CloudMile

CloudMile 雲端費用預測

詹昊庭 魏詩庭 藍知生 蔡鉒驊 徐浚凱 林奕霆

CloudMile — 雲端花費預測報告大綱

產品簡介		本次優化產品為雲端服務管理平台 MileLync
專案目標		分析雲端服務使用，包含使用量、費用預測與異常行為偵測，提升使用者體驗，作為 MileLync 產品的加分項
模型選擇與評估	資料集介紹	三間公司在2022年1月至11月底的雲端使用花費
	模型建立	• Conventional: 以 ACF 觀察週期，再以 auto_arima 選出 AIC 最低的參數組合 • Tree-based: 以 XGBoost 及 Random forest 抽取特徵生成樹，進行預測 • Transformer-based: FEDformer • RNN-based: LSTM • BigQuery: SARIMA 加入 features
模型成效	SARIMA	時間序列分析與資料前處理，對所有 Project 進行初步使用量預測
	Tree-Based	觀察 SARIMA，加入額外變數 (Features) MA
限制		僅以預測目標前 90 天內作為輸入資料，以降低實務上遇公司成本歷史資料不足造成的預測限制

Agenda

1. **MileLync簡介**
2. 資料集與專案設定
3. 模型比較
4. 模型優化：SARIMA、Tree-Based
5. 未來展望

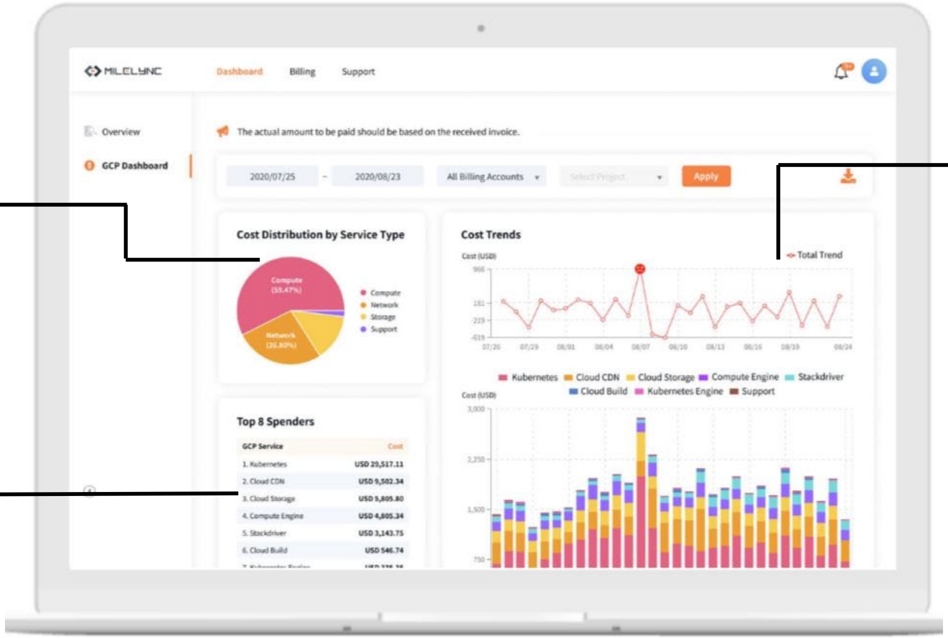
MileLync 為一站式雲端管理平台，可以簡化並整合 GCP Console 複雜的管理平台

Cost Distribution by Service Type

透過圖表和報告的形式呈現成本分佈的詳細信息，用戶可直觀看到各服務類型所佔的成本比例。有助於用戶識別成本較高或不必要的服務，以便進行優化和調整

Top Spenders

以圖表和報告的形式呈現最高成本的詳細資訊，用戶可清楚看到消費最多的使用者和相應金額，有助識別出資源消耗較大的項目



Cost Trends

可按不同的時間範圍顯示成本趨勢，並提供相應的比較和分析。我們本次專案則希望提升成本預測的表現，優化客戶的使用體驗

Agenda

1. MileLync簡介
2. 資料集與專案設定
3. 模型比較
4. 模型優化：SARIMA、Tree-Based
5. 未來展望

共 3 個擁有相同欄位特徵的資料集，並因應實務限制而設定專案資料使用標準

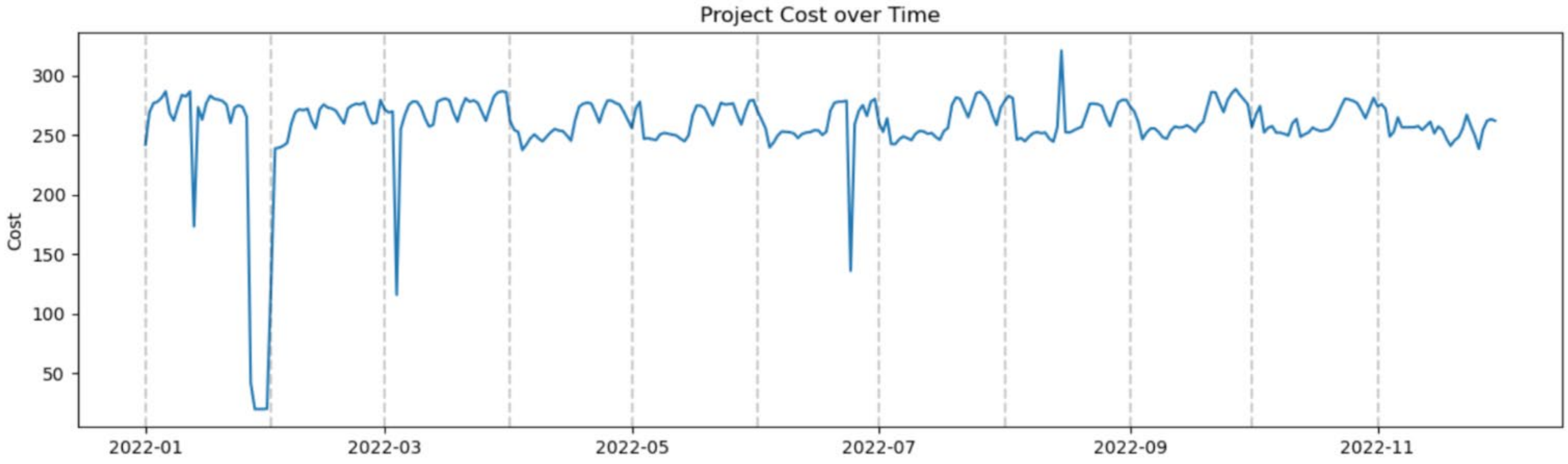
資料集描述

檔案	project_a.csv, project_b.csv, project_c.csv
特徵	共兩個欄位，分別為 date 以及 cost
時間範圍	皆為自 2022/01/01 至 2022/11/30

專案與模型設定

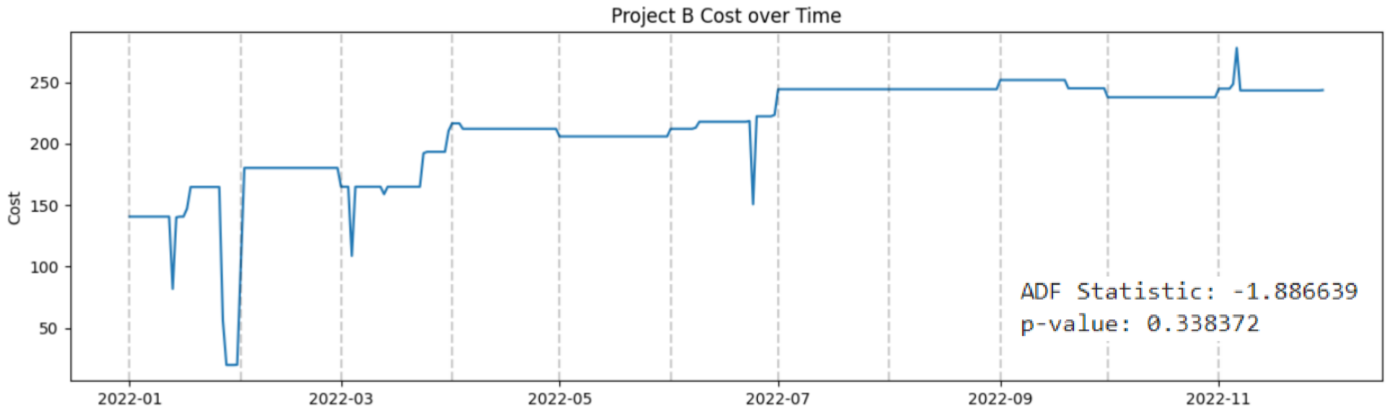
資料切割	訓練資料：2022/01/01 ~ 2022/09/15；驗證資料：2022/09/15 ~ 2022/10/15；測試資料：2022/10/15 ~ 2022/11/30
方法選擇	各模型統一使用 Rolling-base 預測以捕捉序列中的時間相依性

Project A 資料輪廓

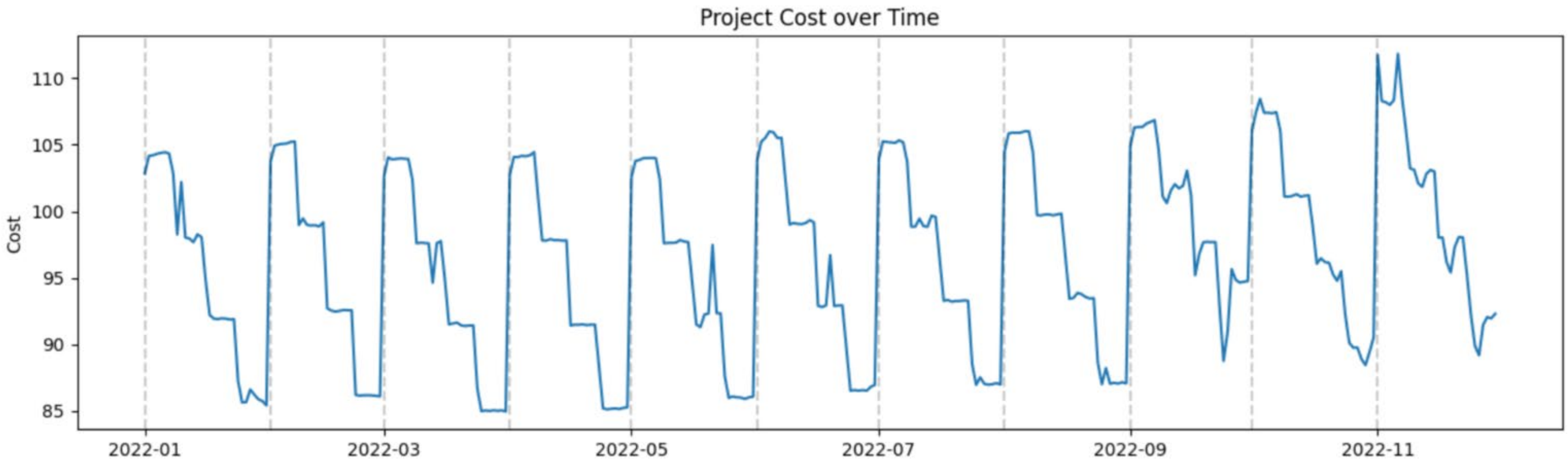



```
ADF Statistic: -6.989692
p-value: 0.000000
```

Project B 資料輪廓



Project C 資料輪廓



ADF Statistic: -5.588818
p-value: 0.000001

Agenda

以 RMSE 衡量 11/01 - 11/30 預測結果

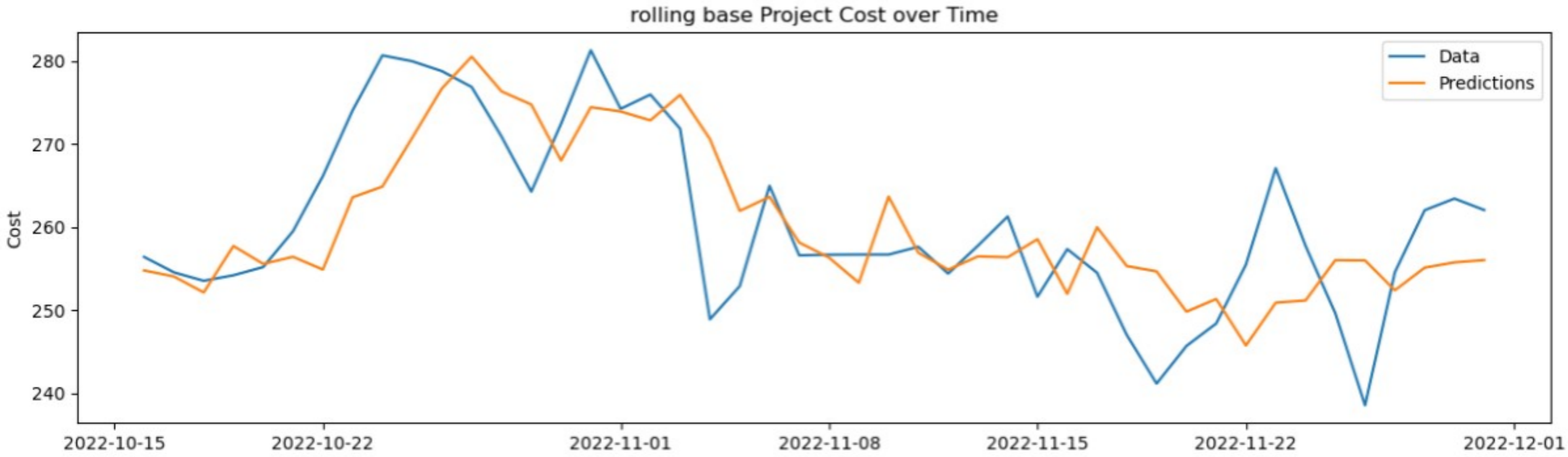
RMSE	Baseline	SARIMA	XGBoost	Random Forest	FEDformer	LSTM	Big Query
Project A	10.36	8.69	17.9	8.88	12.37	7.57	10.79
Project B	3.75	5.71	9.68	6.39	7.70	7.22	6.74
Project C	2.48	1.57	4.16	1.835	3.62	4.30	2.46

以 RMSE 衡量 10/16 - 11/30 預測結果

RMSE	SARIMA	XGBoost	Random Forest	FEDformer	LSTM	Big Query
Project A	6.49	7.61	8.13	11.20	7.90	11.037
Project B	4.72	7.28	5.43	7.06	6.25	21.241
Project C	2.29	2.13	2.91	3.43	3.50	2.725

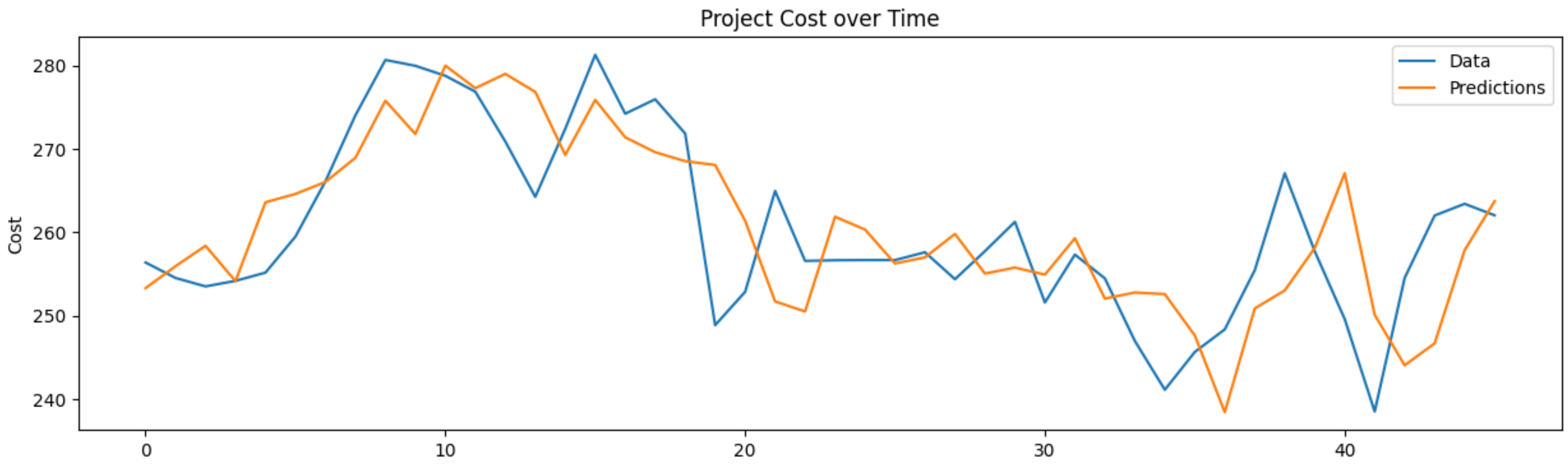
Project A (10/16 - 11/30)

Project A: Sarima(Rolling-based: 90)



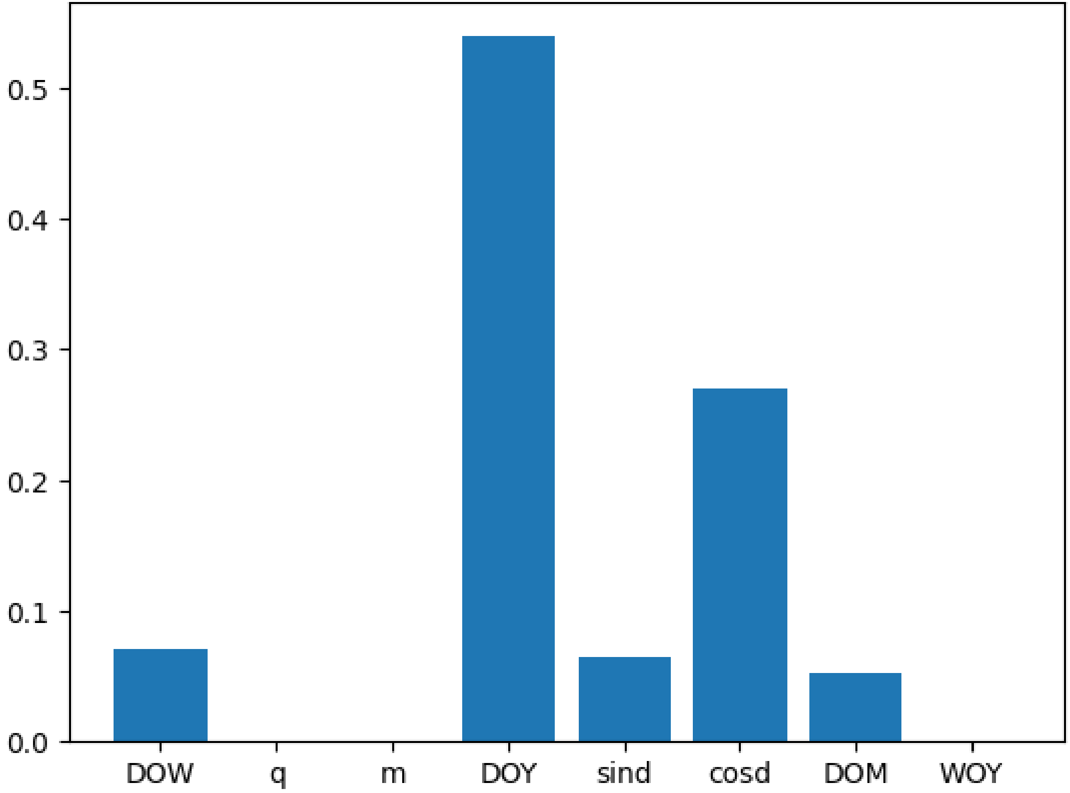
para=[1,0,2][8,1,0,7]
RMSE= 7.87

Project A: XGBoost (Rolling-based: 35)



RMSE= 9.06

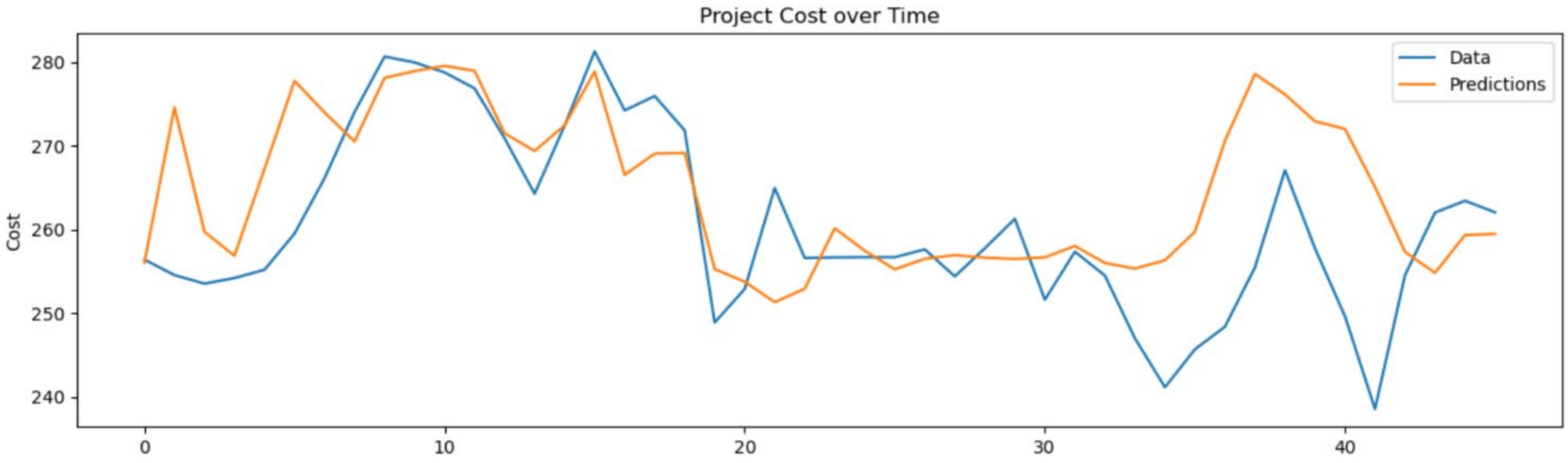
Project A: XGBoost Features (Rolling-based: 35)



參數介紹

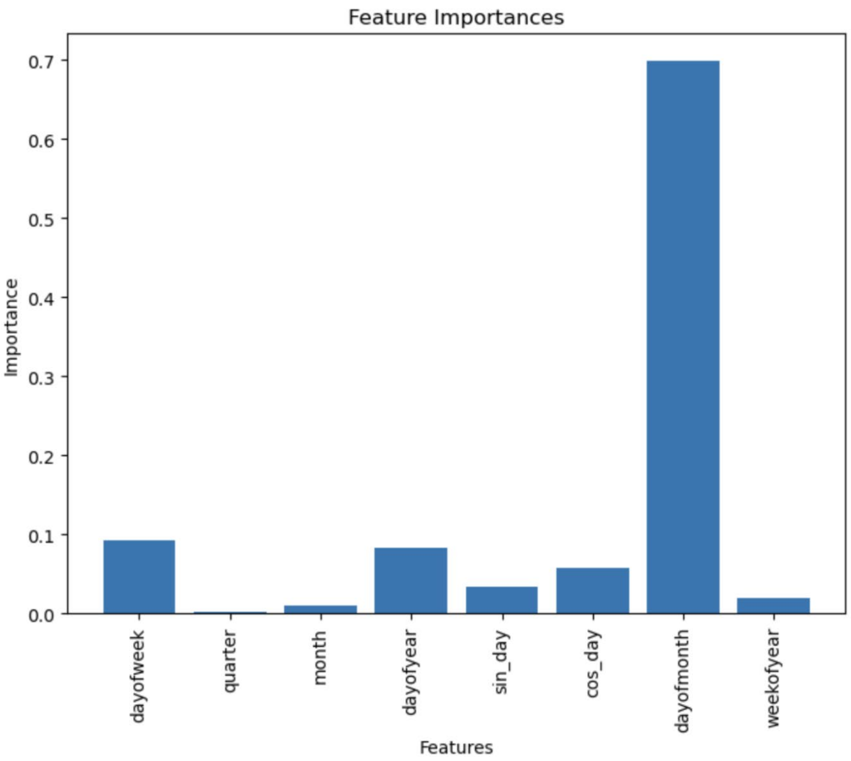
- Dayofweek
- Quarter
- Month
- Dayofyear
- sin_day
- cos_day
- Dayofmonth
- Weekofyear

Project A: RandomForest (Rolling-based: 75)



RMSE = 10.02

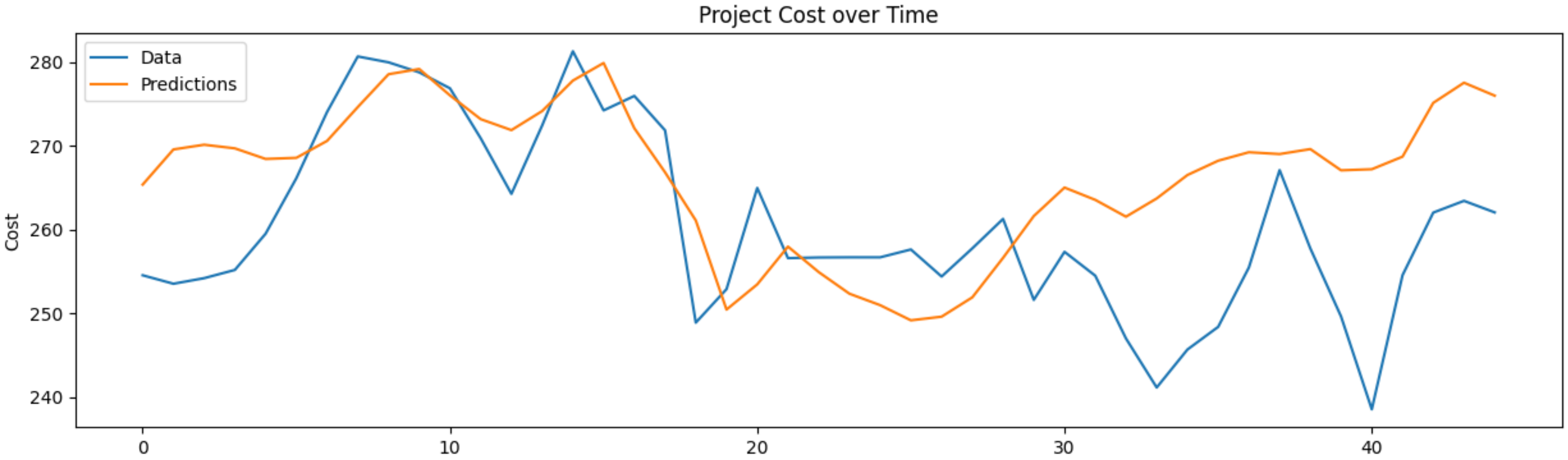
Project A: RandomForest Features (Rolling-based: 75)



Project A: LSTM (Rolling-based: 12)

RMSE = 7.90

Project A: FEDformer (Rolling-based: 72)

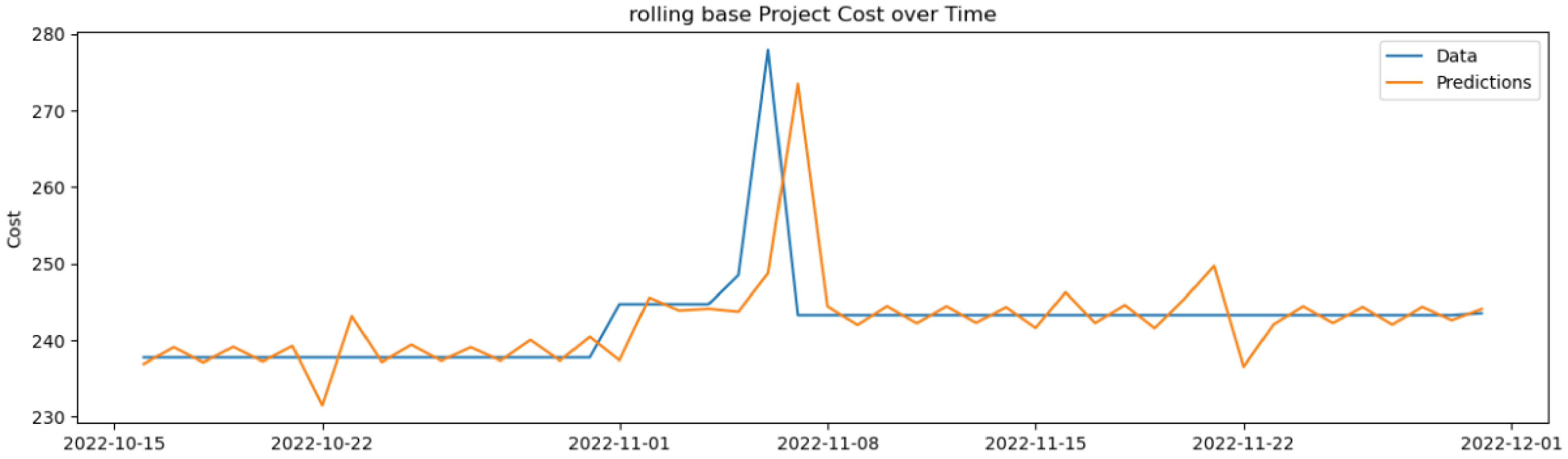


RMSE = 11.20

CloudMile

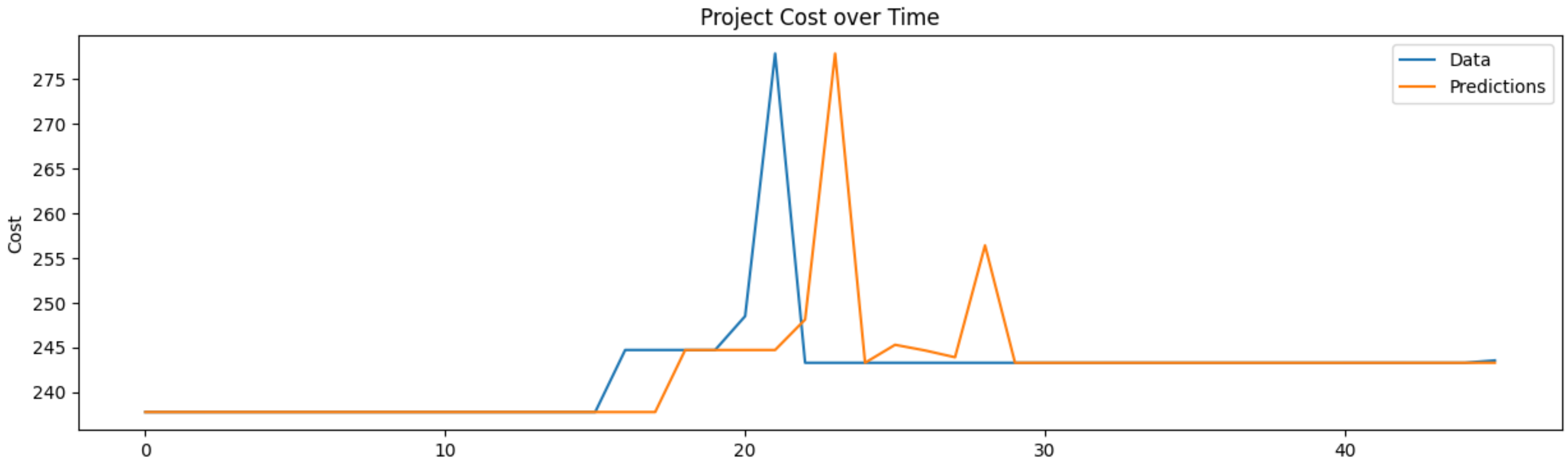
Project B (10/16 - 11/30)

Project B: Sarima (Rolling-Based: 75)



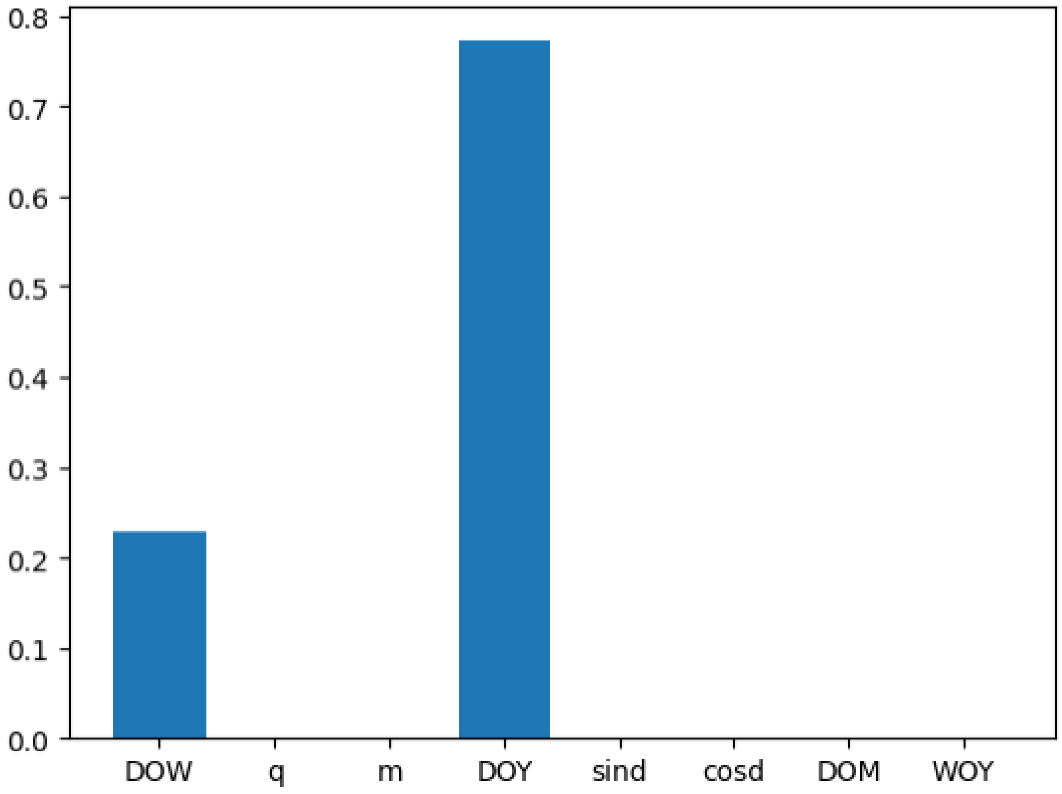
para = [2,1,0],[7,1,0,15]
RMSE = 6.68

Project B: XGBoost (Rolling-based: 25)

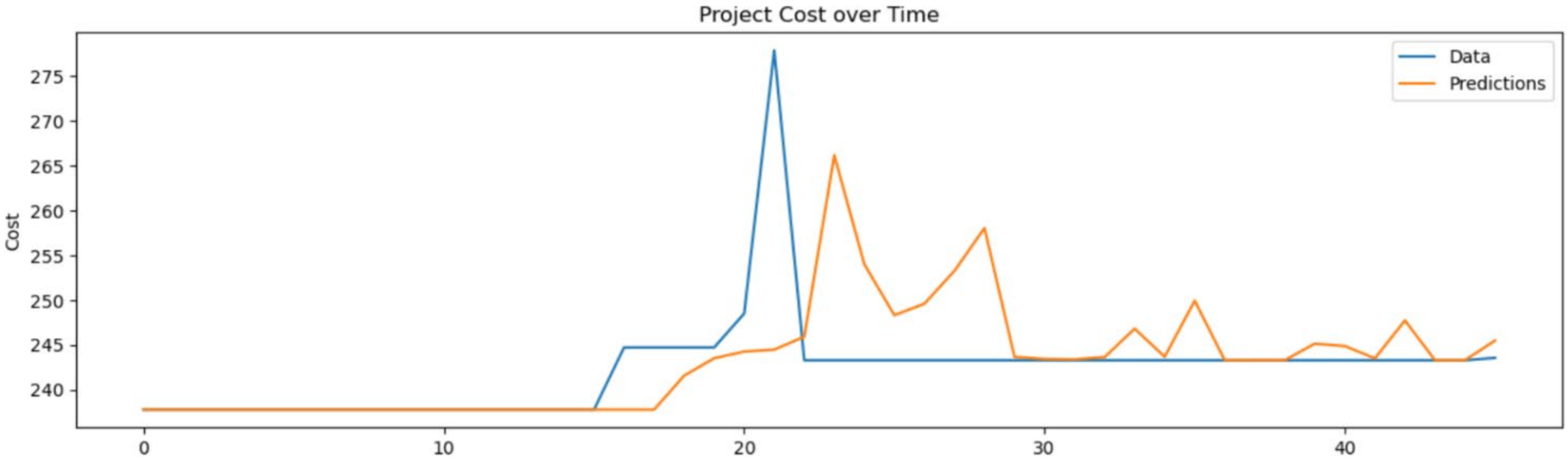


RMSE = 7.53

Project B: XGBoost Features (Rolling-based: 25)

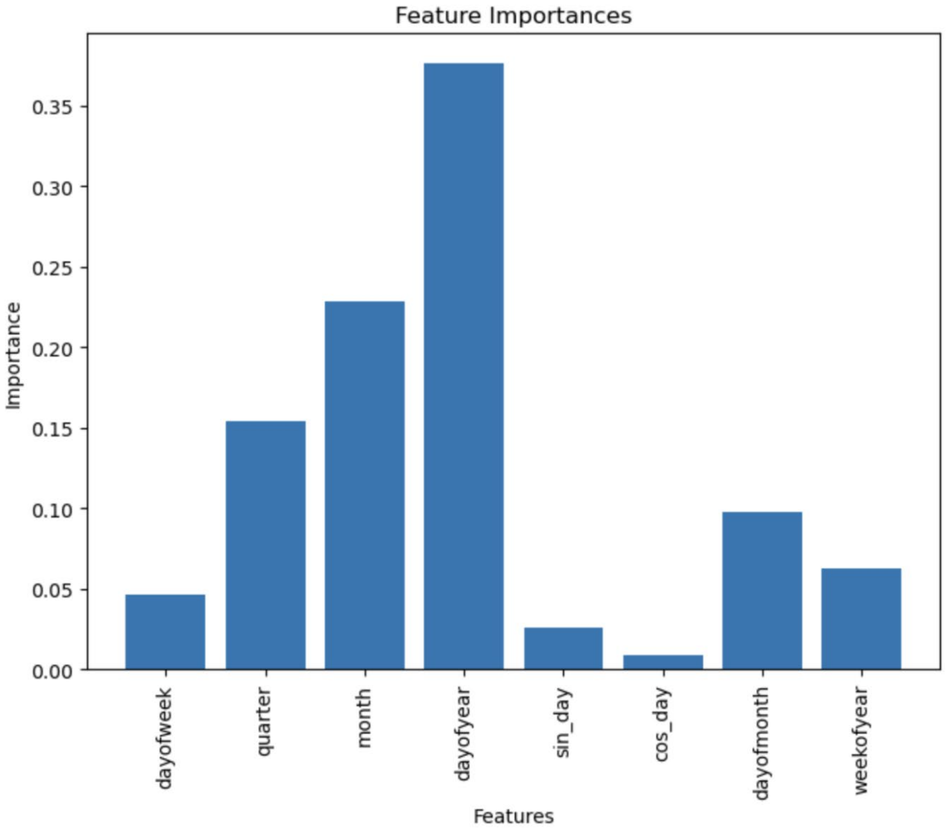


Project B: RandomForest (Rolling-based: 85)

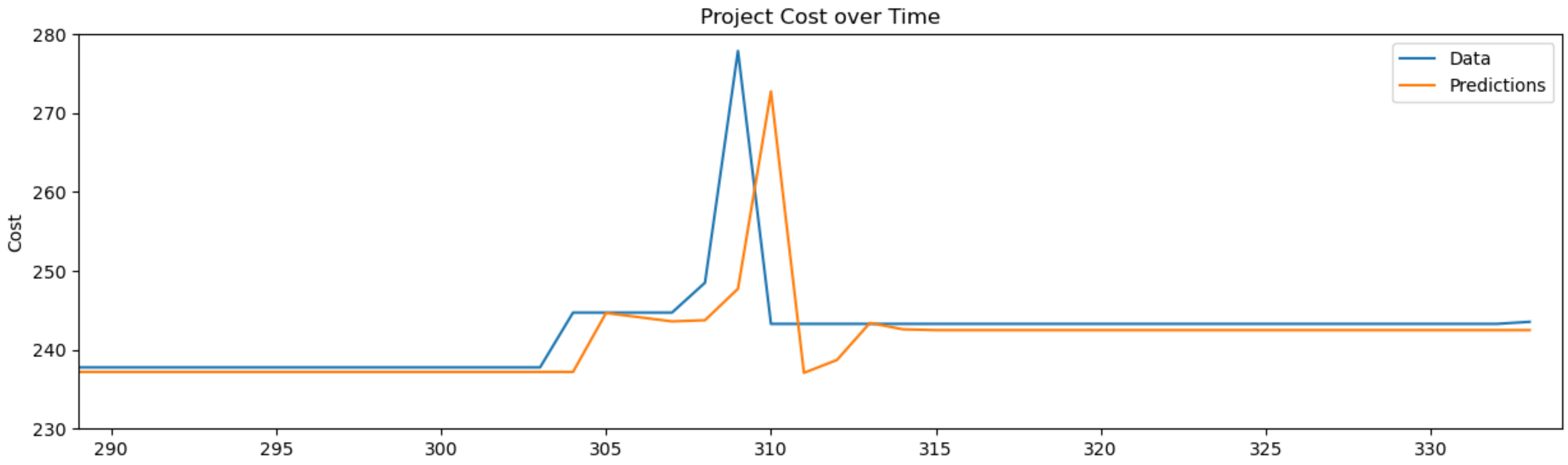


RMSE = 7.16

Project B: RandomForest Features (Rolling-based: 85)

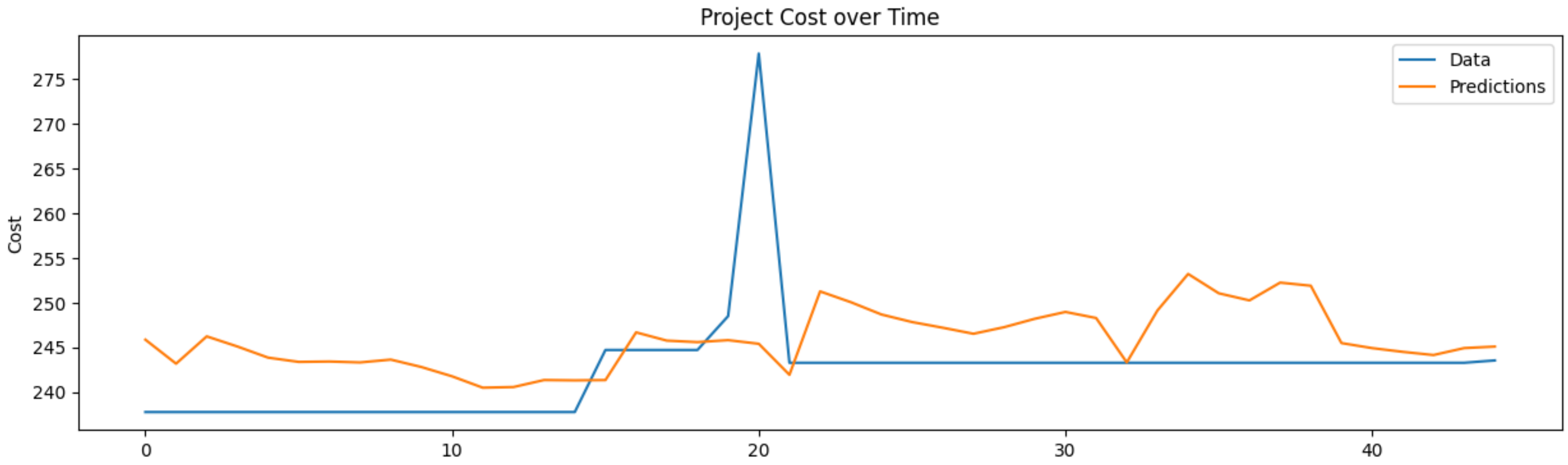


Project B: LSTM (Rolling-based: 12)



RMSE = 6.25

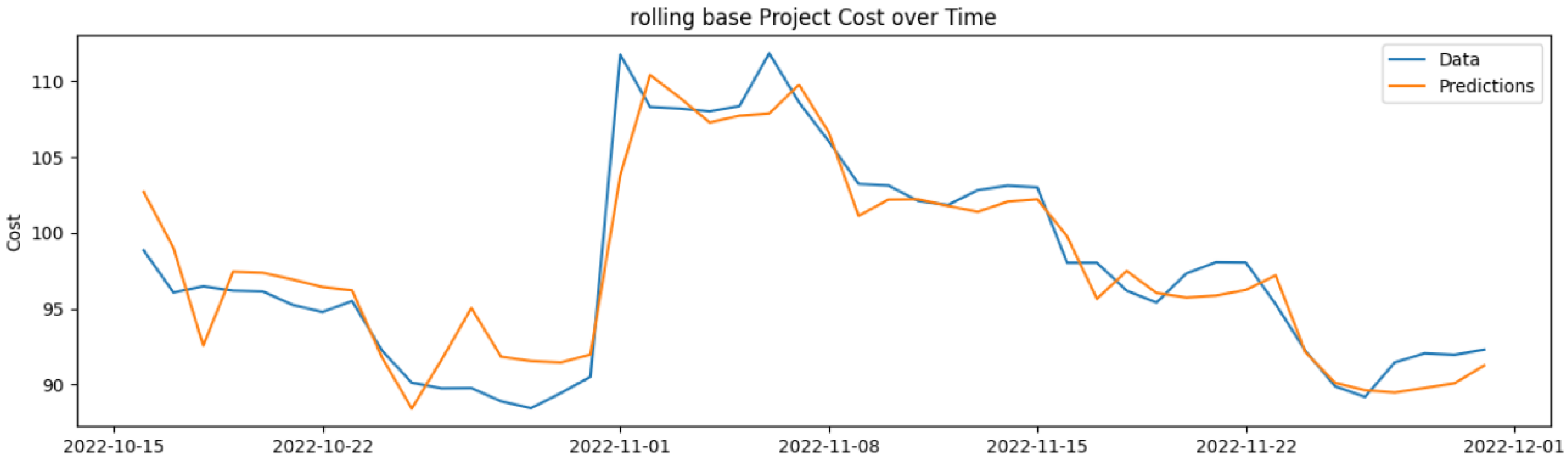
Project B: FEDformer (Rolling-based: 36)



RMSE = 7.06

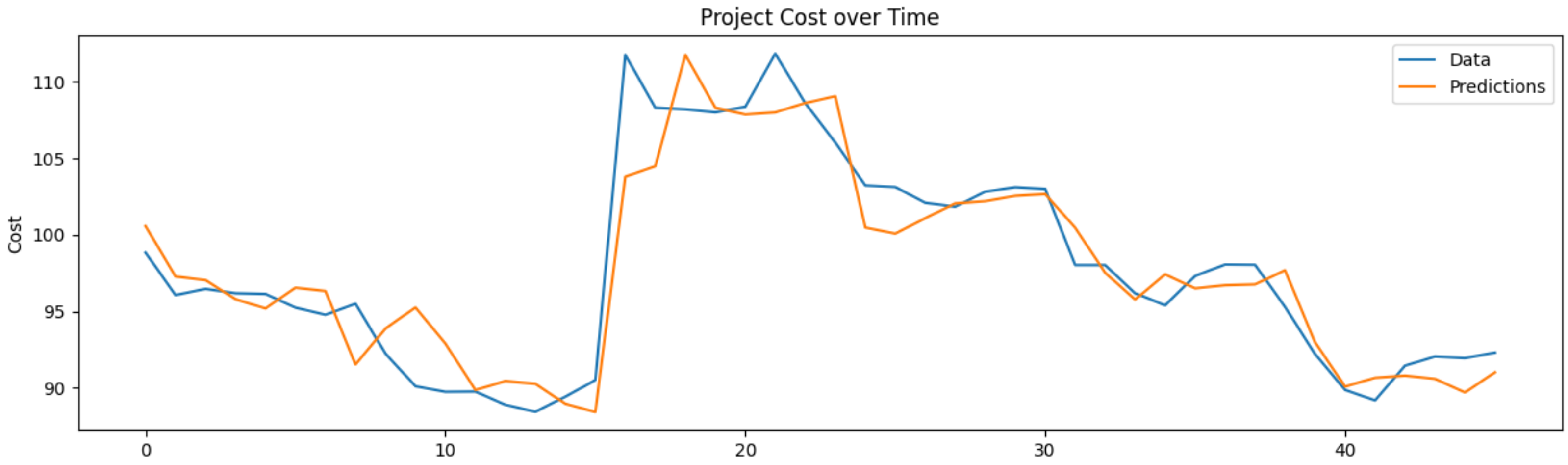
Project C (10/16 - 11/30)

Project C: Sarima (Rolling Base: 90)



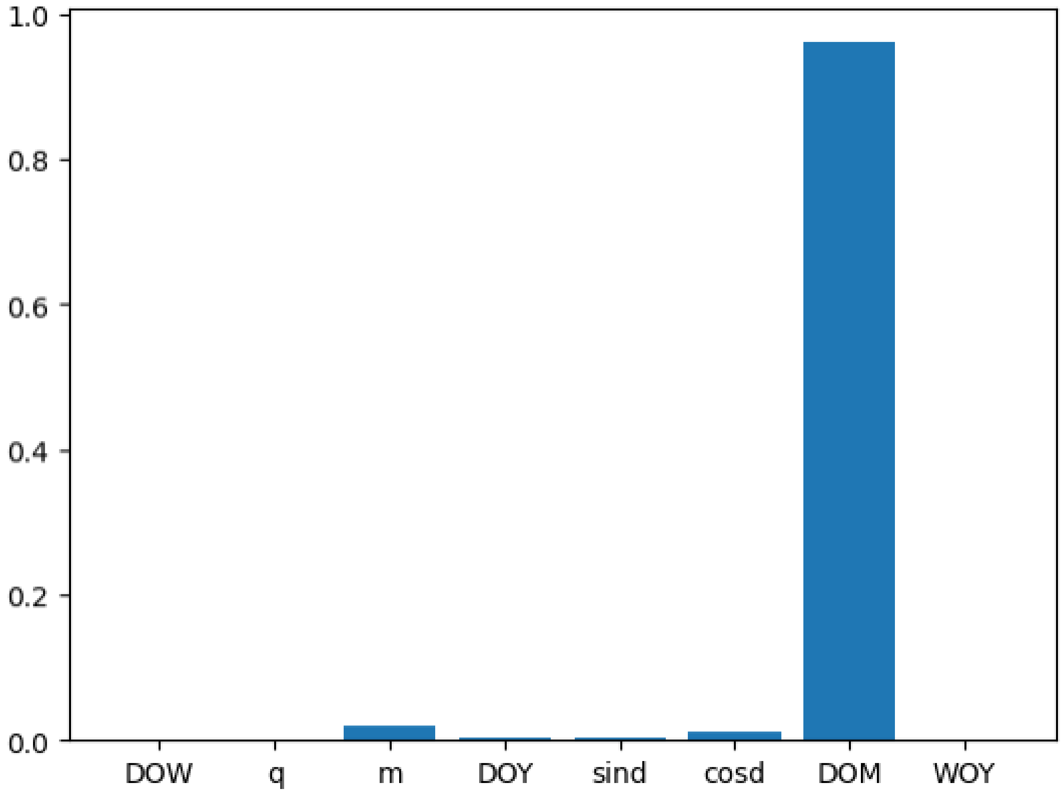
para = [1,0,0],[0,1,0,31]
RMSE = 2.29

Project C: XGBoost (Rolling-based: 45)

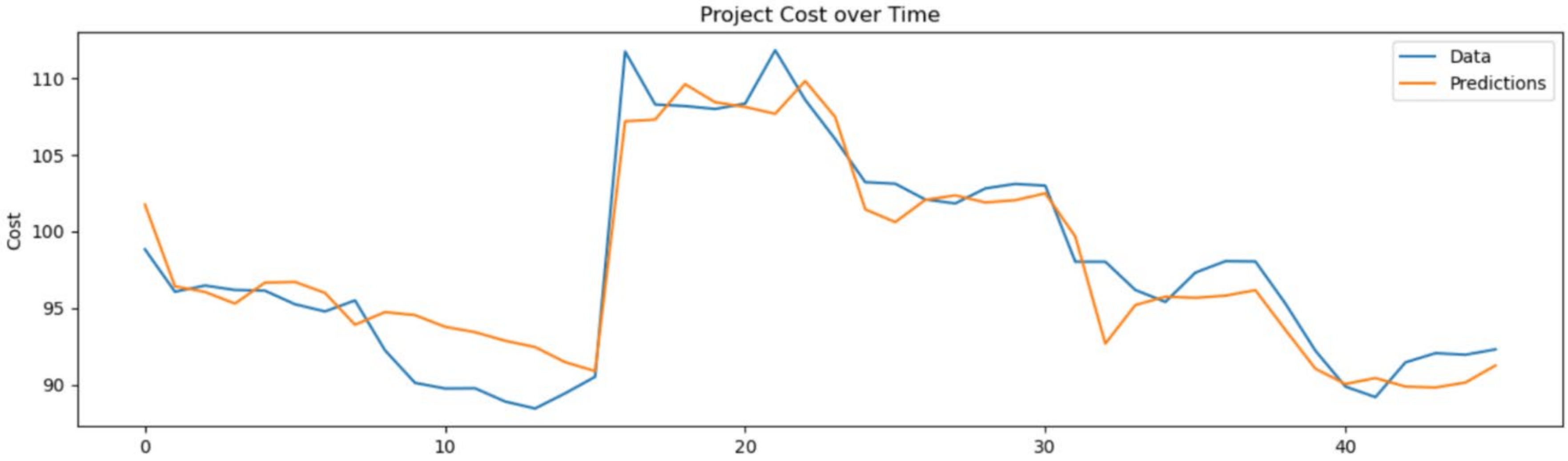


RMSE = 2.287

Project C: XGBoost Features (Rolling-based: 45)

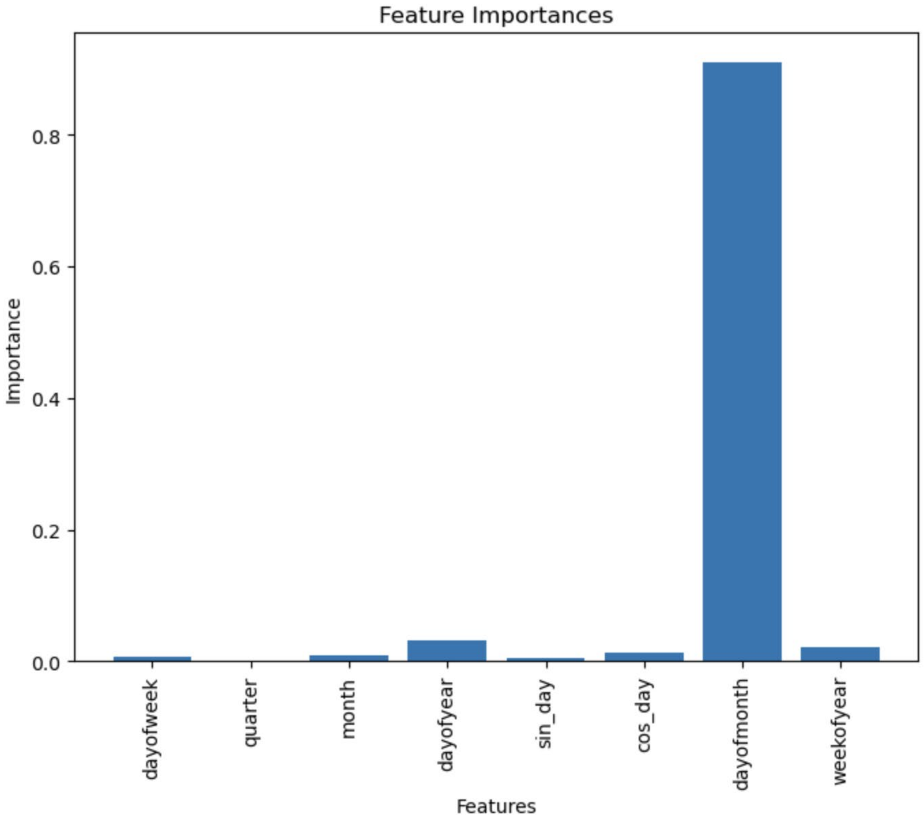


Project C: RandomForest (Rolling-based: 60)

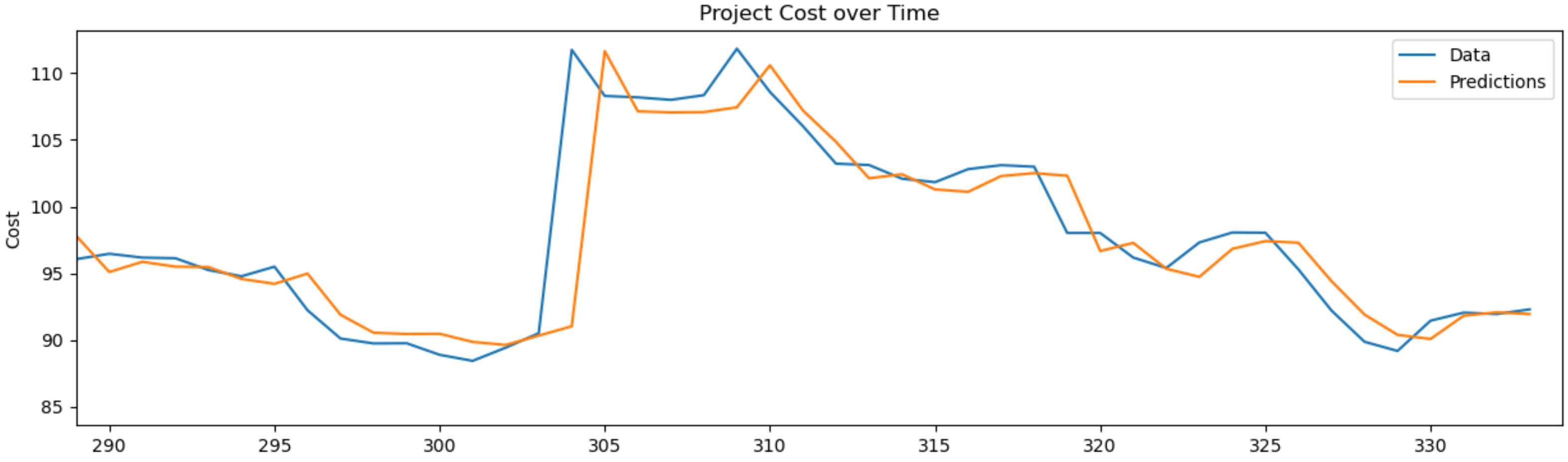


RMSE = 2.91

Project C: RandomForest Features (Rolling-based: 60)

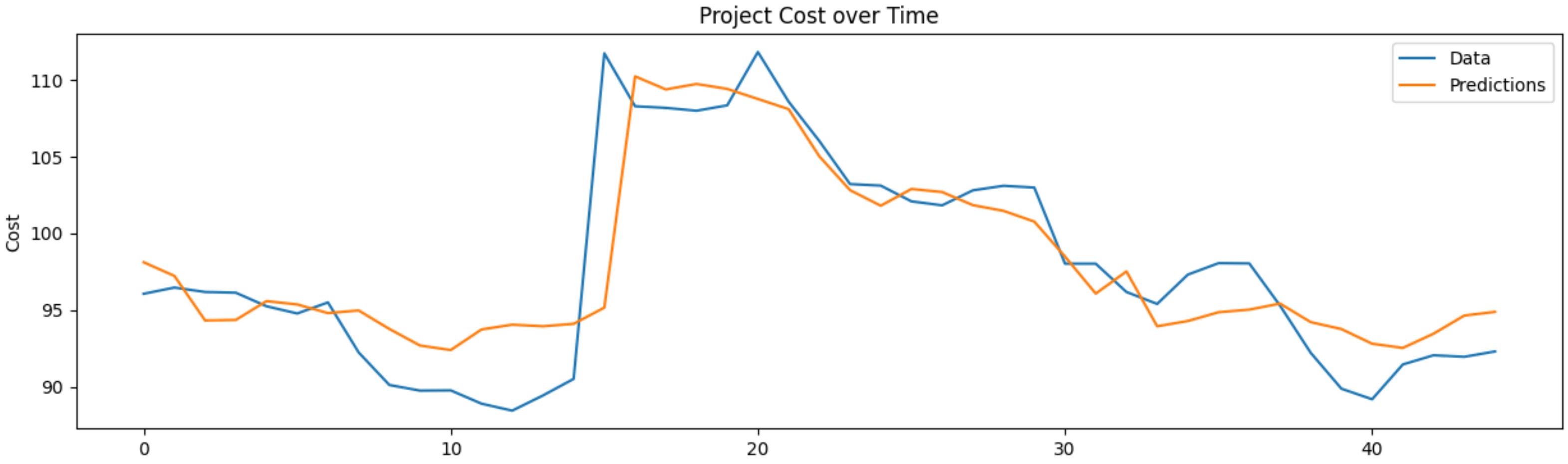


Project C: LSTM (Rolling-based: 60)



RMSE = 3.50

Project C: FEDformer (Rolling-based: 36)



RMSE = 3.43

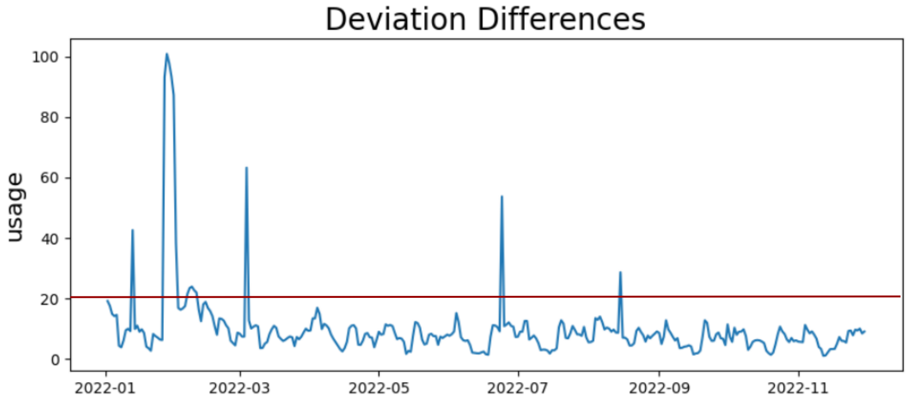
Agenda

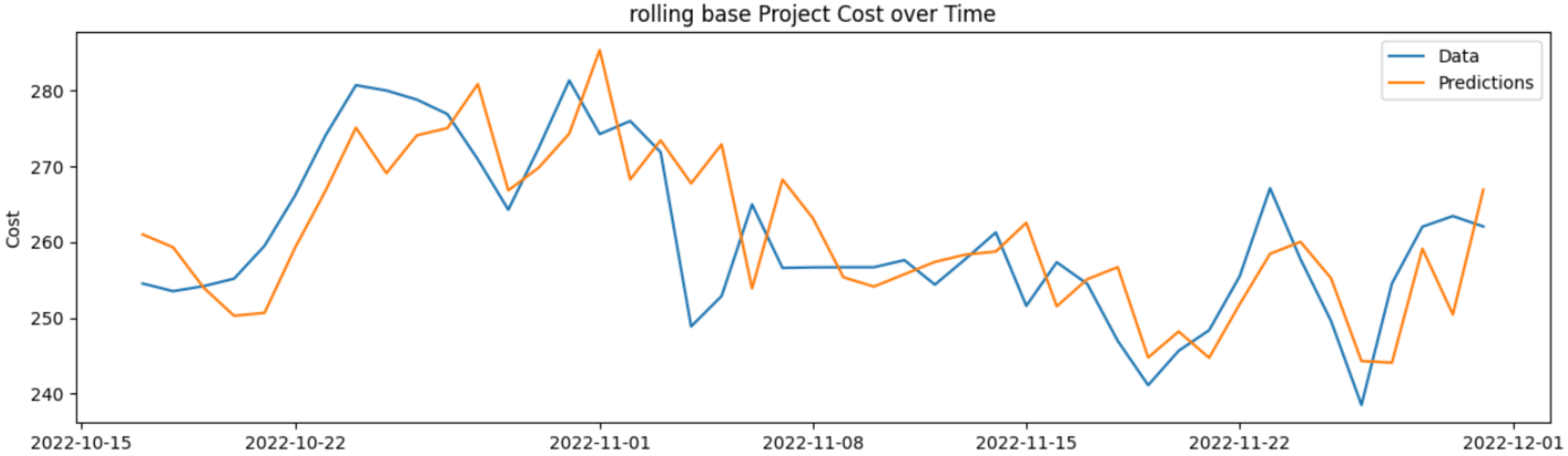
1. MileLync簡介
2. 資料集與專案設定
3. 模型比較
4. 模型優化：SARIMA、Tree-Based
5. 未來展望

Sarima模型優化方法

異常值偵測	• **法一:** 若第t日到第t-4日之平均標準差超過定值則視為異常值。 • **法二:** 若第t日到第t-n日之平均標準差與第t-1日到第t-n-1日平均標準差相減超過定值則視為異常值。
異常值處理	• **平滑化方法:** 為了滿足Rolling需求，將包含當天以及過去四天的**五日平均值**取代異常值。
定期自動選出最佳參數	• 在假設只有三個月的時間序列資料下，於預測資料時，**每15日**以ACF輸出之數值尋找局部最大值，選出最佳的Seasonality參數，並透過auto_ARIMA自動選出(p,d,q)(P,D,Q)之參數。

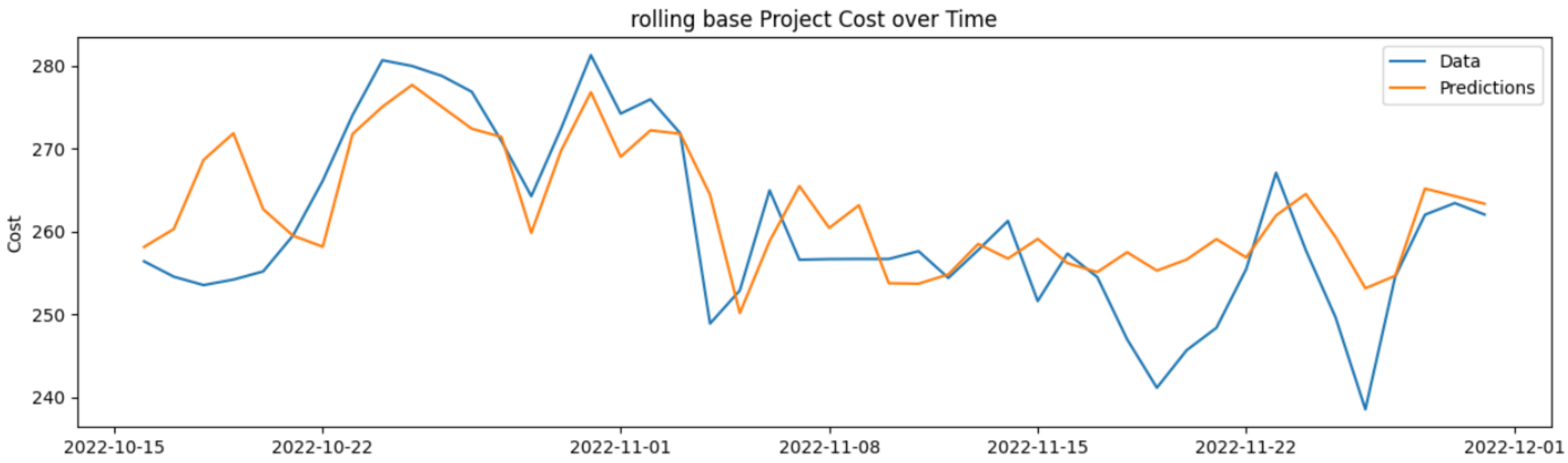
Project A: Sarima
對異常值平滑化 Rolling-based=80





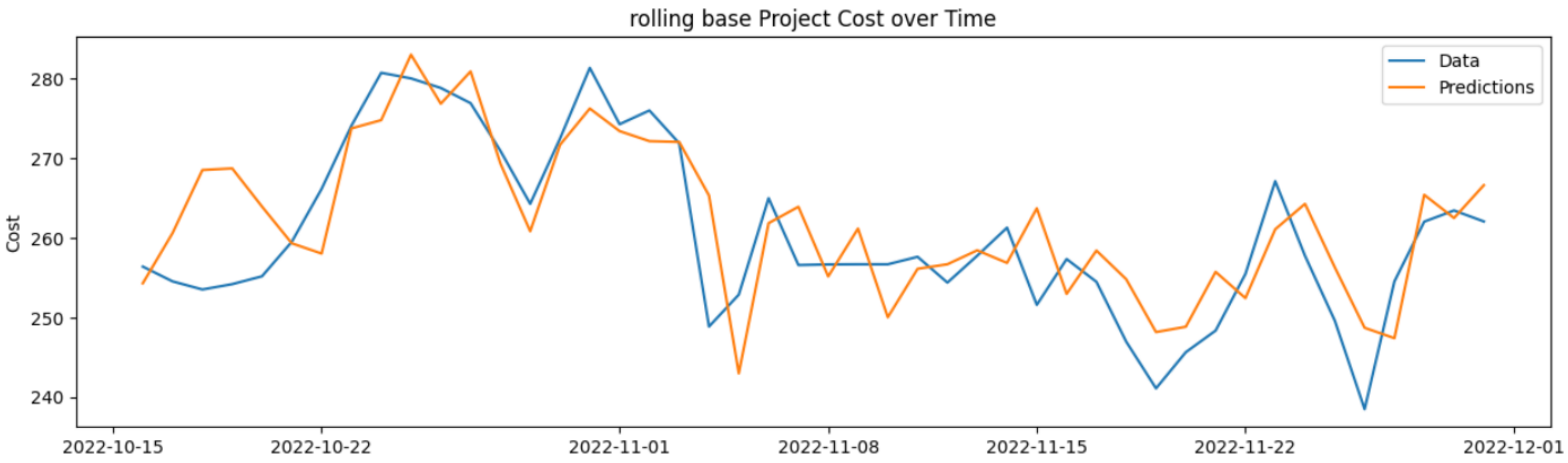
para = [1,0,0],[5,1,0,7]
RMSE = 7.65

Project A: Sarima
定期自動選出最佳參數



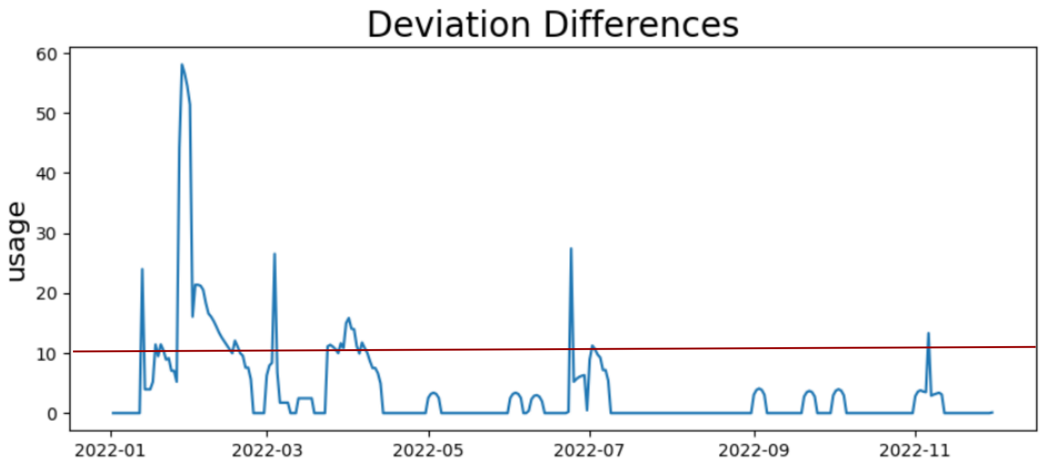
m=7,7,7,8
RMSE = 7.12
windows = 80

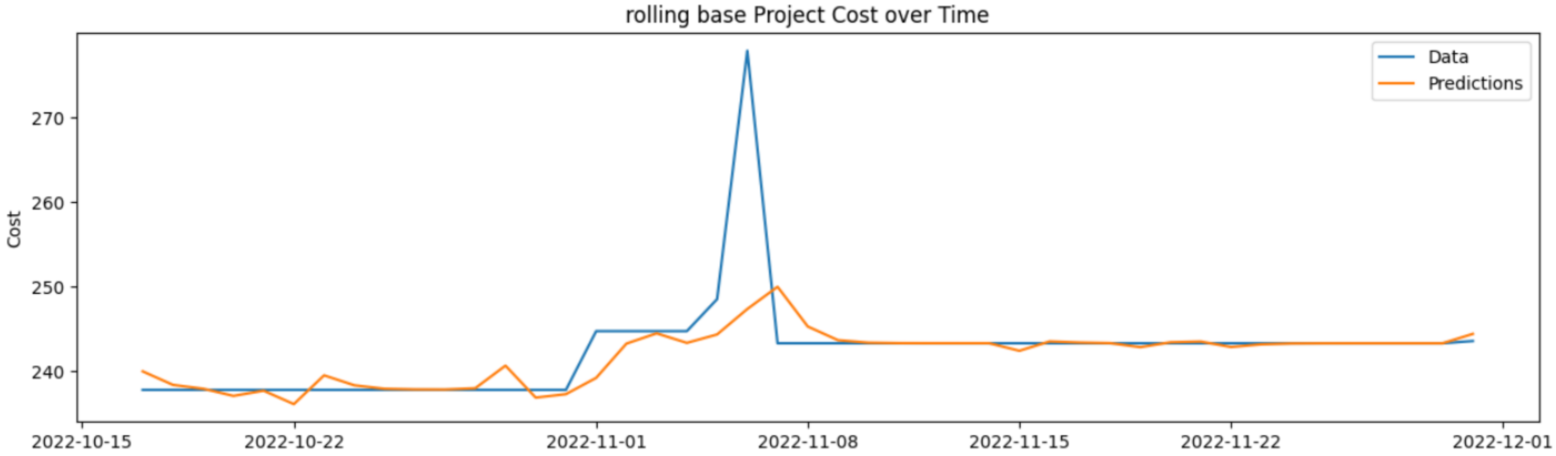
Project A: Sarima
對異常值做平滑化+定期自動選出最佳參數



m=7,7,7,8
RMSE = 6.49
windows = 80

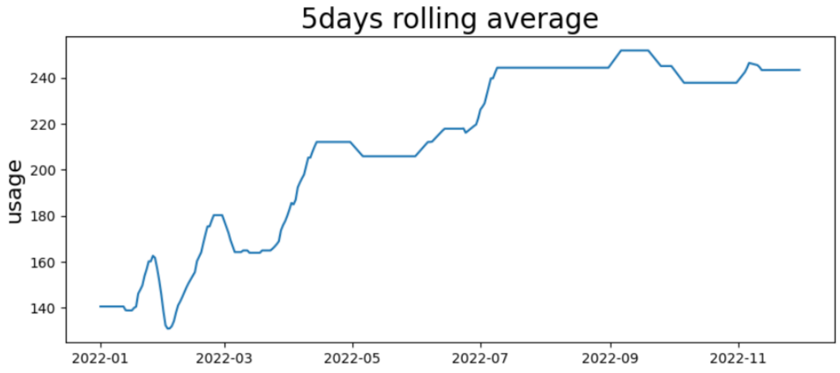
Project B: Sarima
對異常值平滑化

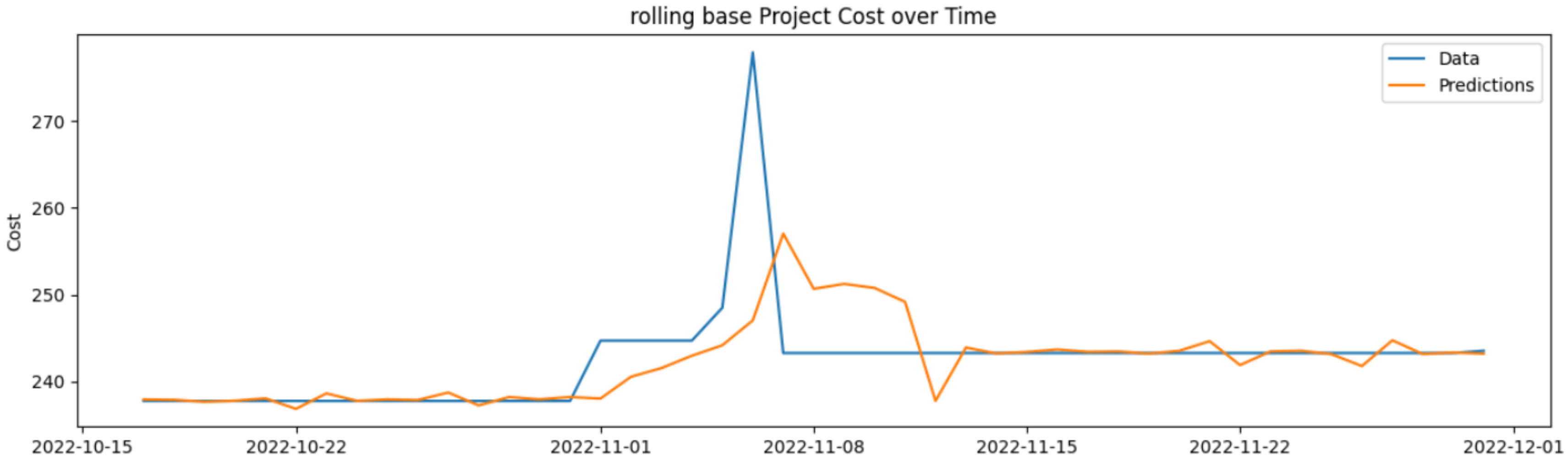




para = [2,0,1],[7,1,0,15]
RMSE = 4.85

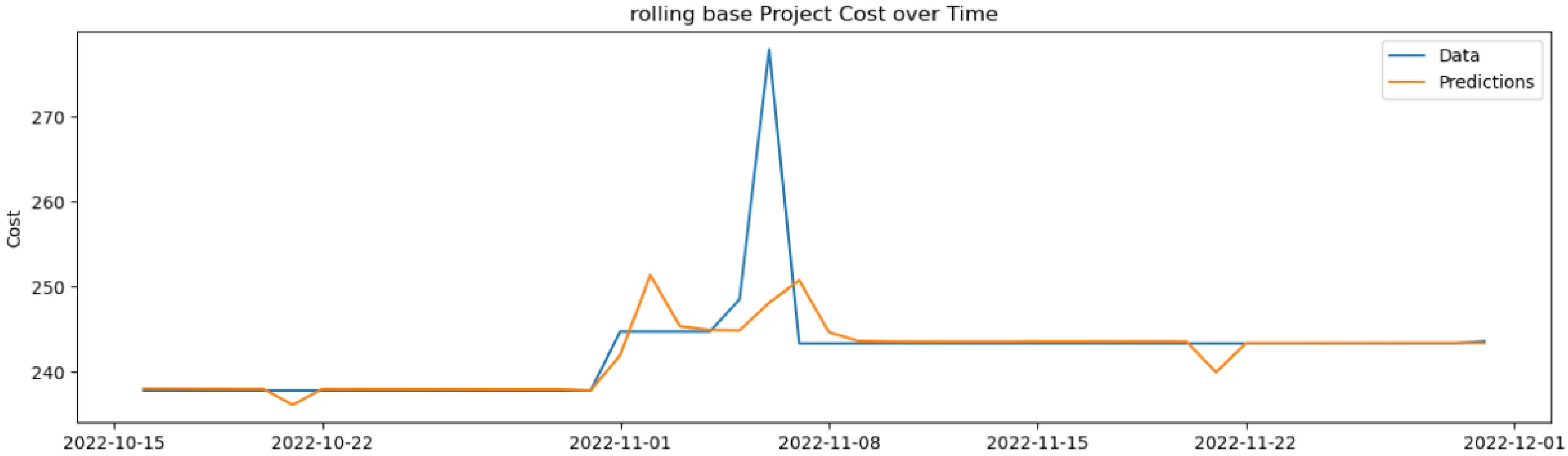
Project B: Sarima
對所有數值以五日平均平滑





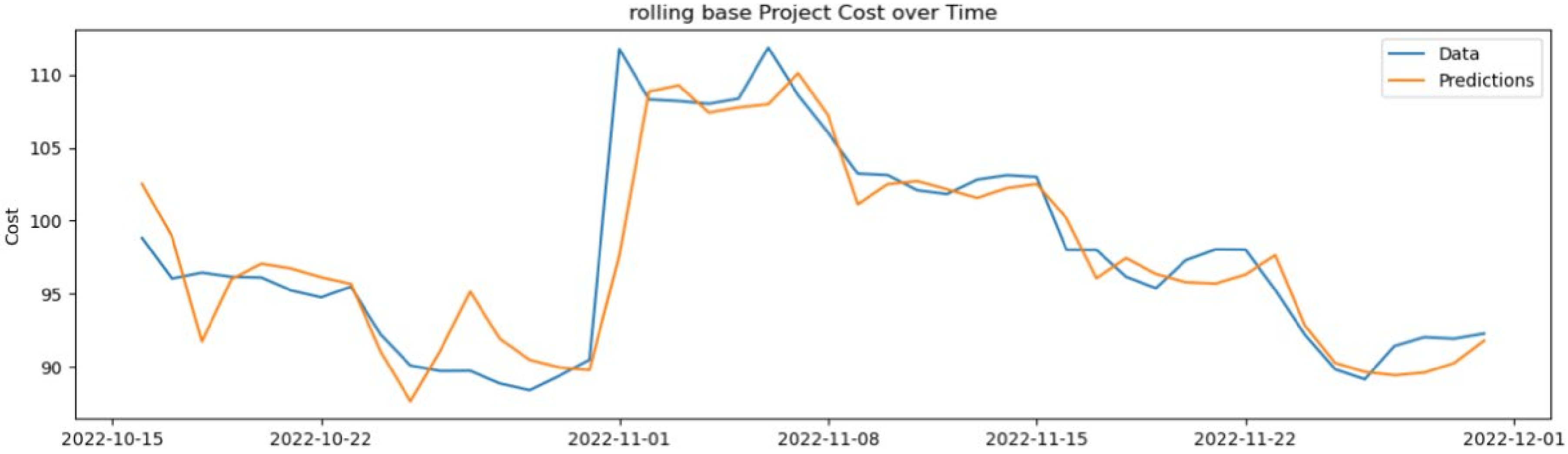
para = [2,0,1],[7,1,0,15]
RMSE = 5.75

Project B: Sarima
對異常值平滑化+定期自動選出最佳參數



m=31,31,31,31
RMSE = 4.72

Project C: Sarima
定期自動選出最佳參數

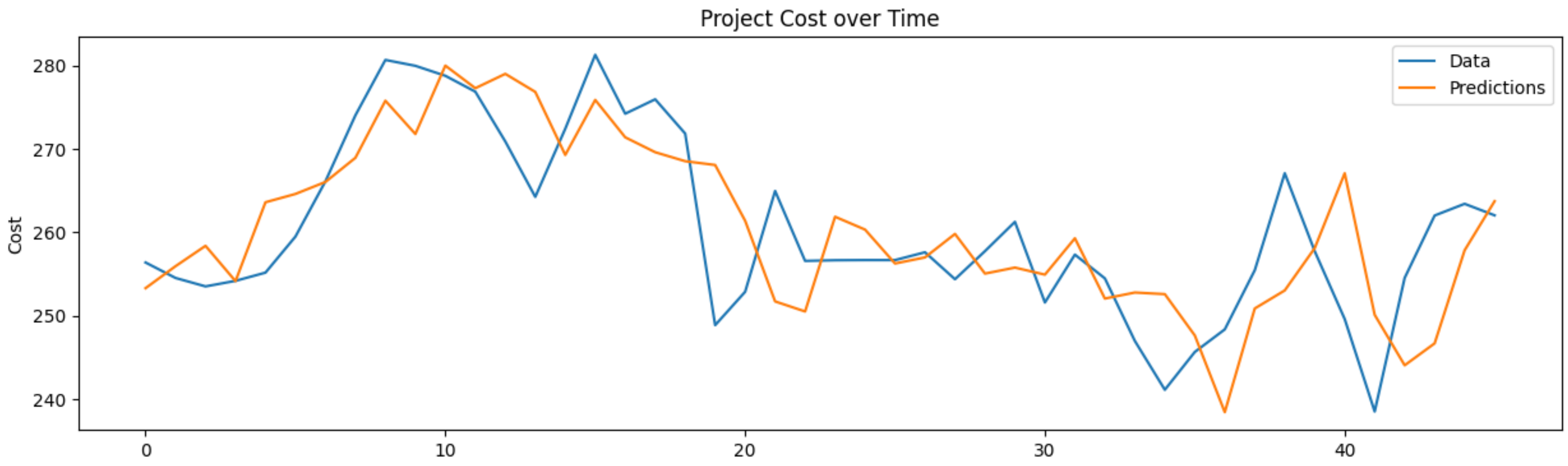


m=31,31,31,31
RMSE = 2.84

Sarima 模型優化結果

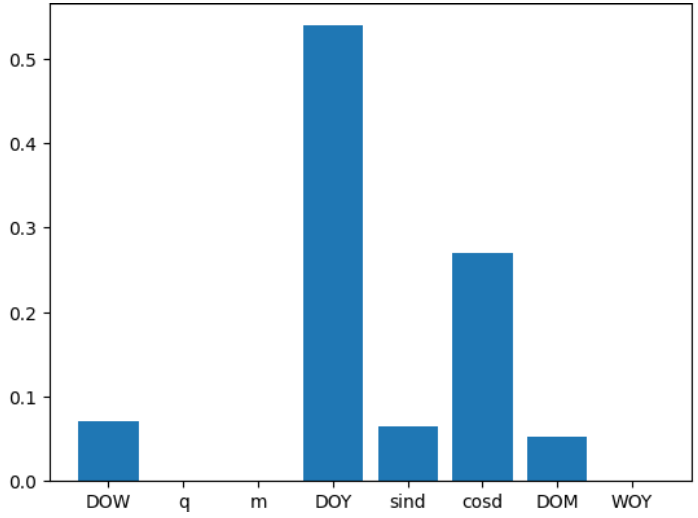
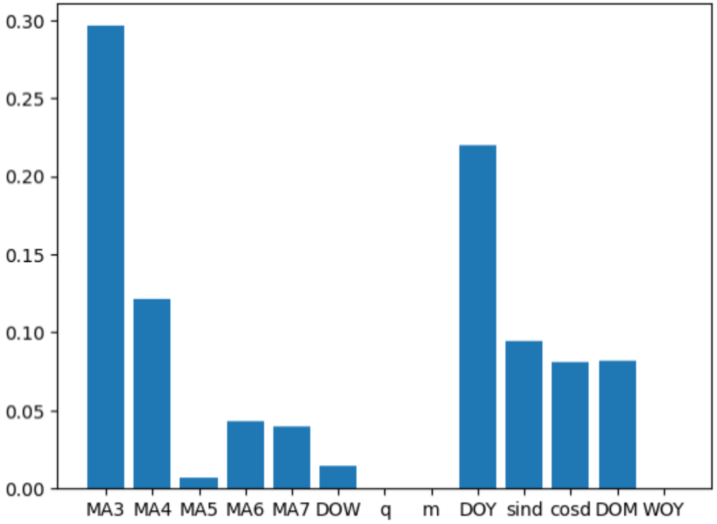
RMSE	SARIMA	優化後SARIMA	Difference
Project A	7.87	6.49	-21%
Project B	6.68	4.72	-29%
Project C	2.29	2.84	+19%

Project A: XGBoost (Rolling-based: 40)

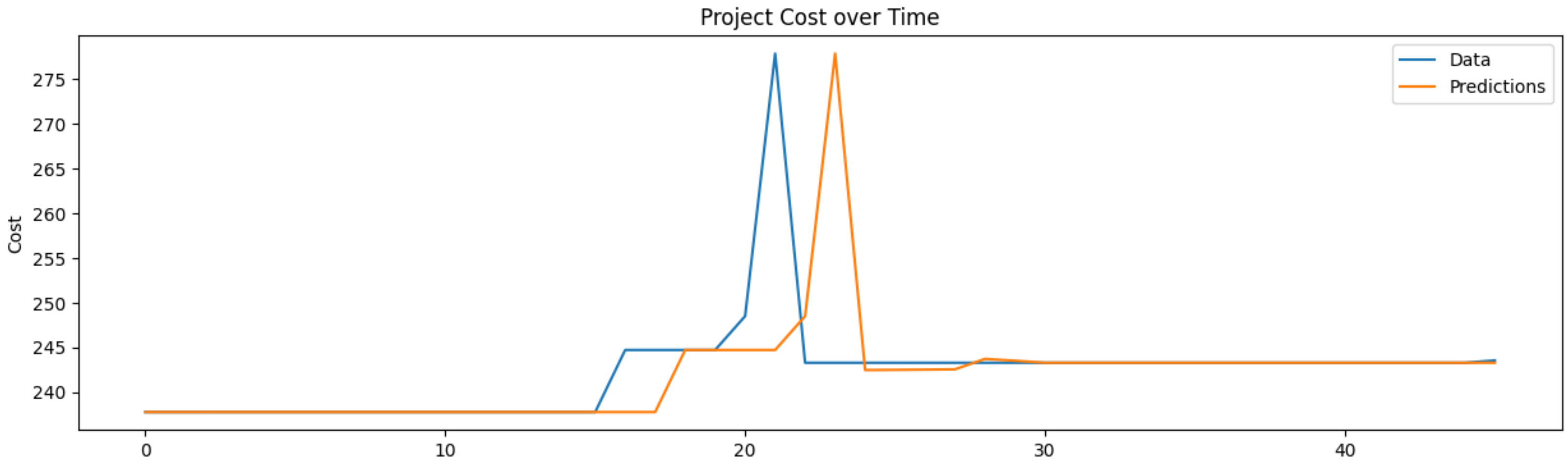


RMSE: 7.61

Project A: XGBoost Features Importance (Rolling-based: 40)

優化前	優化後
	

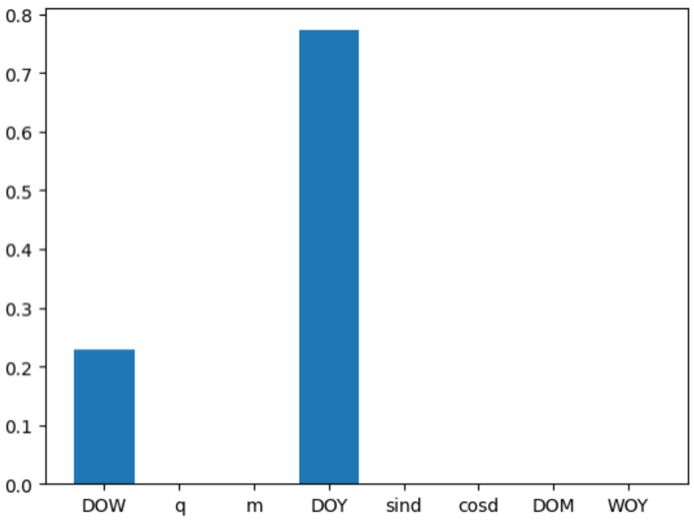
Project B: XGBoost (Rolling-based: 25)

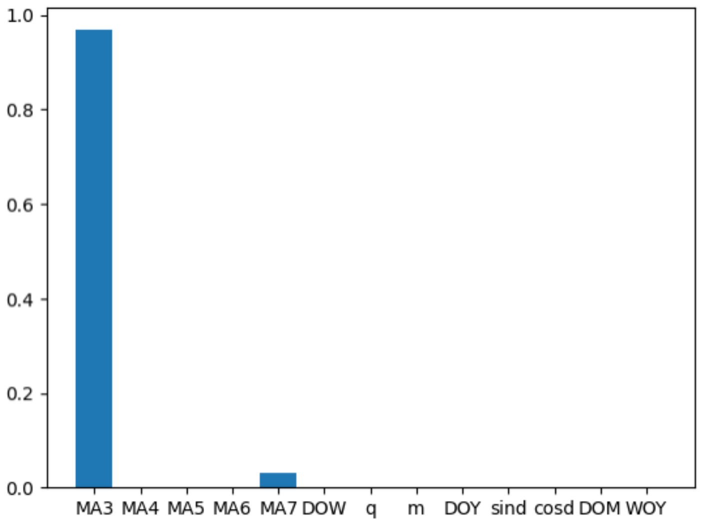


RMSE = 7.28

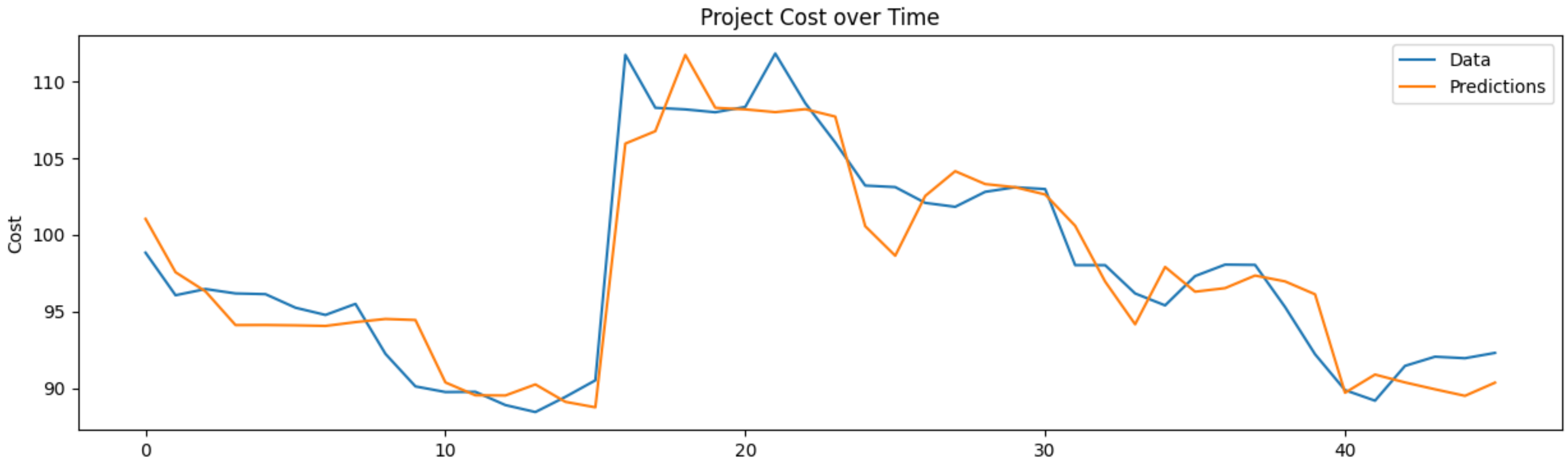
Project B: XGBoost Features Importance (Rolling-based: 25)

優化前	優化後





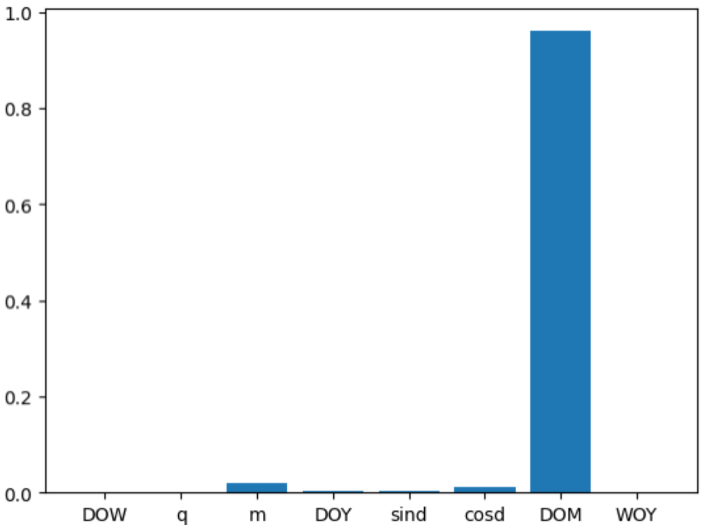
Project C: XGBoost (Rolling-based: 35)

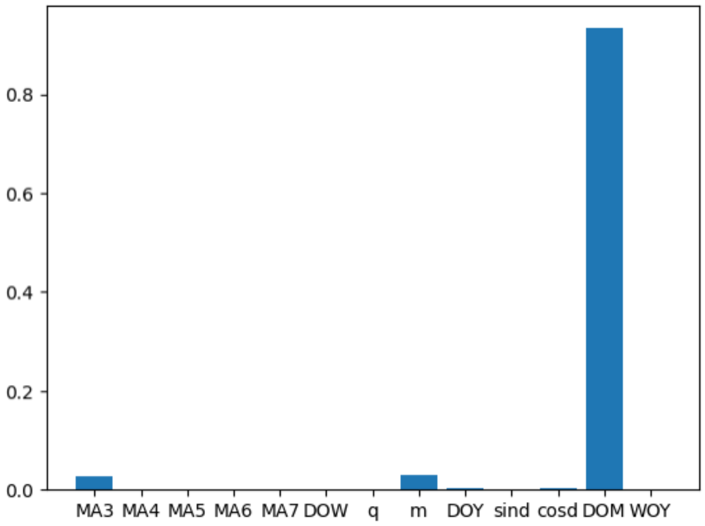


RMSE = 2.29

Project C: XGBoost Features Importance (Rolling-based: 45)

優化前	優化後

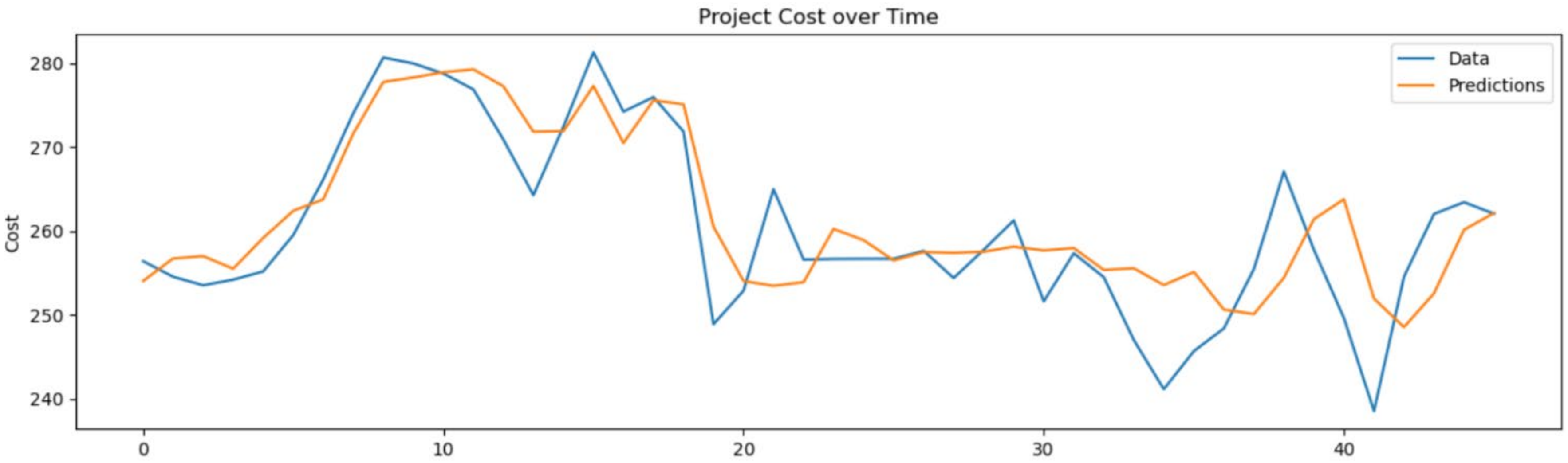




XGBoost 模型優化結果

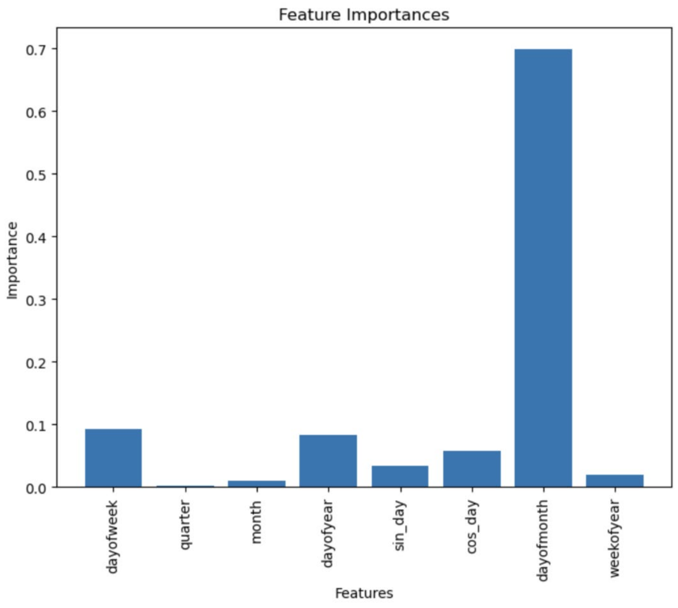
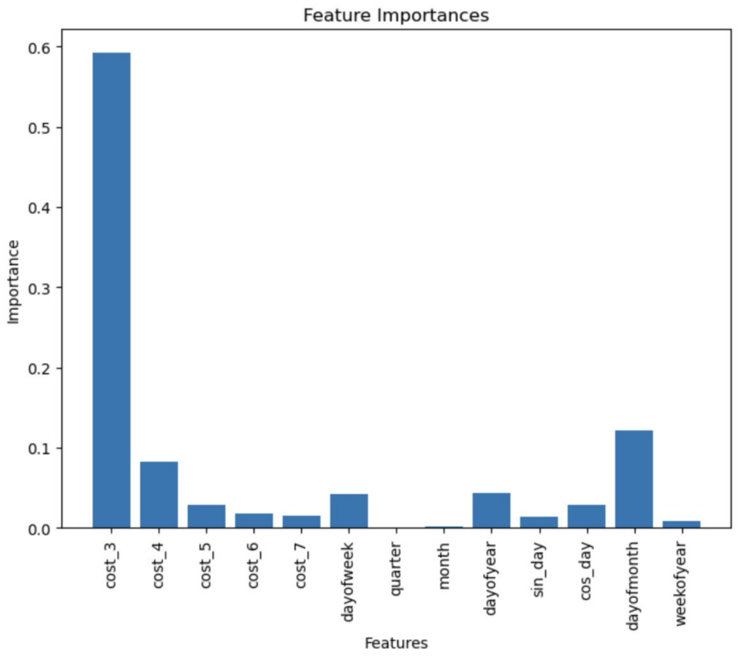
RMSE	XGBoost	優化後 XGBoost	Difference
Project A	9.06	7.61	-16%
Project B	7.53	7.28	-3.3%
Project C	2.29	2.13	-7%

Project A: RandomForest (Rolling-based: 75)

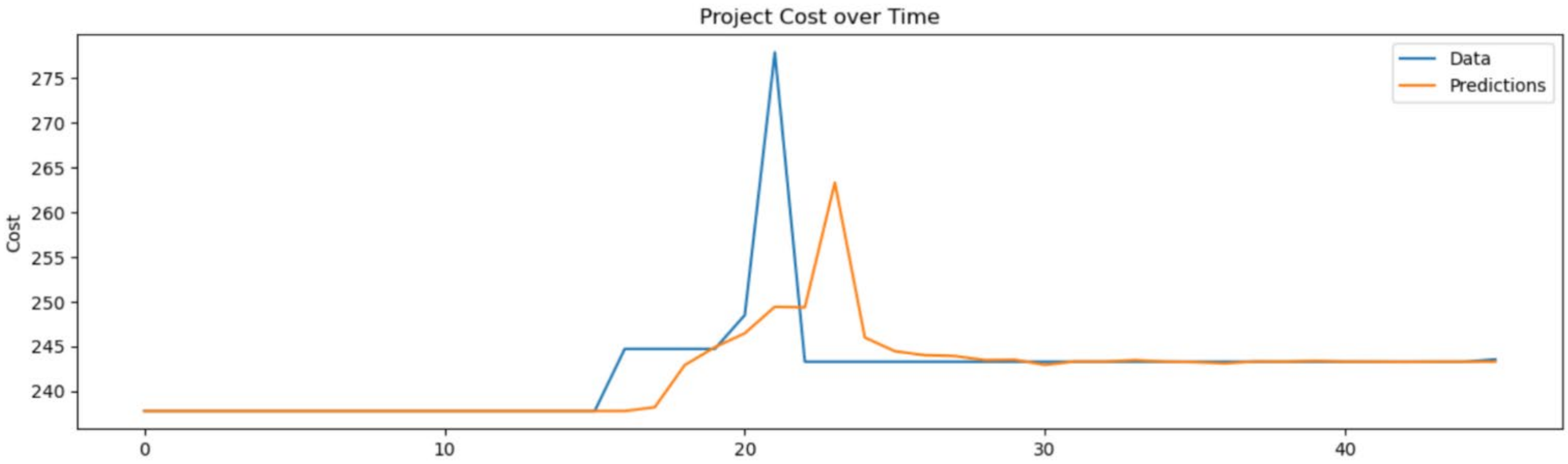


RMSE = 8.13

Project A: RandomForest Features Importance (Rolling-based: 75)

優化前	優化後
	

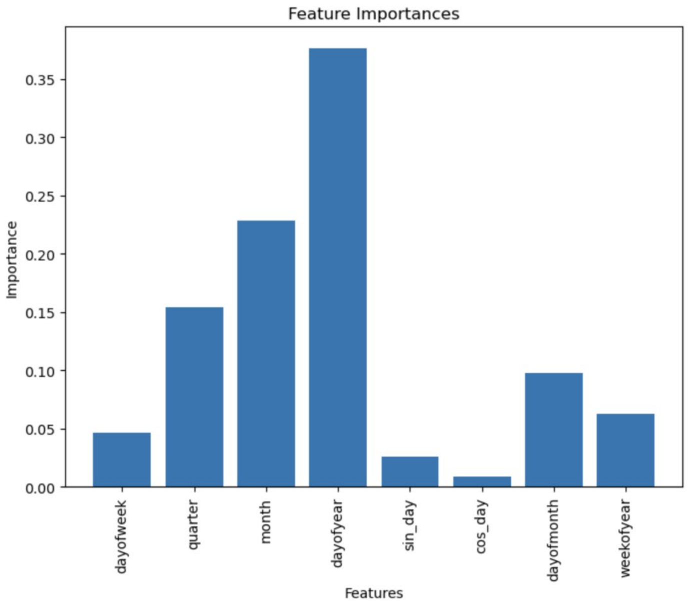
Project B: RandomForest (Rolling-based: 50)

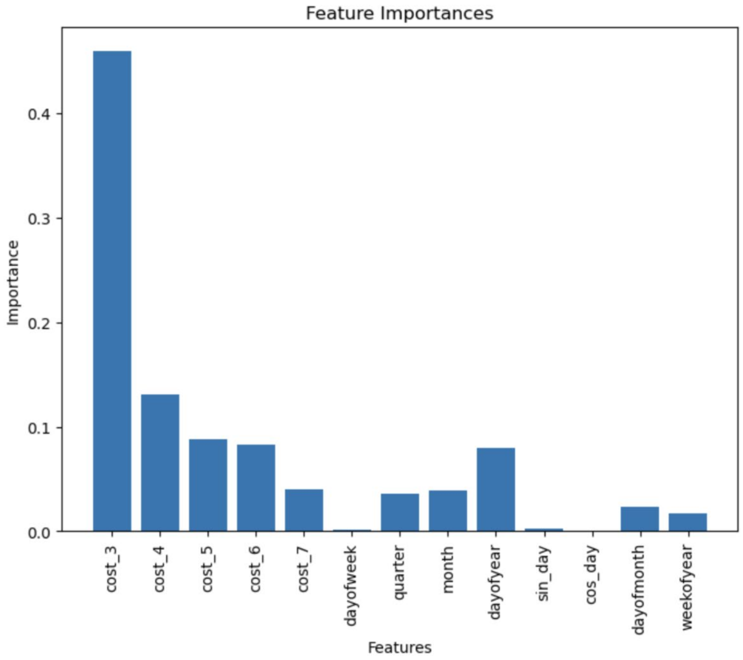


RMSE = 5.43

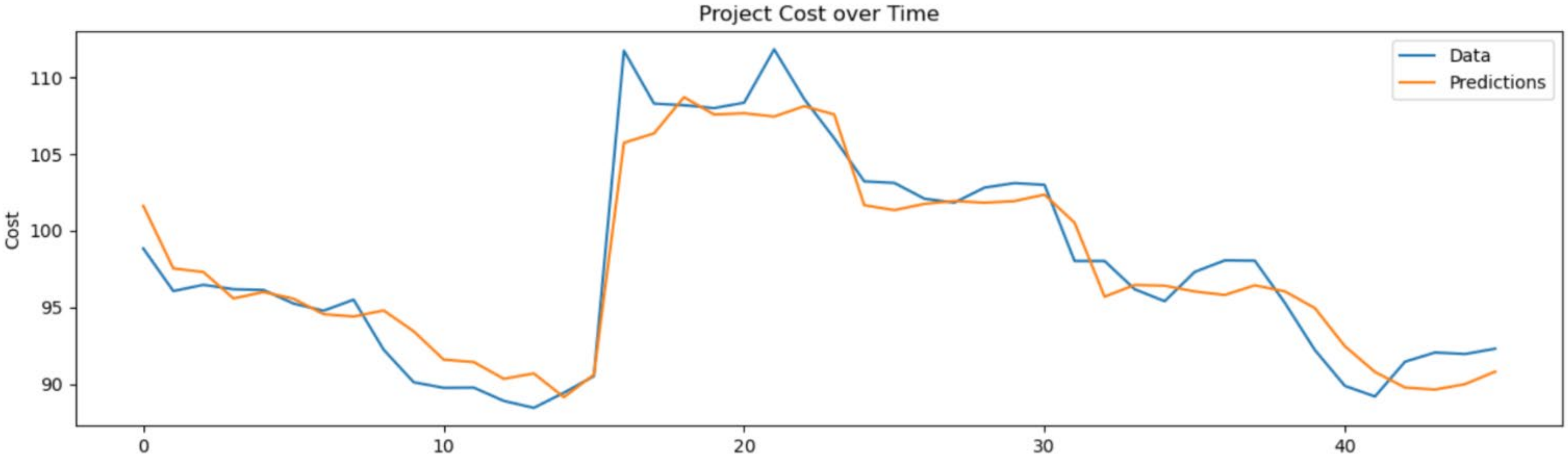
Project B: RandomForest Features Importance (Rolling-based: 50)

優化前	優化後



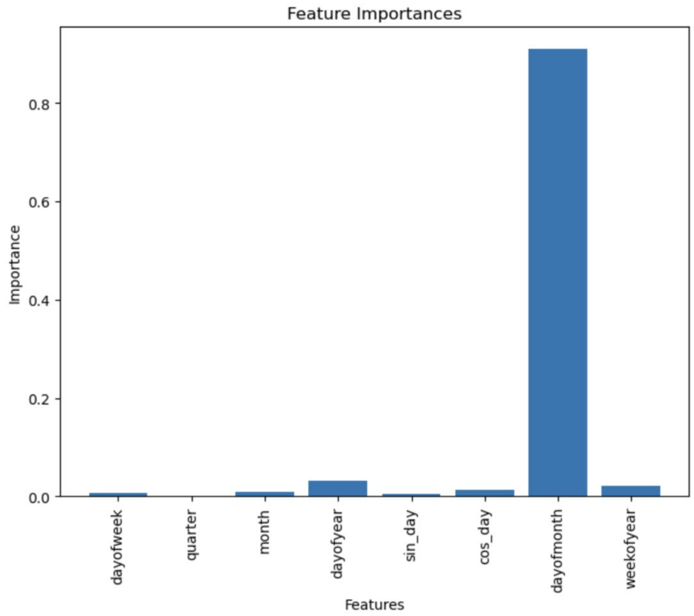
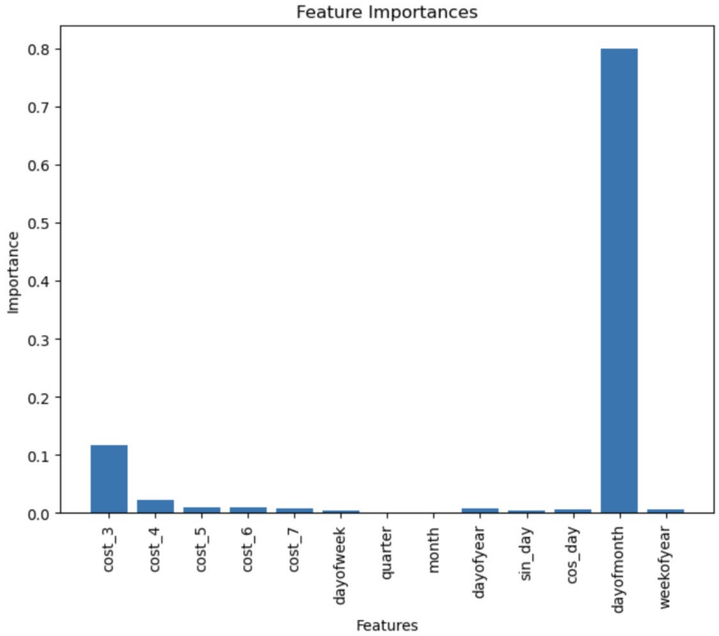


Project C: RandomForest (Rolling-based: 60)



RMSE = 2.91

Project C: RandomForest Features Importance (Rolling-based: 60)

優化前	優化後
	

RandomForest 模型優化結果

RMSE	RandomForest	優化後 RandomForest	Difference
Project A	10.02	8.13	-19%
Project B	7.16	5.43	-24%
Project C	2.21	2.91	+32%

Agenda

分析雲端服務使用量的費用預測，以提升使用者體驗，作為 MileLync 產品的加分項

結論

1. 建議模型篩選上可採用 SARIMA 來預測雲端費用
2. Tree-based 在進行適當特徵工程後表現提升，建議可參考較好模型使用之參數作為特徵
3. 我們使用 SARIMA Moving Average 參數作為 Tree-Based 的特徵優化模型表現

模型效益

1. 有效提升GCP使用量預測功能，打敗 Baseline 分別提升 RMSE 約 27% (A)、37% (C)
2. SARIMA 定期自動選參數與季節性
3. 在有限資料 (90天) 即可達到預測效益

Question

附錄

FEDFormer 加入 MA 後結果

RMSE	FEDFORMER	加入 MA後	Difference
Project A	11.20	12.16	+8%
Project B	7.06	6.93	-2%
Project C	3.43	3.14	-9%