Proyecto Final

Sobreviviendo al Titanic

Link :

https://www.kaggle.com/t/db3638692da445e2a43f23a5b65cf354

Inscripcion:

-No se permiten Grupos.

-Identificar sus usuarios con su primer nombre y primer apellido.

Se abre la competencia encarnizada por determinar quien de los alumnos de análisis de datos 2023 logra desarrollar el mejor modelo de datos y de regresión para determinar si en base a los datos ingresados se lograría predecir si un pasajero sobreviviria al hundimiento del titanic o no.

El set de datos

Set de entrenamiento y validación

Se le harán entrega de dos set de datos. El primer set de datos se denomina dataset de entrenamiento, el cual bajo recomendación, se le solicita dividirlo en dos, un set de datos de entrenamiento puramente (podría ser el 80% de los registros) y un set de datos de validación (podría ser el 20% restante). Esto con la idea de reducir el riesgo del underfitting o overfitting. Determinar que registros van en cada dataset debe de ser aleatorio.

Las columnas incluidas en este dataset son las siguientes:

-PassengerId: Identificador único del pasajero.
-Name: Nombre y “titulo” del pasajero.
-Age: edad del pasajero.
-Sibsp: número de hermanos, hermanas, hermanastros o hermanastras en el barco.
-**Parch**: número de padres e hijos en el barco.
-Ticket: identificador del billete.
- Fare: precio pagado por el billete.
- Cabin: identificador del camarote asignado al pasajero.
- Embarked: puerto en el que embarcó el pasajero.
- Passenger class: clase del ticket del pasajero.
- Passenger sex: Sexo del pasajero.
- Passenger survived: El pasajero sobrevivio Y/N.

Como podemos observar nuestro objetivo es determinar si el pasajero sobrevivio al hundimiento del titanic o no.

Set de Evaluacion

El set entregado como de evaluacion, contiene las mismas columnas que el anterior, unicamente que no incluye la columna de passenger survived. Esta columna debe de ser calculada por su modelo y generada en un archivo csv que incluya las siguientes columnas:

-PassengerId: Identificador del pasajero.
-PassengerSurvived: Identifica mediante un 1 si el pasajero sobrevivio y un 0 si el pasajero no sobrevivio.

Obviamente todo el proceso de feature engineering efectuado sobre el primer set de datos, se debe de efectuar sobre este set de evaluacion y luego procesar este dataset y determinar si el pasajero sobrevivio o no. Esto sera generado en un archivo de texto separando ambas columnas con una coma. El mismo debe de ser subido a la pagina indicada y el mismo evaluara que tanto accuracy tuvo su modelo para predecir.

Evaluacion

La métrica de evaluación para esta competencia es [Mean F1-Score] (*https://en.wikipedia.org/wiki/F-score*). La puntuación F1, mide la precisión utilizando las estadísticas precision y recall.

Precision=TP/(TP+FP)

recall=TP/(TP+FN)

La puntuación F1 viene dada por:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

El F1 pesa por igual la precision y al recall, un buen modelo maximizará tanto la precisión como el recall simultáneamente. Por lo tanto, el desempeño moderadamente bueno en ambos se verá favorecido sobre el desempeño extremadamente bueno en uno y el desempeño deficiente en el otro.

Submission Format

Los archivos para evaluar su modelo deben de contener dos columnas:

PassengerId,passenger_survived
1,1
2,0
3,0
4,1

Evaluacion Final: (60 puntos)

Los entregables del proyecto final son los siguientes:
-Manual tecnico: especificando las decisiones tomadas en el diseño de su modelo de regresion lineal logistica.
-Codigo fuente
-Listado de experimientos efectuados para llegar a su modelo final: incluyendo listado de variables elegidas para cada modelo, score obtenido con la matriz de confusion, precision, recall y F1 score.

Los 40 puntos restantes seran otorgados de acuerdo al lugar obtenido en la competencia.