Universidad de Montevideo – Introducción a la Ciencia de Datos

Trabajo Obligatorio

2021

Suponga que uno de los blogs mas importantes de noticias de Internet contrata a su equipo para realizar un análisis de datos relacionado con la popularidad de sus posteos. El objetivo es predecir la cantidad de *shares* en las redes sociales (popularidad) para un determinado posteo. Con tal objetivo en mente, el encargado de bases de datos les facilitó una planilla con datos históricos con 61 variables (58 atributos predictivos, 2 no predictivos, 1 variable objetivo).

Estructura del proyecto

1. Elaborar una presentación para el equipo directivo donde, en no más de 10 slides, se logre comunicar:

- Relevancia del problema: discutir la importancia del desafío planteado.
- Planteamiento de hipótesis: identificar qué atributos o variables podrían estar relacionados y cómo con la variable de interés.
- Metodología de trabajo: presentar abordajes alternativos desde el punto de vista metodológico, identificando cuál va se va seguir y estableciendo pros y contras en cada caso.
- Resultados del análisis: presentar los resultados que validan o refutan las hipótesis planteadas.
- Plan de acción: en base a los resultados del análisis, hacer recomendaciones al cliente para determinar qué posteos son los más valiosos a la hora de conseguir una mayor cantidad de *shares* en las redes sociales.

2. Como anexo, agregar el script de Python utilizado para la obtención de los resultados.

3. Adicionalmente, su cliente testeara el modelo presentado en datos no conocidos para medir el nivel de precisión.

Nota: Esta es una estructura tentativa de presentación, como consultores son libres de elegir qué y cómo mostrar los resultados de su análisis. En caso de que para realizar el trabajo se adopte algún tipo de supuesto, deberá ser especificado.

Diccionario

0. url: URL of the article (non-predictive)
1. timedelta: Days between the article publication and the dataset acquisition (non-predictive)
2. n_tokens_title: Number of words in the title
3. n_tokens_content: Number of words in the content
4. n_unique_tokens: Rate of unique words in the content
5. n_non_stop_words: Rate of non-stop words in the content
6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
7. num_hrefs: Number of links
8. num_self_hrefs: Number of links to other articles published by the blog
9. num_imgs: Number of images
10. num_videos: Number of videos
11. average_token_length: Average length of the words in the content
12. num_keywords: Number of keywords in the metadata
13. data_channel_is_lifestyle: Is data channel 'Lifestyle'?
14. data_channel_is_entertainment: Is data channel 'Entertainment'?
15. data_channel_is_bus: Is data channel 'Business'?
16. data_channel_is_socmed: Is data channel 'Social Media'?
17. data_channel_is_tech: Is data channel 'Tech'?
18. data_channel_is_world: Is data channel 'World'?
19. kw_min_min: Worst keyword (min. shares)
20. kw_max_min: Worst keyword (max. shares)

21. kw_avg_min: Worst keyword (avg. shares)
22. kw_min_max: Best keyword (min. shares)
23. kw_max_max: Best keyword (max. shares)
24. kw_avg_max: Best keyword (avg. shares)
25. kw_min_avg: Avg. keyword (min. shares)
26. kw_max_avg: Avg. keyword (max. shares)
27. kw_avg_avg: Avg. keyword (avg. shares)
28. self_reference_min_shares: Min. shares of referenced articles in Blog
29. self_reference_max_shares: Max. shares of referenced articles in Blog
30. self_reference_avg_sharess: Avg. shares of referenced articles in Blog
31. weekday_is_monday: Was the article published on a Monday?
32. weekday_is_tuesday: Was the article published on a Tuesday?
33. weekday_is_wednesday: Was the article published on a Wednesday?
34. weekday_is_thursday: Was the article published on a Thursday?
35. weekday_is_friday: Was the article published on a Friday?
36. weekday_is_saturday: Was the article published on a Saturday?
37. weekday_is_sunday: Was the article published on a Sunday?
38. is_weekend: Was the article published on the weekend?
39. LDA_00: Closeness to LDA topic 0
40. LDA_01: Closeness to LDA topic 1
41. LDA_02: Closeness to LDA topic 2
42. LDA_03: Closeness to LDA topic 3
43. LDA_04: Closeness to LDA topic 4
44. global_subjectivity: Text subjectivity
45. global_sentiment_polarity: Text sentiment polarity
46. global_rate_positive_words: Rate of positive words in the content
47. global_rate_negative_words: Rate of negative words in the content
48. rate_positive_words: Rate of positive words among non-neutral tokens
49. rate_negative_words: Rate of negative words among non-neutral tokens
50. avg_positive_polarity: Avg. polarity of positive words

51. min_positive_polarity: Min. polarity of positive words
52. max_positive_polarity: Max. polarity of positive words
53. avg_negative_polarity: Avg. polarity of negative words
54. min_negative_polarity: Min. polarity of negative words
55. max_negative_polarity: Max. polarity of negative words
56. title_subjectivity: Title subjectivity
57. title_sentiment_polarity: Title polarity
58. abs_title_subjectivity: Absolute subjectivity level
59. abs_title_sentiment_polarity: Absolute polarity level
60. shares: Number of shares (target)