“Machine Learning. Algoritmos y modelos de predicción”

Sprint 3

Bernardo Corona Domínguez

NAO ID: 1085

Ciudad de México, México

07 de abril de 2025

Reporte Ejecutivo - Predicción de Demanda de Bicicletas (BikerPro)

1. Información disponible

Se cuenta con un conjunto de datos históricos provistos por BikePro con registros horarios de la demanda de bicicletas en la ciudad de Seúl.
El dataset incluye variables climáticas (temperatura, humedad, velocidad del viento, visibilidad, radiación solar, lluvia y nieve), variables
temporales (fecha, hora, mes, día de la semana) y categóricas (estación del año, día festivo, día de operación).
Estas variables permiten modelar el comportamiento de la demanda bajo condiciones meteorológicas y contextuales reales.

2. Hallazgos del análisis exploratorio

El análisis exploratorio reveló que la demanda de bicicletas presenta fuertes patrones estacionales y horarios.
En días laborales, la demanda se concentra entre las 7-9 am y 5-7 pm, mientras que los fines de semana el patrón es más disperso.
Además, variables como temperatura, radiación solar y días funcionales están altamente correlacionadas con la cantidad de bicicletas rentadas.
Se observó que los días festivos y con lluvia/nieve tienen una reducción notoria en la demanda.

3. Selección de variables

Las variables fueron seleccionadas con base en su relevancia demostrada en el análisis exploratorio. Se incluyeron variables numéricas
relacionadas con el clima, así como variables categóricas codificadas (como estación, día festivo y día funcional) y variables temporales
derivadas (mes y día de la semana). Esta selección asegura que el modelo pueda aprender tanto de patrones climáticos como temporales.

4. Preprocesamiento aplicado

El preprocesamiento consistió en estandarización de variables numéricas usando StandardScaler y codificación one-hot de las variables
categóricas mediante OneHotEncoder. No se aplicó imputación, ya que se verificó que el dataset no contenía valores nulos.
Las transformaciones fueron implementadas mediante un ColumnTransformer dentro de un pipeline para asegurar consistencia.

5. Evaluación del modelo

Para evitar data leakage, la división entre conjunto de entrenamiento y prueba se realizó de forma cronológica: el 80% inicial del dataset
fue usado para entrenar y el 20% restante para validar el modelo. Esta estrategia garantiza que las predicciones simulen un entorno de
predicción real sobre datos futuros. El modelo se evaluó usando RMSE (Root Mean Squared Error) y se seleccionó el modelo con menor
error en el conjunto de prueba.

6. Recursos gráficos utilizados

Se generó una gráfica temporal que compara las predicciones del modelo contra los valores reales a lo largo del conjunto de prueba.
Esto permite identificar visualmente qué tan bien el modelo sigue las tendencias reales de la demanda. También se generaron
gráficos de distribución y correlación durante el análisis exploratorio para fundamentar las decisiones de selección de variables.

7. Consideraciones sobre la API del Backlog

El modelo fue encapsulado y guardado en un archivo pickle
(`model_prediction_bikerpro.pk`) para su futura integración mediante una API,
tal como se indica en el Backlog del proyecto. Este modelo es compatible con frameworks como Flask o FastAPI, lo cual facilitaría su
exposición a través de endpoints para recibir parámetros de entrada (como clima y hora) y devolver la predicción de demanda esperada.

Esquema de modelado:

- Modelo: Bosque Aleatorio (Random Forest Regressor)
- Estrategia de validación: División entrenamiento/prueba (80% entrenamiento, 20% prueba)
- Ingeniería de características: Variables temporales (hora, mes, día de la semana), climáticas (temperatura, humedad, velocidad del viento, visibilidad, radiación solar, lluvia, nieve) y categóricas (estaciones del año, día festivo y día funcional).
- Preprocesamiento: Estandarización de variables numéricas y codificación One-Hot para variables categóricas.

Resultados obtenidos:

- RMSE Entrenamiento: 62.28
- RMSE Prueba: 173.16
- R^2 Entrenamiento: 0.991
- R^2 Prueba: 0.928

Conclusión:

El modelo obtenido ofrece buenas métricas predictivas, indicando un desempeño satisfactorio al capturar la variabilidad de la demanda de bicicletas. Se recomienda considerar posibles ajustes o adiciones de variables para reducir aún más el RMSE y mejorar la estabilidad del modelo.