**DATASCIENCE CAPSTONE**

**FINAL REPORT**

**On**

**Stock predictions using machine learning and sentimental analysis.**

**By**

**Nihar Muniraju**

**Sushma Nagula**

1. **Using one single model-Random Forest**

**Introduction**

This report details the process and results of stock price prediction and classification using machine learning models. The focus was on five major stocks: Apple Inc. (AAPL), Google LLC (GOOGL), Microsoft Corporation (MSFT), Intel Corporation (INTC), and NVIDIA Corporation (NVDA). The goal was to predict the stock price movements and evaluate the classification performance of up and down movements based on historical data and technical indicators.

**Data Collection**

Historical stock price data was collected using the Yahoo Finance API for the period from January 2, 2015, to May 24, 2024. The data was preprocessed to ensure it only included working days, resulting in clean and consistent time series data for analysis.

**Technical Indicators**

Several technical indicators were calculated for each stock to serve as features for the models. These indicators include:

Simple Moving Average (SMA), Exponential Moving Average (EMA), Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI),Williams %R, Commodity Channel Index (CCI), Price Oscillator (PPO), Average True Range (ATR), Momentum, Rate of Change (ROC), Lag Close, Volume Price Trend (VPT), Accumulation/Distribution Line (ADL), Historical Volatility (Hist Vol). These indicators were calculated over various rolling windows to capture different aspects of stock price movements.

**Data Preprocessing**

The data was split into training and testing sets, with the training period ending on January 2, 2021, and the testing period starting on January 3, 2021, and ending on May 24, 2024. The features were standardized using StandardScaler to ensure the models were not biased due to different scales of features.

**Principal Component Analysis (PCA)**

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the feature space while retaining 95% of the variance. This step was crucial to avoid overfitting and improve model performance.

**Model Selection**

RandomForestRegressor

**Classification with SMOTE**

Synthetic Minority Over-sampling Technique (SMOTE) was applied to handle the class imbalance in the classification task. A RandomForestClassifier was used to predict the direction of stock price movements (up or down). The classification reports provide insights into the precision, recall, and F1-score for each class.

**Results**

Yet to put.

**Conclusion**

The RandomForestRegressor consistently performed well across all stocks, demonstrating its robustness in handling different stock price behaviors. The classification performance varied, with better accuracy on training data compared to test data, highlighting the challenges of predicting stock price movements accurately. The use of technical indicators and PCA for feature extraction helped in improving model performance, but further improvements could be achieved by exploring more advanced techniques and hyperparameter tuning.

1. **Using ensemble methods:**

**Introduction**

The goal of this project was to predict the percentage change in stock prices for five major tech companies: Apple (AAPL), Google (GOOGL), Microsoft (MSFT), Intel (INTC), and Nvidia (NVDA). The dataset spans from January 2, 2015, to May 24, 2024, and includes daily adjusted close prices. The ensemble approach combined various regression models to enhance prediction accuracy. Additionally, SMOTE (Synthetic Minority Over-sampling Technique) was used to address class imbalance in predicting positive and negative price changes.

**Data Collection and Preprocessing**

**Data Source:**

Yahoo Finance

**Stocks:**

Apple (AAPL), Google (GOOGL), Microsoft (MSFT), Intel (INTC), Nvidia (NVDA)

**Date Range:**

January 2, 2015, to May 24, 2024

**Features:**

Several technical indicators were calculated for each stock to serve as features for the models. These indicators include:

Simple Moving Average (SMA), Exponential Moving Average (EMA), Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI),Williams %R, Commodity Channel Index (CCI), Price Oscillator (PPO), Average True Range (ATR), Momentum, Rate of Change (ROC), Lag Close, Volume Price Trend (VPT), Accumulation/Distribution Line (ADL), Historical Volatility (Hist Vol). These indicators were calculated over various rolling windows to capture different aspects of stock price movements.

**Preprocessing Steps:**

**Feature Engineering:** Calculation of technical indicators.

**Normalization:** StandardScaler was used to normalize the features.

Handling Missing Values: Dropped rows with missing values after feature calculation.

Splitting Data: Divided into training (2015-01-02 to 2021-01-02) and testing (2022-01-03 to 2024-05-24) sets.

Class Imbalance Handling:

Used SMOTE to oversample the minority class in the training data.

**Ensemble Models Used**

The following regression models were included in the ensemble:

Gradient Boosting Regressor

Ridge Regression

Lasso Regression

ElasticNet Regression

RandomForest Regressor

Support Vector Regressor

Decision Tree Regressor

XGBoost Regressor

AdaBoost Regressor

The ensemble was created using the VotingRegressor from sklearn, which combines the predictions of the individual models.

**Model Training and Evaluation**

**Training:**

Models were trained on the training data.

SMOTE was applied to handle class imbalance for classification tasks.

GridSearchCV was used for hyperparameter tuning of the RandomForestClassifier used for classification.

**Evaluation Metrics:**

Root Mean Squared Error (RMSE)

Mean Absolute Error (MAE)

Classification Report (Precision, Recall, F1-Score)

Cross-Validation Scores

**Classification Report for SMOTE:**

AAPL (Apple Inc.)

Classification Report for Training Set:

precision recall f1-score support

False 0.63 0.78 0.69 531

True 0.84 0.71 0.77 853

accuracy 0.74 1384

macro avg 0.73 0.74 0.73 1384

weighted avg 0.76 0.74 0.74 1384

Classification Report for Test Set:

precision recall f1-score support

False 0.64 0.24 0.35 207

True 0.59 0.89 0.71 255

accuracy 0.60 462

macro avg 0.62 0.57 0.53 462

weighted avg 0.61 0.60 0.55 462

GOOGL (Google LLC)

Classification Report for Training Set:

precision recall f1-score support

False 0.69 0.72 0.70 566

True 0.80 0.77 0.79 818

accuracy 0.75 1384

macro avg 0.74 0.75 0.74 1384

weighted avg 0.75 0.75 0.75 1384

Classification Report for Test Set:

precision recall f1-score support

False 0.41 0.56 0.47 193

True 0.57 0.42 0.49 269

accuracy 0.48 462

macro avg 0.49 0.49 0.48 462

weighted avg 0.51 0.48 0.48 462

MSFT (Microsoft Corporation)

Classification Report for Training Set:

precision recall f1-score support

False 0.59 0.75 0.66 530

True 0.81 0.68 0.74 854

accuracy 0.71 1384

macro avg 0.70 0.71 0.70 1384

weighted avg 0.73 0.71 0.71 1384

Classification Report for Test Set:

precision recall f1-score support

False 0.42 0.47 0.45 184

True 0.62 0.58 0.60 278

accuracy 0.53 462

macro avg 0.52 0.52 0.52 462

weighted avg 0.54 0.53 0.54 462

INTC (Intel Corporation)

Classification Report for Training Set:

precision recall f1-score support

False 0.68 0.75 0.71 586

True 0.80 0.74 0.77 798

accuracy 0.74 1384

macro avg 0.74 0.74 0.74 1384

weighted avg 0.75 0.74 0.74 1384

Classification Report for Test Set:

precision recall f1-score support

False 0.56 0.36 0.44 244

True 0.49 0.68 0.57 218

accuracy 0.51 462

macro avg 0.52 0.52 0.50 462

weighted avg 0.53 0.51 0.50 462

NVDA (Nvidia Corporation)

Classification Report for Training Set:

precision recall f1-score support

False 0.62 0.72 0.67 523

True 0.81 0.73 0.77 861

accuracy 0.73 1384

macro avg 0.72 0.73 0.72 1384

weighted avg 0.74 0.73 0.73 1384

Classification Report for Test Set:

precision recall f1-score support

False 0.37 0.67 0.48 173

True 0.62 0.33 0.43 289

accuracy 0.46 462

macro avg 0.50 0.50 0.46 462

weighted avg 0.53 0.46 0.45 462

**Regression Model Performance**

AAPL (Apple Inc.)

Train RMSE: 0.0138

Test RMSE: 0.0580

Best Model: RandomForest with RMSE 0.0580

Worst Model: AdaBoost with RMSE 0.0974

GOOGL (Google LLC)

Train RMSE: 0.0128

Test RMSE: 0.0644

Best Model: RandomForest with RMSE 0.0644

Worst Model: SVR with RMSE 0.0917

MSFT (Microsoft Corporation)

Train RMSE: 0.0113

Test RMSE: 0.0540

Best Model: RandomForest with RMSE 0.0540

Worst Model: ElasticNet with RMSE 0.0835

INTC (Intel Corporation)

Train RMSE: 0.0127

Test RMSE: 0.0683

Best Model: RandomForest with RMSE 0.0683

Worst Model: AdaBoost with RMSE 0.1014

NVDA (Nvidia Corporation)

Train RMSE: 0.0226

Test RMSE: 0.1379

Best Model: RandomForest with RMSE 0.1379

Worst Model: SVR with RMSE 0.1926

**Plots:**

**Residual Analysis**

Residual analysis for each stock was conducted to examine the differences between the actual and predicted values. The residuals were plotted over time to identify any patterns or anomalies.

**Conclusion**

The ensemble approach effectively combined the strengths of individual models, providing a more robust and accurate prediction for stock price changes. The use of SMOTE helped in handling the class imbalance, improving the classification performance for predicting positive and negative price changes. Overall, the RandomForest model consistently performed well across all stocks, making it a reliable choice for this task.

Notebook: