# Bike Sharing Demand Prediction Using CRISP-DM and Scikit-learn - NotebookLM

Exported on: 2025/10/20 下午11:19:54

將資料中提供的解法 整理出一份摘要

資料中提供了多種共享單車需求預測與分析的解法，涵蓋了傳統時間序列模型、多元迴歸、各種機器學習與深度學習模型，以及結合了圖結構、公共交通資料，甚至是大型語言模型的創新方法。

以下是根據資料整理出的解法摘要：

**一、 經典時間序列分析與統計模型**

1. **ARIMA 模型 (AutoRegressive Integrated Moving Average)**：

◦ ARIMA 模型是一種適用於分析時間序列數據的統計模型，常用於分析小時與自行車需求之間的關係。

◦ 該模型涉及數據處理、**ADF 檢定**（檢測數據是否需要差分以達到平穩性），並透過 ACF 和 PACF 圖決定自迴歸階數 $p$ 和移動平均階數 $q$。

◦ 在一項針對首爾共享單車的預測研究中，經模型評估（RMSE, AIC, BIC），選擇了 **ARIMA(1, 0, 2)** 作為最佳模型，其預測結果與曲線趨勢一致。

2. **泊松分佈 (Poisson Distribution) 與適合度檢定**：

◦ 為了優化調度模式，需要考量顧客抵達的頻率分佈。

◦ 傳統上假設借車與還車的頻率為泊松分佈。

◦ 研究利用**卡方檢定** (Chi-square test) 分析實際數據，結果顯示在不同雨量影響下，借車的使用頻率幾乎不拒絕泊松分佈的假設；還車則除了無雨情況外，也幾乎符合泊松分佈假設。

3. **嶺迴歸 (Ridge Regression)**：

◦ 針對地鐵站周圍的自行車需求，有學者提出基於嶺迴歸的預測方法。

**二、 傳統機器學習與集成模型**

1. **CRISP-DM 流程下的多元線性迴歸 (Multiple Linear Regression)**：

◦ 這類方法將單車租賃預測視為連續數值預測（迴歸）問題。

◦ 流程嚴格遵循 **CRISP-DM** 步驟，包括商業理解、資料理解、資料準備（特徵縮放、獨熱編碼）、建模、評估與部署。

◦ **資料準備**：包括對類別特徵進行 OneHotEncoder 編碼，對數值特徵（如溫度、濕度、風速）進行 StandardScaler 標準化。

◦ **模型精進**：強調必須執行**特徵選擇** (Feature Selection)，例如使用 RFE (Recursive Feature Elimination) 或 LassoCV，並在評估階段提供**預測圖及信賴區間/誤差帶**。

2. **梯度提升模型 (Gradient Boosting Models)**：

◦這些模型（如 XGBoost、LightGBM）被廣泛應用並展現出高效能。

◦**方法論**：包括對目標變量進行對數轉換、對類別特徵編碼，並構建關鍵的**交互特徵**（如溫度與小時的複雜關係）。

◦**XGBoost**：用於評估特徵重要性並選擇最佳特徵子集。在一項針對站點級別小時需求的預測中，XGBoost 在 RMSE、MAE 和 $R^2$ 三項指標上優於 LSTM，並且能更好地預測需求高峰。

◦**LightGBM**：被證實是一種樹狀集成模型，在經過超參數優化後，於共享單車需求預測任務中表現優越。

◦**GBDT + SHAP**：結合 Gradient Boosting Decision Tree (GBDT) 模型和 Shapley Additive Explanation (SHAP) 方法，用於預測需求並分析影響因素，特別是考慮到**建成環境** (built environment) 與需求的交互作用。

3. **數據挖掘方法**：

◦有研究利用數據挖掘技術，發現溫度和小時是小時租賃次數中最顯著的變量。

**三、 深度學習與圖神經網路模型 (GNN)**

1. **LSTM (Long Short-Term Memory) 模型**：

◦這是一種特殊的循環神經網絡 (RNN)，擅長處理序列數據和**長期相關性**。

◦用於建立站點使用量預測系統，協助調度人員安排單車後勤作業。

◦在模型訓練中，會對超參數（如 Dropout, Batch size, Validation split）進行調優，以優化預測性能 (R-sq)。

2. **多尺度時空圖卷積網路 (MSTGCN)**：

◦旨在挖掘共享單車需求在不同尺度上的**多尺度時空特性**，從而準確預測公共交通出行需求。

3. **多通道時空圖卷積網路 (MC-STGCN)**：

◦這是一種新穎的深度學習方法，用於整合**公共交通 (PT) 檢票數據**、自行車需求數據和外部天氣因素。

◦**架構**：將城市交通網絡表示為圖結構，節點代表站點（自行車和地鐵站），邊緣捕捉**模式內 (intra-mode)** 和**模式間 (inter-mode)** 的連接（例如：自行車到自行車，地鐵到自行車）。

◦研究結果表明，整合多模式數據（微移動 + 公共交通 + 天氣，即 **MM-PT-W**）可以顯著增強預測準確性、收斂穩定性，並有效利用長期時間模式。

4. **STGCN 結合大型語言模型 (LLM) - STGCN-L**：

◦一個創新的深度學習框架，旨在將**非結構化的語言數據** (如 Yelp POI 點的文字描述和評論) 整合到結構化移動性模型中。

◦**特徵提取**：利用 **OpenAI GPT-4 Embeddings API** 將 POI 文本轉換為高維度向量 (1536維)，作為每個節點的空間特徵。

◦**性能**：STGCN-L 模型在 MSE 和 MAE 上優於單純的 STGCN 模型，證明了納入 LLM 提取的大眾評價特徵對理解站點間的交通動態有所幫助。

**四、 其他空間與計算方法**

1. **動態時間規整 (DTW) 與聚類分析**：

◦利用 DTW 進行聚類和預測分析，對共享單車需求進行時空分析。

2. **量子計算算法**：

- 提出使用**量子貝葉斯網絡**來預測需求，相比傳統算法，能夠提供計算加速，加快共享單車需求計算速度。