### 1. Data Processing & Modelling

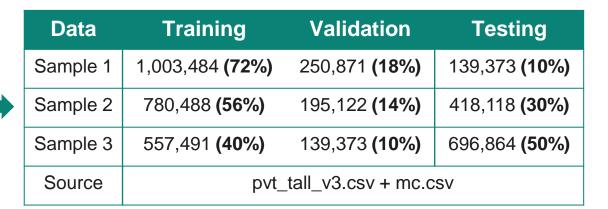
• Combine 2 datasets (in total 1,393,728)

 Randomly select training and testing data at 3 different split sample ratios (9:1, 7:3 and 5:5)



Drop N.A

Minmax Scaler: ['freq','p"]



Training

Validation



Testing

#### **Models**

- Poly Linear Regression
- Random Forest
- XGBoost
- LightGBM

\* Removed SVM due to efficiency and accuracy.

#### **Methods**

- Cross-Validation
- Grid / Random Search
- Early-stopping
- Regularization
- Learning Curve

	Metrics	
MSE	MAE	AIC
RMSE	MAPE	BIC
	R2	

1

# 2. Model Comparison – 9:1 ratio

Train							Test							
Metrics	MSE	RMSE	MAE	MAPE	R2	AIC	BIC	MSE	RMSE	MAE	MAPE	R2	AIC	BIC
Poly LR	0.0016	0.0404	0.0213	7.11E+09	0.9600	-4487257	-4484607	0.0016	0.0404	0.0212	6.67E+09	0.9601	-498765	-496599
Random Forest	0.0029	0.0540	0.0407	3.00E+10	0.9288	-3763416	-3763091	0.0029	0.0539	0.0406	2.02E+10	0.9290	-418593	-418327
XGBoost	0.0006	0.0255	0.0164	6.06E+09	0.9841	-5647211	-5646477	0.0006	0.0254	0.0164	4.47E+09	0.9842	-627865	-627264
LightGBM	0.0003	0.0186	0.0111	1.78E+09	0.9915	-6432913	-6431732	0.0003	0.0186	0.0112	1.68E+09	0.9915	-714577	-713612

Data	Training	Training Validation							
Sample 1	1,003,484 <b>(72%)</b>	250,871 <b>(18%)</b>	139,373 <b>(10%)</b>						
Source	pvt_tall_v3.csv + mc.csv								

### **Conclusion:**

- XGBoost and LightGBM have better performance.
- Poly LR and Random Forest also have good performance.
- MAPE, AIC, and BIC are huge, possibly due to the simulation of the data.

# 2. Model Comparison – 7:3 ratio

	Train							Test						
Metrics	MSE	RMSE	MAE	MAPE	R2	AIC	BIC	MSE	RMSE	MAE	MAPE	R2	AIC	BIC
Poly LR	0.0016	0.0406	0.0214	6.88E+09	0.9597	-3482031	-3479437	0.0016	0.0406	0.0213	7.19E+09	0.9598	-1492671	-1490263
Random Forest	0.0022	0.0464	0.0351	2.55E+10	0.9472	-3220519	-3220083	0.0022	0.0464	0.0351	2.42E+10	0.9474	-1380714	-1380309
XGBoost	0.0007	0.0265	0.0182	3.33E+09	0.9828	-4315439	-4314436	0.0007	0.0266	0.0183	4.96E+09	0.9828	-1847649	-1846719
LightGBM	0.0005	0.0218	0.0130	2.07E+09	0.9884	-4697685	-4696907	0.0005	0.0218	0.0130	2.82E+09	0.9884	-2012256	-2011534

Data	Training	Validation	Testing						
Sample 2	780,488 <b>(56%)</b>	195,122 <b>(14%)</b>	418,118 <b>(30%)</b>						
Source	pvt_tall_v3.csv + mc.csv								

### **Conclusion:**

- XGBoost and LightGBM have better performance.
- Poly LR and Random Forest also have good performance.
- MAPE, AIC, and BIC are huge, possibly due to the simulation of the data.

# 2. Model Comparison – 5:5 ratio

	Train							Test						
Metrics	MSE	RMSE	MAE	MAPE	R2	AIC	BIC	MSE	RMSE	MAE	MAPE	R2	AIC	BIC
Poly LR	0.0017	0.0408	0.0215	7.13E+09	0.9592	-2478971	-2476451	0.0017	0.0408	0.0214	6.68E+09	0.9594	-2481145	-2478625
Random Forest	0.0020	0.0451	0.0329	2.20E+10	0.9502	-2340484	-2340106	0.0020	0.0452	0.0329	2.11E+10	0.9501	-2338432	-2338054
XGBoost	0.0006	0.0236	0.0152	3.05E+09	0.9863	-3241777	-3240815	0.0006	0.0237	0.0153	3.81E+09	0.9862	-3235966	-3235004
LightGBM	0.0012	0.0343	0.0234	9.23E+09	0.9712	-2723613	-2722949	0.0012	0.0344	0.0235	8.44E+09	0.9710	-2716724	-2716060

Data	Training	Training Validation							
Sample 3	557,491 <b>(40%)</b>	139,373 <b>(10%)</b>	696,864 <b>(50%)</b>						
Source	pvt_tall_v3.csv + mc.csv								

### **Conclusion:**

- XGBoost and LightGBM have better performance.
- Poly LR and Random Forest also have good performance.
- MAPE, AIC, and BIC are huge, possibly due to the simulation of the data.