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

Exported on: 2025/10/20 下午11:06:15

整理出所有解決方法，並做出詳細解釋

根據來源資料，解決共享單車需求預測與系統優化的主要方法可分為統計模型、機器學習模型、深度學習與圖神經網路模型，以及基礎數據分析方法。以下將這些方法彙整並提供詳細解釋：

一、 統計學與時間序列模型

1. ARIMA 模型 (AutoRegressive Integrated Moving Average)

• **方法解釋：** ARIMA 模型是一種用於預測和時間序列分析的統計模型。它特別適用於研究小時與自行車需求之間的關係。

• **組成部分：** ARIMA(p, d, q) 包含三個主要部分：

◦ **自迴歸 (AR) 元件 (p)：** 指模型中包含的先前觀測值的線性組合。它代表當前時刻的觀測值與過去時刻的觀測值之間的關係。

◦ **差分 (I) 元件 (d)：** 為了使時間序列達到平穩（即方差和均值在序列過程中不變）所進行的差分運算次數。

◦ **移動平均 (MA) 元件 (q)：** 指模型中包含的先前期間誤差的線性組合。

• **應用目標：** 透過預測 hourly changes（每小時的變化），協助企業及時調整自行車的部署數量，降低維護成本，並避免過多共享單車影響市容。

2. Poisson 分配分析

• **方法解釋：** 這是建構需求預測模型時，考量顧客抵達分布（即需求出現頻率）的基礎分析方法。許多研究根據理論將借車與還車的頻率設為 Poisson 分配。

• **應用方式：** 透過觀察站點，蒐集實際使用者數據，並在考量不同雨量影響下，利用**卡方檢定法** (Chi-Square test) 對借車與還車的使用頻率進行適合度檢定。

• **研究結果：** 某項研究發現，在不同雨量影響下，借車與還車的使用頻率幾乎不拒絕 Poisson 分配的假設。這有助於優化調度模式。

二、 機器學習與梯度提升模型

3. 線性回歸 (Linear Regression)

• **方法解釋：** 線性回歸用於連續數值預測問題，例如根據天氣、溫度、星期幾等特徵預測自行車租借數量 (count)。

• **應用目標：** 透過量化輸入特徵與自行車數量之間的關係，幫助資源規劃和提高使用者滿意度。通常遵循 CRISP-DM 流程，並結合數據準備步驟，例如使用 OneHotEncoder 處理類別特徵和使用 StandardScaler 進行數值特徵縮放。

• **相關模型：Ridge Regression** 曾被提出用於準確掌握捷運站周邊的自行車數量。

4. XGBoost (eXtreme Gradient Boosting)

• **方法解釋：** XGBoost 是一種強大的梯度提升機 (Gradient Boosting Machine) 演算法，在回歸等領域廣泛應用。它通過迭代方式結合多個弱學習器來構建預測模型，並將新的學習器集中於改善先前預測錯誤的樣本，同時引入正規化方法來避免過度擬合。

• **應用比較：** 在一項針對站點層級每小時自行車需求的預測研究中，XGBoost 在 RMSE、MAE 和 R^2 三項指標上優於 LSTM 模型。此外，XGBoost 在預測需求高峰方面表現更好，這對於調度車隊尤為重要。它也常被用來評估特徵重要性，以選擇最佳特徵子集。

5. LightGBM (Light Gradient Boosting Machine)

• **方法解釋：** LightGBM 是一種樹狀集成模型，在共享單車需求預測任務中被證明具有卓越的性能。

• **應用流程：** 該方法通常包含詳細的特徵工程（如對目標變量進行對數轉換、編碼類別特徵、構建溫度與小時等變量之間的關鍵交互特徵）。在經過超參數優化後，LightGBM 的準確度和泛化能力會進一步提升。

三、 深度學習與圖神經網路模型

6. LSTM 模型 (Long Short-Term Memory)

• **方法解釋：** LSTM 是一種特殊的循環神經網絡 (RNN)，擅長處理序列數據和時間序列預測。

• **核心機制：** LSTM 通過引入**門控機制** (Gating Mechanisms)，能夠更好地捕捉和保持序列中的長期相關性。其結構包含細胞狀態 (cell state) 和隱藏狀態 (hidden state)。

• **應用案例：** 在 YouBike 使用量預測系統中，經比較，LSTM 模型被選定為比 XGBoost 模型表現更優異的預測方法，用於建立站點使用量預測系統。

7. STGCN 模型 (Spatio-Temporal Graph Convolutional Network)

• **方法解釋：** STGCN 是一種深度學習框架，用於交通預測，是一種典型的時間序列預測問題。它使用空間圖卷積網路來處理圖結構化的交通數據，直接構建完整的卷積神經網路，以減少參數數量並提高模型訓練速度。

• **數據處理：** 將交通網絡建立為圖結構，節點代表自行車共享站點，邊代表站點間的連接。

8. STGCN-L 模型 (STGCN Combined with Large Language Model)

• **方法解釋：** 這是對 STGCN 架構的增強，引入了**大型語言模型 (LLM)** 區塊。

• **核心創新：** 旨在解決傳統結構化模型難以整合非結構化語言數據的挑戰。LLM 區塊利用 OpenAI GPT-4 Embeddings API 等工具，從**興趣點 (POI) 的文本數據**（如 Yelp 評論和描述）中提取見解，將其轉換為高維度向量 (如 1536 維) 作為每個節點的空間特徵，以增強訓練效果。

9. MC-STGCN 模型 (Multi-Channel Spatio-Temporal Graph Convolutional Network)

• **方法解釋：** 這是一種新穎的深度學習方法，旨在通過整合多種資料流來增強預測準確性，特別是在處理城市移動性時。

• **多源數據整合：** MC-STGCN 將**共享移動性 (MM) 需求數據**、**公共交通 (PT) 檢票數據**（如地鐵刷卡數據）和**外部天氣因素**作為單獨的輸入通道進行處理。

• **圖表示：** 該模型使用圖結構表示城市交通網絡，節點代表站點（自行車和地鐵），邊捕捉**模式內** (bike-to-bike) 和**模式間** (subway-to-bike) 的連接性，並將鄰接矩陣饋送到不同的通道中。研究顯示，整合公共交通數據能顯著提高預測的穩定性和準確性。

四、 其他專門的預測方法

10. MSTGCN 模型 (Multi-scale Spatiotemporal Graph Convolutional Network)

• **方法解釋：** 專門為挖掘共享自行車需求的多尺度時空特性而提出，用於處理旅行需求在不同尺度上具有的不同空間依賴性。

11. GBDT + SHAP (Gradient Boosting Decision Tree + Shapley Additive Explanation)

• **方法解釋：** 使用 GBDT 模型預測共享自行車的出行需求，並結合 SHAP 方法分析影響需求的因素，從而為共享自行車領域的未來發展提供建議。

12. DTW + 聚類分析 (Dynamic Time Warping + Clustering)

• **方法解釋：** 基於動態時間扭曲 (DTW) 的聚類和預測分析方法，用於對共享自行車需求進行時空分析。透過時間序列聚類，可以找到具有不同屬性和使用趨勢的自行車站點群組。

13. 量子計算演算法/量子貝葉斯網路 (Quantum Bayesian Network)

• **方法解釋：** 提出使用量子貝葉斯網路來預測需求，相比傳統演算法，可以提供計算加速，加快共享自行車需求計算的速度。

14. GCN (Graph Convolutional Network)

• **方法解釋：** 一種新的深度模型，用於預測每個自行車站點的借出和歸還人數。

15. Data Mining (資料探勘)

• **方法解釋：** 使用數據挖掘方法預測並發現溫度和小時是每小時租賃自行車數量中最顯著的變數。

整理一份摘要給我

這份摘要將彙整來源資料中關於共享單車需求預測和系統優化的主要方法、應用與核心發現。

共享單車需求預測與優化方法摘要

精確預測共享單車的需求對於城市交通系統至關重要，因為它能有效解決車輛供需失衡、**降低營運成本**（例如調度成本）並**提升使用者滿意度**。為此，研究人員應用了多種從傳統統計學到尖端深度學習的模型。

一、 統計學與基礎機器學習模型

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

◦ **核心功能：** ARIMA 模型是一種時間序列統計模型，適用於研究時間（小時）與自行車需求之間的關係。透過分析和預測每小時的變化 (hourly changes)，協助企業及時調整自行車部署數量。

◦ **方法步驟：** 該方法通常涉及數據預處理、使用 **ADF 檢定**確保時間序列的平穩性 (stationarity)，並透過 ACF 和 PACF 圖來確定模型的自迴歸 (p) 和移動平均 (q) 階數。

2. **泊松分配與檢定 (Poisson Distribution Analysis)**

◦ **應用：** 在構建需求預測模型時，許多研究根據理論將借車和還車的頻率設為泊松分配 (Poisson distribution)。

◦ **驗證：** 透過實際數據蒐集和**卡方檢定法** (Chi-Square test)，發現借車與還車的使用頻率在不同雨量的影響下，**幾乎不拒絕泊松分配的假設**。

3. **線性回歸 (Linear Regression) 與數據探勘**

◦**CRISP-DM 流程：** 線性回歸模型用於連續數值預測問題（如預測租借總數），通常遵循 CRISP-DM（跨產業標準化資料挖掘流程）步驟，包括數據準備（如使用 `OneHotEncoder` 處理類別特徵和 `StandardScaler` 進行數值特徵縮放）和模型評估（如 RMSE, R^2）。

◦**關鍵因素：** 數據探勘的結果顯示，**溫度 (Temperature) 和小時 (Hour)** 是影響每小時租賃自行車數量最顯著的變數。

二、 機器學習與時序模型比較

在多個實證研究中，機器學習和深度學習模型表現出優於傳統模型的趨勢：

1. **XGBoost 與 LightGBM**

◦**模型類型：** 兩者皆為梯度提升樹 (Gradient Boosting) 模型，在回歸問題中表現優異。

◦**LightGBM 優勢：** 經過詳細的**特徵工程**（例如對目標變量進行對數轉換、構建溫度與小時之間的交互特徵）和**超參數優化**後，LightGBM 在預測共享單車需求任務中展現了卓越的準確性和泛化能力。

◦**XGBoost 優勢：** 在一項針對站點層級每小時需求的預測研究中，XGBoost 在預測**需求高峰**方面表現優於 LSTM，這對於車隊調度至關重要。

2. **LSTM 模型 (Long Short-Term Memory)**

◦**核心機制：** LSTM 是一種特殊的循環神經網絡 (RNN)，通過**門控機制**，擅長處理時間序列數據中的長期相關性和記憶問題。

◦**應用比較：** 在一個 YouBike 預測系統的實作中，研究團隊選擇了 **LSTM 模型**作為最終預測方法，因為在經過訓練成果比較後，發現其表現優於 XGBoost 模型。

三、 整合多模態數據的圖神經網絡模型 (GNN)

為了捕捉城市交通網絡複雜的時空依賴性，圖神經網路 (GNN) 被廣泛採用。

1. **STGCN (Spatio-Temporal Graph Convolutional Network)**

◦**架構：** STGCN 是一種用於交通預測的深度學習框架，它將交通網絡構建為圖結構，並直接建立**完整的卷積神經網絡**來處理時空數據，以減少參數數量並提高訓練速度。

2. **MC-STGCN (Multi-Channel STGCN) — 整合公共交通數據**

◦**創新：** 該模型旨在解決傳統模型忽略公共交通 (PT) 刷卡數據對共享單車 (MM) 需求的影響這一研究空白。

◦**多通道整合：** MC-STGCN 是一個多通道架構，將 **共享移動性數據**、**公共交通檢票數據**（如地鐵刷卡數據）和**外部天氣因素**作為**獨立的輸入通道**進行處理，而非簡單地進行特徵增強。

◦**結果：** 整合多模態數據（MM-PT-W）能顯著提高預測的穩定性和準確性，相比僅使用共享移動性數據（MM）的場景，**預測準確性提高了 15%**。

3. **STGCN-L (STGCN + Large Language Model) — 整合非結構化文本**

◦**挑戰與解決方案：** 傳統結構化模型難以整合 POI（興趣點）的**非結構化語言數據**（如 Yelp 評論或描述）。STGCN-L 透過引入**大型語言模型 (LLM) 區塊**來解決此問題。

◦**方法：** LLM 區塊利用 OpenAI GPT-4 Embeddings API 等工具，將 POI 文本數據轉換為高維度向量（例如 1536 維），作為每個節點的**空間特徵**，以提升模型對站點周邊環境的理解和預測能力。

四、 結論與應用

準確的需求預測可以幫助營運者更靈活地安排單車調度後勤作業，大幅降低站點**滿車和空車率**，並可結合後臺管理系統，提供即時天氣預報和未來使用量預測圖。未來的研究方向包括納入更多的外部因素，例如天氣條件和特殊活動的短期影響。