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

Exported on: 2025/10/20 下午11:22:01

將文件中所有解決方法與GPT輔助內容.pdf 使用的方法進行比較、判斷優劣並詳細說明

根據您的查詢，我將文件中提到的所有共享單車需求預測和分析方法，與您在 GPT 輔助內容中使用的**線性迴歸 (Linear Regression, LR)** 方法進行詳細比較、判斷優劣並說明。

GPT 輔助內容中使用的主要方法是：

• **線性迴歸 (Linear Regression, LR)**：一種基礎的機器學習迴歸模型，用於根據天氣、季節等特徵預測連續數值的租賃量。

以下是將文件中提及的其他解決方法，根據模型類型與 LR 進行的比較：

---

一、 傳統統計與時間序列模型

| 方法 | 描述與來源 | 與 LR 的比較與優劣判斷 (相對於 LR) |
|---|---|---|
| **ARIMA (AutoRegressive Integrated Moving Average)** | 一種統計模型，結合了自迴歸 (AR)、差分 (I) 和移動平均 (MA) 三部分。用於時間序列分析和預測，研究中用於分析首爾共享單車的每小時需求。 | **優勢:** 專門為處理時間序列數據而設計，能夠捕捉時間上的趨勢和週期性。**劣勢:** 文獻中使用的 ARIMA 版本是單變量的，僅考慮時間與需求間的關係，**未考慮天氣、特殊事件等外部因素的短期影響**。而 LR 雖然簡單，但很容易納入多變量的外部特徵，如溫度、濕度、風速等。 |
| **Poisson 分配分析** | 一種統計假設方法，用於分析顧客抵達（需求出現頻率）的分布，並使用卡方檢定進行適合度檢驗。 | **優勢:** 具備堅實的理論基礎，特別適用於計數數據（例如在某時間區間內的借車人數）的頻率分析。**劣勢:** 這是一種用於分析**需求頻率分布**的統計方法，**而非直接的預測模型**。LR 則是一個通用的預測模型，用於預測實際的租借量數值。 |

---

二、 機器學習/樹狀基礎模型

| 方法 | 描述與來源 | 與 LR 的比較與優劣判斷 (相對於 LR) |
|---|---|---|
| **XGBoost** | 一種強大的梯度提升機 (Gradient Boosting Machine) 演算法，透過迭代逐步改進預測結果，並引入正規化防止過擬合。用於回歸、分類等領域。 | **優勢 (相對於 LR):** 能夠捕捉**特徵之間的非線性關係和複雜交互作用**，準確度通常遠高於基礎的 LR 模型。LR 假設特徵間是線性關係，而 XGBoost 不受此限制。在一個研究中，XGBoost 模型的 RMSE 和 $R^2$ 表現優於 LSTM，且更擅長預測需求峰值。 |

| | | |
|---|---|---|
| **LightGBM** | 屬於梯度提升模型，經研究比較後，在預測共享單車需求任務中被證明是**性能最優越**的樹狀集成模型。 | **優勢 (相對於 LR):** 與 XGBoost 類似，LightGBM 在處理複雜的特徵工程和非線性關係時，準確度極高。該模型在優化超參數後，進一步提高了預測準確度和泛化能力。 |
| **隨機森林迴歸 (RandomForestRegressor) / 梯度提升迴歸 (GradientBoostingRegressor)** | 屬於機器學習進階模型，在 GPT 輔助內容中被提及作為 LR 的潛在升級選項。 | **優勢 (相對於 LR):** 這些模型在處理多元特徵和非線性關係時，通常比簡單的 LR 更準確、更強大。LR 模型的性能通常較為基礎。 |

---

三、 深度學習/時空圖網路模型

深度學習模型和圖神經網路（GNN）是目前用於共享單車預測的**最前沿和最複雜**的方法。它們的**核心優勢**在於能夠同時捕捉**時間依賴性**和**空間連通性**，這是 LR 模型無法做到的。

| 方法 | 描述與來源 | 與 LR 的比較與優劣判斷 (相對於 LR) |
|---|---|---|
| **LSTM (Long Short-Term Memory)** | 一種特殊的循環神經網絡 (RNN)，擅長處理序列數據和捕捉**長期時間相關性**。 | **優勢 (相對於 LR):** LSTM 專門針對時間序列的長期依賴性進行建模，在預測交通流等序列數據的動態變化時，遠優於簡單的 LR。在清華大學的專案中，LSTM 模型的表現優於 XGBoost。**劣勢:** 模型的計算複雜度更高，需要更多的數據和調參。且在另一個研究中，其性能反而不如 XGBoost。 |
| **STGCN (Spatio-Temporal Graph Convolutional Network)** | 一種深度學習框架，專門用於處理圖結構上的時空數據，將交通網路視為圖形，結合了時間和空間的卷積層。 | **優勢 (相對於 LR):** STGCN 能夠處理**站點之間的空間連通性**和相互依賴關係，這對於 LR 模型來說是無法直接建模的複雜動態。LR 僅能將經緯度作為獨立特徵納入，無法理解網路結構。 |
| **MC-STGCN (Multi-Channel STGCN)** | 一種新穎的方法，透過多通道整合地鐵檢票數據 (公共交通)、天氣數據和單車數據，以增強預測準確性。 | **優勢 (相對於 LR):** 這是處理**多模態城市交通數據**的頂級方法。實驗結果顯示，整合公共交通和天氣數據 (MM-PT-W) 後，預測準確度顯著提高，與僅使用單車數據 (MM) 相比，$R^2$ 提高了 **22%** (周預測)，且模型收斂更快，錯誤率 (MAE) 更低。LR 無法有效整合如此多層次的時空和多模態數據。 |
| **STGCN-L (STGCN + LLM)** | 結合 STGCN 與大型語言模型 (LLM)，利用 GPT-4 Embeddings API 將 POI (興趣點) 的非結構化文本數據轉換為特徵向量，納入模型進行預測。 | **優勢 (相對於 LR):** STGCN-L 解決了 LR 和許多傳統深度學習模型無法有效利用**非結構化文本數據**（如 Yelp 評論、POI 描述）的問題。雖然研究者指出 POI 語言數據帶來的改進較為溫和，但它為未來的多源異構數據整合提供了方向。 |
| **AGCRN** | 引入自適應圖卷積 (Adaptive Graph Convolution) 和遞歸網路，旨在增強捕捉時空數據中複雜關係的能力，並改進非線性與動態變化的建模。 | **優勢 (相對於 LR):** 與其他 GNN 相似，具有強大的時空建模能力。但在實證結果中，其 MSE 卻高於 STGCN 和 STGCN-L。 |

---

總結比較與判斷

| 比較維度 | 線性迴歸 (LR) (GPT 輔助內容) | 其他模型 (ML/DL/GNN) |
|---|---|---|
| **模型複雜度** | **低** | 中至高 (XGBoost/LSTM) 至 極高 (MC-STGCN/STGCN-L)。 |
| **可解釋性** | **高** (易於解釋係數和變量關係)。 | 低 (尤其是深度學習模型，如 LSTM/STGCN)。 |
| **處理關係類型** | **僅限線性關係**。 | 能處理**非線性關係**和**特徵交互作用** (XGBoost, LightGBM)。 |
| **處理時空動態** | 僅能作為多變量特徵納入 (如時間、地點)，**無法建模站點連通性**。 | **極強** (LSTM 捕捉長期時間依賴性；STGCN/MC-STGCN 捕捉空間連通性和時空交互作用)。 |
| **數據整合能力** | 擅長整合結構化的多變量數據 (天氣、季節)。 | **極強** (MC-STGCN 整合公共交通和天氣；STGCN-L 整合非結構化文本數據)。 |
| **預測準確度** | 較低，容易被更複雜的模型超越。 | 通常更高 (尤其 LightGBM、XGBoost、MC-STGCN 在各自的研究中表現優異)。 |

LR 方法的總結：

線性迴歸 (LR) 的**主要優點**在於其**簡單性**和**高可解釋性**，適合作為分析特徵 (如溫度、濕度) 對需求影響的起點。然而，LR 的**主要缺點**是其準確度上限較低，難以捕捉共享單車需求中存在的**複雜非線性、時空依賴性**和**多模態數據**的交互影響。

更優方法的總結：

1. **樹狀模型 (XGBoost/LightGBM)** 提供了 LR 的實用替代方案，它們在不犧牲過多解釋性的前提下，能夠顯著提高預測的**非線性準確度**。

2. **深度學習/圖網路模型 (MC-STGCN, STGCN-L, LSTM)** 代表了預測技術的頂峰。這些模型雖然複雜且難以調參，但它們能夠處理 LR 完全無法處理的**時空依賴性**和**多模態數據**，從而實現最高的預測準確度，對於即時單車調度等實際應用更具價值。