Reporte ejecutivo

Modelo Machine Learning para sistema de Bikepro

Introducción:

Para garantizar que la tarea principal del modelo de predecir la demanda de alquiler de bicicletas en distintas ciudades contempladas para la expansión de BikePro, se utilizaron los datos históricos sobre la demanda de bicicletas en Seul, y se identificaron importantes factores climáticos numéricos como temperatura y lluvia, así como otras variables que afectan considerablemente como el dia de la semana o fechas de vacaciones y días festivos.

El diseño de un pipeline de Machine Learning automatizado, se convirtió en la mejor opción, al integrar las diversas etapas desde transformación de datos hasta entrenamiento y pruebas de modelo, nos aseguramos de que este proceso sea reproducible y permita ajustes de manera eficiente.

Preprocesamiento de datos:

En esta etapa agregamos una parte para limpiar y simplificar el nombre de las variables de cada columna de los datos de BikePro, eliminando elementos como medidas específicas de cada uno y mayúsculas en los nombres. Después implementamos la transformación Yeo-Johnson en las variables numéricas para corregir sesgos y normalizar distribución de variables. Finalmente, usamos StandardScaler para estandarizar las variables e impedir que los rangos más altos dominen el modelo y estropeen resultados

En cuanto a las variables categóricas, utilizamos el One-hot Encoding para convertir variables de texto o etiquetas en vectores numéricos binarios.

Entrenamiento y selección de modelos:

Se usó la Validación Cruzada de Series Temporales con TimeSeriesSplit y 5 cortes como el estándar recomendado al trabajar con validaciones para combatir el sobreajuste y asegurarnos que el RMSE obtenido sea real y confiable. Se usaron tres algoritmos diferentes para buscar los más óptimos hiperparámetros:

Los algoritmos fueron KNN, Regresión de Ridge y por ultimo Random Forest Regressor, que fue el mejor de los tres por ser el modelo más óptimo gracias a su capacidad de manejar interacciones complejas entre las variables numéricas o de clima, y las categóricas o de horario.

Ingienieria de características:

Basándonos en la información obtenida de BikePro, utilizamos los datos para obtener el mes y los fines de semana, para contemplar los cambios de comportamiento social.

La variable más importante que genero el mayor impacto en el desempeño del modelo fue un retraso de una hora con Variable_lag_1. Esto permitió que el modelo aprendiera el comportamiento de la demanda de la hora anterior, y generando una estimación más dinámica y real de la actual.

Conclusiones:

El objetivo principal del prototipo es llegar al máximo de 250 unidades de error RMSE. Para lograrlo se tuvieron que implementar técnicas de pre procesamiento de datos, entrenamientos de tres tipos de modelos, y de manera simultánea una validación cruzada de hiperparámetros. Además, se generaron nuevas variables partiendo de los datos iniciales, que permitían al modelo tener un contexto temporal. Implementado esto se logró reducir el valor de 480.21 a tan solo 374.11 RMSE.

Para llegar al objetivo, se agregó por último la variable count_lag_1, reduciendo drásticamente el valor RMSE a 196.12, una considerable mejoría en la precisión del modelo.