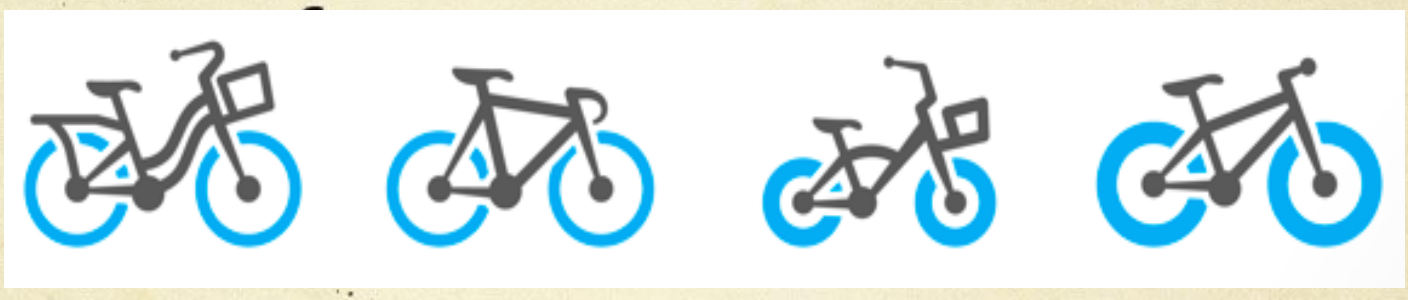
Demanda de uso de bicicletas compartidas



Camilo Yate
Nicolás Lozada
Ricardo Blanco
Alexandra Pinzon

Descripción del problema

El sistema de bicicletas compartidas de Washington, requiere estimar la demanda de uso en las diferentes estaciones del año, a través de patrones de uso históricos con datos meteorológicos, con el fin de poder : presupuestar los recursos, dimensionar la logística de funcionamiento, estimar reparaciones y proyectar ingresos. Se cuentan con las variables:

Data Fields

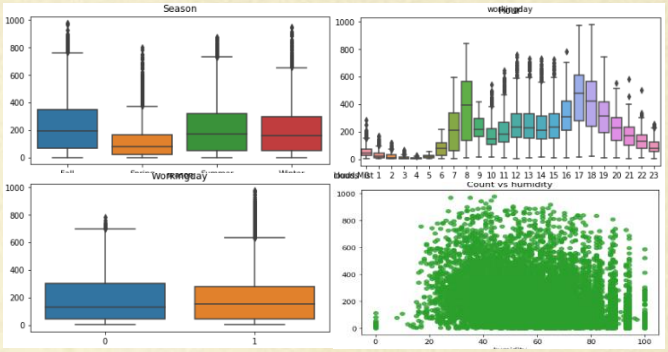
datetime - hourly date + timestamp
season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday - whether the day is considered a holiday
workingday - whether the day is neither a weekend nor holiday
weather - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp - temperature in Celsius
atemp - "feels like" temperature in Celsius
humidity - relative humidity
windspeed - wind speed
casual - number of non-registered user rentals initiated
registered - number of registered user rentals initiated
count - number of total rentals

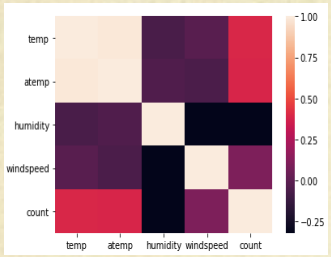
https://www.kaggle.com/c/bike-sharing-demand#evaluation

Metodologías de abordaje

- **Valores ausentes**: Se analiza la estructura de la información proporcionada y se observa que no es necesario aplicar metodologías para valores ausentes.

- **Análisis Exploratorio**: Se analiza la estructura de la información proporcionada y se observa que no es necesario aplicar metodologías para valores ausentes.





- **Correlación de variables**: Analizamos la información redundante y la posibilidad de disminuir la dimensionalidad, desidimos eliminar la avriable “atemp”

- **Transformación de variables**: realizamos cálculos a la base de datos proporcionada con el fin de mejorar el desempeño del modelo. Conversión de categóricas a Dummies, transfomración de varibales de fecha y otros.

https://github.com/camiloyatet/Curso-DeepLearning/blob/master/Proyecto1/DL%20-%20Proyecto1%20-%20Bike%20Sharing%20Demand.ipynb

Metodologías de abordaje

- **Regresiín Lineal**: Utilizando Cross Validation (10 fold) y MSE como medida de desempeño, obtenemos un indicador de error bastante alto al aplicarlo en la base de test

```
count    1.000000e+01
mean     8.323135e+19
std      2.632007e+20
min      6.253787e+03
25%      7.022949e+03
50%      1.124157e+04
75%      1.653182e+04
max      8.323135e+20
dtype: float64
```

```
count       10.000000
mean      5502.648393
std       2611.020893
min       2463.668816
25%       3447.879792
50%       4957.277942
75%       7455.395190
max      10008.328106
dtype: float64
```

- **Random Forest**: Utilizando Cross Validation (10 fold) , MSE como medida de desempeño y 1000 árboles estimadores, obtenemos resultados promedio considerablemente mejores a la regresión lineal (5.502)

- **Gradient Boost**: Utilizando Cross Validation (10 fold) , MSE como medida de desempeño, 1000 árboles estimadores y aplha = 0.01, obtenemos resultados promedio mejores que Random Forest (2.879)

```
count       10.000000
mean      2879.475720
std       1190.456769
min       1563.584886
25%       2064.976839
50%       2385.015914
75%       3871.947289
max       4793.468747
dtype: float64
```

- **Selección Final:** Teniendo en cuenta que el menor MSE se dio en Gradient Boost, lo utilizmaos para participar en la competencia. Quedando en el 30% superior de la clasificación.