Titanic: aprendizaje automático a partir de desastres

Descripción general del dataset

Prediga la supervivencia en el Titanic usando los conceptos básicos de Machine Learning Este problema es extraído del repositorio Kaggle, y hace parte de una competición permanente en el mismo. Los datos se han dividido en dos grupos:

- conjunto de entrenamiento (train.csv)
- conjunto de prueba (test.csv)

El conjunto de entrenamiento debe usarse para construir sus modelos de aprendizaje automático. Para este conjunto, se proporcionan las etiquetas para cada pasajero. Su modelo se basará en "características" como el género y la clase de los pasajeros. También puede utilizar la ingeniería de característica (feature engineering) para crear nuevas características (esto es opcional)**.**

El conjunto de prueba debe usarse para ver qué tan bien se desempeña su modelo con datos no vistos. Para el conjunto de prueba, no se proporcionan las etiquetas para cada pasajero. Es su trabajo predecir estos resultados. Para cada pasajero del conjunto de prueba, utilice el modelo que entrenó para predecir si sobrevivió o no al hundimiento del Titanic.

También se incluye Gender_submission.csv, un conjunto de predicciones que supone que todas y solo las pasajeras sobreviven, como ejemplo de cómo debería verse un archivo de envío a la competencia de Kaggle.

Nota: Aquellos grupos que adjunten evidencia de un envío exitoso y un score mayor o igual a 0.85 tendrán una bonificación del 0.1 sobre su nota final de proyecto.

Diccionario de Datos

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarcation	C = Cherbourg, Q = Queenstown, S = Southampton

Notas sobre las variables

pclass: Un indicador del estatus socioeconómico (SES)
1st = Upper

2nd = Middle
3rd = Lower

age: La edad es fraccionaria si es menor que 1. Si la edad se estima, está en la forma xx.5.

sibsp: El conjunto de datos define las relaciones familiares de esta manera,
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: El conjunto de datos define las relaciones familiares de esta manera.,
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Algunos niños viajaron sólo con una niñera, por lo tanto parch=0 para ellos.

Objetivo

El objetivo del presente proyecto es la supervivencia en el Titanic, así como establecer una caracterización de los sobrevivientes y de aquellos que tuvieron el mismo final que Jack.

Metodología

1. **Limpieza y EDA**: compruebe si hay problemas de calidad de datos.
   Debe evaluar la calidad de los datos, así como comprender la relación entre las características y la variable de destino.

2. **Modelos predictivos**: Entrene al menos tres modelos predictivos distintos que realicen la predicción deseada. Establezca claramente el mejor modelo, de ser posible optimizando sus hiper parámetros. Debe incluir un apartado en el que establezcas los protocolos de evaluación y los procesos de formación y evaluación de los modelos.

3. **Reducción de dimensionalidad**: Considerando todas las variables, realizar un análisis de componentes principales (PCA), eligiendo el número de componentes necesarios para conservar al menos el 80% de la representación original.

4. **Caracterización de los pasajeros**: Con los datos en su nueva representación, realizar una segmentación, estableciendo el mejor número

de conglomerados K, con K > 1. Caracterizar los conglomerados referenciando a las variables originales.

Entregable: Cuaderno Jupyter Notebook con las secciones mencionadas anteriormente de acuerdo con la rúbrica dada a continuación.

Rubrica

Control de calidad de los datos	Visualización	Extracción de información de los datos	Comprensión y limpieza de datos	Protocolos de formación y evaluación	Validación de selección de los tres modelos	Reducción de dimensionalidad con PCA	Caracterización de los pasajeros	TOTAL
0.4	0.6	1.0	0.5	0.5	1.5	0.5	1.0	5.0

Referencia

Will Cukierski. (2012). Titanic - Machine Learning from Disaster. Kaggle. https://kaggle.com/competitions/titanic