

Le Hackaton e la peggiore Hackaton del mondo

Cosa sono le Hackaton

- software
- machine learning

Hackaton commerciali

- per soluzioni industriali
- per scouting

Hackaton non commerciali

- Educative
- AI 4 Good

Kaggle

<https://www.kaggle.com/c/siim-covid19-detection>

The screenshot shows the main landing page of a Kaggle competition. At the top, it says "Featured Code Competition". The title of the competition is "SIIM-FISABIO-RSNA COVID-19 Detection". Below the title, it says "Identify and localize COVID-19 abnormalities on chest radiographs". To the right, it shows a "\$100,000 Prize Money". The logo for "SIIM Society for Imaging Informatics in Medicine" is present, along with the text "194 teams · 2 months to go (2 months to go until merger deadline)". At the bottom, there are navigation links for "Overview", "Data", "Code", "Discussion", "Leaderboard", and "Rules", and a prominent "Join Competition" button.

This is a detailed view of the competition's overview page. It features a sidebar with links for "Description", "Evaluation", "Timeline", "Prizes", "Code Requirements", and "Call For Models". The main content area contains a paragraph about COVID-19's severity and diagnosis, followed by another paragraph about current diagnostic methods. To the right of the text, there is a small image of a chest radiograph showing COVID-19 abnormalities.



[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [Host](#) [My Submissions](#) [Submit Predictions](#)

Overview

Edit

Description

Evaluation

+ Add Page

Questa è una competition privata di School of AI Italia

Al mercato del pesce di Tokyo abbiamo misurato il peso di pesci di diverse specie.
Il vostro compito è stimare con la maggiore precisione possibile il peso dei pesci.

La metrica usata per la competizione è il classico RMSE.
Potete unirvi in team e fare fino a 4 tentativi al giorno.
La competizione terminerà il 3 luglio 2021.

Good Corp vs Evil Corp Hackaton



Netflix Prize hackaton

https://en.wikipedia.org/wiki/Netflix_Prize

Netflix **Cinematch**

A trivial algorithm RMSE of 1.0540

Cinematch uses "straightforward statistical [linear models](#) with a lot of data conditioning

Cinematch scores an RMSE of 0.9514 on the quiz data, roughly a 10% improvement

In order to win the grand prize of \$1,000,000, a participating team had to improve this by another 10%, to achieve 0.8572 on the test set. Such an improvement on the quiz set corresponds to an RMSE of 0.8563.

2006 2007 2008 2009 2010

- fermata per problemi di privacy

Parti principali di un'hackaton

- Titolo e argomento
 - Tempi
 - Metriche
 - I dataset
 - Scelta del modello e competenze di settore
 - Teamwork
-

Titolo e argomento della Hackaton

Tempi

1 o 2 mesi

durante la previsione covid facevamo hackaton a tappe forzate di 1 settimana

Evil Corp 2 settimane ma vedrete che riuscirà a ridurre i tempi all'inverosimile

Metriche

Distinzione tra LOSS e Metrica!!

- La Loss è ciò che minimizzo per addestrare il modello
- La metrica è il mio indicatore di performance finale

Tutti gli algoritmi di machine learning hanno comunemente come funzione obiettivo la minimizzazione dello scarto quadratico medio quindi la Loss è progettata per minimizzare l'RMSE

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

Però spesso l'RMSE non è interessante ai fini pratici e come indicatore finale "adimensionale" e quindi uguale per ogni tipo di task si usano metriche diverse dall'RMSE ad esempio R^2 che abbiamo già visto.

Come faccio se la gara mi premia per la minimizzazione di una metrica diversa dall'RMSE?

Un esempio tipico è quando siamo interessati a grandezze relative

- riscrivere LOSS
- trasformare il target
- usare la classica LOSS RMSE ma validare usando l'indicatore giusto

Cosa fa Evil Corp con le metriche?

Non ti da la formula della metrica giusta ma ti dice a parole che è la differenza tra la metrica relativa e la correlazione. Perché a senso la correlazione tra risultati veri e predetti cresce tanto più sono simili le grandezze e lo scarto relativo decresce su dati corretti.

Ma relativo a cosa?

e QUALE CORRELAZIONE se un Data Scientist conosce per lo meno tre tipi diversi di correlazione?

Fisher, Spearman o Matthews?

Evil Corp non lo dice

Nota interessante

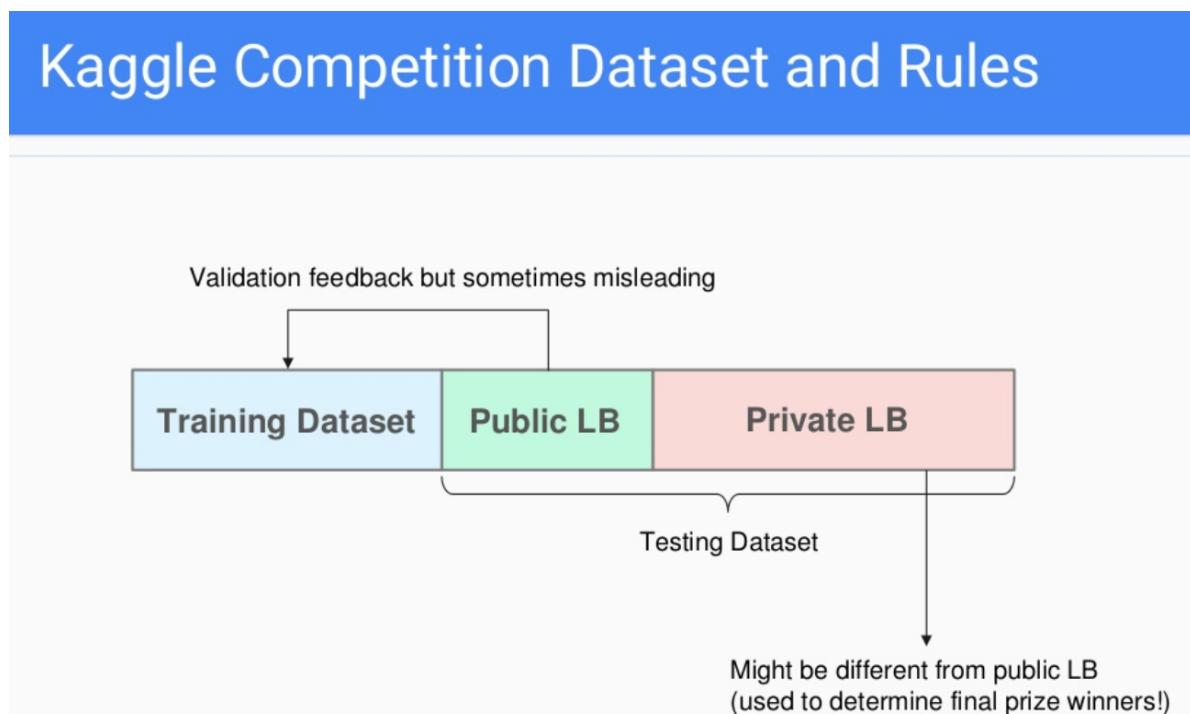
La correlazione di fisher da sola non è una buona LOSS. Essendo invariante per scala se un modello ha rispetto ai dati sperimentali correlazione massima di 1 allora anche un modello che darà risultati esattamente il doppio (o la metà) avrà correlazione massima di 1.

I dataset

Train

Validation

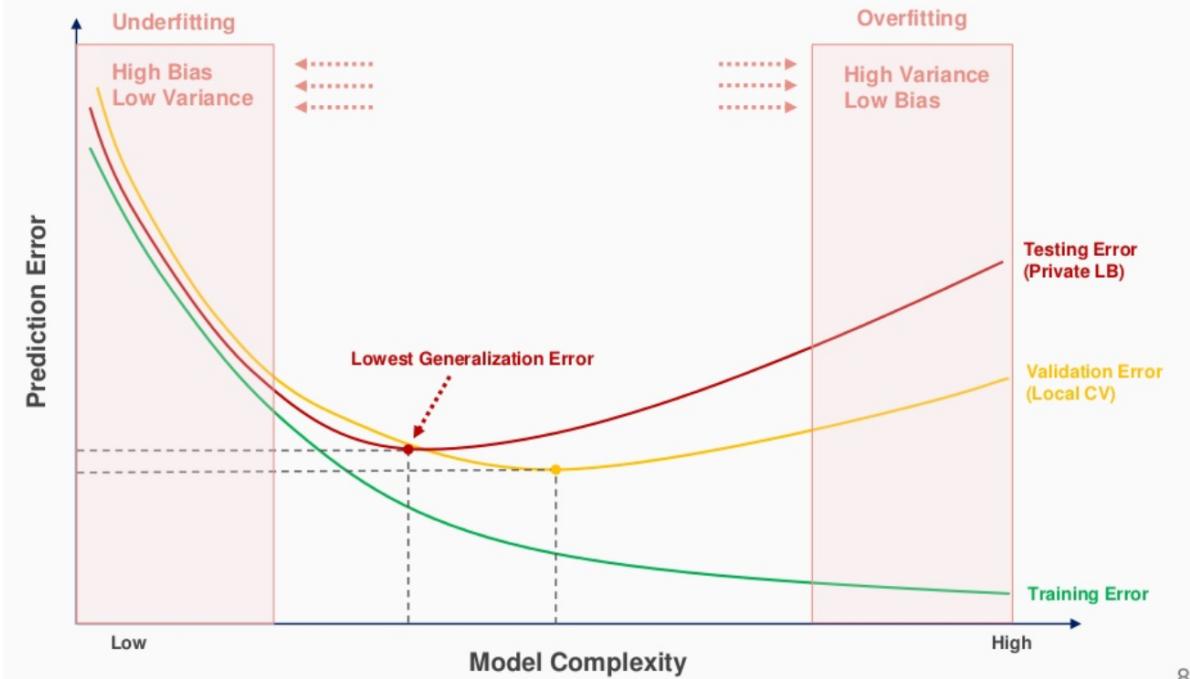
Test



<https://www.slideshare.net/markpeng/general-tips-for-participating-kaggle-competitions>

Underfitting and Overfitting

We want to find a model with lowest generalization error (hopefully)



Good Corp Da Train (x, y) "Test(x ,)" in due parti" non sappiamo come sono miscelate

addestramo il modello su Train (x, y) e uploadiamo i risultati che il modello da su Test(x ,)

su alcuni di questi dati Good Corp valuta la metrica per una leaderboard pubblica

su altri valuta la metrica per una leaderboard privata

il problema intrinseco è l'overfit della leaderboard pubblica tipicamente si limita numero di post giornalieri.. ma in ogni caso vedere continuamente il feedback rovina l'autenticità del lavoro di DS
posti i tuoi risultati

Cosa fa Evil corp con i dataset?

Ti da Train(x, y) e Validation(x, y) dicendo che verificherà periodicamente Validation(x, y) ma di Validation abbiamo già y ... quindi a cosa serve? A indovinare le vere metriche di cui non ti ha dato le formule?

Inoltre non da Test(x ,)

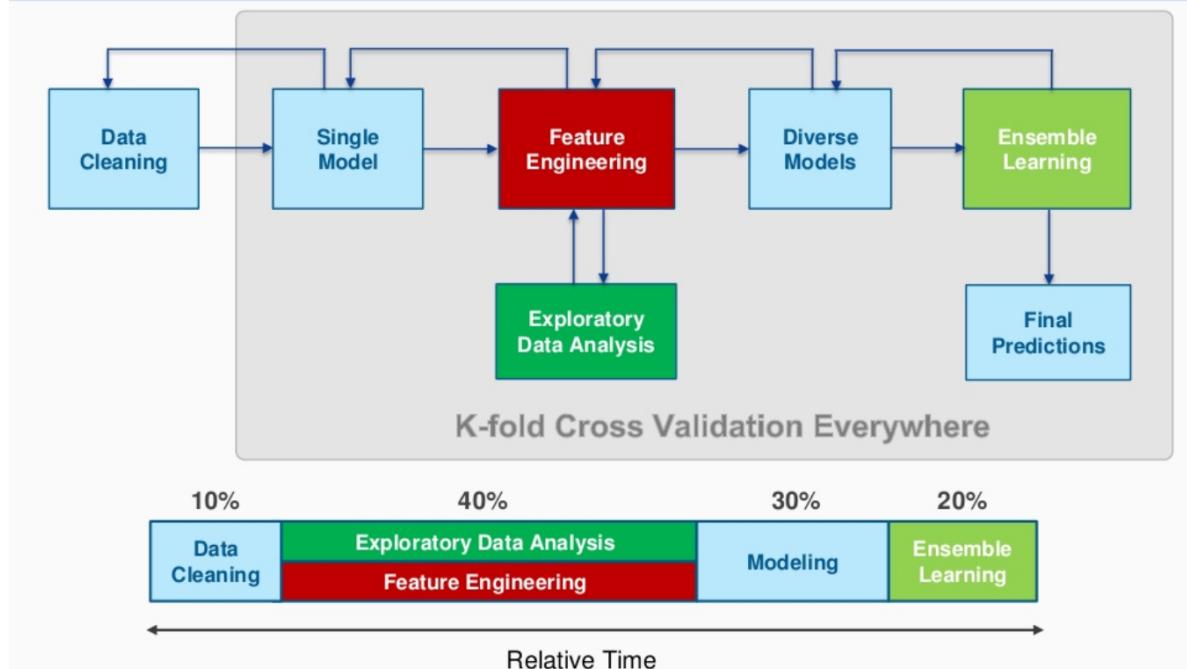
La gara c'è dal 1 maggio al 15 maggio stanco di aspettare il 9 maggio faccio presente che manca il Test(x ,) mi rispondono che vaneggio e poi misteriosamente l'11 maggio viene uploadato il Test(x ,)

Il 13 maggio ci viene chiesto come mai nessuno ha effettuato un post (possibile? in 13 giorni?) io faccio notare che i dati da loro postati l'11 maggio in file tar.gz da 7 giga sono danneggiati

Il 15 maggio alle 23:05 annunciano che "latitudine" e "longitudine" sono scambiati quindi bisogna invertire le due colonne

Scelta del modello e competenze di settore

Recommended Data Science Process (IMHO)



Nella scelta del modello in una competition seria la prima cosa che si va a vedere è un po' di letteratura sull'argomento.

Nel nostro caso bisognava vedere a partire da letture di immagine del satellite Sentinel 2 l'inquinamento atmosferico a terra misurato da stazioni AERONET.

Should I Trust Public LB?

- Yes if you can find a *K*-fold CV that follows the same trend with public LB
 - High positive correlation between local CV and public LB
 - The score increases and decreases in both local CV and public LB
- Trust more in your local CV!

Teamwork

The Advantages of Team Up

- Fewer work loads for each one if divides up the work well
 - A focuses on feature engineering
 - B focuses on ensemble learning
 - C focuses on blending submissions
 -
- Each member can contribute some single models for blending and stacking



Data Leakage e Adversarial Features

I'unico insegnamento dall competition di Evil Corp

- Data Leakage: informazioni del test set che filtrano nel training set

Feature Extraction: Hidden Features

- Sometimes there has some information leakage in the dataset provided by Kaggle competition
 - Timestamp information of data files
 - Some inadvertently left meta-data inside HTML or text
- May lead to unfair results, so normally I skip this kind of competitions!

Tutti i dati della competition di Evil Corp esatti potevano essere scaricati dal sito di AERONET

- Adversarial Features (dal libro Explainable AI with Python di Gianfagna e Di Cecco): features che forniscono alte prestazioni ma non sono affatto robuste spesso sono inaspettate e incontrollabili

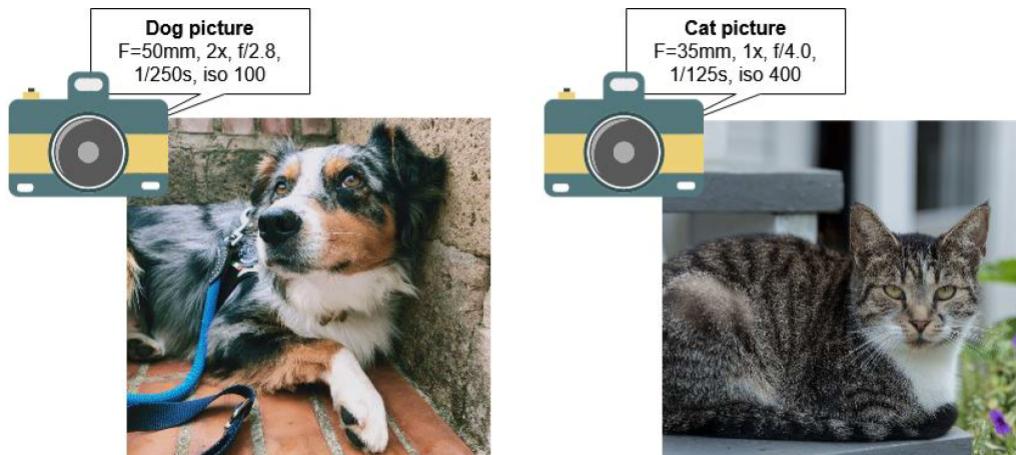


Figure 32 Adversarial Features in images. Courtesy of L.Bottou for the design.

Le immagini sono sporadiche tipo una volta al giorno le misure a terra molto più frequenti quindi le immagini danno un'informazione veramente molto grossolana del fenomeno.

Nel test 19 immagini su una stazione per 1800 misure di inquinamento (considerato che da un'immagine era necessario estrarre il dato del solo pixel sull'immagine).

Il dato X da satellite era ulteriormente ridicolizzato dalla presenza di nuvole nel 50% delle immagini.

Per migliorare la performance aggiungo variabili temporali...

Lo tratto come serie storica...

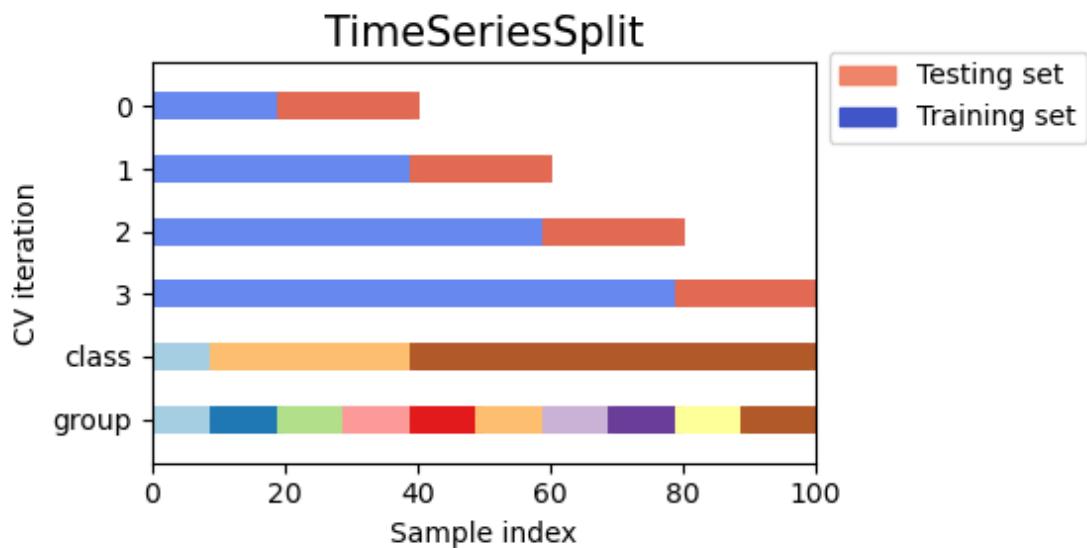
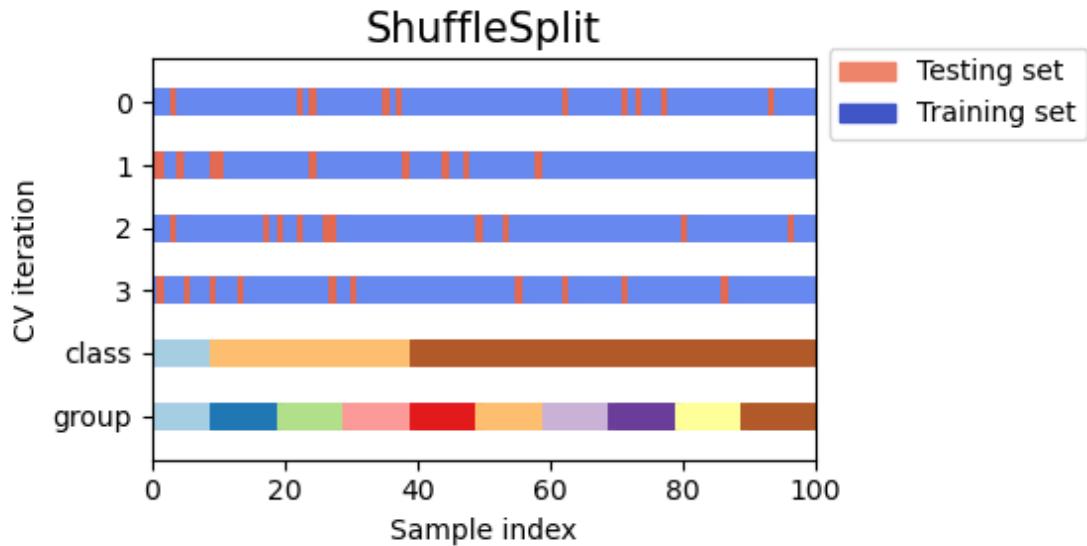
BEWARE se uso direttamente il tempo in una serie storica **overfittato** i dati quindi devo usare features derivate e.g. risultato giorno precedente medie storiche

e in questo caso non posso farlo perché i dati non sono causali ma miscelati

Posso al limite usare giorni della settimana/ mese/ anno / ora come feature cercando di modellare dipendenze ricorrenti

MA il TEST set è miscelato con il TRAIN e VALIDATION set quindi si ottengono i risultati migliori tramite interpolazione che con un modello storico

quindi il risultato migliore per vincere questa hackaton è un modello assolutamente inutile ...



Le nuvole... dati spuri / denoising / pseudo labelling

"Apriti cielo..."

- 50% dei dati non presenti causa nubi ----> il modello usa le variabili temporali e non da satellite ----> è ancora più fragile
- molte immagini non labellate e “immagini vuote” un pixel per vedere la stazione su oltre un milione di pixel di immagine

Cosa ci posso fare?

- denoising e riduzione dimensionale (ad esempio con PCA)
 - pseudo labelling: addestro il modello sulle label che conosco
 - lo applico ai dati di cui non ho label
 - riaddestro con i dati labellati veri + pseudo labellati
 - almeno in teoria in pratica bastava un modello lineare a risolvere la Evil Competition in modo onesto
-

Facciamo una gara noi

Al Fish Market di Tokyo...



InClass Prediction Competition

SOAI21 fish regression

Studenti di SOAI dovete prevedere quanto pesano i pesci del mercato di Tokyo

a month to go

Overview Data Code Discussion Leaderboard Rules Team Host My Submissions Submit Predictions

Overview Edit

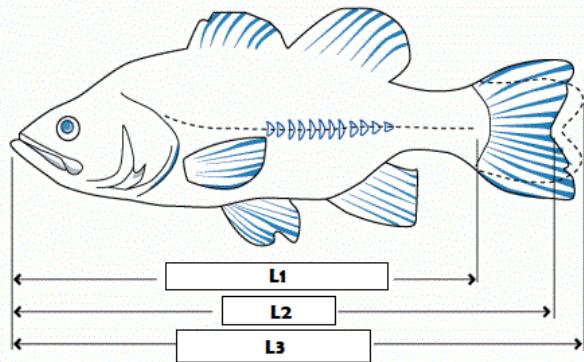
Description Questa è una competition privata di School of AI Italia

Evaluation Al mercato del pesce di Tokyo abbiamo misurato il peso di pesci di diverse specie. Il vostro compito è stimare con la maggiore precisione possibile il peso dei pesci.

+ Add Page La metrica usata per la competizione è il classico RMSE. Potete unirvi in team e fare fino a 4 tentativi al giorno. La competizione terminerà il 3 luglio 2021.

<https://www.kaggle.com/c/soai21-fish-regression>

Measuring Fish Length



https://en.wikipedia.org/wiki/Standard_weight_in_fish