics. On the other hand, the black-box analysis implements the K-means clustering algorithm to cluster driving styles, and feedback regarding the clustering result is provided following the classification metric with driving style clustering. The overall workflow for the driving style analysis is illustrated in Figure 7.1. As a result, the system has great potential in reducing fuel consumption and enhancing drivers' Eco-driving practice.

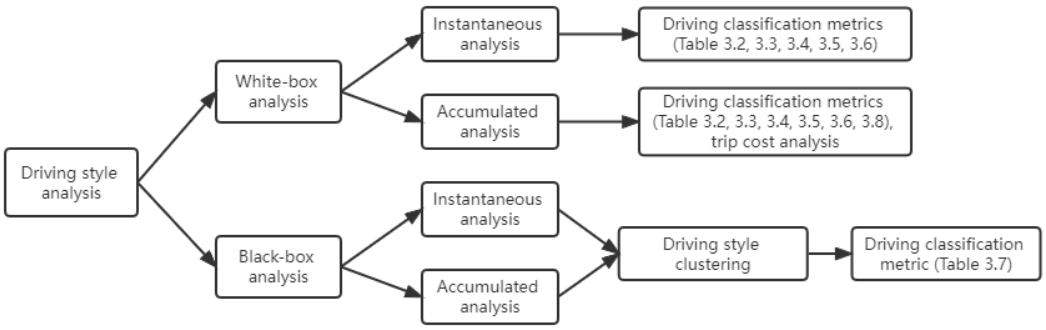

Figure 7.1: Workflow for the driving style analysis

7.1 Answers to the research questions

Regarding the question of how to define Eco-driving performances in a measurable manner, a relative Eco-driving performance score metric which evaluates the driving performance based on the best Eco-driving performance recorded before is introduced to quantify driving performances and return scores ranging from 0 to 100 to represent the level of Eco-driving. The higher the score, the better the corresponding Eco-driving performance. Consequently, this approach not only quantifies the driving performances and thus makes the comparison between driving indices more intuitive, but also ensures that the result is realistic and flexible as it is a relative measure based on history data.

Next, in order to answer the question of how to interpret and classify Eco-driving events with driving data, the thesis selected 7 most important Eco-driving features and built a complex Eco-driving feature space op top of those important features. Then, it introduced a set of classification metrics based on different focus of driving data (Table 3.2, 3.3, 3.4, 3.5, 3.6, 3.8). In short, the metrics specify the connections between specific driving events and the corresponding features in the driving data, they also provide distinguishable thresholds on the driving data to classify driving events into different driving styles. As a result, a comprehensive multi-dimensional evaluation for the driving style which reflects the connections between certain driving events and driving data can be generated according to those metrics. This process is referred to as the white-box analysis in the thesis.

Finally, in order to differentiate different driving styles, K-means clustering is implemented. The algorithm takes into account a complex Eco-driving feature space (Table 3.1), and clusters driving styles to calm, normal, and aggressive. The result of the clustering is intuitive as more aggressive driving indices tends to imply higher fuel consumption. This process is referred to as the black-box analysis in the thesis. After both white-box and black-box analysis are done, comprehensive feedback for both instantaneous and accumulated feedback on various aspects can be provided accordingly following the driving classification metrics (Table 3.2 - 3.8).

7.2 Limitation and future work

Again, one of the most significant limitations to the current system comes from the dataset used for the experiment. As discussed before, the size of the dataset is very restrained for a data-driven system design like this. Ideally, a larger dataset with more features and indices is desired. It is not only important because it may potentially improve the accuracy of the system, but also because the applicability and adaptability of the system also heavily depends on the size of training data. Currently, there are only 4 cars and 23 drivers involved in the case study, not to mention only 1 car and 19 drivers are involved in the driving style analysis. In addition, the 7 important Eco-driving features are also

selected based on the features space of the original dataset. The limitation in the original feature space might cause bias to the selection process as well. In short, it is crucial to make both the sample space and feature space of the training data as large and diverse as possible in the future to reinforce the system to work properly and fit more situations.

Another concern about the system is the current criteria for Eco-driving is completely dependent on one single target value, namely, the fuel consumption. However, there are actually more aspects to be considered in terms of Energy consumption or Carbon emission such as the CO_2 emission. Thus, future work might also evaluate Eco-driving from perspectives of other non-Eco sources.

Moreover, the experiment made an assumption that part of the external conditions are implicitly reflected in the collected driving data. Therefore, the result of the system still has a lot of space for improvements. In the future work, it is possible to take into account more external driving conditions such as the road type, the weather, the altitude, etc. to improve the performance of the system.

Finally, the current system focuses only on gasoline passenger cars. However, hybrid and electric cars are gradually taking more places in today's global vehicle market. Due to the different mechanisms in various power systems, the Eco-driving principles and practice might vary. Therefore, future work can also take these angles to further develop the system.

References

[1] IPCC Climate Change et al. **Mitigation of climate change**. *Contribution of working group III to the fifth assessment report of the intergovernmental panel on climate change*, **1454**:147, 2014. 1

[2] Hannah Ritchie and M Roser. **Fossil fuels. Ourworldindata**, 2019. 1

[3] Jack N. Barkenbus. **Eco-driving: An overlooked climate change initiative**. *Energy Policy*, **38**(2):762–769, 2010. 1, 2

[4] Christoper Spencer. **Ford Tests Show ECO-Driving Can Improve Fuel Economy by an Average of 24 Percent, 2008**, 2014. 1

[5] Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. **A design science research methodology for information systems research**. *Journal of management information systems*, **24**(3):45–77, 2007. 3

[6] Galih Hermawan and Emir Husni. **Acquisition, modeling, and evaluating method of driving behavior based on OBD-II: A literature survey**. In *IOP Conference Series: Materials Science and Engineering*, **879**, page 012030. IOP Publishing, 2020. 5

[7] Dimitrios Rimpas, Andreas Papadakis, and Maria Samarakou. **OBD-II sensor diagnostics for monitoring vehicle operation and consumption**. *Energy Reports*, **6**:55–63, 2020. 5

[8] Peter Dzhelekarski and Dimiter Alexiev. **Initializing communication to vehicle OBDII system**. *ELECTRONICS*, **5**:21, 2005. 5

[9] Martin Falch. **OBD2 Explained - A Simple Intro [2022]**, 2022. 5

[10] Motor Car. **OBD-II PIDs Codes Explained**. 5

[11] Cephas Barreto. **OBD-II datasets**, 2018. 5

[12] JAVIER E MESEGUER, CARLOS T CALAFATE, JUAN CARLOS CANO, AND PIETRO MANZONI. **Assessing the impact of driving behavior on instantaneous fuel consumption**. In *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*, pages 443–448. IEEE, 2015. 5, 6, 18, 34

[13] F. PEDREGOSA, G. VAROQUAUX, A. GRAMFORT, V. MICHEL, B. THIRION, O. GRISEL, M. BLONDEL, P. PRETTENHOFER, R. WEISS, V. DUBOURG, J. VANDERPLAS, A. PASSOS, D. COURNAPEAU, M. BRUCHER, M. PERROT, AND E. DUCHESNAY. **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research*, **12**:2825–2830, 2011. 6

[14] LARS BUITINCK, GILLES LOUPPE, MATHIEU BLONDEL, FABIAN PEDREGOSA, ANDREAS MUELLER, OLIVIER GRISEL, VLAD NICULAE, PETER PRETTENHOFER, ALEXANDRE GRAMFORT, JAQUES GROBLER, ROBERT LAYTON, JAKE VANDERPLAS, ARNAUD JOLY, BRIAN HOLT, AND GAËL VAROQUAUX. **API design for machine learning software: experiences from the scikit-learn project**. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013. 6, 9, 13, 20, 21, 30

[15] SUJATA KHEDKAR, AKSHAYKUMAR OSWAL, MANJARI SETTY, AND SRINIVAS RAVI. **Driver evaluation system using mobile phone and OBD-II system**. *Int. J. Comput. Sci. Inf. Technol*, **6**(3):2738–2745, 2015. 7, 8

[16] MATTHIAS FEURER, AARON KLEIN, JOST EGGENSPERGER, KATHARINA SPRINGENBERG, MANUEL BLUM, AND FRANK HUTTER. **Efficient and Robust Automated Machine Learning**. In *Advances in Neural Information Processing Systems 28 (2015)*, pages 2962–2970, 2015. 8, 9

[17] SHRAVANKUMAR HIREGOUDAR. **Ways to evaluate regression models**, Mar 2022. 9

[18] PAUL M SWAMIDASS. *Encyclopedia of production and manufacturing management*. Springer Science & Business Media, 2000. 9

[19] LEO BREIMAN. **Random forests**. *Machine learning*, **45**(1):5–32, 2001. 9

[20] RANA MASSOUD, FRANCESCO BELLOTTI, RICCARDO BERTA, ALESSANDRO DE GLORIA, AND STEFAN POSLAD. **Eco-driving profiling and behavioral shifts using IoT vehicular sensors combined with serious games**. In *2019 IEEE Conference on Games (CoG)*, pages 1–8. IEEE, 2019. 10, 12, 13, 14, 15, 16, 35

[21] SHI-HUANG CHEN, JENG-SHYANG PAN, KAIXUAN LU, ET AL. **Driving behavior analysis based on vehicle OBD information and adaboost algorithms**. In *Proceedings of the international multiconference of engineers and computer scientists*, **1**, pages 18–20, 2015. 10, 11, 12, 15, 35

[22] RANA MASSOUD, STEFAN POSLAD, FRANCESCO BELLOTTI, RICCARDO BERTA, KAMYAR MEHRAN, AND ALESSANDRO DE GLORIA. **A fuzzy logic module to estimate a driver's fuel consumption for reality-enhanced serious games**. *International journal of serious Games*, **5**(4):45–62, 2018. 11

[23] GOOGLE DEVELOPERS. **K-means advantages and disadvantages**. 13

[24] JANE ULITSKAYA, JOE WIESENFELDER, JOE BRUZEK, JENNIFER GEIGER, AND MIKE HANLEY. **What does RPM mean in cars?**, 2022. 18

[25] SAE INTERNATIONAL. **J1979_200204: E/E diagnostic test modes – equivalent to ISO/DIS 15031-5:April 30, 2002 - SAE international**, 2014. 18

[26] GOOGLE DEVELOPERS. **Feature Crosses: Encoding Nonlinearity**. 20

[27] EDWARD B FOWLKES AND COLIN L MALLOWS. **A method for comparing two hierarchical clusterings**. *Journal of the American statistical association*, **78**(383):553–569, 1983. 30

[28] BART SAERENS, HESHAM RAKHA, KYOUNGHO AHN, AND ERIC VAN DEN BULCK. **Assessment of alternative polynomial fuel consumption models for use in intelligent transportation systems applications**. *Journal of Intelligent Transportation Systems*, **17**(4):294–303, 2013. 34

[29] MENG-HUA YEN, SHANG-LIN TIAN, YAN-TING LIN, CHENG-WEI YANG, AND CHI-CHUN CHEN. **Combining a Universal OBD-II Module with Deep Learning to Develop an Eco-Driving Analysis System**. *Applied Sciences*, **11**(10):4481, 2021. 35

[30] M ZARDOSHT, SS BEAUCHEMIN, AND MA BAUER. **Identifying driver behavior in preturning maneuvers using in-vehicle CANbus signals**. *Journal of Advanced Transportation*, **2018**, 2018. 36

Appendix A

Regarding the reproducibility for the experiment, all codes and datasets in correspondence to this thesis can be found in the public github repository via *https://github.com/Yudong-Fan/OBD-II-Eco-driving-assistant.git*.

Specifications for the files and running environment are shown below and can be found in the README of the repository.

Datasets

- exp1_14drivers_14cars_dailyRoute.csv: original data set for the experiment, acquired from the public dataset published by Cephas Barreto.

- exp2_19drivers_1car_1route.csv: same as the previous one.

- train_shuffled_16features.csv: the train set with 16 features after initial data engineering.

- train_shuffled_7features.csv: the train set with 7 features after selecting important features.

- train_shuffled_43features.csv: the train set with 43 features after feature crosses.

- test_shuffled_16features.csv: the test set with 16 features after initial data engineering.

- test_shuffled_7features.csv: the test set with 7 features after selecting important features.

- test_shuffled_43features.csv: the test set with 43 features after feature crosses.

- total_16features.csv: the total set with 16 features before spliting corresponding train and test set.

- total_7features.csv: the total set with 7 features before spliting corresponding train and test set.

- total_43features.csv: the total set with 43 features before spliting corresponding train and test set.

- case_study_19drivers.csv: dataset used for the case study mentioned in Chapter 4 of the thesis.

Notebooks

- data_preparation_xfeatures.ipynb: data wrangling and engineering to produce the corresponding dataset.

- data_preparation_case_study.ipynb: data wrangling and engineering to produce the dataset used in the case study .

- modelling_xfeatures.ipynb: build regression models according to datasets with different number of features.

- modelling_case_study.ipynb: case study for 19 drivers as mentioned in Chapter 4 of the thesis.

- automl_1h.ipynb: 1h training using automl, the dataset and regressor used can be modified accordingly

Running environment

All codes are written in python and jupyter notebook. Jupyter notebook can be installed following the official installation guide.

After successfully deployed the jupyter notebook, all notebooks can be compiled directly on python3 for a version higher than 3.9. When compiling, make sure that all documents are in the same directory. There is no requirement for the operating system, with the exception of automl_1h.ipynb.

Automl requires python3 and jupyter notebook to be deployed on a Linux operating system. In order to run automl in other operating systems, possible solutions are WSL for windows, virtual machine, docker image, etc..