Predictor de bicicletas prestadas

Análisis Avanzado de Datos - Práctica Final

Alejo Martín, Arias Filippo (NIA: 100487858)

2023

I. Introducción

Los sistemas de bicicletas compartidas se han vuelto muy populares en todo el mundo debido a su capacidad para mejorar la movilidad urbana, la salud y el medio ambiente. Estos sistemas generan grandes cantidades de datos que pueden ayudarnos a comprender mejor cómo se mueven las personas en las ciudades y qué eventos importantes ocurren en ellas. En este proyecto, vamos a utilizar un conjunto de datos de un sistema de bicicletas compartidas para construir un modelo de aprendizaje supervisado que prediga la cantidad de bicicletas alquiladas en función de diferentes factores, como la hora del día, la temperatura, la humedad y las condiciones meteorológicas.

El objetivo principal de este proyecto es mostrar cómo el aprendizaje supervisado puede ser útil para predecir la demanda de bicicletas compartidas según las condiciones ambientales y temporales, lo que podría ayudar a los operadores del sistema a mejorar la disponibilidad de bicicletas y a satisfacer mejor las necesidades de los clientes. Para lograr este objetivo, vamos a seguir un enfoque de aprendizaje supervisado, utilizando un conjunto de datos etiquetados para entrenar un modelo que pueda predecir la cantidad de bicicletas alquiladas según los diferentes factores que hemos mencionado.

II. Descripción del Dataset

El dataset utilizado en este proyecto es un conjunto de registros de préstamos de bicicletas compartidas en un sistema automatizado. Contiene información sobre más de 17,000 préstamos realizados cada hora durante un período de dos años, desde enero de 2011 hasta diciembre de 2012. Cada registro contiene detalles importantes, como la fecha y hora del préstamo, las condiciones climáticas, y la cantidad de bicicletas alquiladas por usuarios registrados y ocasionales.

El dataset consta de 15 columnas que proporcionan información útil sobre los patrones de préstamo de bicicletas. Las primeras columnas incluyen detalles sobre el día y la hora del préstamo, así como la temporada, el año, el mes, el día de la semana y si es un día laborable o festivo. También se registran la temperatura, la sensación térmica, la humedad y la velocidad del viento. Además, la columna "weathersitïndica las condiciones climáticas generales en la hora del préstamo. Finalmente, se incluyen tres columnas que registran el número de bicicletas alquiladas por usuarios ocasionales, usuarios registrados y el número total de bicicletas alquiladas.

- **Instant**: Este es el índice de registro y no tiene un valor informativo por sí mismo.
- **Dteday**: Esta variable indica la fecha en que se realizó el préstamo. La fecha está en formato año-mes-día.
- **Hr**: Esta variable indica la hora del día en que se realizó el préstamo.
- **Season**: Esta variable indica la temporada en la que se realizó el préstamo. Los valores son 1 (primavera), 2 (verano), 3 (otoño) y 4 (invierno).
- **Year**: Esta variable indica el año en que se realizó el préstamo.
- **Mnth**: Esta variable indica el mes en que se realizó el préstamo.
- **Holiday**: Esta variable indica si el día en que se realizó el préstamo era un día festivo (1) o no (0).
- **Weekday**: Esta variable indica el día de la semana en que se realizó el préstamo. Los valores son 0 (domingo) a 6 (sábado).
- **Workingday**: Esta variable indica si el día en que se realizó el préstamo era un día laborable (1) o no (0).
- **Weathersit**: Esta variable indica las condiciones climáticas generales en la hora del préstamo. Los valores son 1 (despejado), 2 (nublado), 3 (lluvia ligera/nieve ligera) y 4 (lluvia intensa/nieve intensa).
- **Temp**: Esta variable indica la temperatura en grados Celsius en el momento del préstamo.
- **Atemp**: Esta variable indica la sensación térmica en grados Celsius en el momento del préstamo.
- **Hum**: Esta variable indica la humedad relativa en el momento del préstamo.
- **Windspeed**: Esta variable indica la velocidad del viento en km/h en el momento del préstamo.
- **Casual**: Esta variable indica el número de bicicletas alquiladas por usuarios ocasionales en la hora del préstamo.
- **Registered**: Esta variable indica el número de bicicletas alquiladas por usuarios registrados en la hora del préstamo.
- **Count**: Esta variable indica el número total de bicicletas alquiladas en la hora del préstamo (es decir, la suma de los valores de las variables Casual y Registered).

III. Descripción del problema

En este proyecto, nuestro objetivo es predecir el número de bicicletas que se alquilarán en una hora determinada utilizando un modelo de aprendizaje supervisado. Para lograr esto, vamos a considerar varias variables que podrían influir en la cantidad de bicicletas alquiladas.

Una de las variables más importantes es la hora del día, ya que se espera que la cantidad de bicicletas alquiladas varíe a lo largo del día debido a los patrones de movilidad de las personas. Es posible que haya más alquileres de bicicletas durante las horas pico de la mañana y la tarde, cuando la gente se dirige al trabajo o regresa a casa.

Otra variable importante es el clima. Las condiciones climáticas pueden tener un gran impacto en la cantidad de bicicletas alquiladas. Es posible que haya menos alquileres de bicicletas en días lluviosos o fríos, mientras que en días soleados y cálidos se pueden alquilar más bicicletas.

Además, las variables de temperatura, humedad y velocidad del viento también pueden tener un efecto en la cantidad de bicicletas alquiladas. Es posible que las personas estén menos dispuestas a alquilar bicicletas en días extremadamente calurosos o húmedos.

Finalmente, es importante tener en cuenta las variables relacionadas con los usuarios, como la cantidad de usuarios registrados y ocasionales. Es posible que los patrones de alquiler difieran según el tipo de usuario y que los usuarios registrados alquilen bicicletas con más frecuencia que los usuarios ocasionales.

IV. Análisis de los datos

	cnt
instant	0.278379
weekday	0.0268999
workingday	0.0302844
weathersit	-0.142426
atemp	0.400929
windspeed	0.0932338
hum	-0.322911
casual	0.694564
registered	0.972151
cnt	1

La tabla muestra la correlación entre la variable dependiente *cnt* (número total de bicicletas alquiladas) y las diferentes variables independientes en el conjunto de datos.

Se puede observar que la variable más fuertemente correlacionada con *cnt* es *registered* (correlación positiva de 0.972), que representa el número de usuarios registrados que alquilan bicicletas. La variable *casual* (correlación positiva de 0.695) también tiene una fuerte correlación con *cnt*, lo que indica que los usuarios no registrados también tienen un impacto significativo en el número total de bicicletas alquiladas.

Además, se puede observar una correlación positiva moderada entre *cnt* y *atemp* (0.401), que representa la temperatura ajustada. Esto sugiere que los usuarios tienen más probabilidades de alquilar bicicletas en días con temperaturas más cálidas.

Por otro lado, se puede observar una correlación negativa moderada entre *cnt* y *hum* (-0.323), que representa la humedad. Esto sugiere que los usuarios tienen menos probabilidades de alquilar bicicletas en días más húmedos.

Finalmente, *weathersit* (condiciones meteorológicas) y *windspeed* (velocidad del viento) tienen una correlación débil y negativa con *cnt*, lo que sugiere que estos factores no tienen un impacto significativo en el número total de bicicletas alquiladas.

Para obtener los histogramas, utilizamos la función hist de la librería matplotlib, que nos permite visualizar la distribución de cada variable. Para obtener los gráficos de líneas correlacional, utilizamos la función plot de la misma librería, pasando como parámetros las variables a graficar y la variable cnt. Finalmente, para obtener la matriz correlacional, utilizamos la función heatmap de la librería seaborn.

Estos gráficos nos permiten visualizar las relaciones entre las variables y la variable cnt. Los histogramas nos muestran la distribución de cada variable, mientras que los gráficos de líneas correlacional nos permiten ver la relación entre cada variable y cnt. La matriz correlacional nos muestra la relación entre todas las variables.

De esta forma, podemos relacionar las conclusiones obtenidas en la última tabla con los gráficos obtenidos. Por ejemplo, la correlación positiva entre la variable atemp y cnt se refleja en el gráfico de líneas correlacional correspondiente, donde podemos ver una tendencia positiva entre ambas variables. Del mismo modo, la correlación negativa entre hum y cnt se refleja en el gráfico correspondiente, donde podemos ver una tendencia negativa entre ambas variables. Los histogramas nos muestran la distribución de cada variable y nos permiten ver si hay valores atípicos o si los datos siguen una distribución normal. La matriz correlacional nos muestra todas las correlaciones entre las variables, lo que nos permite tener una visión general de la relación entre todas las variables en conjunto.

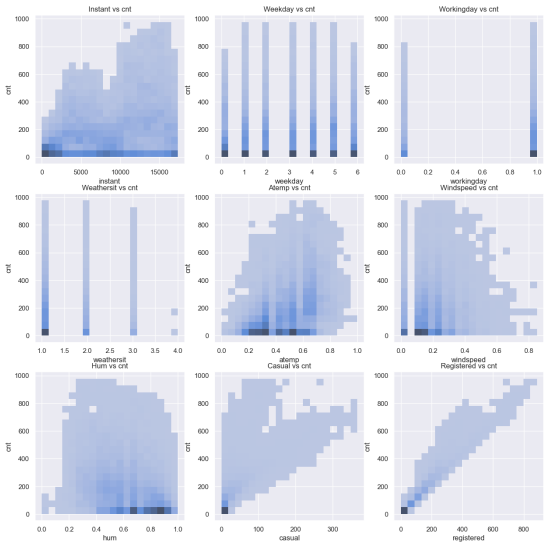

Figura 1. Histogramas

V. Método

1. Aprendizaje supervisado

En primer lugar, dividiremos nuestros datos en un conjunto de entrenamiento y un conjunto de prueba. A continuación, aplicaremos diferentes técnicas de regresión, como la regresión lineal o la regresión polinómica, para modelar la relación entre las variables independientes (día de la semana, hora del día, temperatura, etc.) y la variable dependiente (número de bicicletas alquiladas).

Una vez que hayamos entrenado nuestro modelo de regresión, evaluaremos su rendimiento utilizando diferentes métricas, como el error cuadrático medio o el coeficiente de determinación. Si nuestro modelo tiene un buen rendimiento en el conjunto de prueba, lo utilizaremos para hacer predicciones sobre la demanda futura de bicicletas.

2. Aprendizaje no supervisado

Se utilizaran ténicas de clustering y reducción de dimensionalidad. En el caso del clustering, agruparíamos los datos en diferentes grupos o clústeres según su similitud, lo que nos permitiría identificar patrones y características comunes en los datos. Esto podría ser útil para segmentar la demanda de bicicletas y detectar posibles oportunidades de negocio.

Por otro lado, la reducción de dimensionalidad nos permitiría reducir el número de variables o características en nuestros datos sin perder información importante. Esto podría ser útil si tenemos un gran número de variables y queremos simplificar nuestro modelo de regresión sin perder precisión en nuestras predicciones.

VI. Resultados

1. Regresión lineal

Se ha construido un modelo de regresión lineal para predecir el número de bicicletas prestadas en función de diversas características, como el día de la semana, si es un día laborable, las condiciones meteorológicas, la temperatura, la velocidad del viento, la humedad y la cantidad de usuarios casuales y registrados.

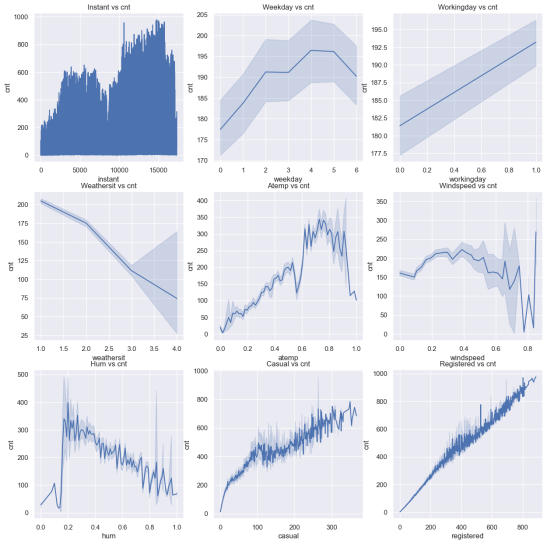

Figura 2. Gráfico de líneas

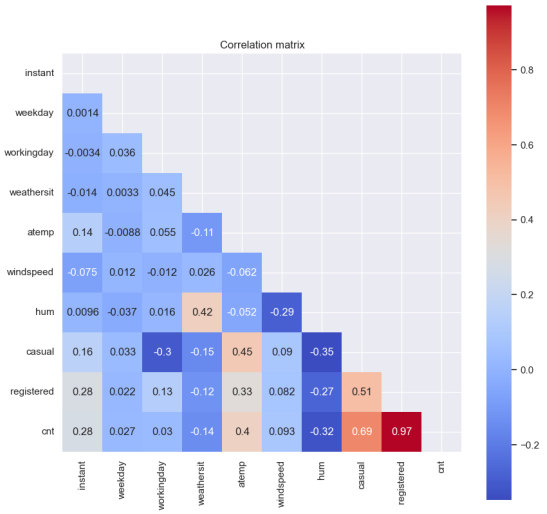

Figura 3. Matriz correlacional

El modelo se entrenó utilizando un conjunto de datos históricos y se evaluó en función de su capacidad para predecir el número de bicicletas prestadas en un conjunto de prueba separado. Los resultados de esta evaluación se resumen a continuación:

Error cuadrático medio (MSE): $8{,}209062916187086e-22$ Coeficiente de determinación (R^2): 1.0

El error cuadrático medio (MSE) es una medida de cuán cerca están las predicciones del modelo a los valores reales. Un MSE más bajo indica un mejor ajuste del modelo. En este caso, el MSE es extremadamente bajo ($8{,}209062916187086e-22$), lo que sugiere que el modelo se ajusta muy bien a los datos.

El coeficiente de determinación (R^2) es una medida de cuánta variación en los datos es explicada por el modelo. Un R^2 de 1.0 indica que el modelo explica toda la variación en los datos, lo que sugiere que las predicciones del modelo son perfectas. En este caso, el R^2 es 1.0, lo que indica un ajuste perfecto del modelo a los datos.

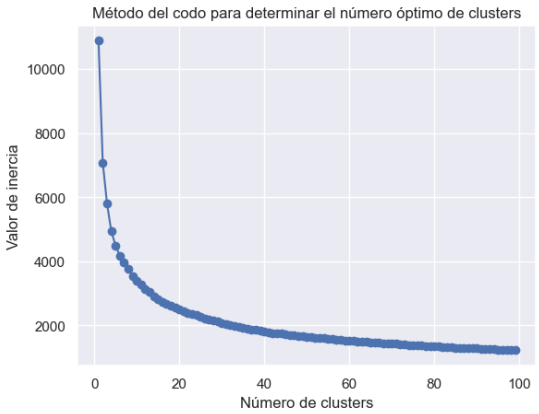

Figura 4. Elbow method

2. Clustering y Reducción de Dimensionalidad

En esta sección, se realiza un análisis de clustering y reducción de dimensionalidad en el conjunto de datos de préstamo de bicicletas. El objetivo de este análisis es encontrar patrones o agrupaciones en los datos que puedan ser útiles para comprender mejor las características de los préstamos de bicicletas.

Para llevar a cabo el análisis de clustering, se utiliza el algoritmo K-Means, que es un método de aprendizaje no supervisado para identificar agrupaciones en conjuntos de datos basados en la similitud de las características. Antes de aplicar el algoritmo, se normalizaron los datos para que todas las características tengan el mismo rango de valores.

Se empleó el método del codo (elbow method) para determinar el número óptimo de clusters. Este método consiste en calcular la suma de las distancias al cuadrado de cada punto al centroide más cercano (inercia) para diferentes valores de k, siendo k el número de clusters. Luego, se grafica la inercia en función de k y se busca un çodo.[en] la gráfica que indique un número óptimo de clusters.

Para la reducción de dimensionalidad, se utilizó el Análisis de Componentes Principales (PCA), que es una técnica para transformar un conjunto de datos de alta dimensión en uno de menor dimensión. Esto se logra proyectando los datos en un nuevo espacio formado por las direcciones (componentes principales) en las que los datos tienen la mayor variación. La reducción de dimensionalidad permite visualizar y analizar los datos de manera más eficiente y puede revelar patrones ocultos en los datos.

Después de aplicar el algoritmo de clustering y reducción de dimensionalidad, se obtuvieron las agrupaciones y se visualizaron en un gráfico de dispersión bidimensional.

VII. Evaluación del modelo de aprendizaje no supervisado

Para evaluar la calidad del modelo de clustering, se utilizaron métricas internas, ya que no se disponía de etiquetas verdaderas para comparar con las etiquetas de clustering.

1. Coeficiente de silueta promedio

El coeficiente de silueta promedio obtenido fue de 0.1853. Esta métrica varía entre -1 y 1, siendo 1 el mejor valor posible. Un coeficiente de silueta cercano a 1 indica que los puntos están bien agrupados y

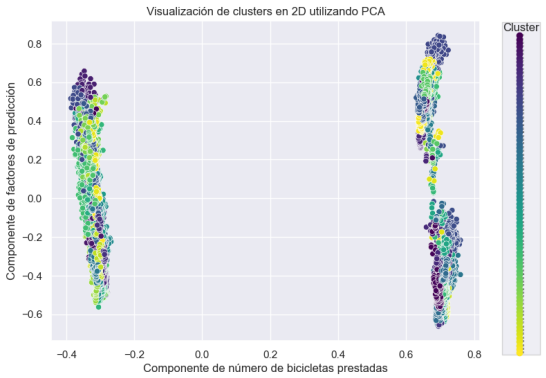

Figura 5. Análisis de componentes principales

los clusters están bien separados. Sin embargo, un valor de 0.1853 sugiere que la calidad del clustering podría mejorarse. La estructura de los clusters podría no ser adecuada o el número de clusters óptimos elegido podría no ser el ideal para los datos en cuestión.

2. Inercia

La inercia, que es la suma de las distancias al cuadrado dentro de los clusters, fue de 1304.70. Cuanto menor sea la inercia, mejor será el modelo de clustering, ya que esto indica que los puntos dentro de cada cluster están más cerca entre sí. Aunque la inercia por sí sola no proporciona una medida absoluta de la calidad del clustering, se puede utilizar para comparar diferentes modelos de clustering o para realizar una búsqueda de parámetros, como seleccionar el número óptimo de clusters.

En resumen, las métricas internas sugieren que el modelo de clustering podría mejorarse. Se podría experimentar con diferentes técnicas de preprocesamiento de datos, ajustar el número de clusters o probar otros algoritmos de clustering para mejorar la calidad del modelo.

VIII. Discusión crítica

Referencias