ADDIS ABABA SCIENCE AND TECHNOLOGY UNIVERSITY (AASTU)

COLLEGE OF ELECTRICAL AND MECHANICAL ENGINEERING

DEPARTMENT OF SOFTWARE ENGINEERING

INTRODUCTION TO MACHINE LEARNING (SWEG 4112)

STOCK PRICE PREDICTION REPORT

SECTION: E

Group Members		ID Number
1.	Samuel Kendie	ETS 1413/14
2.	Samuel Tarekegn	ETS 1420/14
3.	Selamawit Elias	ETS 1452/14
4.	Semira Hussien	ETS 1465/14
5.	Sifen Mekonnen	ETS 1474/14
6.	Yasub Demissie	ETS 1632/14
7.	Yeabtsega Tesfaye	ETS 1660/14

SUBMITTED TO: Instructor Tesfaye

SUBMISSION DATE: Tuesday, April 15, 2025

Stock Price Prediction Report

Introduction

This report outlines the analysis and prediction of Amazon's stock price based on historical data. The analysis employs regression techniques such as Linear Regression, Ridge Regression, and Support Vector Regression (SVR) to forecast future stock prices using lagged features. The workflow covers data preprocessing, model training, evaluation, and visualization of predictions.

Data Overview

The data set used in this analysis is a CSV file that contains historical stock price data for Amazon from May 15, 1995, till February 21,2025. The crucial columns include:

Date: The trading Date.

Open: The opening price of stock.

High: The highest price achieved by the stock during the day.

Low: The lowest price achieved by the stock during the day.

Close: The closing price of the stock.

Volume: Number of shares traded within the day.

The basic information obtained from the dataset includes 6987 entries within the given timeframe, seven data columns, float64(6) and int64(1) datatypes and 436.7+ KB memory usage.

Key Steps

1. Loading Data: The dataset is loaded using pandas and indexed by Date column.
2. Sorting: Data: is sorted chronologically to ensure proper time-series analysis.
3. Exploration: Display the first 5 rows, Dataset information and Descriptive statistics.

Data Processing

To predict future stock prices, lagged features are created based on previous closing prices:

1. Feature Engineering

Lagged features (Close_Lag_1, Close_Lag_2, Close_Lag_3) are generated using the shift() function. Rows with missing values (due to lagging) are dropped.

2. Feature Selection

Independent variables: Close_Lag_1, Close_Lag_2, Close_Lag_3.

Dependent variable: Close.

3. Train-Test Split

Data is split into training (80%) and testing (20%) sets without shuffling to preserve temporal order.

4. Scaling

Features (X) and target variable (y) are scaled using StandardScaler for better model performance.

Model Selection

Three regression models are employed:

Linear Regression: A simple model assuming a linear relationship between input features and target variable.

Ridge Regression: A regularized linear regression model that prevents overfitting by penalizing large coefficients.

Support Vector Regression (SVR): A kernel-based model that uses radial basis function (RBF) for non-linear relationships.

Model Training and Evaluation

Workflow:

1. Training: Each model is trained on scaled training data (X_train_scaled and y_train_scaled).
2. Saving Models: Trained models are serialized into .pkl files using the pickle library for future use.
3. Loading Models: Models are deserialized from .pkl files to demonstrate reusability.
4. Prediction: Predictions are made on the test set (X_test_scaled) and inverse-transformed back to original scale.
5. Evaluation Metrics

Mean Absolute Error (MAE): Measures average prediction error.

Mean Squared Error (MSE): Captures squared prediction error.

Root Mean Squared Error (RMSE): Square root of MSE for interpretability.

R^2 Score: Indicates how well the model explains variance in the target variable.

Results Comparison

A summary table is created using pandas.DataFrame to compare models based on evaluation metrics.

Regression Model	MAE	MSE	RMSE	R^2 Score
Linear Regression	2.198159	9.252202	3.041743	0.993061
Ridge Regression	2.264336	9.667681	3.109289	0.992750
Support Vector Regression	65.934603	7151.441554	84.566196	-4.363236

Visualization

Predicted vs actual stock prices are visualized using matplotlib.

The plot includes:

- Actual closing prices y_test.
- Predictions from each model.
- Features such as gridlines, labels, and legends enhance readability.

Conclusion

This analysis demonstrates how machine learning models can be applied to predict stock prices based on historical data. Key takeaways include:

- Feature engineering with lagged variables improves predictive performance.
- Scaling ensures compatibility with algorithms sensitive to feature magnitudes.
- Linear Regression outperformed both Ridge Regression and SVR as our dataset isn't complex. The highest R^2 value of 0. 993061 and minimal MAE value of 2.198159 is achieved by linear regression making it the best performing model.

Reference

[1] M. S. Ali, “Amazon Stocks 2025,” *Kaggle.com*, 2025. https://www.kaggle.com/datasets/meharshanali/amazon-stocks-2025 (accessed Apr. 14, 2025).