Trabajo Final: calificaciones y sus determinantes en educación secundaria

Navadian | Robaina

Introdución a la Ciencia de Datos

Julio 2023

Presentación del dataset

Para este informe, se tomó un dataset del desempeño académico de dos colegios de secundaria en Portugal. El dataset consta de dos tablas, una contiene los resultados de la asignatura lengua –en este caso, portugués– y otra los resultados de matemática. Este informe se centrará en los datos respectivos a matemática.

Respecto a la calidad de los datos, se sabe que fueron recolectados a través de cuestionarios y reportes. La tabla referente a la asignatura matemática, contiene 395 instancias que representan a cada alumno, y 34 variables, entre ellas de tipo sociodemográficas, acerca de su salud, su tiempo libre, variables académicas que refieren al alumno, y otras respectivas al grupo familiar y las relaciones dentro del mismo. Por último pero no menos importante, la tabla no cuenta con valores faltantes y parece no tener datos atípicos.

A pesar de esta completitud en la tabla, se podrían presentar otros problemas de calidad de los datos. Estos se relacionan a la dificultad de medición de ciertas variables, como el *study time*, que además es relativo al método de estudio y habilidad (inmensurable) de cada alumno. También se pone en duda la veracidad con que habrán respondido los alumnos acerca de su consumo de alcohol, y la cuantificación de *family support*, que refiere a la calidad de la relación familiar, con la dificultad que conlleva medirla y hacerla comparable entre alumnos.

Problemas a resolver y su desarrollo

A pesar de la poca cantidad de datos que contiene el dataset (que restringe la capacidad de realizar generalizaciones o modelos de aprendizaje profundo), podría resultar de interés abordar distintos problemas con el mismo.

Clasificación a través de regresión logística

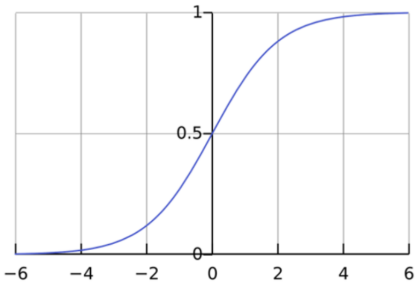
Uno de ellos podría ser un problema de clasificación entre alumnos que aprueben y alumnos que reprueben el curso. Ello se podría abordar a través de distintas metodologías, pero como también puede resultar de interés entender los determinantes de la aprobación se propone realizar una regresión logística, ya que este modelo brinda una buena interpretabilidad a través de los coeficientes β_i de la regresión.

En primer lugar, se debería realizar una revisión de antecedentes para ver qué variables son comúnmente aceptadas como determinantes del buen desempeño estudiantil, lo que se podría complementar con visualizar la matriz de correlación entre las variables del dataset y nuestra variable dependiente.

Las calificaciones en el dataset están entre 0 y 20, por lo que debería definirse un umbral de aprobación y reasignar por ceros y unos. Luego esta variable dependiente, se regresa contra las variables independientes seleccionadas hallando los coeficientes. El modelo de regresión lineal, nos dará predicciones multiplicando el valor de cada variable independiente por los coeficientes β_i y esta será un número continuo, incluso fuera del intervalo $[0; 1]$. Es aquí donde entra en juego la transformación logística:

$$f(x) = \frac{1}{1 + e^{-x}}$$

La cual convierte los valores continuos en una salida entre 0 y 1 con la siguiente forma:



Haciendo que para valores muy negativos $f(x) = 0$ y para valores positivos muy altos $f(x) = 1$. Y dejando en el camino un margen de valores de $f(x)$ entre 0 y 1, para los cuales también se deberá decidir cierto umbral que definirá aprobación o reprobación de los estudiantes.

Diferencias entre géneros

Preguntas a responder también podrían ser respectivas a diferencias por género: qué proporciones de aprobación hay de hombres y de mujeres y con qué calificaciones, o si los determinantes son los mismos o se diferencian. Una primera alternativa, podría ser incluir la variable dummy de sexo en el modelo anterior de regresión logística, y agregar la interacción con el resto de regresores para identificar si los coeficientes cambian cuando se activa la dummy.

Para visualizar los datos, se podría realizar una reducción de dimensionalidad separando por el sexo de los estudiantes y ver si se logra diferenciarlos. Otros modelos a entrenar si lo que se quiere es la clasificación de sexo podrían ser: árboles de decisión (por su interpretabilidad) o random forest que es más preciso y no sobre-ajusta, pero es más complejo de interpretar.

Relaciones entre variables

Se puede aprovechar el dataset para identificar vínculos y realizar predicciones entre las demás variables. Analizar distribuciones de las variables entre los dos colegios (podría ser visualizando los histogramas). Estimar la decisión de no querer ingresar a educación superior a través de variables sociodemográficas, familiares u otros factores que estén relacionados.

Este tipo de problemas de relación de variables con datos cross section, se pueden abordar de forma simple a través de regresiones lineales múltiples. Manteniendo siempre la cautela de no realizar regresiones espurias, ya que correlación no siempre significa causalidad.

Comentarios generales para el desarrollo

Una cuestión importante que se trató bastante en el curso y se comento en las tareas anteriores, fue el balance entre las clases a la hora realizar predicciones de clasificación. En particular en este dataset, notamos que hay una mayor proporción de aprobados que de reprobados, lo que podría sesgar el modelo en detrimento de los últimos.

En el caso del problema de clasificación binaria de aprobación, nos es de sumo interés detectar justamente quiénes son los reprobados con el fin de entender cómo se podría acompañar a dichos estudiantes. Para ello podría ser de utilidad implementar técnicas de submuestreo de los aprobados y balancear así las clases.

En cuanto al modelado de este problema, como se mencionó, resultó de sumo interés la interpretabilidad que nos pueda brindar este modelo con el fin de asistir a los estudiantes que reprueban, por lo que como primera opción se abordarían los modelos de regresión logística y árbol de decisión comparando sus performances y eligiendo así el de mejor desempeño.