

Task 6

Research Canvas

白宸宇



TABLE OF CONTENTS

1.
Problem

2.
Purpose

3.
Research
Questions &
Hypotheses

9.
Drawing
Conclusions

4.
Conceptual/
Theoretical
Framework

5.
Literature
Review

8.
Data
Analysis

7.
Data
Collection

6.
Overall
Approach

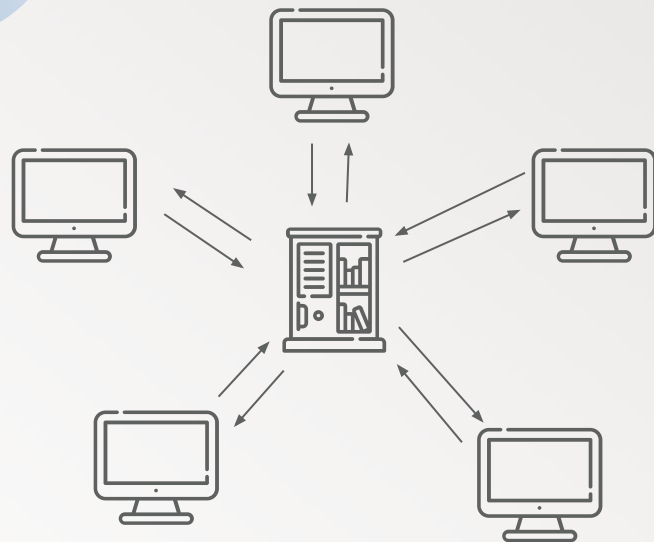
1. Problem

背景：

隨著數據驅動行銷的興起與大數據時代的來臨，企業對高品質且能保護隱私的數據需求日益增加。但礙於法律，公司之間無法直接地共享客戶的資訊，導致數據分析受限，無法充分利用行銷數據

Example: 網路電商公司搜集了客戶的資料，卻無法與廣告行銷公司協作該資料

生成式對抗網絡 (GANs) 和擴散模型提供了一種創新方式，通過合成數據來填補真實數據的不足亦或直接共享的缺點，降低數據共享中的隱私與法律風險。



研究缺口：

生成數據的真實性、代表性以及其在分析和機器學習應用中的有效性仍存在疑慮。企業擔心生成數據可能無法精確模擬真實世界數據，從而影響行銷策略的準確性。



2. Purpose

總體目標：

探討生成式數據在行銷數據分析和機器學習中的應用價值，並驗證其在保護隱私的同時，能否提供與真實數據相當的分析結果。

具體目標：

測試不同生成技術生成的數據在實際行銷應用中的有效性，並提出其在企業數據共享中的最佳實踐。

3. Question & Hypotheses

Q

**真實性
&
代表性**

- 生成數據能多大程度地還原真實數據的趨勢與特徵？
- 生成數據與真實數據在行銷分析和機器學習應用中的表現是否相當？
- 生成數據在保護客戶隱私方面的表現如何？是否能替代直接使用真實數據
- 生成數據是否能有效解決企業面臨的問題？
- 生成數據是否具備足夠的商業價值？如何實現

H

假說

- 生成數據與真實數據在統計分布上有(H1)無(H0)顯著差異，且在行銷分析中的應用結果有(H1)無(H0)顯著差異。
- 不同生成技術(如GANs和擴散模型)生成的數據在質量和應用效果上有(H1)無(H0)顯著差異。

4. Conceptual Framework



核心概念

- 機器學習
- 合成數據
- 行銷數據分析(市場細分、客戶行為預測、產品推薦)
- 數據隱私保護



研究範圍

生成模型技術作為核心工具，生成合成數據，進一步應用於行銷數據分析中。

生成數據在行銷分析中能夠提升數據精確性，並解決數據隱私問題。



應用探討

生成數據在不同行銷應用場景中的有效性，包括市場細分、客戶行為預測等。

生成出的資料在後續ML建模的效果如何

5. Literature review

GAN

簡介

<https://aws.amazon.com/tw/what-is/gan/>

CTGAN

<https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/ctgansynthesizer>

Diffusion Model

<https://medium.com/image-processing-and-ml-note/diffusion-models-b4609ff05ae6>

Copula

<https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/copulagansynthesizer>

Synthetic Data

<https://docs.sdv.dev/sdv>

TVAE

<https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/tvaesynthesizer>

相關先例

過去生成數據應用在行銷數據的先例與研究成果

現有的隱私保護方法和生成數據的應用效果。

6.Overall Approach



7. Data Collection

資料集選用

<https://www.kaggle.com/datasets/xiaojiu1414/digix-global-ai-challenge>

該資料集為紀錄廣告與客戶互動之資料集, 包含:

個人資訊數據: 紀錄不同客戶的PII, 例如年紀、性別、居住地等

廣告展示數據: 紀錄了每個廣告的展示次數、點擊次數、點擊率等

用戶行為數據: 包括用戶與廣告的互動記錄, 如點擊、轉換率等


廣告屬性數據: 包括廣告的類型、目標受眾、投放時間和地點等


後續將資料分為代表非潛在客戶行為的non_potential_customers、
潛在客戶行為的potential_customers、潛在客戶點擊廣告後行為的ads

8.Data Analysis

分別將不同合成器產生的數據做評估, 比較訓練時間、overall score的優劣、以及是否符合檢定等


Generating report ...


(1/2) Evaluating Column Shapes:  21/21 [00:01<00:00, 11.89it/s]
Column Shapes Score: 90.54%

(2/2) Evaluating Column Pair Trends:  210/210 [01:22<00:00, 2.55it/s]
Column Pair Trends Score: 76.24%

Overall Score (Average): 83.39%


Generating report ...


(1/2) Evaluating Column Shapes:  21/21 [00:01<00:00, 12.14it/s]
Column Shapes Score: 90.57%

(2/2) Evaluating Column Pair Trends:  210/210 [01:26<00:00, 2.44it/s]
Column Pair Trends Score: 76.67%

Overall Score (Average): 83.62%

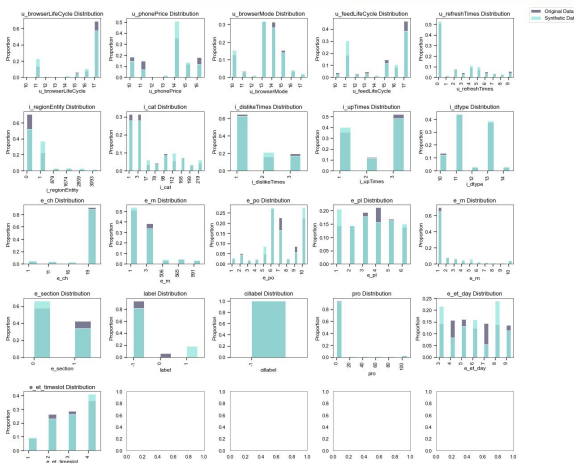
Generating report ...

(1/2) Evaluating Column Shapes:  21/21 [00:01<00:00, 12.53it/s]
Column Shapes Score: 86.96%

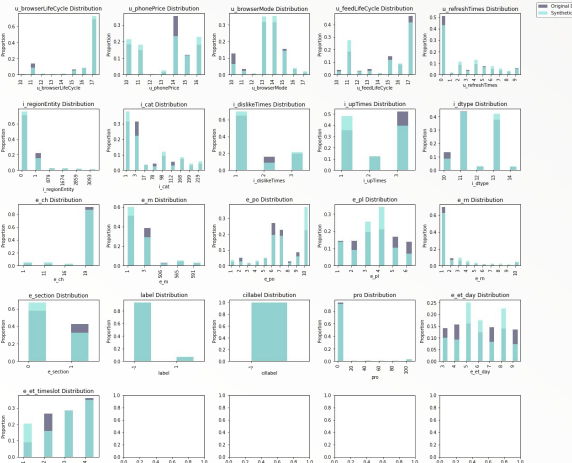
(2/2) Evaluating Column Pair Trends:  210/210 [01:22<00:00, 2.53it/s]
Column Pair Trends Score: 69.59%

Overall Score (Average): 78.28%

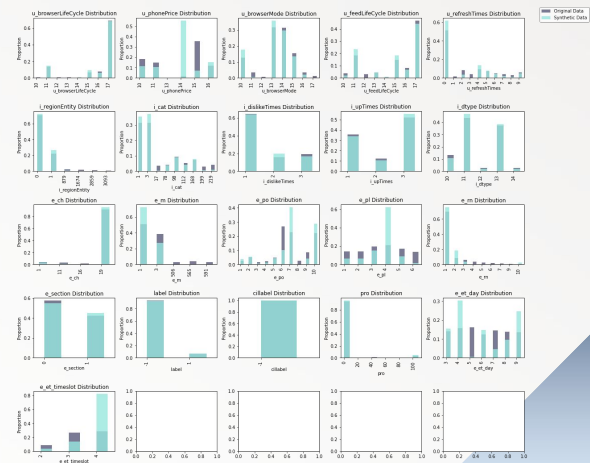
CopulaGAN



CTGAN



TVAE



9. Drawing Conclusion

研究發現

應用建議

後續研究方向



Thanks for watching