Anomaly Detection in Turkish Stock Market : LSTM

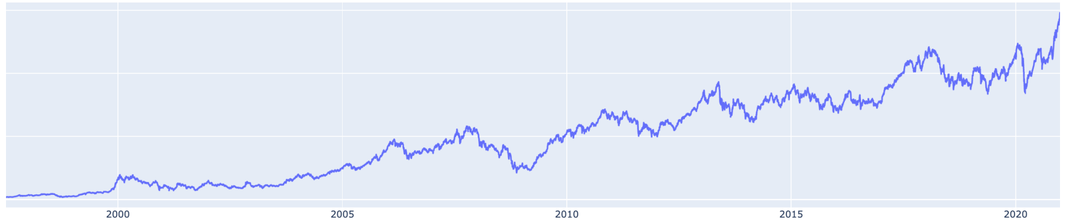
İçimi Demirağ, Semih Engin

Problem

Anomalies in the Turkish stock market, particularly within the BIST100 index, pose significant risks and missed opportunities for investors. Traditional statistical methods for anomaly detection often fall short in capturing the complex dynamics of financial data. This research aims to bridge this gap by developing an LSTM-based anomaly detection model specifically tailored to the BIST100 index.

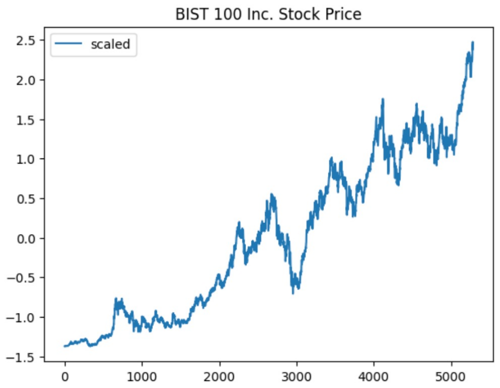
Dataset and Preprocessing

Our project leverages a comprehensive dataset spanning daily BIST100 index prices from 1997 to the present sourced from Yahoo Finance.



We address missing values by removing rows with missing entries, ensuring the time series retains its integrity. Additionally, we normalize the columns for improved training convergence, scaling all values to a similar range without compromising crucial trends.

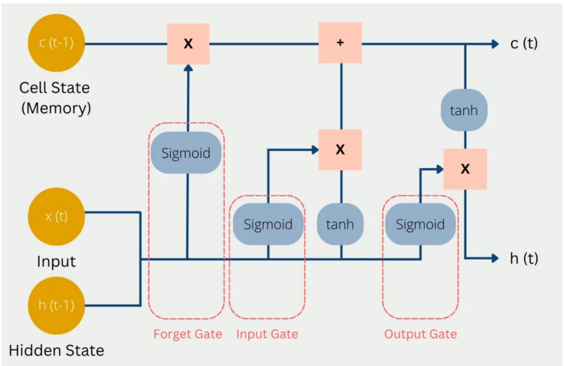
```
df = df.dropna()

df.loc[:5982, 'Open'] = df.loc[:5982, 'Open'] / 100
df.loc[:5982, 'High'] = df.loc[:5982, 'High'] / 100
df.loc[:5982, 'Low'] = df.loc[:5982, 'Low'] / 100
df.loc[:5982, 'Close'] = df.loc[:5982, 'Close'] / 100
df.loc[:5982, 'Adj Close'] = df.loc[:5982, 'Adj Close'] / 100

df = df.loc[df['Date'] < '2020-12-31']
```



Method

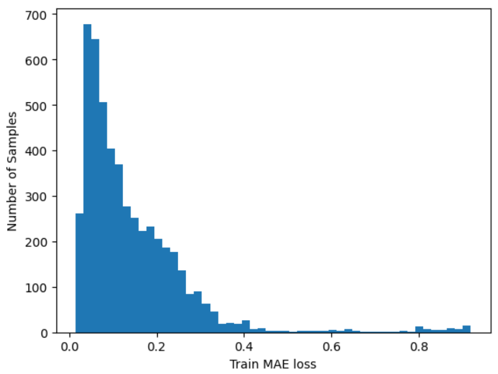
An LSTM-based anomaly detection model has been designed to detect anomalies in the Turkish stock market. The model uses hyperbolic tangent activation function, RepeatVector, Dropout, TimeDistributed (Dense) layer, and Adam optimizer. LSTM is a neural network that can capture long-term dependencies in time series data.

- The hyperbolic tangent activation function allows the model to learn intricate relationships in complex financial data.
- The RepeatVector layer allows the model to see the entire sequence of market movements.
- The Dropout layer helps the model to generalize better by preventing it from becoming overly reliant on specific quirks.
- The TimeDistributed (Dense) layer applies a comprehensive analysis to each individual time step.
- The Adam optimizer helps the model to converge faster and more accurately.



Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 128)	66560
dropout (Dropout)	(None, 128)	0
repeat_vector (RepeatVector)	(None, 30, 128)	0
lstm_1 (LSTM)	(None, 30, 128)	131584
dropout_1 (Dropout)	(None, 30, 128)	0
time_distributed (TimeDistributed)	(None, 30, 1)	129

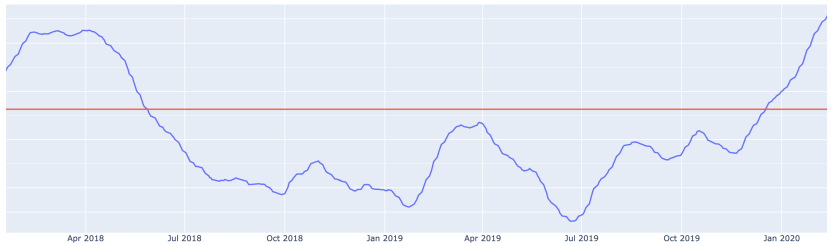
Total params: 198273 (774.50 KB)
Trainable params: 198273 (774.50 KB)
Non-trainable params: 0 (0.00 Byte)

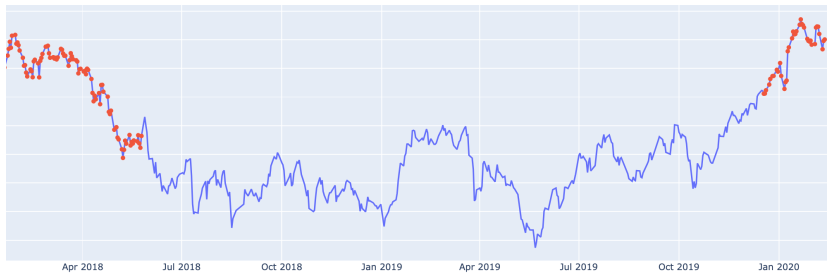


- In our project, we use MAE to evaluate the performance of our model. MAE is a non-parametric measure of the average distance between predicted and observed values. A lower MAE indicates that the model is making more accurate predictions.

Results and Findings

Within our LSTM-based anomaly detection architecture, we utilize mean absolute error (MAE) as a foundational measurement for defining a threshold, enabling the precise recognition of anomalous data points.





The choice of the threshold for the reconstruction error emerged as a critical factor influencing the quality of anomaly detection. A lower threshold correlated with higher reconstruction accuracy, emphasizing the need for meticulous threshold selection for optimal model performance.

Future research may try to improve the performance of anomaly detection models using methods such as retuning hyperparameters and changing evaluation metrics, paving the way for wider application of these models in diverse domains, from industrial process monitoring to financial fraud detection.