MACHINE LEARNING FRAMEWORK FOR FORMULA 1 RACE WINNER AND CHAMPIONSHIP STANDINGS PREDICTOR

HORATIU SICOIE

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

302062

COMMITTEE

Dr. Wendy Powell
Dr. Paula Roncaglia

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 14, 2022

ACKNOWLEDGMENTS

Due to the nature and complexity of the project, I would like to acknowledge my principal thesis supervisor, Dr. Wendy Powell for providing me valuable advice and assisting me in setting up my Thesis goals in a clear and efficient manner. Secondly, I would like to thank all thesis supervisors who have been present during the poster presentation that have challenged myself to step out of my comfort zone while answering their questions and presenting my work. Finally, I must thank all the people to whom I have spoken about this project that have provided interesting and relevant ideas to support my planning for this Thesis submission.

MACHINE LEARNING FRAMEWORK FOR FORMULA 1 RACE WINNER AND CHAMPIONSHIP STANDINGS PREDICTOR

HORATIU SICOIE

Abstract

The proposed frameworks based on machine learning are known to produce additional valuable insights in sport analytics when it comes to performance analysis and assessment, decision-making in strategy and winner prediction, being consistently used in high-level competitions such as Formula 1.The aim of the paper aims to investigate race winner prediction by proposing supervised Machine Learning algorithms as an intelligent software solution. Furthermore, it provides a critical analysis of literature conducted in ML in relation to sport result predictions by highlighting important methodological processes(data sources, collection, implementation and evaluation). In addition, the trajectory leads to integration of new challenges such as enriching the existing historical data available with representative novel features, improving the efficiency of the models proposed in terms of accuracy by also performing cross-validation for selection and fine-tuning. Finally, the content of the paper strives to provide informative content for future research in this area and propose future extensions of the project's applications.

1 INTRODUCTION

There has been an increase of interest towards deploying intelligent software in the field of Sport Analytics based on either Deep Learning (DL) and Machine Learning (ML) techniques .These algorithms come into play in order to challenge the unpredictable nature of sport given by the numerous both external and internal factors that can shift the outcome of a single event. One particular sport which has seen numerous breakthroughs in recent years in terms of innovative technology is Formula 1, highest-class

single-seater racing competition in the world. F1 has recently partnered with Amazon Web Services (AWS) to transform the sport by providing data-driven insights and solutions.

write about AWS in report as it does data analysis in F1

The adoption of such applications has revolutionized the sport as it enabled harnessing to full potential the power of ML techniques and HPCC(High-Performance Cloud Computing). Further, it narrowed down the gap between all F1 historical data and the data collected from hundreds of sensors that a current car is equipped with.As a result, the actions have empowered not only decision-making concerning race strategy but also increased efficiency of cars in terms of aerodynamics and downforce loss. Therefore, considerable progress has been made addressing car performance indicators while also increasing the level of competitiveness, enabling racing teams to deploy in-depth competitor analysis (both car and driver analysis) to exploit weaknesses and to formulate optimal racing strategy. Early work by Jenkins and Floyd (2001) identified technological trajectories and trends in the sport at multiple levels of analysis, suggesting key improvements based on specific case studies, focusing on F1 teams for rapid development. In addition to the constant flow of telemetry data, external factors such as weather and track condition, track information, compound performance and unique race events such as unexpected *DNFs* (Did not finish due to crashes, mechanical failures or disqualification) come into play and can shift the outcome of a race event to a considerable degree.

The motivation behind this research can be addressed from two perspectives. Firstly, the scientific one aims to explain and explore methodological approaches for sport rankings prediction with emphasis on modelling to set grounds for future research in similar fields. Secondly, there is an economic incentive behind this Thesis as Machine Learning has been used widely in sport analytics to gain advantage over bookmakers but also to provide additional insights on top of the already existing domain knowledge to generate rewards for individuals practicing sports betting.

Deriving from the general idea presented above, the scope of this project is to research, formulate, deploy and evaluate supervised learning algorithms in the context of predicting race winners and provisional standings in the current season. The solutions provided will be evaluated by comparing the predicted final standings with the actual championship standings.A brief explanation of the sport's characteristics will be given as an introduction in the following paragraph.

A race-weekend is represented by an initial free practice stage and qualifying, which are followed by the race itself. The practice stage enables drivers and constructors to test the available compounds and monitor performance, while the qualifying stage is crucial for determining standings on the grid prior to the race.The drivers score points at the end of the race

based on their final position, while constructors earn points based on the points gathered by their two appointed drivers.

The project will encapsulate data collection, pre-processing and modelling processes that will be further detailed in the following sections. The main goal of the project is to predict championship standings as accurate as possible that is in this situation, correlated to a high degree with the ground truth. The algorithms selected will be critically evaluated in terms of accuracy, scalability(address the problem of time complexity in training in case of doing cross-validation, additional feature creation or parameter tuning) and robustness(new season race results must be added without affecting model performance).To conclude the introductory section, the main research question of this project is formulated as follows:

> *To what extent supervised learning techniques, with emphasis on ensemble methods and regression, predict championship standings for the 2021 F1 season based on historical data?*

The following section aims to investigate and make connections with previous scientific articles to set up the theoretical foundation of this paper.

2 RELATED WORK

There is sufficient scientific work conducted in the field related to the scope of the project allowing for the adoption of a rigorous methodological framework. Such robust framework will support the investigations performed in order to tackle the research question effectively and to a high validity degree. Bunker and Thabtah (2019) have constructed a theoretical framework for general sports winner prediction using unlabeled data, therefore deploying more sophisticated unsupervised learning techniques (Artificial Neural Networks). The performance achieved using such technique accounted for 71 %, having greater reported accuracy than domain experts.The CRISP-DM framework proposed is composed of the following: getting applicable domain knowledge (implies understanding of the characteristics of the sport of interest), data understanding and collection (decide on the granularity of the data that is to be collected), preparation and feature extraction (feature selection algorithms to establish most important feature variables), modelling (deployment of researched algorithms suitable for the type of data you are dealing with) and evaluation of performance by choosing the most suitable metrics. In addition, Haghighat, Rastegari, Nourafza, Branch, and Esfahan (2013) review data mining techniques for result prediction in various sports.The researchers are indicating what type of data needs to be collected in a classification problem for result prediction. The collection process sets the foundation

for this paper as well, indicating what type of data needs to be collected to satisfy the given goal, converting it to applied knowledge through appropriate extraction and interpretation. The research outlines the challenge for researchers in the field to investigate valid sport websites from where the data must be retrieved using appropriate web-scraping techniques together with emphasizing the importance of hybrid modeling techniques(ensemble methods) to increase the predictive accuracy.

Besides the previously mentioned scientific publications, a more suitable framework which aligns with the scope of the project is developed by Ofoghi, Zeleznikow, MacMahon, and Dwyer (2010). Their work is more closely related as they are deploying and evaluating supervised learning as opposed to unsupervised learning. The algorithm chosen is used for classifying (Naïve Bays with k=10-fold cross-validation) each instance into a predefined category labeled as final standing which has to be unique so that the final results will not overlap. Furthermore, having been applied to a similar sport in terms of structure, namely omnium cycling, lap times are also recorded and have significant variable importance in outcome prediction, similarly to F1. As a result, such proceedings are contributing significantly to predictions' accuracy based on their evaluation on feature importances. In addition, given the vast amount of historical data being publicly available, well-integrated and robust models for race winner prediction in Greyhound racing have been used to gain speculative advantage over bookmakers in a system prediction scheme designed by Schumaker (2013) built on top of SVR. It is therefore to use as a candidate for modelling this counterpart of classification algorithm as it performs well with sparse data. Furthermore, Edelman (2007) emphasizes the integration of feature maps (vector-valued transformation of the input into a feature space) for instance a Gaussian Kernel, as a "trick" for SVMs resulting in an effective stratified analysis by race in horse racing. Similarly,in Lessmann, Sung, and Johnson (2009) the concept of maximal margin separation of SVM classification to construct non-linear decision surfaces is proposed. As a result, the algorithm is built on top of an RBF (Radial Basis Function Kernel) to account for non-linearity and produces considerable better results on race winner betting returns than the linear model. The hyper-parameters C and γ are being manipulated by the means of k-fold cross-validation, splitting the data in equal partitions of size K while the classifier is recursively deployed on K-1 partitions and assessed on the remaining one. Such model will be considered and evaluated as the data presents multiple common characteristics and the parameters chosen will be the highest scoring ones in terms of correct predictions made on the training set.

Drawing connections towards the data pre-processing section of this thesis, an import lesson is given in Ofoghi, Zeleznikow, Macmahon, Re-

hula, and Dwyer (2016).The researchers discovered that conversion the time variable from the format $HH : MM : SS$ to raw seconds in order to compute differential times carry more significance and therefore emphasize dominance of racing participants in terms of performance. Such procedure will also be used in this Thesis, where the race time for the first position will be replaced by a null value while for the other drivers, their corresponding times will be replaced by the difference in time to the first place. As an incentive for this procedure, the solution provided also accounts for variations in race distances (track length) as well as environmental conditions (granularity of the weather data) that result in robustness of final estimations. Crucially, the findings provided by this approach in the previously mentioned research indicate that first places finishers have statistically significant lower times than the rest of the participants.

Recent research has been conducted in other sports for championship standings prediction after model deployment. Multiple ML solutions used in match winner prediction for NHL (National Hockey League) used as ensemble learning methods are highlighted in Gu, Foster, Shang, and Wei (2019) together with exploratory analysis of PCA (Principal Component Analysis) integration for composite rankings of players and teams have achieved great results (>90% accuracy). As important take-aways for modelling in sport analytics, SVR,RFR and GBR will be the primary algorithms that this paper aims to investigate. The decision making in choosing the appropriate model given the data is supported by early work of Moore and Lee (1994) where the use of cross-validation methods (LOOCV=Leave-one-out Cross Validation) for model selection, not only provides insights into what algorithm is able to generalize and predict for future events, but also aims to reduce over-fitting of the training set. Such techniques will be deployed in case poor results will be obtained using the classical learning approach.

As a supporting tool for model selection in order to produce better results, for instance in regression to reduce the prediction errors, gradient boosting in Drucker (1997) is used as an ensemble learning method to combine sets of weak learners in stronger ones to minimize the training error. A more recent work covering this topic produced by Bühlmann and Yu (2003) investigates the L2-loss function that can deal with both linear and non-linear patterns in the data where the MSE (Mean squared error) in regression is computed in terms of L2-boosting procedure). To integrate this approach, the ensemble method GBR will be trained and evaluated with the same loss function. An advantage of this technique is tractability, as a variable is being controlled as a smoothing/regularization parameter for non-linear data while for the linear ones, the same variable plays the role in investigation of the bias-variance trade-off.

More recent work in the field by Horvat and Job (2020) tackle the importance of feature selection prior to model deployment by providing two alternatives. Non-technical approach involves feature selection based on expert experience, deployment and evaluation against non-expert, approach which has achieved greater predictive accuracy than non-expert in studies on football game winner prediction(68.8%). Contrary, classical techniques such as PCA as a tool for dimensionality reduction or dropping one feature iteratively while observing the effect on the prediction scores have both been successfully used in previous studies in sport analytics.

This section contextualized the literature conducted on similar topics in the field while setting up the methodological procedure that this Thesis is built upon. The research papers analysed have been reviewed to provide efficient decision-making during every stage of the project's pipeline. The following section will provide additional insights into each stage, with in-depth explanation and reasoning for the implementation chosen.

3 METHOD

The first step that has been conducted was to research for historical data sources of this particular motorsport. Ergast is an open-source API(Application Programming Interface) web-based database from where users can freely extract information using specific conditions. The data is compiled from http://ergast.com/mrd/. An API allows communication between software applications and websites to support additional functionality dependence for a specific service.

The API stores historical F1 data in a database illustrated in Appendix A (page 19) for an overall idea of how the data will be merged together after all the processing is complete. Python has been used for retrieval of such information using request queries. The data has been retrieved into a dictionary data structure where the keys are represented by the attributes of each table and the dictionary values are the numerical/categorical variables retrieved based on given conditions.

To assure the relevancy of the data, queries to retrieve the data have been constructed with constraints of data retrieval starting from 2014, when a major change has happened to the sport known as the hybrid era (2014-present). Confronting literature, ML with application in sport analysis is traditionally represented by having previous seasons for training and the season of interest for testing. Upon retrieval of data that addresses information about circuits, drivers, races, results,qualifying times, driver standings and constructor standings with corresponding attribute values depicted in Appendix A, the next procedure was to enrich it with weather information about each particular race. For each race collected during the

retrieval of the queries, an URL link from Wikipedia has been appended as a unique identifier column in the data. By accessing the specific link, the Python library *BeautifulSoup* has been used to extract information about the weather during the race.The column retrieved has been further split into 6 categories: *WARM*, *COLD*,*WET*,*DRY*,*WINDY*,*CLOUDY* based on matching key-words in weather description. For example, the value of 1 was associated for *WET* if certain keywords such as 'rainy', 'wet','pouring','slippery' are mentioned in the string description retrieved, otherwise 0 . Additional features of the race tracks have also been scraped using the same method, describing the length of each track present in the dataset as well as the orientation of the racing line(clock-wise or anti-clockwise)

Data manipulation has been conducted with the use of libraries such as *Pandas* and *Numpy* to merge together all the information dispersed in tables based on a common "key", basically reverse-engineering the initial database structure after processing of each individual table component is complete. Further preliminary steps include creation of additional feature variables, such as drivers age during race which was calculated by substracting the race date from the drivers' respective date of birth(after converting both numerical values to date-time objects).Such feature creation is extremely relevant as more mature drivers correlate with experience and therefore higher rankings. Additional feature "split times" has been created as inspiration from Ofoghi et al. (2016), where the final time for drivers during race has been replaced by the difference in time to the driver on the first place position. As previously mentioned, this technique is very useful to emphasize the idea that some drivers can get lapped and therefore their race performance needs to be accounted for as being considerably poor.

In addition, given the multiple events in the sport throughout the years, the data had undergone thorough examination. Issues arrived for constructors' name and circuit names that had to be unified for all existing entries as they have changed multiple times from 2014-2021. Furthermore, the variables describing qualifying times required modification as there are numerous missing values for drivers that fail to qualify to the last stages. As a solution to this problem, the features have been replaced by variables describing the best, worst and the average qualifying time.

With the usage of pre-processing library from $Scikit-learn$, variable encoding using *OneHotEncoder* has been to tackle the problem of multiple categorical variables (Circuit Name, Driver Name, Constructor Name etc.) present in the merged data while *StandardScaler* has been used for feature scaling of variables concerning qualifying times.

After all processing as been completed, modelling has been performed using modelling modules from *Scikit − learn* library that support deployment and evaluation of machine learning algorithms. A pipeline has been built by merging together the column transformations(encoding&scaling) with the models. To improve accuracy scores for poor performing algorithms, 5-Fold Cross Validation has been conducted. All the race data prior to the current season (2014-2020) has been used for training while the data accounting for the current season has been used for testing. As previously mentioned, an additional data have been collected that represents the final standings for drivers in order to enable direct comparison with the predicted standings.

To support adequate model evaluation, the methodology proposes the challenge of converting the regression models' output into the final championship standings.A list containing tuple relations of races and locations for the current season has been constructed and iterated through, generating predicted final positions for each active driver on the grid this season. Furthermore, the final standing has been built by converting those predicted race positions into points and storing the numerical values at each iteration. The structure of the entire project has been summed up in the workflow chart in Figure 1.

Overall, the entire methodology has been presented with all the implications and limitations.The chosen methodological procedures is supported by evaluation of Richter, O'Reilly, and Delahunt (2021) where challenges at each level of analysis for a research in sport analytics using Machine Learning techniques are presented.In terms of data capturing, exploratory data analysis has been conducted to prevent erroneous retrieval. Furthermore, feature selection is conducted to prevent loss of relevant information while also normalization was used to prevent redundancy. Challenges in modelling have been avoided through balanced data retrieval and parameters have been optimized using *RandomizedSearchCV* on the pipeline. Due to the time complexity growing exponentially and becoming a serious issue to overcome, such parameter tuning technique has been used instead of *GridSearchCV*. Lastly, evaluation challenges involve correct performance measures through regression analysises as well as deploying ranking correlation measures on the predicted final driver standings. The resulting data has been therefore imported in *RStudio* for statistical analysis and interpretation of results using correlation metrics to determine to what extent the research question is satisfied.

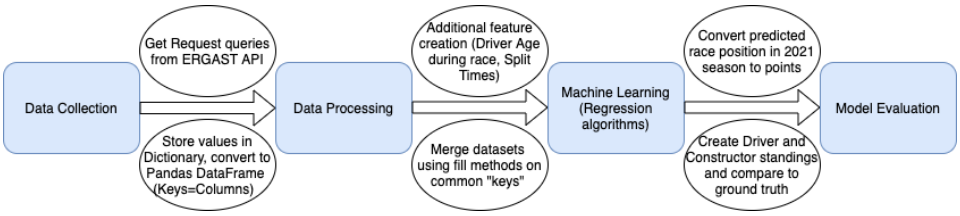


Figure 1: Project workflow with intermediate steps

4 MODELS

This section focuses on the usage of regression and ensemble learning algorithms chosen followed up with reasoning criteria and investigation of the hyper-parameters performance, choosing the ones that best suit the data after cross-validation inspection. As mentioned, the main method was to construct a principal pipeline component composed of the python object $ColumnTransformer$ which compresses numerical and categorical operations(feature scaling and encoding). The second constituent of the pipeline is the model which was firstly initialized with default parameters.The constituents have been firstly cross-validated using *RandomizedSearch* to get the best performing parameters in terms of accuracy scores.The models have been afterwards adjusted with the best estimator specification, fitted on the training set followed by predicted rankings generation for each driver for each race in the test set. Furthermore, each following sub-section aims to provide insights into the model's internal structure and reasoning for having been chosen.

4.1 *Random Forest Regressor*

The algorithm proposed is an ensemble learning solution which has performed slightly better than the others that have been chosen. The main reasons for choosing this algorithm rely on it's ability to handle noise and to deal with categorical variables present in abundance in the collected data. Moreover,the algorithm is characterized by fast run time and is not prone to over-fitting. On the other hand, it is proven to be sensitive to outliers, phenomena that occurred to a moderate degree and will be discussed in the discussion section. Intuitively, it is composed of tree-structure classifiers which are used as collective learners (Liu, Wang, & Zhang, 2012). Figure 2 depicts the performance of the regression line(in blue) for the final driver points together with the shaded area that according to *Seaborn* documentation, represents the 95% confidence interval.

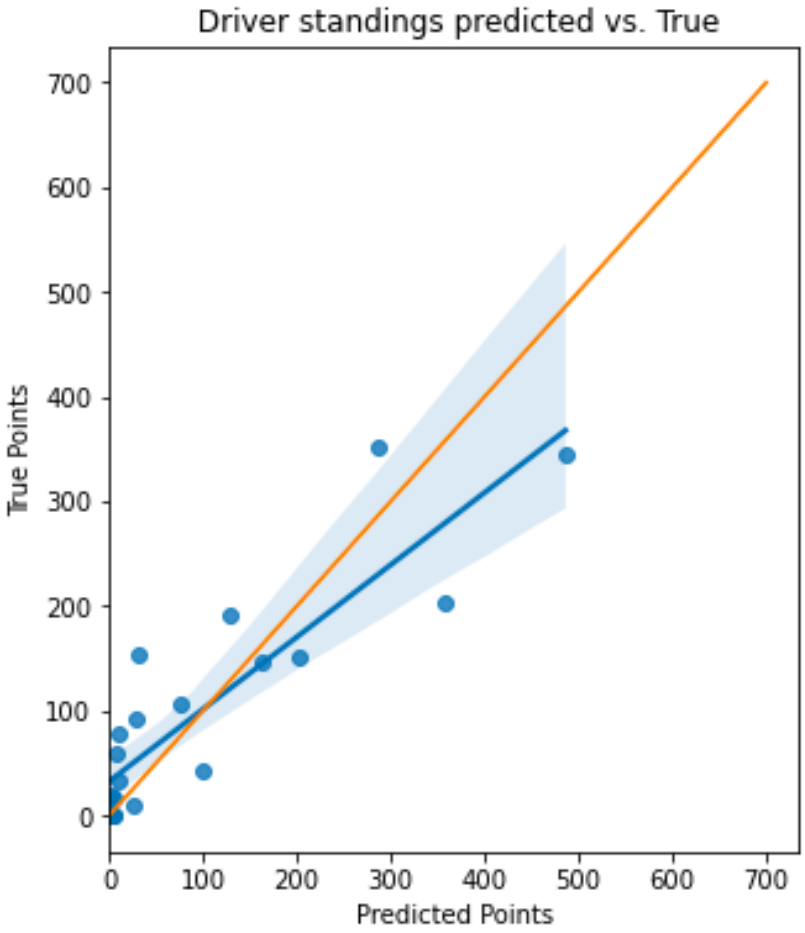


Figure 2: RFR regression line, Predicted vs True

Applied to the current scenario of regression problem, the average is collected over k of the trees h(X,θ_k) with the corresponding regression function (Equation 1).

$$\hat{Y} = E_\theta(X, \theta_k) \tag{1}$$

4.2 *Gradient Boosting Regressor*

GBR has been chosen as an extension to the ensemble solution, as it usually outperforms the random forests by optimization of the multiple parameters that it is designed upon. Friedman (2017) provides an overview of the algorithmic implementation of gradient boosting together with most commonly used loss-functions in regression problems. In addition, regularization is introduced as it allows manipulation of the meta-parameter

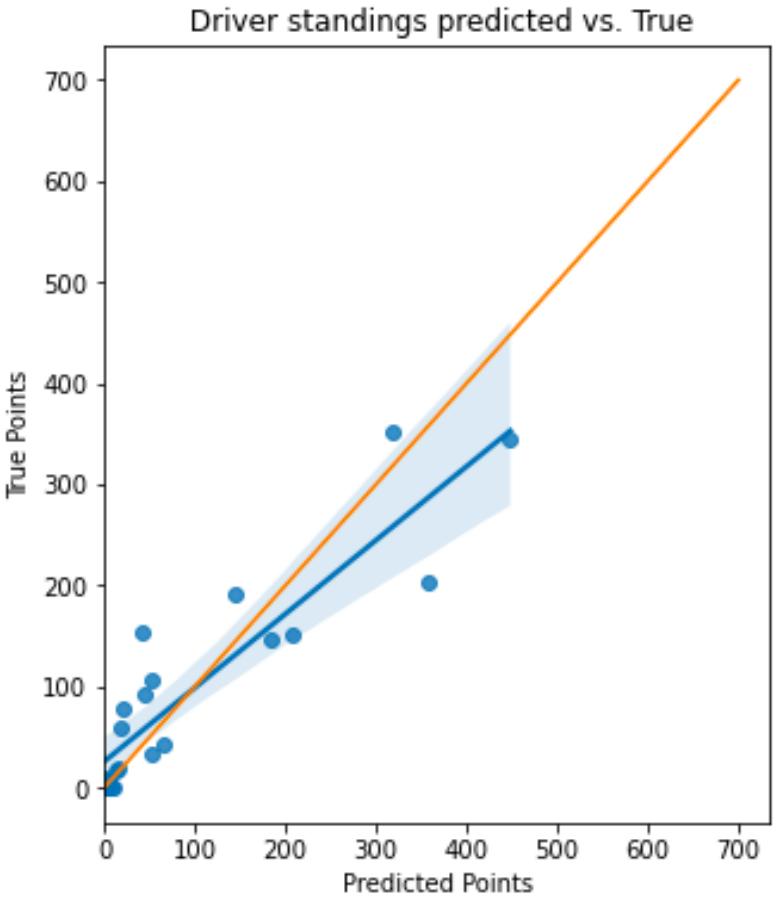


Figure 3: GBR regression line, Predicted vs True

M which accounts for the number of boosting iterations in the training process, that can be controlled through shrinkage and bagging. Figure 3 depicts model performance against true driver standings.

4.3 *Support Vector Regressor*

The proposed model used as counterpart for the classical classification algorithm was used as direction from literature. Cross-validation has indicated that similarly, the best kernel function to minimize the margin function is indeed the Gaussian Radial Basis function with formula in Equation 2. However, the in-question algorithm has performed worse than the other two, having the lowest R^2 value among the three.

$$k(x_i, x_j) = -exp(-\frac{|(x_i - x_j)|^2}{2\sigma^2}) \tag{2}$$

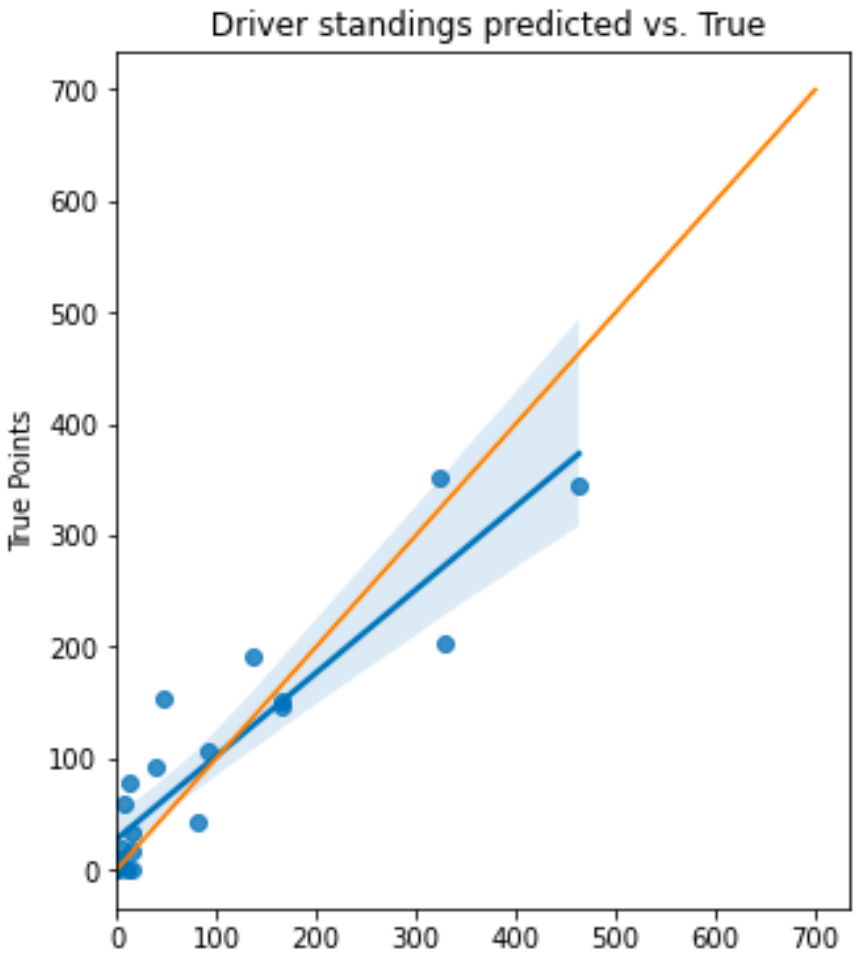


Figure 4: SVR regression line, Predicted vs True

5 RESULTS

This section gives an overview of the core findings of the paper, provides tables and charts for thorough understanding of the methodological procedure and also critical analysis of the results given by the evaluation metrics deployed in modelling performance assessment.

In terms of evaluating the rankings for each individual race in the testing season, the initial expectations of not having high accuracy has unfortunately been fulfilled, as it is intrinsically difficult to correctly determine the standings of the driver lineup for each race in the season and such implication is beyond the scope of this project. Table 1 provides scores (R^2) for the chosen models together with the fine tuning of the parameters that were computed using *RandomizedSearchCV*.As a convenient solution which tackles the research question directly, the results for all the races

have been aggregated in order to suppress the estimation errors and points corresponding to the finish positions have been generated and accumulated at each iteration.

Model	ne	mf	mss	msl	lr	kernel	C	Train Score	Test Score
RFR	834	'sqrt'	10	4	X	X	X	0.573	0.217
GBR	2000	X	X	X	0.01	X	X	0.410	0.178
SVR	X	X	X	X	X	'rbf'	1000	0.404	0.189

Table 1: Model performances with tuned parameters after cross-validation. Notation:ne=number of estimators, mf= maximum features, mss= min samples split, msl= min samples leaf, lr=learning rate

After aggregation of the results and composition of the final standings, the results have become more satisfactory. Predicted final standings have been constructed and presented in Appendix B, C and D for Random Forest, Gradient Boosting and Support Vector regressors respectively in order to facilitate more comprehensive visualisation.

Evaluation metrics is used on the predicted rankings to evaluate the degree to which the algorithms are able to fulfill the research question. To determine to what extent the modelling algorithms are able to predict championship standings, correlation metrics (Spearman's ρ more preferred alternative as the case implies ranking data) have been computed. The values indicate high correlation that compensate for the scoring in training and testing. Spearman's rank correlation was instantiated to assess the relationship between predicted position and true position.There was a positive correlation between the two variables for all 3 models, rho_r = 0.902 , rho_g=0.903 rho_s=0.883 suggesting promising results. Additional metrics such as the coefficient of determination(R^2), MSE(Mean Squared Error) and RMSE(Root Mean Squared Error) have been used to gain insight into how close the regression line is to the data points and are summarised in Table 2.

Model	Spearman ρ	Pearson r	R squared	MSE	rMSE
RFR	0.902	0.880	0.616	4163.32	64.524
GBR	0.903	0.906	0.589	3166.80	56.274
SVR	0.883	0.917	0.630	2778.27	52.709

Table 2: Evaluation scores

Additional table was constructed to measure effectiveness of predictions for a region of interest, namely the top 10 drivers. The motivation behind this is that after each race, points are awarded only to drivers who finish in this category and ideally, we would like not to be too large discrepancies

for scoring drivers. A "relaxation" margin of error is implied with varying values. From Table 3 we can observe that if an error of at least 2 places is allowed, all 3 chosen models perform better than chance while for a larger margin the results improve to .64.

MODEL	TOP 10 +-1		TOP 10 +-2		TOP 10+-3	
	True	False	True	False	True	False
RFR	0.38	0.61	0.55	0.44	0.63	0.36
GBR	0.35	0.64	0.53	0.46	0.64	0.35
SVR	0.394	0.60	0.54	0.45	0.64	0.35

Table 3: Expected correct predictions for top 10 finishers, with 1,2 and 3 margin of error in prediction.

The pipeline component of features represented by the column transformer encapsulates all the features retrieved and processed from the API together with the additional features that were created and mentioned in the methods section. The feature importance for each model has been analysed and a plot has been created in Figure 5. Further analysis of feature importance will be made in the discussion section.

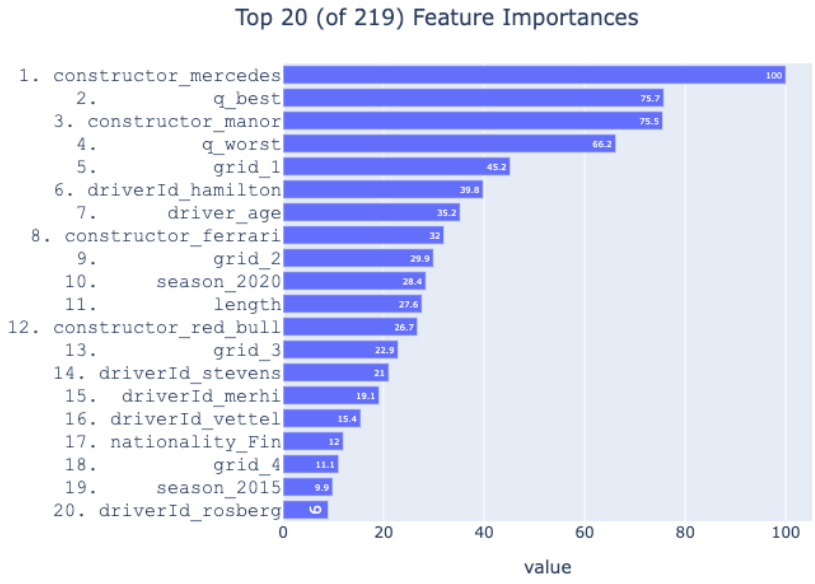


Figure 5: Feature importance for RFR

6 DISCUSSION

The discussion section will aim to conduct analysis on the approach and results. In addition, challenges imposed by the methodology will be emphasized as well as other threats to validity.

Firstly, one issue rose in the initial stages of data collection. Given the abundance of the data available on the Database API, difficult choices were made in terms of what is the most relevant data that carry most relevancy. Certain information has been left out, including information about race incidents, lap times, pit stops that would negatively impact the modelling section in terms of time constraints. The main motivation behind feature selection lies in the cited literature, where research in sport analysis traditionally implies selection through domain knowledge.

Furthermore,the pipeline component has represented a challenge in terms of coding and time complexity requirements. For each pipeline, the initial proceeding was to perform exhaustive *GridSearch* for fine tuning, however due to numerous features and parameters to be tuned during cross-validation,even with resources that involve cloud computing (*GoogleColab*) the run time surged to a considerable degree as grid search is exponential time. As a solution to satisfy the constraints, a more "superficial" approach *RandomizedCV* was used to set iteration constraints(maximum 100). A downside for this approach is intuitively missing out optimal pair-wise combinations of hyper-parameters that can increase the models' accuracy. Such procedure is going to be proposed as suggestions for future research.

After results have been analysed and feature importance table has been created, one could initially observe the driving feature of the models. Unsurprisingly,regardless of the 3 models that were chosen, the feature indicating whether the driver belongs to Mercedes is dominant. Given the fact that the previous years data(2014-2020) shows Mercedes supremacy over other constructors, the model developed a bias towards predicting Mercedes drivers(*Bottas* and *Hamilton*) to be winning almost every single race of the current season.Such critical observation can be seen from another perspective as well, the 6^{th} most important feature is *driverIDhamilton* who has been World Champion for 6 years out of the 7 included in the training set. Furthermore, a solution might arise by performing coefficient shrinkage to penalise the before-mentioned features as this season, the other constructors have caught up in terms of performance and race pace, especially *RedBullRacing*.Driver age also seems to be a relevant feature, the top 4 grid positions and the last season of the training set. The latter might be caused due to the higher overlapping of race locations but also due to the presence of all the drivers on the grid in both of the seasons.

Another important insight that is known in the sport to be an important factor in determining a race winner is the best qualifying time which directly influences the starting grid position of the drivers.This fact has been confirmed by looking at the graph.In terms of the standings analysis, the matching accuracy provided in Table 3 provide satisfactory results for the first 10 drivers. Further inspection to the middle section of the grid reveals a pattern where the algorithms failed to capture the relationship between the drivers and the constructors they are belonging to. For example, drivers with previous excellent years in the sport (*Vettel* and *Raikkonen* who are at their end of their career)are now driving for teams that do not excel in the sport(Aston Martin and Alfa Romeo respectively). In the past years, they were both drivers for championship contenders *Ferrari* which highlighted in the feature importance table, was a driving feature in prediction. For both drivers, the discrepancies were the largest of the ones observed, especially for *Vettel* who has left *Ferrari* only last year and therefore most of the data used in training captured him as a *Ferrari* driver and superior results were attributed to him.

7 CONCLUSION

In the covered research, the extent to which ensmeble learning methods and regression are able to predict championship standings in F1 has been investigated. The approach to provide a scientific answer to the research question has been supported by the literature review conducted within the field of sport analytics. Methodologies from studies with application in other race related sports such as cycling, greyhound racing, triathlon but also from general sports like NFL and Basketball of game-predicting expert systems have been thoroughly investigated and key findings were adopted.

Data collection through API retrieval and web-scraping techniques has been conducted directly from reliable data sources and decision-making in terms of data selection has been made based on domain knowledge as a traditional approach in the field. Furthermore, all information has undergone pre-processing and additional feature creation before being merged into the final version. Data modelling was achieved through the use of pipelines(end-to-end construct that facilitate flow of data from input to output of the machine learning model) and results have been generated. A process of iteration and point generation has been used at every race event in the current season, as the models were estimating the predictive rankings. Initial results were not satisfactory in terms of the coefficient of determination as a performance assessment metric. On the other hand, when the results have been aggregated along the whole season,

it has been revealed that the generated rankings are highly correlated with the ground truth. All 3 fine-tuned models have produced high ranking correlation scores (RFR ρ : 0.902, GBR ρ : 0.903, SVR ρ : 0.883) with the actual standings. Furthermore, additional evaluation of the first 10 places(drivers who are scoring points) at each race event was performed to discover that with a margin error of 3-positions, 63% of the predictions were correct while for a more strict error margin of 2-positions, greater than chance prediction(53% has been achieved for all three deployed models.

For future research on this area, the limitations of the project point towards fine-tuning algorithms using exhaustive search at the expense of time complexity. Predicted rankings for each race have room for improvement and better optimization of the algorithms. Another suggestion would be to look into Deep Learning solutions for neural networks as they have been used in mostly all of the literature reviewed and could provide new insights on top of the already chosen models. In addition, selection of additional feature such as individual lap-times, racing incidents or pit stops can play the role of a source of data enrichment as well as deploying specialised tools for dimensionality reduction and feature selection. As outlined in the results and discussion section, models have failed to account the relationship between the driver and the corresponding constructor team. A solution to investigate for this issue can be solved by additional constructor data collection.

To conclude, the proposed framework has proven to predict championship standings of the current F1 season to a considerable degree. Further work as suggested in the paragraph above as well as evaluation of novel models can result into even more robust predictions. As suggested, more granular data about driver performance in terms of lap times for a specific racing event but also about constructor performance can be introduced to support this statement.Collaborative efforts have already converged at a larger scale towards F1 predictive analytics, with data being analysed and modeled from multiple sources, not only historical data that this project encompasses, but also engineering data from factories and telemetry data from car performance indicators.

REFERENCES

Bühlmann, P., & Yu, B. (2003). Boosting with the l 2 loss: regression and classification. *Journal of the American Statistical Association*, *98*(462), 324–339.

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied computing and informatics*, *15*(1), 27–33.

Drucker, H. (1997). Improving regressors using boosting techniques. In

Icml (Vol. 97, pp. 107–115).

Edelman, D. (2007). Adapting support vector machine methods for horserace odds prediction. *Annals of Operations Research*, *151*(1), 325–336.

Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. springer open.

Gu, W., Foster, K., Shang, J., & Wei, L. (2019). A game-predicting expert system using big data and machine learning. *Expert Systems with Applications*, *130*, 293–305.

Haghighat, M., Rastegari, H., Nourafza, N., Branch, N., & Esfahan, I. (2013). A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal*, *2*(5), 7–12.

Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(5), e1380.

Jenkins, M., & Floyd, S. (2001). Trajectories in the evolution of technology: A multi-level study of competition in formula 1 racing. *Organization studies*, *22*(6), 945–969.

Lessmann, S., Sung, M.-C., & Johnson, J. E. (2009). Identifying winners of competitive events: A svm-based classification model for horserace prediction. *European Journal of Operational Research*, *196*(2), 569–577.

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. In *International conference on information computing and applications* (pp. 246–252).

Moore, A. W., & Lee, M. S. (1994). Efficient algorithms for minimizing cross validation error. In *Machine learning proceedings 1994* (pp. 190–198). Elsevier.

Ofoghi, B., Zeleznikow, J., MacMahon, C., & Dwyer, D. (2010). A machine learning approach to predicting winning patterns in track cycling omnium. In *Ifip international conference on artificial intelligence in theory and practice* (pp. 67–76).

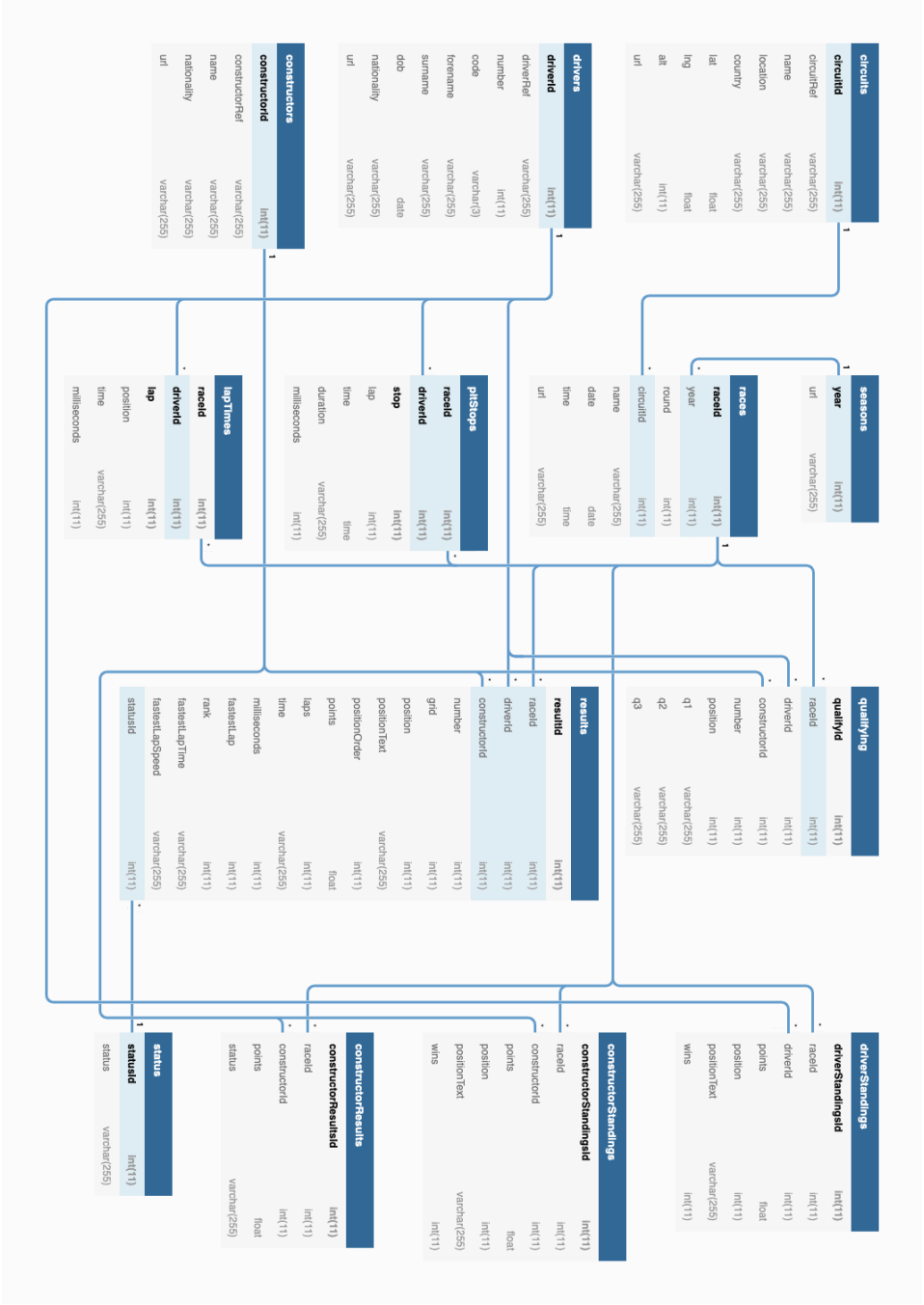
Ofoghi, B., Zeleznikow, J., Macmahon, C., Rehula, J., & Dwyer, D. B. (2016). Performance analysis and prediction in triathlon. *Journal of sports sciences*, *34*(7), 607–612.

Richter, C., O'Reilly, M., & Delahunt, E. (2021). Machine learning in sports science: challenges and opportunities. *Sports Biomechanics*, *0*(0), 1-7. Retrieved from https://doi.org/10.1080/14763141.2021.1910334 (PMID: 33874846) doi: 10.1080/14763141.2021.1910334

Schumaker, R. P. (2013). Machine learning the harness track: Crowdsourcing and varying race history. *Decision Support Systems*, *54*(3), 1370–1379.

APPENDIX A

Ergast API database structure from where queries are retrieved. The data is compiled from http://ergast.com/mrd/.



APPENDIX B

Predicted Driver Standings with Random Forest Regressor

	driver	pred_points	true_points	pred_positions	true_positions	Position error in prediction
0	hamilton	486	343.5	1	2	-1
1	bottas	356	203.0	2	3	-1
2	max_verstappen	287	351.5	3	1	2
3	leclerc	199	152.0	4	6	-2
4	sainz	157	145.5	5	7	-2
5	perez	126	190.0	6	4	2
6	vettel	106	43.0	7	12	-5
7	ricciardo	74	105.0	8	8	0
8	norris	31	153.0	9	5	4
9	gasly	30	92.0	10	9	1
10	raikkonen	21	10.0	11	16	-5
11	giovinazzi	10	1.0	12	18	-6
12	stroll	10	34.0	12	13	-1
13	alonso	9	77.0	14	10	4
14	ocon	7	60.0	15	11	4
15	russell	5	16.0	16	15	1
16	mick_schumacher	4	0.0	17	19	-2
17	tsunoda	1	20.0	18	14	4
18	mazepin	0	0.0	19	19	0
19	latifi	0	7.0	19	17	2
20	kubica	0	0.0	19	19	0

APPENDIX C

Predicted Driver Standings with Gradient Boosting Regressor

	driver	pred_points	true_points	pred_positions	true_positions	Position error in prediction
0	hamilton	448	343.5	1	2	-1
1	bottas	358	203.0	2	3	-1
2	max_verstappen	319	351.5	3	1	2
3	leclerc	208	152.0	4	6	-2
4	sainz	184	145.5	5	7	-2
5	perez	144	190.0	6	4	2
6	vettel	66	43.0	7	12	-5
7	stroll	52	34.0	8	13	-5
8	ricciardo	52	105.0	8	8	0
9	gasly	44	92.0	10	9	1
10	norris	42	153.0	11	5	6
11	alonso	19	77.0	12	10	2
12	ocon	17	60.0	13	11	2
13	tsunoda	16	20.0	14	14	0
14	russell	12	16.0	15	15	0
15	mick_schumacher	9	0.0	16	19	-3
16	giovinazzi	8	1.0	17	18	-1
17	latifi	5	7.0	18	17	1
18	mazepin	1	0.0	19	19	0
19	raikkonen	0	10.0	20	16	4
20	kubica	0	0.0	20	19	1

APPENDIX D

Predicted Driver Standings with Support Vector Regressor

	driver	pred_points	true_points	pred_positions	true_positions	Position error in prediction
0	hamilton	463	343.5	1	2	-1
1	bottas	329	203.0	2	3	-1
2	max_verstappen	324	351.5	3	1	2
3	leclerc	166	152.0	4	6	-2
4	sainz	164	145.5	5	7	-2
5	perez	135	190.0	6	4	2
6	ricciardo	91	105.0	7	8	-1
7	vettel	82	43.0	8	12	-4
8	norris	47	153.0	9	5	4
9	gasly	38	92.0	10	9	1
10	stroll	16	34.0	11	13	-2
11	russell	16	16.0	11	15	-4
12	giovinazzi	15	1.0	13	18	-5
13	alonso	13	77.0	14	10	4
14	mick_schumacher	9	0.0	15	19	-4
15	ocon	6	60.0	16	11	5
16	tsunoda	4	20.0	17	14	3
17	raikkonen	1	10.0	18	16	2
18	mazepin	0	0.0	19	19	0
19	latifi	0	7.0	19	17	2
20	kubica	0	0.0	19	19	0