Detyra me regresion @ DAIT Datathon

Edis Hasaj & Silva Bashllari

February 21, 2025

1 Përshkrimi i detyrës

Ju është dhënë një grup të dhënash i cili përmbledh disa cilësi të artikujve të postuara 2 vitet e fundit nga faqja Mashable. Detyra juaj do jetë të parashikoni numrin e shpërndarjeve të një artikulli bazuar në cilësitë e tij.

2 Përshkrimi i të dhënave

Në këtë pjesë do përshkruhen me detaje cilësitë e artikujve të grupit të te dhënave.

1. url: URL-ja e artikullit (jo për parashikim)
2. timedelta: Ditët midis publikimit të artikullit dhe marrjes së të dhënave (jo për parashikim)
3. n_tokens_title: Numri i fjalëve në titull
4. n_tokens_content: Numri i fjalëve në përmbajtje
5. n_unique_tokens: Përqindja e fjalëve unike në përmbajtje
6. n_non_stop_words: Përqindja e fjalëve jo-ndaluese (non-stopword) në përmbajtje
7. n_non_stop_unique_tokens: Përqindja e fjalëve unike jo-ndaluese (non-stopword) në përmbajtje
8. num_hrefs: Numri i lidhjeve (links)
9. num_self_hrefs: Numri i lidhjeve me artikuj të tjerë të publikuar nga Mashable

10. num_imgs: Numri i imazheve
11. num_videos: Numri i videove
12. average_token_length: Gjatësia mesatare e fjalëve në përmbajtje
13. num_keywords: Numri i fjalëve kyçe në metadatë
14. data_channel_is_lifestyle: A është kanali i të dhënave 'Lifestyle'?
15. data_channel_is_entertainment: A është kanali i të dhënave 'Entertainment'?
16. data_channel_is_bus: A është kanali i të dhënave 'Business'?
17. data_channel_is_socmed: A është kanali i të dhënave 'Social Media'?
18. data_channel_is_tech: A është kanali i të dhënave 'Tech'?
19. data_channel_is_world: A është kanali i të dhënave 'World'?
20. kw_min_min: Fjala kyçe më e dobët (minimumi i shpërndarjeve)
21. kw_max_min: Fjala kyçe më e dobët (maksimumi i shpërndarjeve)
22. kw_avg_min: Fjala kyçe më e dobët (mesatarja e shpërndarjeve)
23. kw_min_max: Fjala kyçe më e mirë (minimumi i shpërndarjeve)
24. kw_max_max: Fjala kyçe më e mirë (maksimumi i shpërndarjeve)
25. kw_avg_max: Fjala kyçe më e mirë (mesatarja e shpërndarjeve)
26. kw_min_avg: Fjala kyçe mesatare (minimumi i shpërndarjeve)
27. kw_max_avg: Fjala kyçe mesatare (maksimumi i shpërndarjeve)
28. kw_avg_avg: Fjala kyçe mesatare (mesatarja e shpërndarjeve)
29. self_reference_min_shares: Shpërndarjet minimale të artikujve të referuar në Mashable
30. self_reference_max_shares: Shpërndarjet maksimale të artikujve të referuar në Mashable
31. self_reference_avg_sharess: Shpërndarjet mesatare të artikujve të referuar në Mashable
32. weekday_is_monday: A është publikuar artikulli të hënën?
33. weekday_is_tuesday: A është publikuar artikulli të martën?

34. weekday_is_wednesday: A është publikuar artikulli të mërkurën?
35. weekday_is_thursday: A është publikuar artikulli të enjten?
36. weekday_is_friday: A është publikuar artikulli të premten?
37. weekday_is_saturday: A është publikuar artikulli të shtunën?
38. weekday_is_sunday: A është publikuar artikulli të dielën?
39. is_weekend: A është publikuar artikulli gjatë fundjavës?
40. LDA_00: Afërsia me temën LDA 0
41. LDA_01: Afërsia me temën LDA 1
42. LDA_02: Afërsia me temën LDA 2
43. LDA_03: Afërsia me temën LDA 3
44. LDA_04: Afërsia me temën LDA 4
45. global_subjectivity: Subjektiviteti i tekstit
46. global_sentiment_polarity: Polariteti i ndjesisë së tekstit
47. global_rate_positive_words: Përqindja e fjalëve pozitive në përmbajtje
48. global_rate_negative_words: Përqindja e fjalëve negative në përmbajtje
49. rate_positive_words: Përqindja e fjalëve pozitive ndër fjalët jo-neutrale
50. rate_negative_words: Përqindja e fjalëve negative ndër fjalët jo-neutrale
51. avg_positive_polarity: Polariteti mesatar i fjalëve pozitive
52. min_positive_polarity: Polariteti minimal i fjalëve pozitive
53. max_positive_polarity: Polariteti maksimal i fjalëve pozitive
54. avg_negative_polarity: Polariteti mesatar i fjalëve negative
55. min_negative_polarity: Polariteti minimal i fjalëve negative
56. max_negative_polarity: Polariteti maksimal i fjalëve negative
57. title_subjectivity: Subjektiviteti i titullit
58. title_sentiment_polarity: Polariteti i titullit

59. abs_title_subjectivity: Niveli absolut i subjektivitetit të titullit
60. abs_title_sentiment_polarity: Niveli absolut i polaritetit të ndjesisë së titullit
61. shares: Numri i shpërndarjeve (objektivi i parashikimit) (Nuk gjendet si kolonë tek skedari i grupit të të dhënave të testimit)

2.1 Skedarët ku gjendet grupi i të dhënave

Grupi i të dhënave është i ndarë në 2 skedarë.

1. `trajnim.csv`: Artikuj të cilat e kanë kolonën e numrit të shpërndarjeve, mund të përdoren për trajnim dhe validim të modeleve.
2. `testim.csv`: Artikuj të cilat nuk e kanë kolonën e numrit të shpërndarjeve, dhe kanë një kolonë me numer identifikues shtesë.

3 Menyra e vlerësimit të parashikimit

Duke qenë se kjo është një detyrë regresioni, matja e gabimit të modelit tuaj do bëhet nëpërmjet funksionit të rrënjës së mesatares së katrorëve të gabimeve (Root mean squared error ose RMSE).

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

ku n është numri i predikimeve të bëra, $\hat{y}_i$ është parashikimi i objektivit në artikullin e i-të, dhe y_i është vlera reale e objektivit të parashikimit në artikullin e i-të.

Në dokumentimin e scikit-learn: sklearn.metrics.root_mean_squared_error

4 Si të dorëzojmë zgjidhjen

Në përfundim të detyrës, duhet të dorëzoni një arkivë ZIP që do përmbajë:

1. Skedarin `.ipynb` ku gjendet kodi juaj.
2. Parashikimet e bëra tek grupi e të dhënave të testimit në format CSV.

4.1 Si të shkarkoni kodin tuaj nga Colab

Tek shiriti i menusë, klikoni `File > Download .ipynb`

4.2 Si të shkarkojmë parashikimet

Parashikimet duhen dorëzuar në format CSV, me kolonat:

1. `id`: Numri rendor identifikues i elementit te grupi i të dhënave i testimit
2. `parashikimi`: Vlera e parashikuar e objektivit për elementin

Shembull:

```
id, parashikimi
1, 30.0
2, 14.0
3, 15.0
```

Sigurohuni që skedari që do dorëzoni i përshtatet formatit më sipër.

Mbasi ta keni ruajtur skedarin, klikoni ikonën në formë dosjeje në shiritin vertikal në anë të majtë të ekranit, dhe gjeni skedarin me parashikimet. Vendoseni mouse-in siper skedarit, klikoni butonin ⋮, dhe më pas klikoni `Download`.

4.3 Raporti shkencor

48 orë pas dorëzimit të kodit dhe parashikimeve, duhet të dërgoni një raport shkencor në format `.pdf` tek email-i ynë "dait.meetup@gmail.com" ku do të përshkruani rrjedhën logjike të punës suaj. Qëllimi nuk është të dokumentohet kodi, por të dokumentoni qasjen e përdorur për të përmirësuar cilësinë e parashikimeve dhe arsyetimin për përdorimin e teknikave që keni përdorur.

Këshille: Mbani shënime gjatë punës në menyrë që ti kurseni vetes kohë në shkrimin e raportit.

Struktura e këtij raporti do jetë si më poshtë:

1. Hyrje dhe përshkrim i detyrës që bëtë.
2. Përshkrim i përgjithshëm i qasjes suaj ndaj detyrës.
3. Para-procesimi i të dhënave
4. Modelet e përdorura

5. Qasja ndaj kërkimit të hiper-parametrave (Hyperparameter tuning) dhe parametrat e gjetur (Opsionale)
6. Rezultatet
7. Diskutimi

Sigurohuni që skedari i raportit të ndjekë këtë skemë emërtimi: `RAPORT - ${EMRI I UNIVERSITETIT TUAJ} - ${EMRAT E ANËTARËVE TË SKUADRËS, TË NDARË ME PRESJE}`.

Shembull: "RAPORT - Universiteti Universitar - Filan Fisteku, Arben Hoxha.pdf"