Tecnológico de Monterrey

Campus Puebla

Work's name:

Actividad 4.2 Regresión Logística

Course:

Analitica de Datos II

Student:

Ivanna Maldonado Cervantes
Paula Simonetta Madrid Pérez
Ania Diaz Gonzalez
Miranda Eugenia Colorado Arróniz
Omar Alejandro Quinn Toledo

En esta actividad aplicamos regresión logística para explorar cómo ciertas características de los proyectos (tipo, alcance, organización, etc.) Se relacionan con resultados operativos como On-hold, State, Project size, Project Health y Percent complete.

Fuente: dataset interno de proyectos (Forvia).
Limpieza básica: estandarización de formatos (por ejemplo, convertir "Percent complete" de "85%" a 85.0), manejo prudente de nulos.
Codificación: las columnas categóricas clave —*Project Type, Geographical scope, Project manager, State, Project size, Project organization, BG, Project Health, On-hold*— se codificaron para poder usarlas en el modelo.

Definición X/Y. Para cada modelo, definimos la Y (variable objetivo) y las X (predictores) según las preguntas de negocio.

Split 70/30. train_test_split con 30% para prueba.

Estandarización. StandardScaler sobre las X (la logística es sensible a escala).

Entrenamiento. LogisticRegression, probando versión balanceada (class_weight='balanced') cuando hay desbalance.

Evaluación. Matriz de confusión + precisión, recall, accuracy y F1

Matriz de confusión: muestra dónde se equivoca el modelo (qué clase confunde con cuál).

Precisión (precision): "cuando digo 'positivo', ¿qué tanto atino?".

Sensibilidad (recall): "de todos los positivos reales, ¿cuántos detecto?".

Accuracy: aciertos totales. (Se busca balancear las clases donde se requiere)

F1: promedio entre precisión y recall..

Modelo 1 — *On-hold* (sí/no)

Y: On-hold

X: Project Type, Geographical scope, Project size

Propósito: alerta temprana de pausas.

Lectura de métricas:El modelo muestra una precisión aceptable al identificar casos como "On-hold = sí", lo que indica que cuando enciende la alerta suele acertar. Sin embargo, su recall es bajo, lo que sugiere que deja fuera varios proyectos que también están en pausa.

Modelo 2 — *State* (binario)

Y: State (agrupado a 2 niveles)

X: Project size, Project Type, On-hold

Propósito: entender si tamaño/tipo y pausas se reflejan en el estado.

Lectura de métricas: El modelo presenta una precisión aceptable al identificar la clase operativa del estado, por lo que sus aciertos en esa clase son correctos. No obstante, su recall es bajo para la otra clase, indicando que omite varios casos que también deberían haberse clasificado en esa categoría.

Modelo 3 — *Project size* (SMALL vs LARGE/MEDIUM)

Y: Project size (binario: SMALL vs LARGE/MEDIUM)

X: Project manager, State, Project organization

Propósito: planeación de recursos (anticipar demanda).

Lectura de métricas: El modelo alcanza una precisión aceptable al identificar “LARGE/MEDIUM”, de modo que cuando predice ese tamaño suele atinar. Sin embargo, su recall para “SMALL” es bajo, indicando que se le escapan varios proyectos de tamaño pequeño.

Modelo 4 — *Project Health* (binario o multiclase)

Y: Project Health

X: Project organization, State, On-hold

Propósito: relacionar estructura y estatus con la salud del proyecto.

Lectura de métricas: El modelo muestra una precisión aceptable al identificar casos como "Mala", lo que significa que sus aciertos son correctos. Sin embargo, su recall es bajo, indicando que omite varios casos que también deberían haber sido clasificados como "Mala".

Modelo 5 — *Percent complete* (alto vs bajo)

Y: Percent complete (arriba vs abajo del promedio/umbral)

X: Project Health, State, On-hold

Propósito: identificar si salud/estado/pausas anticipan avance.

Lectura de métricas: El modelo ofrece una precisión aceptable cuando identifica proyectos “Arriba del umbral”, por lo que la mayoría de esos aciertos son válidos. No obstante, su recall es bajo, lo cual implica que deja fuera varios proyectos que en realidad también están por encima del umbral.

Conclusiones:

Ponderar clases ayuda. Con class_weight='balanced' suele subir el recall de la clase chica

Las variables organizacionales sí tienen correlación. *Project organization*, *manager* y *state* impactan resultados.