PROYECTO DE CURSO
Web Traffic Time Series Forecasting

Valentina Echavarria Porras [a], David Santiago Arcila Alvarez[b], Felipe Cadavid Rincon[c].

[a] *Bachiller académico, estudiante perteneciente a la Escuela ambiental, Facultad de Ingeniería, Ingeniería Civil, Universidad de Antioquia, cédula:1214743266, Medellín (A), Colombia. valentina.echavarria1@udea.edu.co*

[b] *Bachiller académico con énfasis en pedagogía, estudiante perteneciente a la Escuela ambiental, Facultad de Ingeniería, Ingeniería Ambiental, Universidad de Antioquia, cédula: 100008537, Medellín (A), Colombia.* davids.arcila@udea.edu.co

[b] *Bachiller académico, estudiante perteneciente a la Escuela de ingeniería, Facultad de Ingeniería, Ingeniería de Sistemas, Universidad de Antioquia, cédula:1128389605, Medellín (A), Colombia.* felipe.cadavidr@udea.edu.co

Aprueba
Raúl Ramos Pollán
Julián David Arias Londoño

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA CIVIL, AMBIENTAL Y SISTEMAS
INTELIGENCIA ARTIFICIAL
MEDELLÍN, ANTIOQUIA
JUNIO DE 2022

PROBLEMA PREDICTIVO A RESOLVER

Con base a los requerimientos del reto, el objetivo de este es implementar un modelo que nos permita realizar un pronóstico y predicción de tráfico web basado en series temporales; en este caso, del tráfico web de artículos de Wikipedia.

La característica de una serie temporal es proporcionar una representación, la cual realiza mediciones específicas en un momento dado de forma cronológica. Este mismo está compuesto usualmente de datos por intervalos de tiempo, pero no siempre proporciona vistas acertadas, para predecir y estimar con precisión, tal como se ve en algunas predicciones del clima o la bolsa de valores.

Por lo tanto, además de implementar un modelo para predicciones, también se busca llegar a determinar un modelo adecuado y acertado, mitigando problemas de análisis de datos e inferencia en la clasificación de datos, en este caso, para tráfico web.

DATASET A UTILIZAR

El dataset que ha sido seleccionado para este proyecto, el cual reside en Kaggle. (Ver Dataset aquí). El dataset consiste en 145.000 series temporales, que van desde el primero de Julio de 2015 hasta el 31 de Diciembre de 2016.

Adicionalmente incluye datos de prueba para entrenamiento que van desde el primero de enero de 2017 hasta el primero de marzo de 2017 para hacer pruebas de primera fase, y datos de prueba hasta septiembre de 2017 para la segunda fase de pruebas.

Sin embargo, se identificó que existe complejidad en la interpretación de algunos datos de tráfico, ya que si no hubo tráfico o hay datos faltantes, ambos casos están representados por un cero (0)

El proyecto contiene varias carpetas con diferentes archivos indicados a continuación:

1. Keys (key_1 y key_2): Contine informacion concerniente a las paginas y los identificadores de las páginas
    a. ***page***: Contiene el nombre de la página visitada en la serie temporal concatenado en snake case con otros datos que pueden ser importantes para el análisis. Potencialmente se tendrá que masajear y transformar esta columna para tener una mejor representación de su contenido. Datos contenidos en esta columna:
        i. Nombre del artículo
        ii. Version del artículo de Wikipedia por idioma
        iii. Origen/Agente, es decir, desde donde se abrió el articulo (Mobile, Desktop, etc)
        iv. Fecha del registro
    b. ***id***: identificador de la página visitada en la serie temporal. Establece la correlacion de los datos de la columna page con las visitas en los archivos mencionados más adelante
2. Submissions (sample_submission_1 y sample_submission_2) :
    a. ***id***: Identificador, potencialmente la clave primaria que correlaciona las visitas con la página visitada (en los archivos keys)
    b. ***visits***: Contador, representado como un entero, indica la cantidad de visitas para el identificador respectivo.

3. Train (train_1 y train_2):
    a. ***page***: El nombre de la página visitada únicamente, concatenado con el agente en snake case
    b. ***columnas con las fechas en año calendario***: contiene la cantidad de visitas por cada fecha del calendario comprendido en los datos entregados

Al ser una serie temporal nosotros definimos la organización de las 30 columnas necesaria en fechas para la utilización de los datos según se considere necesario.

MÉTRICA DE DESEMPEÑO

Para el modelo, se tendrá en cuenta el modelo recomendado por el reto en Kaggle, en este caso el modelo SMAPE, la cual es una métrica de porcentual de regresión que veremos a continuación:
valores reales (yi) y los valores predichos (yi^).

$$SMAPE = (\frac{1}{N}\Sigma\frac{|y_i - \hat{y}_i|}{\left(|y_i| - |\hat{y}_i|\right)/2}) * 100$$

La 'S' en SMAPE significa simétrica, 'M' significa media que toma el valor promedio de una serie, 'A' significa absoluta que usa valores absolutos para evitar que los errores positivos y negativos se cancelen entre sí, ' P' es el porcentaje que hace que esta métrica de precisión sea una métrica relativa, y 'E' significa error, ya que esta métrica ayuda a determinar la cantidad de error que tiene nuestro pronóstico.

El SMAPE nos permite superar las limitaciones con la medición del error de pronóstico MAPE. En contraste con el error porcentual absoluto medio, SMAPE tiene un límite inferior y un límite superior, por lo tanto, se lo conoce como simétrico, esta métrica porcentual nos facilita medir el error de manera escalonada, es decir, se busca acotar el error entre valores de 0 a 1, donde 0 significa que el ajuste es perfecto, mientras que 1 sería un mal ajuste. Cabe destacar que muchas veces las métricas porcentuales pueden tener valores mayores a 1.

DESEMPEÑO

Se desearía poder acotar el error en 0.6, es decir en un 60% de acierto, ya que se busca predecir valores certeros sobre el tráfico futuro de una página para poder valorar su visibilidad, ya que estos datos se pueden usar para definir la relación que existe entre las páginas y los eventos, o sucesos que se dén en ciertas etapas temporales, como lo son por ejemplo, temporadas vacacionales, temporadas de estudio en calendario académico, eventos públicos, entre otros, y poder establecer una posible clasificación.

REFERENCIAS

Anapedia. (n.d.). *Understand advanced metrics - Anaplan Technical Documentation*. Anapedia. Retrieved July 5, 2022, from https://help.anaplan.com/685ff9b2-6370-46ba-af10-679405937113-Understand-advanced-metrics

GeeksforGeeks. (2021, November 28). *How to Calculate SMAPE in Python?* GeeksforGeeks. Retrieved July 5, 2022, from https://www.geeksforgeeks.org/how-to-calculate-smape-in-python/

Indeed Editorial Team. (2021, October 21). *How To Use the SMAPE Formula (4 Methods With Examples)*. Indeed. Retrieved July 5, 2022, from https://www.indeed.com/career-advice/career-development/smape-formula

Villanueva Ocampo, B., López Sarmiento, D., & Rivas Trujillo, E. (2012, Julio- Diciembre). Métodos de modelamiento y predicción de tráfico orientados a plataformas de transmisión de video e IPTV usando series de tiempo. *REVISTA CIENTIFICA*, (16), 11-21.

Web Traffic Time Series Forecasting. (2017, Julio 13). Web Traffic Time Series Forecasting | Kaggle. Retrieved July 5, 2022, from https://www.kaggle.com/competitions/web-traffic-time-series-forecasting/data