Il mio primo progetto con Kaggle[1]

0,78468 (accuracy: top 14%)

di

Giuseppe Sinatra

La sfida proposta dal famoso sito Kaggle[2] viene presentata come didattica, nel senso è pensata per essere un primo step nel mondo delle sfide relative all'ambito del *Machine Learning*, nonostante questo si presenta come sfidante e come un'ottima introduzione ai problemi reali.

Il progetto prevede la costruzione di un modello che possa predire quali passeggeri del Titanic si sono salvati e quali no, per fare questo abbiamo a disposizione un training dataset dove è indicato se il passeggero è sopravvissuto o no e alcune informazioni come: classe di appartenenza, nome, età, costo del biglietto, se viaggiava da solo oppure no, ecc…
Il nostro obiettivo è usare questi dati per applicare il nostro modello sul test dataset e valutare l'accuratezza della nostra previsione.

Senza una struttura ben chiara di quali siano gli step da svolgere, sarà molto complesso spostarsi dalla previsione donne salve e uomini no, questo richiede molti step preliminari. Vi racconto la mia esperienza visto che questo è stato il primo progetto affrontato dopo un intenso periodo di studio.

Il primo problema che ho affrontato e cercare d'individuare le feature che mostrassero una correlazione con la variabile target, per alcune variabile questo l'ho potuto fare direttamente perché non presentavano valori mancanti, per altre invece prima ho dovuto riflettere su come meglio avrei potuto riempire questi valori.
Costruire diversi tipi di grafici è stata un'ottima occasione di apprendimento in quanto mi ha permesso di utilizzare rappresentazioni che nell'ambito fisico sono poco usate come per esempio i boxplot. Per capire come leggere questi grafici mi sono fatto aiutare da

---

[1] https://www.kaggle.com/giuseppesinatra

[2] www.kaggle.com

ChatGPT, la cui chat è stata anche un modo interessante per tenere traccia dei vari tentativi fatti e alcuni canali YouTube come quello di Ken Jee[3].

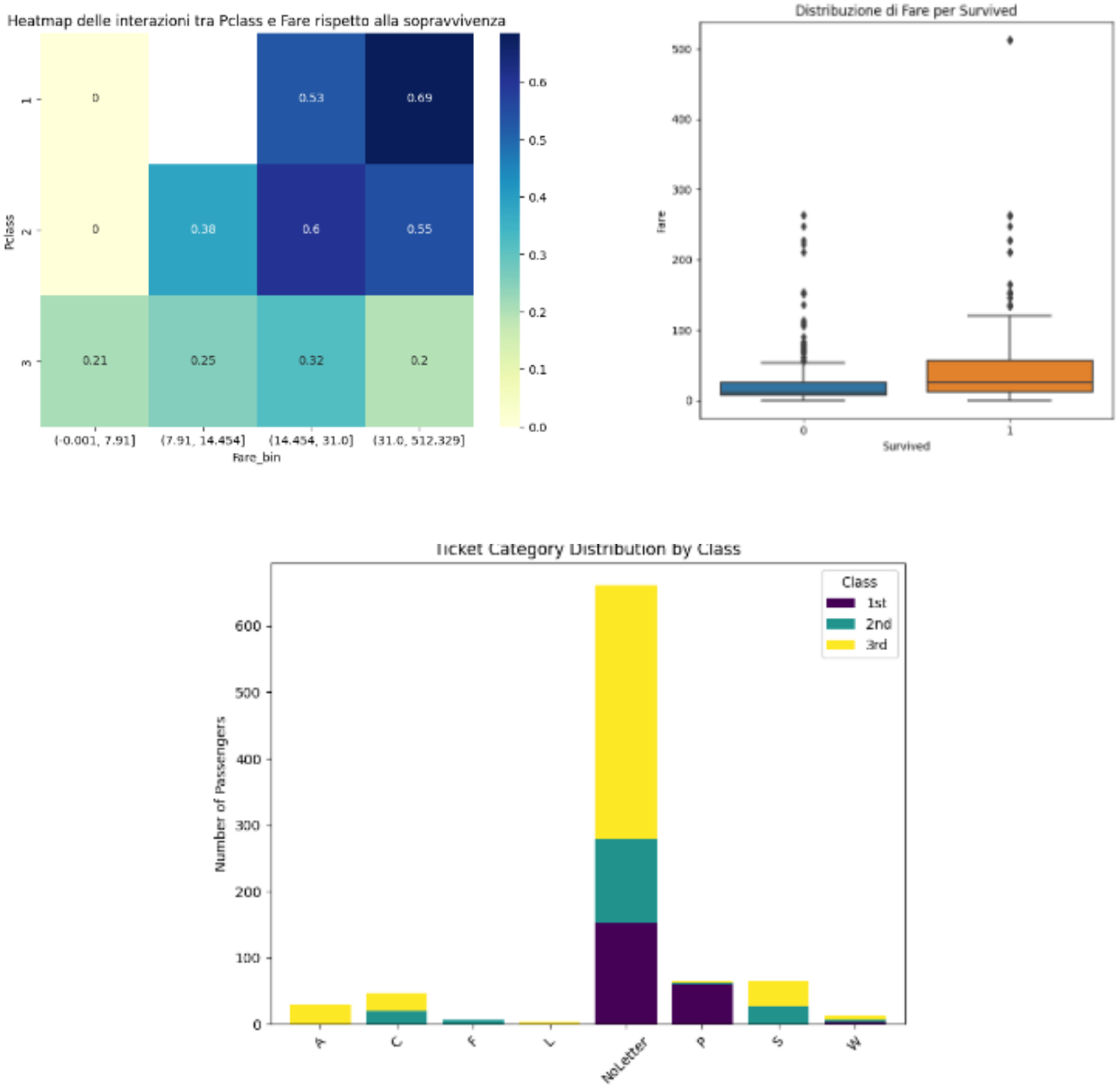

Fig. 1-2-3: Alcuni grafici per ispezionare i dati

Finita questa fase d'ispezione mi sono occupato di queste attività: divisione in bin delle variabile *fare* e *age*, per poter mettere in evidenza come anche all'interno della stessa variabile sono presenti delle categorie che hanno maggior probabilità di salvarsi, ho anche deciso su quali variabili doveva concentrarsi il mio modello, questa è solo una prima

[3] https://www.youtube.com/@KenJee_ds

selezione anche perché alcune quantità sono state sostituite dai *get_dummies*, che sostituisce la variabile iniziale.

La fase successiva è stata quella di definire le *pipeline* dei modelli che volevo utilizzare, nel mio caso ne ho usati 5: *Logistic Regression, Decision Tree, KNeighbors, Random Forest, Xgboost.* Definiti i modelli ho individuato quali feature per ogni modello influenzassero di più l'accuratezza, come si può vedere in *Fig. 4* ogni modello è sensibile ad un sottoinsieme diverso.

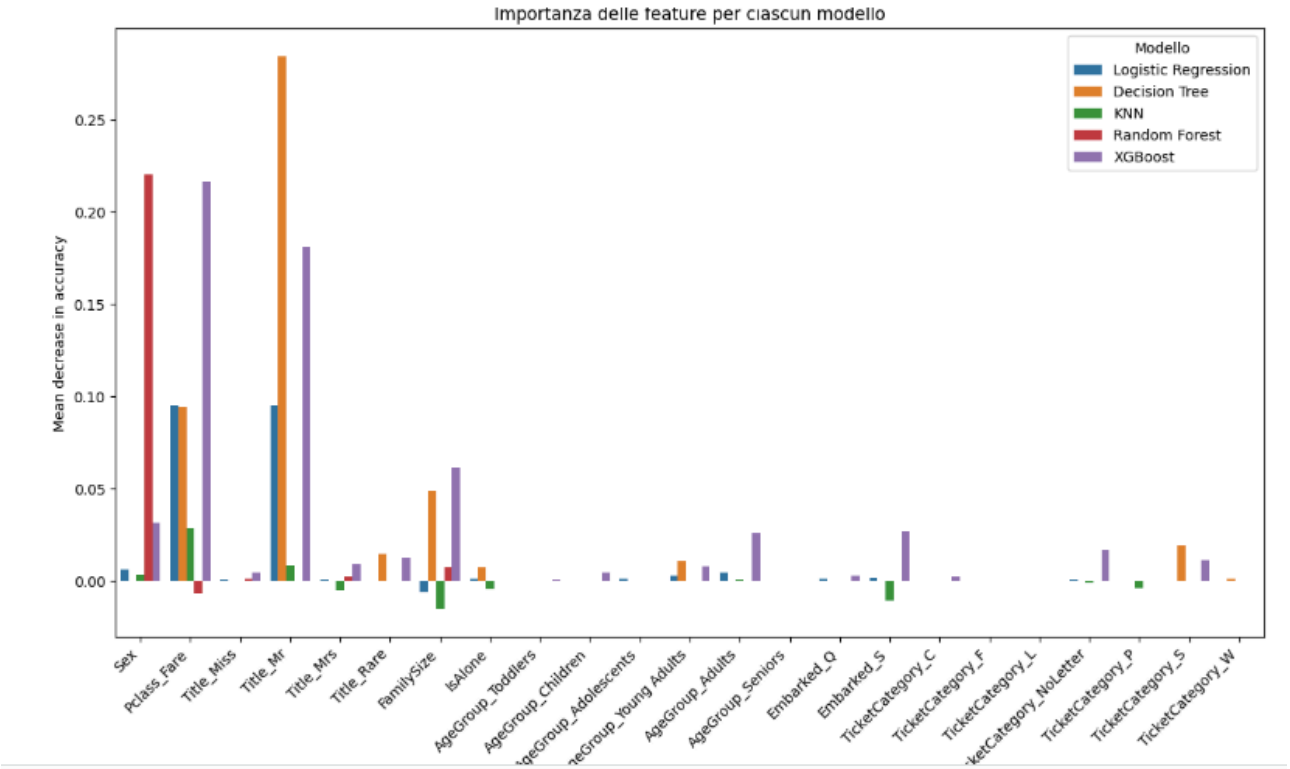

Fig 4: Importanza delle feature per ciascun modello

Come si può notare vale la pena allenare ogni modello solo sulle specifiche feature che risultano più impattanti. Questo però ha come conseguenza la necessità di dover simulare un *MajoringVote* perché quello implementato su *Scikit-Learn* prevede che tutti i modelli siano allenati sulle stesse variabili.

Gli ultimi passaggi prima di arrivare al risultato finale sono stati quelli di eseguire una *GridSearch* per trovare i parametri più adatti per ogni modello e la *10-fold cross validation* per stimare il rischio di overfitting. Per quanto riguarda l'ensemble dei modelli visto che dobbiamo prevedere 0 o 1 ho definito una media delle predizioni dei singoli modelli pesata sull'accuratezza determinata dalla *cross validation*.

Il lavoro da fare è ancora lungo e migliorabile ma sono rimasto molto stupido di quanto ho dovuto imparare per poter affrontare un caso reale, effettivamente c'è molta differenza rispetto a quelli che si possono trovare in un libro. Il mondo del *ML* è veramente affascinante in quanto ti pone delle sfide interessanti e ti richiede di andare sempre un pochino più in profondità.