Fraud detection and Forcsating
Case Study at Aviation Industry

by
Omid Ghorbani

OVERVIEW

01 **Explorary data analysis**

02 **Fraud detection**

03 **Forcasting with base line model**

04 **Forcasting using ML**

05 **Adding feature and hyper parameter tuning**

06 **Future Work**

Flight departure: 3pm

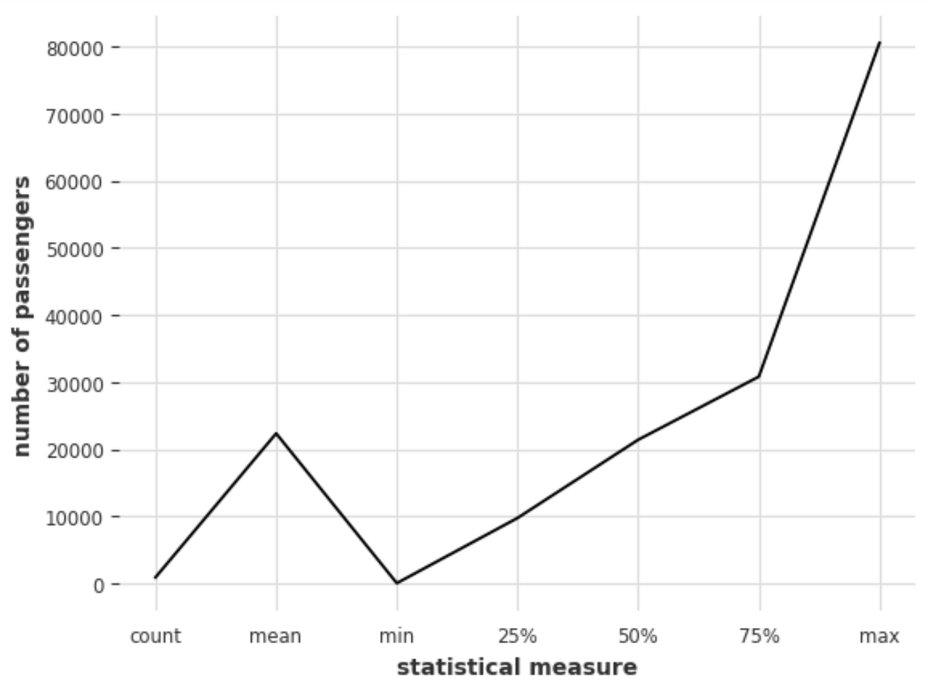
My parents at 6am

4.5B

E D A

E D A

	YYMM	AIRLINE	ORG	DST	PAX	SEATS	SEAT_FACTOR
0	2023-10-01	B6	EWR	LAX	14572	18285	0.7969
1	2023-10-01	B6	JFK	LAX	42028	50549	0.8314
2	2023-10-01	B6	JFK	MCO	32501	38233	0.8501
3	2023-10-01	B6	LAX	EWR	14862	18285	0.8128
4	2023-10-01	B6	LAX	JFK	41442	50072	0.8276

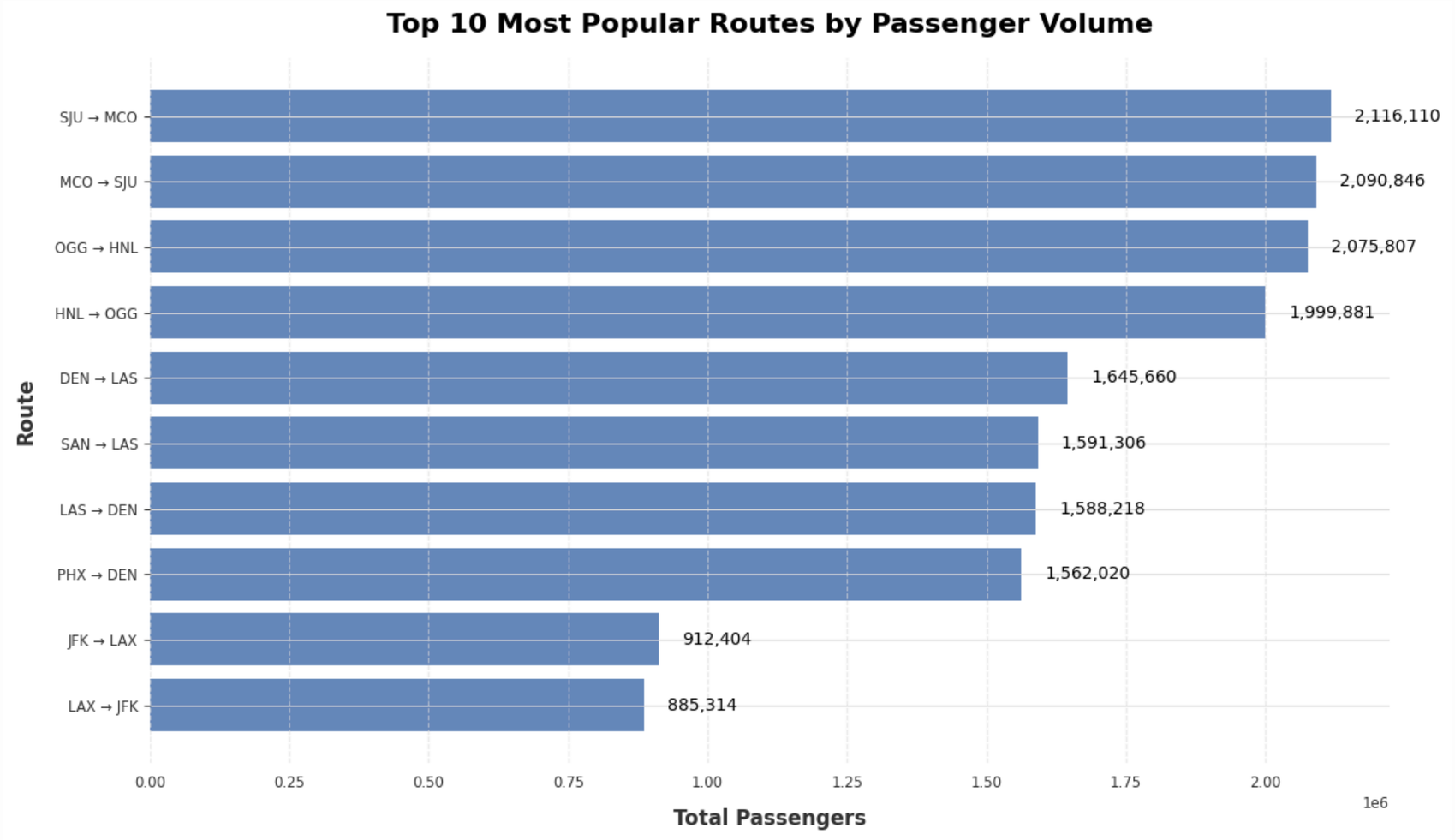


Average : 20k passengers per month for all airlines

POPULAR DESTINATION

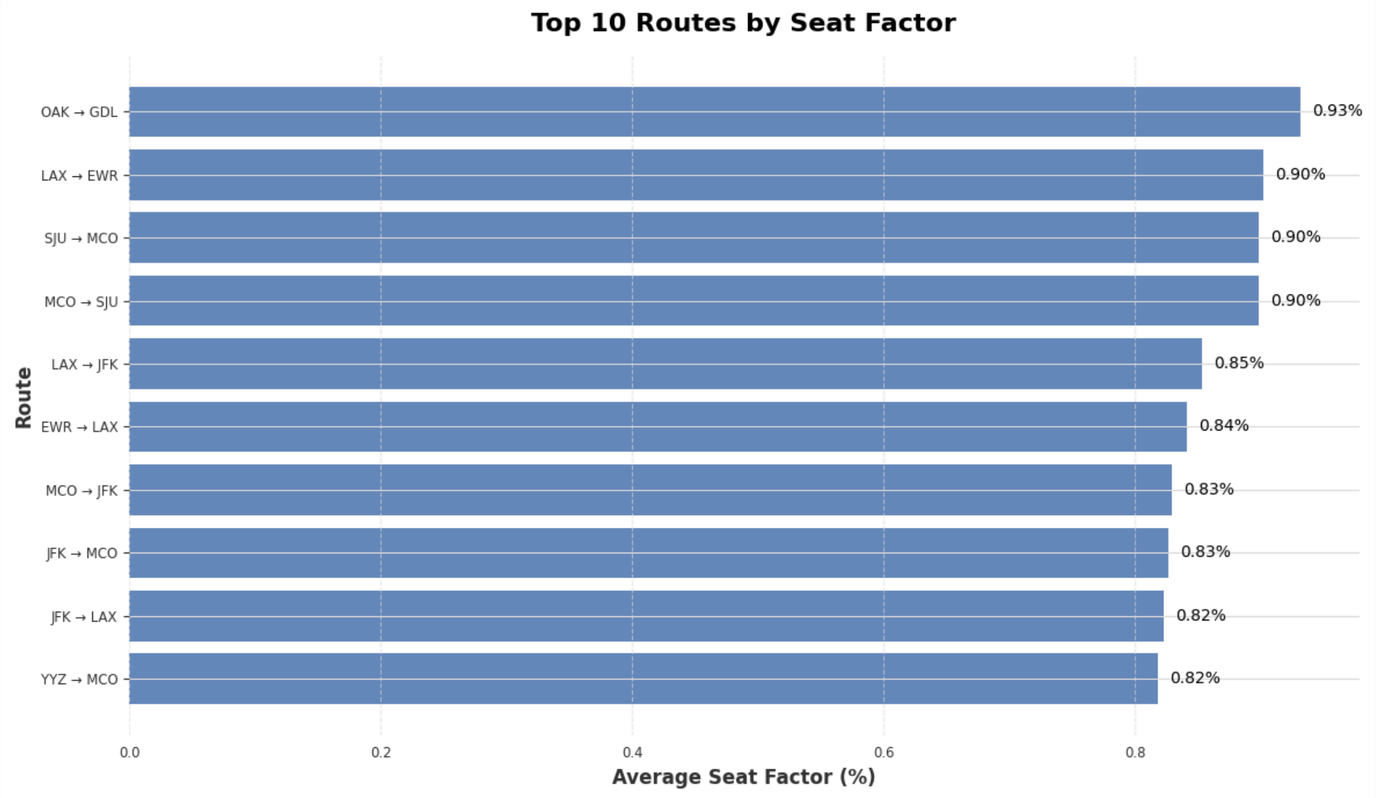
close route

Route	Origin	Destination
SJU → MCO	Puerto Rico (San Juan)	United States (Florida -
MCO → SJU	United States (Florida -	Puerto Rico (San Juan)
OGG → HNL	United States (Hawaii - Maui)	United States (Hawaii -



EFFICIENT ROUTE

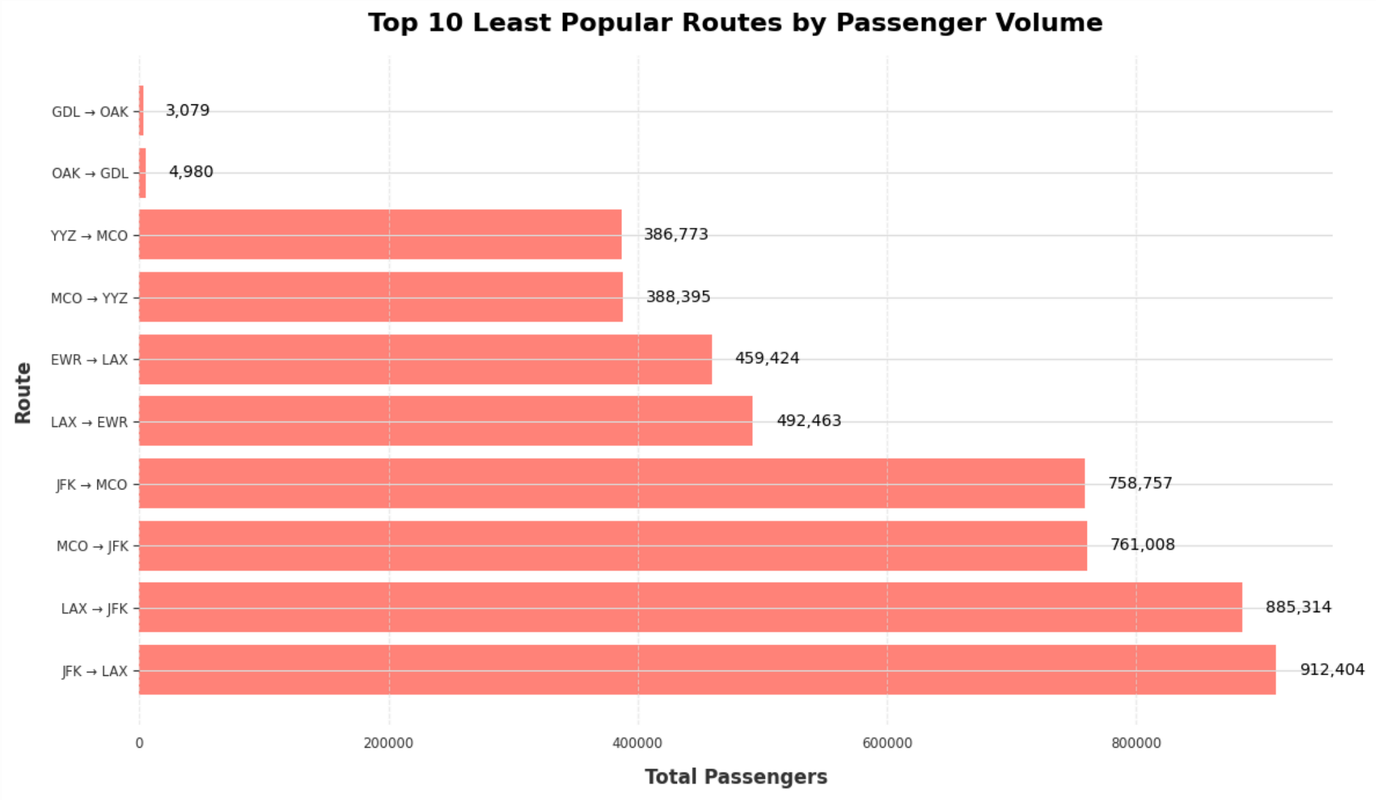
Route	Origin	Destination
OAK → GDL	United States (California)	Mexico (Guadalajara)
LAX → EWR	United States (California)	United States (New Jersey)
SJU → MCO	Puerto Rico (San Juan)	United States (Florida)



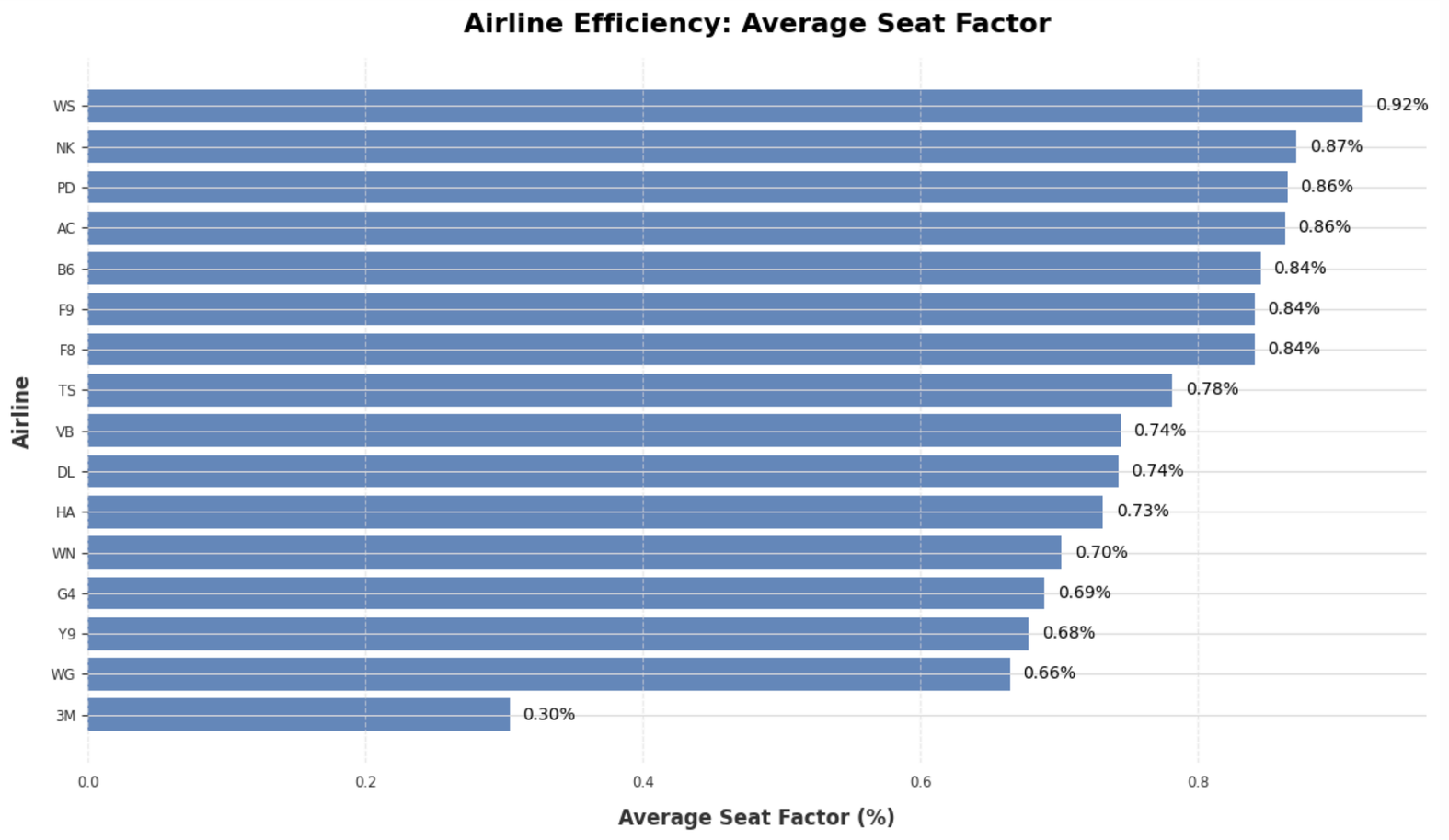
Lest common route

far route

GDL → OAK	Mexico (Guadalajara)	United States (California)
OAK → GDL	United States (California)	Mexico (Guadalajara)
YYZ → MCO	Canada (Toronto)	United States (Florida)

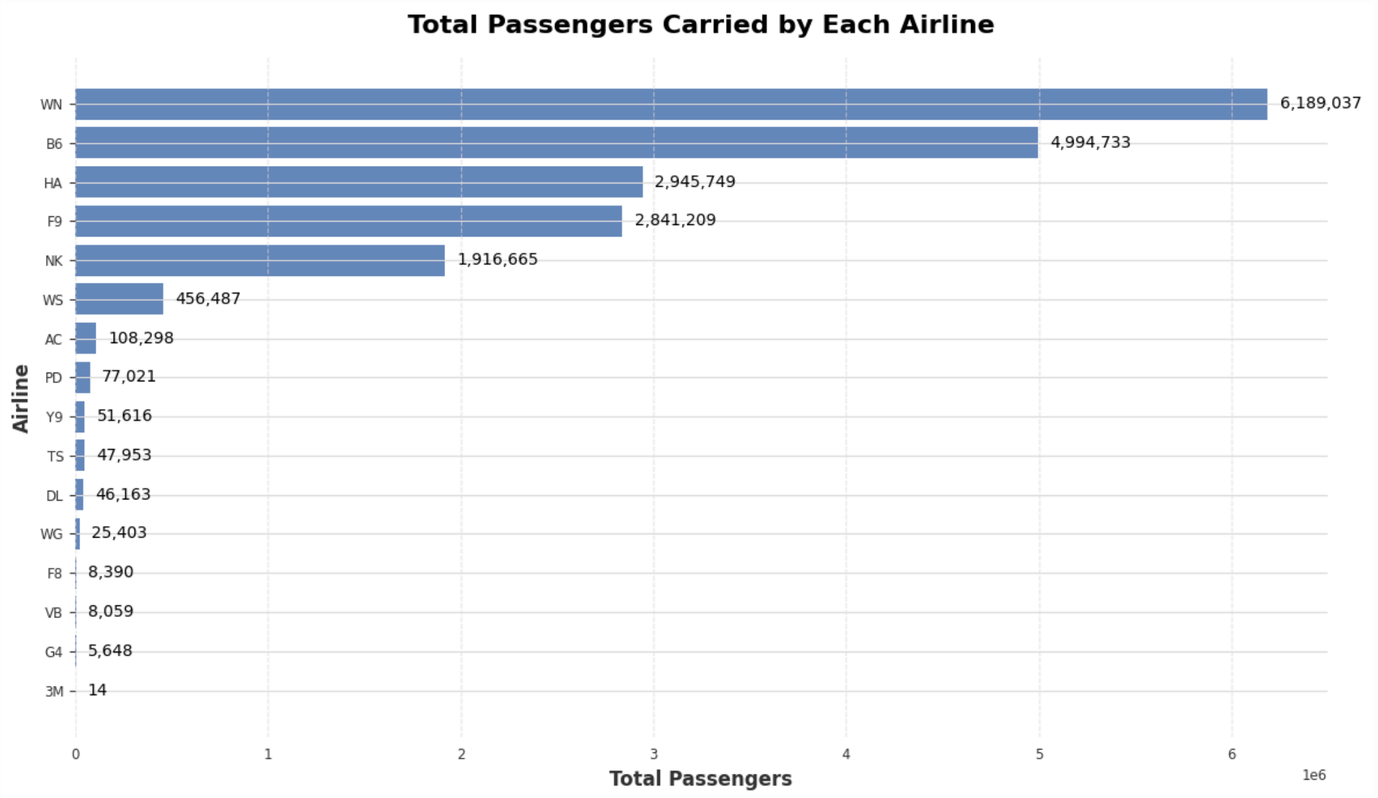


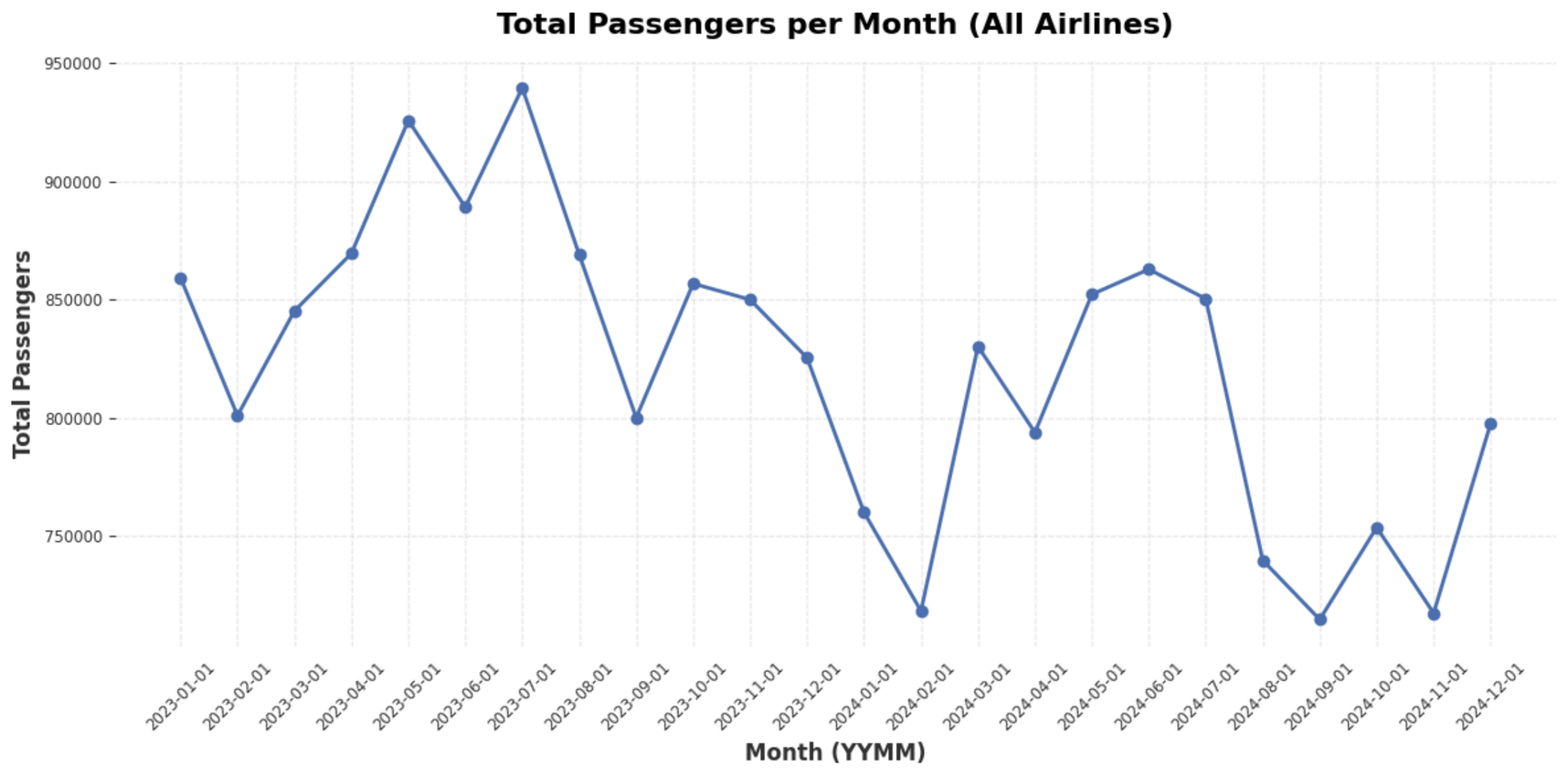
Code	Airline Name	Country
WS	WestJet Airlines	Canada
NK	Spirit Airlines	United States



Code	Airline Name	Country
WN	Southwest Airlines	United States
B6	JetBlue Airways	United States
HA	Hawaiian Airlines	United States

Exp dist



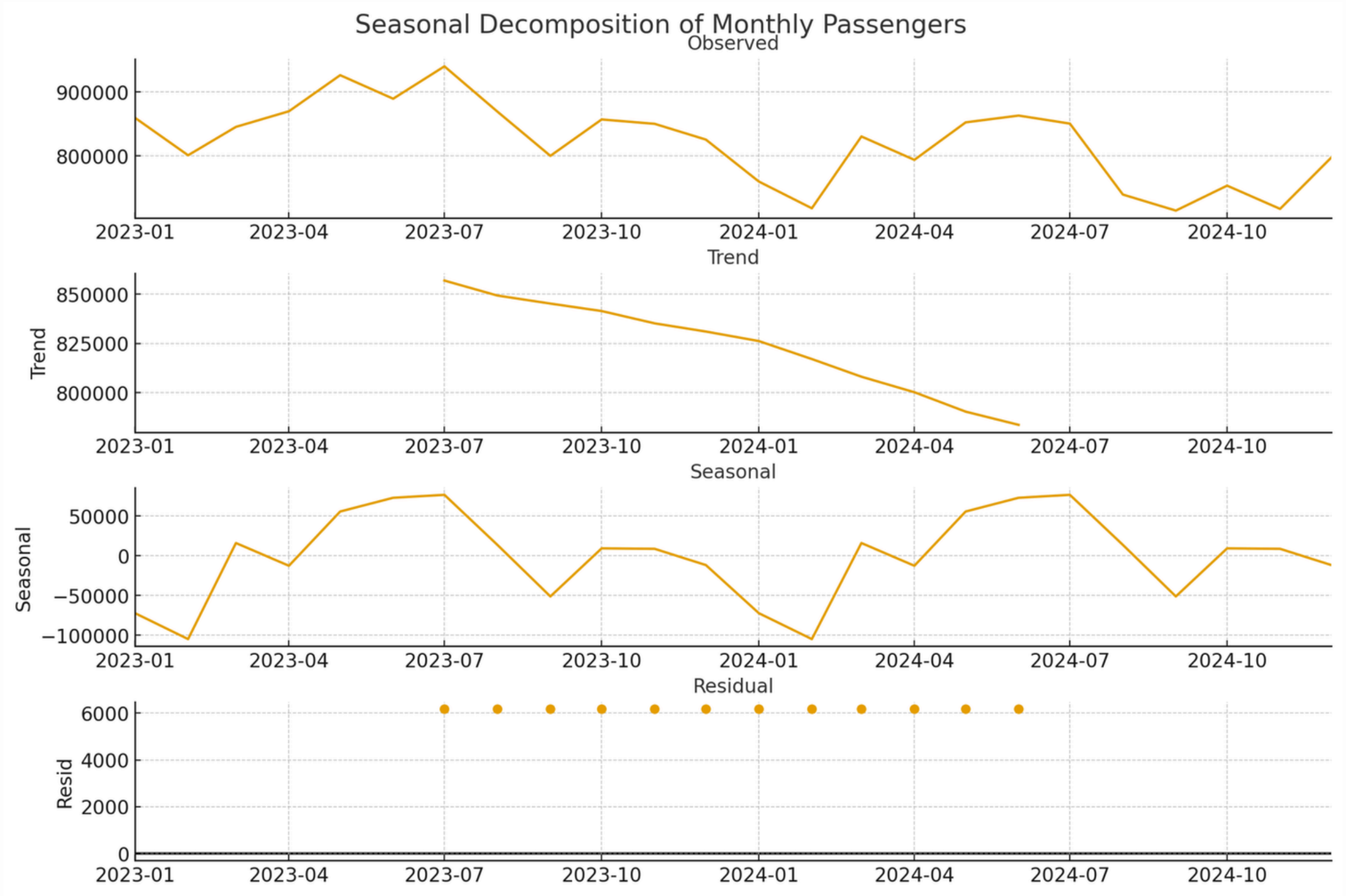


Peaks in May–July 2023 (early summer travel).

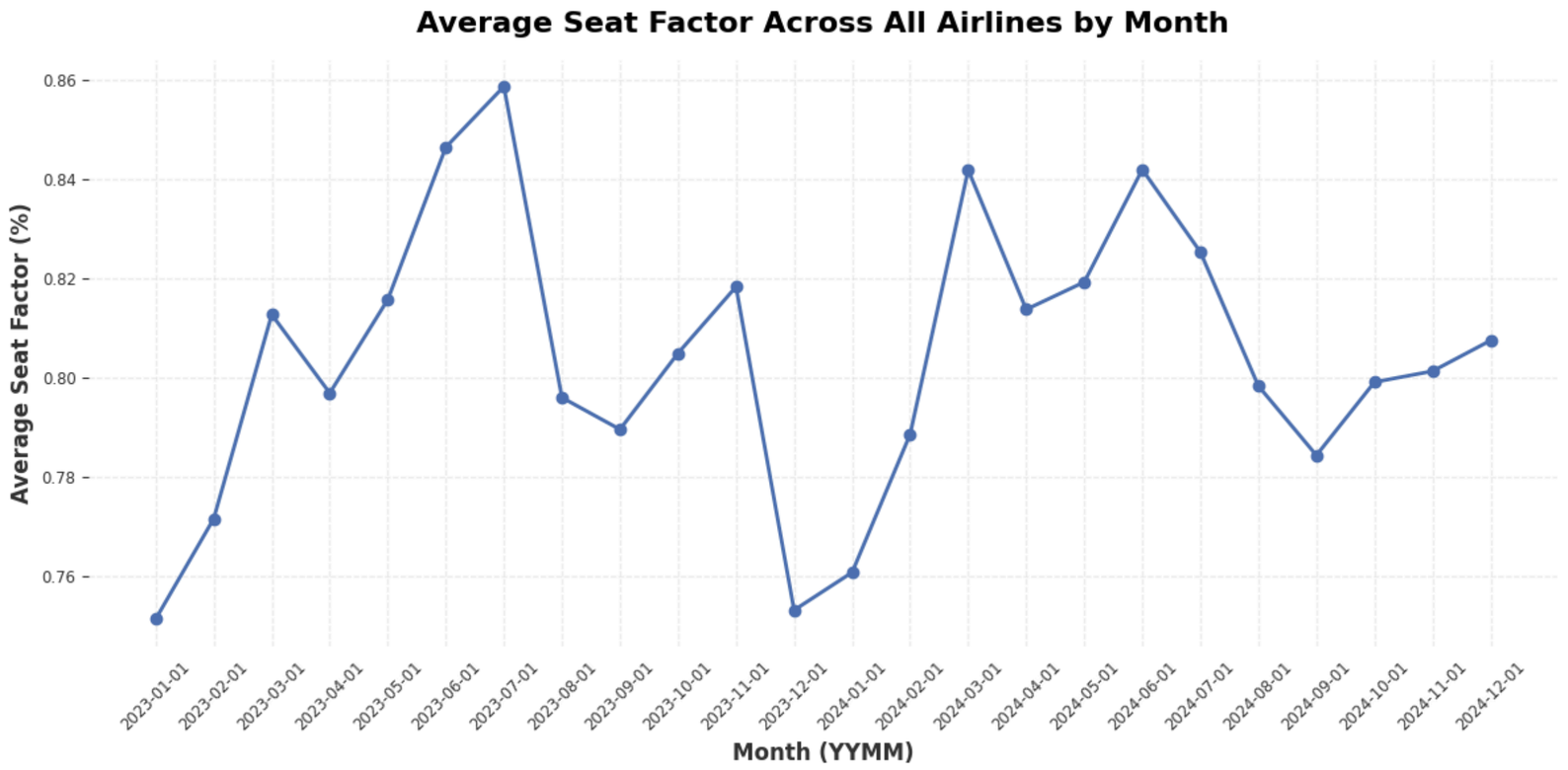
Dips around January–February 2024. (after holiday - no money)

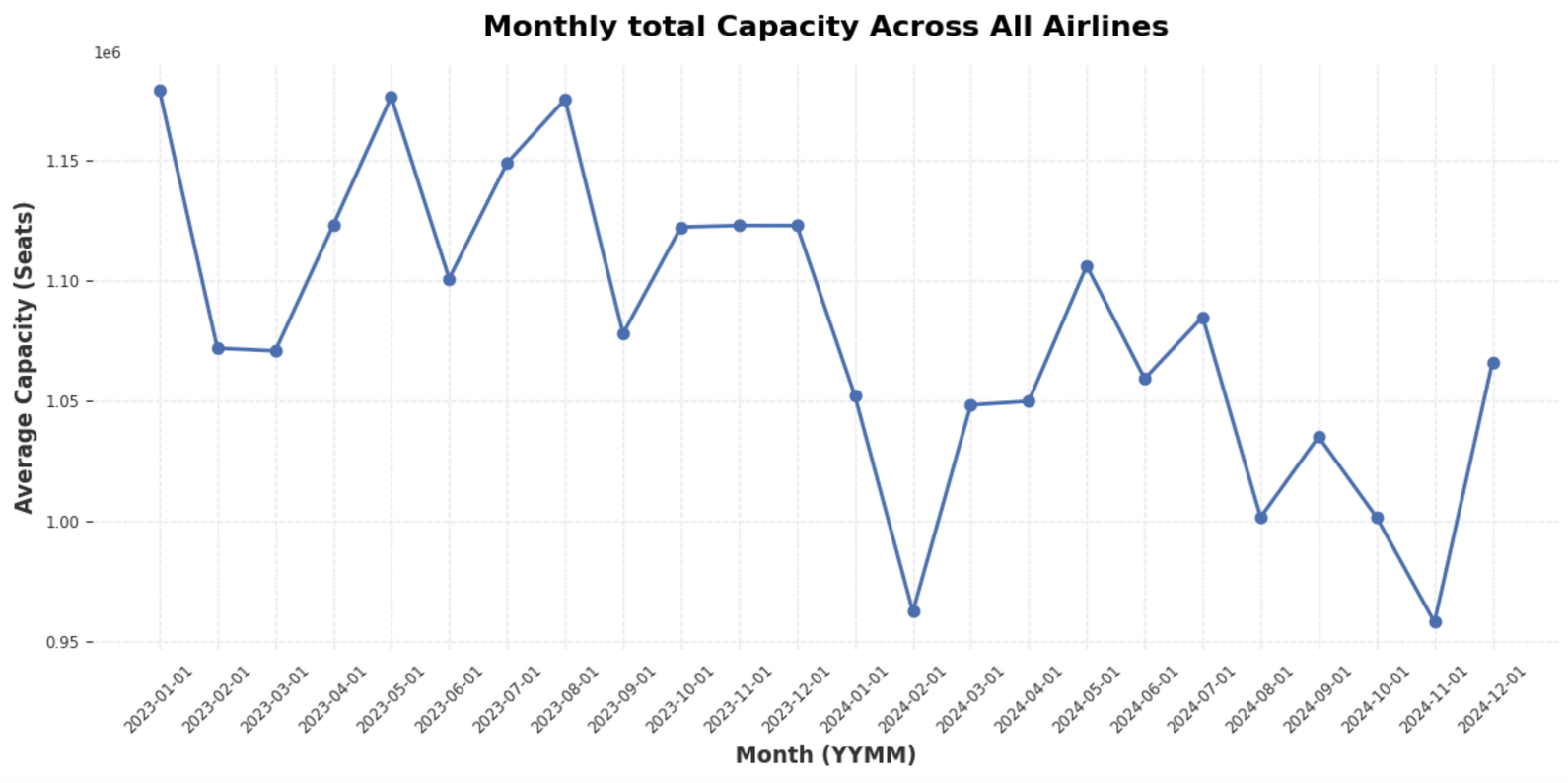
downward trend in total passengers over time.

Seasonal decomposition



SEAT FACTOR


Average Seat Factor Across All Airlines by Month
Average Seat Factor (%)
0.86
0.84
0.82
0.80
0.78
0.76
2023-01-01
2023-02-01
2023-03-01
2023-04-01
2023-05-01
2023-06-01
2023-07-01
2023-08-01
2023-09-01
2023-10-01
2023-11-01
2023-12-01
2024-01-01
2024-02-01
2024-03-01
2024-04-01
2024-05-01
2024-06-01
2024-07-01
2024-08-01
2024-09-01
2024-10-01
2024-11-01
2024-12-01
Month (YYMM)



RESONABLE, SINCE THE TREND IS DOWNWARD FOR PASSENGERS

Fraud detection

RULE BASE ANOMALY DETECTION

IF seat factor > 1
Then Anomaly

	YYMM	AIRLINE	ORG	DST	PAX	SEATS	SEAT_FACTOR
60	2024-06-01	WN	DEN	LAS	49840	49096	1.0151
678	2024-09-01	NK	LAX	EWR	15400	12097	1.2730
743	2023-11-01	DL	SAN	LAS	6321	4130	1.5305

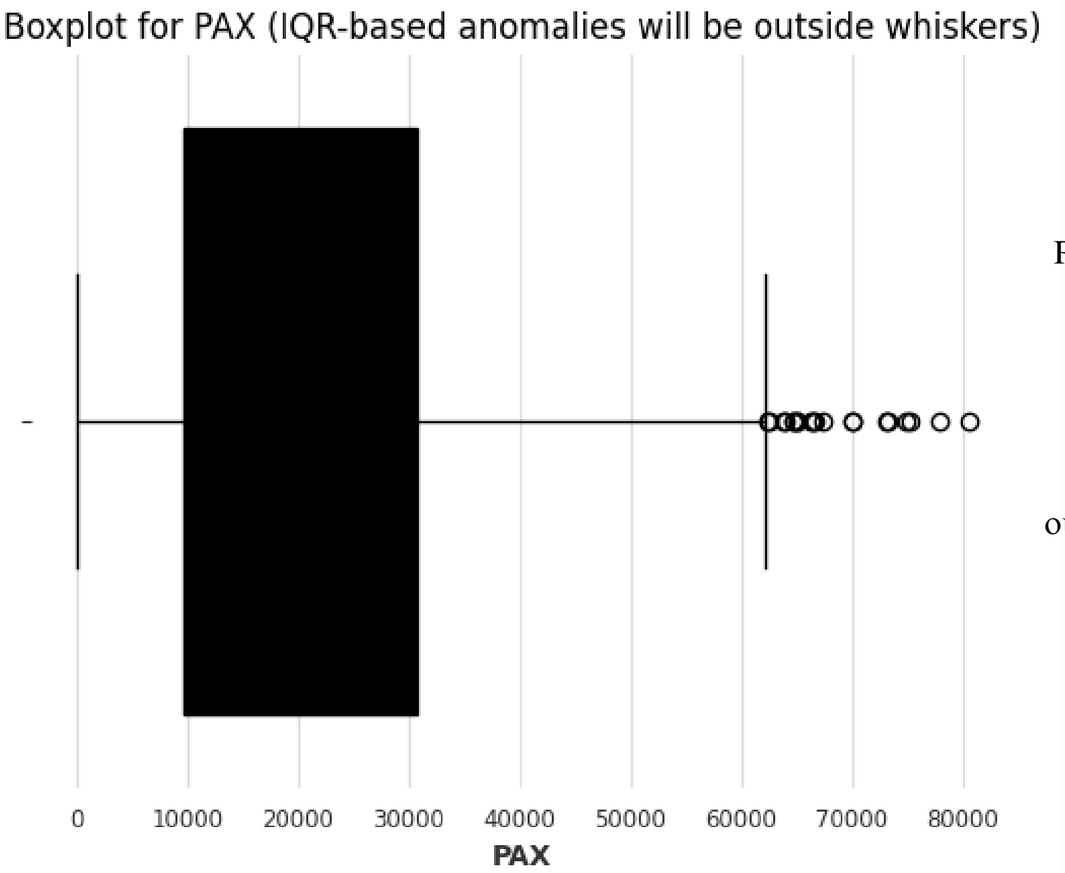
Total anomalies detected: 3

filled all missing value with forward fill

filling jan base on december

UNIVARIATE ANOMALY DETECTION

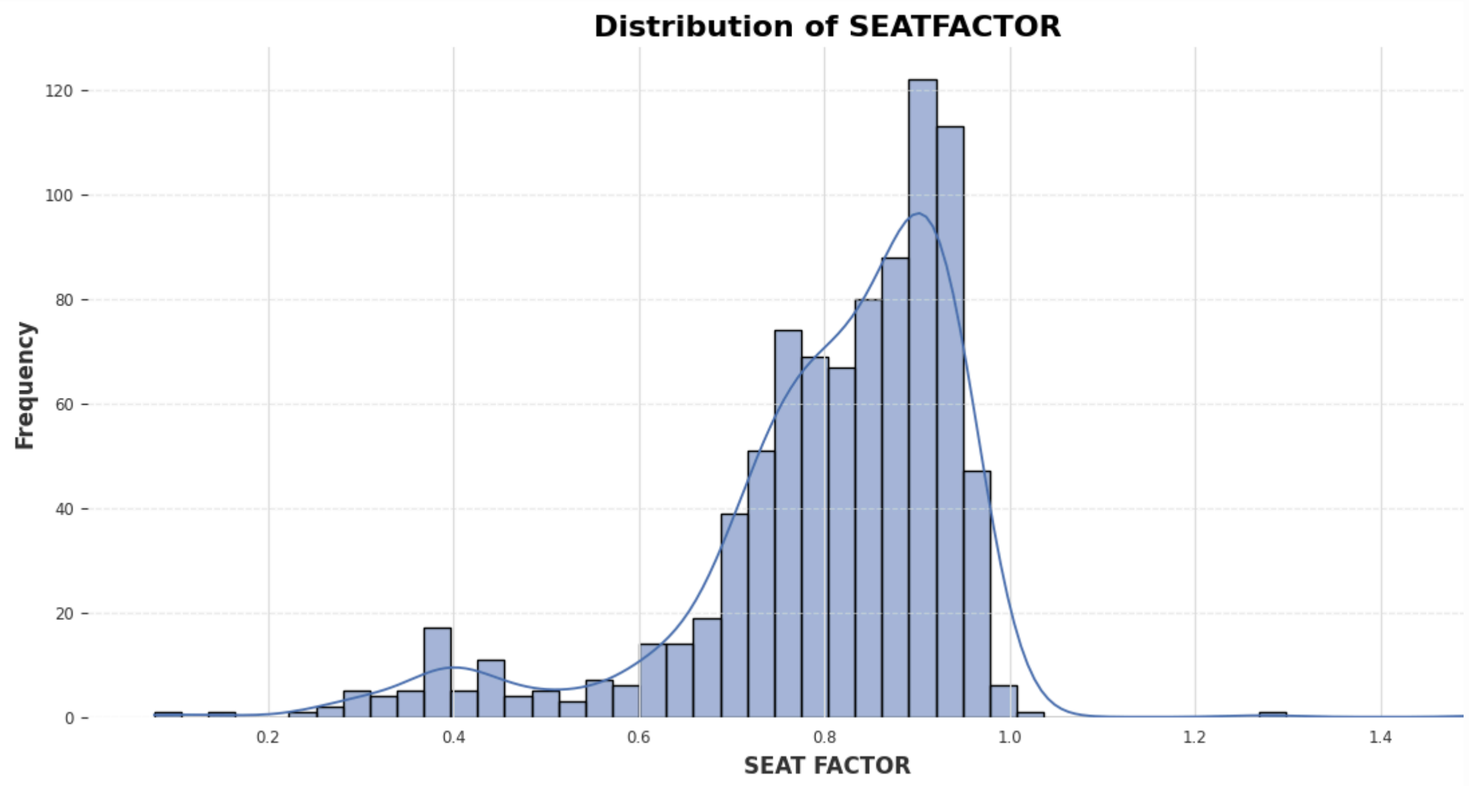
	YYMM	AIRLINE	ORG	DST	PAX	SEATS	SEAT_FACTOR
15	2023-10-01	HA	HNL	OGG	64705	90368	0.7160
16	2023-10-01	HA	OGG	HNL	62594	90540	0.6913
85	2023-04-01	HA	HNL	OGG	70024	92093	0.7604
86	2023-04-01	HA	OGG	HNL	66411	91693	0.7243
126	2023-05-01	HA	HNL	OGG	73089	95273	0.7672
127	2023-05-01	HA	OGG	HNL	70047	94801	0.7389
165	2023-06-01	HA	HNL	OGG	74952	93611	0.8007
166	2023-06-01	HA	OGG	HNL	73345	93355	0.7857
243	2023-01-01	HA	HNL	OGG	66661	88375	0.7543
244	2023-01-01	HA	OGG	HNL	65092	88308	0.7371
325	2023-03-01	HA	OGG	HNL	75196	97168	0.7739
398	2024-07-01	HA	HNL	OGG	67366	83005	0.8116
399	2024-07-01	HA	OGG	HNL	63997	82966	0.7714
463	2023-08-01	HA	HNL	OGG	66220	97168	0.6815
464	2023-08-01	HA	OGG	HNL	62428	98326	0.6349
537	2024-08-01	HA	HNL	OGG	65032	84179	0.7725
571	2024-03-01	HA	HNL	OGG	64644	82682	0.7818
572	2024-03-01	HA	OGG	HNL	63839	82704	0.7719
860	2023-07-01	HA	HNL	OGG	80535	97491	0.8261
861	2023-07-01	HA	OGG	HNL	77947	97174	0.8021

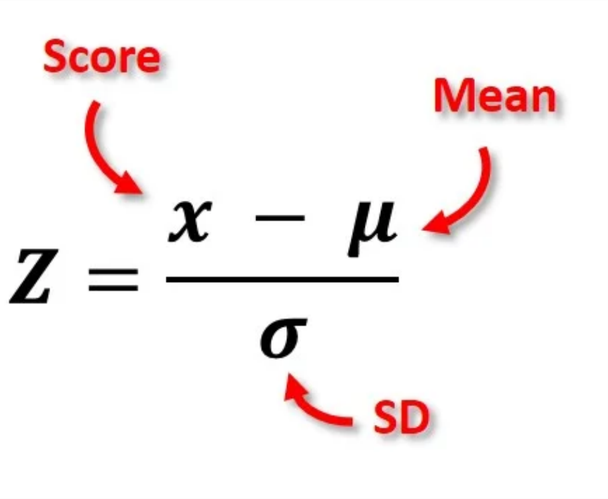


Pros : dont assume a distribution for data
simple ,
work with one variable

cons : doesnt consider all features -
outlier might be normal osscilation of data

UNIVARIATE ANOMALY DETECTION- ZSCORE





Anomalies detected using Z-score for Seat Factor:

	YYMM	AIRLINE	ORG	DST	SEAT_FACTOR	SEAT_FACTOR_Z
385	2024-07-01	B6	JFK	MCO	0.2395	-3.730167
668	2024-09-01	DL	SAN	LAS	0.0775	-4.802144
688	2024-09-01	WN	OGG	HNL	0.2694	-3.532315
743	2023-11-01	DL	SAN	LAS	1.5305	4.812566
789	2023-12-01	NK	EWR	LAX	0.1411	-4.381294
795	2023-12-01	WG	MCO	YYZ	0.2576	-3.610397

The probability that Z-score is bigger than 3.5 or smaller than -3.5 is 5/1000

MULTIVARIATE MACHINE LEARNING ANOMALY DETECTION

```
iso_forest = IsolationForest(n_estimators=200, contamination=0.01, random_state=42)
iso_forest.fit(df_model)
```

Cons and pros : Detects anomalies across all routes combined.
Useful if you want to catch "outliers compared to the whole industry."
Risk: may biased toward one routes with anomalies.

best approach : hybrid

	YYMM	AIRLINE	ORG	DST	PAX	SEATS	SEAT_FACTOR	SEAT_FACTOR_Z
126	2023-05-01	HA	HNL	OGG	73089	95273	0.7672	-0.238300
165	2023-06-01	HA	HNL	OGG	74952	93611	0.8007	-0.016626
205	2024-12-01	HA	HNL	OGG	25600	81378	0.3146	-3.233220
463	2023-08-01	HA	HNL	OGG	66220	97168	0.6815	-0.805390
661	2024-09-01	B6	JFK	LAX	38348	95400	0.4020	-2.654881
668	2024-09-01	DL	SAN	LAS	2788	35959	0.0775	-4.802144
795	2023-12-01	WG	MCO	YYZ	633	2457	0.2576	-3.610397
860	2023-07-01	HA	HNL	OGG	80535	97491	0.8261	0.151449
861	2023-07-01	HA	OGG	HNL	77947	97174	0.8021	-0.007362

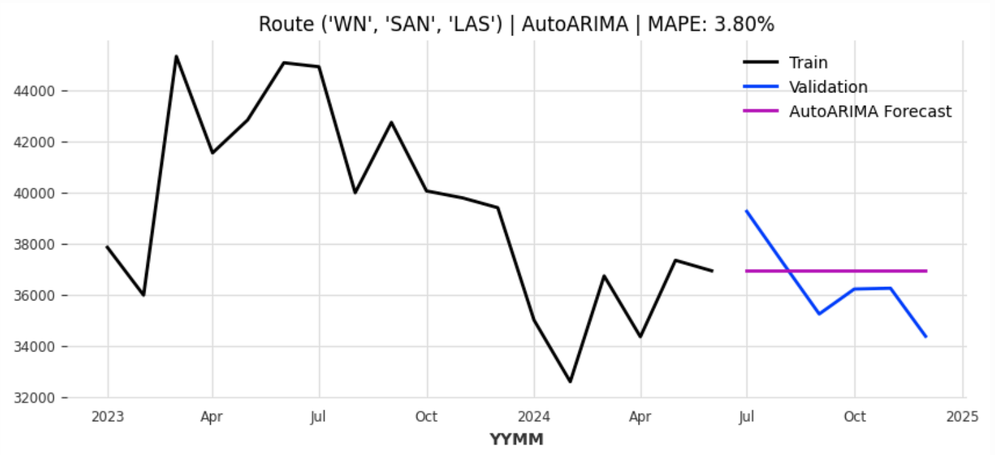
FORCASTING

RESULT ON 4 BASE MODEL

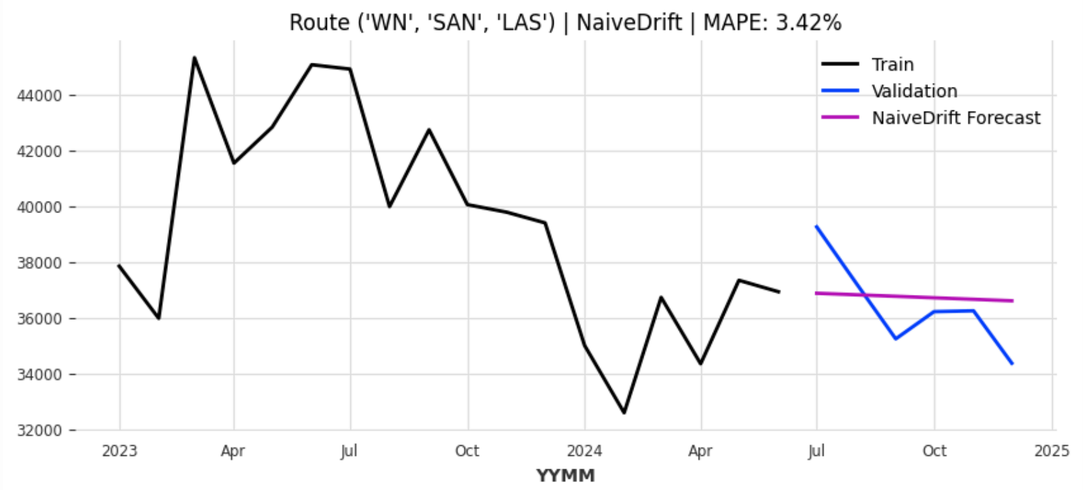
```
=== Average MAPE across all routes ===
             Model  Average_MAPE
3   AutoARIMA_MAPE     19.680876
2       Theta_MAPE     25.322377
0    ETS_None_MAPE     25.471893
1  NaiveDrift_MAPE     25.661375

✅ Best baseline model overall: AutoARIMA_MAPE (Average MAPE = 19.68%)
```

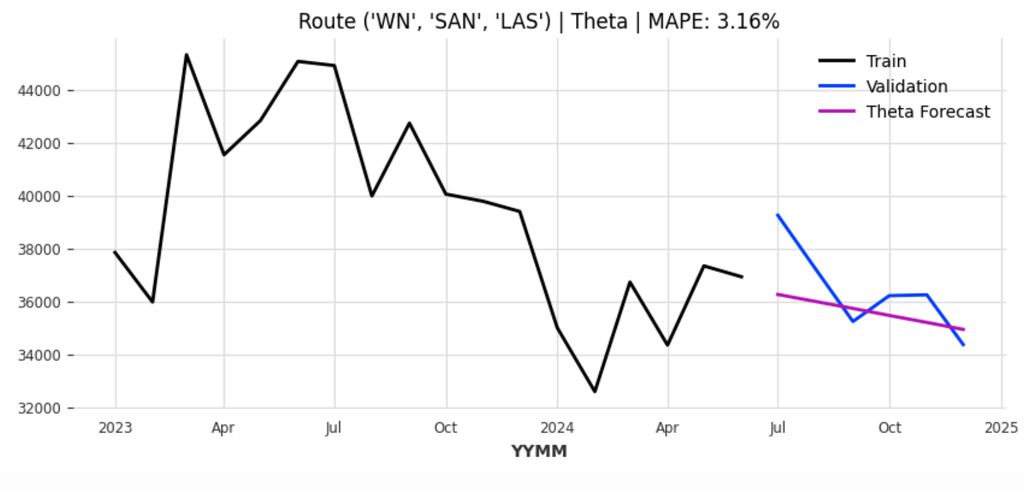
AutoARIMA (seasonal, m=12)
Fits an ARIMA model with yearly seasonality.
Captures trend + seasonality explicitly.

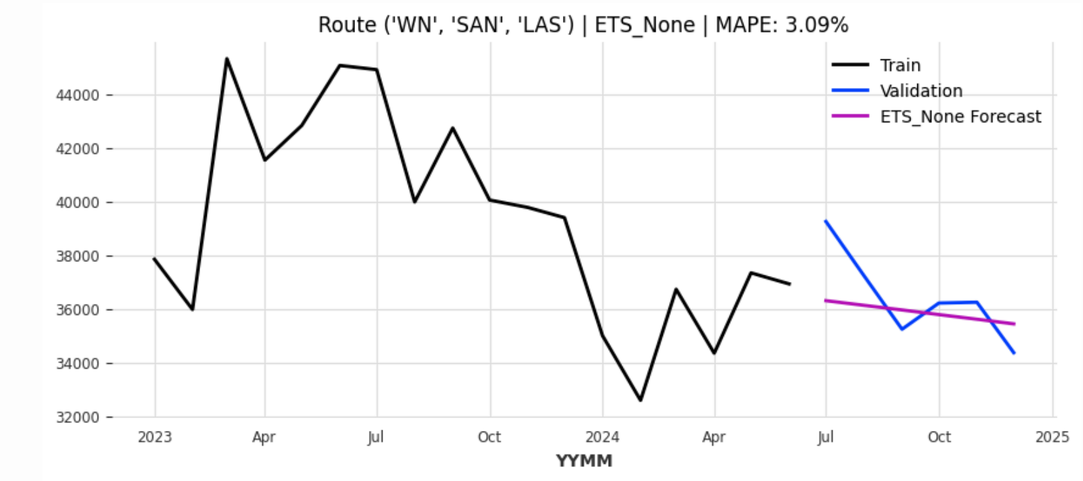
	Route	AutoARIMA_MAPE
27	(WN, SAN, LAS)	3.804845
15	(HA, OGG, HNL)	4.169653
4	(B6, LAX, JFK)	5.274487
19	(NK, SAN, LAS)	7.009181
28	(WN, SJU, MCO)	8.743351
1	(B6, JFK, LAX)	8.769753
30	(WS, YYZ, MCO)	10.424511
11	(F9, PHX, DEN)	11.814831
24	(WN, MCO, SJU)	12.118851
6	(B6, MCO, SJU)	12.627646

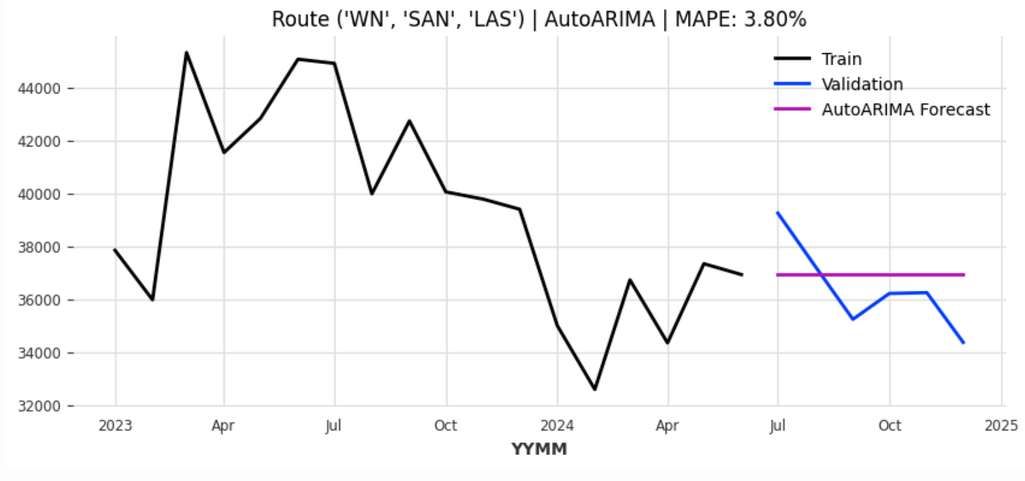


COMPARISON FOR ONE ROUTE









ML FORCASTING MODEL

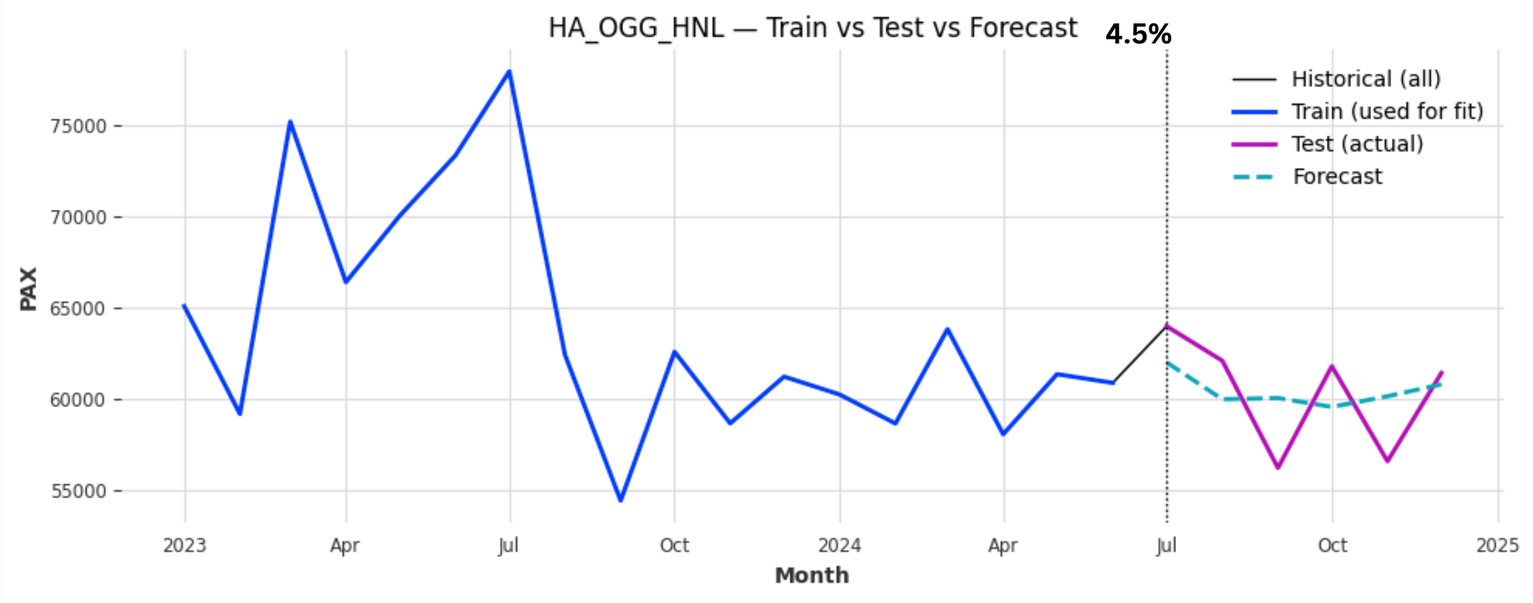
TCN (TEMPORAL CONVOLUTIONAL NETWORK)

Overall MAPE (zero-aware) ≈ 25.05% on 6-month horizon

not better than base line
no hyperparameter tuning.

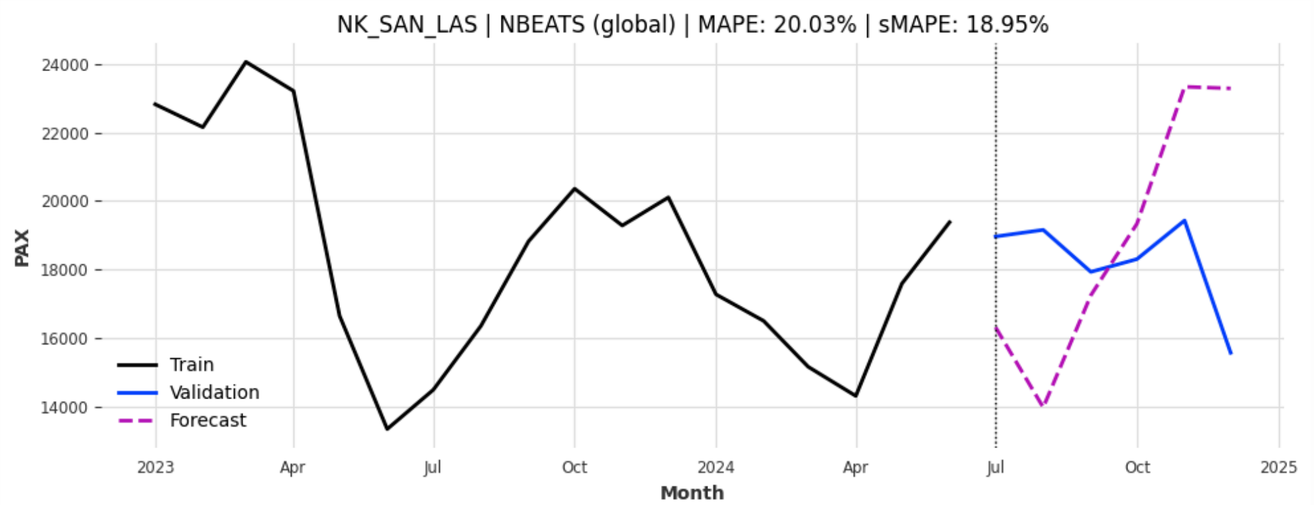
Target: PAX (pivoted to multi-route matrix).

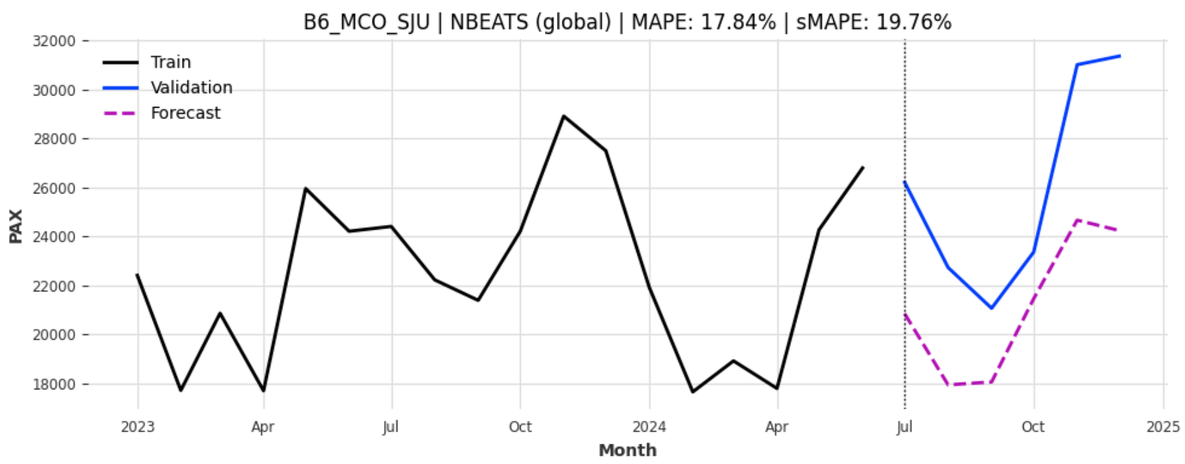
Past covariate: SEAT_FACTOR (no future and static covariates used).



N B E A T S

MAPE : 23 percent static and past covariance being used (future not allowed)





LIGHT GBM MODEL

MAPE:18.6 %

Used static and past and future covariate

```
model = LightGBMModel(
    lags=12,
    # use last 12 months of past covariates; your earlier version used the full -1..-12
    lags_past_covariates=[-1, -2, -3, -4, -5, -6, -7, -8, -9, -10, -11, -12],
    # use inside-horizon future covariates (6-month block)
    lags_future_covariates=[0, 1, 2, 3, 4, 5],
    output_chunk_length=6,
    random_state=42,
    # slight regularization to handle many sparse columns
    num_leaves=60, max_depth=10, min_data_in_leaf=5, max_bin=512,
    learning_rate=0.05, n_estimators=300
)
```

FEATURE ENGINEER

Future covariates + seats

Past covariates : seat factor

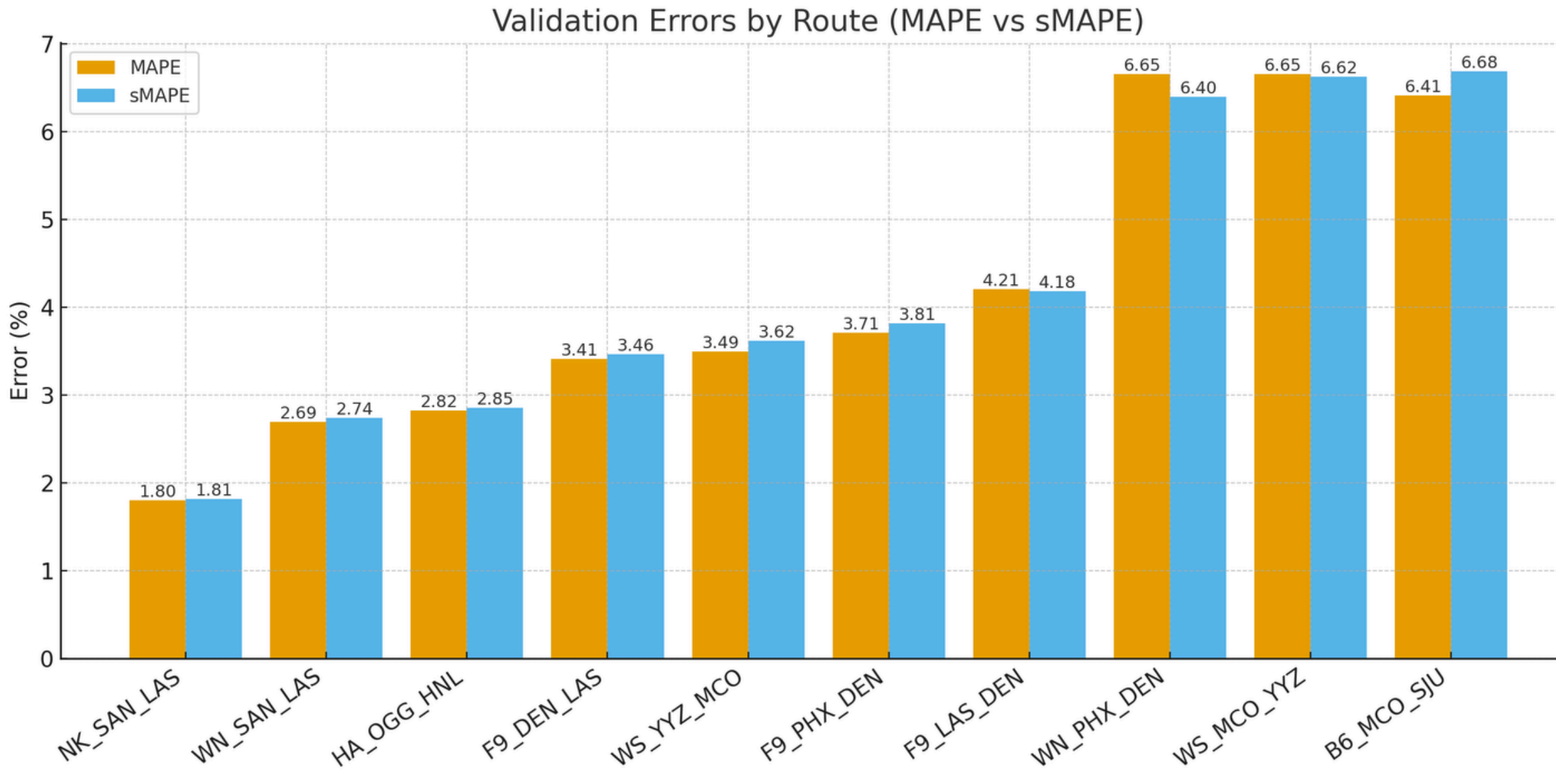
static : airline and origin / destination

	wkfirst_0	wkfirst_1	wkfirst_2	wkfirst_3	wkfirst_4	wkfirst_5	wkfirst_6	days_in_month	weekend_days	has_new_year	has_christmas	has_easter	holiday_any
2023-01-01	False	False	False	False	False	False	True	31	9	1	0	0	1
2023-02-01	False	False	True	False	False	False	False	28	8	0	0	0	0
2023-03-01	False	False	True	False	False	False	False	31	8	0	0	0	0
2023-04-01	False	False	False	False	False	True	False	30	10	0	0	1	1
2023-05-01	True	False	False	False	False	False	False	31	8	0	0	0	0
2023-06-01	False	False	False	True	False	False	False	30	8	0	0	0	0
2023-07-01	False	False	False	False	False	True	False	31	10	0	0	0	0
2023-08-01	False	True	False	False	False	False	False	31	8	0	0	0	0
2023-09-01	False	False	False	False	True	False	False	30	9	0	0	0	0
2023-10-01	False	False	False	False	False	False	True	31	9	0	0	0	0
2023-11-01	False	False	True	False	False	False	False	30	8	0	0	0	0
2023-12-01	False	False	False	False	True	False	False	31	10	0	1	0	1
2024-01-01	True	False	False	False	False	False	False	31	8	1	0	0	1
2024-02-01	False	False	False	True	False	False	False	29	8	0	0	0	0
2024-03-01	False	False	False	False	True	False	False	31	10	0	0	1	1
2024-04-01	True	False	False	False	False	False	False	30	8	0	0	0	0
2024-05-01	False	False	True	False	False	False	False	31	8	0	0	0	0
2024-06-01	False	False	False	False	False	True	False	30	10	0	0	0	0
2024-07-01	True	False	False	False	False	False	False	31	8	0	0	0	0

RESULT FOR LIGHT GBM

MAPE avg : 14.2 %
after hypertunning

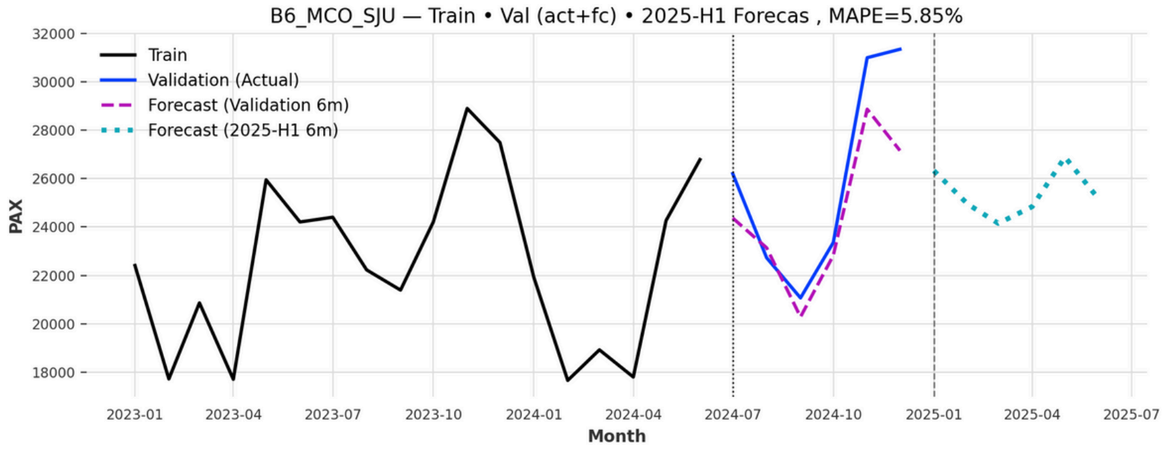
Parameter	Value
lags	6
lags_past_covariates	[-1]
lags_future_covariates	[0, 1, 2, 3, 4, 5]
n_estimators	300
learning_rate	0.02
num_leaves	31
max_depth	8
min_data_in_leaf	5
feature_fraction	1
bagging_fraction	0.6
bagging_freq	1

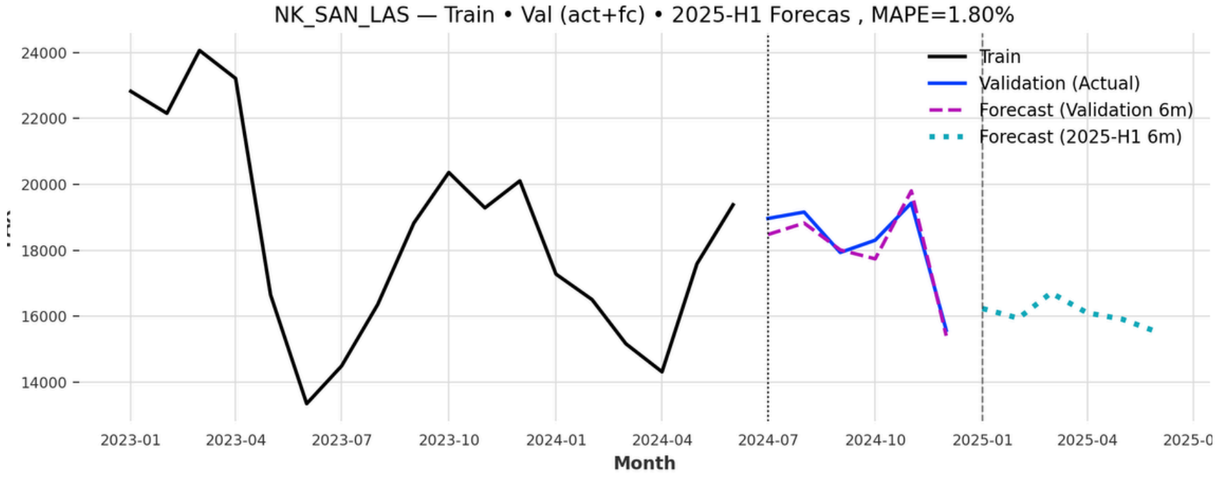


RESULT FOR LIGHT GBM

MAPE avg : 14.2 %

After hypertunning





FUTURE WORK

1. **Adding weather data as past covariate and weather prediction as future variable for each route https://openweathermap.org/**
2. **Ensembles & stacking Blend LightGBM + NBEATS.**
3. **national breaks by country pair.**
4. **More advance method for missing value**

REFERENCES

YouTube Playlist: Time Series Forecasting with Darts
🔗 https://www.youtube.com/watch?v=cKJKHGlbSi4&list=PLxzpq1ouJ5mchBJPUmbIxB7CNoX4T4ZKz

Darts Library
– Unit8 SA. “Darts: Time Series Made Easy in Python.” [Online]. Available: https://unit8co.github.io/darts/

THANK YOU

Omid Ghorbani
Agust 2025