Máster en Big Data Analytics

Data Science

Sesión práctica: Proyecto en Data Science

Lorena Muñoz Albors
Curso 2016-2017

ÍNDICE

1. Introducción **3**

1.1 Enunciado 3

2. Planteamiento de resolución **4**

2.1 Desarrollo del modelo 4

2.2 Test 4

2.3 Otros planteamientos 5

3. Conclusiones **5**

1. Introducción

El objetivo de esta memoria es mostrar el trabajo realizado en la asignatura Data Science del master de Big Data Analytics. El trabajo consiste en comprender un problema predictivo y desarrollar soluciones. Implementar dichas soluciones y refinar el proceso predictivo en un proyecto real.

Dicho problema consiste en la predicción del número de bicicletas en una estación del sistema 'Valenbisi' de la ciudad de Valencia.

Los datos que se proporcionan son los correspondientes a:

- Datos de 8 estaciones [1-6,9,10] de Junio del 2012 a Julio del 2014.
- Datos de explotación (deployment) en 2 estaciones [7,8] de Mayo del 2014.
- Fichero para las predicciones (2 estaciones [7,8] de Junio del 2014 a Julio del 2014).

1.1 Enunciado

La tarea a realizar es la predicción con 3 horas de adelanto del número de bicicletas disponibles en las dos estaciones de test.

Para construir nuestro modelo disponemos de distancias entre estaciones, datos meteorológicos y días festivos además de otros atributos derivados que muestran el *profile* de cada estación.

2. Planteamiento de resolución

Partiendo de los datos del problema planteado, se decide en primer lugar ejecutar la función load_data que se nos proporciona para cargar todos los datos de las 8 estaciones (excepto la 7 y la 8) en un dataframe.

Una vez obtenidos todos los datos de todas las estaciones con la matriz de correlación comprobamos cual es la estación más cercana a las estaciones 7 y 8, y elegimos los datos de la estación 9 para entrenar el modelo.

Para entrenarlo vamos a usar validación cruzada de 10 pliegues y 1 repetición.

2.1 Desarrollo del modelo

En primer lugar en la ***submission1***, elegimos un modelo de regresión lineal para entrenar los datos de la estación 9 con todos los parámetros del perfil *full* (*full_profile_3h_diff_bikes, full_profile_bikes, bikes_3h_ago)* y el parámetro *bikes*.

Más tarde, realizamos la predicción del modelo junto con los datos de deployment de la estación 8 y observamos que en dicha predicción hay datos negativos y coherentes como los datos del parámetro numDocks que provocan un MAE elevado.

Por tanto, se decide pasar los negativos a un valor 0 y que en la predicción no se supere el número máximo de numDocks del estación 8.

Con este modelo obtenemos un MAE de **3.232334**

2.2 Test

Aplicamos el anterior modelo al test y para llegar a una mejor resolución del modelo aplicamos una función que redondea el número de bicicletas predicho.

2.3 Otros planteamientos

- En la ***submission2***, se decide realizar el mismo planteamiento anterior, pero elegimos un modelo de regresión lineal para entrenar los datos de la estación 9 con todos los parámetros del perfil *short* (*short_profile_3h_diff_bikes, short_profile_bikes, bikes_3h_ago)* y el parámetro *bikes*.

Con este modelo obtenemos un MAE de ***3.323583***.

- En la ***submission3***, se plantea elegir otro tipo de modelo, el K-Nearest Neighbors para entrenar los datos de la estación 9 con todos los parámetros del perfil *short* (*short_profile_3h_diff_bikes, short_profile_bikes, bikes_3h_ago)* y el parámetro *bikes*.

Con este modelo obtenemos un MAE de ***3.3317***.

3. Conclusiones

Como conclusión de este proyecto de Data Science cabe destacar que ha sido en general muy satisfactorio, ya que personalmente se han cumplido los objetivos planteados.

A través de este proyecto basado en datos reales he podido asimilar y aplicar el lenguaje R ante un problema que se me podría plantear en el ámbito profesional.