Report on Stock Price Prediction using Principal Component Analysis

Team Members

Ujjwal Raj(ME22B1072)
Aman Gupta(ME22B1071)
Prashant Tyagi(ME22B1069)
Vinay(ME22B1070)

1 Abstract

The categorization of high-dimensional data presents a fascinating challenge to machine learning models, as the frequent presence of highly correlated dimensions or attributes can affect the accuracy of classification models. In this paper, we investigate the problem of high dimensionality in stock exchange data to predict market trends by applying Principal Component Analysis (PCA) with linear regression. PCA can help improve the predictive performance of machine learning methods while reducing redundancy in the data. We conduct experiments on high-dimensional spectral data from three stock exchanges: New York Stock Exchange, London Stock Exchange, and Karachi Stock Exchange. We compare the accuracy of linear regression classification models before and after applying PCA. The experiments demonstrate that PCA can enhance the performance of machine learning models, provided that the relative correlation among input features is carefully investigated, and principal components are selected judiciously. We use Root Mean Square Error (RMSE) as an evaluation metric to assess the classification model.

2 Introduction

Prediction is a process to make assumption of future based on existing data. The more precise the prediction, the more it could be easier to make decision for future. Prediction of stock exchange trends has been an interesting topic in the field of pattern recognition and machine learning because of its possible monetary profit. A stock market is an organized set-up with a regulatory body and has registered members who can buy or sell shares. It's a public market, where different companies invest high capital and do trading of their shares. Stock market prediction provides information about stock market, which can help the shareholders to make decision about trading. It may serve as warning system for long-term shareholders while for short-term investors may serve as recommender system. One of the problems with Stock market data is that the available data is highly volatile with very high dimensions. Many of the attributes are highly correlated and makes prediction of stock market a highly challenging, complicated and daunting task. The best solution for above problem is to reduce dimensions of data. Measuring correlation between data dimensions can do reduction and discarding those attributes, or dimensions which have least impact on overall prediction model. Principal component analysis is one of effective technique, which can be applied to reduce the dimensionality of data. In this paper, we have predicted the trend of Verastem Inc exchanges by using linear regression as a classification model. The past values are used to build a classification function and then uses this function to predict about future values. Principal component analysis is used with linear regression model to check that whether PCA has improved the accuracy of model or not.

3 Problem Definition

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction. In this project, our objective is to harness the power of PCA to predict stock prices. By reducing the dimensionality of the data, we aim to extract meaningful patterns and improve the accuracy of our predictions.

4 Motivation

Predicting even slight fluctuations in stock movements can translate into substantial financial gains. Yet, grappling with voluminous datasets teeming with diverse features presents a formidable obstacle. The curse of dimensionality further complicates this scenario by inundating the analysis with redundancy and noise, impeding accurate predictions. However, Principal Component Analysis (PCA) emerges as a robust antidote to this predicament. By distilling complex datasets into more manageable dimensions, PCA effectively mitigates the curse of dimensionality while preserving essential information critical for accurate forecasting. Through dimensionality reduction, PCA streamlines the analysis process, facilitating clearer insights into stock market trends. Thus, PCA stands as a powerful tool in the arsenal of predictive analytics, enabling investors to make informed decisions amid the intricacies of financial markets.

5 Approach

To combat the curse of dimensionality, we employ two primary strategies: feature selection and feature extraction (dimensionality reduction). Feature selection involves identifying the most influential features, thereby enhancing model interpretability and performance. Dimensionality reduction techniques, such as PCA, mitigate the curse of dimensionality by transforming the data into a lower-dimensional space. This not only improves computational efficiency but also facilitates data visualization and understanding.

6 Example: Housing Dataset

Consider a scenario where a homebuyer must evaluate numerous features before making a purchase decision, such as proximity to amenities, neighborhood safety, and property size. Despite the abundance of factors, the ultimate decision often revolves around a few critical features. Similarly, in stock prediction, PCA helps identify key factors amidst a multitude of variables, simplifying the prediction process and improving decision-making.

7 Mathematical Method: Eigen Decomposition

Eigen decomposition is a fundamental concept in PCA. It involves decomposing a covariance matrix into its constituent eigenvalues and eigenvectors. These eigenvectors represent the principal components, with each capturing a certain amount of variance in the data. By retaining the principal components with the highest variance, we effectively reduce the dimensionality of the dataset while preserving the most significant information. Moreover, the concepts of linear regression seamlessly integrate with PCA, enabling us to model stock price trends accurately.

8 Mathematical Steps In PCA

8.1 Standardize the Data

First, we standardize the data to have a mean of 0 and a standard deviation of 1. This is given by:

$$X_{std} = \frac{X - \mu}{\sigma}$$

where X is the original feature vector, μ is the mean of the feature vector, and σ is the standard deviation of the feature vector.

8.2 Compute the Covariance Matrix

Next, we compute the covariance matrix of the standardized data. The covariance matrix is given by:

$$\Sigma = \frac{1}{n-1}(X_{std} - \mu)(X_{std} - \mu)^T$$

where n is the number of observations.

8.3 Compute the Eigenvalues and Eigenvectors

We then compute the eigenvalues and eigenvectors of the covariance matrix. The eigenvector v and eigenvalue λ

of a matrix A satisfy the equation:

$$Av = \lambda v$$

The eigenvectors represent the directions or components for the reduced subspace of X_{std}, while the eigenvalues represent the magnitudes for the directions.

8.4 Sort Eigenvalues and their Corresponding Eigenvectors

Next, we sort the eigenvalues in decreasing order and choose the first k eigenvectors, which correspond to the k largest eigenvalues, where k is the number of dimensions of the new feature subspace ($k \leq d$).

8.5 Transform the Original Dataset

Finally, we transform the original standardized dataset via the selected eigenvectors to obtain a k-dimensional feature subspace.

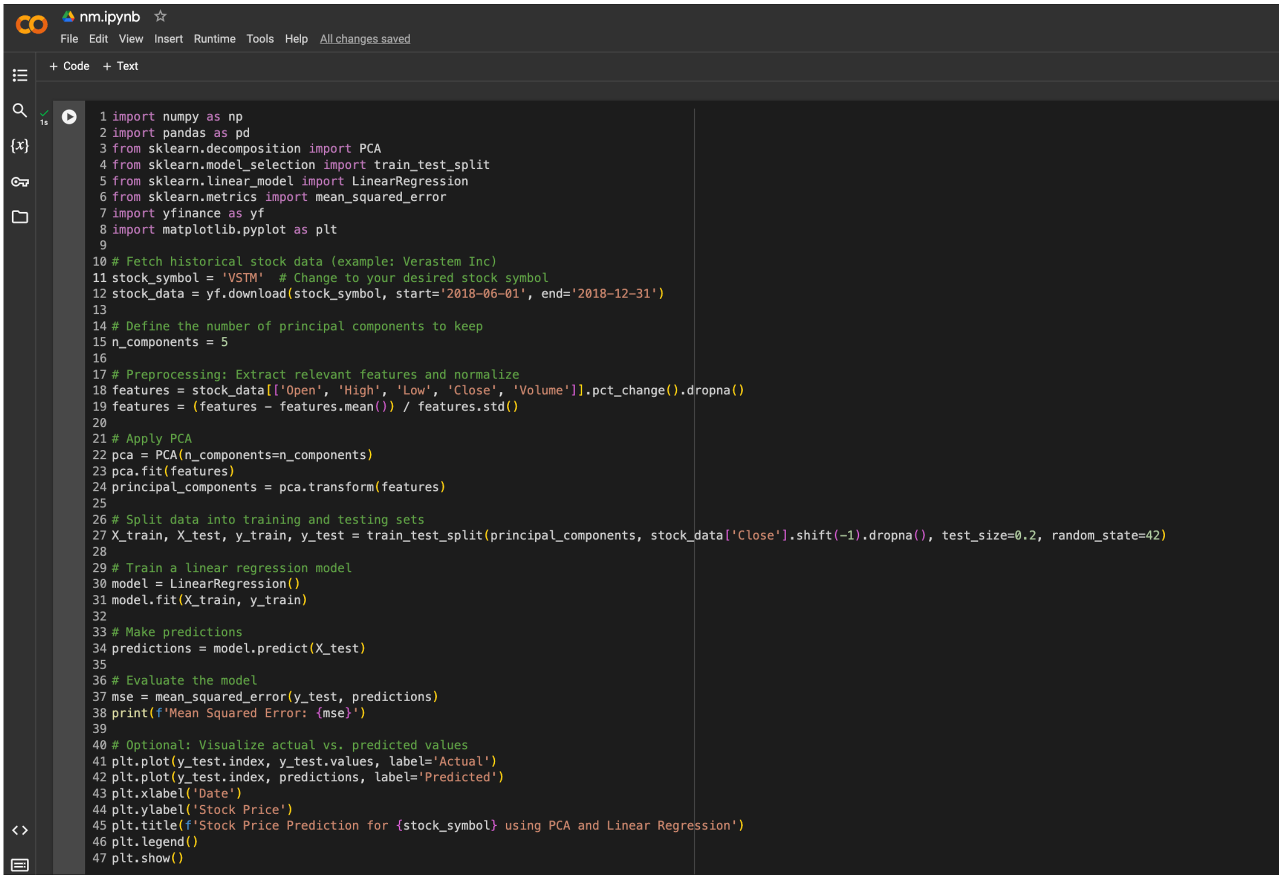
$$Y = X_{std} \cdot W$$

where Y is the transformed k-dimensional feature subspace and W is the matrix of the top k eigenvectors.

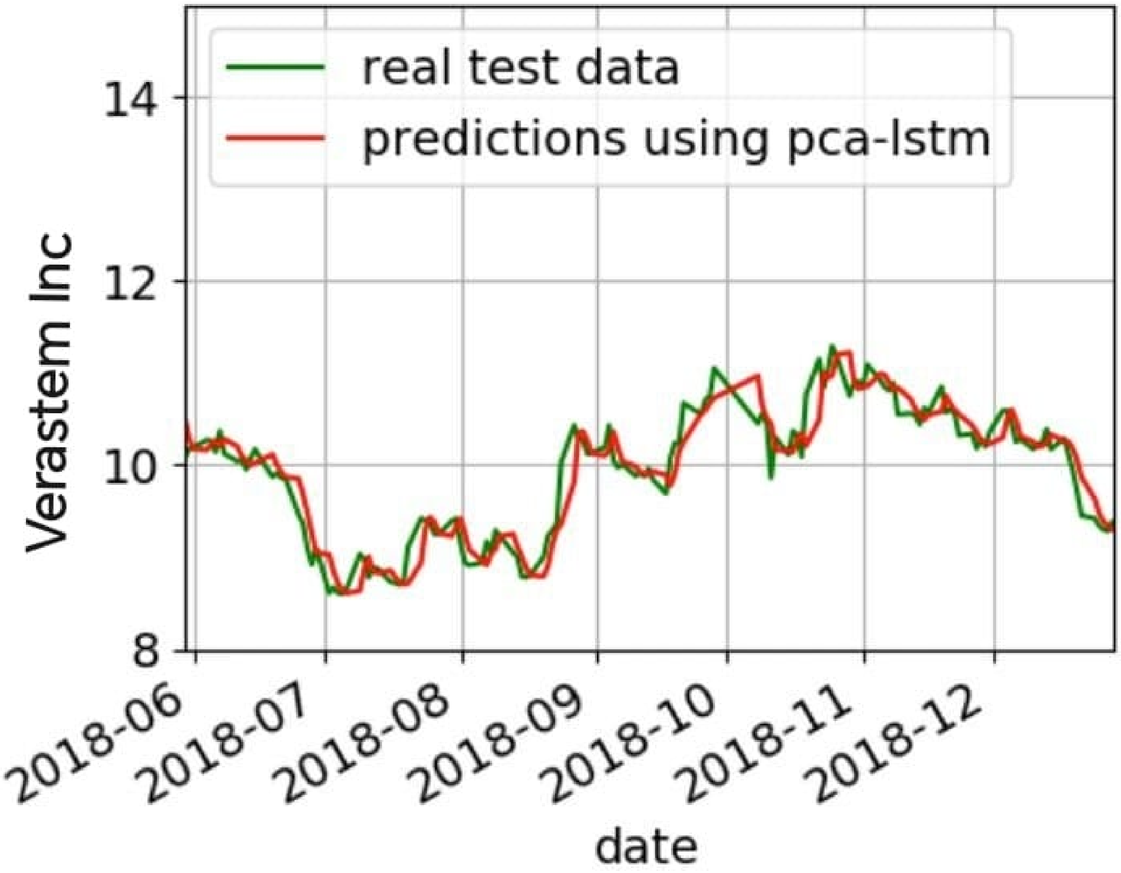
9 Finding Principal Components with Max Variance

To identify the principal components that capture the maximum variance in the data, we compute the eigenvalues associated with each eigenvector. The eigenvector corresponding to the highest eigenvalue represents the principal component with the most significant variance. By projecting all components onto this principal component, we effectively reduce the dimensionality of the dataset while retaining the most critical information. This process transforms the data into a lower-dimensional space, facilitating more efficient analysis and prediction.

10 Stock Price Prediction with PCA



```python
import numpy as np
import pandas as pd
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import yfinance as yf
import matplotlib.pyplot as plt

# Fetch historical stock data (example: Verastem Inc)
stock_symbol = 'VSTM'  # Change to your desired stock symbol
stock_data = yf.download(stock_symbol, start='2018-06-01', end='2018-12-31')

# Define the number of principal components to keep
n_components = 5

# Preprocessing: Extract relevant features and normalize
features = stock_data[['Open', 'High', 'Low', 'Close', 'Volume']].pct_change().dropna()
features = (features - features.mean()) / features.std()

# Apply PCA
pca = PCA(n_components=n_components)
pca.fit(features)
principal_components = pca.transform(features)

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(principal_components, stock_data['Close'].shift(-1).dropna(), test_size=0.2, random_state=42)

# Train a linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions
predictions = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, predictions)
print(f'Mean Squared Error: {mse}')

# Optional: Visualize actual vs. predicted values
plt.plot(y_test.index, y_test.values, label='Actual')
plt.plot(y_test.index, predictions, label='Predicted')
plt.xlabel('Date')
plt.ylabel('Stock Price')
plt.title(f'Stock Price Prediction for {stock_symbol} using PCA and Linear Regression')
plt.legend()
plt.show()
```



11 Dataset Explanation

The code starts by fetching historical stock data for Verastem Inc(an American pharmaceutical company that develops medicines to treat certain cancers) using the yfinance library. It then preprocesses the data by calculating the percentage change for the 'Open', 'High', 'Low', 'Close', and 'Volume' features and standardizing these features to have a mean of 0 and a standard deviation of 1. The code then applies Principal Component Analysis (PCA) to reduce the dimensionality of the data, keeping only the top 5 principal components. The data is then split into training and testing sets, and a linear regression model is trained on the training set.

12 Stock Price Prediction and Evaluation

The trained model predicts stock prices. Its performance is evaluated using Mean Squared Error and visualized by plotting actual vs predicted prices.

13 Summarizing the Project

We have successfully explored and demonstrated the efficacy of PCA in predicting stock prices, highlighting its role in mitigating dimensionality issues. We believe that our work can serve as a foundation for future research in this area.

Thank You

We extend our sincere gratitude for the opportunity to undertake this project and for your support and guidance throughout the process.