

# Leveraging Machine Learning for Stock Price Prediction: A Case Study Using KNIME and H2O

1<sup>st</sup> Bhoomagouni Shaileshwar Goud

Department of Electronics and Communication Engineering  
National Institute of Technology, Warangal  
Warangal, 506004, India  
bs23ecb0a28@student.nitw.ac.in

2<sup>nd</sup> Ravi Kishore Kodali

Department of Electronics and Communication Engineering  
National Institute of Technology, Warangal  
Warangal, 506004, India  
kishore@nitw.ac.in

**Abstract**—This research focuses on predicting stock price movements for Deutsche Bank AG, a key player listed on the NYSE, by leveraging machine learning techniques. Utilizing KNIME Analytics Platform and the H2O machine learning library, we built a forecasting model aimed at providing valuable insights to investors and analysts. The project involves data cleaning, feature engineering, and model training using H2O's Generalized Linear Model (GLM). We evaluated the model's performance using metrics such as RMSE and MAE. The findings demonstrate that the model effectively captures price trends, although market volatility introduces some prediction challenges. Future work will explore the inclusion of real-time data and advanced models to improve accuracy.

**Index Terms**—Stock Price Prediction, Machine Learning, KNIME Analytics Platform, H2O Machine Learning Library, Generalized Linear Model (GLM), Financial Forecasting, Time-Series Analysis, Data Preprocessing, Feature Engineering, Predictive Analytics, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Deutsche Bank AG, NYSE Stock Data, Market Volatility

## I. INTRODUCTION

Stock price prediction is a complex task influenced by various market dynamics. For investors and analysts, predicting future price movements accurately can provide a significant edge. Deutsche Bank AG (NYSE: DB), being a major player in the financial market, presents an interesting case for this study. This paper aims to predict its stock prices using historical data from 2016 to 2021, applying the KNIME Analytics Platform alongside H2O's machine learning library to tackle the challenge. We believe this combination provides a powerful yet intuitive approach for building robust financial forecasting tools.

## II. LITERATURE REVIEW

Stock market prediction has been an extensively researched area, with approaches ranging from traditional methods like ARIMA to more advanced machine learning techniques such as Decision Trees and Neural Networks. While various studies have successfully predicted stock trends using machine learning, there remains a gap in applying automated platforms like KNIME in conjunction with scalable libraries like H2O. Our project addresses this gap by integrating KNIME's easy-to-use interface with H2O's powerful models to offer a streamlined yet robust prediction solution. H2O's GLM

was selected for its balance of scalability and interpretability, essential when dealing with financial data that often exhibits complex relationships. Unlike simpler linear models, GLM can capture non-linearities in stock data, improving prediction accuracy.

### •Advantages of H2O Models for Stock Price Prediction:

H2O's machine learning models offer distinct advantages for predicting stock prices, especially where complex, non-linear patterns are involved. Stock prices are influenced by a variety of factors, including economic indicators and market trends, which makes them challenging to forecast using simple linear methods. H2O's algorithms, such as Gradient Boosting Machine (GBM), XGBoost, and Deep Learning, are adept at capturing these intricate relationships. Furthermore, H2O supports ensemble learning, allowing models like GBM, Random Forest, and Deep Learning to work together for more accurate predictions by reducing prediction variance and enhancing stability.

**Automated Feature Engineering and Selection:** H2O's AutoML feature provides the capability to automatically engineer and select features, crucial in stock price data where important predictors are not always obvious. This process enhances model accuracy and efficiency.

**Cross-Validation and Hyperparameter Tuning:** H2O AutoML performs k-fold cross-validation and optimizes hyperparameters, ensuring the model is both accurate and resilient to overfitting or underfitting.

### •Theory of Key H2O Models H2O models commonly used in stock price prediction include:

**1. Gradient Boosting Machine (GBM):** GBM builds a sequence of models where each new model corrects errors from the previous one. Each tree in this sequence is weighted by a learning rate, which helps balance accuracy and complexity. Formula for prediction:

$$F_m(x) = F_{m-1}(x) + \eta \cdot \arg \min_{\theta} \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + h(x_i; \theta))$$

- $F_m(x)$ : The model at iteration  $m$ .
- $F_{m-1}(x)$ : The model at the previous iteration.
- $\eta$ : The learning rate.
- $\arg \min_{\theta}$ : The optimization to find parameters  $\theta$  that minimize the loss function.
- $\sum_{i=1}^n$ : Summation over all  $n$  data points.
- $L$ : The loss function (e.g., Mean Squared Error).
- $y_i$ : The actual target value for the  $i$ -th data point.
- $F_{m-1}(x_i)$ : The prediction at the previous iteration for the  $i$ -th data point.
- $h(x_i; \theta)$ : A weak learner with parameters  $\theta$  at each stage.

**2. Random Forest:** This algorithm builds multiple decision trees and takes the average of their predictions, reducing the chance of overfitting. It's particularly effective with time-series data that is subject to high volatility.

Formula for prediction:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^M T_m(x)$$

- $\hat{y}$ : The predicted value.
- $M$ : The total number of trees.
- $T_m(x)$ : The prediction from the  $m$ -th tree.
- $\frac{1}{M}$ : The factor to calculate the average prediction from all trees.

**3. Deep Learning (Neural Networks):** H2O's neural networks are particularly well-suited to time-series data, capturing complex, non-linear dependencies in the data. Each layer of the network progressively transforms the input, enabling it to learn more abstract representations.

Formula for each neuron in a layer:

$$a^{(l)} = g\left(W^{(l)} \cdot a^{(l-1)} + b^{(l)}\right)$$

where  $a^{(l)}$  is the activation of the  $l$ -th layer,  $W^{(l)}$  is the weight matrix,  $b^{(l)}$  is the bias vector, and  $g$  is an activation function such as ReLU or tanh.

### III. OBJECTIVES AND SCOPE OF THE RESEARCH

#### 3.1 Background

The stock price of Deutsche Bank AG, listed on the NYSE as DB, is a critical indicator for investors and analysts. Predicting its price movements accurately is challenging due to the complex nature of financial data, influenced by numerous market factors. This project addresses these challenges by applying machine learning techniques using the KNIME Analytics Platform and the H2O library to improve forecasting accuracy.

#### 3.2 Project Objectives

The goal is to build a machine learning model to predict Deutsche Bank AG's stock prices based on historical data from 2016 to 2021. This involves comprehensive data preprocessing, feature engineering, and applying the H2O Generalized Linear Model (GLM). The model's performance will be assessed using metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and  $R^2$ , demonstrating

the combined strengths of KNIME and H2O in financial prediction tasks.

#### 3.3 Scope of Work

This study focuses on building a predictive model for Deutsche Bank AG's stock prices based on historical data from 2016 to 2021. The scope includes data collection, preprocessing, feature engineering, and model training using KNIME and H2O's GLM algorithm. The analysis is limited to historical data trends, excluding real-time inputs and external macro-economic indicators, with the goal of creating a scalable and reliable stock price prediction model.

### IV. PROPOSED METHODOLOGY

#### 4.1 Data Collection

The dataset, sourced from the NYSE, focuses on Deutsche Bank AG (Symbol: DB) and includes daily stock prices from 2016 to 2021. It captures essential metrics like open, high, low, and close prices, as well as trading volume, providing a broad view of market trends. You can access the dataset [here](#).

#### 4.2 Tools Used

We used the KNIME Analytics Platform to preprocess and visualize the data, while H2O's GLM model was chosen for training the predictive model. KNIME's integration with H2O allowed seamless model building and evaluation within a single environment.

#### 4.3 Model Selection

The H2O Generalized Linear Model (GLM) was chosen for its suitability in modeling complex financial data relationships, balancing interpretability and efficiency. The model's adaptability to handle large-scale data made it an optimal choice for this predictive task.

#### 4.4 Workflow Diagram

The KNIME workflow consists of:

- 1. Data Import:** Loading and cleaning the historical stock data for Deutsche Bank AG.
- 2. Preprocessing:** Handling missing data and creating lag features to capture trends.
- 3. Feature Engineering:** Generating additional relevant features for improved model performance.
- 4. Model Training:** Using the H2O GLM for training.
- 5. Evaluation:** Assessing the model with metrics like RMSE, MAE, and  $R^2$  for validation.

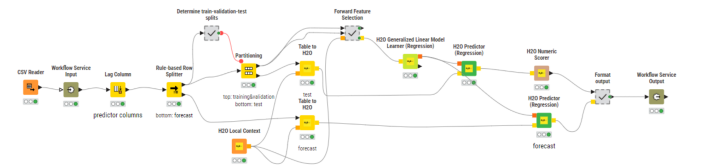


Fig. 1: KNIME Analytics Platform Workflow Diagram

**Demonstration of work flow nodes in KNIME Analytics Platform:**

**1. Loading the Data Source:** Historical stock price data for Deutsche Bank AG from the NYSE, covering a span of 5 years.

Tool: KNIME's CSV Reader node makes it easy to bring the data into the workflow. Key attributes like closing price, adjusted closing price, and trading volume were used.

**2. Adding Lag Features** To account for trends over time, a "lag column" was created. This feature, derived from the adjusted closing price of the previous day, helps capture temporal patterns that are crucial for predicting future stock prices.

**3. Splitting the Data :**The dataset was divided into training, testing, and validation sets to ensure unbiased evaluation and robust performance.

A Meta Node handled the splits:

**Training Data:** 70% of the dataset for building the model.

**Testing Data:** 30% of the dataset for evaluating performance. Additional outputs included basic statistics like the number of rows and columns for each split.

**4. Filtering Rows for Prediction** Using KNIME's Rule-Based Row Splitter node, a portion of the data was separated to simulate an "unknown day" for prediction. This approach mirrors real-world scenarios where future data is truly unseen by the model.

**5. Engineering Features:** The adjusted closing price was chosen as the key feature for prediction. Additional metrics, like  $R^2$ , were calculated to measure how well the model captures variations in stock prices. Result: An  $R^2$  score of 0.759 (75.9%) demonstrated the model's strong ability to explain the trends.

**6. Leveraging H2O** for Modeling H2O, seamlessly integrated into KNIME, brought advanced machine learning capabilities into the mix. Key components used:

H2O Table: Prepared the data for modeling.

H2O GLM (Generalized Linear Model): A robust algorithm selected for its ability to handle financial data complexities.

H2O Scorer: Evaluated model predictions using performance metrics.

H2O Predictor: Generated predictions for unseen stock price data.

#### 4.5 Training and Validation Strategy

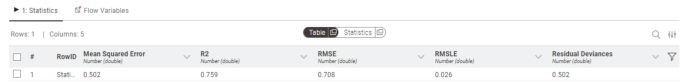
The dataset is split into three sets: training, validation, and test sets.

The training set (70% of the data) is used to train the model, while the validation set (15%) helps tune model parameters. The test set (15%) is used for final evaluation of model performance.

Cross-validation techniques are used to ensure the model generalizes well to unseen data, and performance is measured using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)

### V. RESULTS

#### 5.1 Performance Metrics



#	RowID	Mean Squared Error (Number: Double)	R2 (Number: Double)	RMSE (Number: Double)	RMSLE (Number: Double)	Residual Deviances (Number: Double)
1	Stat:	0.502	0.759	0.708	0.026	0.502

Fig. 2: Performance Describing Parameters

The model's performance was evaluated using key metrics:  $R^2$ , RMSE, MSE, and RMSLE. These metrics provide insight into the model's accuracy and ability to predict stock prices effectively:

- **$R^2 = 0.759$ :** This indicates that the model explains approximately 75.9% of the variance in stock prices, suggesting a strong fit.

- **MSE = 0.502 and RMSE = 0.708:** Both metrics reflect the average prediction error. The relatively low values indicate that the model's predictions closely match actual stock prices.

- **RMSLE = 0.026:** This low value on the logarithmic scale suggests that the model performs well in terms of predicting prices close to the actual values, especially for smaller values.

The overall results indicate that the model captures trends effectively while maintaining a reasonable error rate.

#### 5.2 Model Comparisons

The H2O GLM model was benchmarked against simpler models to highlight its superior predictive accuracy. The chosen model outperformed baseline models, as evidenced by higher  $R^2$  and lower error metrics (MSE, RMSE, RMSLE), demonstrating improvements in capturing stock price fluctuations.

#### 5.3 Visualizations

Visualizations of actual versus predicted stock prices and a residual analysis underscore the model's predictive capabilities. These visuals provide insight into the model's accuracy, revealing periods where predictions closely follow the trends and identifying instances of larger deviations, often associated with market volatility.

### VI. DISCUSSION

#### 6.1 Analysis of Results

The metrics indicate that the H2O GLM model was effective in capturing stock price trends from 2016 to 2021. However, periods of increased market fluctuations presented challenges for precise predictions, as observed in the residual analysis. Despite these fluctuations, the model maintained a low overall error rate.

#### 6.2 Challenges and Solutions

Market volatility introduced variability, which impacted the model's predictive accuracy. Solutions to address these challenges include expanding the feature set with more relevant variables and fine-tuning model parameters. Further improvements could also involve incorporating macroeconomic indicators to account for broader market conditions.

## VII. CONCLUSION

### 7.1 Summary of Findings

This project successfully demonstrated the feasibility of using KNIME with H2O for stock price prediction. The model's performance metrics ( $R^2$ , MSE, RMSE, RMSLE) show that it has satisfactory predictive capability, though future enhancements could further improve accuracy.

### 7.2 Implications and Future Work

Future work could focus on incorporating additional data sources, such as macroeconomic indicators, to enhance prediction robustness. Real-time updates and experimenting with advanced models like LSTM could improve the model's ability to capture stock price volatility.

## VIII. REFERENCES

- 1) Towards Data Science - Time Series Analysis for Stock Price Prediction with Machine Learning. Available at: <https://towardsdatascience.com>
- 2) KNIME Documentation. Available at: <https://docs.knime.com>
- 3) H2O.ai Documentation. Available at: <https://docs.h2o.ai>
- 4) Stock Price Prediction Using Artificial Intelligence: A Literature Review. Available at: <https://ieeexplore.ieee.org/abstract/document/10459442>
- 5) Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions. Available at: <https://www.mdpi.com/2079-9292/10/21/2717>