Modelos Estadísticos Aplicados I

Reporte de Proyecto Final

Apupalo Laura, Espinel Hellen, Zúñiga Pablo

2023-02-05

Introducción

El continuo avance de las tecnologias invita a estudiantes y docentes a desarrollar herramientas de aprendizaje y a adaptarse, entre estas herramientas están los módulos virtuales ya que estos brindan una posibilidad de preparación educativa en distintos niveles. A menudo se requiere conocer el rendimiento de los estudiantes en este tipo de plataformas, para saber en dónde se puede mejorar la calidad de educación.

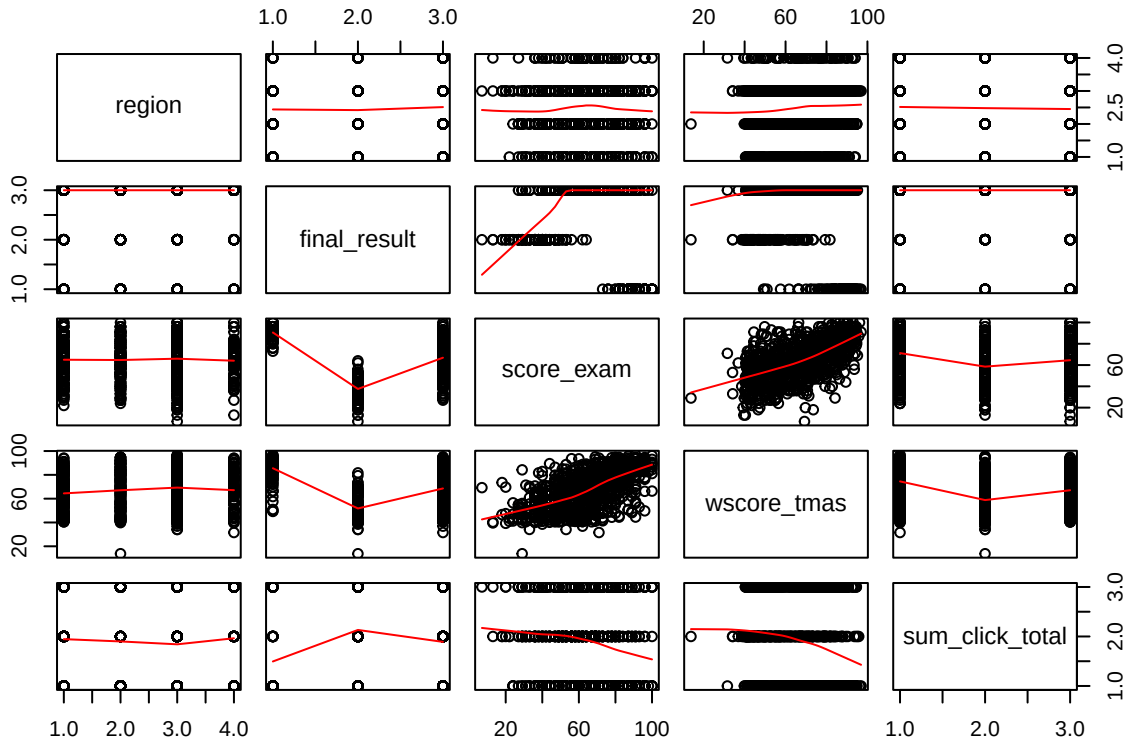
Estos módulos han permitido reconocer a los estudiantes por medio de interacciones con la plataforma. Sin embargo, se debe considerar que las calificaciones o el puntaje de un estudiante no solo depende del conocimiento que se tenga sobre el uso de estos módulos sino también existen aspectos que pueden influenciar tales como el género, la región de donde proviene, status económico, entre otros. Por medio de esta investigación se pretende obtener cuales son las variables más influyentes en la nota final de un estudiante.

Metodología

Con los conocimientos adquiridos durante el desarrollo del curso se busca crear un modelo para explicar el puntaje en la calificación final(score_exam).

Identificando variables influyentes para el modelamiento, además de verificar el cumplimiento de los supuestos, para que este sea validado y permitirnos realizar inferencias en base a un modelo final.

Se poseen los datos de la interacción de los estudiantes con el Entrono de Aprendizaje Virtual(VLE) durante los cursos impartidos por Open University Learning Analytics Dataset (OULAD) correspondiente al módulo DDD para el año 2014 durante su presentación a artir de Octubre, el conjunto de datos incluye 11 variables de las cuales mediante un graficos scatterplot se identificaron que las variables wscore_tmas, wscore_tmas, final_result son influyentes para explicar el resultado final, sus intercacciones se pueden ver en el siguiente scatterplot:



Modelos Propuestos

Se proponen los siguientes Modelos una vez se han retirado las observaciones atípicas no influyentes:

Modelo 1

El cual contiene las variables: region, final_result, wscore_tmas y sum_click_total.

$\widehat{\text{score\_exam}} = 60.002 - 1.546(\text{region}_{\text{North}}) - 2.291(\text{region}_{\text{South}}) - 2.25(\text{region}_{\text{West}}) - 40.542(\text{final\_result}_{\text{Fail}}) - 17.303(\text{final\_result}_{\text{Pass}}) + 0.381(\text{wscore\_tmas}) - 0.495(\text{sum\_click\_total}_{\text{Low}}) + 0.017(\text{sum\_click\_total}_{\text{Moderate}})$

Modelo 2

El cual contiene las variables region, final_result wscore_tmas, sum_click_total,agregando efectos de interacion entre y final_result con wscore_tmas y también entre wscore_tmas y sum_click_totalLow.

$\widehat{\text{score\_exam}} = 83.73 - 1.639(\text{region}_{\text{North}}) - 2.178(\text{region}_{\text{South}}) - 1.316(\text{region}_{\text{West}}) - 64.502(\text{final\_result}_{\text{Fail}}) - 51.867(\text{final\_result}_{\text{Pass}}) + 0.104(\text{wscore\_tmas}) + 14.88(\text{sum\_click\_total}_{\text{Low}}) + 8.72(\text{sum\_click\_total}_{\text{Moderate}}) + 0.232(\text{final\_result}_{\text{Fail}} \times \text{wscore\_tmas}) + 0.427(\text{final\_result}_{\text{Pass}} \times \text{wscore\_tmas}) - 0.233(\text{wscore\_tmas} \times \text{sum\_click\_total}_{\text{Low}}) - 0.118(\text{wscore\_tmas} \times \text{sum\_click\_total}_{\text{Moderate}})$

Modelo 3

En este se icluye un efecto cuadrático en la variable wscore_tmas y se obtiene:

$\widehat{\text{score\_exam}} = 88.135 - 2.009(\text{region}_{\text{North}}) - 2.243(\text{region}_{\text{South}}) - 1.349(\text{region}_{\text{West}}) - 74.232(\text{final\_result}_{\text{Fail}}) - 56.856(\text{final\_result}_{\text{Pass}}) + 0.037(\text{wscore\_tmas}) + 12.892(\text{sum\_click\_total}_{\text{Low}}) + 8.262(\text{sum\_click\_total}_{\text{Moderate}}) + 0.004(\text{wscore\_tmas}^2) + 0.385(\text{final\_result}_{\text{Fail}} \times \text{wscore\_tmas}) + 0.493(\text{final\_result}_{\text{Pass}} \times \text{wscore\_tmas}) - 0.203(\text{wscore\_tmas} \times \text{sum\_click\_total}_{\text{Low}}) - 0.114(\text{wscore\_tmas} \times \text{sum\_click\_total}_{\text{Moderate}})$

Obs: se representa a la variable cuadratica centrada en la media como $wscore\_tmas^2$.

Validación de los Supuestos

Para las pruebas de Normalidad se uso Jarque-Bera test mientras que para homocedasticidad se usó Breusch-Pagan Test.

En la siguiente Tabla se presentan las pruebas de normalidad, heterocedasticidad, y la media de los errores.

Table 1: Comprobación de los supuestos

	Media	Linealidad	Valor p N	Normalidad	Valor p H	Homocedasticidad
Modelo 1	0	Si	0.093	Si	0	No
Modelo 2	0	Si	0.230	Si	0	No
Modelo 3	0	Si	0.135	Si	0	No

Para corregir problemas de Heterocedasticidad se utiliza la estimación por Mínimos Cuadrados Ponderados(WLS) luego la validación de los supuestos se pueden ver en la siguiente tabla:

Table 2: Comprobación de los supuestos para Modelos(WLS)

	Media	Linealidad	Valor p N	Normalidad	Valor p H	Homocedasticidad
WLSModelo 1	0	Si	0.094	Si	0.753	Si
WLSModelo 2	0	Si	0.234	Si	0.938	Si
WLSModelo 3	0	Si	0.135	Si	0.954	Si

Ya que todos los modelos propuestos tienen los supuestos en los errores formalmente validados se procede a usar el criterio AIC() para seleccionar el "mejor" modelo.

Se obtuvieron los siguientes AIC:

Table 3: AIC de los Modelos

	Modelo1 WLS	Modelo2 WLS	Modelo3 WLS
AIC	6710.105	6614.882	6559.814

Basandonos en el criterio de selección AIC nos quedamos con el modelo Modelo3 WLS.

Modelo 3 WLS:

$\widehat{\text{score\_exam}} = 88.127 - 2.112(\text{region}_{\text{North}}) - 2.268(\text{region}_{\text{South}}) - 1.423(\text{region}_{\text{West}}) - 74.38(\text{final\_result}_{\text{Fail}}) - 56.949(\text{final\_result}_{\text{Pass}}) + 0.037(\text{wscore\_tmas}) + 13.094(\text{sum\_click\_total}_{\text{Low}}) + 8.324(\text{sum\_click\_total}_{\text{Moderate}}) + 0.004(\text{wscore\_tmas}^2) + 0.388(\text{final\_result}_{\text{Fail}} \times \text{wscore\_tmas}) + 0.495(\text{final\_result}_{\text{Pass}} \times \text{wscore\_tmas}) - 0.206(\text{wscore\_tmas} \times \text{sum\_click\_total}_{\text{Low}}) - 0.114(\text{wscore\_tmas} \times \text{sum\_click\_total}_{\text{Moderate}})$

Interpretación de los coeficientes

- $\hat{\beta}_0$: La calificación final promedio para un estudiante que pertenece al Este de la región, aprobó con distinciones, obtuvo 0 en las asignaciones marcadas por el tutor, y tuvo una participacion alta en la plataforma es 88.127.
- $\hat{\beta}_1$: La calificación final promedio para un estudiante que pertenece al Norte de la región, comparada con uno que pertenece al Este bajo las mismas condiciones es menor en 2.11 puntos.
- $\hat{\beta}_2$: La calificación final promedio para un estudiante que pertenece al Sur de la región, comparada con uno que pertenece al Este bajo las mismas condiciones es menor en 2.26 puntos.
- $\hat{\beta}_3$: La calificación final promedio para un estudiante que pertenece al Oeste de la región, comparada con uno que pertenece al Este bajo las mismas condiciones es menor en 1.42 puntos.
- $\hat{\beta}_4$: La calificación final promedio para un estudiante que no aprobó el modulo comparada con uno que aprobó con distinciones, bajo las mismas condiciones, es menor en 74.37 puntos.
- $\hat{\beta}_5$: La calificación final promedio para un estudiante que aprobó el modulo comparada con uno que aprobó con distinciones, bajo las mismas condiciones, es menor en -56.94 puntos.
- $\hat{\beta}_6$: Cuando se aumenta en una unidad la calificación en las asignaciones marcadas por el tutor bajo las mismas condiciones, la calificación final promedio aumenta en 0.03 puntos.
- $\hat{\beta}_7$: La calificación final promedio para un estudiante que tuvo una participacion baja comparada con uno que tuvo una participacion alta bajo las mismas condiciones difiere en 13.09

- $\hat{\beta}_8$: La calificación final promedio para un estudiante que tuvo una participacion moderada comparada con uno que tuvo una participacion alta bajo las mismas condiciones es mayor en 8.32.

Inferencias

Se realizan Intervalos de Cofianza Post-Seleccion y se obtiene:

	LI	LS
(Intercept)	58.5887959	117.6648329
regionNorth	-5.5401372	1.3163177
regionSouth	-5.4881661	0.9527178
regionWest	-5.9229013	3.0759705
final_resultFail	-110.1625776	-38.5973914
final_resultPass	-86.5229389	-27.3749892
wscore_tmas	-0.3178636	0.3913731
sum_click_totalLow	-0.5678047	26.7550925
sum_click_totalModerate	-5.0968732	21.7452871
I((wscore_tmas - mean(wscore_tmas))^2)	-0.0022540	0.0098015
final_resultFail:wscore_tmas	-0.1216630	0.8968611
final_resultPass:wscore_tmas	0.1401753	0.8492409
wscore_tmas:sum_click_totalLow	-0.4070985	-0.0039767
wscore_tmas:sum_click_totalModerate	-0.2992834	0.0708015

Mediante Intervalos de Post Selección se´tiene que los coeficientes significativos son los de final_resultFail, final_resultPass, el efecto de interacción final_resultPassxwscore_tmas y wscore_tmasxsum_click_totalLow.

Predicción

se desea predecir el puntaje promedio en el examen para un estudiante con baja interacción en el entorno virtual vs uno con alta participación en el entorno virtual.

Se predice que para un estudiante que tuvo una alta participación de la Región Este que aprobó y que obtuvo una calificación de 90 en las asignaciones marcadas por el tutor, tendrá una calificación en el examen final entre [75.56,82.45] el 95% de las veces, mientras que para aquel que tenga una participación baja, bajo las mismas condiciones obtendrá una calificación final entre [69.57,77.63] el 95% de las veces.