Otros temas para análisis EDA

Eva Carbón, Cintia García, Paloma Panadero, Pedro Sánchez, Emilio Delgado

15-agosto-2018

1. Distribuciones estadísticas de las variables

Actualmente, tenemos los siguientes análisis en el notebook `EC_Seattle_EDA.ipynb` :

- Distribución de transacciones en horario de funcionamiento de los parquímetros
- Distribución de ocupación en horario de funcionamiento de los parquímetros
- Distribución de transacciones por parquímetro
- Distribución de ocupación por parquímetro
- Frecuencia media de actualización de ocupación

Si lo veis bien, podemos añadir algo de inferencia estadística. Por ejemplo:

- **Test Z** para saber si la *distribución de transacciones en horario de funcionamiento de los parquímetros* es una variable gaussiana, calculando el p-valor.
- **Test Z** para saber si la *distribución de temperaturas* es una variable gaussiana, calculando el p-valor.
- **Test Z** para saber si la *distribución de precipitaciones* es una variable gaussiana, calculando el p-valor.
- **Test KS y/o Test de Mann-Whitney** para saber si la *distribución de transacciones por parquímetro* se parece a la *distribución de ocupación por parquímetro*, calculando el p-valor. Esto puede ser interesante para alegar que no pertenecen a la misma distribución, y entonces los algoritmos que vamos a utilizar probablemente puedan funcionar correctamente. O, en caso de que salga significativa la diferencia, podría ayudarnos a mejorar el algoritmo. También podemos aplicarlo sobre otras variables que se nos ocurra que puedan ser parecidas.

2. Correlaciones

Podemos estudiar si hay correlaciones entre algunas de las variables y el *target*: el porcentaje de ocupación. Por ejemplo, se puede estudiar:

- Correlación entre el porcentaje de ocupación y la variable precipitación.
- Correlación entre el porcentaje de ocupación y la temperatura máxima.
- Correlación entre el porcentaje de ocupación y el número de puntos críticos cercanos.
- Frecuencia media de actualización de ocupación

Para las correlaciones, como casi todas las variables son continuas, podemos utilizar el **Test de Pearson**:

$$\rho_{X,Y} = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y}$$

3. Regresión Lineal

Enlazando con la sección anterior, también podemos hacer una regresión lineal directa entre las variables no localizadas y el porcentaje de ocupación, calculando el F-score y los valores de significación para cada coeficiente.

$$y = \beta X_1 + \beta X_2 + \cdots + \beta X_n$$

Esto nos podría aportar más información sobre qué variables son las más importantes, y podríamos intentar lanzar alguna hipótesis sobre por qué lo son.

4. Importancia de las variables

En línea con el apartado anterior, podemos arrojar más luz sobre el papel que juegan las variables estáticamente (en el tiempo y en el espacio). Podemos, por ejemplo, quitando las coordenadas espacio-temporales, hacer un *fit* con un **Random Forest** al porcentaje de ocupación, considerando cada transacción como un ejemplo más del conjunto de entrenamiento/test. Con esto, podríamos obtener la importancia de cada una de las variables lineales, antes de hacer el ajuste espacio-temporal. Con ello, quizás saquemos algún *insight* de variables importantes, que puedan ayudar a mejorar el *score*.