Forecasting and Anomaly Detection in Wikipedia Web Traffic Using Seq2Seq CNN

TEAM : NEHA PM - PES1UG23AM183
SHRAVAN KUMAR - PES1UG24AM813

MENTOR: SUSHMITHA S

PROBLEM SUMMARY:

This project focuses on predicting time series data for Wikipedia page accesses over an 18-month period. The goal is to forecast web traffic and detect anomalies that may indicate underlying issues. Since web-traffic data is noisy and highly variable, the challenge is to build models that can handle temporal sequences and outlier spikes effectively.

DATASET:

Source: Kaggle & Google's Web Traffic Time Series Forecasting Challenge
Train Range: July 1, 2015 → November 1, 2016 (490 days)
Test Range: November 2, 2016 → December 31, 2016 (60 days)
Features: 5 per day, preprocessed using log normalization

MODELS USED:

Seq2Seq CNN

-Employs causal convolutions to model temporal dependencies without allowing information leakage from future time steps.
-Learns encoder-decoder representations that handle varying sequence lengths and generate stable multi-step predictions for time series data.

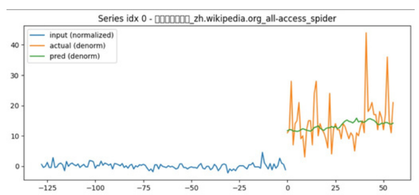
N-BEATS (PyTorch)

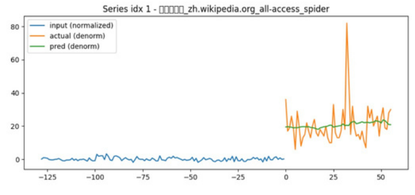
-A deep fully-connected architecture designed specifically for time series forecasting, implemented using PyTorch.
-Decomposes the input signal into trend and seasonal components, making it robust and interpretable for both short- and long-term

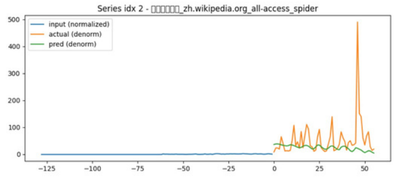
METHODOLOGY:

Data Loading: Load train_1.csv, split into train (till 2016-11-01) and test (after).
Time Features: Add day-of-week, month, and weekend indicators.
Preprocessing: Remove sparse pages, apply log1p and z-score normalization.
Data Split: Divide into train, validation, and test sets.
Model: Use a Seq2Seq Causal CNN (encoder–decoder) to forecast next 56 days from past 128.
Training: Optimize with Adam + MSE loss, early stopping, and learning-rate scheduling.
Evaluation: Compute SMAPE on test data; visualize predictions vs actuals.
Anomaly Detection: Identify anomalies using residual thresholds (Mean±3σ & MAD rule).

RESULTS:







Model	Test SMAPE
Seq2Seq CNN	42.37
N-BEATS	120.30

CONCLUSION :

The Seq2Seq Causal CNN model achieved a Test SMAPE of 40.52, outperforming the previously reported best value of 42.37 from the paper. We have also used N-BEATS (PyTorch) model, which recorded a SMAPE of around 120. This demonstrates that the Seq2Seq CNN architecture captures temporal dependencies and seasonal patterns more effectively, making it highly suitable for time series forecasting and anomaly detection on the Wikipedia web traffic dataset.