MWS Cup 2022 課題3
Post Meeting

株式会社ＦＦＲＩセキュリティ
https://www.ffri.jp

作問メンバーについて

紹介

今年の作問体制

今年は以下のメンバーにご協力いただいた（敬称略）
あらためて感謝申し上げます

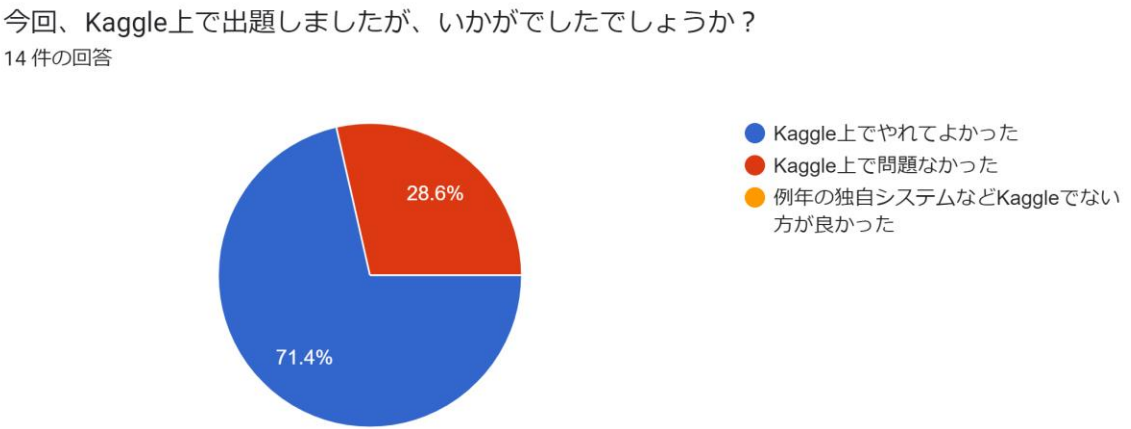
所属・氏名		役割
株式会社FFRIセキュリティ	茂木裕貴	作問チームリーダー
	中川恒	MWS Cup副委員長
LINE株式会社	愛甲健二	作問委員
	小野颯真	作問委員
	草間好輝	作問委員
	華徳凱	作問委員

今年の問題について

アンケート結果

Kaggleの利用は継続したい

Kaggleでやれてよかったという声が自由記述でも多かった
各チーム代表者1人制は不評だったため来年以降は複数人可にしたい



Stringsに絞ったことは良かった

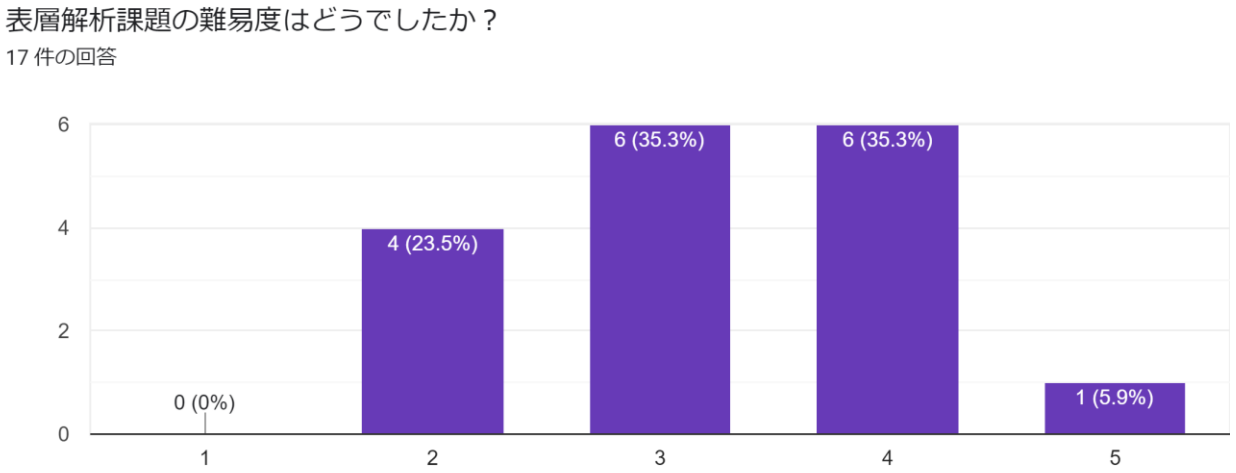
以下の理由からStringsに絞った検知にした

特徴	概要
試行錯誤しやすい	特徴量を絞ることで方向性が定まる NLPの各種手法を試せる
データ量を増やす	データサイズの削減により検体の件数 を昨年の約2倍に
論文に基づく設定	Stringsを用いた検知は複数の論文が 存在する

**取り組みやすい、試行錯誤できた
という声が多く、好評だった
来年以降は要検討**

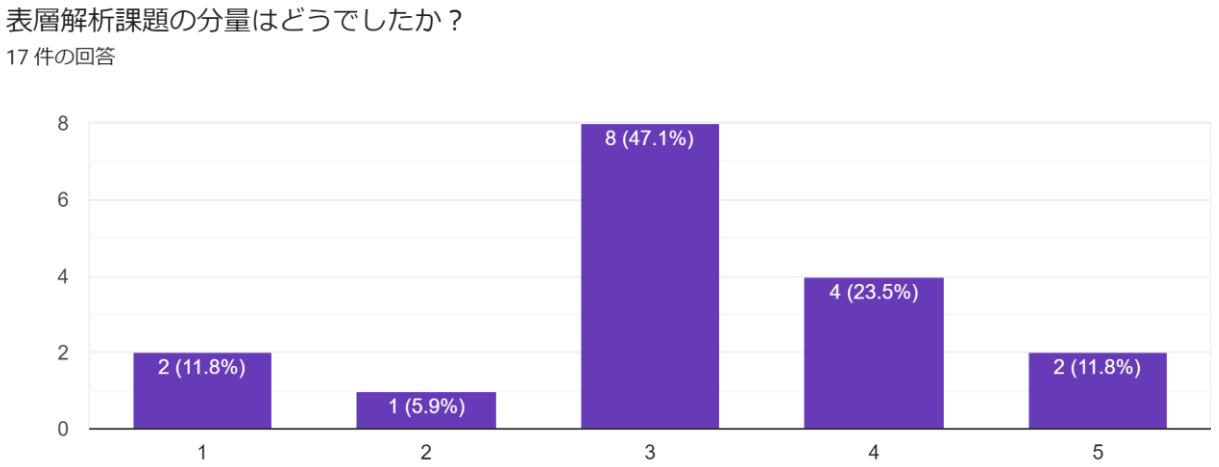
難易度はちょうど良かった

試行錯誤でスコアを上げられたという声が多かった



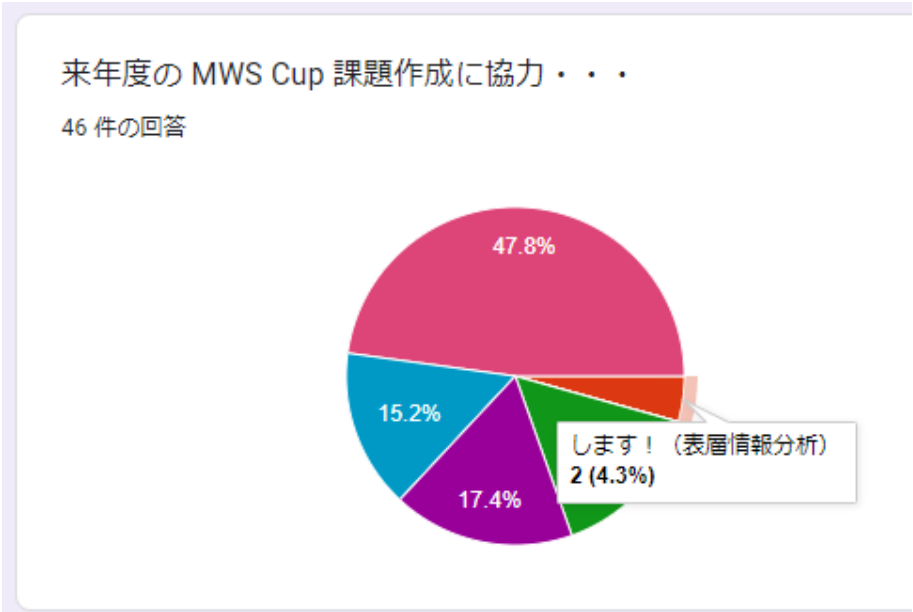
分量は少ないという声があった

「当日課題」がないことを惜しむ声もあった



一方で分量は解く人次第とも言える
当日課題については答えがない、
ので…

作問協力者募集中です



課題

Stringsは今年使ってしまったが、どうする？
Kaggle Notebookで動くレベルの模範といえる解答
答えが決まる一方でバラけて分類に貢献する当日課題

MWS Cup 2022 課題3

Write-up 解説

1st 解法のDiscussionが模範的

KOHs
Topic Author
1st place

1st place solution (public LB 2nd, private LB 1st)
Posted in mws-cup-2022-3 14 days ago
11

はじめに

チーム「we love 松尾」のマルウェア表層解析担当です。

非常に有意義で楽しいコンペを開催していただきありがとうございます！

僭越ながら、1st place solutionとして我々の解法を紹介します。何かの参考になれば幸いです。

アンサンブルの構築の関係で、多数のNotebookに分かれておりまして、分かりやすさの観点から、取り組んだ内容を以下にまとめることにしました。（誤字脱字などあったらすみません...）

設定

- seed値 50
- ハイパーパラメータチューニング時のseed値 1050

特徴量作成

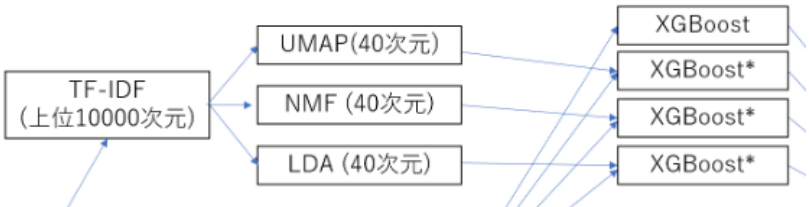
- TF-IDFで特徴抽出 → 上位10000次元を UMAP, NMF, LDAでそれぞれ40次元に次元圧縮
  - UMAPで圧縮した特徴量 → xgb01, NMFで圧縮した特徴量 → xgb02, LDAで圧縮した特徴量 → xgb03 のように次元削減手法ごとに別のモデルを作成した
- fexrdが提供していた特徴量 + 同じように追加した特徴量（文字列xxの数を数える系）

上記Discussion[1]を是非読んでほしいが、出てくる単語をいくつか抜粋してDSコンペに慣れていない人向けに解説する

embeddingの次元削減

TF-IDFで得られたベクトルを次元削減してLightGBM等への入力としている

図は[1]より抜粋



こうした高次元への埋め込みを次元削減して使用する方法はNLPでは広く行われている

例：事前学習済みBERTで埋め込んだあとPCA

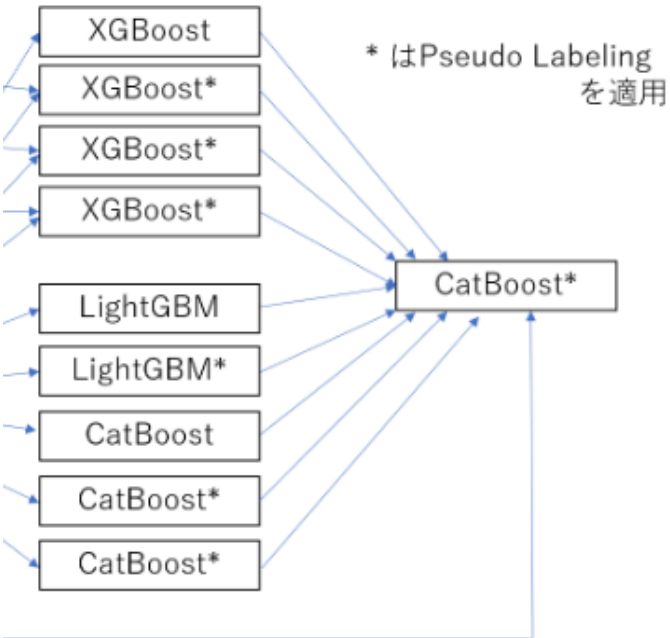
Feature Engineering - Principal Component Analysis on News Headline Embeddings[2]

注：今回は「自然言語」というわけではないので、「自然言語」で事前学習されたBERT等で埋め込んだ場合有効かというと微妙

Stacking

モデルを「積み重ねる」手法

モデルの予測値を特徴量として予測する



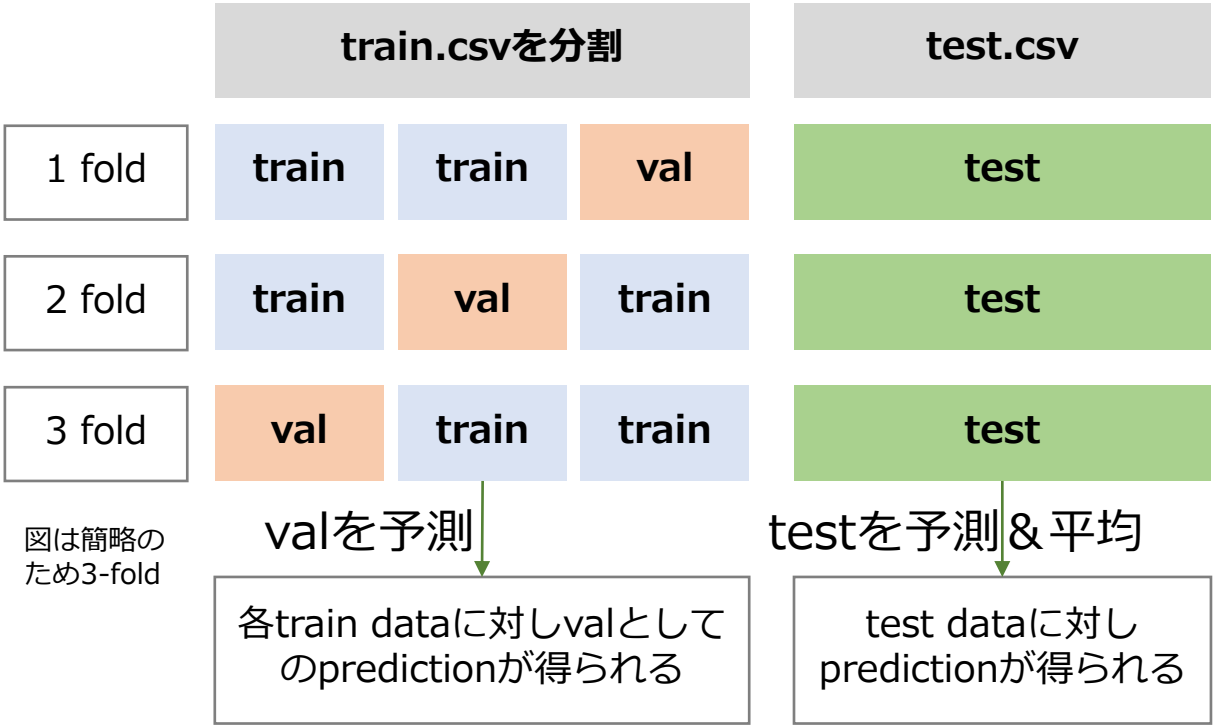
図は[1]より
抜粋

以下[3]に沿って解説する

Stacking（続き）

例：1層目がLightGBM, CatBoost, XGBoostの3つ、2層目がCatBoost

まず1層目の各モデルで以下を行う

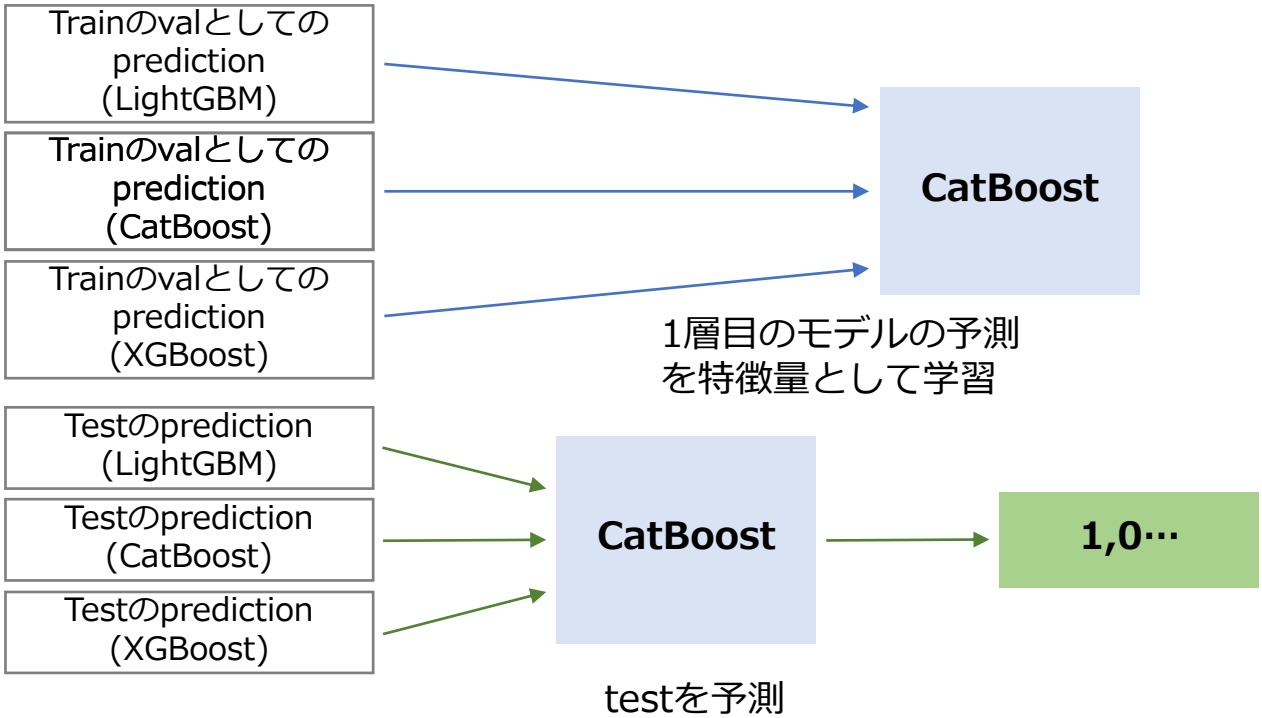


図は簡略のため3-fold

Stacking（続き）

例：1層目がLightGBM, CatBoost, XGBoostの3つ、2層目がCatBoost

2層目のモデルで以下を行う



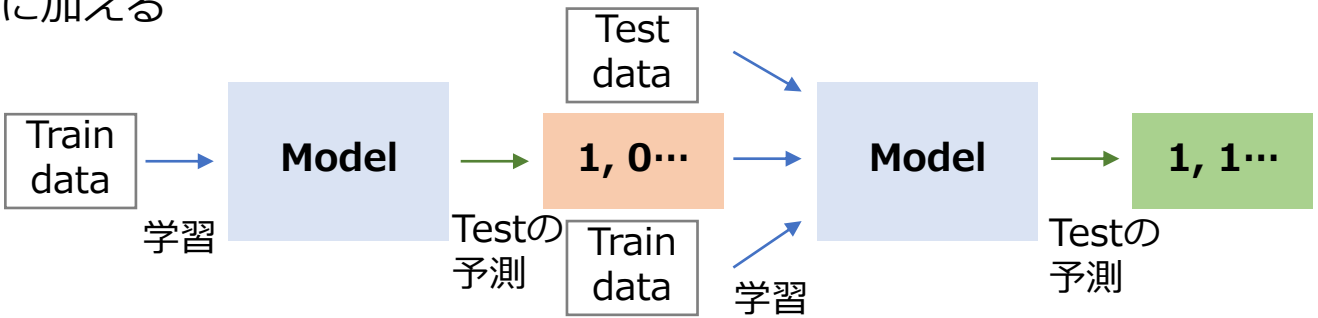
Adversarial Validation

主に2つの使われ方がある
いずれにせよTrain DataとTest Dataに乖離があるときに使用される

論文[4]の該当箇所	概略	方法
3.1 Automated Feature Selection	特徴ベクトルがTrain DataとTest Dataであまり乖離しないように特徴量を選択する	Train DataとTest Dataを見分ける分類器を作り、見分けられないように特徴量を削っていく
3.2 Validation Data Selection	Test Dataに近いデータでValidation Dataを作成し、正確に精度を見積もることを目指す	Train DataとTest Dataを見分ける分類器を作り、Test Dataである予測確率が高いものをValidation Dataとする

Pseudo Labeling

Test Dataの予測値をそのデータのラベルとしてみなし、Train Data に加える



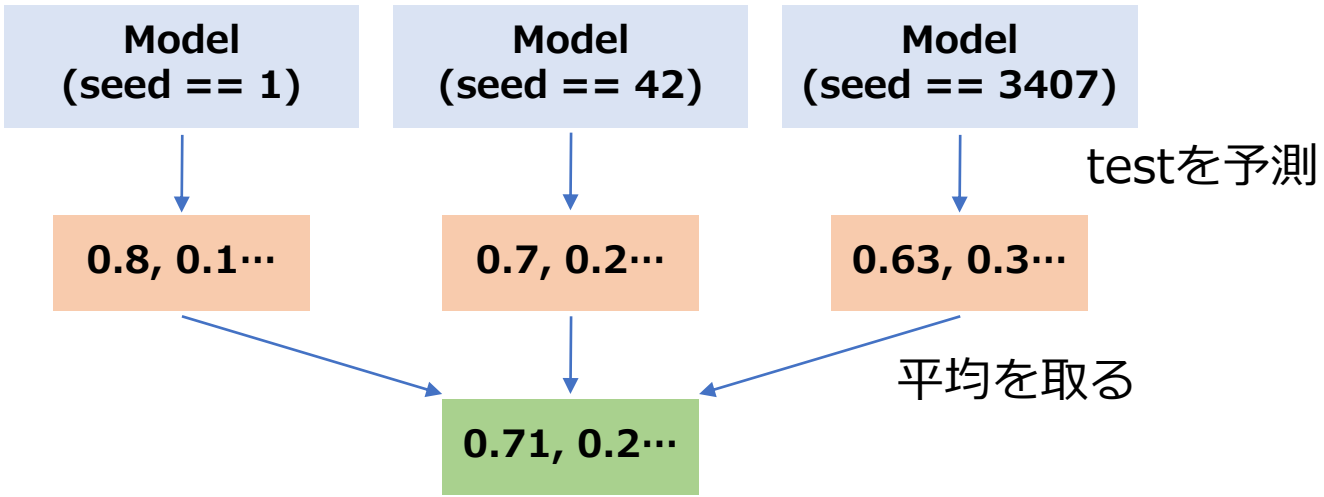
繰り返して行う事も多い

【Kaggle】NBME - Score Clinical Patient Notesコンペにチーム参加し、4位となり金メダルを獲得しました[5]

[3]では工夫の仕方について言及がある

Random Seed Averaging

Random Seedを変えただけのモデルでアンサンブルする（平均を取るだけ。Seedは例）



アンサンブル用に手軽にモデルを増やせるためKaggleでも使用されている

2019 Data Science Bowl 1st Solution[6] など

基本は特徴量エンジニアリング

いろいろテクニックを紹介したが、進め方としてはやみくもに（最初から）Random Seed Averagingやハイパラチューニングをし続けるより特徴量エンジニアリングをやるのがよい

「分析コンペでモデルの精度を上げるために最も重要な要素である」[3]

“Feature engineering is the art part of data science.”[7]

Write-upのすすめ

今回は3チームにWrite-upを公開していただいた

コンペのページ
https://www.kaggle.com/competitions/mws-cup-2022-3/
のCode及びDiscussionから閲覧できる

特に今回の課題3のようなDSコンペは決まった正解がなく、各チームのWrite-upによる議論は重要

課題3以外でも何らかの形でWrite-upを取り上げ、何らかの形で評価できるようにしていきたい

再掲：次回以降

作問にご協力いただける方を募集しています！！！

	2018 - 2021	2022
分類問題のサステナビリティ	• Fexrdの出力を食わせるだけで高精度 • 次回は模範解答からスタートされうる	• Stringsのみを用いた • trainデータとtestデータの時間をずらした

来年以降は未定

問題を「解く」ではなく「作る側」

- DSコンペに参加者ではなく作問側として参加する機会は希少
- より良い問題、より良い模範解答ができればMWSコミュニティそしてサイバーセキュリティ業界全体にプラス

参考文献

[1] 1st place solution (public LB 2nd, private LB 1st), Available at https://www.kaggle.com/competitions/mws-cup-2022-3/discussion/362177 (Accessed 14 November 2022)

[2] Feature Engineering - Principal Component Analysis on News Headline Embeddings, Available at https://developers.refinitiv.com/en/article-catalog/article/ai-feature-engineering-pca-on-news-headlines-embeddings (Accessed 10 November 2022)

[3] 門脇大輔，阪田隆司，保坂桂佑，平松雄司著『Kaggleで勝つデータ分析の技術』，2019, 技術評論社

[4] Jing Pan, Vincent Pham, Mohan Dorairaj, Huigang Chen,

and Jeong-Yoon Lee. “Adversarial Validation Approach to

Concept Drift Problem in User Targeting Automation Systems at Uber.” ACM AdKDD 2020 (2020).

[5]【Kaggle】NBME - Score Clinical Patient Notesコンペにチーム参加し、4位となり金メダルを獲得しました, Available at https://blog.recruit.co.jp/data/articles/kaggle-nbme-score-clinical-patient-notes/ (Accessed 10 November 2022)

[6] 1st Place Solution, Available at https://www.kaggle.com/c/data-science-bowl-2019/discussion/127469 (Accessed 10 November 2022)

[7] Feature Engineering for Automated Machine Learning | Dataset Features, Available at https://www.datarobot.com/wiki/feature-engineering/ (Accessed 14 November 2022)