Final Project Report: AI Model Comparison

Course: Introduction to AI  
Project Title: Stock Price Predictor  
Submitted by:

* Cisco Dacanay
* Zach Greenhill

Date: 12/11/24

# 1. Introduction and Objective (5 Marks)

**Objective:**The objective of this project is to create several different predictive models for stock price forecasting to identify the most accurate approach for predicting future stock prices based on historical market data and other relevant features.

**Problem Statement:**

Accurately predicting stock prices can be difficult and inaccurate. Most prediction techniques focus on short-term predictions. Our goal is to try to predict the value of a given stock a month in the future. We will explore several different regression techniques to find methods to try to accurately predict stock prices.

**Overview of AI Models Chosen:**

|  |  |  |
| --- | --- | --- |
| **Model Number** | **Model Name** | **Purpose** |
| 1 | XGBoost | Uses gradient boosting to create an ensemble of decision trees. |
| 2 | Linear Regression | Fits a linear equation to predict values. |
| 3 | Neural Network | Used to find complex relationships in multi-dimensional data |
| 4 | Support Vector Regression (SVR) | Attempts to find a hyperplane that generalizes data. |
| 5 | Random Forest | Creates an ensemble of decision trees with bagging and random feature subsets. |

# 2. Justification of Model selection (2 Marks)

Justification for Model Selection:

|  |  |
| --- | --- |
| **Model Name** | **Reason for Selection** |
| XGBoost | Required |
| Linear Regression | Required |
| Neural Network | Required |
| Support Vector Regression | Good for regression tasks and avoids overfitting by attempting to create an optimal hyperplane. This model also requires few hyperparameters. |
| Random Forest | Good for tabular data and avoids overfitting by creating uncorrelated trees with randomly selected feature subsets. |

# 3. Model Descriptions (1 Marks)

Model Overview:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Number** | **Model Name** | **Architecture Details** | **Key Features** |
| 1 | XGBoost |  |  |
| 2 | Linear Regression |  |  |
| 3 | Neural Network |  |  |
| 4 | Support Vector Regression |  |  |
| 5 | [Model 5 Name] | Architecture details | Key features of Model 5 |

# 4. Dataset Description (2 Marks)

Dataset Information:

|  |  |
| --- | --- |
| Dataset Attribute | Description |
| Name | [Description] |
| Source | [Description] |
| Size | [Description] |
| Class Distribution | [Description] |
| Preprocessing Steps | [Description] |
|  |  |

Example

|  |  |
| --- | --- |
| Name | CIFAR-10 |
| Source | Available from https://www.cs.toronto.edu/~kriz/cifar.html |
| Size | 60,000 images (50,000 for training, 10,000 for testing) |
| Class Distribution | 10 classes, with 6,000 images per class |
| Preprocessing Steps | Normalization to range [0, 1], data augmentation (random cropping, horizontal flipping) |

**Dataset Justification:**  
Explain why this dataset is suitable for the models.

# 5. Experimental Setup (10 Marks)

Experimental Design: If any other metric is used, add it to the table

|  |
| --- |
| **Metric** |
| Mean Squared Error |
| Root Mean Squared Error |
| Mean Absolute Error |
| R2 Score |

Parameter Settings:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Name | Learning Rate | Epochs | Optional Hyperparameter | Optional Hyperparameter | Additional Details |
| XGB | 0.1 | 100 | Max Depth = 6 |  |  |
| Linear Regression | 0.001 | 120 | Batch Size = 128 |  |  |
| Neural Network | 0.0005 | 30 | Batch Size = 128 | 2 hidden layers | * 128 nodes at first hidden layer * 64 nodes at second hidden layer |
| SVR | N/A | N/A | C (Regularization Parameter) = 1.0 | Epsilon = 0.1 | Using a Linear kernel |
| Model 5 | Value | Value | Value | Value | Any other notes |

Environment Details:

|  |  |
| --- | --- |
| Component | Specification |
| Operating System | [Specification] |
| Software Version | [Specification] |
| Hardware | [Specification] |
| Link to the code base | https://github.com/GreenhillZachary/CS482\_Final\_Project |

# 6. Results and Analysis (50 Marks)

Performance Metrics:

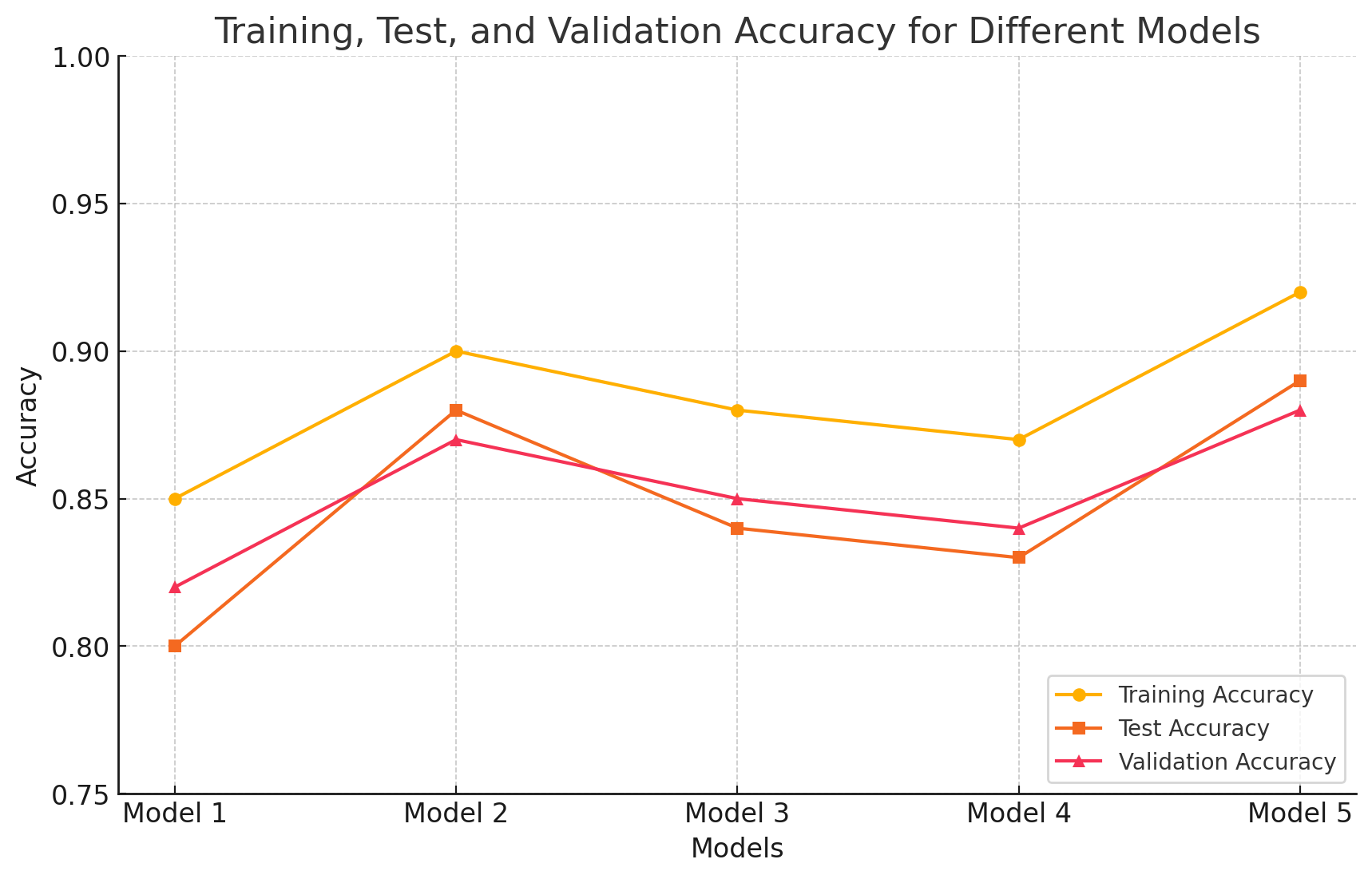
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Name | Mean Squared Error (MSE) | Root MSE | Mean Absolute Error | R2 Score |  |
| XGBoost | 38.2049 | 6.1810 | 3.3770 | 0.9846 |  |
| Linear Regression | 310.3066 | 17.6155 | 14.3498 | 0.8746 |  |
| Neural Network | 1088.0126 | 32.9850 | 24.4299 | 0.5604 |  |
| SVR | 62.6183 | 7.9132 | 4.1178 | 0.9747 |  |
| Random Forest | 31.9539 | 5.6528 | 3.2857 | 0.9871 |  |

Comparative Analysis:  
Summarize and analyze differences in performance.

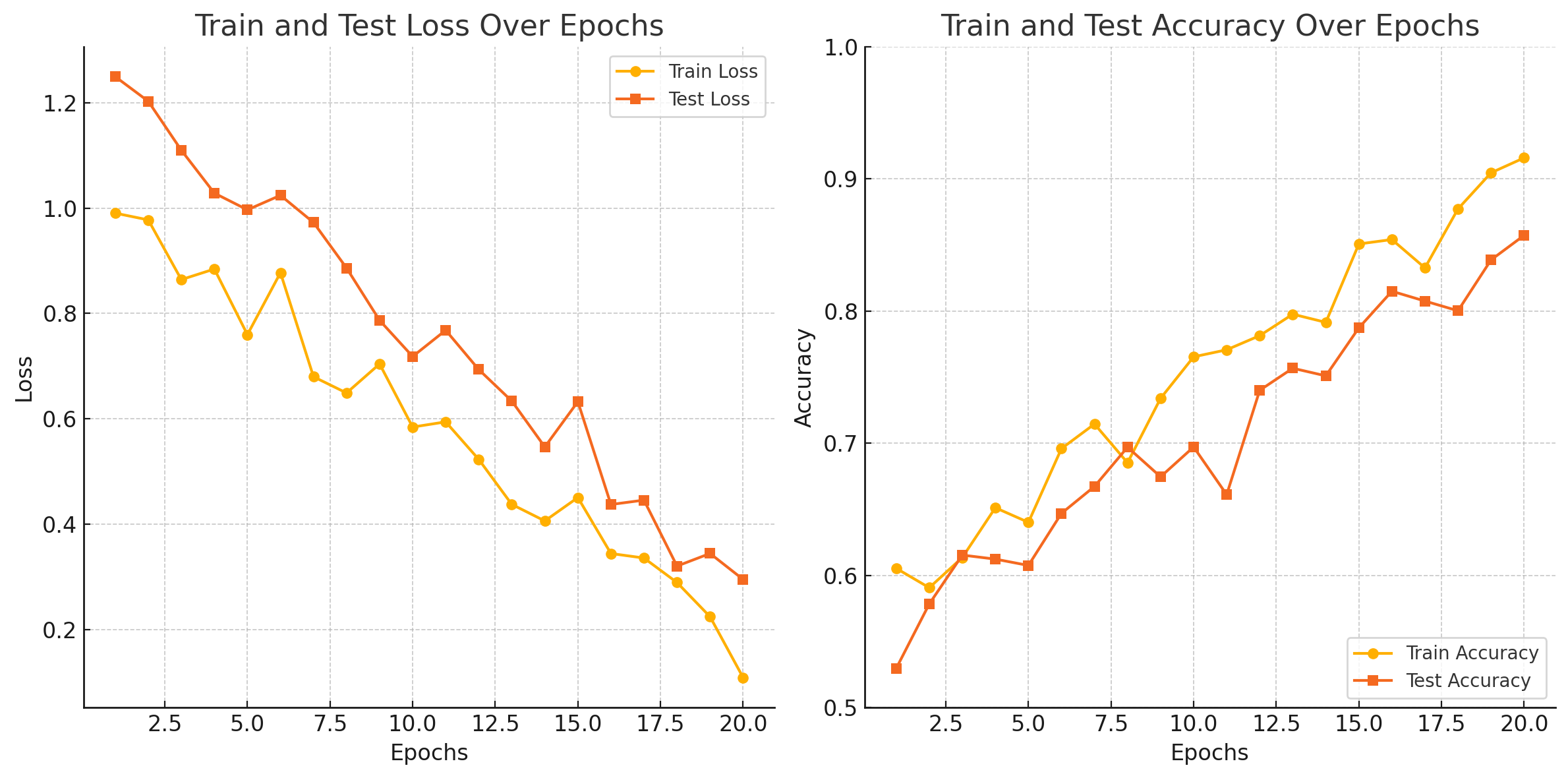
Error Analysis:  
Identify main sources of error and discuss possible reasons.

**Plots**

**- Training and Test accuracies**



**For neural network-based model, plot the following**



# 7. Discussion and Insights (10 Marks)

Interpretation of Results:  
Discuss key insights gained from the model comparison.

Limitations:  
Mention any limitations in the experimental design, data, or model performance.

Future Directions:  
Provide suggestions for further improvements or research.

# 8. Conclusion (5 Marks)

Summarize the project findings and the strengths and weaknesses of each model.