PROGETTO DI MACHINE LEARNING

Come battere una leggenda: Kobe Bryant vs Machine Learning

Gabriele Carrara | Alberto Filosa | Davide Garavaldi | Simone Tufano

Gabriele Carrara
Matricola: 814720
e-mail:g.carrara12@campus.unimib.it

Alberto Filosa
Matricola: 815589
e-mail: a.filosa1@campus.unimib.it

Davide Garavaldi
Matricola: 818308
e-mail:d.garavaldi@campus.unimib.it

Simone Tufano
Matricola: 816984
e-mail: s.tufano1@campus.unimib.it

In questo studio sono stati analizzati i tiri tentati da Kobe Bryant nel corso della sua carriera, al fine di costruire un modello in grado di prevedere se il tiro entri o meno nel canestro. In particolare, attraverso l'utilizzo di tecniche di Machine Learning, sono stati applicati e confrontati diversi modelli di predizione binaria e tecniche di valutazione della Variable Importance. I risultati sono stati proposti secondo un'ottica di utilizzo reale dei dati nel corso di una partita, al fine di limitare il rendimento del giocatore avversario.

KEYWORDS:
Kobe Bryant - Basket – Machine Learning – Classificazione Binaria

1 | INTRODUZIONE

Il 26 gennaio 2020 negli Stati Uniti scompare Kobe Bryant[1], famoso ex atleta NBA (National Basket Association), considerato da molti uno dei migliori giocatori di basket di tutti i tempi. Per più di vent'anni è stato un'icona sportiva a livello mondiale, nonché un idolo per molti ragazzi come noi che sono cresciuti guardandolo giocare a basket; per questo, infatti, si è deciso di omaggiarlo studiando un dataset relativo ai tiri da lui tentati nell'arco della sua vincente carriera.

L'idea di partenza dell'analisi è stata quella di mettersi nei panni di un ipotetico allenatore avversario per cercare di capire come affrontare Kobe Bryant, analizzando i suoi punti di forza e quelli di debolezza nel tirare a canestro. Le domande di ricerca in particolare sono state:

1. Da quale posizione è più favorevole far tirare Kobe Bryant per evitare di subire un canestro?
2. L'efficacia nel tiro dipende solo dalla distanza o dipende anche da altre circostanze?

2 | DESCRIZIONE DEL DATASET

Per rispondere alle domande dell'analisi è stato preso in considerazione il dataset "Kobe Bryant Shot Selection", relativo ad una competition del sito di Kaggle[2]. Il dataset è composto da 30698 osservazioni e 25 variabili, che vengono spiegate di seguito:

1. *combined_shot_type*: tipologie di tiro generiche (6 livelli);
2. *action_type*: tipologie di tiro specifiche (57 livelli);
3. *game_event_id*: identificatore di ogni singola azione della partita;
4. *lat*: latitudine dello stadio;
5. *lon*: longitudine dello stadio;
6. *loc_x*: coordinata sull'asse x da cui viene tentato il tiro (-250 indica la linea laterale destra, 250 indica la linea laterale sinistra);

7. *loc_y*: coordinata sull'asse y da cui viene tentato il tiro (-40 indica la linea di fondo dietro al canestro verso cui viene effettuato il tiro, 900 la linea di fondo opposta);
8. *minutes_remaining*: minuti rimanenti in un periodo;
9. *period*: periodo di gioco (4 periodi più eventuali supplementari);
10. *playoff*: variabile binaria, indica se la partita si giochi nei Playoff o in Regular Season;
11. *season*: stagione in corso;
12. *seconds_remaining*: secondi rimanenti nel periodo;
13. *shot_distance*: distanza da cui è stato tentato il tiro (in piedi);
14. *shot_type*: variabile binaria, indica se il tiro è da 2 o da 3 punti;
15. *shot_zone_basic*: zona del campo da cui viene tentato il tiro (7 livelli);
16. *team_id*: identificativo della squadra (costante "LA");
17. *team_name*: nome della squadra (costante "Los Angeles Lakers");
18. *game_date*: data in cui si è giocata la partita;
19. *matchup*: squadre che si affrontano;
20. *opponent*: nome della squadra avversaria;
21. *game_id*: identificativo della partita;
22. *shot_id*: identificativo del tiro effettuato;
23. *shot_made_flag*: variabile binaria, indica se il tiro sia stato messo a segno o no;
24. *shot_zone_area*: area del campo da cui viene tentato il tiro (6 livelli);
25. *shot_zone_range*: range di distanza da cui viene tentato il tiro (5 livelli).

Nello specifico, è stata considerata *shot_made_flag* come variabile risposta. Su quest'ultima erano presenti 5000 osservazioni con valori mancanti, poiché erano i valori da prevedere ai fini della competition sopra citata. Esclusa la variabile *shot_made_flag*, il dataset non presentava altri valori mancanti. Di seguito vengono riportati i grafici relativi alle variabili *shot_zone_area* e *shot_zone_range*, in modo da rappresentare più chiaramente le diverse zone di un campo da basket (Figura 1)[3].

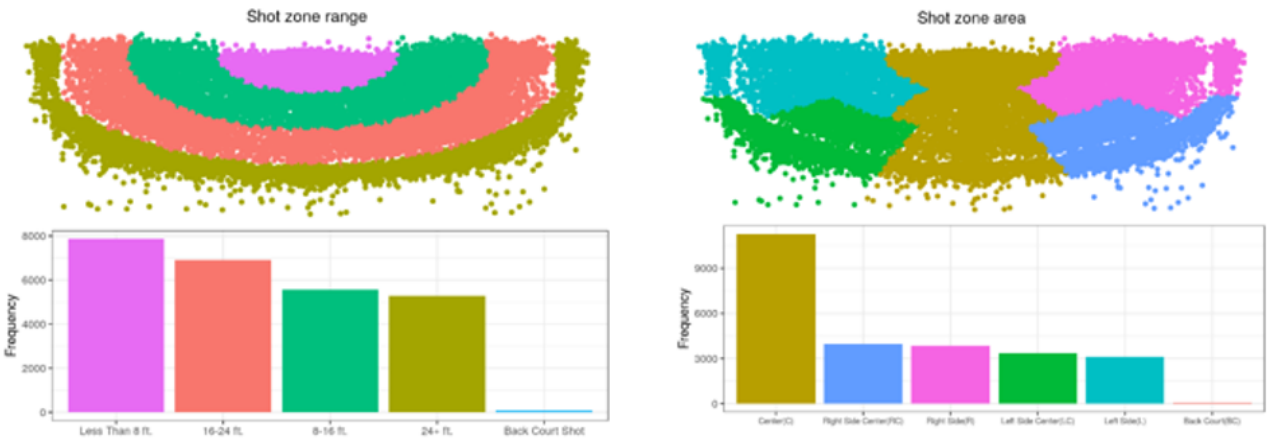

Figura 1 Distribuzione Shot Zone Range e Shot Zone Area

Infine, per realizzare una breve analisi esplorativa, è stato inserito il nodo "Statistics" che permette di vedere la distribuzione univariata delle variabili e le loro statistiche descrittive basilari. Concentrandosi sulla variabile risposta, si può osservare che 14232 tiri (55,38%) non sono stati segnati, mentre 11465 (44,62%) si sono conclusi con un canestro; questo fatto porta, quindi, ad affermare che non ci si trova di fronte ad un problema di sbilanciamento delle modalità della variabile risposta.

3 | DATA PREPARATION

Come primo passo, è stata svolta una selezione preliminare delle variabili di interesse, in quanto alcune facevano riferimento al medesimo argomento, ma con diversa granularità. In particolare, gli insiemi di variabili che presentavano questo problema sono stati due. Nel primo caso, le variabili in questione sono state *game_id*, *matchup* e *game_date* e sono state escluse, in quanto non significative per lo studio. Nel secondo caso, invece, le variabili in questione sono state *action_type* e *combined_shot_type* e, poiché erano interessanti per l'analisi, sono state create due diverse partizioni del dataset: nella prima, è stata inserita la variabile *action_type* e nella seconda la variabile *combined_shot_type*, mantenendo costanti tutti gli altri attributi. Inoltre, in questa fase preliminare sono state eliminate le variabili a modalità costante (*team_id* e *team_name*). Date le dimensioni del dataset (più di 30000 righe), si è optato per la procedura di "Holdout". Entrambi i dataset ottenuti (sia quello contenente *action_type* sia quello contenente *combined_shot_type*) sono stati sottoposti a due differenti divisioni:

1. è stato estratto un dataset di Test da quello totale (20% delle osservazioni) e la restante parte è stata suddivisa in due porzioni, denominate Train e Validation (rispettivamente costituite dal 80% e dal 20% delle osservazioni della parte in questione);
2. il dataset totale è stato suddiviso esclusivamente in Train e Test (80% e 20% delle osservazioni).

Di seguito si riporta lo schema riassuntivo che illustra le divisioni effettuate.

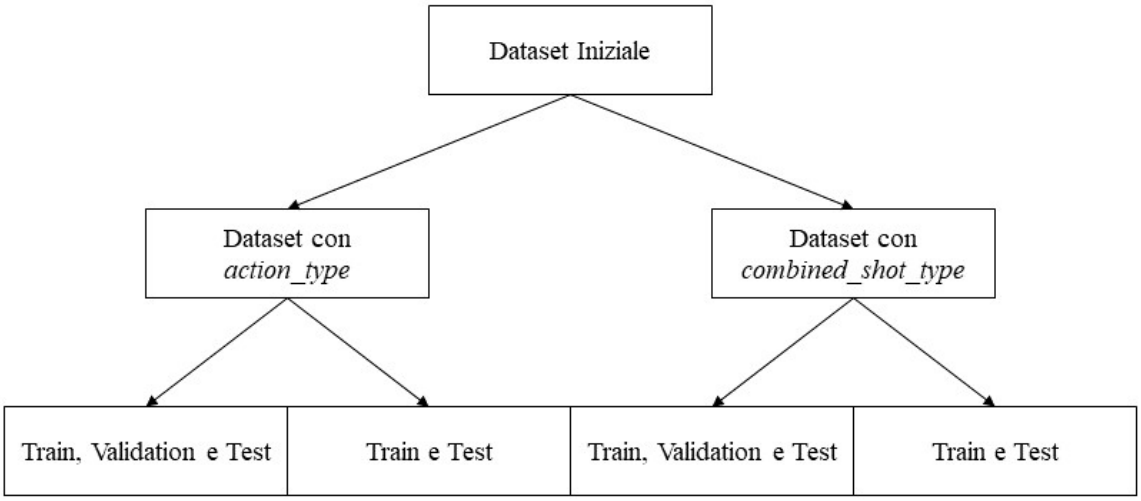

Figura 2 Divisione del dataset

4 | FEATURE SELECTION

Nelle due divisioni in cui è presente il dataset di Validation è stata svolta una selezione delle variabili tramite la procedura di Variable Selection. Quest'ultima è stata applicata utilizzando tre tecniche diverse: Wrapper J48, Wrapper NB Tree e Wrapper Random Forest. Una volta confrontati i diversi output, è stata fatta un'unica selezione tenendo conto dei risultati delle tre tecniche. Per quanto riguarda il dataset contenente *action_type* le variabili più importanti sono state *playoffs*, *seconds_remaining* e *action_type*; invece, per quanto riguarda il dataset contenente *combined_shot_type*, le variabili più importanti sono state *seconds_remaining* e *combined_shot_type*. Al contrario, la Feature Selection non è stata applicata ai casi in cui non fosse presente il dataset di Validation; per questi ultimi, si è provveduto esclusivamente ad eliminare le variabili collineari, poiché avrebbero potuto creare problemi nella fase di modellistica. A tale scopo, infatti, è stata costruita la matrice di correlazione tra variabili ed è stata scelta una soglia di 0.80 per l'eliminazione di queste ultime. In entrambi i dataset è stata rimossa la variabile *shot_zone_range*, dato che era particolarmente correlata con la variabile *shot_zone_basic*.

5 | MODELLI

Ai quattro dataset precedentemente ottenuti sono stati applicati i seguenti modelli tramite l'estensione di Knime Weka 3.7:

- J48, albero decisionale;
- Random Forest con il parametro num_trees pari a 10;
- SVM (Support Vector Machine);
- Naive Bayes Tree;
- Simple Logistic;
- Naive Bayes.

Tutti i modelli sono stati lanciati con i parametri di default di Knime. Inoltre, per i dataset in cui non è stata applicata la Feature Selection, dove il numero di covariate era maggiore, è stato utilizzato anche il modello KNN (K Nearest Neighbour), con numero di vicini pari a 10. I modelli sono stati applicati prima al dataset di Train (dove si è svolta la fase di "apprendimento") e, successivamente, al dataset di Test, dove sono state valutate le relative performance tramite i nodi "Weka Predictor" e "Scorer". La metrica scelta per il confronto tra i modelli è stata l'accuracy, il cui valore puntuale è stato estratto per ogni modello con il relativo intervallo di confidenza (α=0.95).

6 | ASSESSMENT

Di seguito si riportano le accuracy di tutti i modelli con i relativi intervalli di confidenza.

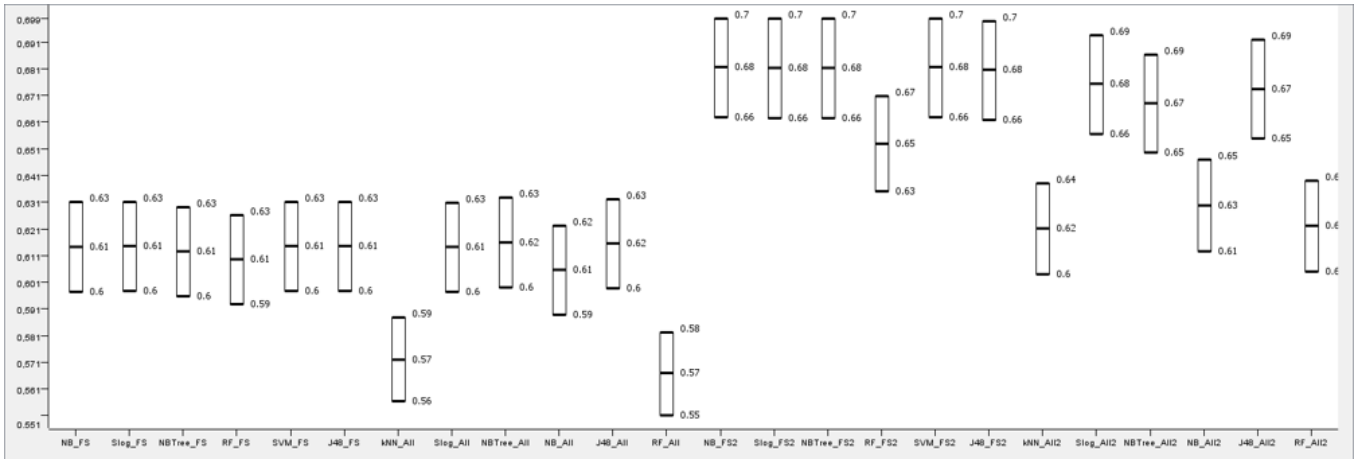

Figura 3 Intervalli di confidenza per l'accuracy

I modelli che risultano statisticamente più significativi sono otto e provengono tutti dalla partizione di dataset contenente *action_type*. Nello specifico, i primi cinque modelli sono relativi al dataset a cui era stata applicata Feature Selection e sono:

- NB Tree;
- Simple Logistic;
- Naive Bayes;
- J48;
- SVM.

Invece, i restanti 3 modelli sono relativi al dataset a cui non era stata applicata Feature Selection e sono i seguenti:

- Simple Logistic;
- NB Tree;
- J48.

Come si vede dal grafico (Figura 3), tutti questi modelli hanno un'accuracy che si aggira attorno a 0.68 e differiscono poco l'uno dall'altro. Perciò, per eseguire un ulteriore confronto, sono state costruite le curve ROC e Lift per gli otto modelli precedentemente considerati. Di seguito, vengono riportate le curve ROC per i modelli applicati sia al dataset che presentava Feature Selection (Figura 5), sia al dataset che non presentava quest'ultimo procedimento (Figura 4).

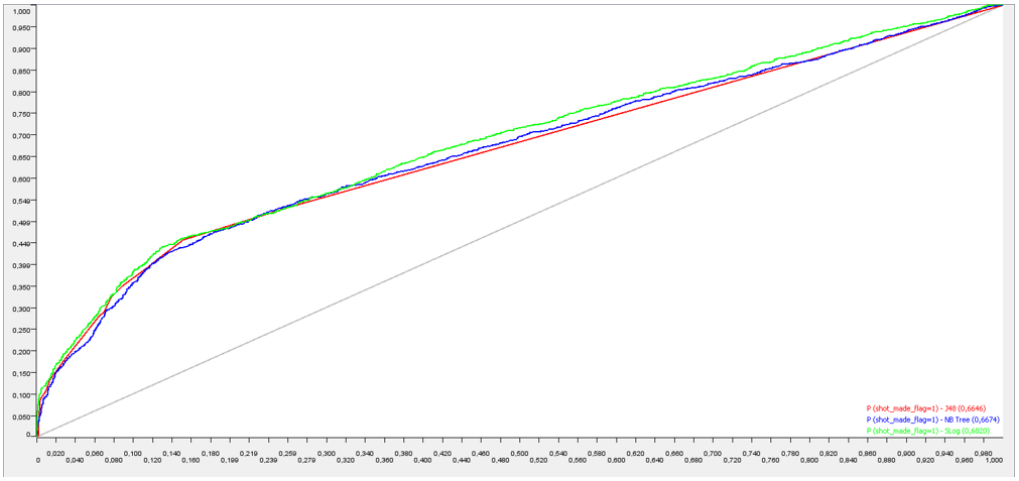

Figura 4 ROC dei modelli senza Feature Selection

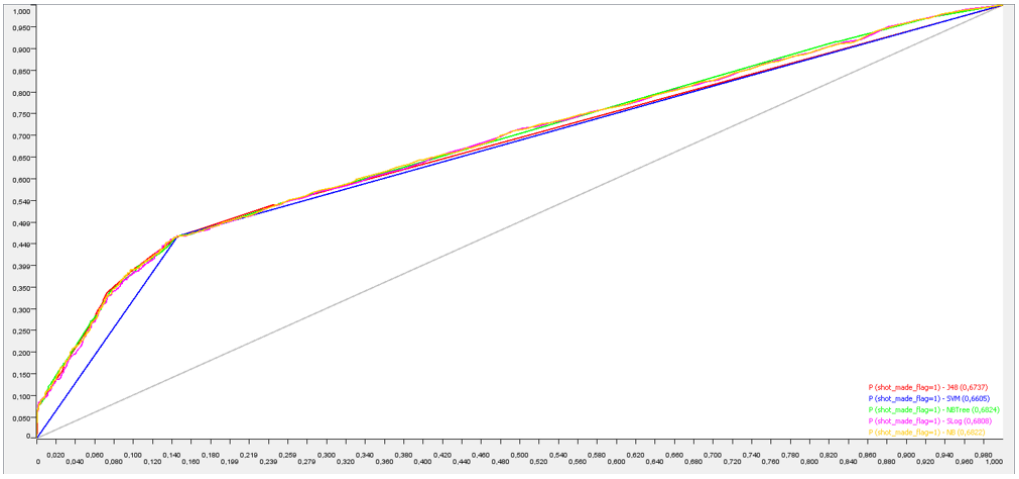

Figura 5 ROC dei modelli con Feature Selection

Dal confronto fra queste curve e da quello delle curve Lift risulta, seppur di poco, che il modello più performante è il Simple Logistic applicato al dataset senza Feature Selection. Di seguito si riporta la curva Lift del suddetto modello (Figura 6).

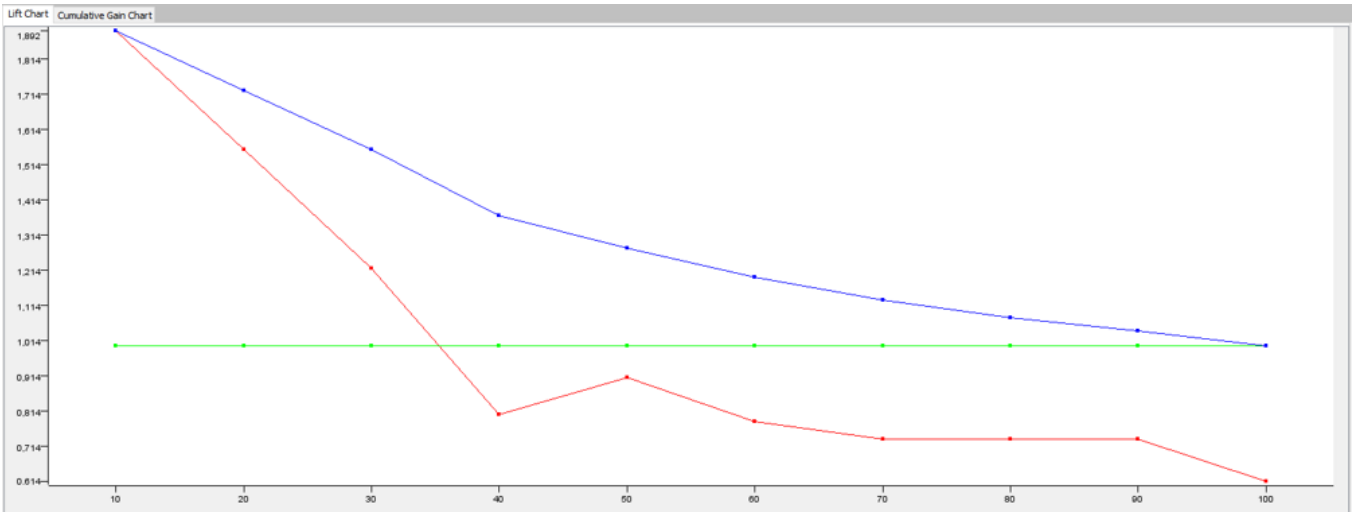

Figura 6 Lift chart del modello Simple Logistic

7 | SCORING E VISUALIZZAZIONI

Il modello vincente è stato quindi utilizzato per fare scoring sulle 5000 righe della variabile risposta su cui inizialmente non erano presenti valori. I risultati sono stati poi rappresentati graficamente tramite il nodo "R View (Workspace)"[4]; il grafico è stato costruito tramite il pacchetto ggplot2 di R[5]. Di seguito si riporta il risultato.

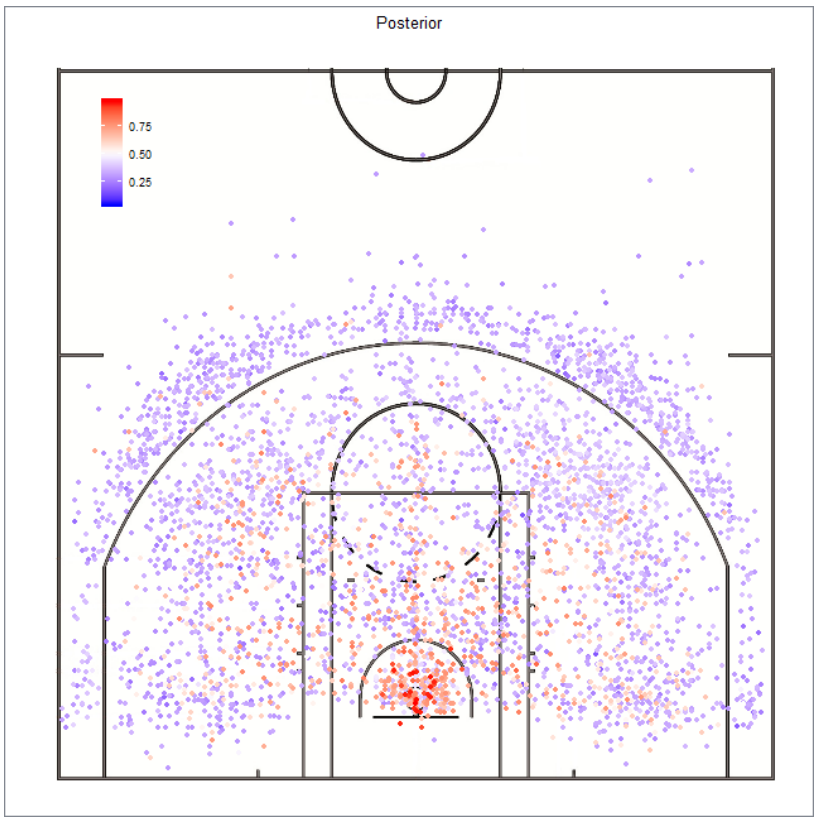

Figura 7 Shot chart con le posterior per ogni tiro

Dal grafico (Figura 7) si nota che i tiri con la probabilità di successo più alta sono quelli più vicini al canestro. Tuttavia, mettendosi nei panni di un ipotetico allenatore avversario, è necessario anche considerare il valore del tiro (da 2 o da 3 punti). Si è quindi proceduto a calcolare una nuova variabile *Expected Points*, ottenuta moltiplicando la posterior predetta dal modello, per il valore del tiro (se da 2 o da 3 punti). Questo procedimento è stato svolto poiché i tiri da 3 possono portare, anche in presenza di percentuali realizzative leggermente inferiori rispetto ai tiri da 2, una maggior quantità di punti segnati (ad esempio, a parità di numero di tiri, segnare con il 40% da 3 punti equivale a segnare con il 60% da 2 punti). La nuova variabile *Expected Points* è stata visualizzata nella Figura 8.

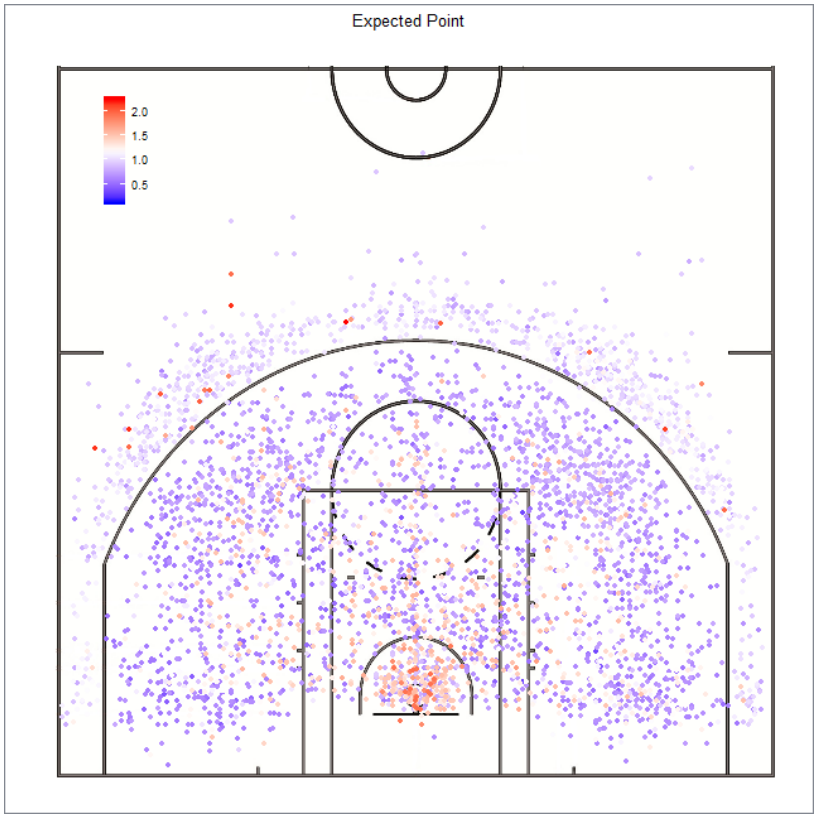

Figura 8 Shot chart con expected point per ogni tiro

Dal grafico si nota che, in generale, i tiri da 3 punti risultano essere migliori rispetto alla maggior parte dei tiri da 2, sebbene presentino una bassa probabilità di essere segnati. Inoltre, i tiri vicino al canestro rimangono quelli con il valore di punti attesi più alto, nonostante la differenza sia meno netta rispetto al grafico precedente.

8 | CONCLUSIONI

In risposta alle domande su cui si basa lo studio, si può affermare che:

1. Kobe Bryant risulta particolarmente efficace nel segnare i tiri effettuati vicino al canestro quindi, dal punto di vista della squadra avversaria, è consigliabile tenerlo lontano da questa zona;
2. Anche nei tiri da 3 punti tentati dal lato destro del campo (parte sinistra dell'immagine) Kobe Bryant risulta abbastanza performante, perciò è sconsigliabile concedergli anche questa tipologia di tiro;
3. Il tiro ideale da permettere è il cosiddetto "Mid-range shot" (tiro da 2 punti ma lontano dal canestro), sia per la minor probabilità di segnare, sia per il valore del tiro tentato;
4. La variabile che risulta più importante affinchè un tiro venga segnato o meno non è la distanza dal canestro, bensì il tipo di tiro. Tuttavia, è altrettanto vero che la tipologia di tiro dipende in parte dalla posizione del campo da cui viene effettuato;
5. Negli ultimi secondi dei tempi di gioco così come nei Playoff, momenti in cui la partita si fa più difficile, anche i tiri normalmente più semplici risultano essere più complessi, tant'è che le sue percentuali (e di conseguenza le posterior) si abbassano.

In generale, si può dire che un modello non potrà mai prevedere perfettamente se un tiro verrà segnato o meno. Infatti, un tiro tentato più volte dalla medesima posizione, sebbene facile, non sarà mai segnato con il 100%, perché anche un campione come Kobe Bryant può sbagliare in certe situazioni; al contrario, tiri ad altissimo coefficiente di difficoltà possono essere segnati nonostante le basse probabilità di successo, fatto accaduto più volte durante la carriera del giocatore. Inoltre, nel dataset non erano presenti altre variabili rilevanti per ottenere una previsione più accurata, come la distanza del marcatore, la stanchezza del giocatore o alcuni aspetti psicologici, ad esempio la sicurezza nel tiro o il momento della partita.

RIFERIMENTI

1. Wikipedia . Kobe Bryant — Wikipedia, The Free Encyclopedia. 2020. [Online; accesso Febbraio-2020].

2. Kaggle . Notebook di xvivanco. 2020. [Online; accesso Febbraio-2020].

3. Kaggle . Kobe Bryant Shot Selection. 2020. [Online; accesso Febbraio-2020].

4. Knime Hub . Basic R Usage. 2020. [Online; accesso Febbraio-2020].

5. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York . 2009.

6. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in StatisticsSpringer New York . 2013. [Approccio Metodologico].

7. Stella FA. *Machine Learning [Lecture notes or PowerPoint slides]* . 2019-2020. [Approccio Metodologico].