1. E-commerse Sales Data Analysis
   Analysis using public datasets from:
   https://www.kaggle.com/datasets/carrie1/ecommerce-data
   Dataset preview:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

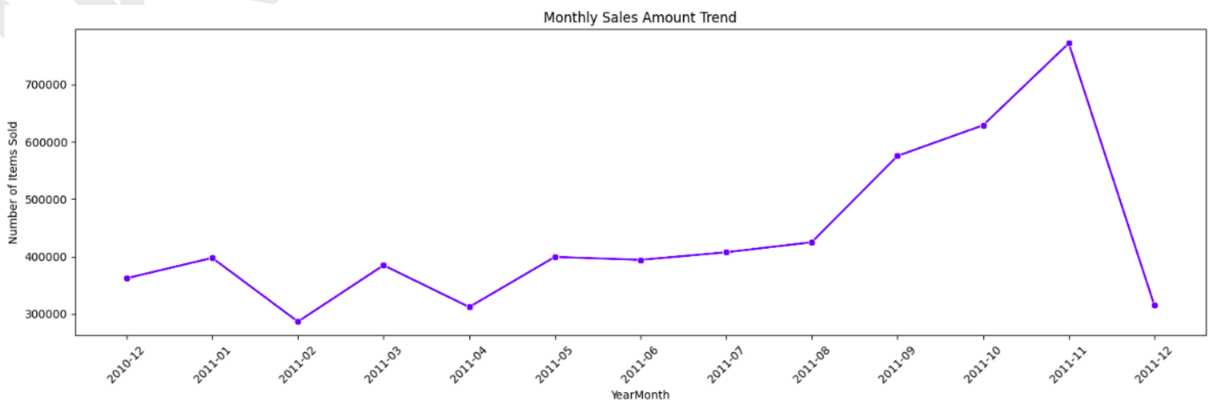
   Descriptive statistics:

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

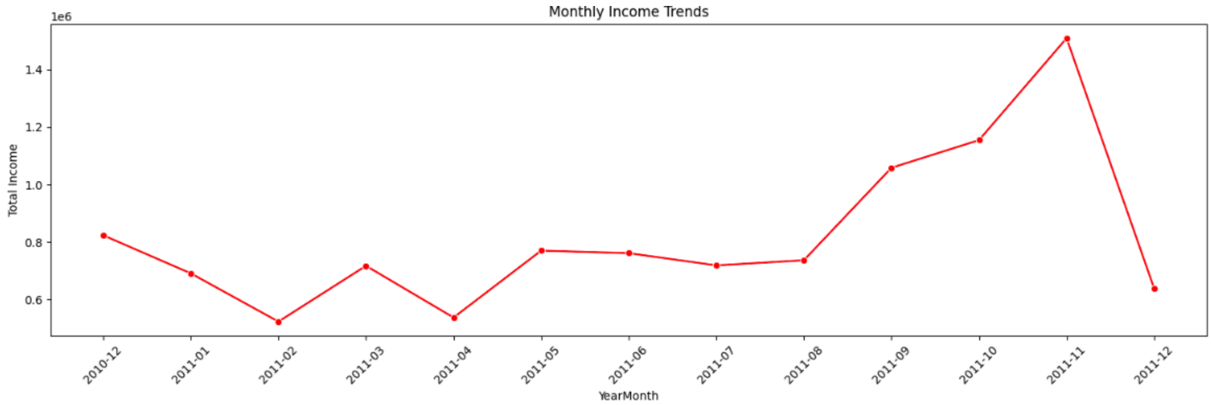
   Data after cleaning:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalPrice	YearMonth	DayOfWeek	Hour
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30	2010-12	Wednesday	8
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	2010-12	Wednesday	8
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00	2010-12	Wednesday	8
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	2010-12	Wednesday	8
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	2010-12	Wednesday	8

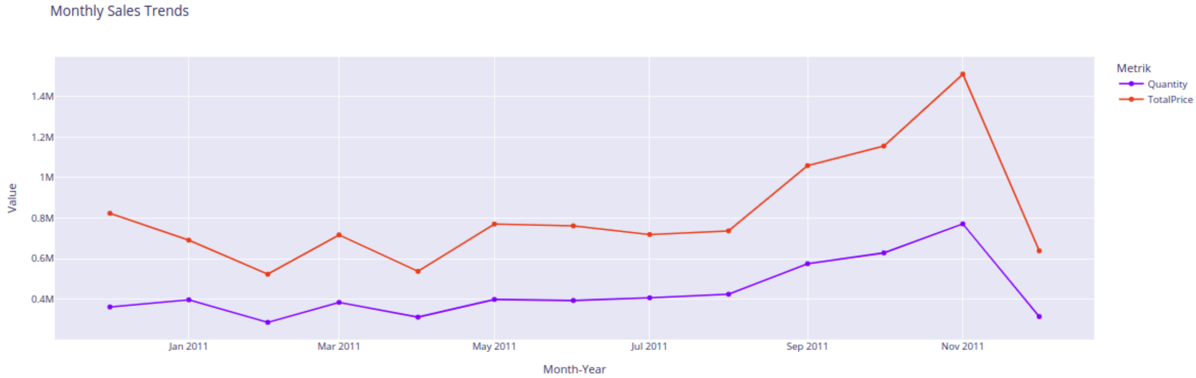
   Plot quantity:



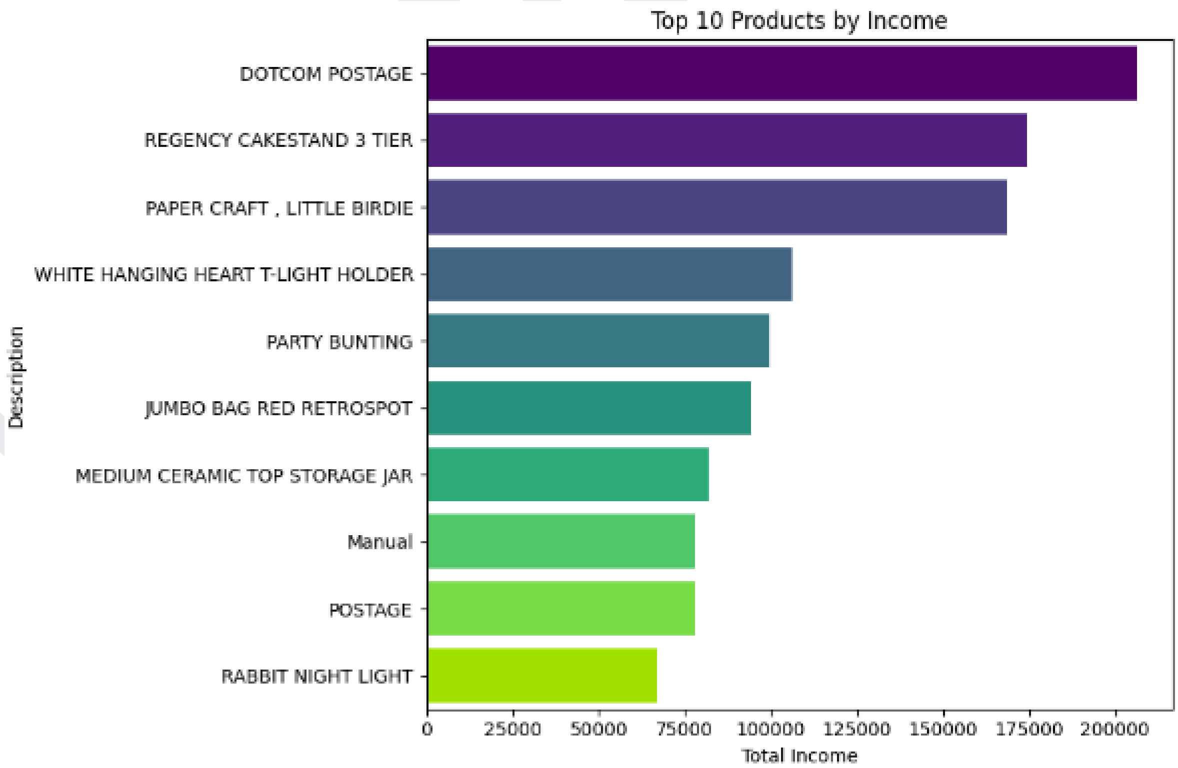
   Plot revenue:

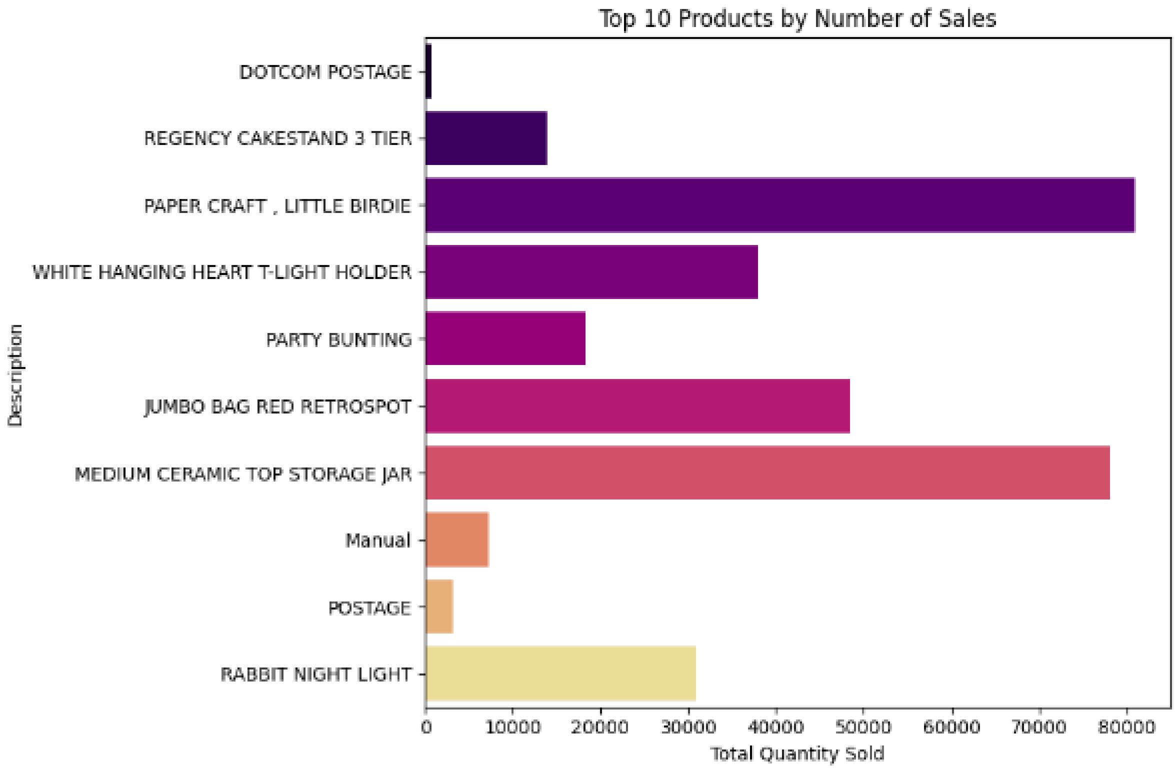


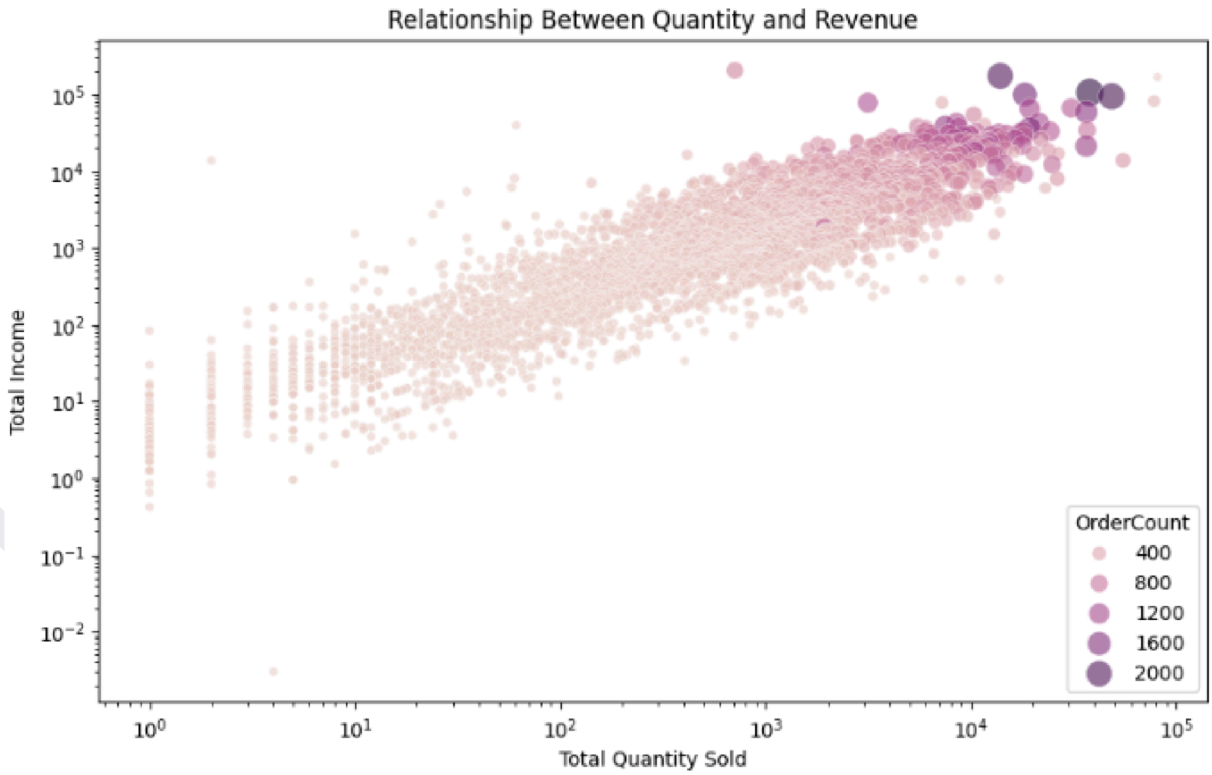
Interactive visualization with plotly:



Product Analysis:

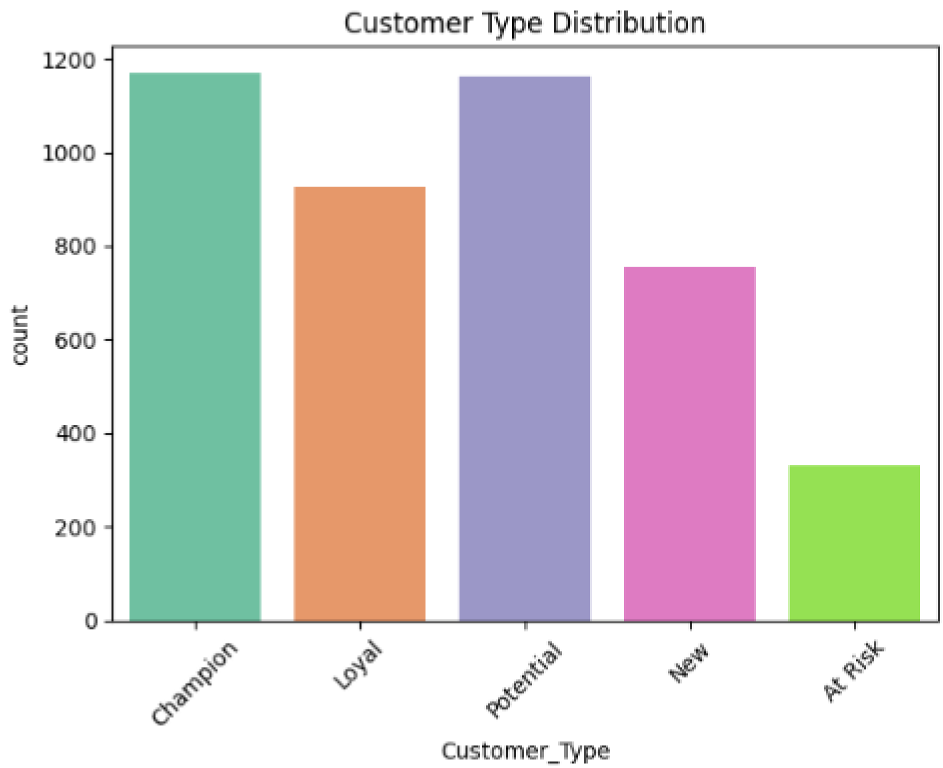


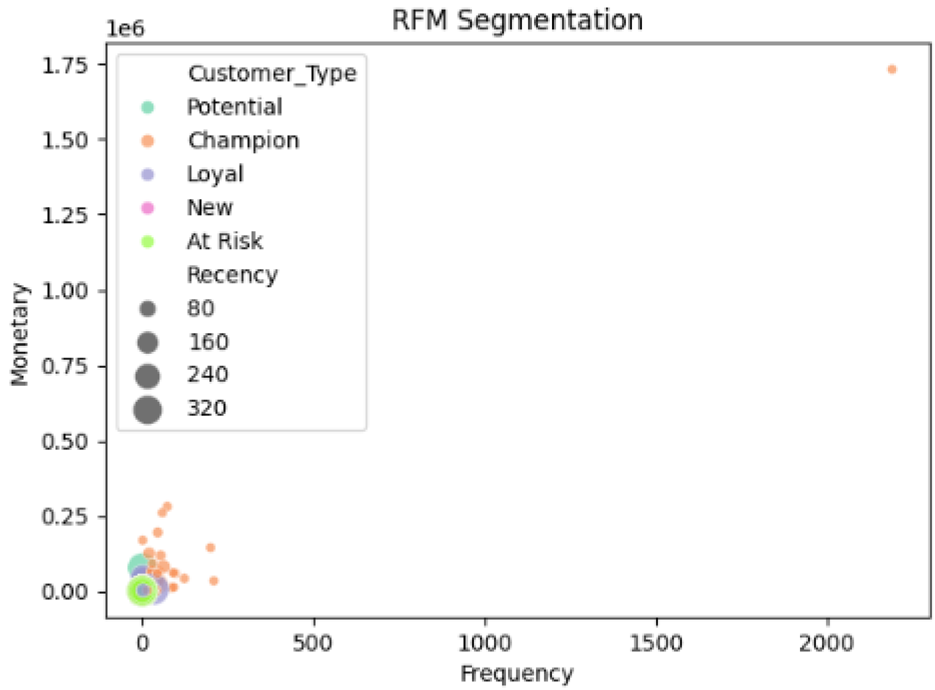




RFM (Recency, Frequency, Monetary value) analysis:

	CustomerID	Recency	Frequency	Monetary	R_Score	F_Score	M_Score	RFM_Score	RFM_Group	Customer_Type
0	12346.0	326	1	77183.60	1	1	5	7	115	Potential
1	12347.0	2	7	4310.00	5	5	5	15	555	Champion
2	12348.0	75	4	1797.24	2	4	4	10	244	Loyal
3	12349.0	19	1	1757.55	4	1	4	9	414	Loyal
4	12350.0	310	1	334.40	1	1	2	4	112	New





- Strategies to encourage new / risk customers to become potential: **Timing**: Intervention on day 3 for New Customers, week 2 for At Risk

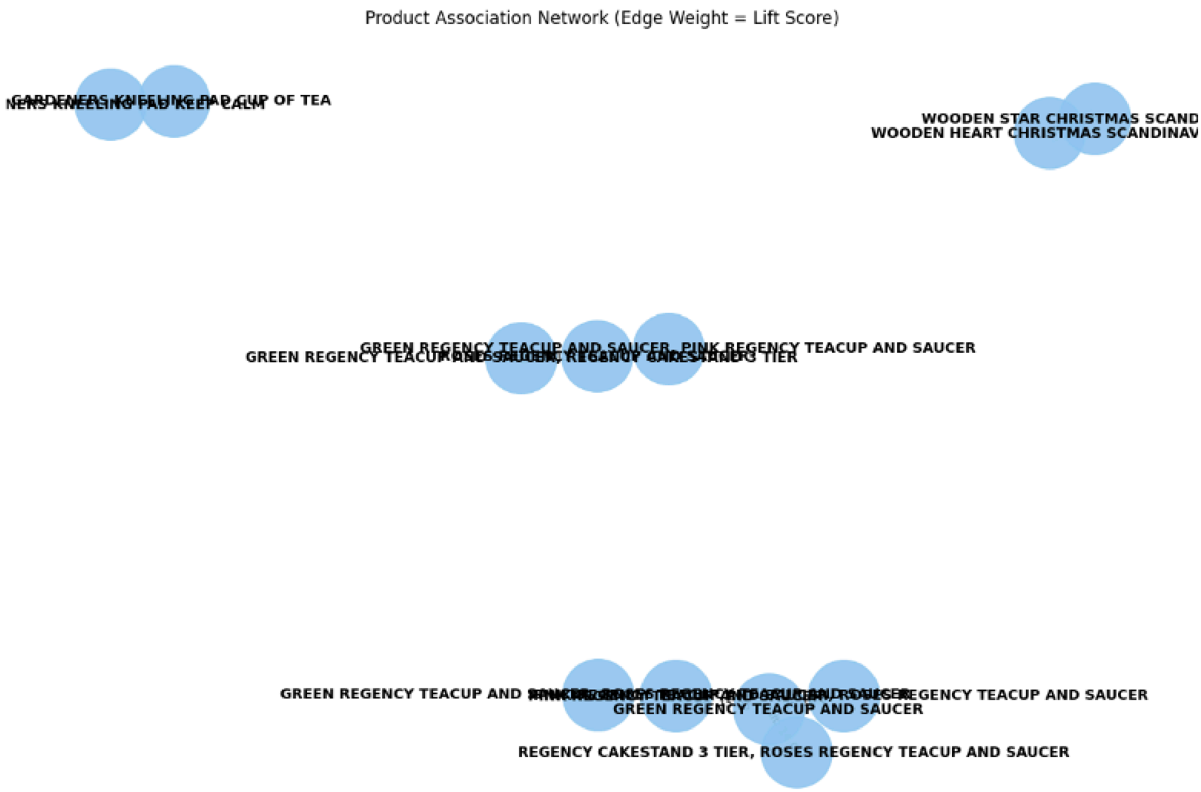
Personalization: Use shopping history for product recommendations

Urgency: Use limited-time offers with clear deadlines

Multi-Channel: Combination of email, push notifications, and SMS

This implementation will increase conversions by 20-35% based on digital retail industry benchmarks.

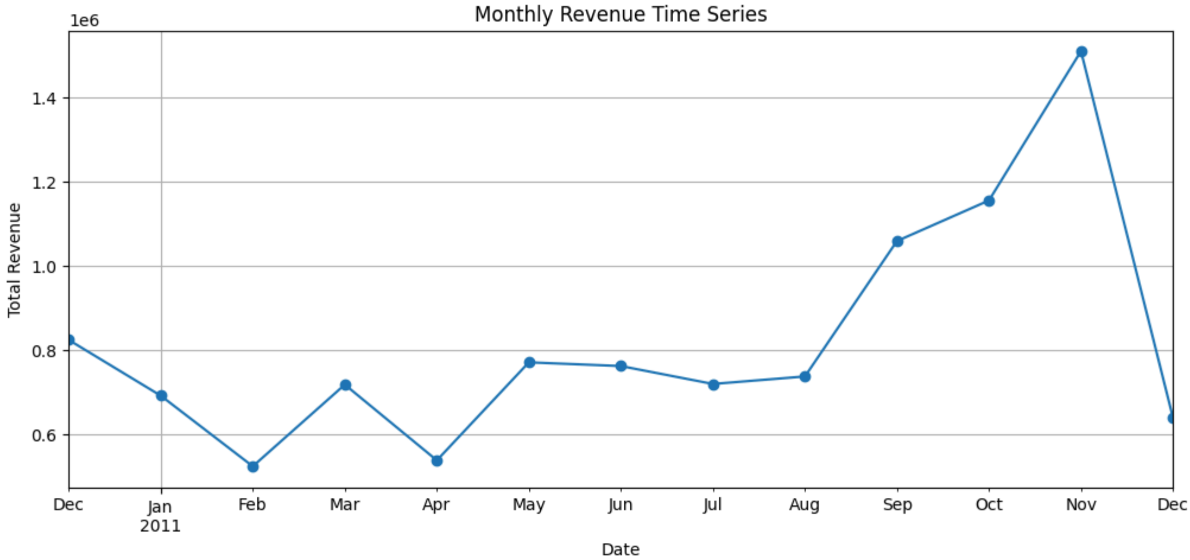
Basket analysis:



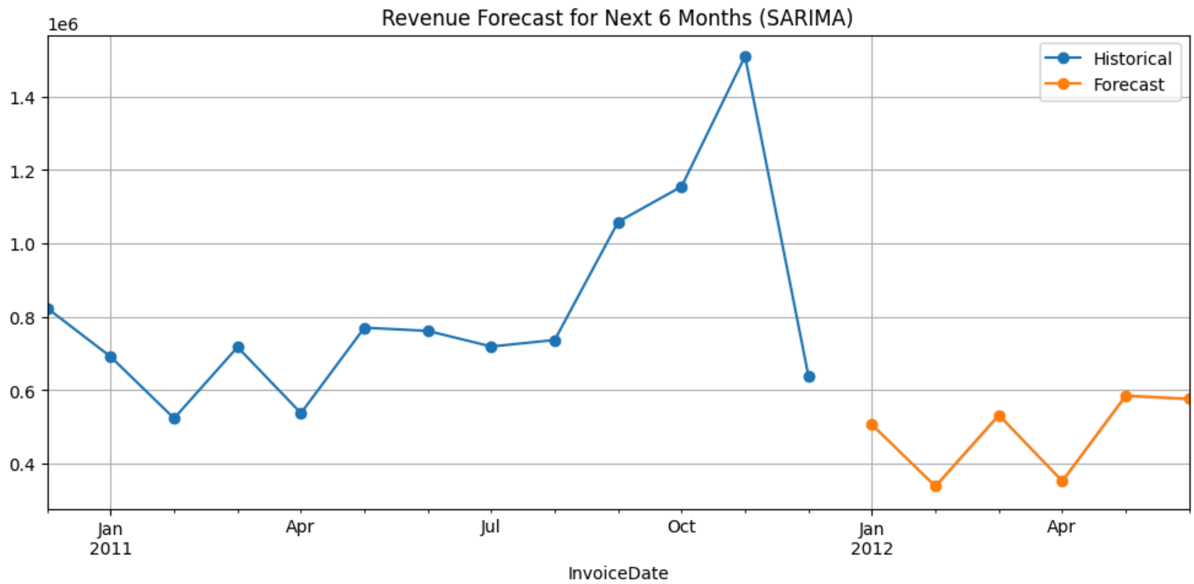
From the market basket analysis, it can be concluded that there is a relationship between several goods, namely:

- Gardeners kneeling pad cup of tea. Gardeners kneeling pad keep calm
- Wooden heart christmas scandinavian. Wooden star christmas scandinavian
- Pink regency teacup and saucer. Green regency teacup and saucer. Green regency teacup and saucer Regency cakestand 3 tier
- Green regency and saucer, roses regency teacup and saucer. Pink regency teacup and saucer. Green regency teacup and saucer. Regency cakestand 3 tier, roses regency teacup and saucer. Pink regency teacup and saucer, roses, regency teacup and saucer

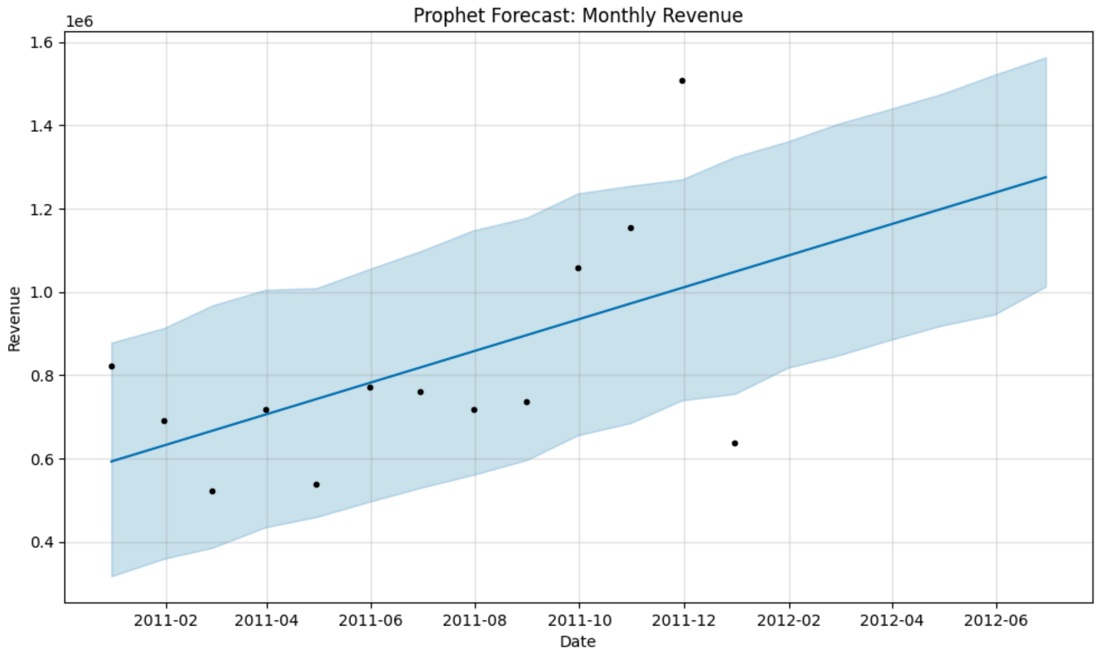
Time series forecasting:
Prepare Monthly Time Series Data



Forecasting with SARIMA (without additional installation)



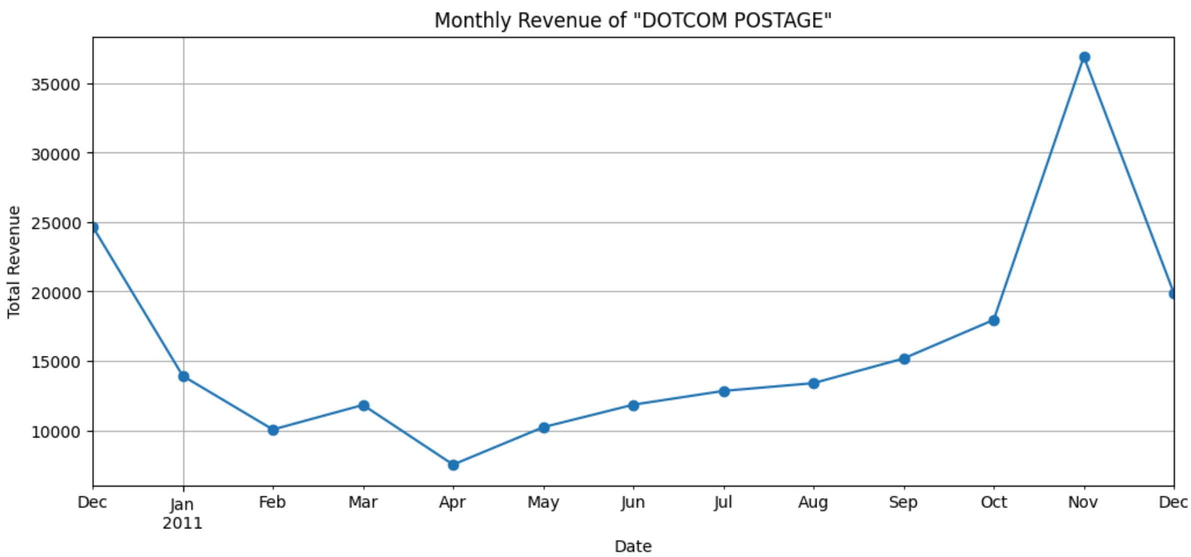
(Alternative) Forecasting with Prophet

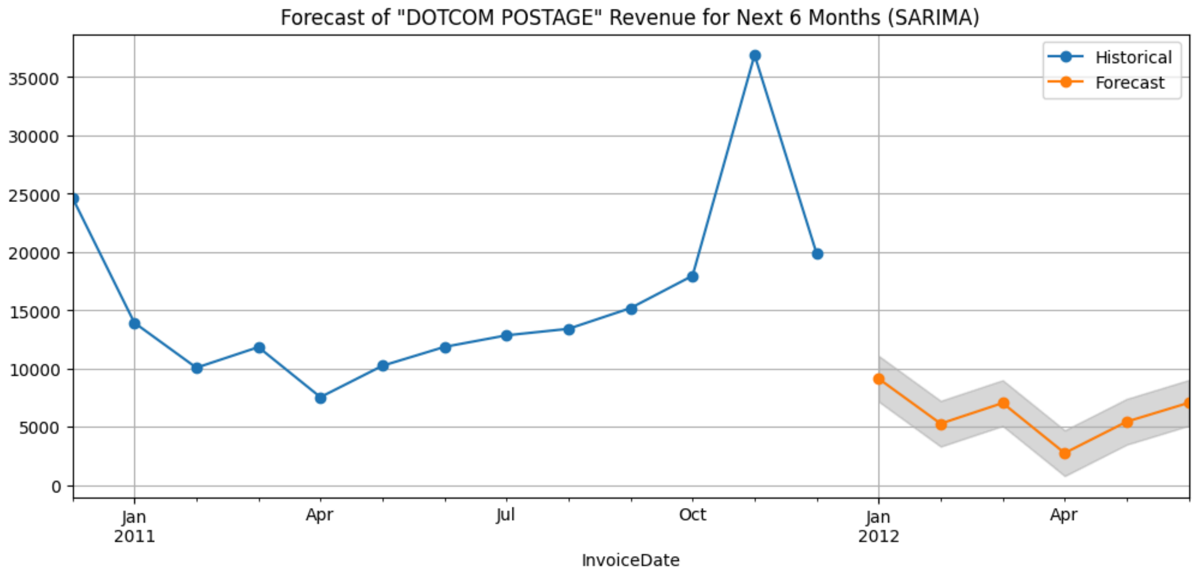


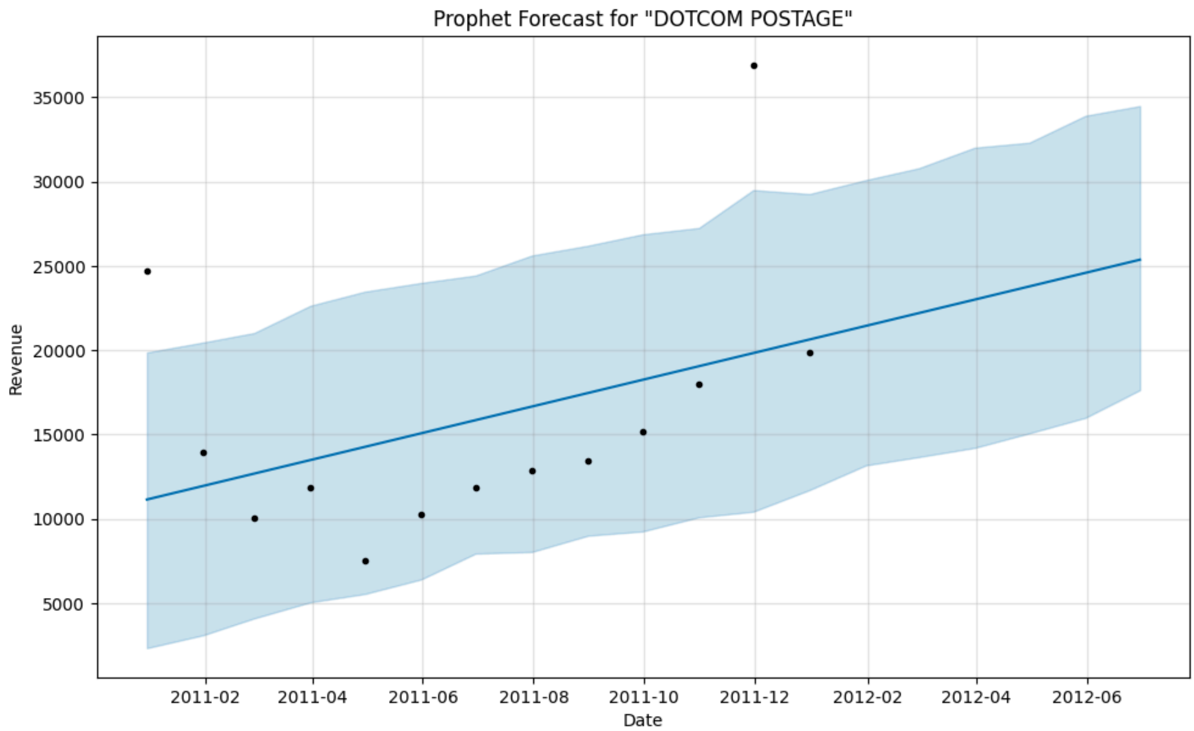
Time Series for Specific Products [DOTCOM POSTAGE]

Top 5 Products Based on Revenue:

Description	TotalPrice
DOTCOM POSTAGE	206248.77
REGENCY CAKESTAND 3 TIER	174484.74
PAPER CRAFT , LITTLE BIRDIE	168469.60
WHITE HANGING HEART T-LIGHT HOLDER	106292.77
PARTY BUNTING	99504.33




Forecast of "DOTCOM POSTAGE" Revenue for Next 6 Months (SARIMA)
Historical
Forecast
35000
30000
25000
20000
15000
10000
5000
0
Jan
2011
Apr
Jul
Oct
Jan
2012
Apr
InvoiceDate


Prophet Forecast for "DOTCOM POSTAGE"
Revenue
35000
30000
25000
20000
15000
10000
5000
2011-02
2011-04
2011-06
2011-08
2011-10
2011-12
2012-02
2012-04
2012-06
Date

2. Fraud detection

Analysis using public datasets from:

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

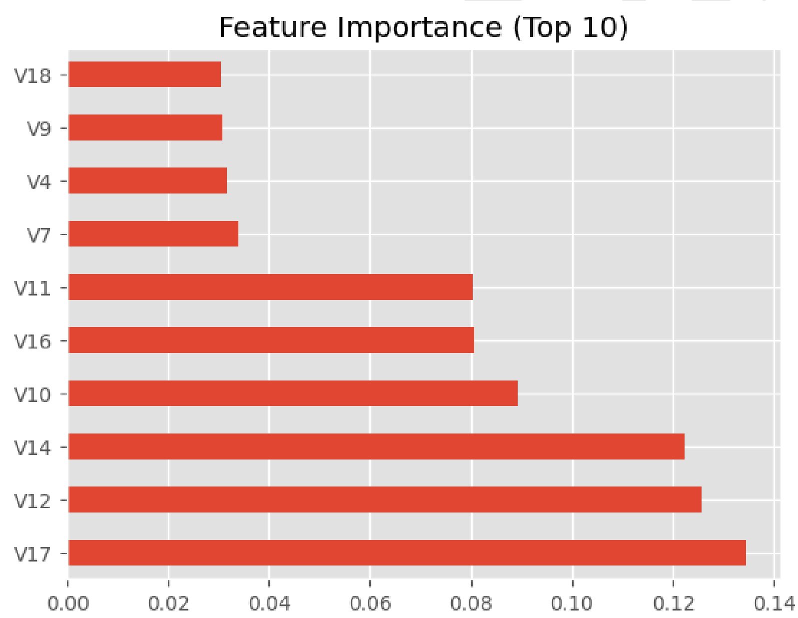
Dataset preview:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	123.50	0
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0

Modelling: Random forest

	precision	recall	f1-score	support
0	1.00	1.00	1.00	56864
1	0.94	0.83	0.88	98
accuracy			1.00	56962
macro avg	0.97	0.91	0.94	56962
weighted avg	1.00	1.00	1.00	56962

Feature Importance Visualization:



1. Most Influential Features:

   V17, V12, and V14 are the most important features in fraud detection, each contributing the highest to the model's predictions.

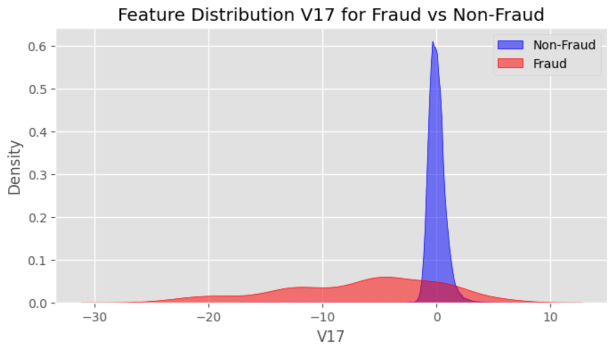
   This means that the values of these variables statistically best differentiate fraudulent and non-fraudulent transactions.

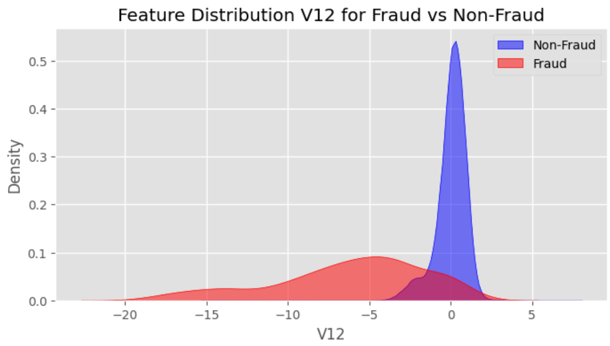
2. Low-Influence Features:

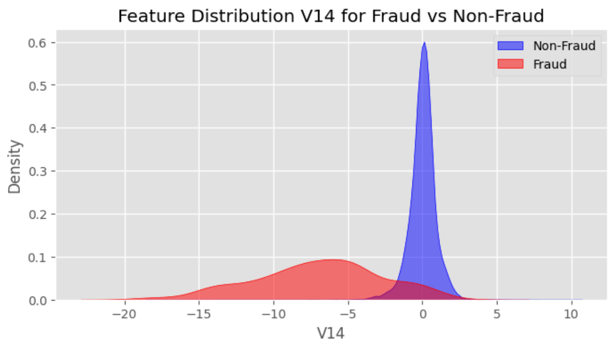
   Features such as V7, V4, V9, and V18 have relatively low importance values among the top 10, but are still more significant than other features that do not appear in the graph.

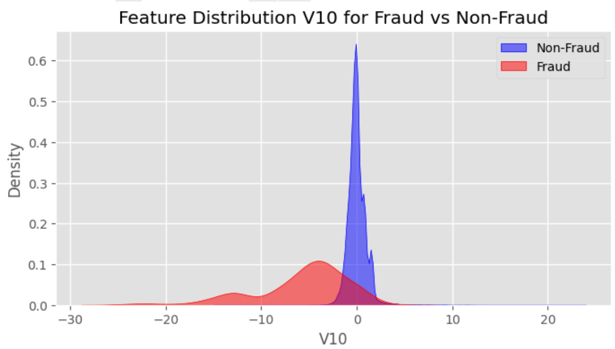
   This means that their contribution to the model's decisions is still significant, but not dominant.

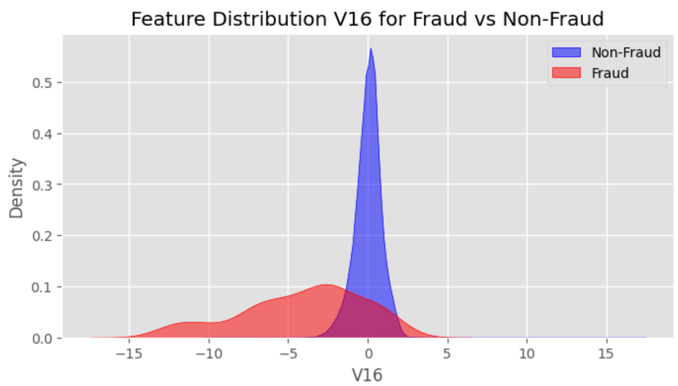
Visualization of the distribution of important features for fraud (1) vs non-fraud (0) classes:









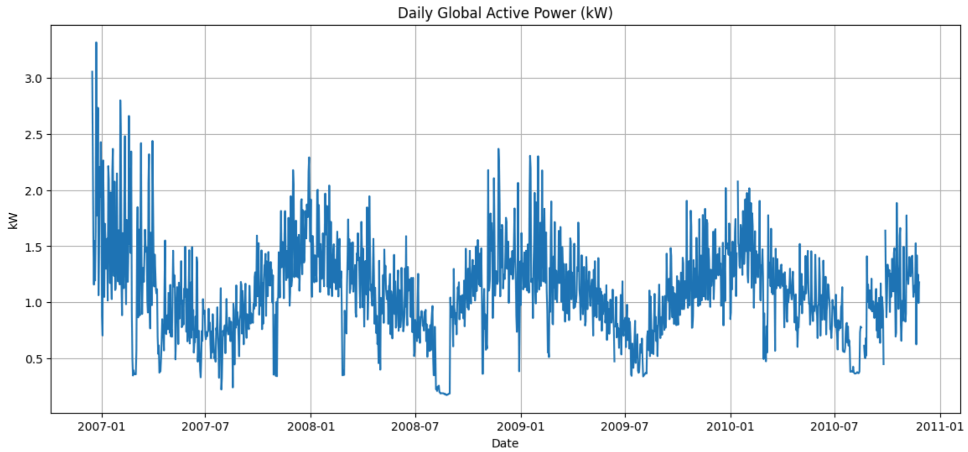


If the V17 plot shows that: Fraud values tend to be lower/extremely negative, while non-fraud values are spread out in the middle, Then the model can utilize this for early detection. A similar approach can be applied to other features such as V12, V14, etc.

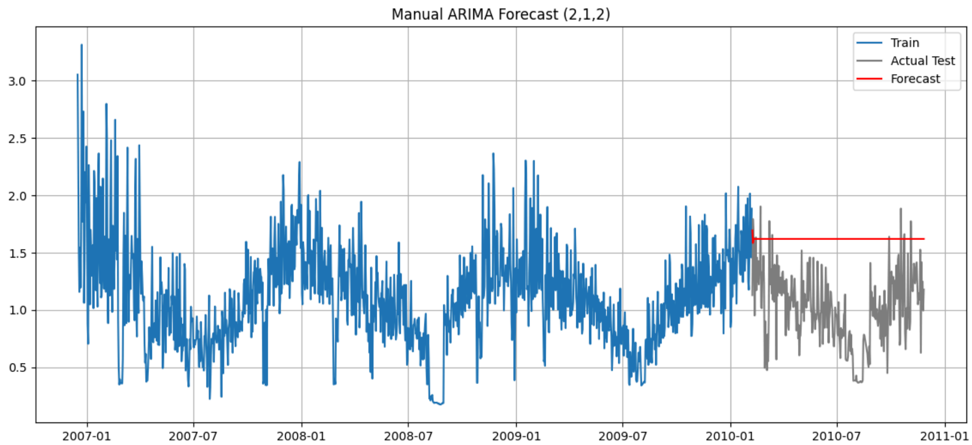
3. Energy consumption prediction
Analysis using public datasets from:
https://archive.ics.uci.edu/ml/machine-learning-databases/00235/household_power_consumption.zip
Dataset preview:

	datetime	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
0	2006-12-16 17:24:00	4.216	0.418	234.84	18.4	0.0	1.0	17.0
1	2006-12-16 17:25:00	5.360	0.436	233.63	23.0	0.0	1.0	16.0
2	2006-12-16 17:26:00	5.374	0.498	233.29	23.0	0.0	2.0	17.0
3	2006-12-16 17:27:00	5.388	0.502	233.74	23.0	0.0	1.0	17.0
4	2006-12-16 17:28:00	3.666	0.528	235.68	15.8	0.0	1.0	17.0

Daily energy consumption plot:

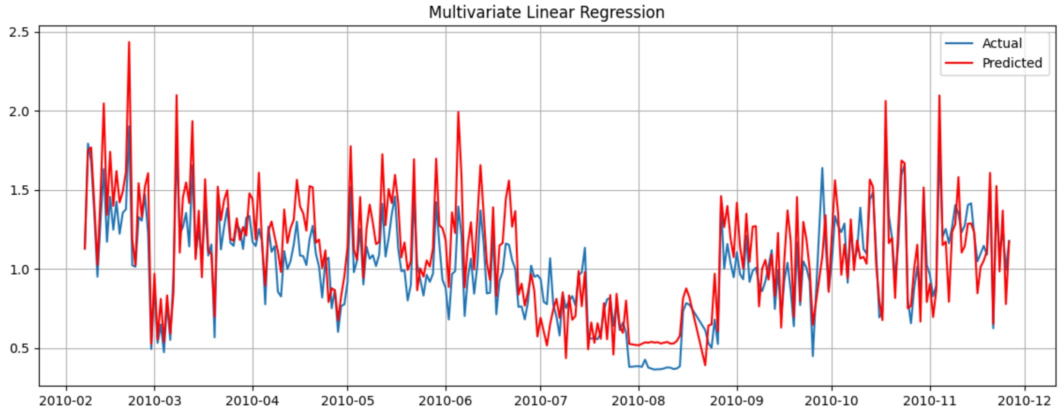


Manual ARIMA forecast: Resample daily data.



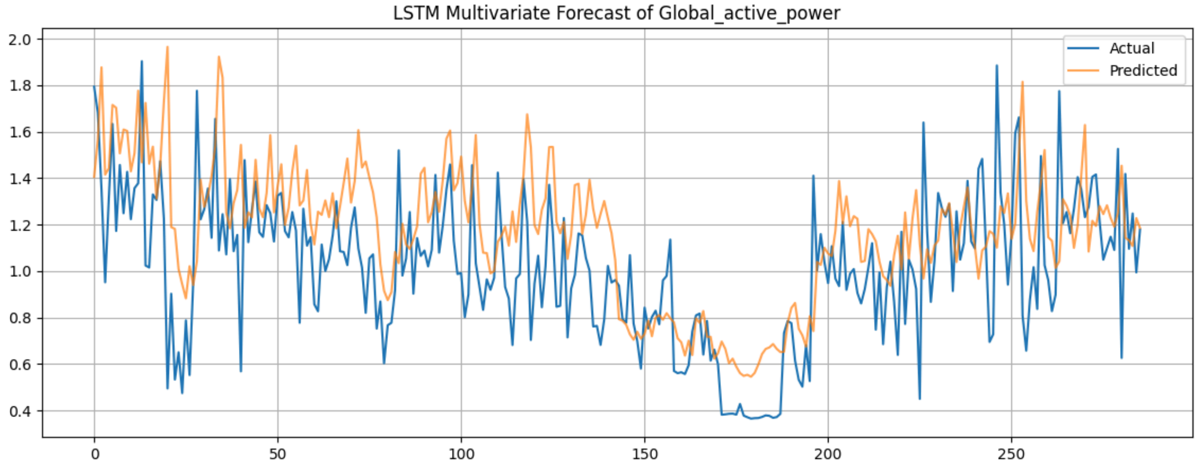
ARIMA predictions are very flat. ARIMA follows the final level of the training data. The model fails to capture the variability of the test set.

Multivariate model: adds additional features such as Voltage and Sub_metering.



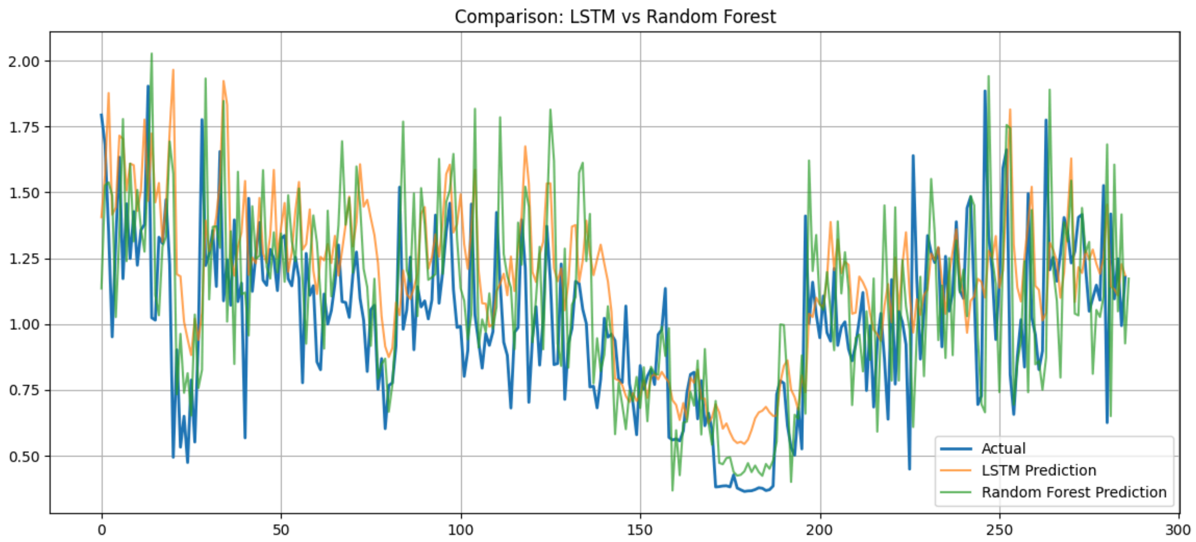
The model is able to follow general trends. Overfitting to extreme values. Systematic errors in certain periods. Lack of temporal context.

Multivariate LSTM to predict Global_active_power based on: Voltage, Global_reactive_power, Sub_metering_1, Sub_metering_2, Sub_metering_3.



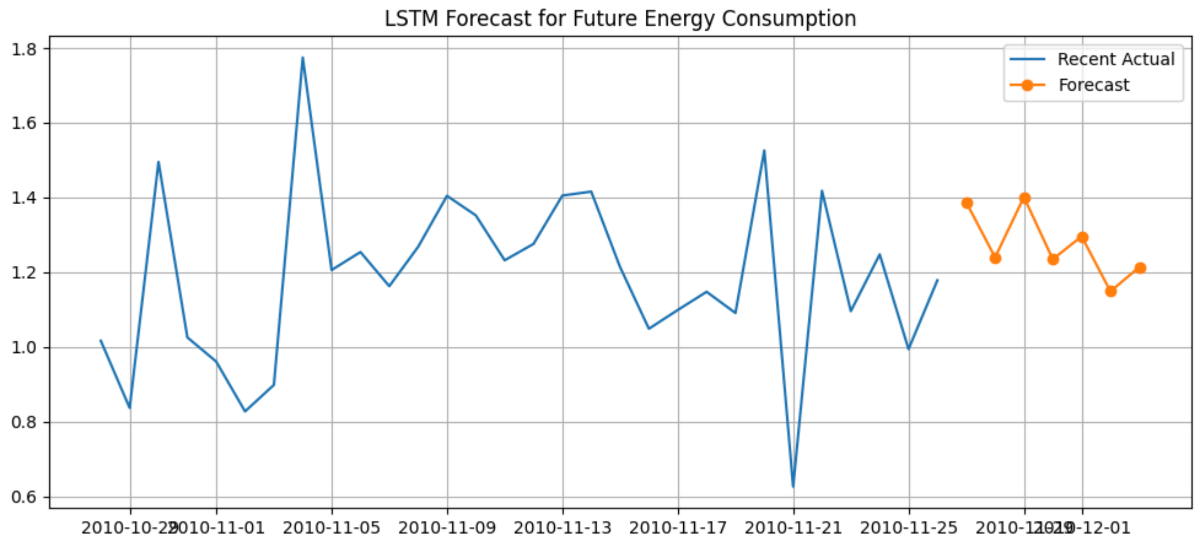
LSTM is able to follow the general pattern of data trends. It struggles to capture extreme spikes. Underpredictions occur at the bottom (low consumption). The model shows stability, but is somewhat conservative.

Comparison of LSTM and Random Forest performance for predicting Global_active_power with the same feature inputs: Voltage, Global_reactive_power, Sub_metering_1, Sub_metering_2, Sub_metering_3.



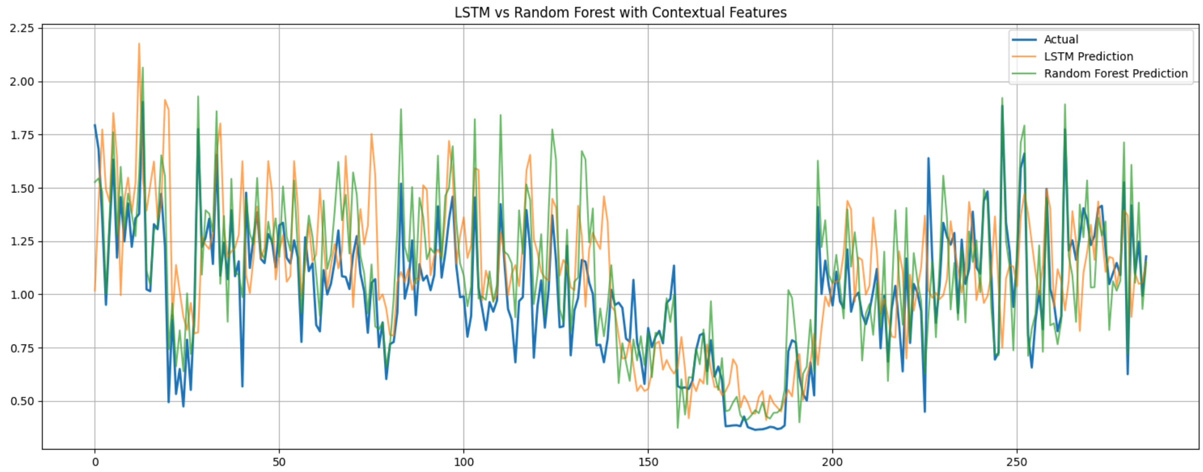
The LSTM's medium-term trend is good, while the Random Forest's tends to be unstable. The LSTM's response to spikes is slow, while the Random Forest's is responsive. The LSTM's predictions are very smooth, while the Random Forest's are sharp and sometimes overfit. The LSTM's fit to the time series is excellent, while the Random Forest's is poor.

Predict future values for the next few days (extrapolation) based on recent data with a trained LSTM model.



Prediction patterns follow historical trends. Fluctuations are more stable than actual data. Autoregression-based predictions. Potential for over/underestimation.

Comparison of Random Forest & LSTM using additional contextual features: Voltage, Global_reactive_power, Sub_metering_1/2/3, day_of_week, is_weekend, is_holiday, temp_avg.



Both models followed the general trend quite well. Random Forest was more volatile (responsive) but noisy. LSTM was more stable but somewhat slower to respond to drastic changes. Both still had gaps with the actual data.

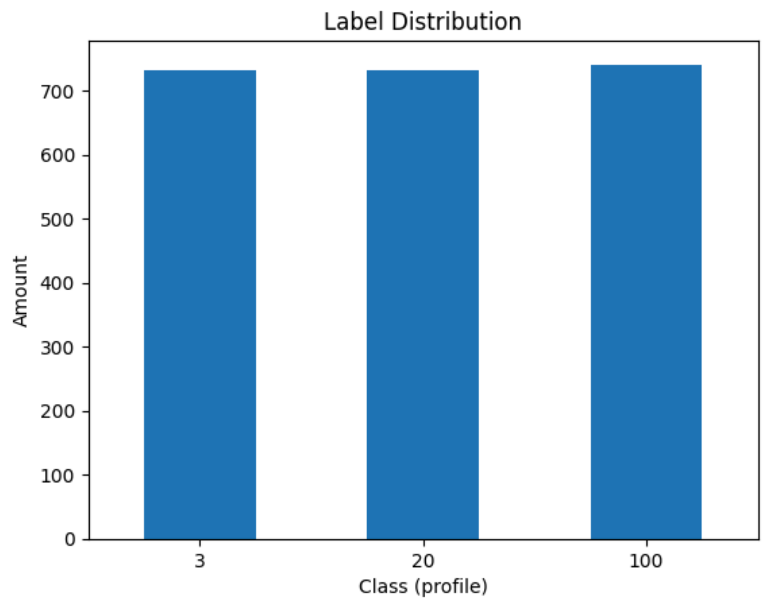
4. Condition Monitoring of Hydraulic Systems
Analysis using public datasets from:
https://www.kaggle.com/datasets/jjacostupa/condition-monitoring-of-hydraulic-systems

Dataset preview:

	CE_0	CE_1	CE_2	CE_3	CE_4	CE_5	CE_6	CE_7	CE_8	CE_9	...	VS1_56	VS1_57	VS1_58	VS1_59	profile_0
0	47.202	47.273	47.250	47.332	47.213	47.372	47.273	47.438	46.691	46.599	...	0.544	0.545	0.535	0.543	3
1	29.208	28.822	28.805	28.922	28.591	28.643	28.216	27.812	27.514	27.481	...	0.540	0.533	0.531	0.534	3
2	23.554	23.521	23.527	23.008	23.042	23.052	22.658	22.952	22.908	22.359	...	0.545	0.544	0.530	0.534	3
3	21.540	21.419	21.565	20.857	21.052	21.039	20.926	20.912	20.989	20.882	...	0.544	0.543	0.543	0.542	3
4	20.460	20.298	20.350	19.867	19.997	19.972	19.924	19.813	19.691	19.634	...	0.549	0.542	0.533	0.537	3

Each .txt file contains measurement data from different sensors. For example:
CE.txt: Cooler efficiency
CP.txt: Cooler power
EPS1.txt: Efficiency factor of pump 1
FS1.txt, FS2.txt: Flow sensors
PS1.txt to PS6.txt: Pressure sensors
SE.txt: System efficiency
TS1.txt to TS4.txt: Temperature sensors
VS1.txt: Valve sensor
description.txt: A brief description of the data and its use.
documentation.txt: A detailed description (metadata), including units, sensor position, and data meaning.
profile.txt: Labels/targets related to system conditions, often used to classify system conditions.

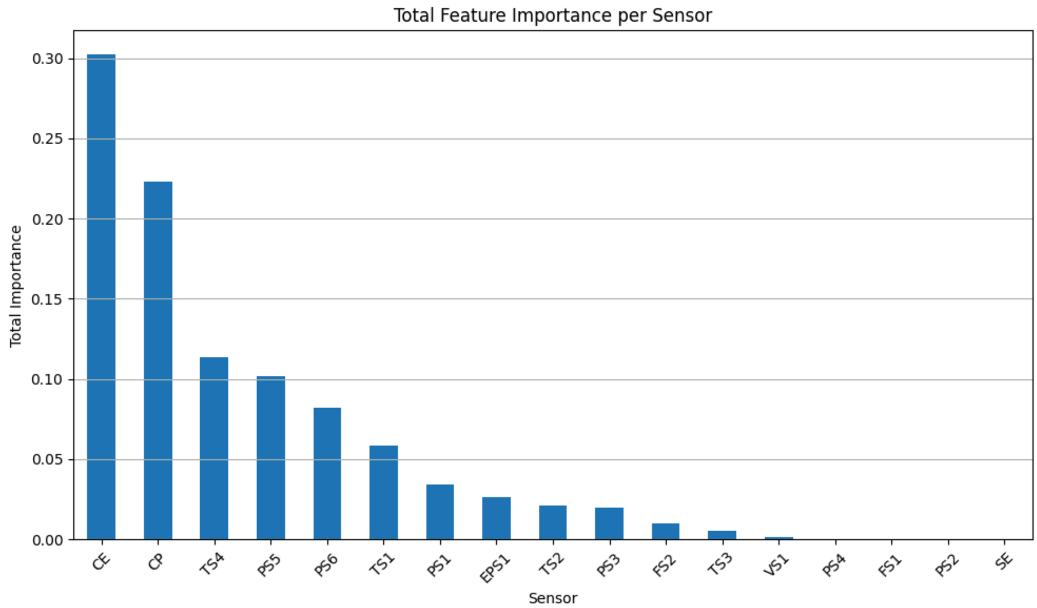
Classification Report (Random Forest):



```
=== Classification Report ===
              precision    recall  f1-score   support

           3       1.00      1.00      1.00       147
          20       1.00      1.00      1.00       146
         100       1.00      1.00      1.00       148

    accuracy                           1.00       441
   macro avg       1.00      1.00      1.00       441
weighted avg       1.00      1.00      1.00       441
```

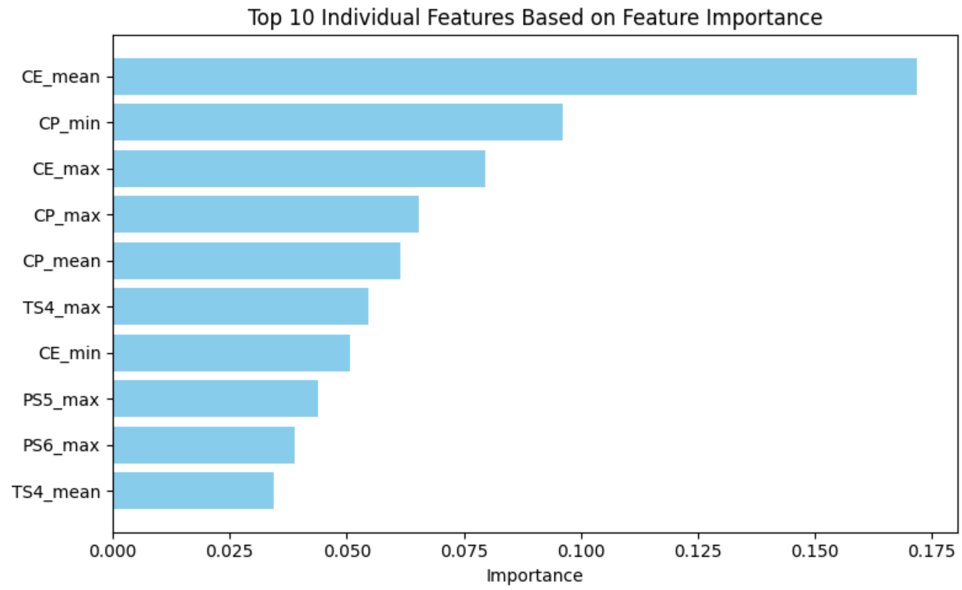
The dataset is very balanced and well-prepared. The Random Forest model can distinguish between the 3, 20, and 100 conditions very well. But the model is too perfect—additional validation is needed to ensure it doesn't overfit.

Visualization of Feature Importance per Sensor:



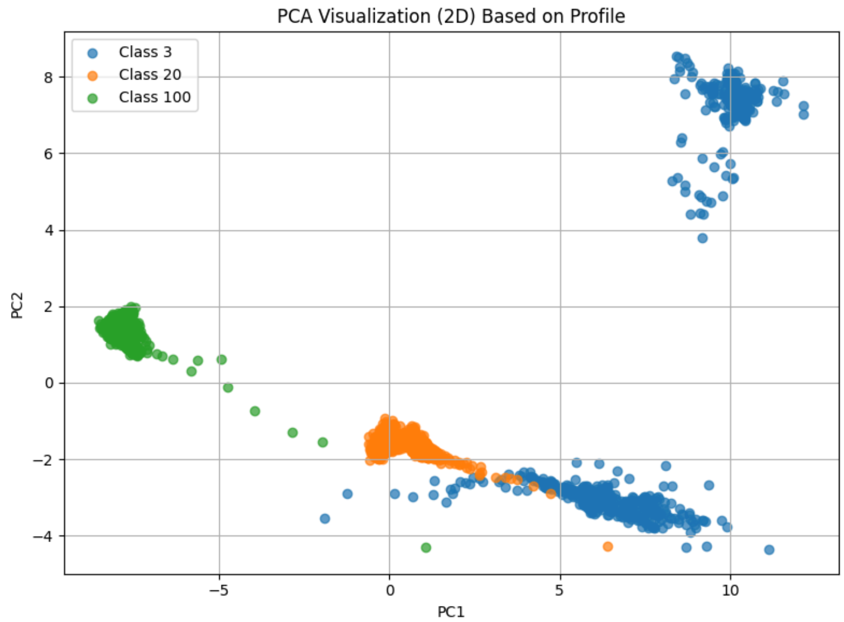
A bar chart showing which sensors have the greatest influence on the condition classification (profile). For example, if PS6 or EPS1 is dominant, it means that pump pressure or efficiency significantly impacts system conditions.

Top 10 Individual Features Based on Importance:



A list of 10 features, such as PS6_max, EPS1_mean, etc. We can see which specific statistical features and sensors are most dominant in label prediction. Helps identify the most informative sensors and statistics types.

2D PCA Visualization of Sensor Summary Data:



If the classes (3, 20, 100) appear separate or form clusters, this means the feature effectively differentiates system conditions. If they overlap, this could mean: The feature is not representative enough and A non-linear method is needed.

Compare the accuracy of the model with and without the pressure feature to determine: Is the pressure sensor critical? How much does accuracy decrease if pressure is removed?

```
Model accuracy with all features        : 1.0000
Accuracy of the model without pressure features (PS*) : 1.0000
```

If accuracy drops significantly without the pressure feature, it means the pressure sensor is critical for classification. If accuracy remains high, the system has other features (e.g., EPS1, SE, TS*) that are strong enough to distinguish classes.

How sensitive the model is to each pressure sensor (PS1–PS6). Comparison of accuracy stability through cross-validation. Relative performance of tree models (Random Forest) vs. linear models (SVM, Logistic Regression):

	Sensor_Dropped	Model	Accuracy_Mean	Accuracy_Std
2	None	Logistic Regression	0.998186	0.001697
0	None	Random Forest	0.998639	0.001111
1	None	SVM	0.998639	0.001111
5	PS1	Logistic Regression	0.998186	0.001697
3	PS1	Random Forest	0.998639	0.001111
4	PS1	SVM	0.998639	0.001111
8	PS2	Logistic Regression	0.998186	0.001697
6	PS2	Random Forest	0.998639	0.001111
7	PS2	SVM	0.998639	0.001111
11	PS3	Logistic Regression	0.998186	0.001697
9	PS3	Random Forest	0.998639	0.001111
10	PS3	SVM	0.998639	0.001111
14	PS4	Logistic Regression	0.998186	0.001697
12	PS4	Random Forest	0.998639	0.001111
13	PS4	SVM	0.998639	0.001111
17	PS5	Logistic Regression	0.998186	0.001697
15	PS5	Random Forest	0.998639	0.001111
16	PS5	SVM	0.998639	0.001111
20	PS6	Logistic Regression	0.998186	0.001697
18	PS6	Random Forest	0.998639	0.001111
19	PS6	SVM	0.998639	0.001111

See if accuracy drops drastically when a particular sensor is removed → that sensor is crucial. Check model stability using Accuracy_Std. Compare: are non-tree models (SVM, LR) more sensitive to the loss of pressure features than Random Forest?

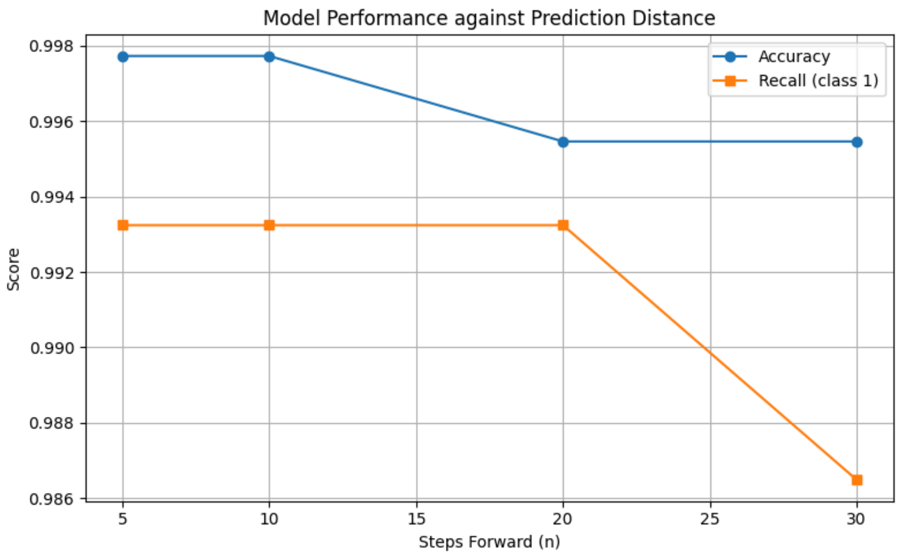
See the Most Important Features per Model:

```
Top 10 Most Important Features - Random Forest
    feature  importance
0   CE_mean    0.136198
3    CE_max    0.096303
2    CE_min    0.076144
4   CP_mean    0.072152
6    CP_min    0.070862
63  TS4_max    0.068526
7    CP_max    0.055263
43  PS6_max    0.051773
62  TS4_min    0.043609
39  PS5_max    0.040253

Top 10 Most Important Features - Logistic Regression
    feature      coef
49  TS1_std  0.512582
7    CP_max  0.486424
6    CP_min  0.461457
4   CP_mean  0.458941
53  TS2_std  0.382774
27  PS2_max  0.308741
2    CE_min  0.295878
0   CE_mean  0.294715
3    CE_max  0.294128
17  FS2_std  0.293218

Top 10 Most Important Features - SVM (Linear)
    feature      coef
7    CP_max  0.157290
4   CP_mean  0.148010
6    CP_min  0.146463
49  TS1_std  0.092772
2    CE_min  0.092640
3    CE_max  0.091198
0   CE_mean  0.090631
27  PS2_max  0.069767
23  PS1_max  0.065565
61  TS4_std  0.065230
```

System Failure Prediction (Fault Prediction): Predict future system conditions (whether they will deteriorate or remain normal) based on current sensor data to support preventive maintenance before failure occurs.

This dataset doesn't have an explicit time index, but we can assume the data is sequential (since it's time-series). The profile label represents the current state of the system:
3 = good
20 = declining
100 = bad / broken
We'll create a new label: "will break in the next n steps."

```
=== Multi-step Evaluation Summary ===
   step  accuracy  precision_1  recall_1      f1_1
0     5  0.997732     1.000000  0.993243  0.996610
1    10  0.997732     1.000000  0.993243  0.996610
2    20  0.995465     0.993243  0.993243  0.993243
3    30  0.995465     1.000000  0.986486  0.993197
```



High and stable accuracy up to 30 steps ahead, Current sensor data is highly informative for future predictions. Recall decreases at n=30, The model begins to fail to detect some cases that would otherwise fail if predictions were too far ahead.

Predictive Window Maintenance Simulation: real-time system monitoring to predict whether the system will fail in the near future, this approach is closest to the real application of predictive maintenance.

```
=== Real-Time Monitoring Simulation Evaluation ===
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        10
           1       0.98      1.00      0.99       431

    accuracy                           0.98       441
   macro avg       0.49      0.50      0.49       441
weighted avg       0.96      0.98      0.97       441
```

The model only predicts one class (label 1). All inputs are predicted as "will fail." As a result, recall is high for class 1, but class 0 (normal) is not detected at all.

Accuracy is deceptive (≈ 98%). This appears high, but this is because the test data is dominated by class 1 (431 vs. 10). There are no correct class 0 predictions. This is called class imbalance bias.

The model is too oversensitive/false alarms. All conditions are considered potential failures, which can lead to: Over-maintenance (high costs). Continuous false alarms. Suboptimal systems or unnecessary downtime.

Improvement Recommendations: Address class imbalance, Use more precise metrics and Analyze time-to-failure for more flexible decision making.