Jak wygrać konkurs Kaggle

czyli o sposobach poprawy jakości modelu predykcyjnego, który (wydaje się, że) jest już najlepszy

Stanisław Kaźmierczak

Agenda

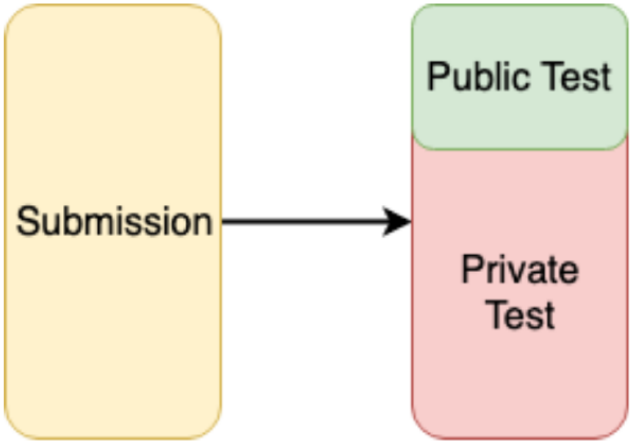
1. Motywacja
2. Zasady Kaggle
3. Inżynieria cech
4. Modelowanie
5. Metody optymalizacji
6. Case
7. Sztuczki

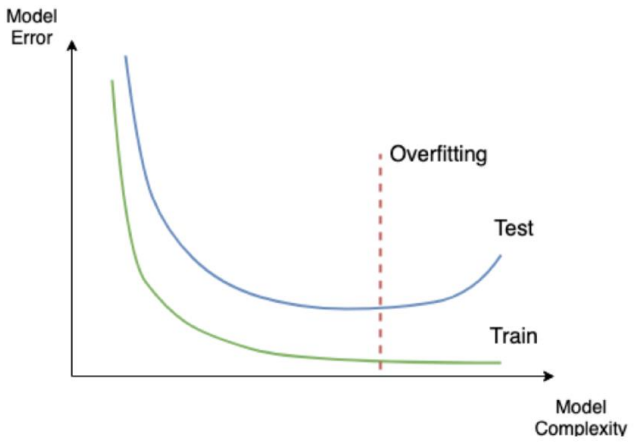
Motywacja

- Wbrew tytułowi, główną motywacją nie jest zwycięstwo w konkursie Kaggle
  - Choć nagrody bywają wysokie…

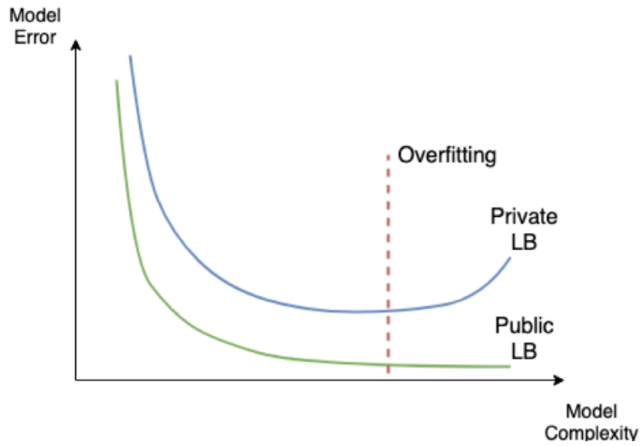
- Budowa wysokiej jakości modelu na konkurs Kaggle stanowi analogię do budowy wysokiej jakości modelu działającego na danych produkcyjnych (np. danych, które dopiero się pojawią), których poprawnych wartości nie znamy na etapie tworzenia.
- Od niewielkiej poprawy jakości modelu (wyrażonej np. w ułamku procenta badanej metryki) mogą zależeć zdrowie/życie konkretnych ludzi lub bardzo duże pieniądze.

Tablica wyników



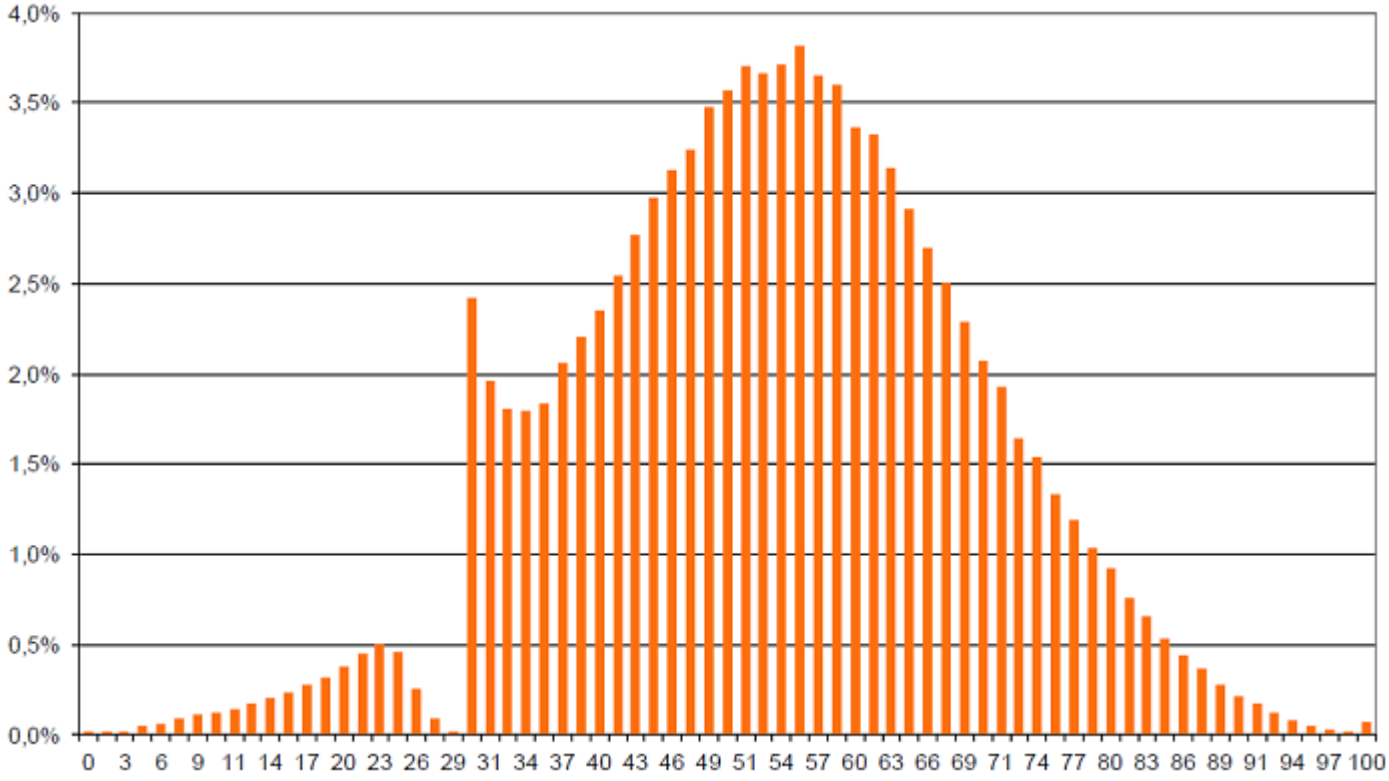


[5]



shake-up – różnica pomiędzy miejscem w rankingu publicznym i prywatnym.

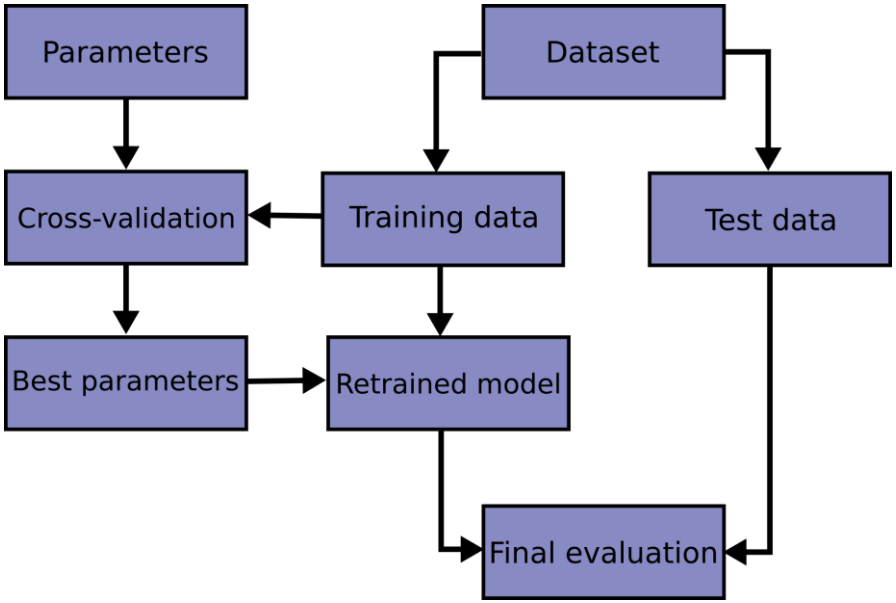
Eksploracyjna analiza danych

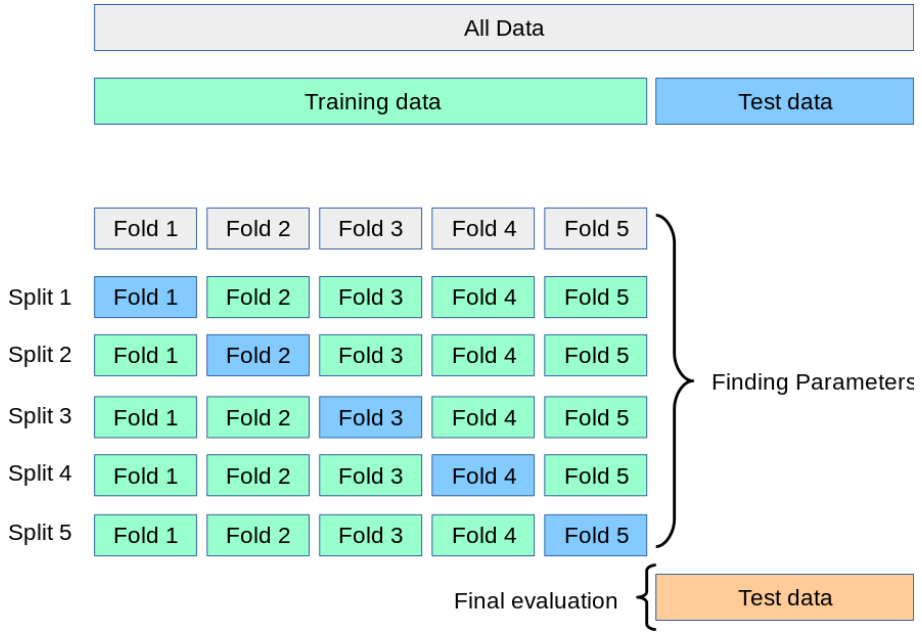


[11]

Standardowa kroswalidacja

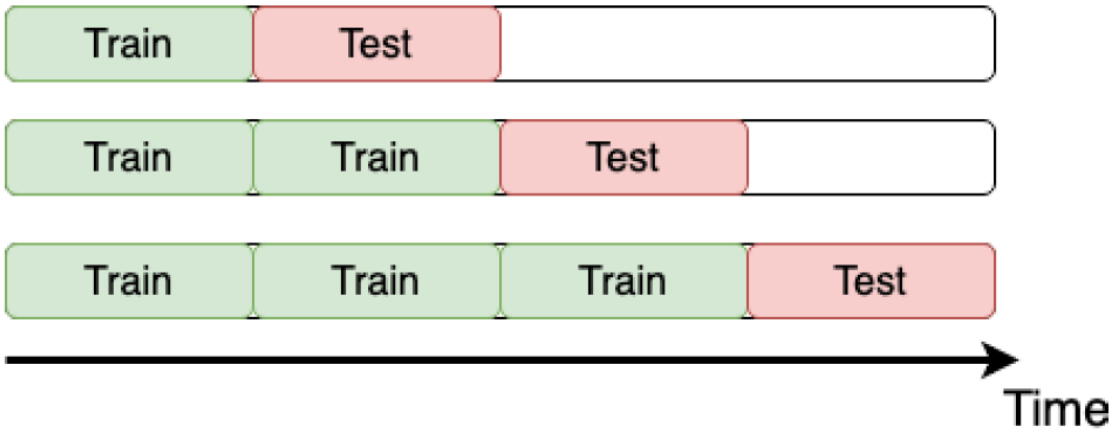
Kroswalidacja zamiast ciągłego wysyłania rozwiązań.





[6]

Kroswalidacja dla szeregów czasowych



[5]

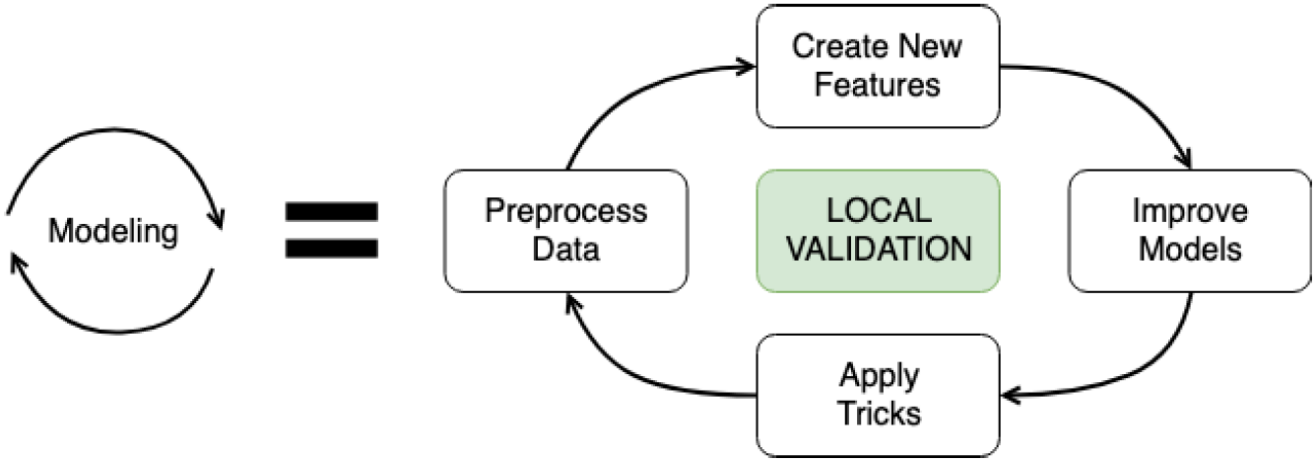
Kroswalidacja – interpretacja wyników

```
# Overall validation score
overall_score_minimizing = np.mean(fold_metrics) + np.std(fold_metrics)
# Or
overall_score_maximizing = np.mean(fold_metrics) - np.std(fold_metrics)
```

Fold number	Model A MSE	Model B MSE
Fold 1	2.95	2.97
Fold 2	2.84	2.45
Fold 3	2.62	2.73
Fold 4	2.79	2.83
Mean	**2.80**	**2.75**
Overall	**2.919**	**2.935**

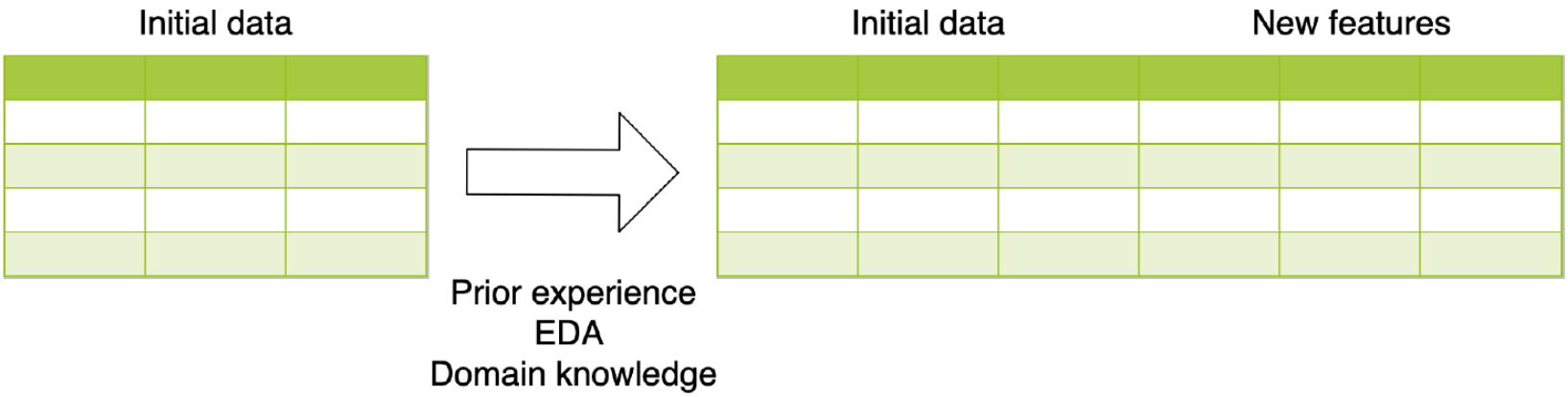
[5]

Modelowanie

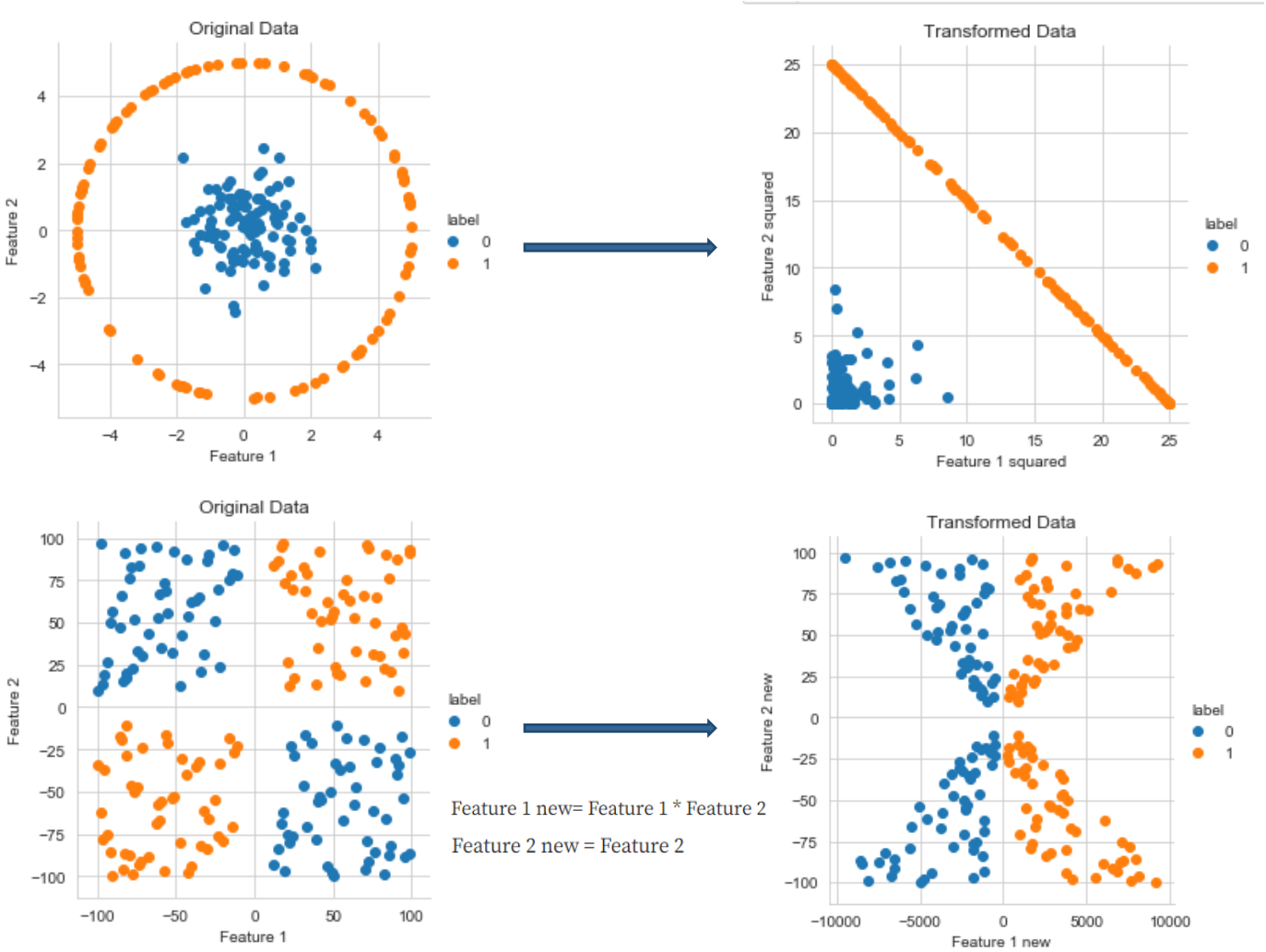


- W praktyce na danym zbiorze większość metod nie poprawia rezultatu.
- Znajomość szerokiego wachlarza technik zwiększa szansę na poprawę wyników.
- Ze względu na wysoką wymiarowość problemu optymalizacyjnego z reguły jednocześnie sprawdzamy jedną, maksimum dwie metody

Inżynieria cech (1)



Inżynieria cech (2)



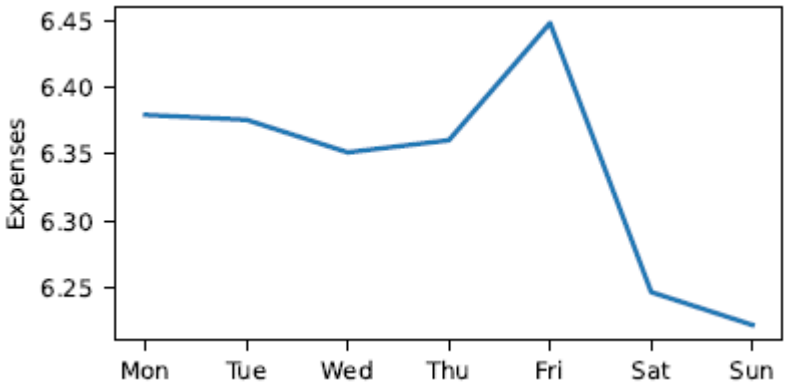
[7]

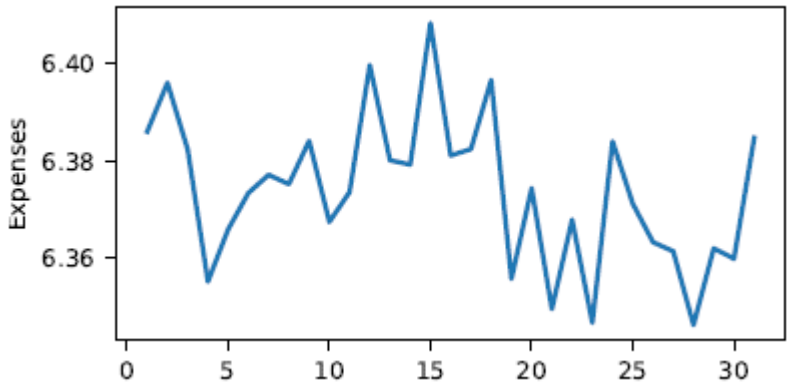
Dodatkowe cechy numeryczne

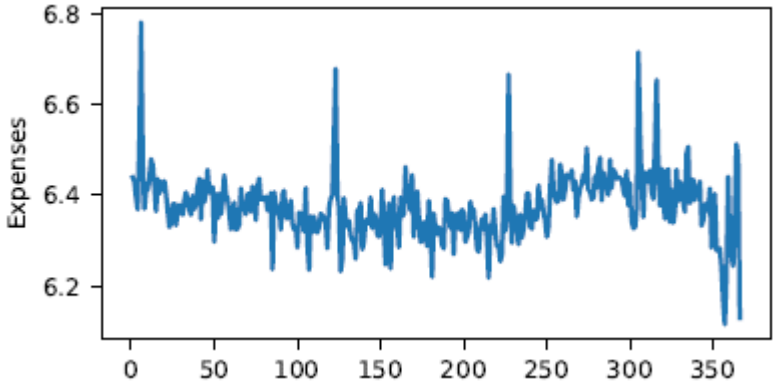
1. **Polynomial Transformations :** f1**2, f1*f2 , f1+f2 , 3*f1+2f2 etc.,
2. **Trigonometric Transformations:** sin(f1), tanh(f2**2), cot(f1/f2), etc.,
3. **Boolean Transformations:** AND, OR, NOT, NAND, XOR, etc.,
4. **Logarithmic Transformations:** log(f1), log(f1*f2), etc.,
5. **Exponential Transformations:** pow(e,f1), pow(e, 2*f2), etc.,

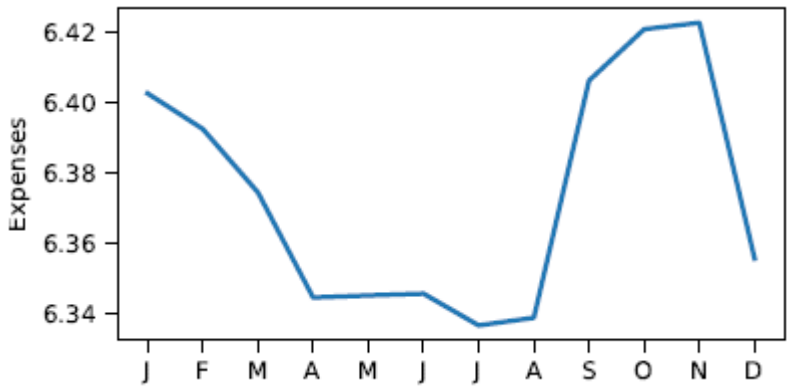
[7]

Cechy data/czas

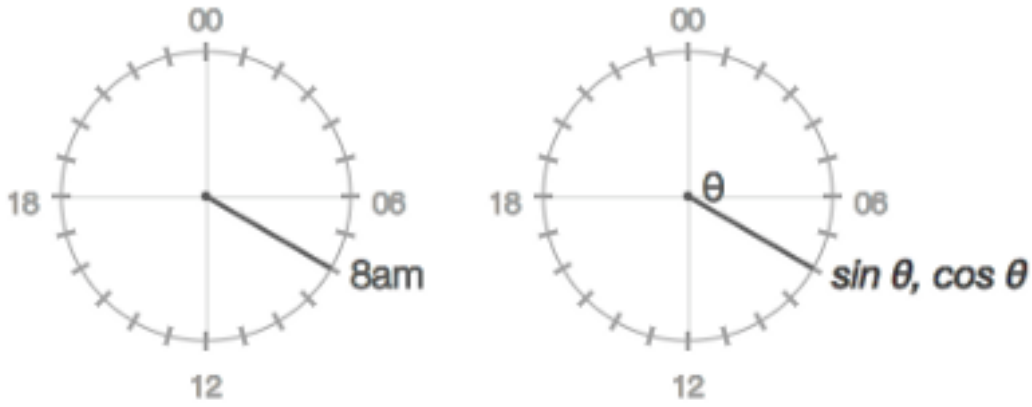








Cechy cykliczne



[8]

A co jeśli organizujemy mundial?

Cechy kategoryczne

ID	Categorical feature
1	A
2	B
3	C
4	A
5	D
6	A

ID	Label-encoded
1	0
2	1
3	2
4	0
5	3
6	0

ID	Cat == A	Cat == B	Cat == C	Cat == D
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	1	0	0	0
5	0	0	0	1
6	1	0	0	0

- Label/order encoding tworzy hierarchię kategorii
- Dla modeli drzewiastych jest to bez znaczenia, dla modeli liniowych ma istotne znaczenie
- Daną cecha może być różnie kodowana w zależności od zadania predykcyjnego
- Nie zawsze sposób kodowania jest oczywisty

Cechy kategoryczne – inne podejścia

- Backward Difference Coding
- BaseN
- Binary
- CatBoost Encoder
- Hashing
- Helmert Coding
- James-Stein Encoder
- Leave One Out
- M-estimate
- One Hot
- Ordinal
- Polynomial Coding
- Sum Coding
- Target Encoder
- Weight of Evidence

[5]

Brakujące dane

- Trzeba zaadresować problem, to nie jest tylko kwestia jakości modelu
- Tradeoff pomiędzy szumem i mniejszą ilością danych
- Korelacja na poziomie kolumn/wierszy
- Cechy numeryczne
  - uzupełnianie średnią/medianą
  - uzupełnianie stałą (modele drzewiaste)
- Cechy kategoryczne
  - Najczęstsza wartość
  - Nowa kategoria
- Dodatkowa kolumna z informacją, gdzie dane były imputowane
- Uzupełnianie kontekstowe (*fancyimpute*) – idea z algorytmu kNN

Modelowanie

- Gradient boosting jest elementem zwycięskiego rozwiązania w większości zadań z danymi tabelarycznymi
  - Uwaga na możliwość przeuczenia zbyt dużą liczbą drzew
- Las losowy daje z reguły gorsze rezultaty, ale brak możliwości przeuczenia (w aspekcie liczby drzew)
  - Dobry uniwersalny model
- Ze względu na limity zgłoszenie następuje z reguły po kilku eksperymentach z lokalną kroswalidacją.

Competition type	Feature engineering	Hyperparameter optimization
Classic Machine Learning	+++	+
Deep Learning	-	+++

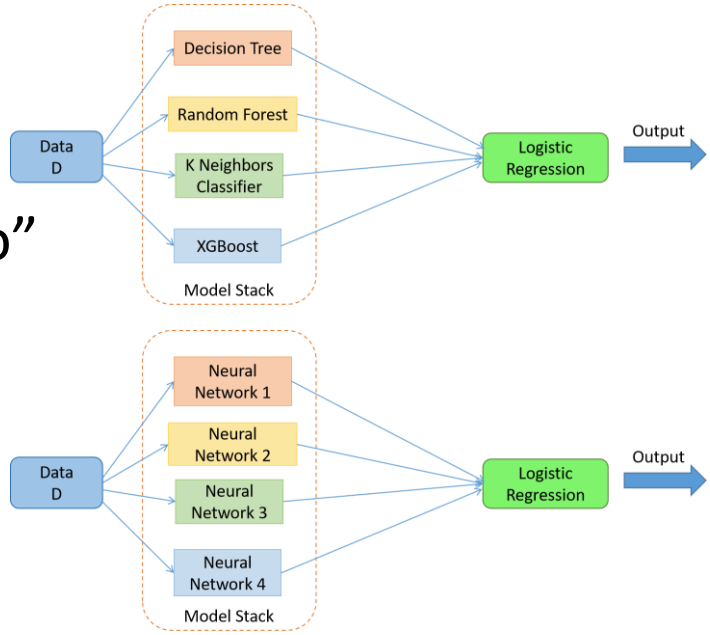
[5]

Optymalizacja hiperparametrów

- Grid search
  - Wszystkie konfiguracje
- Randomized grid search
  - Gdy przestrzeń konfiguracji jest duża
- Optymalizacja bayesowska
  - Uwzględnia dotychczasowe ewaluacje
- Auto ML

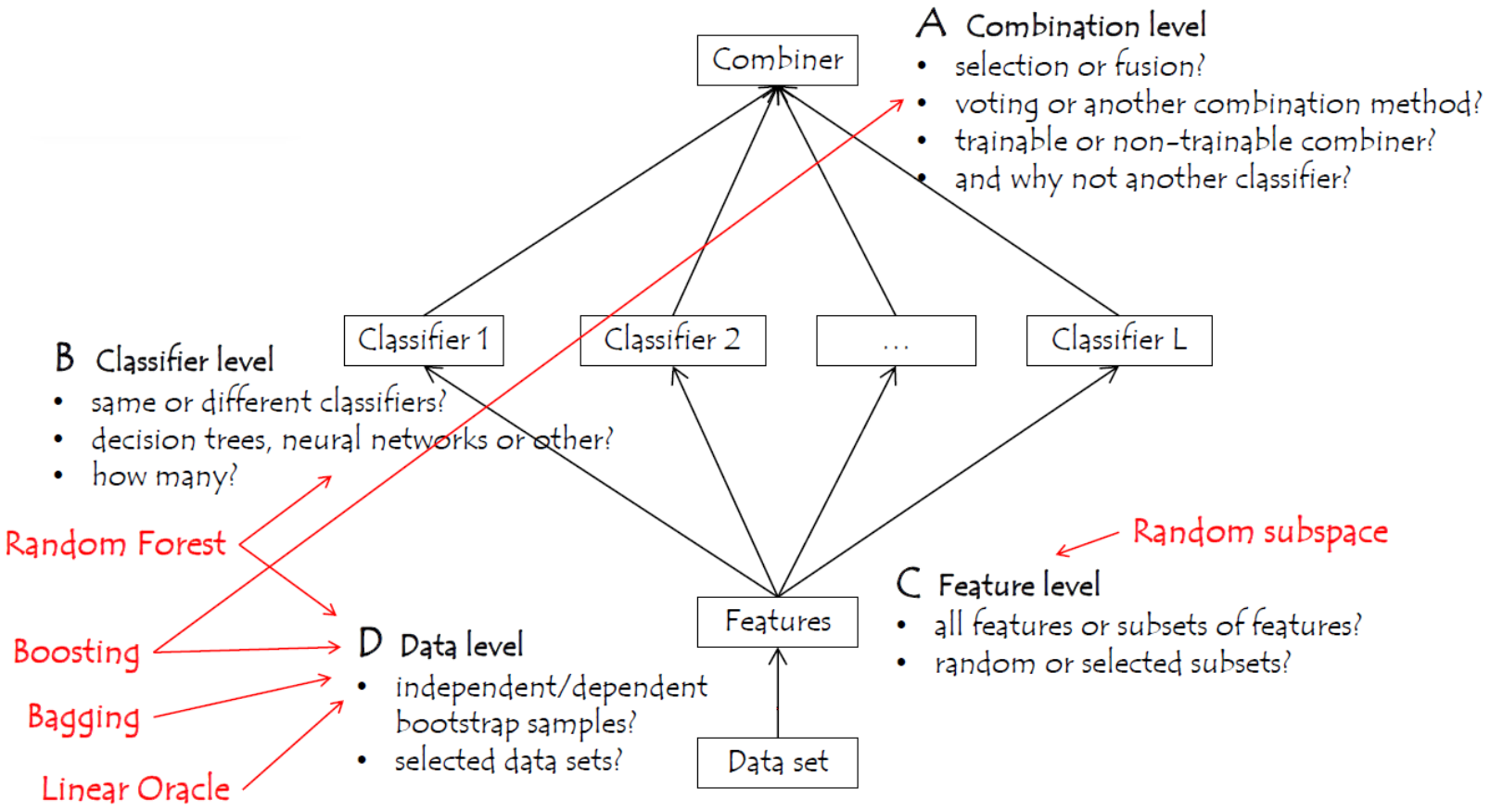
Ensembling (1)

- Blending
  - Klasyfikacja: Hard/soft voting
  - Regresja: uśrednianie
  - Wersje ważone
  - Tani w kontekście nakładu pracy
  - Można stosować niemalże „w ciemno”
- Stacking
- Główne czynniki wpływające na wynik:
  - Siła modeli składowych
  - Ich zróżnicowanie



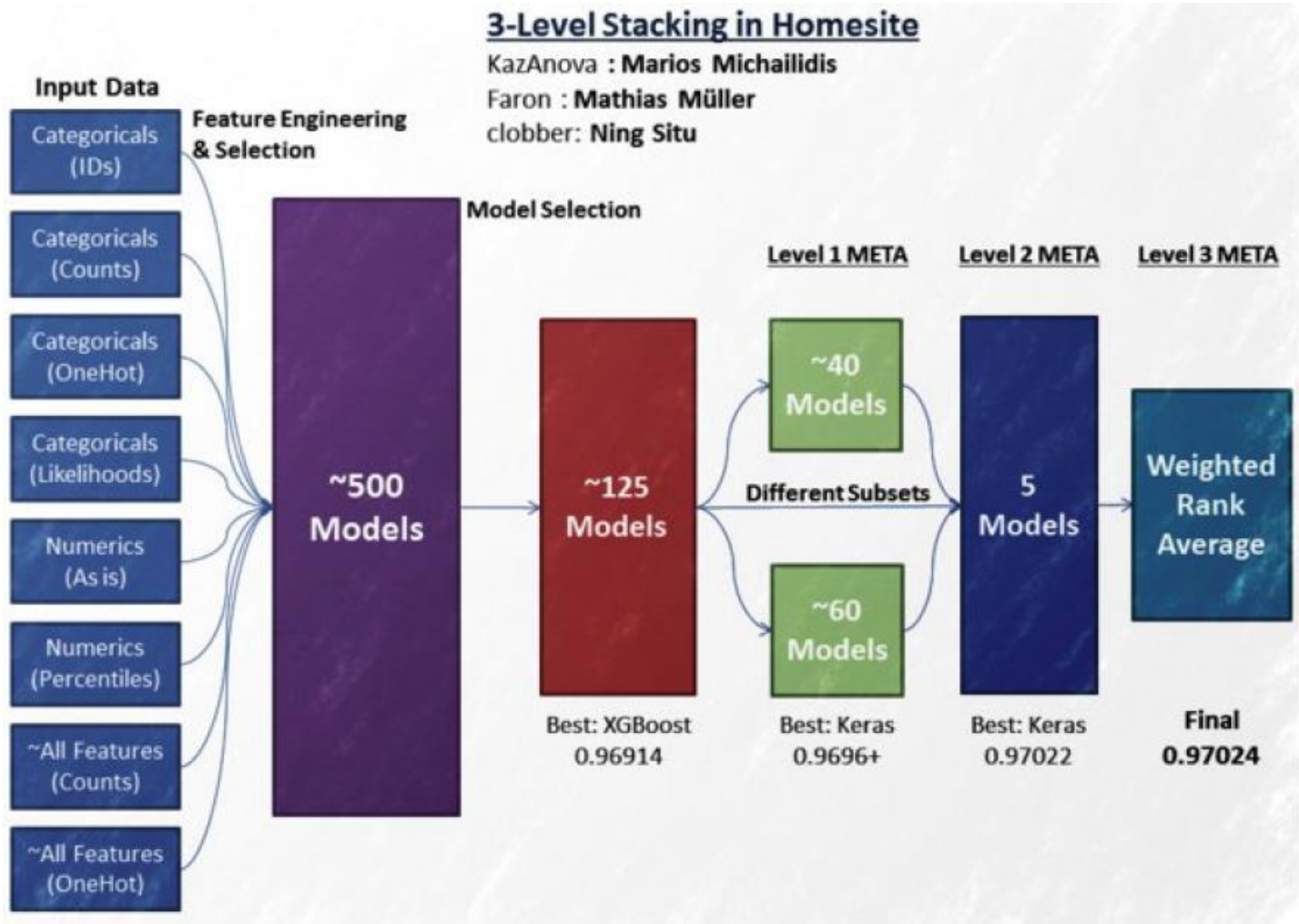
[9]

Ensembling (2)



[10]

Homesite Quote Conversion Challenge



Niebezpieczeństwo przeuczenia – case (1)

- Konkurs badający poziom psychopatii użytkowników Twittera [1]
- Ewaluacja na podstawie średniej precyzji (zakres wartości [0, 1])
- Publiczna i prywatna tablica wyników
- Możliwość wybrania 5 wyników do finalnej ewaluacji
- Uczestnik wykonał 42 zgłoszenia [2].
- Na koniec konkursu uczestnik zajmował drugie miejsce (na 111 uczestników) na publicznej tablicy wyników.

Niebezpieczeństwo przeuczenia – case (2)

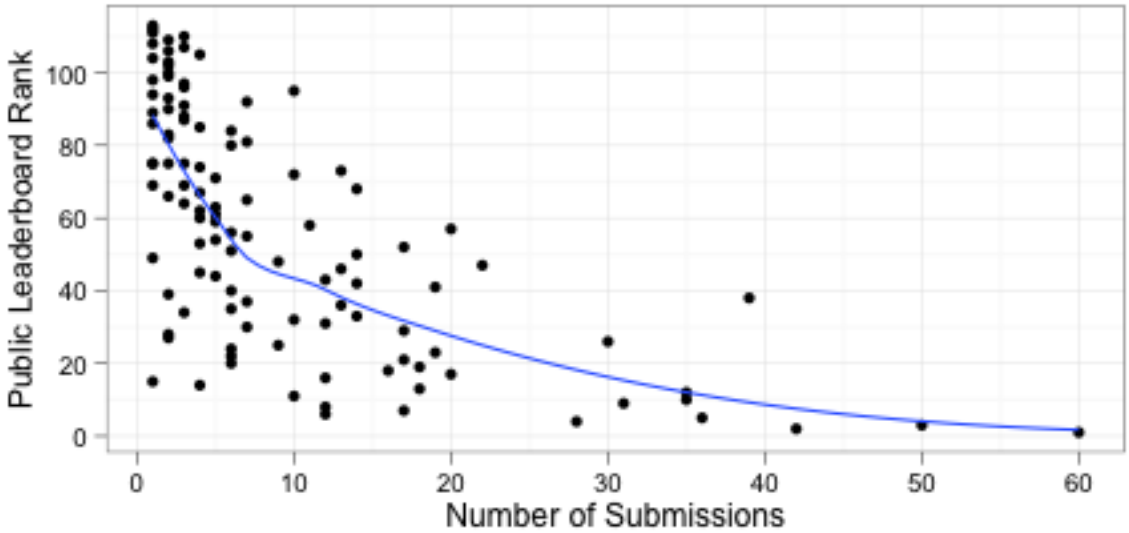
- W ostatecznej klasyfikacji uczestnik spadł z 2. na 52. miejsce.
- Znalazł się m.in. za benchmarkowym lasem losowym – rozwiązaniem publicznie dostępnym od początku konkursu.
- Pierwsza piątka publicznej tablicy wyników znalazła się ostatecznie na odpowiednio 64., 52., 58., 16., 57. miejscach.

	↓8	Random Forest Benchmark	0.86141		
45	↓8	dickoa	0.86141	1	Tue, 22 May 2012 12:07:36
45	↓8	Rohit	0.86141	2	Fri, 25 May 2012 21:00:14
45	↓8	squawkboxed	0.86141	1	Fri, 08 Jun 2012 14:57:28
45	new	BLetson	0.86141	3	Fri, 29 Jun 2012 14:49:38
50	↓9	testing	0.86135	4	Sat, 16 Jun 2012 05:18:44 (-26.1h)
51	↓9	schappi	0.86130	7	Sat, 16 Jun 2012 12:53:13
52	↓8	Greg Park	0.86116	42	Fri, 29 Jun 2012 01:08:38 (-14.1d)
53	↓8	Glen	0.86111	35	Tue, 05 Jun 2012 23:44:06 (-3.3d)

Niebezpieczeństwo przeuczenia – case (3)

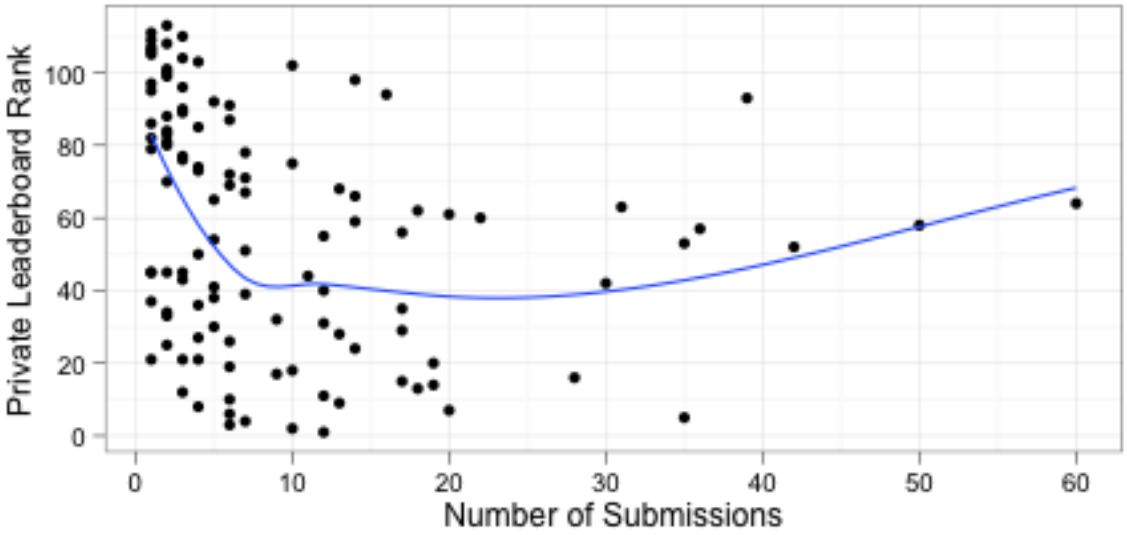
Publiczna tablica wyników

- Pierwsza piątka ma dużo zgłoszeń
- Więcej zgłoszeń przekłada się na lepszy wynik
  - Publiczna tablica wyników odzwierciedla po części podejście siłowe wynikające z liczby prób, a nie rzeczywistą zdolność do generalizacji



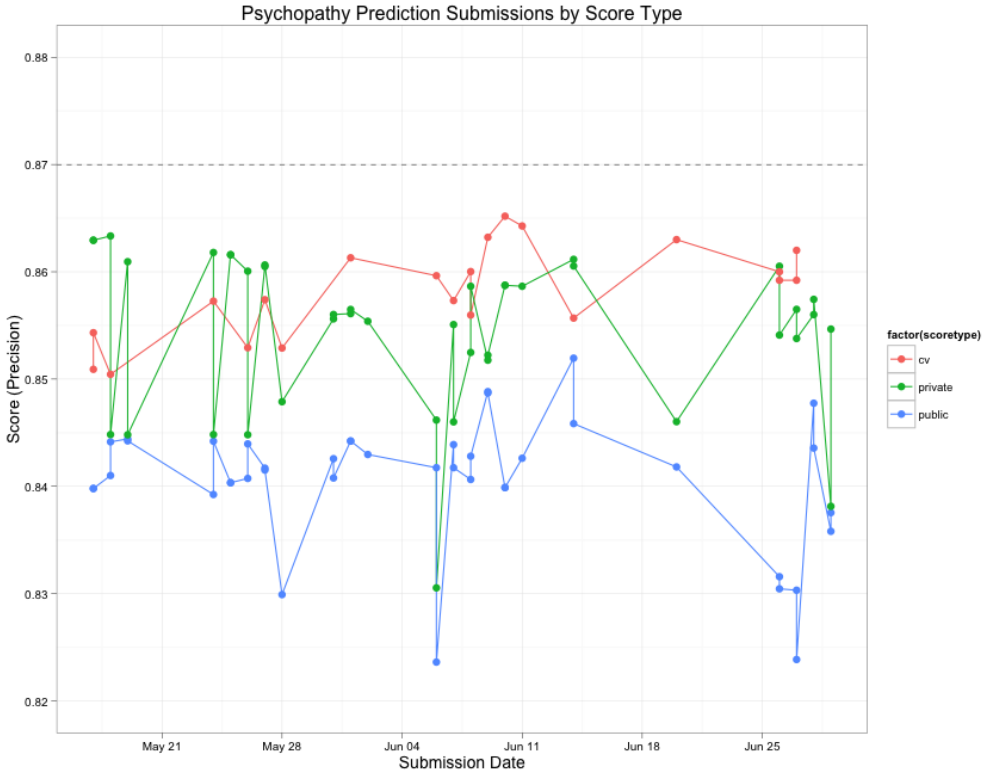
Niebezpieczeństwo przeuczenia – case (4)

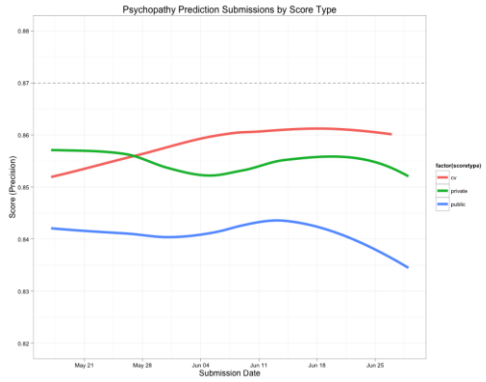
Prywatna tablica wyników



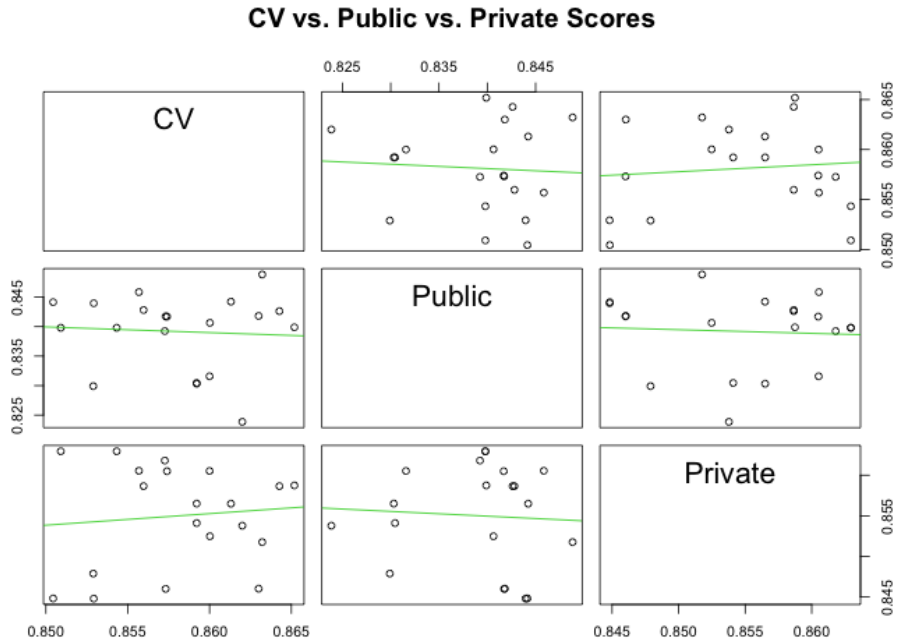
Niebezpieczeństwo przeuczenia – case (5)

- Wielu uczestników przed wykonaniem zgłoszenia używa kroswalidacji
  - Cel: niemarnowanie zgłoszenia oraz dodatkowa weryfikacja
- Wynik na kroswalidacji generalnie stopniowo się poprawia
- Wynik prywatny się pogarsza (dwa pierwsze zgłoszenia są najlepsze ze wszystkich)
- Wynik publiczny osiąga szczyt, a potem się pogarsza
- Nie widać jasnej zależności między powyższymi wynikami



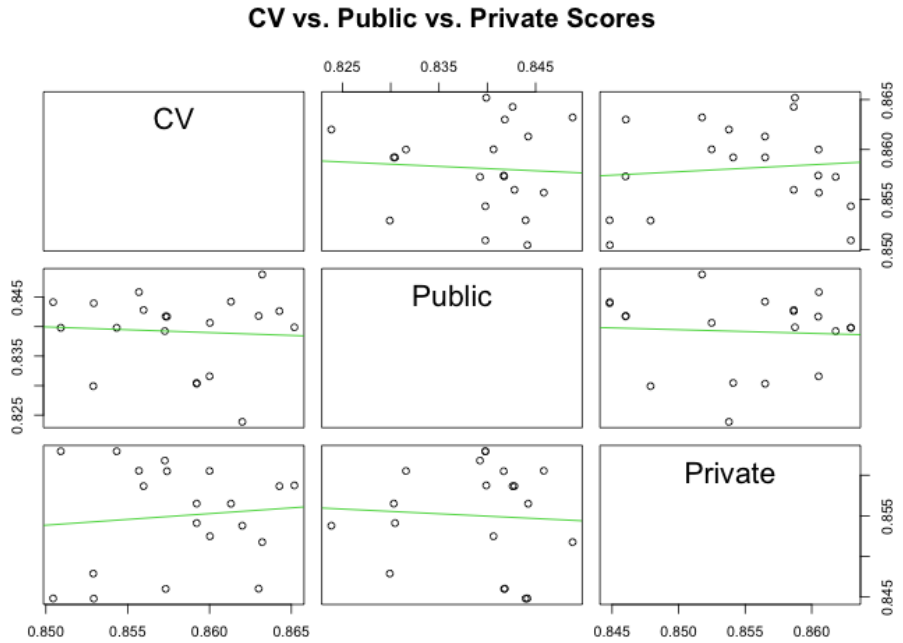


Niebezpieczeństwo przeuczenia – case (6)



Kroswalidacja nie jest skorelowana z ostatecznym wynikiem. Dlaczego?

Niebezpieczeństwo przeuczenia – case (6)



Kroswalidacja nie jest skorelowana z ostatecznym wynikiem. Dlaczego?

Została metodologicznie błędnie wykonana. Problem ten dotyczy wielu publikacji w prestiżowych czasopismach [3].

Niebezpieczeństwo przeuczenia – case (7)

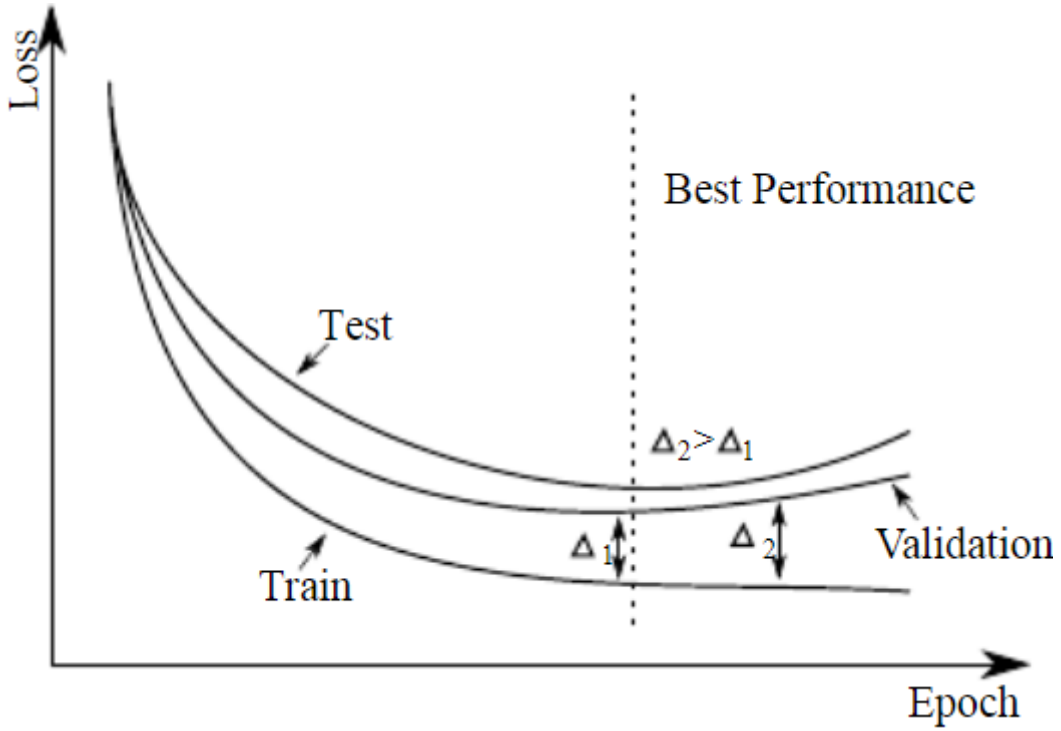
Źle

1. Screen the predictors: find a subset of "good" predictors that show fairly strong (univariate) correlation with the class labels
2. Using just this subset of predictors, build a multivariate classifier.
3. Use cross-validation to estimate the unknown tuning parameters and to estimate the prediction error of the final model.

Dobrze

1. Divide the samples into K cross-validation folds (groups) at random.
2. For each fold $k = 1, 2, \ldots, K$
   (a) Find a subset of "good" predictors that show fairly strong (univariate) correlation with the class labels, using all of the samples except those in fold k.
   (b) Using just this subset of predictors, build a multivariate classifier, using all of the samples except those in fold k.
   (c) Use the classifier to predict the class labels for the samples in fold k.

Parametr *patience* w sieciach neuronowych



[4]

Inne wskazówki/metody (1)

- Finalne zgłoszenie: najlepszy wynik lokalny i najlepszy wynik w publicznej tablicy wyników
- Oversampling/augmentacja danych
- Postprocessing, np. adresowanie (concept/data drift)
- Sztuczki i *hakowanie* Kaggla
  - Jawne wprowadzanie w błąd na forum ☹
  - Wykorzystanie instancji testowych na etapie preprocessingu
  - Używanie zbioru testowego na etapie treningu (uczenie częściowo nadzorowane)
  - Szukanie instancji publicznych
    - Ustalanie ich etykiet i użycie do treningu
    - Podbicie wyniku publicznego celem zmylenia innych zawodników

Inne wskazówki/metody (2)

Rank	Team Name	Preliminary Score	Final Score	Submissions
1	v	0.7204	**0.694600**	188
2	kubapok	0.6919	**0.691500**	41
3	Stan	0.7086	**0.689400**	244
4	dragon	0.8035	**0.689200**	285
5	bottomline	0.5933	**0.593400**	12
6	baseline	0.5593	**0.564400**	4
7	HBKU AI	0.5514	**0.548200**	1
8	AMFAD	0.5380	**0.540900**	11
9	ML	0.5165	**0.515300**	29
10	amy	0.5117	**0.511600**	20

Showing 1 to 10 of 11 entries

Previous 1 2 Next

Źródła (1)

1. https://www.kaggle.com/competitions/twitter-psychopathy-prediction
2. https://gregpark.io/blog/Kaggle-Psychopathy-Postmortem/
3. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: Springer.
4. Salehinejad, H., Sankar, S., Barfett, J., Colak, E., & Valaee, S. (2017). Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078.*
5. Y. Babakhin, *Winning a Kaggle Competition in Python,* DataCamp course.
6. https://scikit-learn.org/stable/modules/cross_validation.html

Źródła (2)

7. https://gnana97.medium.com/importance-of-feature-engineering-in-machine-learning-and-deep-learning-a6ba2df3f0d4
8. https://medium.com/life-at-hopper/ai-in-travel-part-2-representing-cyclic-and-geographic-features-4ada33dd0b22
9. https://www.analyticsvidhya.com/blog/2021/08/ensemble-stacking-for-machine-learning-and-deep-learning/
10. Kuncheva, L. (2016). Getting Lost in the Wealth of Classifier Ensembles?. In *ICPRAM* (p. 7).
11. http://smarterpoland.pl/index.php/2013/03/czy-polonisci-sa-mniej-obiektywni-a-matematyki-jest-za-malo-w-liceum/


Q&A