Adattömörítés szubmoduláris kiválasztással

Készítette: Bartalis Dávid

Témavezetők:
Bérczi-Kovács Erika
ELTE, Operációkutatási Tanszék
Béres Ferenc
SZTAKI, Informatikai Kutatólaboratórium

Budapest, 2021

Tartalomjegyzék

A félév során elvégzett feladatok

- Az előző félévben írt kód könnyen adaptálhatóvá tétele
- Két újabb adathalmaz elemzése, vizsgálata
- Az Apricot módszer kipróbálása ezen adathalmazokon
- Sorok és oszlopok redukálása az Apricot módszerrel
- Megismert módszerek, modellek: neptune.ai, lightGBM, tfidf

Apricot - Szubmodularitás

Cél: reprezentatív részhalmaz kiválasztása.
Felhasználás: Tanítási halmaz redukálása, tanulási folyamat felgyorsítása.
Módszer: szubmoduláris kiválasztás.

Definíció

Egy $\mathcal{F} : 2^V \to \mathbb{R}$ halmazfüggvény szubmoduláris, ha $\forall B \subseteq A \subseteq V$ halmazokra és $x \in \overline{\mathrm{A}}$ esetén

$$\mathcal{F}(A \cup x) - \mathcal{F}(A) \leq \mathcal{F}(B \cup x) - \mathcal{F}(B)$$

Használt függvények

Feature-based / Tulajdonság alapú függvény:

$$\mathcal{F}(X) = \sum_{d=1}^{D} w_d \phi \left(\sum_{x \in X} m_d(x) \right)$$

Facility location / Szolgáltató elhelyezési függvény:

$$\mathcal{F}(X) = \sum_{v \in V} \max_{x \in X} \delta(x, v)$$

Max Coverage / Maximális fedés függvény:

$$\mathcal{F}(X) = \sum_{i=1}^{d} \left(\left(\sum_{x \in X} x_i \right) > 0 \right)$$

Disaster Tweets

- NLP (Natural Language Processing) feladat
- Feladat: egy adott tweet valóban katasztrófáról szól-e
- Tanítási halmaz mérete: (5264, 10000)

	id	text	target
3806	5408	Former Township fire truck being used in Phili...	0
3444	4922	The Dress Memes Have Officially Exploded On Th...	0
3443	4920	Well as I was chaning an iPad screen it fuckin...	0
6219	8875	So does Austin smoke too since he agreed to th...	0
3440	4917	Im Dead!!! My two Loves in 1 photo! My Heart e...	0
...	...	...	...
3663	5213	@Truly_Stings Yo Dm me	1
3660	5210	Driver fatalities down on Irish roads but pede...	1
3659	5209	Message boards will display updated traffic fa...	1
3673	5228	Kosciusko police investigating pedestrian fata...	1
7612	10873	The Latest: More Homes Razed by Northern Calif...	1

Disaster Tweets

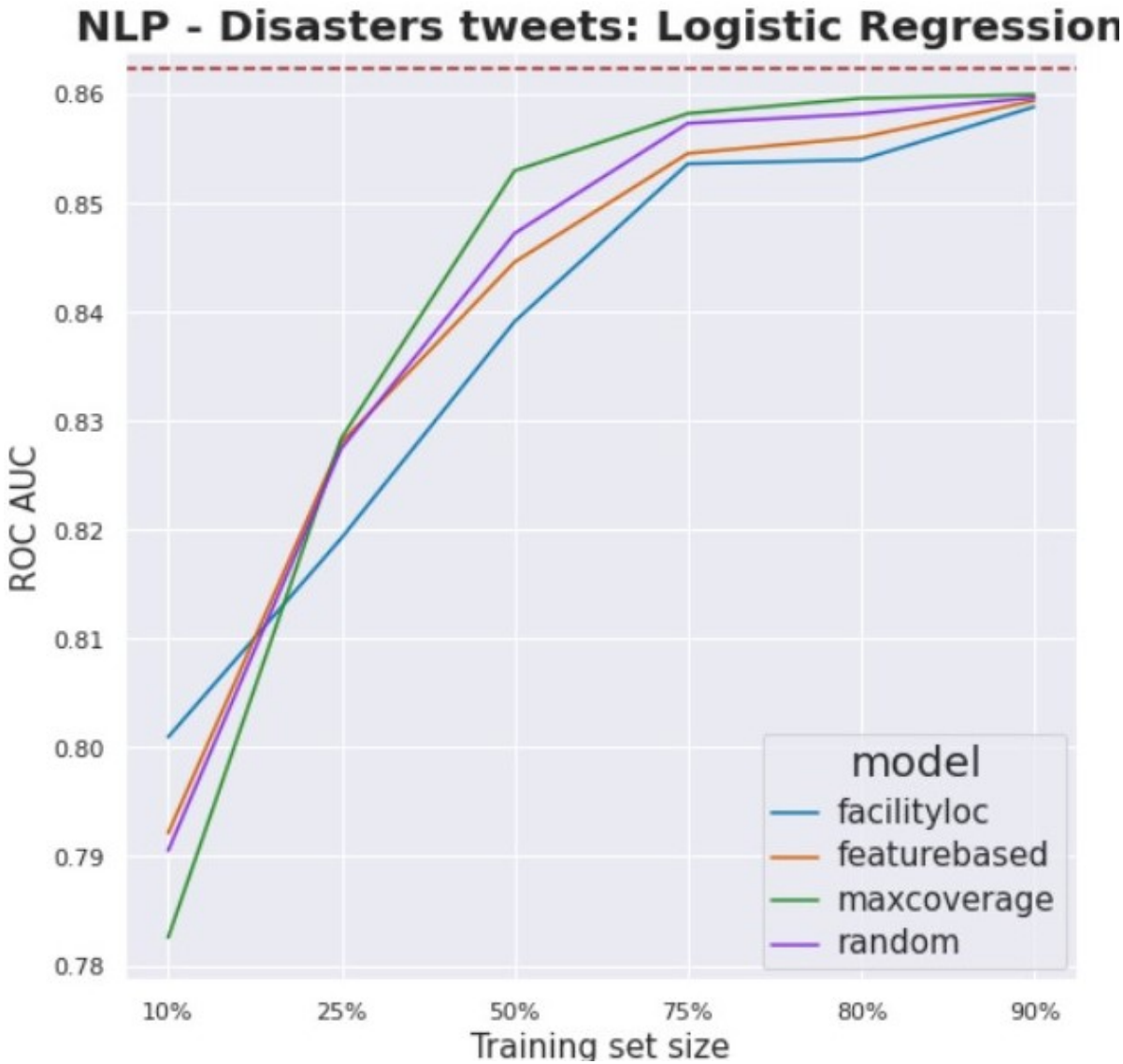
Tanításhoz használt modellek:

- LogisticRegression
- GradientBoostingClassifier
- RandomForestClassifier

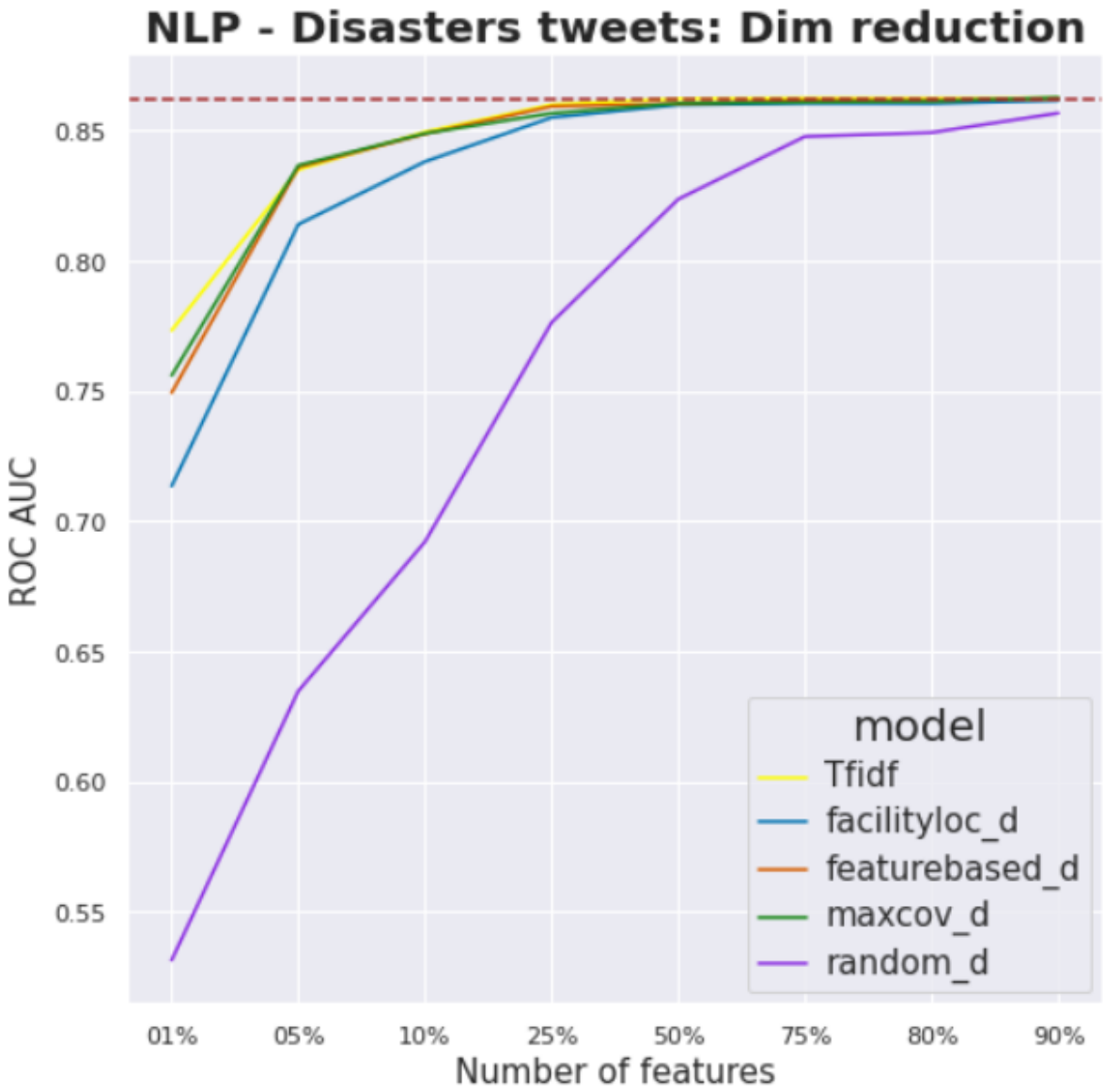
Kiértékeléshez használt metrikák:

- Accuracy
- Precision
- Recall
- Roc Auc

Sorok redukálása



Oszlopok redukálása



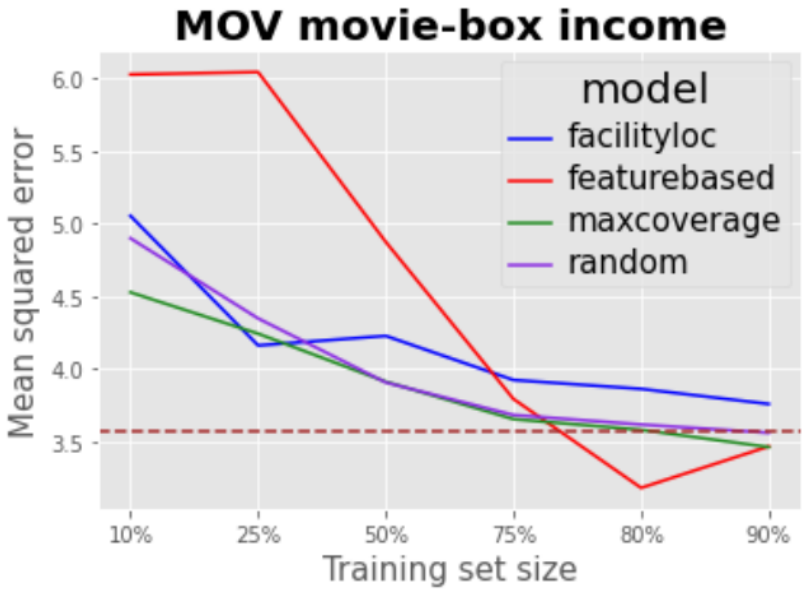
MOV movie-box income

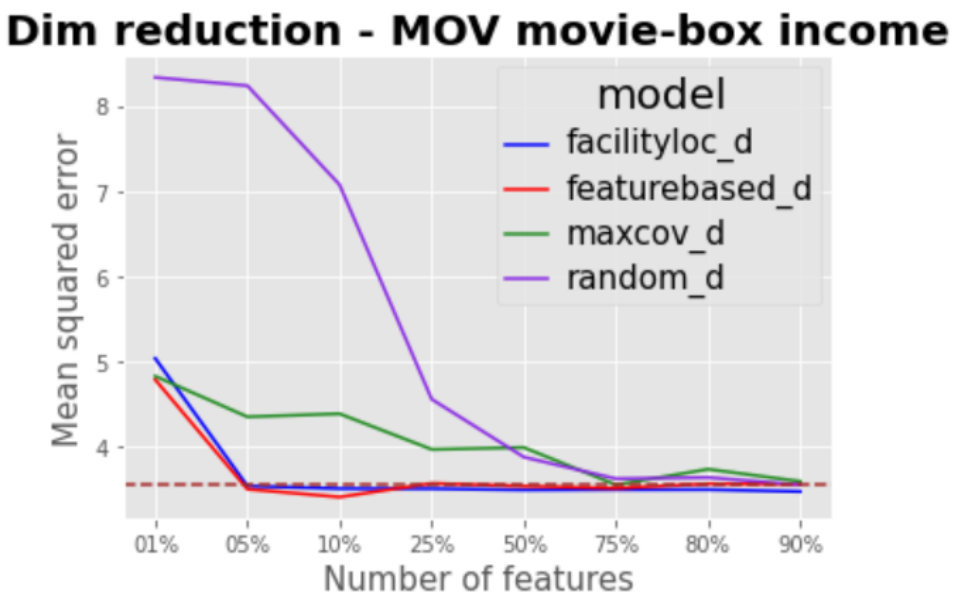
- Feladat: egy adott filmből származó bevételt megjósolni
- Tanítási halmaz mérete: (2700, 243)

```
[RangeIndex(start=0, stop=3000, step=1),
 Index(['id', 'belongs_to_collection', 'budget', 'genres', 'homepage',
        'imdb_id', 'original_language', 'original_title', 'overview',
        'popularity', 'poster_path', 'production_companies',
        'production_countries', 'release_date', 'runtime', 'spoken_languages',
        'status', 'tagline', 'title', 'Keywords', 'cast', 'crew', 'revenue'],
       dtype='object')]
```

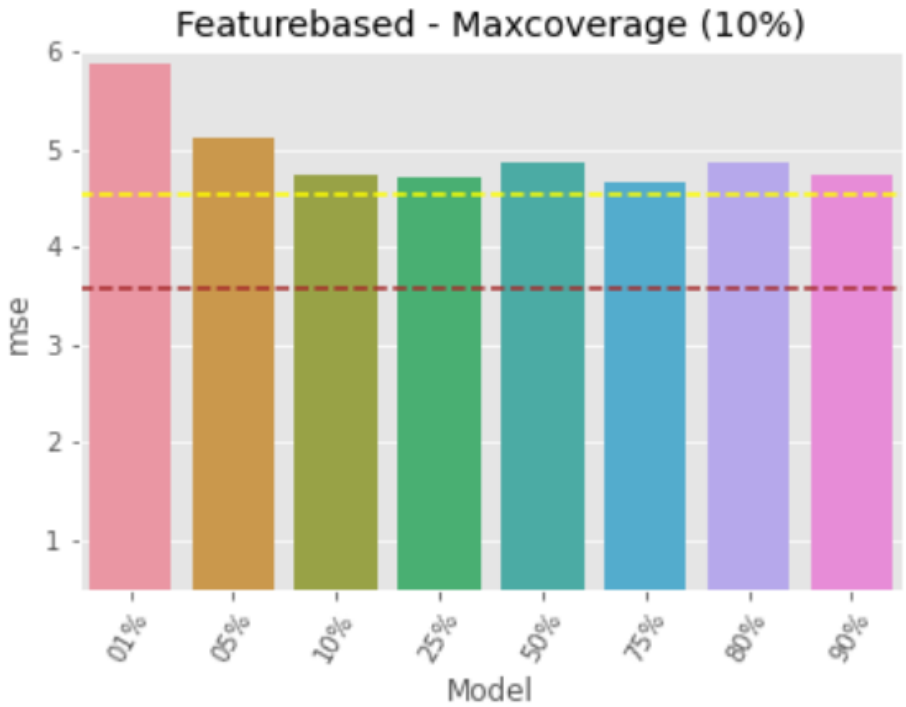
- Tanításhoz használt modell: lightGBM
- Veszteségfüggvény: mean squared error

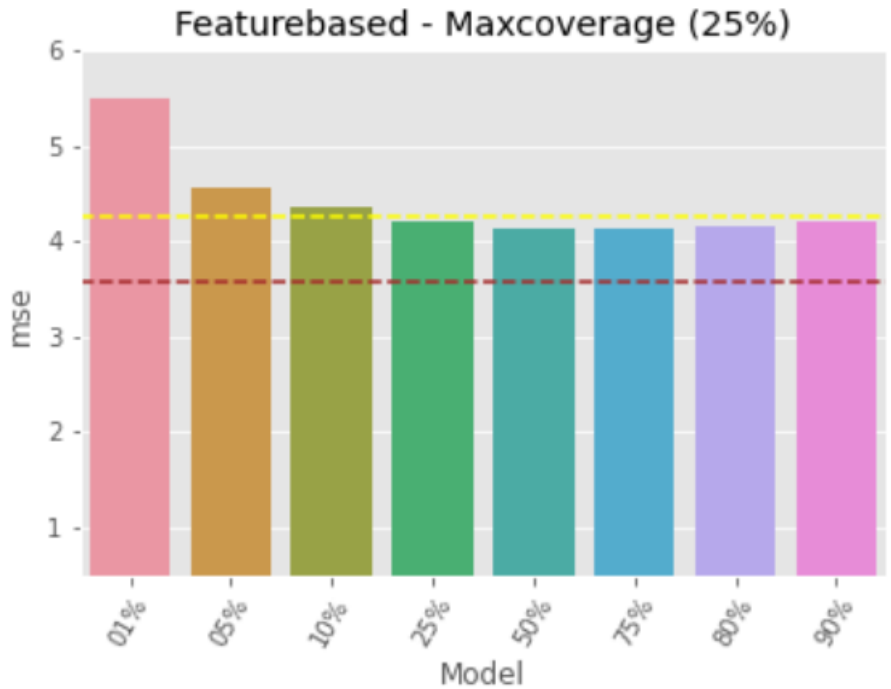
Sorok, oszlopok csökkentése





Dupla Apricot





Összegzés, tervek

Megállapítás:

- Az Apricot a sorok redukálása során nem sokkal múlta felül a random módszert, viszont a feature selection feladatoknál rendkívül jól teljesített.
- A tapasztalatok alapján a feature-based függvény segíthet a túltanulás elkerülésében.

Tervek:

- Szubmoduláris függvények alkalmazása oszlopredukciókhoz.
- Pénzügyi adaton való kipróbálása a módszernek.
- Egy, még nem használt szubmoduláris függvény implementálása a csomagba.

Köszönöm a figyelmet!