Evaluation

In order to carry out a comparative analysis of our Machine Learning models, it is necessary to set up an evaluation procedure that will be identical for all the models produced. This evaluation procedure is a crucial step in defining the performance of our models. Indeed, when tackling classification problems in Data Science, several evaluation methods exist. The variety of these methods is often confusing for those who want to understand these concepts or have their say in the evaluation of projects. However, these different methods are not always adapted to the context of the study in hand. It is therefore necessary to take into account all of the elements surrounding our subject.

1. Evaluation metric

In order to measure the performance of our models in predicting podium finishes or the winner of a race, we need to determine an evaluation metric that best meets our objective. We had several choices:

* ***Accuracy:***

This is the most common evaluation metric used to assess the performance of a binary classification. The metric in question gauges the proportion of accurate predictions made by the model relative to the total number of predictions. A high accuracy score indicates that the model is making a substantial proportion of correct predictions, whereas a low accuracy score indicates that the model is making an excessive number of incorrect predictions. Mathematically, accuracy is calculated as followed:

Where TP correspond to the true positive (correct predictions), TN is the number of true negative (correct predictions), FP is the number of false positive (incorrect predictions) and FN is the number of false negative (incorrect predictions). So, the objective here is very clear, and is to get as many good predictions as possible.

* ***Precision***:

Precision is a measure of the accuracy of the positive predictions made by a model. It is defined as the ratio of true positive predictions to the total number of positive predictions (i.e., the sum of true positives and false positives). A high precision indicates that the model produces more relevant results with fewer irrelevant ones. Mathematically, precision can be expressed as:

Where TP correspond to the true positive (correct predictions and FP is the number of false positive (incorrect predictions).

* ***Recall***:

The recall metric, also known as sensitivity, is a performance measure used to evaluate the ability of a classification model to identify all relevant instances in a dataset. It is particularly important in scenarios where missing a positive instance (false negative) can be critical. Recall is defined as the ratio of true positive predictions to the total number of actual positive instances (i.e., the sum of true positives and false negatives). A high recall indicates that the model captures most of the positive instances, minimizing the number of false negatives.

Where TP correspond to the true positive (correct predictions and FN is the number of false negative (incorrect predictions).

Now that these metrics have been presented to us, we need to see which one is best suited to our problem.

Firstly, as mentioned earlier in the data understanding section, our target variable is fairly unbalanced. In fact, because of our objective of predicting the fact of being on the podium or the fact of being the winner of the race, our target variable is necessarily made up of more negative events (not being on the podium / winner) than positive events (being on the podium / winner). As a result, the use of accuracy as a metric poses a problem for reliably evaluating the performance of our models. If all the predictions of our models correspond to the majority class, then the accuracy will be high. It is therefore not viable to use accuracy as an evaluation metric here.

Secondly, precision and recall seem to be much more suitable metrics for our purposes. Recall allows us to see the capacity of our models not to wrongly predict the non-podium or the non-winning of a driver. Accuracy, on the other hand, allows us to measure the ability of our models not to predict a driver's victory or podium finish when these events have not actually occurred. In the context of our study, it seems necessary to have these two inputs in order to evaluate our models correctly. The most judicious method would therefore be to find a metric that can combine these two evaluation methods. As such, we can refer to a very specific metric to fulfill this: the F1 score.

* ***F1 score:***

The F1 score is a performance metric that combines precision and recall into a single measure, providing a balance between the two. It is particularly useful when you need to consider both false positives and false negatives and want a single metric to evaluate the performance of a classification model. The F1 score is the harmonic mean of precision and recall, and it gives a more balanced view when the data is imbalanced or when one metric is significantly more important than the other.

Therefore, in this study, we will use the F1 score as an evaluation metric in order to have an overall vision of the performance of our model. We will then be able to observe in detail the precision and recall of our models.

1. Model validation

Bias is the inability for a machine learning method to capture the true relationship for our data.

In machine learning, the difference in fits between data sets is called variance.

Once we have identified models that perform well on various evaluation metrics, it's crucial to conduct an additional step to ascertain their statistical correctness. This step involves assessing the models' ability to generalize effectively to new, unseen data beyond the training set. Through techniques such as cross-validation, where the data is partitioned into multiple subsets for training and validation, we can simulate how well the model will perform on unseen data. By rigorously testing our models on diverse datasets and employing statistical validation methods, we can confidently determine their robustness and suitability for real-world deployment. This validation process not only ensures the reliability of our models but also enhances trust and confidence in their predictive capabilities.

Furthermore, the validation process allows us to delve deeper into the bias-variance tradeoff, a critical aspect of model assessment. By analyzing the performance of our models across different datasets, we gain insights into their behavior concerning bias and variance which are respectively the inability for a machine learning method to capture the true relationship and the difference in fits between data sets. Models with a high degree of bias may demonstrate a consistent tendency to underperform across a range of datasets, indicating an oversimplified representation of the underlying patterns. Conversely, models with a high degree of variance may exhibit erratic performance, suggesting that they have been overfitted to the training data. Through a process of careful analysis and iteration, it is possible to fine-tune our models to achieve an optimal balance between bias and variance. This entails adjusting the model complexity, regularisation parameters, or feature selection methods in order to achieve the desired level of generalisation without sacrificing predictive accuracy. By navigating the bias-variance trade-off during the validation process, we ensure that our models not only perform well on the training data but also generalise effectively to new and unseen data, thus enhancing their reliability and practical utility.

Therefore, to gain insights into whether our model is overfitting, we will use learning curves, which are graphical representations illustrating the relationship between a model's performance and the amount of training data it has been exposed to. Learning curves provide a visual tool for assessing how the model's performance evolves as the training dataset size increases. By plotting performance metrics such as accuracy against the size of the training dataset, we can observe patterns that indicate potential overfitting or underfitting. In particular, if the model achieves high performance on the training data but poor performance on the validation data as depicted in the learning curves, it suggests overfitting.