A Machine Learning Approach for Modeling and Analyzing of Driver Performance in Simulated Racing

Fazilat Hojaji(✉), Adam J. Toth, and Mark J. Campbell

Esports Science Research Lab, Lero Irish Software Research Centre, University of Limerick, Limerick, Ireland
{Fazilat.Hojaji,Adam.Toth,Mark.Campbell}@ul.ie

Abstract. The emerging progress of esports lacks the approaches for ensuring high-quality analytics and training in professional and amateur esports teams. In this paper, we demonstrated the application of Artificial Intelligence (AI) and Machine Learning (ML) approach in the esports domain, particularly in simulated racing. To achieve this, we gathered a variety of feature-rich telemetry data from several web sources that was captured through MoTec telemetry software and the ACC simulated racing game. We performed a number of analyses using ML algorithms to classify the laps into the performance levels, evaluating driving behaviors along these performance levels, and finally defined a prediction model highlighting the channels/features that have significant impact on the driver performance. To identify the optimal feature set, three feature selection algorithms, i.e., the Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost) and Random Forest (RF) have been applied where out of 84 features, a subset of 10 features has been selected as the best feature subset. For the classification, XGBoost outperformed RF and SVM with the highest accuracy score among the other evaluated models. The study highlights the promising use of AI to categorize sim racers according to their technical-tactical behaviour, enhancing sim racing knowledge and know how.

Keywords: Telemetry · Sim racing · Artificial intelligence · Machine learning

1 Introduction

Esport has become more and more popular, and this trend has been going on for a while [1]. According to [2], more and more fans are anticipated to tune in to see some of the top players in the world compete in their favourite games. Over 640 million people are anticipated to watch esports worldwide by 2025. Due to such popularity, huge amounts of data on games and players is being produced. In recent years, many new data analysis techniques have been widely used for processing and analysing data to extract insights, which are of great significance for improving players' performance levels [3]. However, there is lack of tools that offer player performance feedback and suggestions for how to

L. Longo and R. O'Reilly (Eds.): AICS 2022, CCIS 1662, pp. 95–105, 2023.
https://doi.org/10.1007/978-3-031-26438-2_8

improve [4]. This leads to many new opportunities for esports research to discover what makes a gamer deserving of winning.

Within data science, artificial intelligence (AI) has become a new technique for data analysis and sports performance prediction [5, 6]. AI is a branch of computer science that simulates human intelligence processes by machines specifically computer systems [7]. Machine learning is a form of artificial intelligence that automatically enhances the performance of computer systems by identifying data patterns [8]. The benefit of AI is that it can rapidly process huge volumes of data, and data analysis techniques are continually evolving, enabling users to gain crucial information that is challenging to obtain manually [9].

The application of AI in simulated racing (sim racing), which is relevant to the research reported in this work, leads to technological enhancement in the computer-based simulator and contributes to the direct improvement of the team and sim racer performance [10]. In this case, solutions, and strategies for becoming the best and the fastest driver are of utmost importance, with various methods of data analysis and data collection tools being used for sim racers. In terms of prediction and analytics, most of the existing studies rely exclusively on the in-game data analysis [11–13]. However, using only in-game data for estimating the driver's performance is a limitation for giving the team and drivers optimal feedback. Although it can give the fundamental information on the characteristics and behaviour of the driver, a huge amount of data that can be gathered from the physical world and sensors is neglected. In sim racing, the physical and control parameters of the simulation may be tracked and saved as telemetry files [10]. It allows sim racers to gather all the information provided by the vehicle and to analyze the data captured during a race or session [14]. Insights from telemetry data lead to a better understanding of the corresponding strengths and weaknesses of the car and the drivers' behaviour and can improve their performance by accurately tuning their car setup as well as informing on driving strategies and techniques [15]. Such information can supplement logs obtained from in-game data, providing additional information for the design of predictive models.

In this study, we report on predicting the driving performance in sim racing using the telemetry data collected from Assetto Corsa Competizione (ACC) and different ML methods for data analysis. While there are a few studies dealing with the prediction of a driver performance in general [14, 16], there is still a lack of research on the evaluation of driver performance relying on sim telemetry data. The abundance of telemetry data produced by sim telemetry tools enables the execution of fundamental analysis via ML methods. However, the existing research utilized only the limited number of parameters (i.e., steering wheel, throttle, pedal and brake pedal) in their analysis and most of the telemetry in-game data are omitted. To the best of our knowledge, this is the first study that applies AI techniques to telemetry data obtained from web data sources and relies on different features of telemetry to predict the performance level of the driver in sim racing. This approach may cover the lack of sufficient training data and achieve better accuracy as opposed to existing approaches which only rely on the small amounts of data collected in lab.

2 Data and Methods

In this section, we describe the data used in this research, data pre-processing, and analysis helping predict the driver's performance.

2.1 Telemetry Data

From the time a virtual gaming session is started until it is over, data which is known as game telemetry, is generated, and drivers may use this data to analyze and understand in-game behaviour [17]. Several sim racing telemetry tools have been developed and they can log, display and analyze data from control vehicle systems. For this study, we chose MoTec i2 Standard (v1.1.2.0473, Melbourne, Victoria, Australia), as it is a professional telemetric data analysis application, well known in all kinds of actual motorsport competitions and more data is available on-line with this tool.

The basis of this work is a dataset from ArisDrives MoTec Server[1], an online repository of MoTec data with different car/track combinations, freely available for anyone to upload and download files. All telemetric data have been obtained from the ACC simulator and logged through the MoTec data analysis package. We describe the analysis of the Brands hachs track in this paper mainly because we have access to more data for this track. Besides data downloaded from ArisDrives MoTec Server, we included data from 710 GTRL Racing (https://disboard.org/), a sim racing server hosting for AC and ACC races. The data were gathered from servers prior to September 2022, the time of preparing this article. These data are totally de-identified, available to everyone, and simply retrieved from the public domain. Additionally, all General Data Protection Regulations (GDPR) requirements have been met.

2.2 Data Processing

To extract telemetry data from MoTec log files, we have used MoTec i2 Pro (V1.1.5) available on MoTec website[2]. Following the guideline to setup ACC workspace on MoTec [18], we configured the software and defined particular settings for section time, channel and row data using built-in maths and filter functions. To gain a better insight of data, we created three data files from each MoTec log file, exported as.csv files: 1) time report including sector/lap time in a tabular form, 2) channel report containing match statistics of different channels, and 3) time series data containing general descriptions of the event (e.g., venue, track name, vehicle, duration) as well as 84 columns corresponding to driver and vehicle in-game metrics. We used Python (3.9) as our programming language on Anaconda 3 (Spyder 5.2) platform for the implementation of the pre-processing and analysis steps. We have provided a brief description of these steps below.

The first data pre-processing step involved removing invalid laps (zero lap times caused by MoTec disconnection, and pit laps (i.e., in-lap and out-lap). A total of 802

[1] http://motec.ascaroth.de.

[2] https://www.motec.com.au/i2/i2downloads/.

laps remained that were subjected to additional criteria for outlier removal using the z-score normalization method [19], and those laps were temporally isolated. To determine the optimal z-score threshold, data were analysed by applying different range of values (± 1.0, ± 2.0, ± 3.0), and finally chose Z-score $= +3.0$ as we observed better results with such value. Note that we did not eliminate laps with Z-scores lower than -3.0 because those laps represent the very fast laps. After removing the outliers, 782 laps remined for the further analysis. In addition, we made some general descriptive analysis to find the distribution of data and to identify trends in data. Moreover, graphical and correlation analyses were performed to find highly co-related features.

3 Results

Resulting from the pre-processing step, 782 laps were used for extensive analysis. Table 1 summarizes the statistics per lap. There are some slow laps, as seen by the variance between the maximum and median (slow laps). These laps are not eliminated in the analysis that follows; instead, they are taken into account such that a certain number of slow laps is also present.

Table 1. Overall lap statistics

Number of laps	Min	Max	Mean	Std	Median
782	82.353	169.009	108.277	18.04	111.778

3.1 Performance Level Analysis

In order to identify the most important metrics affecting racing performance, first we attempted to categorize the laps into the performance levels. To do this, we used two different data sets resulting from pre-processing step, 1) laptime data; 2) channel data including laptime plus all channel data. We analysed two K-value selection algorithms, namely Elbow method [20], and Silhouette Coefficient [21] to determine the optimal number of clusters for a given track data set, then used k-means method, the most commonly used clustering algorithm in both sport science (e.g., [14, 17]) and the research outside of the context of gaming literature [22, 23]. The K-means algorithm selects the number of clusters (k) and initializes each cluster centroid in a different location within the dataset. Following initialization, centroids iteratively move and begin clustering the data based on the Euclidian distance between the data and the cluster's mean until no further movements are required and the clusters are established [24]. The results from each dataset were consistent, representing that lap time is the most important indicator in sim racing performance.

Table 2 presents the results of clustering as well as the statistics of the corresponding groups. The cluster names refer to the lap-time, i.e., the SLOW means the slow lap-time and FAST means a fast lap-time. Violin plots displaying the means and distributions of

the three clusters are shown in Fig. 1. The thick line in the centre of each plot represents the median, while the two red lines represents interquartile range. On each side of the red line is a kernel density estimation to show the distribution shape of the data. Wider areas of the violin plot represent a higher density of laps in the cluster for the given value; whilst a less population is represented by smaller sections. All groups have a normal distribution as we observe that the values of mean and median are approximately close.

Table 2. Lap time statistics for performance levels for Brands hatches track

Group	Number of laps	Mean	Std	Min	Max	Median
SLOW	91	147.685	11.398	119.0915	169.099	146.265
MIDDLE	219	117.272	6.934	107.031	132.350	116.966
FAST	475	96.580	5.157	82.353	106.990	96.099

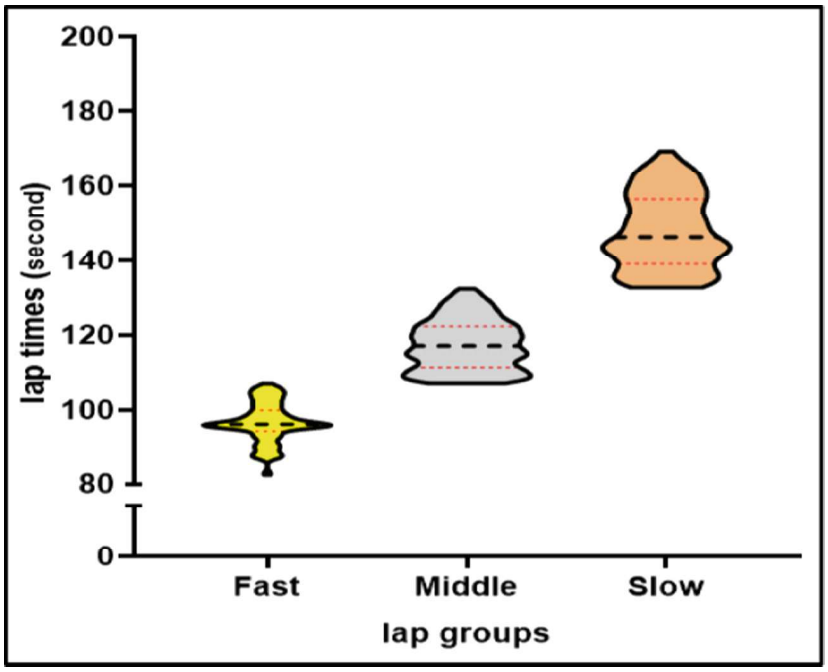

Fig. 1. Violin plots displaying the means and distributions, within each group

3.2 Feature Selection

Feature selection is the process of creating a subset from an initial feature set using ML algorithms, which removes the redundant and irrelevant features and picks the relevant features of the dataset [25]. Here, we relied on telemetry data retrieved from MoTec that contained 84 channels, including math channels (e.g., lane deviation) which we defined using built-in maths functions in MoTec. Considering the result of pre-processing step on correlation analysis, 38 channels were eliminated. The 46 remaining channels were

used for applying ML algorithms to find the most important metrics that have significant impact on the final performance. To do so, we calculated mean, median, max, standard deviation for each data channel, then conducted a bootstrapping analysis among various supervised machine learning algorithms using scikit-learn Python library. The algorithms we used are Extreme Gradient Boosting (XGBoost) [26], Support Vector Machine [27], and Random Forest [28]. These algorithms are effective for various classifications, and depending on various datasets, they each have unique attributes and performances. Figure 2 shows scatterplots depicting the accuracy of different classification methods. In comparison to other algorithms, the Extreme Gradian Boosting delivered the highest precision score among three methods in terms of mean absolute error. In addition, all algorithms got better accuracy across all included ranks (blue color in Fig. 2) compared to the results of individual rank groupings, which is reasonable given that more data yields more accurate classification outcomes.

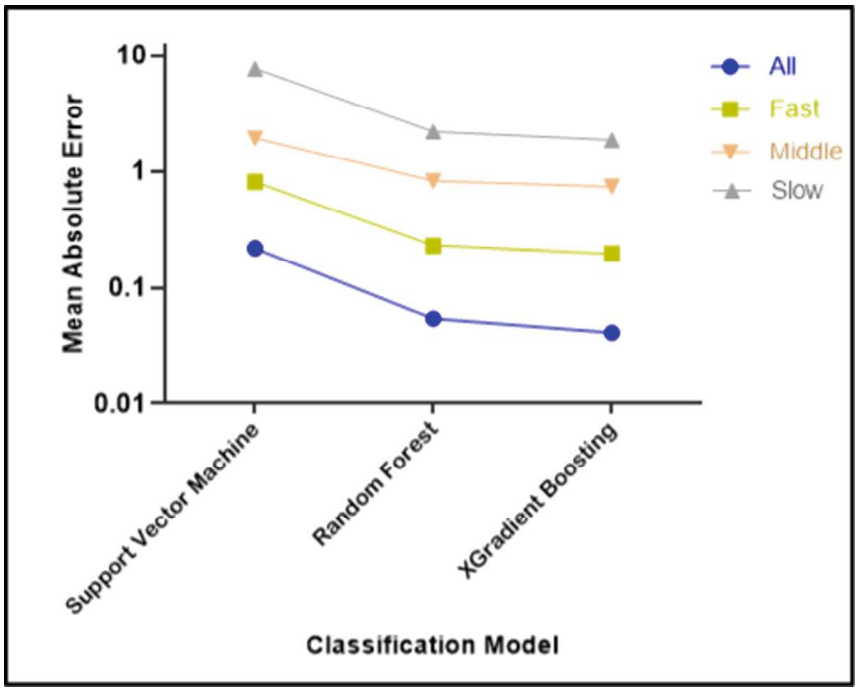

Fig. 2. Scatterplots depicting the accuracy of the classification methods for feature selection

For the classification, we divided the data into a training set (70%) and testing set (30%) in order to train the model. The model was trained using the training sets, and the accuracy of the predictions was assessed using the testing sets. The XGBoost model was able to predict the lap time with an absolute accuracy of 92.2% and an absolute error of 7.8%. A backward elimination method was used to compare classification accuracy before and after each feature was eliminated in order to assess the contribution of each feature to the classifier. The chosen features were utilized to train the classifier in phase two on the same dataset, increasing classification accuracy.

Figure 3 shows the bar graph of feature ranking for the ten most important metrics. The weights of each feature demonstrate how each feature affects the predicted lap time.

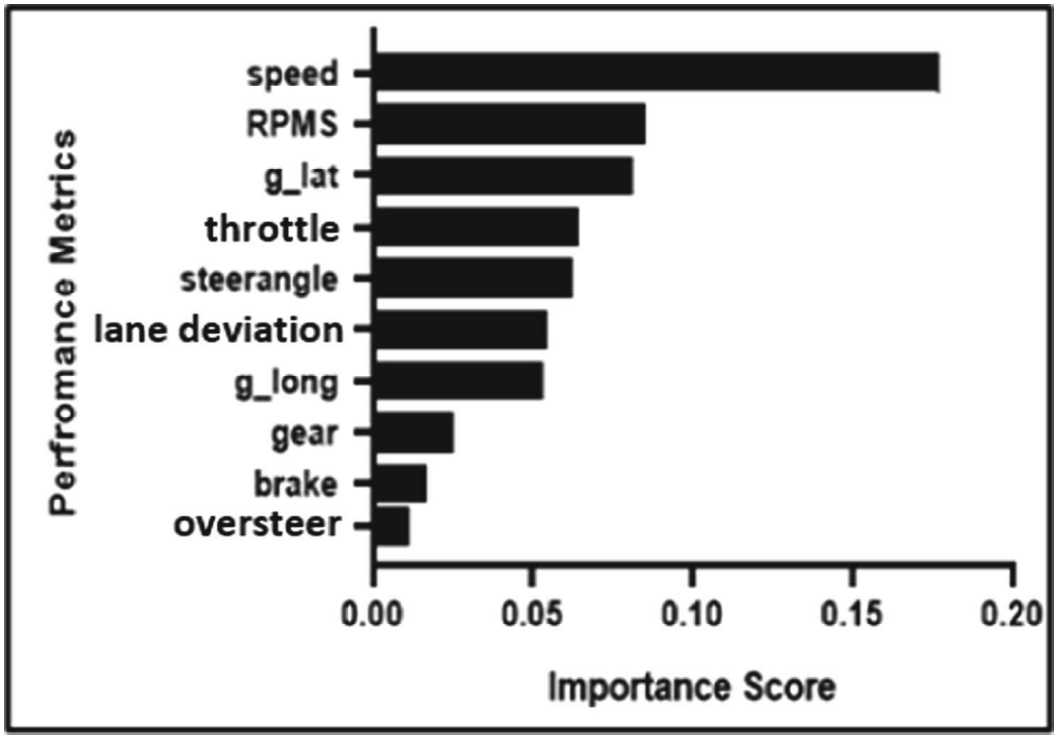

Fig. 3. Metrics found to be significant to the classification model created to predict the driver performance in sim racing

Importantly, we show the order of importance that each metric has for the prediction of lap time within our model. Here we see that all metrics that significantly were important in our classification model, are the parameters that the driver can directly control. Speed, throttle, brake and gear refer to the vehicle features. The engine RPMS describes the engine's rotations per minute and is a function of the gear used by the driver as an indication of when to change gear. Steer angle displays the angle of the steering wheel that is being input into the car at any given time. Steering error is calculation of speed, steering angle, g lateral, oversteer, brake, and throttle, indicating understeer, negative, or oversteer, positive. Lastly **g lat** and **g long** indicate the level of acceleration of the car in a specific direction, Longitudinal (forward and back) or lateral (side to side). More accurately, the higher the longitudinal g-forces are the more extreme acceleration the car has undergone which means the car has more grip when accelerating. The same can be applied through a corner using the lateral g-forces, the higher the g-force the more cornering grip the car has. From the results we observe that speed, RPMs and acceleration are the most important factors for predicting performance. These findings aid in our analysis of the different categories of drivers' driving styles. It would be also possible to focus on a specific segment rather than the data for the full lap to estimate the lap time. We defer this work as the future work.

3.3 Analysing Driving Patterns

A typical chart for analyzing driving behavior in racing is shown in Fig. 4. The figure shows the **speed**, **steering angle**, **brake pedal position**, **throttle** and **g lat** as a function of

lap distance travelled by combining the fastest and slowest laps. In order to make a better sense of the track, we incorporated sector lengths of the track into the telemetry data obtained from MoTec. To do so, we used the built-in MoTec Track Editor to determine the official sector and corner division. The sector names for each lap were then determined by processing the timeseries lap data for each lap. The vertical grey lines in Fig. 4 depict the sectors' boundaries. It is clear that all groups of drivers race in the same manner on straight sectors, while performing differently in corners. As we observe, the FAST drivers accelerate earlier and more quickly after each corner, with a sharp throttle and higher brake and stable steering control. We can observe how quickly the steering decreases when the throttle is increased in the fast laps. Additionally, how little turning is done while the brakes are fully applied. It is also obvious that fast drivers press the throttle earlier and stronger while releasing the brake later. A significant consistency can be shown when comparing driving behaviours with the feature rankings shown in

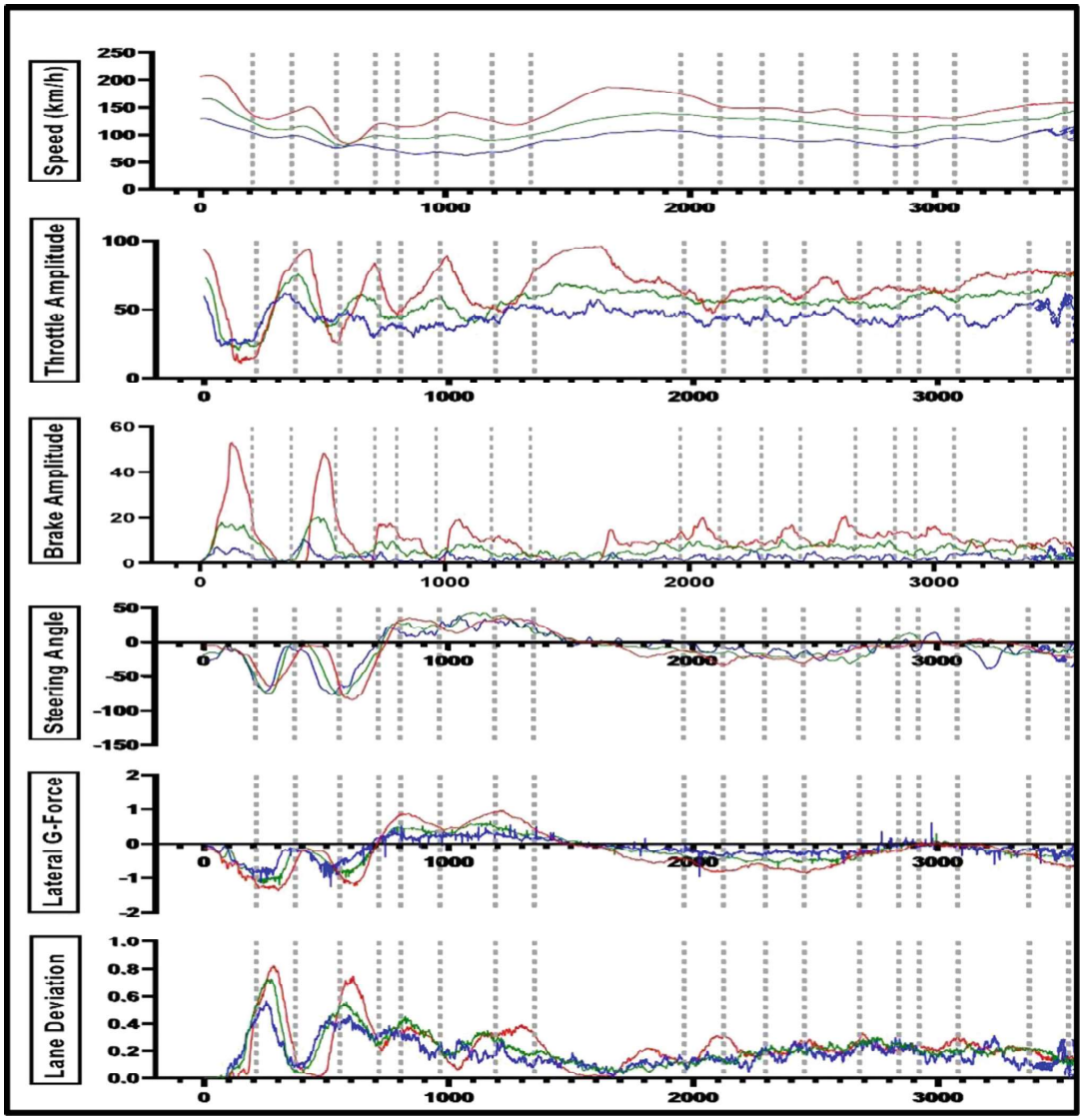

Fig. 4. Different features of driving behaviour for Fast, Slow and Middle performance group.

Fig. 4. It demonstrates that a fast driver maintains the maximum throttle for a longer period of time (producing a higher mean value) and brakes less frequently (producing a lower mean value). The brake maximum and median are greater for laps with shorter time. The similar trend is followed by the brake mean. There was no discernible trend in the acceleration characteristics (i.e., **g lat**).

A deeper investigation needs to be carried out about the connections between all metrics that define the parameters to describe driver behaviour. It would be interesting to determine the maximum turning angle, the maximum brake, the length of full braking and the gap between the first application of the brake and the release of the throttle. This kind of analysis would also be very helpful in determining corner segments. *The breaking part of a corner*, where the car must sufficiently slow down to prepare for the turn-in point; *the racing line at a corner*, which is the segment between the turn-in point; *the apex point*, which is the inside midpoint of the corner, and *the outside apex*, where the driver must gradually accelerate out of the corner, can all be identified by studying the driving styles of professional racers.

4 Conclusion

In this work, we provided an AI enabled solution for predicting sim racing performance using telemetry data. Given telemetry data from different sources, a cluster analysis was used to divide the resulting laps into three groups based on the performance (lap-time) and then XGBoosting model was used to determine the key metrics that have more impact on the driver's performance. Overall, speed, level of acceleration, the angle of the steering wheel, RPM and number of times the driver failed in vehicle control-related (steering), were all identified as important factors that impacted driving performance across all ranked laps. The findings from our analysis provides researchers with key metrics to develop more efficient training tools and techniques to improve sim racing performance.

Further research should seek to understand more deeply the analysis of the driving style to help metrics that impact lap-time. Moreover, it would be interesting to predict the lap-time by only examining the telemetry data for a specific segment rather than the data for the entire lap. For instance, it would be interesting to explore the possibility of determining whether some parts of the lap are essential for the performance across the entire lap.

References

1. Kovács, J.M., Szabó, Á.: Esport and simracing markets–the effects of COVID-19, difficulties and opportunities. Soc. Econ. **44**, 498–514 (2022)
2. Statista. Esports market revenue worldwide. https://www.statista.com/statistics/490522/global-esports-market-revenue/
3. Chu, W.C.-C., et al.: Artificial intelligence of things in sports science: weight training as an example. Computer **52**(11), 52–61 (2019)
4. Roose, K.M., Veinott, E.S.: Leveling up: using the tracer method to address training needs for esports players. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications Los Angeles (2020)

5. García-Aliaga, A., et al.: In-game behaviour analysis of football players using machine learning techniques based on player statistics. Int. J. Sports Sci. Coach. **16**(1), 148–157 (2021)
6. Mittal, H., et al.: A study on machine learning approaches for player performance and match results prediction. arXiv preprint arXiv:2108.10125 (2021)
7. Lichtenthaler, U.: Integrated Intelligence: Combining Human and Artificial Intelligence for Competitive Advantage, Plus E-Book Inside (ePub, Mobi Oder Pdf): Campus Verlag GmbH (2020)
8. Russell, S.J.: Artificial Intelligence a Modern Approach. Pearson Education, Inc. (2010)
9. Li, B., Xu, X.: Application of artificial intelligence in basketball sport. J. Educ. Health Sport **11**(7), 54–67 (2021)
10. de Frutos, S.H., Castro, M.: Assessing sim racing software for low-cost driving simulator to road geometric research. Transp. Res. Procedia **58**, 575–582 (2021)
11. Cristina de Angelo, J., et al.: Video game simulation on car driving: analysis of participants' gaze behavior and perception of usability, risk, and visual attention. Strateg. Des. Res. J. **12**(3), 312–322 (2019)
12. van Leeuwen, P.M., et al.: Differences between racing and non-racing drivers: a simulator study using eye-tracking. PLoS ONE **12**(11), e0186871 (2017)
13. Shechtman, O., et al.: Comparison of driving errors between on-the-road and simulated driving assessment: a validation study. Traffic Inj. Prev. **10**(4), 379–385 (2009)
14. Remonda, A., Veas, E., Luzhnica, G.: Comparing driving behavior of humans and autonomous driving in a professional racing simulator. PLoS ONE **16**(2), e0245320 (2021)
15. Sim Racing Telemetry. https://www.simracingtelemetry.com/
16. Bugeja, K., Spina, S., Buhagiar, F.: Telemetry-based optimisation for user training in racing simulators. In: 2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games). IEEE (2017)
17. Odierna, B.A., Silveira, I.F.: MMORPG player classification using game data mining and K-means. In: Arai, K., Bhatia, R. (eds.) FICC 2019. LNNS, vol. 69, pp. 560–579. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-12388-8_40
18. Forum, K.: MoTeC telemetry and dedicated ACC workspace. https://www.assettocorsa.net/forum/index.php?threads/motec-telemetry-and-dedicated-acc-workspace.55103/
19. Smiti, A.: A critical overview of outlier detection methods. Comput. Sci. Rev. **38**, 100306 (2020)
20. Géron, A.: Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Unsupervised Learning Techniques. O'Reilly Media, Incorporated (2019)
21. McCallum, A., Nigam, K., Ungar, L.H.: Efficient clustering of high-dimensional data sets with application to reference matching. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2000)
22. Abdullah, D., et al.: The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. Qual. Quant. **56**(3), 1283–1291 (2022)
23. Ashari, I.F., et al.: Application of data mining with the K-means clustering method and Davies Bouldin index for grouping IMDB movies. J. Appl. Inform. Comput. **6**(1), 07–15 (2022)
24. Maheshwari, A.: Data Analytics Made Accessible. Amazon Digital Services, Seattle (2014)
25. Cai, J., et al.: Feature selection in machine learning: a new perspective. Neurocomputing **300**, 70–79 (2018)
26. Chen, T., et al.: XGBoost: extreme gradient boosting. R package version 0.4-2, vol. 1, no. 4, pp. 1–4 (2015)

27. Pisner, D.A., Schnyer, D.M.: Support vector machine. In: Machine Learning, pp. 101–121. Elsevier (2020)
28. Prasad, R., et al.: Designing a multi-stage multivariate empirical mode decomposition coupled with ant colony optimization and random forest model to forecast monthly solar radiation. Appl. Energy **236**, 778–792 (2019)