UTPL

La Universidad Católica de Loja

Maestría en Inteligencia Artificial Aplicada

Proyecto Final Grupo 1

Análisis de datos y visualización

Docente: PhD. Janneth Chicaiza Espinosa

Integrantes:

-Aizprua Barrios Jaris Surya
-Ramírez Velastegui Mónica Alexandra

1 Introducción

En este proyecto, se implementan diversas técnicas adquiridas a lo largo del módulo, tales como:

- Análisis exploratorio de datos,
- Visualización de datos
- Preparación de datos,
- Creación de modelos,
- Evaluación e interpretación de resultados.

2 Elección del dataset

Elegimos un dataset disponibles en el catálogo de la Universidad de Irving

https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand

El conjunto de datos contiene el recuento de bicicletas públicas alquiladas por hora en el Sistema de bicicletas compartidas de Seúl, con los datos meteorológicos e información de vacaciones.

8760 registros

3 Dataset

Número de bicicletas rentadas o alquiladas

Hora del día, Temperatura, Humedad, Velocidad del viento, Visibilidad, Punto de rocío, Radiación solar, Cantidad de lluvia, Cantidad de nieve, Temporada, Dia festivo, Dia hábil

Date	ented Bike C	Hour	Temperature(	Humidity(%)	Wind speed (r	Visibility (10m	Dew point ter	Solar Radiatic	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0	0	0	Winter	No Holiday	Yes
01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0	0	0	Winter	No Holiday	Yes
01/12/2017	173	2	-6	39	1	2000	-17.7	0	0	0	Winter	No Holiday	Yes
01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0	0	0	Winter	No Holiday	Yes
01/12/2017	78	4	-6	36	2.3	2000	-18.6	0	0	0	Winter	No Holiday	Yes
01/12/2017	100	5	-6.4	37	1.5	2000	-18.7	0	0	0	Winter	No Holiday	Yes
01/12/2017	181	6	-6.6	35	1.3	2000	-19.5	0	0	0	Winter	No Holiday	Yes
01/12/2017	460	7	-7.4	38	0.9	2000	-19.3	0	0	0	Winter	No Holiday	Yes
01/12/2017	930	8	-7.6	37	1.1	2000	-19.8	0.01	0	0	Winter	No Holiday	Yes
01/12/2017	490	9	-6.5	27	0.5	1928	-22.4	0.23	0	0	Winter	No Holiday	Yes
01/12/2017	339	10	-3.5	24	1.2	1996	-21.2	0.65	0	0	Winter	No Holiday	Yes
01/12/2017	360	11	-0.5	21	1.3	1936	-20.2	0.94	0	0	Winter	No Holiday	Yes
01/12/2017	449	12	1.7	23	1.4	2000	-17.2	1.11	0	0	Winter	No Holiday	Yes
01/12/2017	451	13	2.4	25	1.6	2000	-15.6	1.16	0	0	Winter	No Holiday	Yes
01/12/2017	447	14	3	26	2	2000	-14.6	1.01	0	0	Winter	No Holiday	Yes
01/12/2017	463	15	2.1	36	3.2	2000	-11.4	0.54	0	0	Winter	No Holiday	Yes
01/12/2017	484	16	1.2	54	4.2	793	-7	0.24	0	0	Winter	No Holiday	Yes
01/12/2017	555	17	0.8	58	1.6	2000	-6.5	0.08	0	0	Winter	No Holiday	Yes

Variable a predecir

Variables independientes

4 ANALISIS EDA

- Carga de datos de dataset
- Verificar tipos de datos del datafrase y modificar la variable fecha a Date
- Ver un resumen estadístico del dataframe
- Verificar valores faltantes y obtiene el total de registros y el porcentaje
- Detección y visualización de valores atípicos
- Aplicación de técnicas EDA orientadas a determinar problemas de calidad en los datos
- Verificar valores faltantes y obtiene el total de registros y el porcentaje y elimina los registros nulos
- Visualizar datos

5 Aplicación de modelos

- Modelo de predicción mediante Regresión Lineal
- Modelo de predicción mediante Regresión Múltiple
- Modelo Random Forest
- Modelo Gradient Boosting
- Modelo Tuned Random Forest

6 Conclusiones

Análisis de modelos

A partir de los resultados obtenidos, podemos hacer las siguientes observaciones sobre los modelos utilizados para predecir la variable 'Rented Bike Count':

•Linear Regression: El modelo de regresión lineal tiene el peor rendimiento entre todos los modelos evaluados, con una MAE de 330.39, un MSE de 194288.21 y un R2 de 0.53. Esto sugiere que el modelo no captura bien la complejidad de los datos.

•Random Forest: El modelo Random Forest muestra una mejora significativa con una MAE de 144.72, un MSE de 57734.51 y un R2 de 0.86. Esto indica que el modelo es capaz de capturar mejor las relaciones no lineales en los datos.

•Gradient Boosting: El modelo Gradient Boosting tiene un rendimiento ligeramente inferior al de Random Forest con una MAE de 173.77, un MSE de 69732.22 y un R2 de 0.83. Aunque es mejor que la regresión lineal, no supera a Random Forest.

•Tuned Random Forest: Después de ajustar los hiperparámetros, el modelo Random Forest ajustado tiene el mejor rendimiento con una MAE de 100.13, un MSE de 30608.01 y un R2 de 0.93. Esto muestra que la optimización de los hiperparámetros puede mejorar significativamente el rendimiento del modelo.

7 Recomendaciones

Se sugiere:

•Utilizar Modelos No Lineales: Para problemas similares, se recomienda utilizar modelos no lineales como Random Forest y Gradient Boosting, ya que pueden capturar relaciones complejas en los datos mejor que los modelos lineales.

•Optimización de Hiperparámetros: Siempre considere la optimización de hiperparámetros para obtener el mejor rendimiento del modelo. Herramientas como RandomizedSearchCV y GridSearchCV pueden ser muy útiles para este propósito.

•Feature Engineering: Investigar y crear nuevas características (features) podría ayudar a mejorar aún más el rendimiento del modelo.

•Evaluación Continua: Es importante reevaluar y ajustar los modelos periódicamente con nuevos datos para asegurarse de que el rendimiento del modelo sigue siendo óptimo.

•Ensemble Methods: Considerar la combinación de múltiples modelos (por ejemplo Bagging) puede ofrecer mejoras adicionales en la precisión y robustez de las predicciones.

Gracias por su Atención

UTPL

La Universidad Católica de Loja