電動車資料分析報告

目錄

組員分工:

找資料: 邱威誠 邱榮材

資料前處理: 邱威誠

視覺化: 邱威誠

數據轉換: 邱威誠 邱榮材

Encoding & 跑 model: 邱榮材

後續分析: 邱榮材

Ppt: 邱威誠 邱榮材

書面一到四: 邱威誠

書面五到七: 邱榮材

一、前言

本報告旨在分析電動車的數據，從而深入了解年分、製造商、電動里程、建議零售價等多個變數對電動車市場的影響。我們將探討如何使用標準化和編碼技術處理資料，並利用機器學習模型進行預測，以助於未來電動車的市場需求評估。

1. 目標:

利用過去的電動車數據來了解電動車現在有多普遍，預測它們在市場上將增長多少，並發現這種成長背後的主要趨勢和動力

2.資料來源:

https://www.kaggle.com/datasets/rajkumarpandey02/electric-vehicle-population-data/code

3.欄位介紹:

VIN （1-10）	車輛識別號	Postal Code	郵遞區號
County	縣	Model Year	型號年分
City	城市	Make	製造商
State	州	Model	車型
Electric Vehicle Type	電動車類型	Clean Alternative Fuel Vehicle	清潔替代燃料車輛計劃的資格狀態
Legislative District	立法區	DOL Vehicle ID	許可車輛識別部門
Electric Range	電力續航里程(一顆電池可以跑多遠)	Vehicle Location	車輛位置
Base MSRP	建議零售價	Electric Utility	電力公司
2020 Census Tract	2020 年人口普查區域		

二、資料前處理

1. 資料清理:

首先我們先檢查每個欄位的缺失值情況，發現少數的缺失值，部分欄位與我們想預測的無關，所以選擇刪除

- 過少缺失值: 建議售價 電力里程
- 較不相關: 郵遞區號 車輛立法區 車輛位置 2020 年人口普查區域
- 刪除整個欄位: County

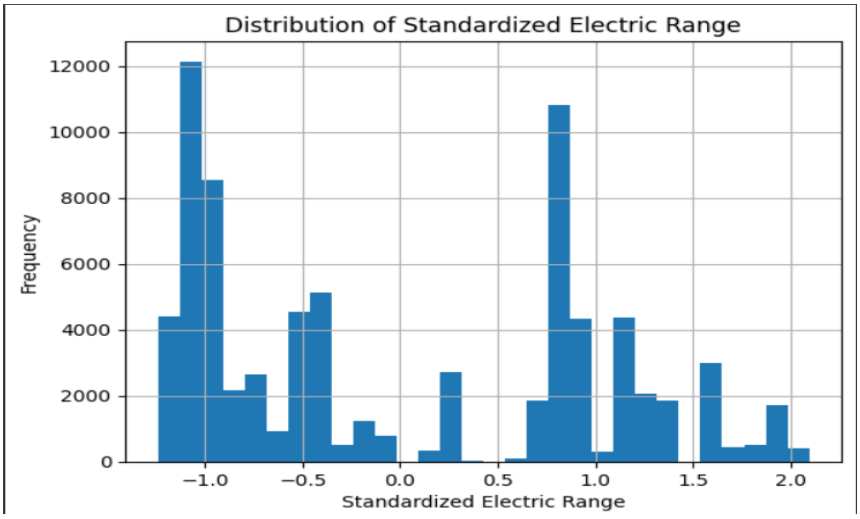
而 Model 欄位我們選擇用眾數「model 3」 去填補。再來發現到「City」及「Model」有著過多的列，我們把數量小於 5 的合併為「Other」減少過多不必要的資訊。

下一步檢查各欄位的數據範圍，以確保沒有異常值存在。例如，電動里程應該在合理範圍內，過高或過低的值可能意味著數據輸入錯誤或異常值。而我們找出五萬多筆 0，又根據程式碼結果可以知道「Electric Range」= 0 的時候，「Base MSRP」也會是 0，可能是新車或是沒有相關數據，所以選擇除去「Electric Range」= 0 的 index。

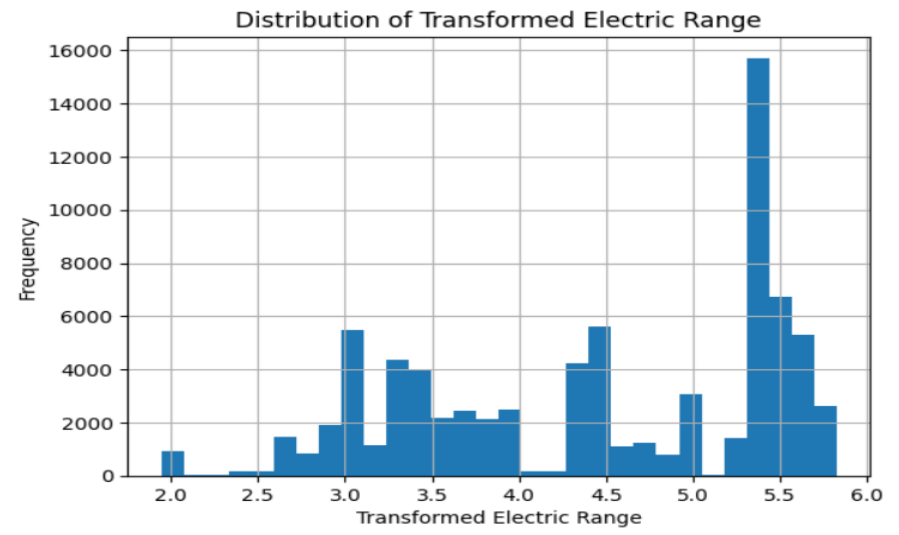
2. **數據轉換：**

對清理過後的「Electric Range」畫圖，發現為左偏的型態，所以我們對它進行轉換

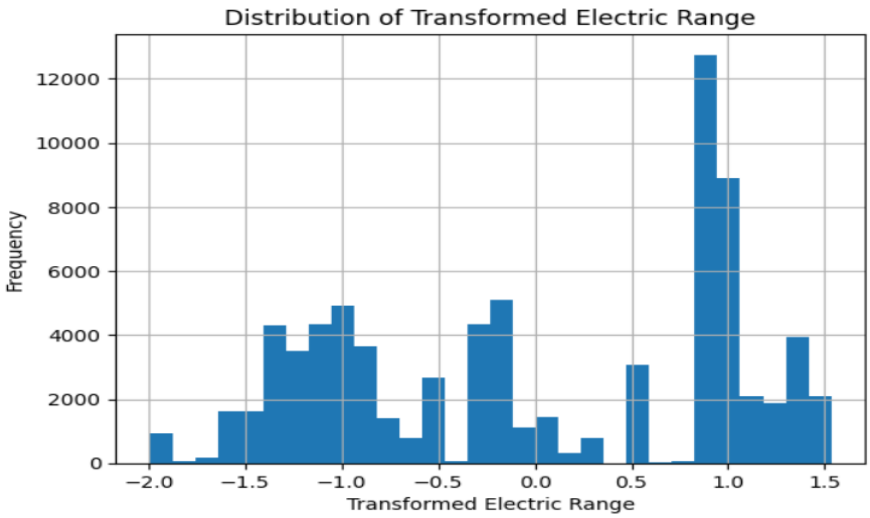
(1). 標準化



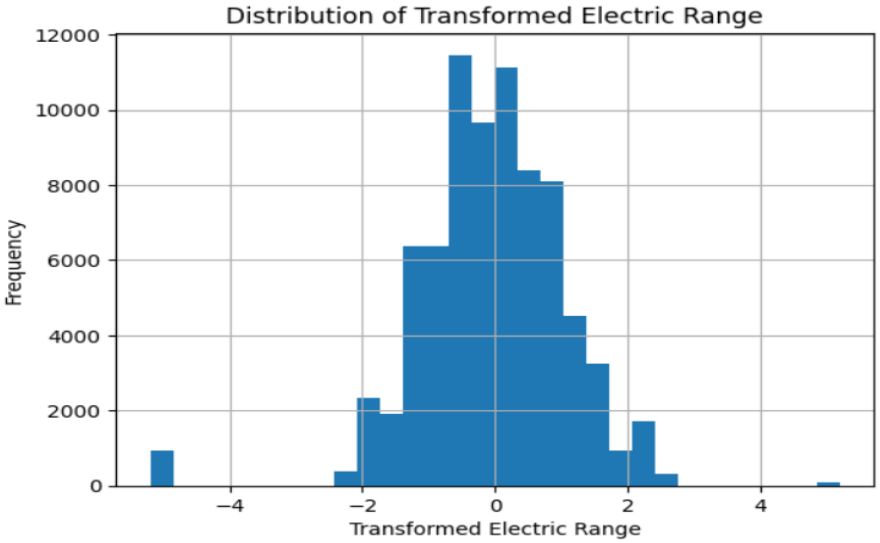
(2). 取 Log



(3). PowerTransform



(4). QuantileTransformer



由上述傳換可以看出 QuantileTransformer 最接近鐘形分布

三、資料 Encoding

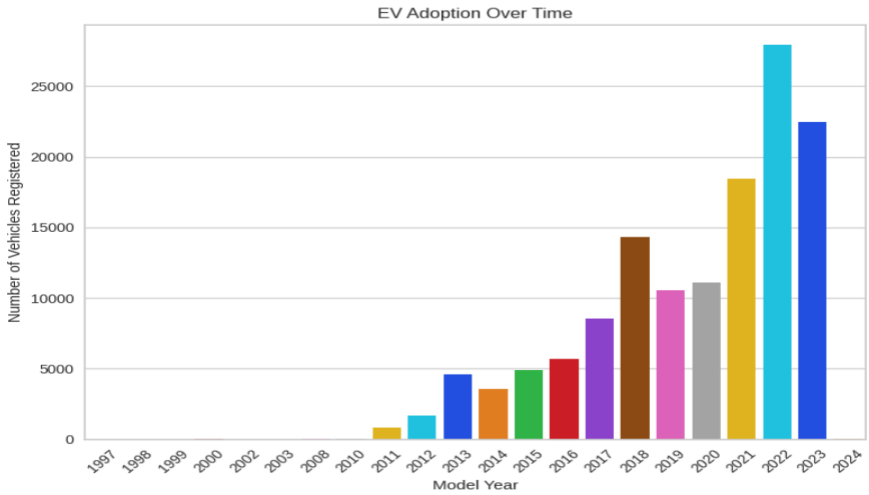
在這裡，我們採用 Target Encoding 方法，將類別變數（Make 和 Model）轉換為數值變量並作為特徵，並使用 ElectricRange 列作為目標變數，corss-validation 為 5，使它們更適合用於回歸分析以及後續的模型。

四、資料視覺化

通過數據視覺化，幫助我們更容易理解數據間的關係，以及每個欄位的趨勢及成長等關聯。

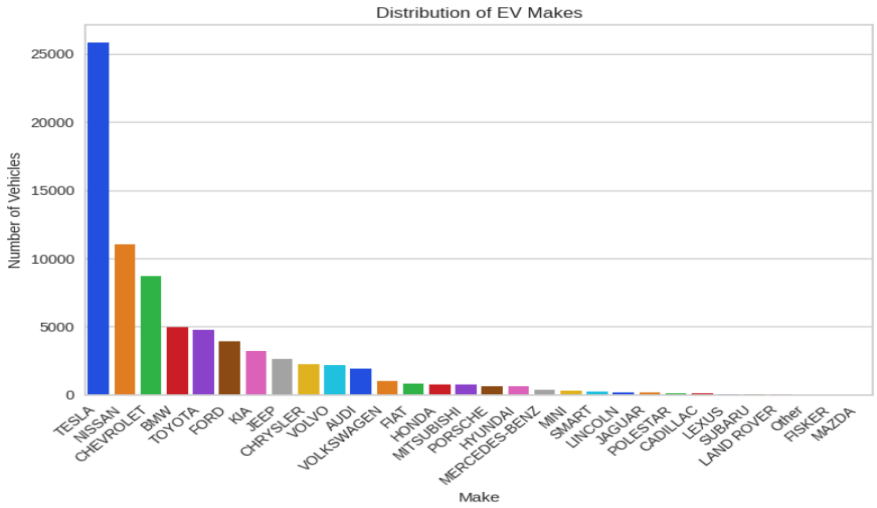
1. **電動車年分:**

可以清楚看出電動車成長的趨勢，從 2017 前的緩慢成長，到 2018~2020 穩定成長，以及現在 2021~2023 的爆發性成長



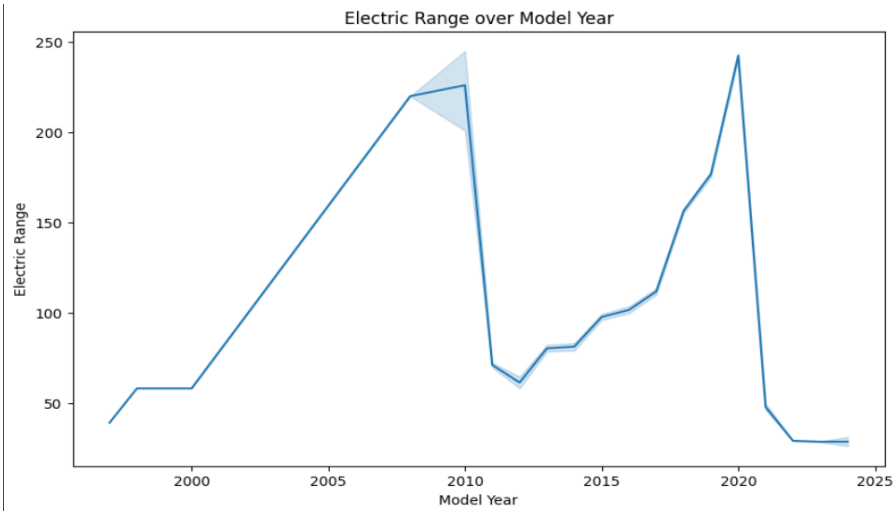
2. **製造商分布:**

明顯看出 TESLA 為近幾年最流行的電動車品牌，市占率遠大於其他，是顧客最愛的電動車廠商



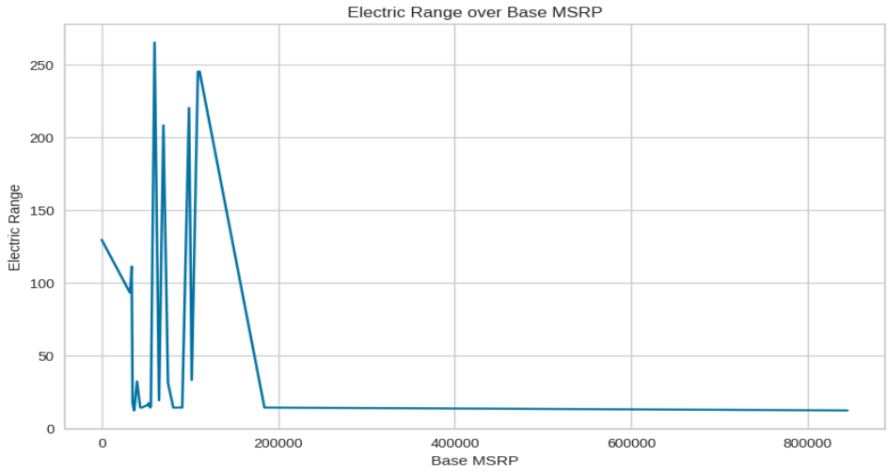
3. **年分 VS 電力里程:**

我們從年份圖得知 2011 年才開始慢慢出現電動車，但圖形 2010 年前就出現大量的電力里程，可以看出參考價值偏低



4. **建議零售價 VS 電力里程:**

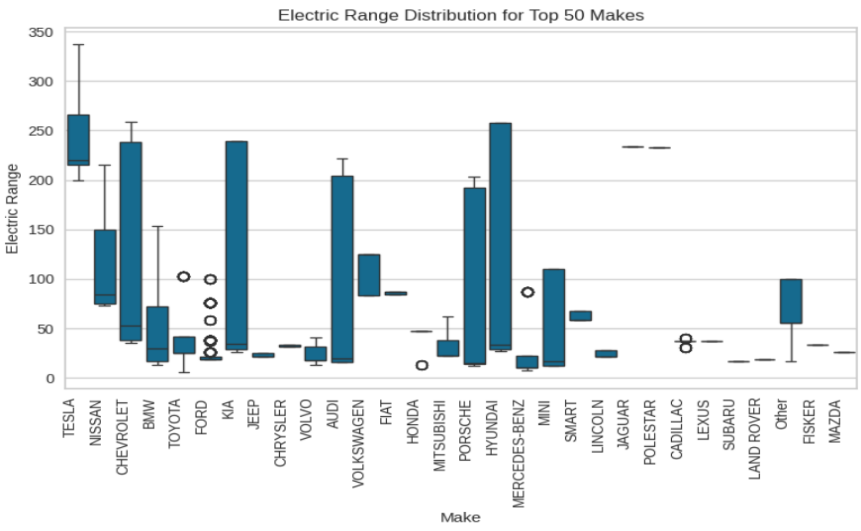
我們預想里程較長通常伴隨較高的售價，因為行駛里程的增加通常伴隨更高容量的電池技術，而電池成本占電動車整體成本的比重較大



但可以發現上圖無發發現到我們預想的觀點，因為建議售價有過多的 0 值，造成預想跟實際有落差。

5. **前 50 大製造商 VS 電力里程:**

我們抓出前 50 大製造商，判斷他們跟電力里程的關係，可以明顯看出 TESLA 電池效能大於其他製造商，也說明了 TESLA 在市場上的競爭力



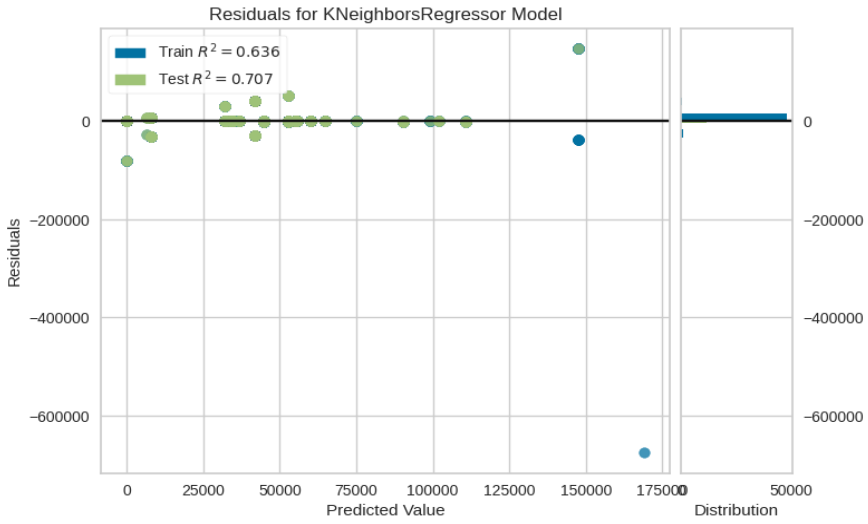
五、模型訓練

1.建立模型:

將整理過的 Make(製造商)、Model(模型)、electric range(電力里程)和 Base MSRP 整理成新的 Dataframe，再把資料分割成訓練集和測試(test_size=0.2)，之後讓 Base MSRP 為我們要預測的 Y 值，使用 pycaret 來比較使用哪種模型，最後根據 MAPE 的最小值決定使用 KNN。

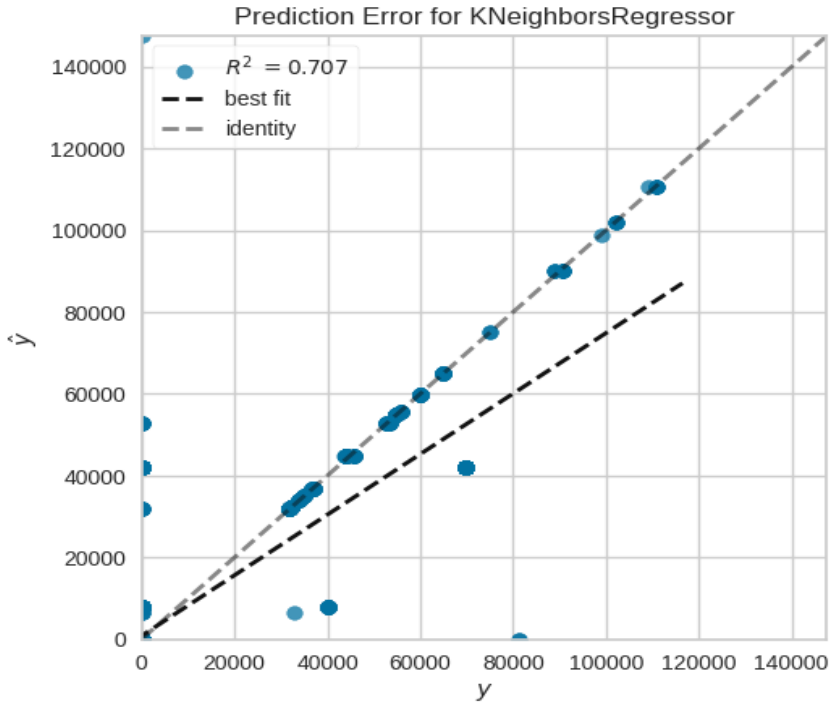
	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
knn	K Neighbors Regressor	1310.3102	72857472.0000	8488.5933	0.5552	1.6915	0.1427	0.4367
dt	Decision Tree Regressor	1343.1386	58984269.2877	7572.8048	0.6480	1.7531	0.2354	0.0900
et	Extra Trees Regressor	1348.7605	60661608.3091	7671.6131	0.6387	1.7575	0.2354	0.9133
xgboost	Extreme Gradient Boosting	1368.8053	72041702.6667	8381.2757	0.5613	1.8517	0.2355	0.2533
rf	Random Forest Regressor	1359.4464	65656429.4809	8007.0396	0.6041	1.7595	0.2359	0.9967
catboost	CatBoost Regressor	1359.3720	61425821.1533	7709.6886	0.6350	2.6263	0.2360	6.1033
lightgbm	Light Gradient Boosting Machine	1515.4294	64087424.3712	7879.5929	0.6189	3.8425	0.2580	0.9100
gbr	Gradient Boosting Regressor	2212.1759	74897830.0114	8580.4852	0.5477	5.6959	0.4136	1.1033
ada	AdaBoost Regressor	5948.2429	149777001.2104	12151.2615	0.0764	6.9045	0.6684	0.3833
lar	Least Angle Regression	4704.7153	159059194.6667	12572.3587	0.0240	7.3650	0.9364	0.1200
lasso	Lasso Regression	4704.6003	159059216.0000	12572.3555	0.0240	7.3657	0.9364	0.1167

2.檢查模型的殘差是否有 trend:



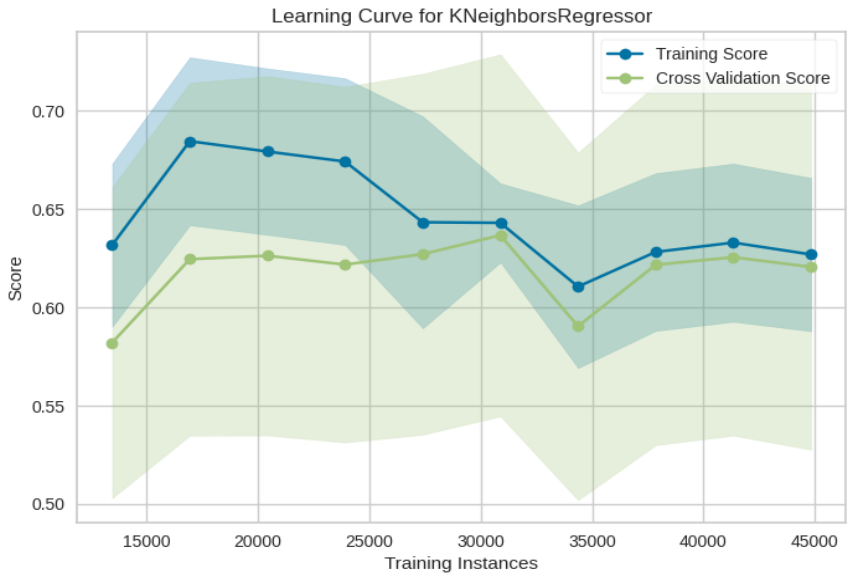
無明顯的 trend

3.查看模型解釋的比例(best fit 是否接近 identify):



X 軸為真實的 Base MSRP 資料，Y 軸為模型預測的 Base MSRP 資料
由於 base MSRP 為 0 的值還是有存在(no data 或者製造商隱藏數值)，導致一部分的值沒有辦法估計，但在 base MSRP 不為 0 的時候基本上都落在 identify 上。

4.查看正確率:



不管是 train score 還是交叉驗證 score 的 learning curve 最後都落在 63%左右

5.小結:

由上述圖表來看，正確率不算特別高，應該是原始 base MSRP 為 0 的值的影響，但除開 y= 0 的值，其實有相當一部份的值有落在 y=y_hat 的線上。

六、結論:

綜合上述分析與預測，我們得出的結論:

1. **電力里程與售價的正相關性:** 售價較高的車型通常配備更大容量的電池和更先進的技術，這也解釋了為何行駛里程越長，售價越高。
2. **市場中的主要品牌**：Tesla 的領先地位顯而易見，其車型數量明顯多於其他品牌。Nissan 和 Chevrolet 等品牌則為中低價位市場提供了競爭性產品，滿足不同消費者需求。
3. **模型預測具有一定準確性**：線性回歸模型在預測行駛里程方面取得了合理的準確性。未來可以考慮加入更多變量（如車輛重量、電池容量等）來提升模型的預測效果。

七、未來改進的方向:

1. 即使模型是有準確性的，但應該使用 correlation map 來增加更多變數，而不是只挑我們想要的。

2. 使用 electric range 對類別變數做 encoding 時應該將 electric range 設置為模型的 Y，這點在做的時候沒有想到。會再做一次以 electric range 為 Y 的模型。