Predicting 2024 Paris Olympics Men's 100m

Hypothesis:

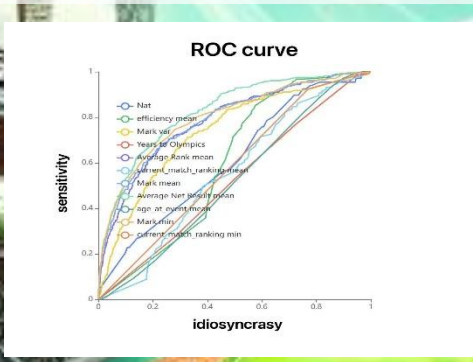
Whether an athlete reaches the final is determined by the form (performance) of the previous four years of competition, ignoring the effects of possible missed plays and opponent specificity in the actual Olympic competition.

Web Mining:

Data were collected from https://worldathletics.org

Influencing Factors:

Age at event
Average Net Result mean
Mean mark
Best mark(min)
Variance of the mark
Mean match ranking
Best match ranking(min)
Average Rank mean
Mean efficiency
Nation Score
Years to Olympics



Characteristic engineering:

1. Handling of outliers and missing values
2. Data formatting
3. Normalization of data
4. Pearson correlation analysis

	Olympic Ranking	Finalist_weight	Average Net Result mean_pro	age_at_event mean_pro	Mark mean_pro	Years to Olympics_pro	Nat_pro	efficiency mean_pro	current_match_ranking mean_pro	Average Rank mean_pro	current_match_ranking min_pro	Mark min_pro	Mark var_pro
Olympic Ranking	1(0.000***)	-0.893(0.000***)	0.282(0.000***)	0.022(0.135)	0.319(0.000***)	0.023(0.133)	-0.094(0.000***)	-0.107(0.000***)	0.076(0.000***)	0.255(0.000***)	0.075(0.000***)	0.308(0.000***)	-0.199(0.000***)
Finalist_weight	-0.893(0.000***)	1(0.000***)	-0.295(0.000***)	-0.026(0.085*)	-0.322(0.000***)	-0.02(0.177)	0.101(0.000***)	0.114(0.000***)	-0.05(0.001***)	-0.26(0.000***)	-0.067(0.000***)	-0.316(0.000***)	0.21(0.000***)

Note: ***, **, * represent 1 per cent, 5 per cent and 10 per cent significance levels, respectively.

Conclusion:

· Our logistic regression model, although scientifically sound and robust, has shown limited accuracy.
· Conversely, the random forest model displays exceedingly high accuracy. However, its simulation does not perfectly align with real-world conditions, indicating a need for further optimization and adjustments.
· The image on the right highlights the random forest model's predictions for strong contenders likely to reach the finals.

Athlete
Fred KERLEY
Oblique SEVILLE
Kishane THOMPSON
Noah LYLES
Akani SIMBINE
José GONZáLEZ
Louie HINCHLIFFE
Ferdinand OMANYALA
Pablo MATEO
Emmanuel ESEME

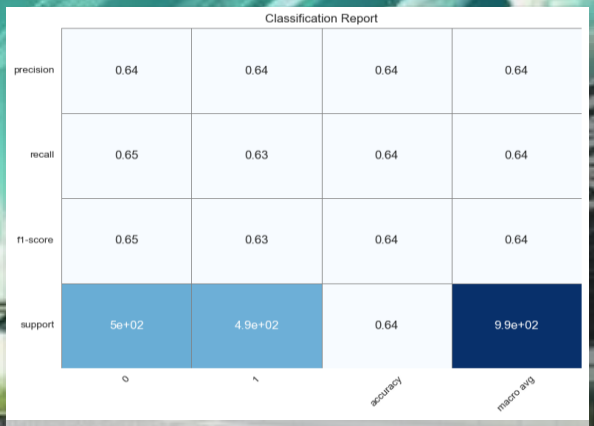
Further Research:

For future work, we plan to refine our models using additional data and advanced feature engineering techniques. Continuous improvement and validation of the models will help in achieving more accurate predictions.

Group3:
Youyao Gao, Yu Liang,
Haosong Sun, Xinyi Zeng

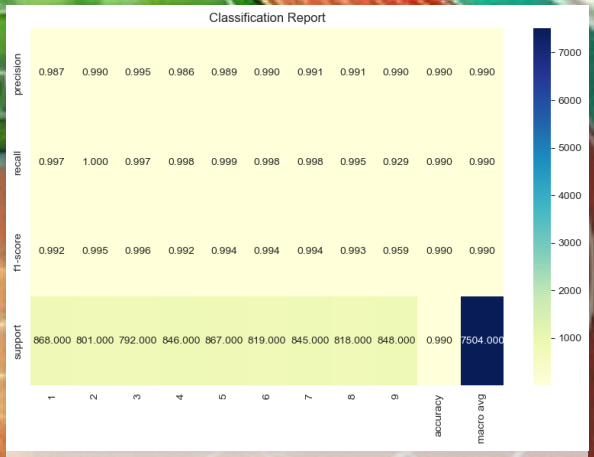
More information can be found on Github
https://github.com/dashen2004/NUS_Group3

Logistic Regression Model



Our classification report for logistic regression model provides balanced precision, recall, and F1-scores around 0.64-0.65. The precision for both classes (0 and 1) is 0.64, indicating that 64% of the predicted positives are true positives. The recall values are 0.65 for class 0 and 0.63 for class 1, showing the model's ability to identify actual positives. The overall accuracy is 0.64, suggesting the model correctly classifies 64% of instances. Overall, the model shows balanced performance with areas for improvement in feature engineering and model tuning.

Main Model



The classification report for main model provides an overview of the model's performance across nine classes (1 to 9). The precision values are all very high, ranging from 0.867 to 0.990, indicating that the majority of positive predictions for each class are correct. Recall values are similarly high, from 0.997 to 0.929, suggesting that the model successfully identifies most of the actual positive cases. The F1-scores, which balance precision and recall, reflect this consistency, with all values above 0.85. The overall accuracy of the model is 0.990, demonstrating excellent performance and reliability across all classes. This classification report highlights a robust model with consistent and reliable performance in predicting outcomes for each class.