UNIVERSIDAD COMPLUTENSE DE MADRID

ntic master
revolucionamos la comunicación

Machine learning

Práctica de evaluación

UCM

La idea es que participéis en un concurso de datos y que la evaluación del módulo consista justamente en ver los resultados que obtenéis en esta competición. Es una competición que está abierta de forma constante.

El conjunto de datos que hay que utilizar para hacer vuestras predicciones se basa en determinar el estado de unas bombas de agua a partir de unas 50 variables (numéricas, categóricas, etc). Es un conjunto de datos que tiene todas las dificultades que uno se puede encontrar en un conjunto de datos: missings, variables categóricas con una alta cardinalidad (muchos valores diferentes), etc.

La variable target determina los tipos de estados de las bombas (son tres). Por tanto, no vamos a aplicar modelos binarios (Si/No, Bueno/Malo, 1/0), si no multiclase. Y veréis que el número de elementos de cada clase no está balanceado. Las bombas pueden estar en estos tres estados: Funcional, no-funcional y necesita_reparación.

Para participar en este concurso tenéis que daros de alta en esta página:

https://www.drivendata.org/

Y el concurso en el que tendréis que presentar vuestros modelos es este:

https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/

Los datos de este concurso están aquí:

https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/data/

La descripción de las variables que encontraréis son el conjunto de entrenamiento está aquí:

https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/25/

La forma de concursar es la siguiente. De los datos os bajaréis tres conjuntos:

- el conjunto entrenamiento ("train")
- las etiquetas (la variable target) que hay que pegar al conjunto de entrenamiento ("labels")
- y el conjunto sobre el que haréis las predicciones ("test")
- También os podréis bajar un conjunto de datos que os muestra cómo deberéis remitir vuestras predicciones ("Submission")

Por tanto, la idea es que uséis los datos de "train/labels" para entrenar vuestros modelos. Con un modelo entrenado, haréis una predicción sobre el conjunto de "test" que os dará el estado "previsto" de las bombas.

El fichero de previsiones veréis que simplemente incluye dos columnas: los ids de las bombas "test" y vuestra previsión.

Las previsiones se suben en esta página:

https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/submissions/

Tan solo podréis subir 3 previsiones diarias, así que si lo dejáis para el final, solamente podréis hacer pocas iteraciones.

> Nota: Como aparecen en las propias condiciones del concurso, no se permite el uso de datos externos al concurso. Incluir datos de APIs, fuentes externas, etc.

Creemos que de esta forma podréis poner en práctica todo el proceso de creación de un modelo, probar con diferentes algoritmos, enfrentaros a los problemas de variables con múltiples categorías como tiene este conjunto, de completar datos faltantes, etc. El proceso de generación de un modelo en las empresas no diferirá mucho del que experimentaréis en este concurso.

Entregables:

Lo que se pide en este caso es que incluyan en un documento Notebook de Python de forma ordenada y clara las diferentes etapas (texto y código) que han desarrollado durante el concurso (sobre el modelo con el que consiguieron el mejor resultado). Además, podéis adjuntar la puntuación (score) obtenido en el concurso. Es importante que **además del código se incluyan los comentarios necesarios** que expliquen los pasos que vais realizando y las explicaciones de las transformaciones. Pensad en el trabajo como un informe de resultados que revisará una persona de Negocio de vuestra organización (queremos que tenga un cierto orden y presentación).