TALLER 2

Objetivo

Evaluar las capacidades del estudiante para construir modelos de Machine Learning supervisado contemplando las etapas de la metodología ASUM-DM, como son el entendimiento y la preparación de datos, entrenamiento y evaluación de modelos, y análisis de resultados.

Descripción

La alcaldía de una ciudad está interesada en implementar un nuevo servicio de patinetas para incentivar la movilidad sostenible. Como parte de este proceso, se encuentra evaluando la viabilidad económica e impacto de dicho servicio. Dado lo anterior, le ha encargado a usted la construcción de un modelo que permita predecir la demanda de patinetas al día con base en los datos de una ciudad vecina. Usted debe construir el mejor modelo de regresión que le permita predecir el número de usuarios promedio por día del nuevo servicio, al mismo tiempo de poder entender la importancia y relación de las variables consideradas. A continuación, se relaciona el diccionario de la base de datos de la ciudad vecina:

Field	Description
Instant	Record Index
Date	Date (Format: YYYY-MM-DD)
Season	Season of the year
Holiday	Is it holiday?
Weather	Description of the weather situation
Temp	Temperature in Celsius
Feel_Temp	Feeling temperature in Celsius
Hum	Normalized humidity
Wind	Wind speed in m/s
Casual	Count of casual users
Registered	Count of registered users
cnt	Count of total rental bikes including both casual and registered

Actividades

A continuación, se describen los hitos esperados por la alcaldía:

Limpieza y preparación de datos (25 pts)

Búsqueda y corrección de valores atípicos, valores faltantes y duplicados. Debido a que la base de datos no es muy grande, deberá abstenerse de eliminar registros. Así mismo, busque la oportunidad de generar nuevas variables con base en la información suministrada.

Análisis de datos (15 pts)

Analice las variables que conforman la base de datos. Realice análisis univariados y bivariados según considere. No olvide utilizar técnicas de análisis visual.

Desarrollo de modelos de regresión (30 pts)

Implemente al menos 3 modelos basados en el algoritmo de regresión lineal: uno simple, uno polinomial y uno con algún tipo de regularización.

Evaluación de modelos (20 pts)

Con base en el desempeño de cada uno de los modelos, concluya cuál es el modelo que se le debe presentar a la alcaldía. Justifique su respuesta.

Interpretación (10 pts)

El día de la presentación de resultados, una persona de la alcaldía le hace las siguientes preguntas:

- ¿Cuáles son las 3 variables más importantes para la predicción de la cantidad de usuarios?
- Describa cual es el escenario ideal para el incremento de usuarios.
- ¿Qué pasos adicionales deberían tener en cuenta para una próxima iteración/mejora del modelo?

Para tener en cuenta:

- Al consultar a un meteorólogo, este sugiere agrupar las precipitaciones en un grupo, la neblina en otro grupo y el resto en otro.
- Es de interés particular el comportamiento de los usuarios durante la semana.
- Realice imputación y corrección de variables según la lógica del negocio.

Criterios de aceptación

- El taller debe ser desarrollado individualmente.
- Debe ser entregado en los tiempos estipulados y solo a través de BloqueNeón. No se admiten entregas por otros medios como correo electrónico.
- El entregable debe consistir de un notebook subido a un repositorio público de GitHub, el cual debe incluir los outputs de la ejecución de cada celda pero también deberá poder ser ejecutado en su totalidad. En BloqueNeón se debe subir solo la URL del repositorio, no se admitirán commits posteriores a la fecha máxima de entrega.
- Dentro del notebook, haga uso de celdas de texto tipo markdown para exponer sus resultados y/o conclusiones de cada punto. También puede utilizar el archivo Readme del repositorio para concluir lo que considere necesario.
- Debe utilizar únicamente el dataset provisto en este taller.