Probabilistic Machine Learning

final project

Anomaly Detection with Variational Autoencoders for Financial Data

Andrea Buscema & Tanja Derin

June 2024

Table of contents

01

VAE

02

Problem Statement

03

LSTM model

04

CVAE model

05

Conclusions

Variational Autoencoders

Encoder

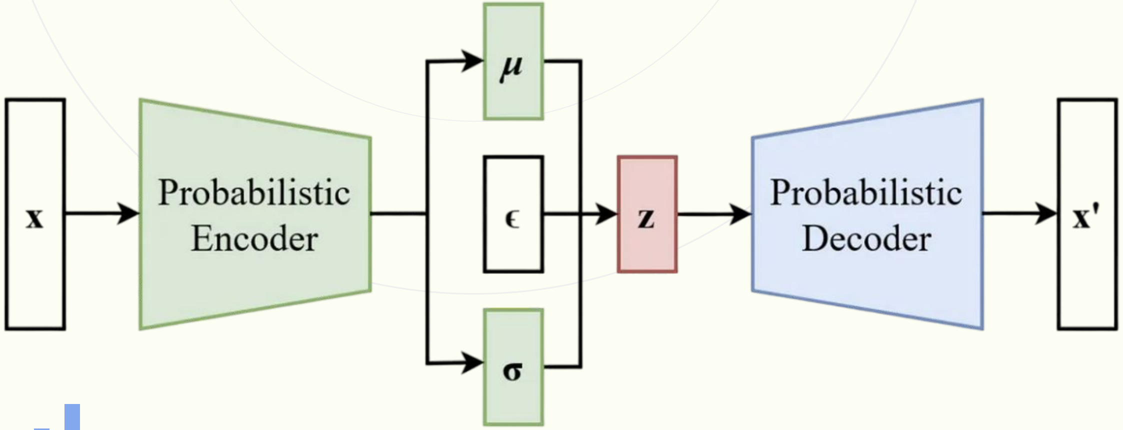
Compresses the input into a lower-dimensional representation. This encoding captures the most significant features of the input data

Latent Space

Encoded representation of input data, utilizing the reparametrization trick to facilitate model training and data generation

Decoder

It takes the compressed data and tries to reconstruct the original data as closely as possible



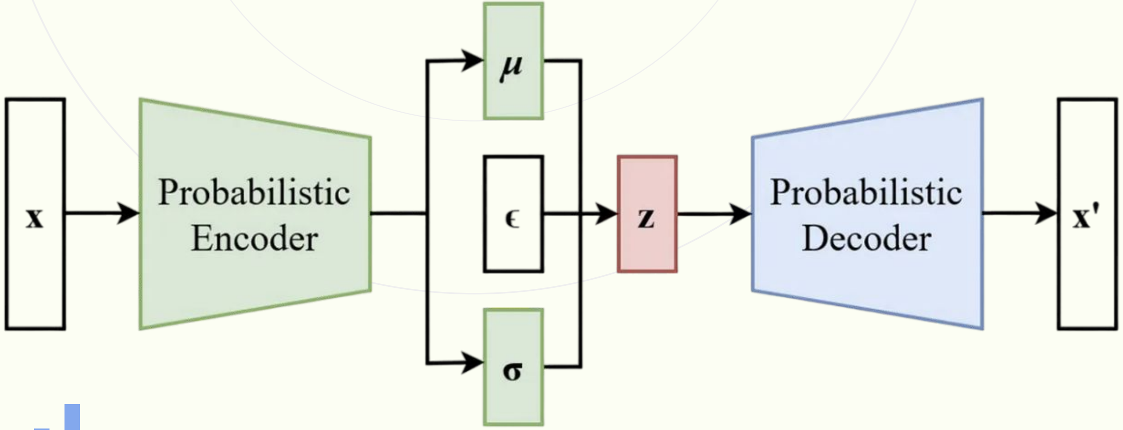
Variational Autoencoders

The input is **encoded** as distribution over the latent space

A point from the latent space is **sampled** from that distribution

The sampled point is **decoded** and the reconstruction error can be computed

The **reconstruction error** is back propagated through the network



02

Problem Statement

Anomaly detection

Anomaly Detection Techniques

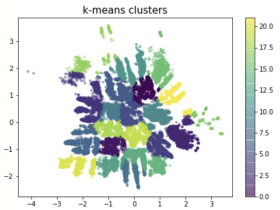
SUPERVISED

Training and testing data are both labeled as either normal or anomalous

UNSUPERVISED

SEMI-SUPERVISED

Some of the data are labeled and some are unlabeled



Representation in Latent Space

Clustering

Wasserstein Distance-Based Detection

Reconstruction Quality

Anomalies are likely to have higher reconstruction errors because the model is trained predominantly on normal data and thus struggles to reconstruct outliers

Problem Statement

For this project we tried two applications of Variational Autoencoders to detect anomalies in stock data

- First approach utilizes an **LSTM-VAE model** analyzing 30 years of AAPL stock data to capture deeper, more complex temporal patterns in anomalies
- The second approach utilizes **Conditional VAE** on the past five years of Dow 30 stocks data, conditioned on stock symbols

Comparing the findings from these two models helps validate the **effectiveness** of VAEs in recognizing anomalies in stock behavior over different periods and different dataset

03

LSTM Model

LSTM

Long-Short Term Memory it's a type of recurrent neural network (RNN) architecture designed to address the vanishing problem that standard RNNs can face when dealing with long sequences. LSTMs are particularly effective at capturing and learning from **long-term dependencies** in **sequential data**

- **Memory Cell**: maintain information over long periods
- **Architecture**: An LSTM unit consists of Cell state, Hidden state, Forget Gate, Input Gate and Output Gate

At each time step, the LSTM decides what to forget, what new information to store and what to output. This selective memory mechanism allows LSTMs to learn which information is important to keep or discard over long sequences

VAE-LSTM

The **encoder LSTM** processes the input sequence and captures its temporal dependencies into a fixed-size representation (the final hidden state)

The **decoder LSTM** takes the latent representation and generates a sequence of the same length as the input, maintaining temporal coherence in the reconstruction

Dataset

The dataset refers AAPL (Apple Inc.) daily stock prices, from January 2 1990 to December 29 2023

`Date` column was converted to a datetime format

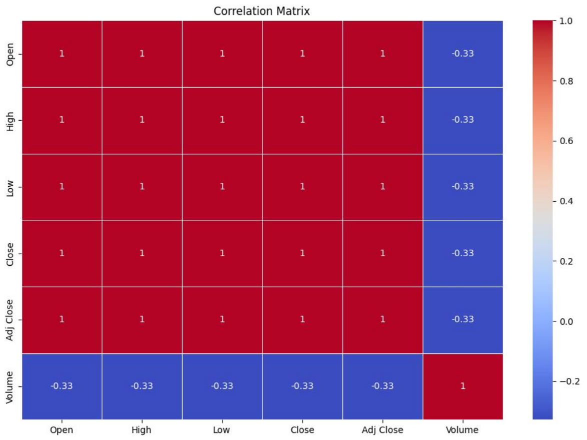
All numerical features were **normalized** to a range of 0 to 1 to ensure suitable for training the VAE-LSTM model

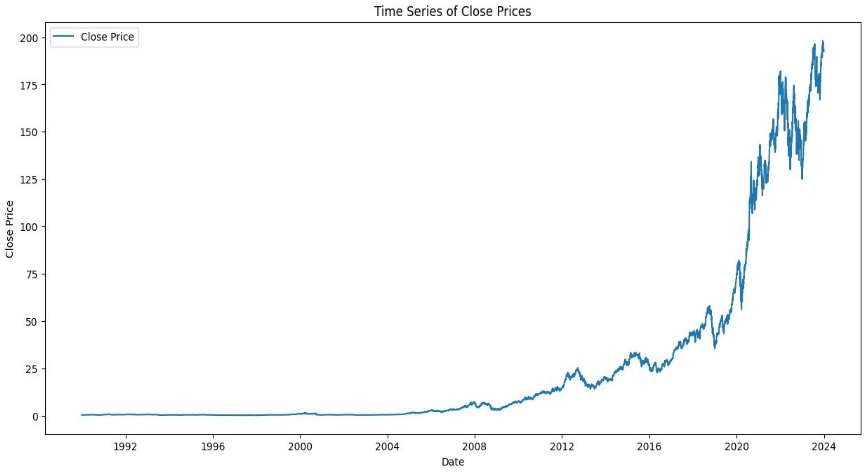
In this VAE-LSTM model, it was implemented sequences of a fixed length, which involves creating overlapping windows of data that the model can use to learn temporal patterns

→ X (features) shape (8505, 60, 5)

→ Y (labels) shape (8505, 5)

8505 sequences, each with a length of 60 time steps and 5 features





Model Building VAE-LSTM

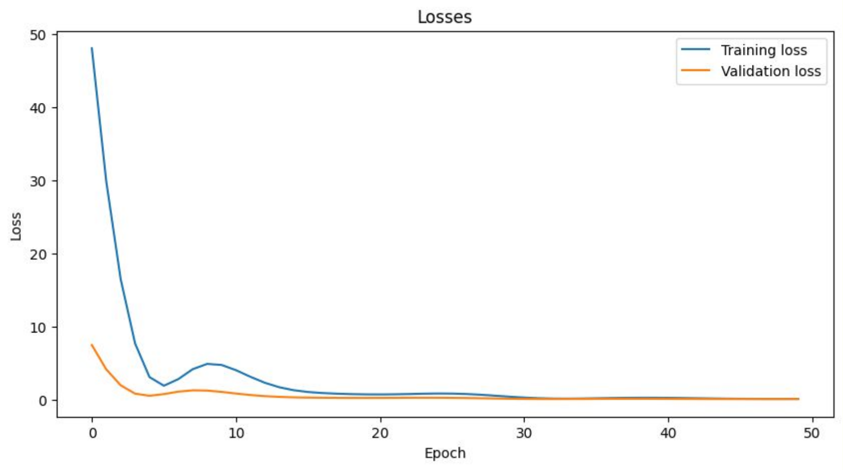
Overview

The VAE LSTM model developed integrates both the LSTM architecture and Variational Autoencoder principles

- **Encoder** uses an LSTM to process the input sequence and the final hidden state of the LSTM is used to parameterize the latent space

Latent Space with two fully connected layers that output the mean and log-variance of the latent distribution

Decoder uses another LSTM to reconstruct the sequence from the latent representation and a final fully connected layer maps the LSTM output back to the original input dimensions



Model Architecture

- **Initializes** the encoder LSTM, latent space fully connected layers, decoder LSTM and output layer

- **Encode method** which processes the input sequence through the encoder LSTM and maps the final hidden state to μ and σ (log variance) of the latent space

- **Reparameterization trick** to handle the latent space in probabilistic terms, enabling generation and manipulation

- **Decode method** repeats the latent vector for each time step in the sequence and processes this through the decoder LSTM and output layer to reconstruct the sequence

- **Forward method** combines encode, reparameterize, and decode to process input data through the entire model

The **Loss function** combines **MSE** for reconstruction accuracy and the **Kullback-Leibler divergence** to regulate the distribution of the latent space, which helps in learning efficient and meaningful representations of data

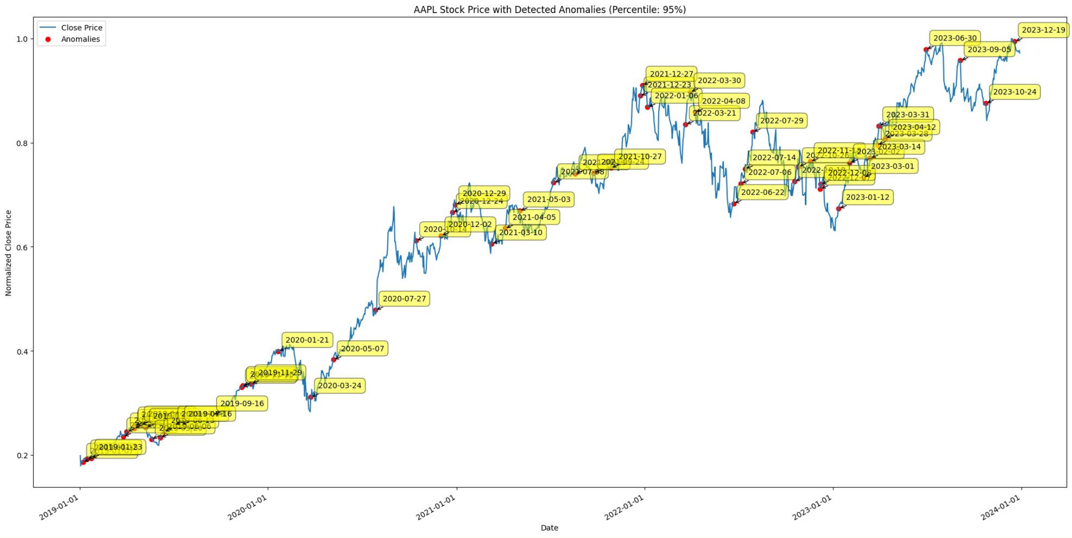
Anomaly Detection

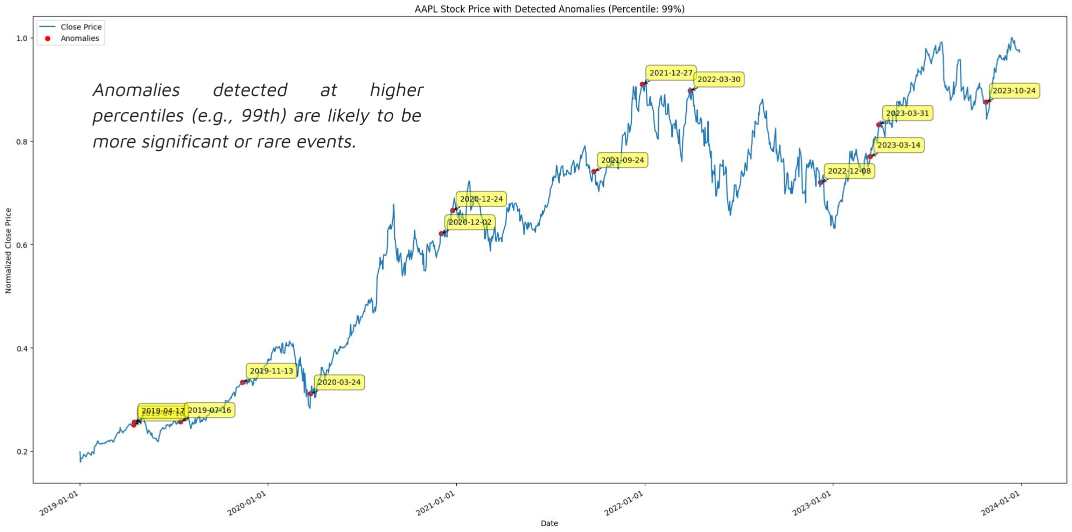
Reconstruction Error

The reconstruction error in this context is calculated as the **mean squared error (MSE)** between the original input data and the reconstructed data for each sequence. This error quantifies the model's ability to accurately reconstruct the input data after it has been encoded into a lower-dimensional latent space and subsequently decoded back into the original space in the VAE-LSTM framework

Threshold Selection

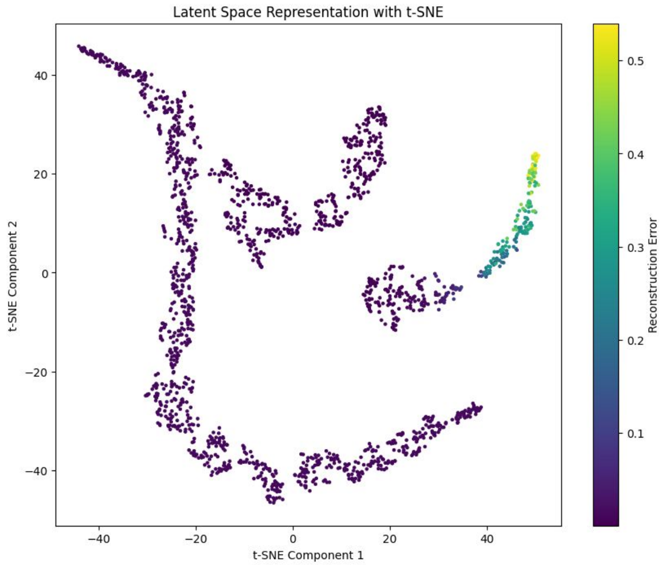
Multiple thresholds (90th, 95th, 97th and 99th percentiles) are used to detect anomalies. For each threshold, anomalies are identified as points with errors above the threshold. The contextual analysis around anomalies implemented helps in understanding whether these are genuine anomalies or false positives





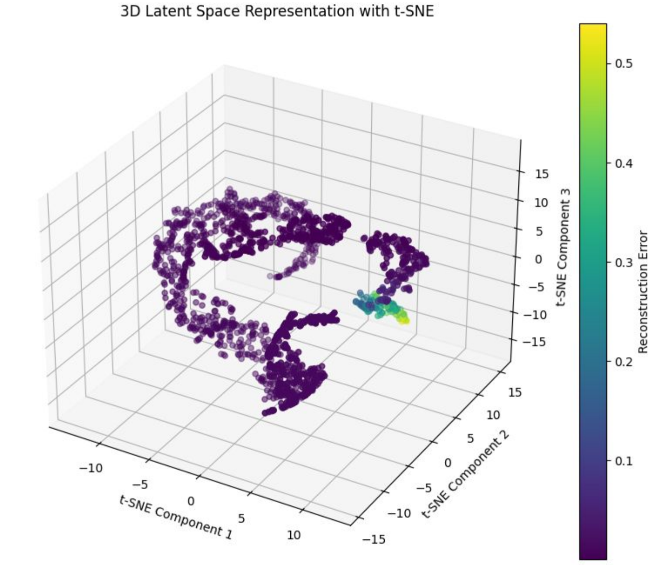
Anomalies detected at higher percentiles (e.g., 99th) are likely to be more significant or rare events.

Latent Space Representation



Visualizing the latent space helps in understanding how well the model has learned to represent the data. It can also help in identifying clusters in the data

In these plots, each point represents a stock price data point that has been encoded into a latent space by the VAE-LSTM model



The color gradient represents the reconstruction error, with higher values indicating potential anomalies. The presence of distinct clusters in the latent space suggests that the VAE has learned to group similar sequences together

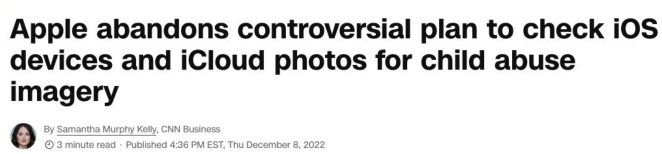
The purple points represent normal data points where the reconstruction error is low, indicating that the model has accurately reconstructed these points from the latent space

The yellow and green points represent anomalies where the reconstruction error is higher, suggesting unusual or abnormal patterns in the stock prices

Some dates - 99th percentile

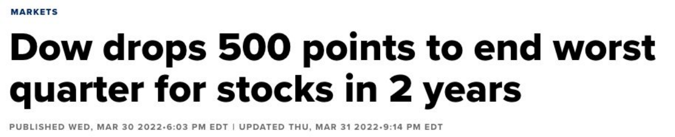
Anomalies detected on the following dates:

2019-04-16
2019-04-17
2019-07-16
→ 2019-11-13
2020-03-24
2020-12-02
2020-12-24
→ 2021-09-24
2021-12-27
2022-03-30
→ 2022-12-08
2023-03-14
→ 2023-03-31
→ 2023-10-24



Apple abandons controversial plan to check iOS devices and iCloud photos for child abuse imagery

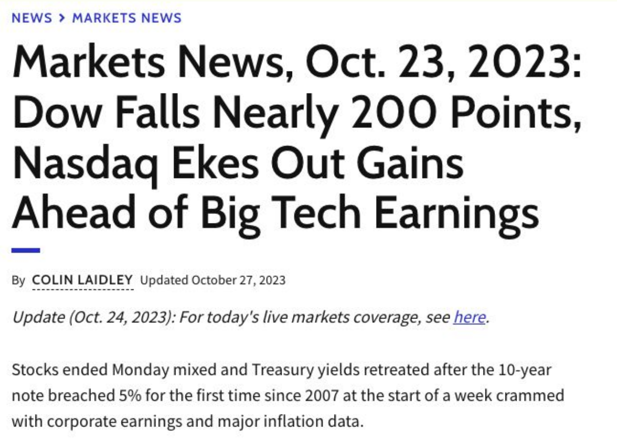
By Samantha Murphy Kelly, CNN Business
3 minute read · Published 4:36 PM EST, Thu December 8, 2022



MARKETS

Dow drops 500 points to end worst quarter for stocks in 2 years

PUBLISHED WED, MAR 30 2022·6:03 PM EDT | UPDATED THU, MAR 31 2022·9:14 PM EDT



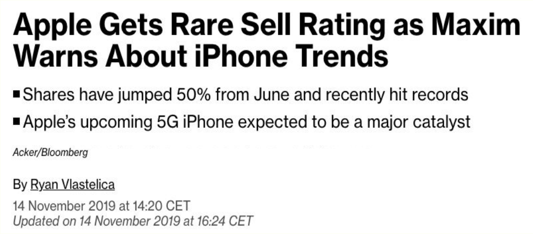
NEWS › MARKETS NEWS

Markets News, Oct. 23, 2023: Dow Falls Nearly 200 Points, Nasdaq Ekes Out Gains Ahead of Big Tech Earnings

By COLIN LAIDLEY Updated October 27, 2023

Update (Oct. 24, 2023): For today's live markets coverage, see here.

Stocks ended Monday mixed and Treasury yields retreated after the 10-year note breached 5% for the first time since 2007 at the start of a week crammed with corporate earnings and major inflation data.

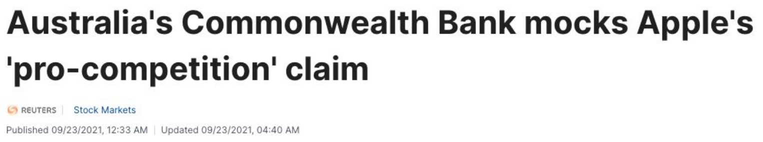


Apple Gets Rare Sell Rating as Maxim Warns About iPhone Trends

- Shares have jumped 50% from June and recently hit records
- Apple's upcoming 5G iPhone expected to be a major catalyst

Acker/Bloomberg

By Ryan Vlastelica
14 November 2019 at 14:20 CET
Updated on 14 November 2019 at 16:24 CET



Australia's Commonwealth Bank mocks Apple's 'pro-competition' claim

REUTERS | Stock Markets
Published 09/23/2021, 12:33 AM | Updated 09/23/2021, 04:40 AM

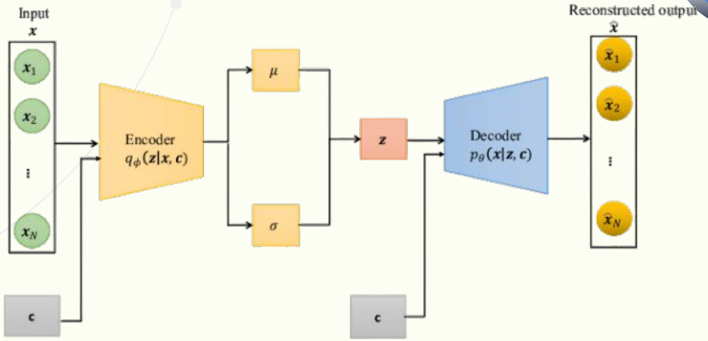
04

CVAE model

Conditional VAE

Conditional Variational Autoencoders extend the traditional one by **conditioning** both the encoder and the decoder on additional variables , representing attributes or labels of the data

- Encoder: Models $Q(z|X, c)$, taking into account both the input data and the conditioning variable.
- Decoder: Decoder becomes $P(X|z, c)$, aiming to reconstruct the input based on the latent variables and the conditioning context.
- ELBO: The objective function in CVAEs is adjusted to account for the conditioning variable, and is expressed as



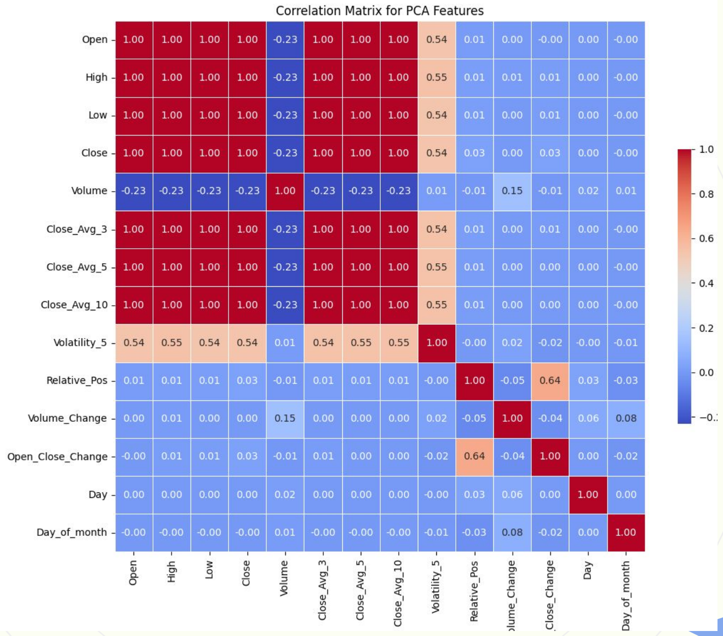
$$\log P(X|c) - D_{KL}[Q(z|X, c)\|P(z|X, c)] = \mathbb{E}[\log P(X|z, c)] - D_{KL}[Q(z|X, c)\|P(z|c)]$$

For the stock prediction model, the CVAE is **conditioned on stock symbols**, which enables it to capture unique characteristics and patterns specific to each stock

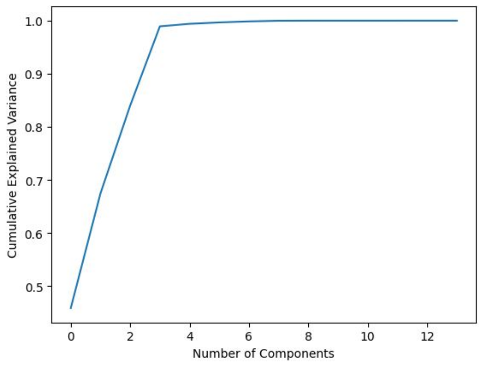
Dataset

The dataset is stock market data from the S&P 500 index covering a period of five years from January 2, 2019, to December 29, 2023

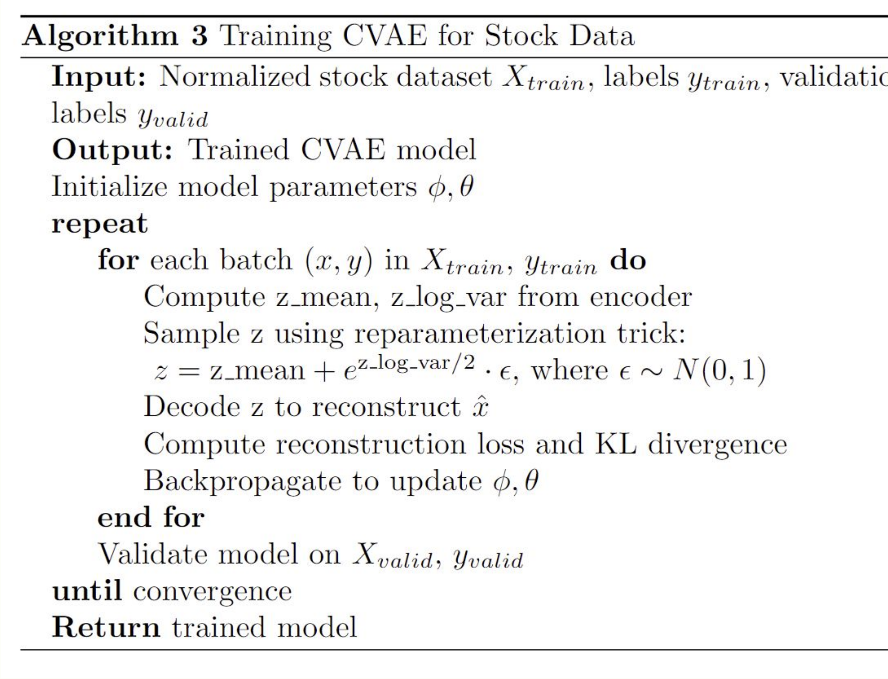
- The original data was enriched through the extraction of **additional features**
- All numerical features were **normalized** to a range of 0 to 1
- Transforming the **`Symbol`** column into a one-hot encoded format


Correlation Matrix for PCA Features

	Open	High	Low	Close	Volume	Close_Avg_3	Close_Avg_5	Close_Avg_10	Volatility_5	Relative_Pos	olume_Change	Close_Change	Day	Day_of_month
Open	1.00	1.00	1.00	1.00	-0.23	1.00	1.00	1.00	0.54	0.01	0.00	-0.00	0.00	-0.00
High	1.00	1.00	1.00	1.00	-0.23	1.00	1.00	1.00	0.55	0.01	0.01	0.01	0.00	-0.00
Low	1.00	1.00	1.00	1.00	-0.23	1.00	1.00	1.00	0.54	0.01	0.00	0.01	0.00	-0.00
Close	1.00	1.00	1.00	1.00	-0.23	1.00	1.00	1.00	0.54	0.03	0.00	0.03	0.00	-0.00
Volume	-0.23	-0.23	-0.23	-0.23	1.00	-0.23	-0.23	-0.23	0.01	-0.01	0.15	-0.01	0.02	0.01
Close_Avg_3	1.00	1.00	1.00	1.00	-0.23	1.00	1.00	1.00	0.54	0.01	0.00	0.01	0.00	-0.00
Close_Avg_5	1.00	1.00	1.00	1.00	-0.23	1.00	1.00	1.00	0.55	0.01	0.00	0.00	0.00	-0.00
Close_Avg_10	1.00	1.00	1.00	1.00	-0.23	1.00	1.00	1.00	0.55	0.01	0.00	0.00	0.00	-0.00
Volatility_5	0.54	0.55	0.54	0.54	0.01	0.54	0.55	0.55	1.00	-0.00	0.02	-0.02	-0.00	-0.01
Relative_Pos	0.01	0.01	0.01	0.03	-0.01	0.01	0.01	0.01	-0.00	1.00	-0.05	0.64	0.03	-0.03
Volume_Change	0.00	0.01	0.00	0.00	0.15	0.00	0.00	0.00	0.02	-0.05	1.00	-0.04	0.06	0.08
Open_Close_Change	-0.00	0.01	0.01	0.03	-0.01	0.01	0.00	0.00	-0.02	0.64	-0.04	1.00	0.00	-0.02
Day	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	-0.00	0.03	0.06	0.00	1.00	0.00
Day_of_month	-0.00	-0.00	-0.00	-0.00	0.01	-0.00	-0.00	-0.00	-0.01	-0.03	0.08	-0.02	0.00	1.00

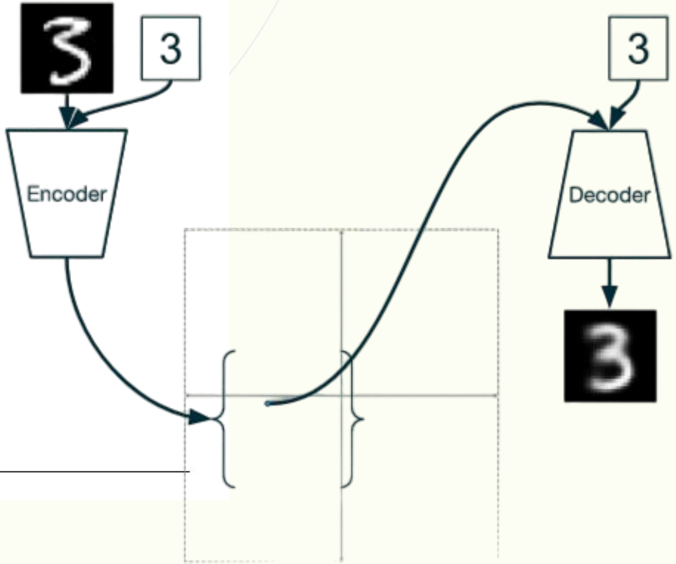


The model



Algorithm 3 Training CVAE for Stock Data

```
Input: Normalized stock dataset $X_{train}$, labels $y_{train}$, validation data $X_{valid}$,
labels $y_{valid}$
Output: Trained CVAE model
Initialize model parameters $\phi, \theta$
repeat
    for each batch $(x, y)$ in $X_{train}$, $y_{train}$ do
        Compute z_mean, z_log_var from encoder
        Sample z using reparameterization trick:
         $z = \text{z\_mean} + e^{\text{z\_log\_var}/2} \cdot \epsilon$, where $\epsilon \sim N(0, 1)$
        Decode z to reconstruct $\hat{x}$
        Compute reconstruction loss and KL divergence
        Backpropagate to update $\phi, \theta$
    end for
    Validate model on $X_{valid}$, $y_{valid}$
until convergence
Return trained model
```



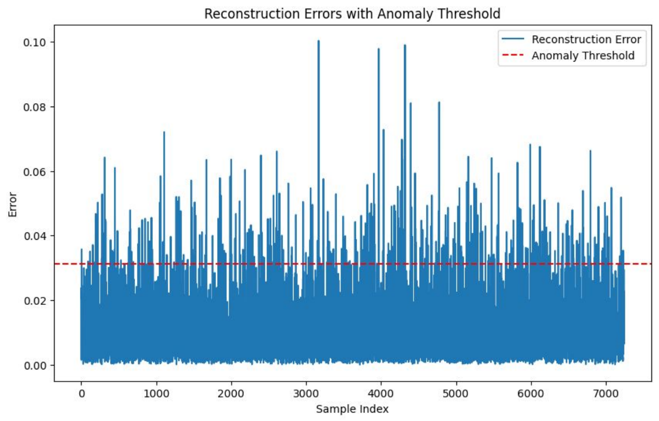
Anomaly Detection

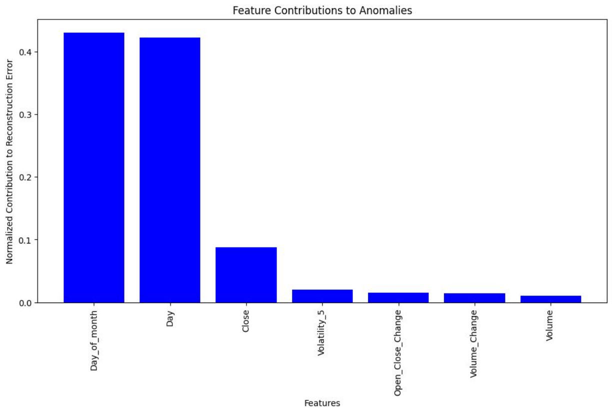
01

For each batch of test data, the model **predicts** the reconstructed outputs along with the latent variables mean and variance

02

Error Calculation: The reconstruction error for each sample is computed as the mean squared error between the original and reconstructed data





03

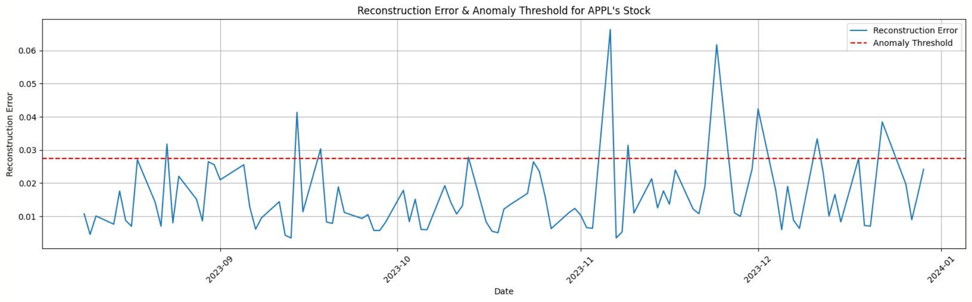
Thresholding for Anomalies: An anomaly threshold is set at the 95th percentile of the reconstruction errors, identifying the top 5% errors as anomalies

04

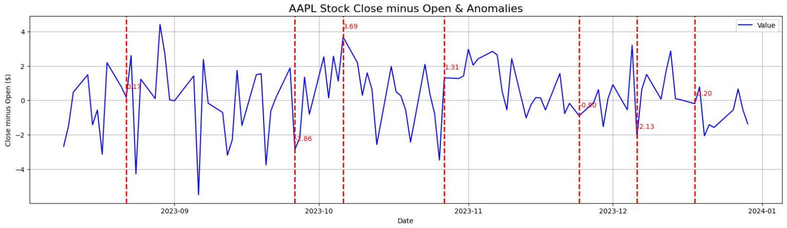
Visually assess the reconstruction errors against the anomaly threshold, highlighting the distribution of errors and the identified anomalies

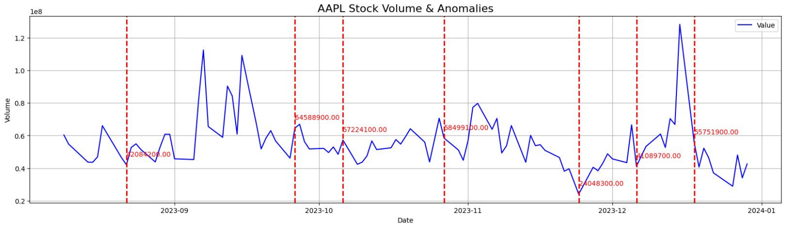
Anomaly detection for APPL

With the CVAE, we can ask the model to recreate data for a particular label. In the example of stock market data, we can ask it to recreate data for a particular stock symbol- AAPL



To **monitor anomalies** across different stock features such as closing prices, trading volumes, and the differences between opening and closing prices a series we generated these plots



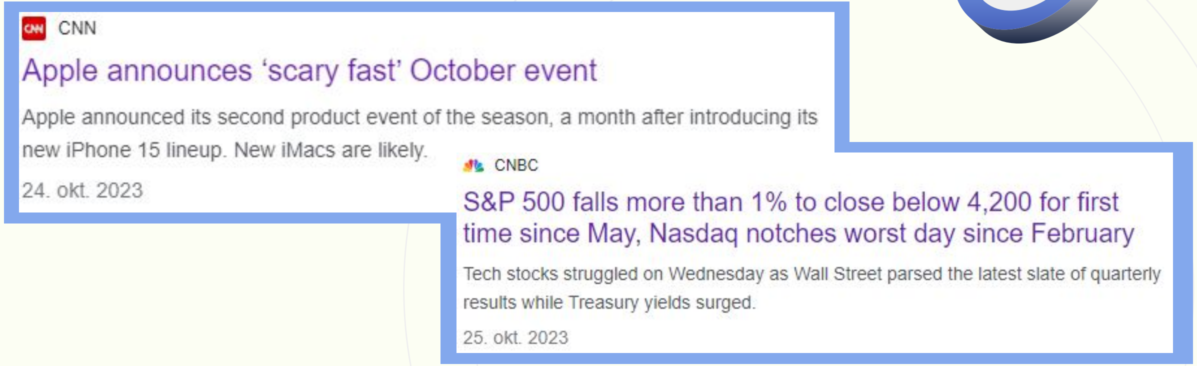


05
Conclusions

Conclusions

Detected anomalies:

CVAE	LSTM
2023-08-25	2023-06-30
2023-08-29	2023-09-05
2023-08-30	2023-10-24
2023-09-15	2023-12-19
2023-10-23	
2023-10-24	
2023-11-21	
2023-11-30	
2023-12-07	
2023-12-28	



CNN

Apple announces 'scary fast' October event

Apple announced its second product event of the season, a month after introducing its new iPhone 15 lineup. New iMacs are likely.

24. okt. 2023

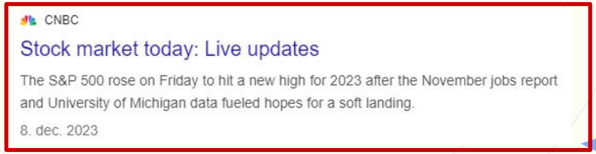
CNBC

S&P 500 falls more than 1% to close below 4,200 for first time since May, Nasdaq notches worst day since February

Tech stocks struggled on Wednesday as Wall Street parsed the latest slate of quarterly results while Treasury yields surged.

25. okt. 2023

CVAE allows for a **broader perspective**, potentially making it better suited for applications needing a generalized model across various stocks

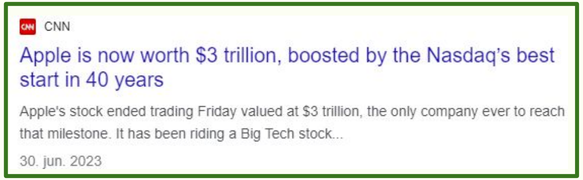


CNBC

Stock market today: Live updates

The S&P 500 rose on Friday to hit a new high for 2023 after the November jobs report and University of Michigan data fueled hopes for a soft landing.

8. dec. 2023



CNN

Apple is now worth $3 trillion, boosted by the Nasdaq's best start in 40 years

Apple's stock ended trading Friday valued at $3 trillion, the only company ever to reach that milestone. It has been riding a Big Tech stock...

30. jun. 2023

LSTM's focus on long-term data from a single stock provide **deeper insights** into specific stock behaviors

DATA SCIENCE & ARTIFICIAL INTELLIGENCE

Thanks!

Any questions?

UNIVERSITÀ DEGLI STUDI DI TRIESTE