Formula 1 Overtake Prediction

Analisi Predittiva con Machine Learning su Dati di Telemetria

Relazione Tecnica di Progetto

Autori:

Angelo Fusco & Mattia Fanzini

Università:

Università degli Studi di Salerno

Data:

14 Gennaio 2026

Indice

1 Introduzione

La Formula 1 moderna è definita dai dati. Ogni decisione strategica, dal pit stop alla gestione gomme, si basa su modelli predittivi. Questo progetto risponde alla domanda: *"È possibile prevedere la probabilità di un sorpasso utilizzando esclusivamente dati telemetrici storici?"*.

Utilizzando dati estratti dalla libreria open-source `FastF1`, abbiamo sviluppato una pipeline di Machine Learning completa per analizzare le dinamiche di gara del Gran Premio di Monza (2023). Sono stati confrontati tre algoritmi (Logistic Regression, Random Forest, XGBoost) per classificare l'evento binario "Sorpasso" (*IsOvertake*). Il modello migliore è stato infine integrato in una Dashboard interattiva.

2 Il Dataset e Analisi Esplorativa

2.1 Caratteristiche del Dataset

Il dataset di partenza è stato costruito estraendo la telemetria giro per giro. Dopo le fasi di pulizia, il dataset finale presenta le seguenti caratteristiche dimensionali:

- **Totale Campioni (Righe):** 16.145 eventi (coppie di giri pilota-tracciato).
- **Features (Colonne):** 12 variabili, incluse metriche grezze e ingegnerizzate.
- **Sbilanciamento Classi:** Il dataset presentava un forte sbilanciamento (Imbalance), dove gli eventi di sorpasso reale costituivano meno del 5% del totale dei giri, rendendo necessario l'uso di tecniche di oversampling.

2.2 Data Engineering e Feature Relative

I dati grezzi (velocità, giri motore) non sono predittivi se presi isolatamente. È stato sviluppato il modulo `relative_feature_builder.py` per calcolare metriche "differenziali" che catturano il duello tra attaccante e difensore:

- **Delta LapTime:** Differenza temporale sul giro tra i due piloti.
- **Delta Tyre Life:** Differenza di usura degli pneumatici (in giri).
- **Compound Advantage:** Vantaggio di mescola codificato numericamente (es. Soft=3 vs Hard=1).
- **Estimated Gap:** Stima del distacco fisico (metri/secondi) basata sul delta tempo.

2.3 Preprocessing e Gestione Outlier

Per garantire la qualità del training, sono stati filtrati eventi non competitivi (Pit Stop, Safety Car) utilizzando un filtro statistico basato sulla distribuzione dei tempi ($T_{limit} = \mu + 2\sigma$). I valori mancanti sono stati gestiti tramite `fillna(0)` per mantenere la consistenza temporale delle serie.
Per mitigare lo sbilanciamento delle classi, è stata applicata la tecnica **SMOTE** (Synthetic Minority Over-sampling Technique), generando campioni sintetici per la classe minoritaria "Sorpasso" durante il training.

3 Analisi dei Modelli e Risultati

La fase di training ha confrontato tre classificatori utilizzando uno split stratificato 80/20. Di seguito i risultati ottenuti.

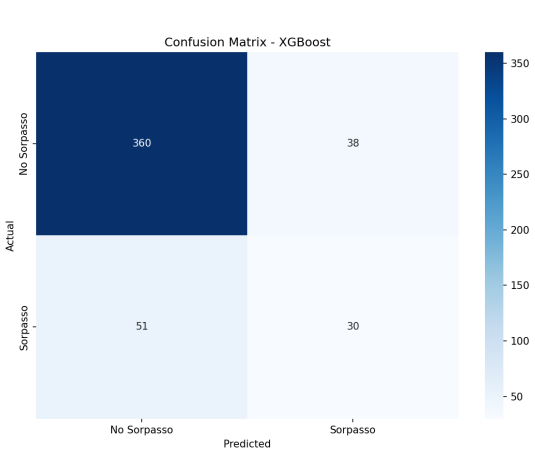
3.1 Confronto Metriche

Tabella 1: Performance dei Modelli (Dati da training_report.json)

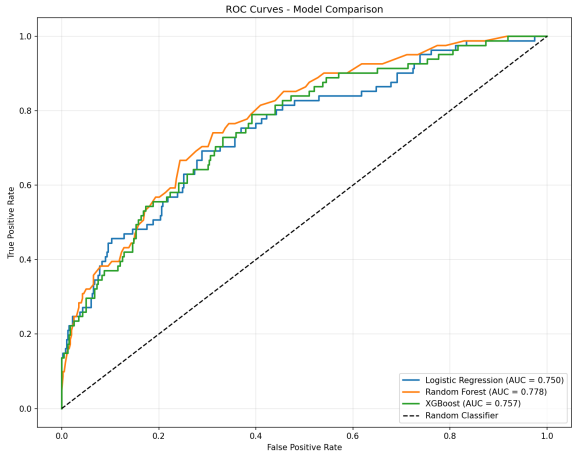
Modello	**Accuratezza**	**Precision**	**Recall**	**ROC-AUC**
Logistic Regression	71.2%	0.32	0.63	0.750
Random Forest	80.0%	0.40	0.39	0.778
XGBoost	**81.4%**	**0.44**	**0.37**	**0.757**

3.2 Selezione del Modello: XGBoost

XGBoost è stato selezionato come modello finale per la produzione. Sebbene la Logistic Regression avesse una Recall più alta, la sua Precisione era inaccettabile (troppi falsi positivi). XGBoost offre il miglior bilanciamento, minimizzando i "falsi allarmi" (solo 38 falsi positivi contro i 108 della regressione logistica), caratteristica fondamentale per non suggerire strategie di gara errate.



(a) Matrice di Confusione XGBoost



(b) Curve ROC a confronto

Figura 1: Visualizzazione delle performance di classificazione

4 Interpretabilità del Modello

Per comprendere quali fattori influenzano maggiormente la previsione di un sorpasso, è stata estratta l'importanza delle feature dal modello XGBoost.

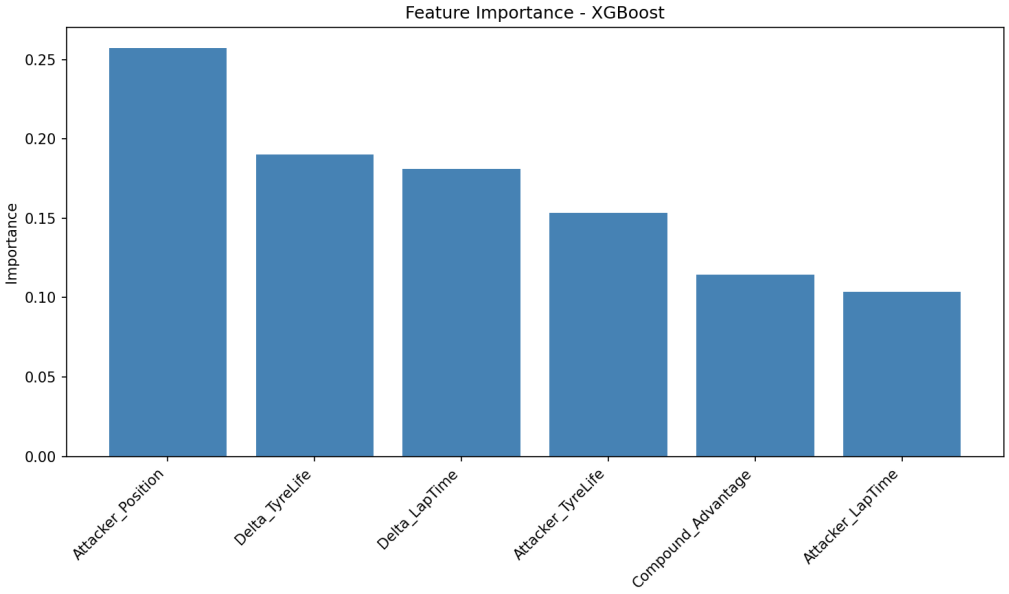


Figura 2: Feature Importance (XGBoost). Si nota come il "Delta LapTime" (differenza di passo gara) sia il predittore dominante, seguito dall'usura gomme.

5 Applicazione: F1 Strategy Room

Il culmine del progetto è l'applicazione web sviluppata con Streamlit (`app.py`), che carica il modello addestrato (`best_model.pkl`) per effettuare previsioni in tempo reale. Le

funzionalità principali includono:

1. **Simulazione Scenario:** L'ingegnere di pista può definire lo stato delle gomme (usura, mescola) per calcolare la probabilità immediata di sorpasso.

2. **Interactive Dashboard:** Visualizzazione grafica delle posizioni stimate sul tracciato di Monza tramite mappa SVG.

6 Conclusioni e Sviluppi Futuri

Il progetto ha dimostrato che è possibile prevedere i sorpassi con un'accuratezza superiore all'81% utilizzando esclusivamente dati storici e feature ingegnerizzate come il *Delta LapTime* e il *Delta TyreLife*. L'uso di SMOTE è stato determinante per gestire la rarità degli eventi di sorpasso.

6.1 Sviluppi Futuri

Per migliorare ulteriormente l'affidabilità del sistema in ottica di produzione, si propongono i seguenti sviluppi:

- **Integrazione Dati Meteo:** Includere variabili come temperatura asfalto e pioggia, che alterano drasticamente il grip.

- **Modelli Sequenziali (LSTM):** Utilizzare reti neurali ricorrenti per considerare lo storico dei 3-4 giri precedenti, catturando il "setup" del sorpasso.

- **Generalizzazione:** Estendere il training a tutti i circuiti del calendario, aggiungendo una feature categorica per la "difficoltà di sorpasso" del tracciato.