Máster en Big Data Data Science & Inteligencia Artificial

Universidad Complutense de Madrid

Minería de datos y Modelización Predictiva

El conjunto de datos DatosEleccionesEspaña.xlsx contiene información demográfica sobre los distintos municipios de España junto con los resultados que se obtuvieron en las últimas elecciones. Existen 7 posibles variables objetivo:

AbstentionPtge: Porcentaje de abstención

Izda Pct: Porcentaje de votos a partidos de izquierda (PSOE y Podemos)

Dcha Pct: Porcentaje de votos a partidos de derecha (PP y Ciudadanos)

Otros Pct: Porcentaje de votos a partidos distintos de PP, Ciudadanos, PSOE y Podemos

AbstencionAlta: Variable dicotómica que toma el valor 1 si el porcentaje de abstención es superior al 30 % y, 0, en otro caso.

Izquierda: Variable dicotómica que toma el valor 1 si la suma de los votos de izquierdas es superior a la de derechas y otros y, 0, en otro caso.

Derecha: Variable dicotómica que toma el valor 1 si la suma de los votos de derecha es superior a la de izquierda y otros y, 0, en otro caso.

El objetivo de esta práctica es obtener dos modelos de regresión (lineal y logística) seleccionando, de entre las 7 variables anteriores, una variable objetivo continua y otra variable objetivo binaria. El resto de las variables objetivo que no han sido seleccionadas se eliminan del conjunto de datos, no se utilizan como variables explicativas. La variable objetivo continua se utiliza para la construcción del modelo de regresión lineal y la variable objetivo binaria se utiliza para el modelo de regresión logística. Antes de desarrollar los modelos de regresión, es necesario llevar a cabo un proceso de depuración de los datos. Por tanto, los pasos a seguir para la realización de la práctica son:

1. Introducción al objetivo del problema y las variables implicadas. Se debe explicar cual es el objetivo del problema que se va a analizar según las variables objetivo seleccionadas. Además, explicar brevemente como son las variables implicadas en el estudio de forma general, no es necesario describir cada una de ellas.
2. Importación del conjunto de datos y asignación correcta de los tipos de variables.

3. Análisis descriptivo del conjunto de datos. Número de observaciones, número y naturaleza de variables, datos erróneos etc.
4. Corrección de los errores detectados.
5. Análisis de valores atípicos. Decisiones.
6. Análisis de valores perdidos. Imputaciones.
7. Detección de las relaciones entre las variables input continuas, así como las relaciones entre todas las variables input y cada una de las variables objetivo.
8. Construcción del modelo de regresión lineal. NO hacer el modelo manual. NO utilizar transformaciones. Utilizar interacciones solo entre las variables continuas.
    - Mediante los métodos de selección clásica de variables.
    - Mediante el método de selección aleatoria de variables. Realizarla solo con uno de los métodos de selección clásica, el que consideréis más adecuado, justificarlo. El número de iteraciones a utilizar en la selección aleatoria que sea bajo para que el coste computacional no sea muy elevado. Determinar este número de iteraciones libremente según la capacidad computacional de vuestro ordenador.
    - Selección del modelo ganador de entre todos los modelos construidos.
    - Interpretación de los coeficientes de dos variables incluidas en el modelo ganador, una binaria y otra continua.
    - Justificar por qué es el modelo ganador y medir la calidad del mismo.

9. Construcción del modelo de regresión logística. NO hacer el modelo manual. NO utilizar transformaciones. Utilizar interacciones solo entre las variables continuas. En este apartado deben explicarse y justificarse todos los pasos a realizar para responder a los apartados solicitados en la tarea. Aunque los pasos y códigos a utilizar son muy similares en regresión lineal y logística, debe indicarse lo que es igual, haciendo referencia al apartado correspondiente en regresión lineal. Los códigos que son distintos en regresión lineal y logística, aunque muy similares, se puede destacar solo la parte del código que es diferente respecto al código detallado anteriormente en regresión lineal.
    - Mediante los métodos de selección clásica de variables.
    - Mediante el método de selección aleatoria de variables. Realizarla solo con uno de los métodos de selección clásica, el que consideréis más adecuado, justificarlo. El número de iteraciones a utilizar en la selección aleatoria que sea bajo para que el coste computacional no sea muy elevado. Determinar este número de iteraciones libremente según la capacidad computacional de vuestro ordenador.
    - Selección del modelo ganador de entre todos los modelos construidos.
    - Determinar el punto de corte óptimo.
    - Interpretación de los coeficientes de dos variables incluidas en el modelo ganador, una binaria y otra continua.
    - Justificar por qué es el modelo ganador y medir la calidad del mismo.

Se entregará un informe en PDF (máximo 25 páginas, la portada y el índice no están incluidas, cualquier página adicional no se tendrá en cuenta) en el que se explicarán

detalladamente los pasos seguidos incluyendo los códigos y salidas más relevantes. Imprescindible mostrar el summary del modelo ganador. Es muy importante comentar y justificar razonadamente las decisiones que se toman en cada uno de los apartados. En el documento GuíaElaboraciónTarea.docx se muestra una pequeña guía sobre como elaborar el informe correspondiente a la tarea.

La puntuación de la tarea está dividida en tres partes:

- Depuración de datos (3,3 puntos). Todos los apartados de esta parte tienen la misma puntuación.
- Regresión Lineal (3,3 puntos). Todos los apartados de esta parte tienen la misma puntuación.
- Regresión Logística (3,4 puntos). Todos los apartados de esta parte tienen la misma puntuación.