Entrega 1 Proyecto de inteligencia artifical

Presentado por:

Andrea Sánchez Castrillón (1001420939)
Alejandro Vargas Ocampo (1088298091)

Profesor:
Raul Ramos Pollan

Universidad de Antioquia
Facultad de Ingeniería
2023-2

Entrega 1 proyecto de inteligencia artificial

1. Problema predictivo a resolver

Como se aprecia en el dataset reportado en el punto 2, se tienen variables relacionadas con la encuesta de clasificación en el sistema de potenciales beneficiarios de 2019 (Sisbén 3); en el siguiente link se presentan las variables evaluadas con la encuesta, dentro de las que vale la pena mencionar aspectos como electrodomésticos con los que cuenta el hogar, ingresos mensuales del hogar, el material con el que se construyó la vivienda y si la vivienda es propia o alquilada, servicios públicos con los que cuenta la vivienda, entre muchas otras variables y, finalmente, se obtiene un puntaje del Sisbén 3, en una escala de 0 a 100 puntos.

Variables evaluadas en la encuesta: Se encuentran en la sección “TableSchema”, relacionada en el siguiente link:

http://medata.gov.co/dataset/base-de-datos-sisben-2019/resource/be0bbbf8-9c87-494e-b6ae-ca2d03b858c5#{view-graph:{graphOptions:{hooks:{processOffset:{},bindEvents:{}}}},graphOptions:{hooks:{processOffset:{},bindEvents:{}}}}

Con base en lo anterior, el problema a resolver consiste en generar un modelo predictivo que, con base en todas las variables evaluadas en la encuesta de Sisbén 3, permita obtener el puntaje específico para un individuo.

2. Dataset a emplear

El dataset original a emplear, se encuentra disponible en:

http://medata.gov.co/dataset/base-de-datos-sisben-2019

El nuevo dataset, con las condiciones especificadas en el proyecto (al menos 5000 instancias, al menos 30 columnas, al menos el 10% han de ser categóricas, al menos un 5% de datos faltantes en al menos 3 columnas), se encuentra disponible en:

https://drive.google.com/file/d/1nU5vyvDo_OtqGZuwadCQkjMGGGF_ccFf/view?usp=sharing

3. Métricas de desempeño requeridas

Dado que se desea predecir un valor numérico (puntaje del Sisbén que va de 0 a 100), se deben emplear métricas de regresión; en este sentido, se pretenden evaluar dos métricas de desempeño, relacionadas a continuación:

a. Error absoluto medio (Mean Absolute Error): Dado que el puntaje del Sisbén va de 0 a 100 y es la salida que se pretende predecir, los datos de salida no se ven enfrentados a diferentes escalas, lo que permite que emplear el error absoluto medio sea adecuado; adicionalmente MAE tiene las ventajas de que su evaluación es fácil de calcular, además de que proporciona una medida de qué tan bien está funcionando el modelo.

   El MAE se calcula como:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{Y_i}\right|$$

   Donde n corresponde a el número de datos, Yi representa el vector de etiquetas de la i-ésima muestra, y Yiˆ son las predicciones correspondientes de la muestra utilizando el método propuesto.

b. Raíz del error cuadrático medio (RMSE): El RMSE permite penalizar errores grandes, lo que se convierte en una ventaja a la hora de valorar los puntajes de asignación del sisbén, pues errores grandes, implican una incorrecta categorización, lo que propicia no garantizar una focalización adecuada para que las personas más vulnerables accedan a la ayuda estatal; en este sentido, el RMSE resulta ser adecuado, pues penalizar fuertemente errores grandes puede dar cuenta de que se requiera una modificación o cambio en el modelo predictivo.

   El RMSE se calcula como:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} Y_i - \widehat{Y_i}^2}$$

Donde N corresponde a el número de datos, Yi representa el vector de etiquetas de la i-ésima muestra, y Yiˆ son las predicciones correspondientes de la muestra utilizando el método propuesto.

Se evaluará el modelo predictivo planteado inicialmente, con las anteriores dos métricas y de acuerdo a los resultados y su análisis, se decidirá si fueron adecuados o si se deberían buscar otro tipo de métricas.

Como métrica de negocio, se puede definir la tasa de clasificación de usuarios versus el tiempo.

4. Criterio sobre desempeño deseable en producción

Si la tasa de clasificación de usuarios en un periodo de tiempo determinado no es superior en, por lo menos, un 20% de la tasa de clasificación de usuarios con la anterior metodología, no valdrá la pena emplear el modelo determinado, pues el modelo estaría encaminado hacia la reducción de trabajo y tiempo por encuestador.

5. Referencias bibliográficas

- “Mean Absolute Error”. ScienceDirect. Retrieved 20 september 2023, from https://www.sciencedirect.com/topics/engineering/mean-absolute-error
- “Root Mean Squared Error”. ScienceDirect. Retrieved 20 september 2023, from https://www.sciencedirect.com/topics/engineering/root-mean-squared-error