Immagine che contiene emblema, simbolo, cerchio, logo

Descrizione generata automaticamenteUniversità degli Studi di Padova – Dipartimento di Scienze Statistiche

**Insegnamento di Metodi statistici per big data**

**A.A. 2022/2023**

*Metodi di classificazione applicati alla Formula 1*

Riccardo Schiavon 2003666

Francesco Ciucevich 2003662

Giacomo Filippin 2003009

**Presentazione del progetto**

I dati utilizzati per questa analisi provengono dalla pagina web al seguente link <http://ergast.com/mrd/db/>, a cura di Chris Newell, e riportano tutte le informazioni e dati principali sui campionati di Formula 1 dal 1950, anno della creazione della competizione, all’ultima gara svolta del 2023.

L’obiettivo dell’indagine è quello di classificare correttamente ciascuna vettura in podio/non podio applicando alcuni metodi di classificazione differenti e comprendere quali siano i fattori che maggiormente influiscono su questo risultato. Si è deciso di svolgere l’analisi dall’inizio della cosiddetta “era ibrida” del motore, ossia dal 2014 alla stagione appena conclusa 2022.

Per ottenere ciò, è stato realizzato un dataset con più informazioni possibili ricavate sulla base dei vari file in formato .csv disponibili nel sito.

**Pre-processing e pulizia dei dati**

Nel corso della prima fase del lavoro, una prima “ossatura” del dataset è stata costruita eseguendo svariate unioni sequenziali tra i vari files tramite la funzione “merge” di R. I files utilizzati sono i seguenti: *‘driver\_standings.csv*’ e ‘*drivers.csv’*, contenenti variabili relative ai piloti, come ad esempio l’anno di nascita e la nazionalità, la classifica del campionato piloti di gara in gara e i punti cumulati di gara in gara; ‘*races.csv’* con variabili relative alle singole gare, come ad esempio l’orario e il circuito; ‘*circuits.csv’* con variabili relative ai circuiti, come ad esempio le coordinate; ‘*qualifying.csv’* con variabili relative alle qualifiche, come i giri veloci e la posizione di partenza in griglia; ‘*constructors.csv’* e ‘*constructor\_standings.csv’* contenenti variabili relative alle scuderie, come la classifica del campionato costruttori e i punti cumulati di gara in gara da quest’ultime; ‘*pit\_stops.csv’* contenente i vari tempi di sosta in pit lane e il numero di soste.

Sono state usate come chiavi le variabili ‘*driverId*’, ‘*constructorId*’ e ‘*raceId*’, che identificano univocamente ogni osservazione del nostro futuro dataset (ovvero ogni auto in ciascun gran premio tra il 2014 e il 2022 compresi).

Una volta creata l’ossatura, è stato fatto un lavoro di pulizia sulle variabili ottenute: sono state rimosse alcune colonne duplicate e alcune variabili non rilevanti per i nostri scopi, mentre le variabili di interesse sono state rinominate per facilitarne l’interpretazione.

A questo punto, dopo aver incorporato tutte le informazioni direttamente disponibili nei dataset a disposizione, è stato necessario ricavare molte variabili aggiuntive, tra cui, per fare degli esempi, il miglior tempo sul giro in qualifica, il tempo medio dei pitstop, la classifica piloti e la classifica costruttori dell’anno precedente e anche la variabile risposta, qualitativa dicotomica con valore 1 se una macchina è andata a podio in una particolare gara e 0 altrimenti.

Per la variabile risposta *podio*, dicotomica, sono state eseguite diverse operazioni successive: per ciascuna macchina in ciascuna gara, è stato sottratto al valore dei punti cumulati fino alla gara in questione (compresa) il valore dei punti cumulati alla gara precedente (compresa), ottenendo così i punti guadagnati da ciascuna macchina in ogni gara. Successivamente, usando i punti come indicatori della classifica delle gare, le auto sono state ordinate per ogni gara in base ai punti ottenuti, ottenendo l’ordine esatto delle prime 10 posizioni per ogni gara (gli ultimi 10 non ricevono punti e sono quindi impossibili da ordinare). Una volta ottenuto l’ordine di arrivo, è bastato dare valore 1 ai primi tre e 0 agli altri.

Questa strategia, che sarebbe stata perfetta in condizioni ideali, non è stata sufficiente in quanto spesso ci sono state gare a cui hanno partecipato meno piloti a causa di squalifiche, danni o problemi di altro tipo, e questo ha reso non sottraibili i punti cumulati. In particolare, ogni gara con anche solo una macchina in meno ha reso impossibile questa strategia sia per la gara stessa sia per la gara successiva, per un totale di 31 gare non calcolabili. I valori della risposta per le gare mancanti sono stati quindi inseriti manualmente in modo da completare il vettore e poter procedere con l’analisi e la modellazione.

Successivamente è stata creata la variabile *quali\_lap* selezionando il miglior giro tra le tre sessioni di qualifica (Q1, Q2, Q3) per ogni pilota dal file "*qualifying.csv*". Il miglior tempo sul giro, ai fini di agevolare l'analisi e permettere confronti tra i valori della variabile, è stato poi convertito in formato numerico.

Sono state poi rilevate due osservazioni mancanti, causate probabilmente da guasti o incidenti, le quali sono state sostituite con un valore di default di 120 secondi per evitare la perdita di osservazioni e mantenere così l'integrità del dataset.

Per le variabili "*pit\_medio\_anno*" e "*pit\_medio\_anno\_scud*", è stato utilizzato principalmente il file "*pit\_stops.csv*". Le due variabili, che rappresentano i tempi di sosta medi relativi all'anno precedente per pilota e per scuderia rispettivamente, sono state ottenute tramite un'operazione di "standardizzazione" delle medie dei tempi. Nello specifico si è sottratta la media dei tempi di sosta per quell'anno ai tempi medi ottenuti quell'anno da pilota e scuderia. Così facendo è stato ottenuto un valore indicativo del guadagno o della perdita effettivi di tempo ottenuti per ogni pilota e per ogni scuderia.

In fase di creazione delle variabili sono stati riscontrati alcuni problemi quali la presenza di valori anomali di tempi di pit stop; questi tempi eccessivamente lunghi sono dovuti dal ritiro della vettura in quella gara e quindi si è deciso di non considerarli al fine dell'analisi poiché non significativi.

Infine è stata rilevata la presenza di alcuni valori mancanti in una delle ultime versioni delle variabili; questa mancanza di dati è dovuta dalla presenza di un numero ristretto di piloti "rookie" al primo anno assoluto nella competizione, oppure di piloti tornati a gareggiare dopo qualche anno di pausa, e che quindi non hanno permesso di registrare i tempi relativi alla stagione precedente rispetto a quella presa in esame per ogni osservazione del dataset. Per risolvere questo problema è stato deciso di assegnare a quei piloti il valore medio dei tempi di sosta standardizzati della stagione precedente.

Per tenere in considerazione della posizione finale in classifica dei piloti e delle varie scuderie nell'anno precedente, sono state create due variabili rispettivamente *last\_position* e *last\_team\_position*. La variabile *last\_position* è stata creata calcolando prima la classifica finale dei piloti all'ultima gara dell'anno precedente ottenuta grazie a una variabile presente nel file "*races.csv*" che tiene conto dei punti cumulati dei piloti a ogni gara durante l'anno. Analogamente per la variabile *last\_team\_position*, grazie a una variabile che tiene in considerazione dei punti cumulati delle scuderie gara per gara durante l'anno, è stata presa la classifica finale delle scuderie all'ultima gara dell'anno precedente.

Come ci si aspettava però, queste variabili hanno creato parecchi valori mancanti dovuti a tutti i piloti e scuderie "rookie". Per gestire questi valori senza perdere osservazioni importanti, si è deciso di impostare dei valori di default. Poiché in media i piloti durante gli anni sono circa 20, per i piloti "rookie" è stato impostato un valore di *last\_position*=21 e analogamente poiché le scuderie in media sono state circa 10, per le scuderie nuove è stato impostato un valore di *last\_team\_position*=11. Questa gestione ha potuto preservare il numero totale di osservazioni disponibili per l'analisi.

Un’altra variabile creata è stata *Podiums*, presa in considerazione per tenere conto del numero di podi per ciascun pilota fino alla gara precedente (compresa), in quanto è di interesse valutare se questa variabile possa influenzare, e come, la variabile risposta.

Per la creazione di alcune di queste variabili si è deciso di non considerare le osservazioni riguardo a due scuderie nell’anno 2014, Marussia e Catheram, in quanto per problemi economici e finanziari non hanno potuto correre gli ultimi Gran Premi. È stato tuttavia considerato per la variabile *last\_team\_position* dell’anno successivo il fatto che la Marussia ottenne due punti nel 2014, al contrario della Catheram che decise di abbandonare la Formula 1 al termine di quell’anno.

Infine, è stata presa una decisione riguardo a un numero di osservazioni decisamente non cospicuo. Infatti, erano precedentemente presenti circa una decina di osservazioni riguardo a singole corse effettuate da terzi piloti delle scuderie che hanno sostituito i piloti principali per qualche specifico avvenimento in quella specifica gara. È stato deciso quindi di sostituire questi terzi piloti con il relativo pilota principale per mantenere l’integrità dei dati e permettere una più semplice interpretazione.

**Dataset finale**

Il dataset finale è quindi composto dalle seguenti variabili:

•*Podio* (variabile risposta)

•*Last\_position*

•*Last\_team\_position*

•*Quali\_lap*

•*Pit\_medio\_anno*

•*Pit\_medio\_anno\_scud*

•*Podiums*

•*Grid*

•*Year* (stagione di riferimento)

•*Round* (numero della gara stagionale)

Inoltre, per tenere in considerazione la presenza di alcune fondamentali variabili qualitative sconnesse, quali i nomi dei piloti, i nomi delle scuderie e i nomi dei circuiti, sono state create delle matrici di dummy con cui abbiamo sostituito i vettori originali delle variabili in modo da poter includere anche queste in fase di modellazione (modelli di tipo elastic net e alberi di classificazione necessitano infatti matrici numeriche). Successivamente abbiamo creato delle matrici di interazione utilizzando queste variabili dummy, in particolare, sono state aggiunte le interazioni pilota:scuderia, scuderia:anno, pilota:circuito, scuderia:round. È importante notare che quest’ultime richiedono un altissimo numero di nuove colonne (ad esempio, la matrice di interazione tra pilota e scuderia ha p = 748 colonne, quella tra pilota e circuito 1628), di conseguenza, non avendo la possibilità di adottare un approccio backward ‘alla cieca’ su tutte le interazioni possibili (che renderebbe p > n creando sovraparametrizzazione), abbiamo deciso di scegliere solo quelle per noi più sensate.

**Modellistica/modellazione dei dati**

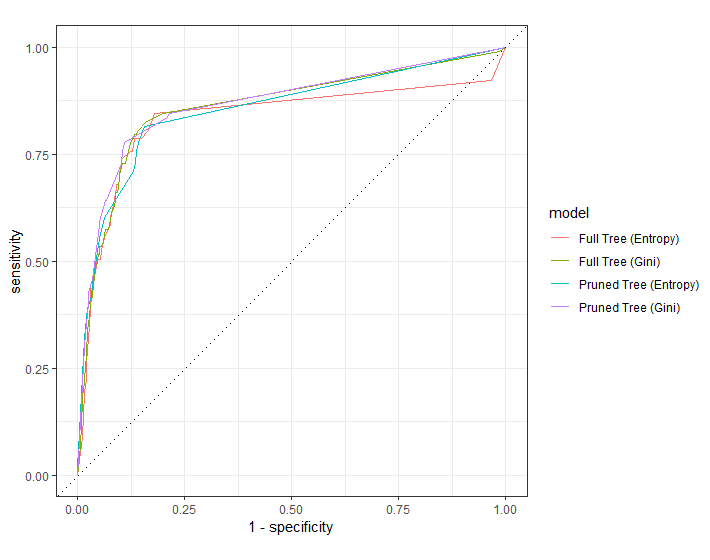
Per poter riprodurre gli stessi risultati e stabilire una base comune per la generazione di numeri casuali, è stato impostato un seme randomico (*set.seed()*). Il valore selezionato per il seme randomico è 123. Ciò significa che ogni volta che viene eseguito il codice o l'algoritmo, verranno generati gli stessi numeri casuali, consentendo una ripetibilità dei risultati ottenuti. Inoltre, per valutare l'efficacia dei nostri modelli, il dataset è stato suddiviso in un insieme di addestramento (train) e un insieme di test (test). È stato scelto di assegnare l'80% dei dati al train set, mentre il restante 20% sarà utilizzato per il test. Questa proporzione ci permette di utilizzare una quantità significativa di dati per l'addestramento del, mantenendo al contempo una porzione adeguata di dati separati per la valutazione delle prestazioni (724 osservazioni su 3618 totali).

Per valutare la bontà di previsione del nostro modello abbiamo deciso di utilizzare tre differenti approcci, rispettivamente un metodo di regolarizzazione (*elastic net*), un modello lineare generalizzato per risposte binarie ed un albero di classificazione.

Il primo modello valutato è stato un elastic *net*, che combina i modelli Ridge e Lasso attraverso il parametro alfa di mistura, in questo caso fissato a 0.5. L’idea, non potendo costruire un glm con 2500 coefficienti, è di selezionare solo quelli più rilevanti allo studio; sono stati quindi calcolato attraverso gli errori di stima per i vari lambda quale fosse quello ottimale, ed eseguendo il ‘taglio’ in corrispondenza del valore trovato (0.033), si sono selezionati 45 dei 2519 coefficienti. In fase di valutazione, per ogni modello utilizzato, sono stati confrontati ai podi reali dell’insieme di test le previsioni di ciascun modello, che sono state ottenute fissando una soglia sulle probabilità da esso stimate, e assegnando 0 o 1 a seconda che i valori superassero o meno questa soglia.È importante notare che la soglia utilizzata non è quella di default, ovvero di 0.5, ma è stata abbassata a 0.3; questo perché il valore della corretta previsione dei podi è molto maggiore a quello della corretta previsione dei non-podi, e in questo modo si è fatto sì che i vari modelli ‘rischiassero’ delle previsioni utili in più, anche a costo di un lieve calo di accuratezza.

Per quanto riguarda l’approccio con il modello GLM, è stato deciso di stimare un modello con 17 tra le 45 variabili risultate maggiormente significative tramite l’analisi del modello *elastic net*. Questo perché tra i 45 coefficienti stimati dall’*elastic net*, sono presenti delle variabili molto specifiche e quindi è stato deciso di selezionare le 17 variabili risultanti dopo un successivo taglio ad un lambda diverso da quello ottimale. Successivamente, tramite una procedura *backward* (stepAIC) è stato selezionato un modello con 9 variabili esplicative.

Durante la costruzione dell'albero, sono state utilizzate due diverse funzioni per misurare l'impurità dei nodi: l'indice di Gini e l'entropia. Sperimentando con entrambe le funzioni, si sono ottenuti risultati significativi che hanno permesso di valutare la bontà dell'albero di classificazione e di confrontarne le performance dei modelli basati su Gini e sull'entropia entrambi con albero completo e tagliato. Per quanto riguarda l’indice di Gini, l’albero potato prevede meglio rispetto a quello completo, e analogamente utilizzando l’entropia. Confrontando i vari errori stimati e guardando le curve ROC sembra preferibile la curva viola quindi l’albero potato utilizzando l’indice di Gini come misura per l’impurità dei nodi.



**Confronto tra i metodi**

Per confrontare i tre metodi utilizzati vengono riportati nella seguente tabella alcuni strumenti di valutazione. Oltre alle metriche comunemente considerate viene riportato anche l’F1-Score, particolarmente utile quando ci sono classi sbilanciate, come nel nostro caso di studio. L'F1 score è una media armonica tra la precisione e la sensibilità del modello. Questo significa che un F1 score elevato indica una buona bilanciata capacità del modello di ottenere sia pochi falsi positivi (alta precisione) che pochi falsi negativi (alta sensibilità).

|  |  |  |  |
| --- | --- | --- | --- |
|  | *ELASTICNET* | *GLM* | *ALBERO* |
| Accuratezza | *0.888* | *0.863* | *0.904* |
| Precisione | *0.604* | *0.513* | *0.660* |
| Sensibilità | *0.722* | *0.737* | *0.680* |
| Specificità | *0.917* | *0.884* | *0.942* |
| F1-Score | *0.658* | *0.605* | *0.670* |

Osservando la tabella, si nota che il modello lineare generalizzato con famiglia binomiale performa consistentemente meno in tutte le metriche, eccetto per la sensibilità, rispetto agli altri due metodi di classificazione. Si può quindi considerare meno efficace rispetto all’albero e all’elastic net che hanno valutazioni più simili. D’altro canto, possiamo invece notare come l’albero di classificazione batta l’elastic net in tutto tranne la sensibilità, che analogamente al GLM risulta più elevata ma non può da sola giustificare una preferenza verso questo metodo, in quanto può anche essere solo una conseguenza di stime più ‘conservative’ delle probabilità stimate da parte dei modelli elastic net e GLM.

Alla luce di queste valutazioni, si è considerato l’albero come il metodo più adatto a gestire questo caso di studio, ed è stato di conseguenza scelto per le interpretazioni e la conclusione del progetto.

Viene riportato di seguito l’immagine dell’albero finale ottenuto:

Immagine che contiene diagramma, testo, Piano, linea

Descrizione generata automaticamente

**Considerazioni finali**

Dall’albero ottenuto si nota che ciò che influisce maggiormente è partire o meno tra le prime quattro posizioni nella griglia di partenza. L’unico caso in cui si va a podio partendo oltre le prime quattro posizioni è partire quinto o sesto, guidando una Mercedes tra l’ottava e la quattordicesima gara di campionato. Viene inoltre fatta un’ulteriore distinzione tra le prime tre posizioni in griglia e la quarta. In particolare, partendo in quarta posizione si arriva a podio solo se la vettura non è una Mercedes antecedente al 2018. Per le prime tre posizioni, invece, risulta influente la posizione finale nella classifica Costruttori della scuderia nell’anno precedente, in particolare se la scuderia è arrivata quinta o peggio la macchina non andrà a podio. Qualora invece la posizione della scuderia fosse tra le prime quattro, la macchina andrà a podio se parte dalla pole position o se è una Red Bull dalla quindicesima gara di Campionato in poi o se il pilota è Hamilton.

Si sarebbe potuto interpretare l’albero più in profondità, tuttavia l’informazione presente negli split successivi è fornita da variabili che presentano collinearità con variabili già utilizzate.

Da questa interpretazione possiamo trarre delle conclusioni più generiche su alcuni fattori. In particolare, risulta che la scuderia Red Bull tende ad essere più veloce nella fase finale di Campionato, mentre la Mercedes nella fase centrale. Questo potrebbe essere un indicatore di una differente strategia sulla distribuzione degli sviluppi della vettura durante le stagioni. In conclusione, risulta evidente che parte fondamentale del risultato in gara sia determinato durante la qualifica, come evidenziato anche dal seguente grafico.

