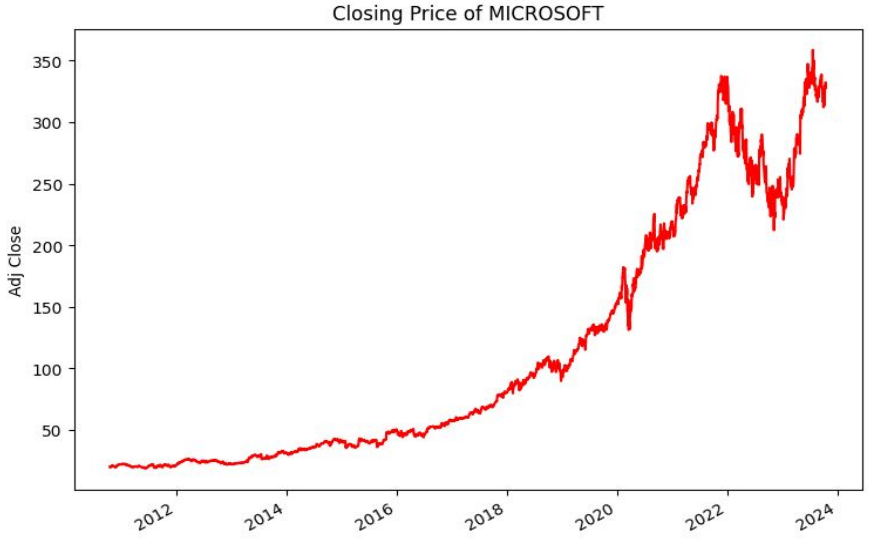
CSE 343 : ML Project Monsoon 2023

Aryan Dhull- 2021520
Deepanshu- 2021524
Pranav Aggarwal- 2021551
Prerak Gupta- 2021552

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Finance Forecaster:

Using Machine Learning to predict stock prices



Motivation

- Precise stock price prediction remains a challenging yet highly valuable venture in today's fast-paced financial markets.

- As traders and investors seek to make knowledgeable decisions, the ability to forecast the ups and downs of the market plays an important role.

- Eventually, successful price prediction could certify market participants with magnified insights, likely reducing risks and optimizing investments in an ever-evolving economic landscape.

Motivation

- This project aims to grasp algorithms and data analysis techniques to unlock patterns hidden within historical stock data.

- Machine learning techniques, such as time series forecasting and regression can be applied to historical stock market data to predict future price movements.

- Developing accurate predictive models can be precious for traders, investors, and financial institutions looking to make informed decisions.

Literature Review

1. Research on stock trend prediction method based on optimized random forest [1]

Authors– Lili Yin, Benling Li, Peng Li, Rubo Zhang

- This paper proposes a method for stock trend prediction based on optimized random forest.
- It emphasis the use of exponential smoothing to optimise the decisions made by random forest which involves calculating the average of a set of data points within a specified window or time period.
- The main advantage of using this method is to smooth out the fluctuations and address both recent and long term trends.

Literature Review

2. Stock Price Prediction Using the ARIMA Model [2]

Authors- Ayodele A. Adebiyi, Aderemi O. Adewumi, Charles K. Ayo

- The literature reveals that the ARIMA model is widely used in stock price prediction due to its ability to handle time series data effectively.

- Results obtained revealed that the ARIMA model has a strong potential for short-term prediction and can compete favourably with existing techniques for stock price prediction.

- ANN models which are very popular due to its ability to learn patterns from data and infer solution from unknown data are used to improve the performance of the predictions.

Dataset

We've harnessed yfinance, a library that taps into the Yahoo Finance API, for real-time stock price data. It's a crucial asset for making informed financial decisions.

Our dataset spans from 2010 to the present day, and it revolves around four technology giants: Google, Apple, Microsoft, and Amazon. These companies form the focal point of our analysis.

Dataset

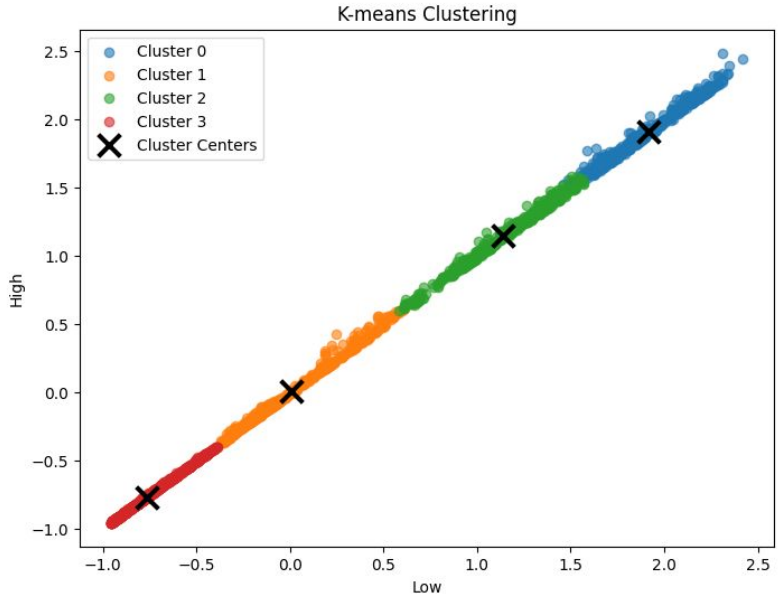
After augmenting the dataset with necessary features and conducting a comprehensive analysis of the return and risk metrics for each company, we have successfully compiled a refined dataset specifically focused on Microsoft. This dataset comprises 14 key features and encompasses over 3000 entries, offering a detailed perspective on Microsoft's stock prices.

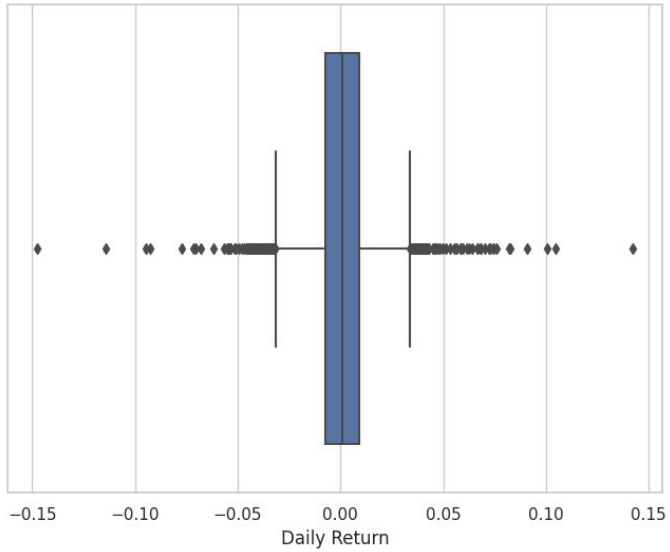
Date	Open	High	Low	Close	Adj Close	Volume	MA for 10 days	MA for 20 days	MA for 50 days	Daily Return	Close1	Close_2	Close_3	Next_Day_Closing_Price
2010-12-28	27.969999	28.170000	27.959999	28.010000	21.731487	23042200	21.708984	21.243864	20.579771	-0.002137	28.010000	28.070000	28.299999	27.969999
2010-12-29	27.940001	28.120001	27.879999	27.969999	21.700445	19502500	21.736138	21.348991	20.615578	-0.001428	27.969999	28.010000	28.070000	27.850000
2010-12-30	27.920000	28.000000	27.780001	27.850000	21.607349	20786100	21.736138	21.419206	20.660629	-0.004290	27.850000	27.969999	28.010000	27.910000
2010-12-31	27.799999	27.920000	27.629999	27.910000	21.653900	24752000	21.729932	21.458774	20.703371	0.002154	27.910000	27.850000	27.969999	27.980000
2011-01-03	28.049999	28.180000	27.920000	27.980000	21.708214	53443800	21.736139	21.496015	20.745503	0.002508	27.980000	27.910000	27.850000	28.090000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2023-10-06	316.549988	329.190002	316.299988	327.260010	327.260010	25645500	317.263004	323.117001	325.002360	0.024737	327.260010	319.359985	318.959991	329.820007
2023-10-09	324.750000	330.299988	323.179993	329.820007	329.820007	19891200	318.491003	322.711002	324.845659	0.007823	329.820007	327.260010	319.359985	328.390015
2023-10-10	330.959991	331.100006	327.670013	328.390015	328.390015	20557100	320.116003	322.542003	324.709253	-0.004336	328.390015	329.820007	327.260010	332.420013
2023-10-11	331.209991	332.820007	329.140015	332.420013	332.420013	20063200	322.079004	322.360004	324.645065	0.012272	332.420013	328.390015	329.820007	331.160004
2023-10-12	330.570007	333.630005	328.720001	331.160004	331.160004	19313100	323.831003	321.983003	324.732104	-0.003790	331.160004	332.420013	328.390015	327.730011

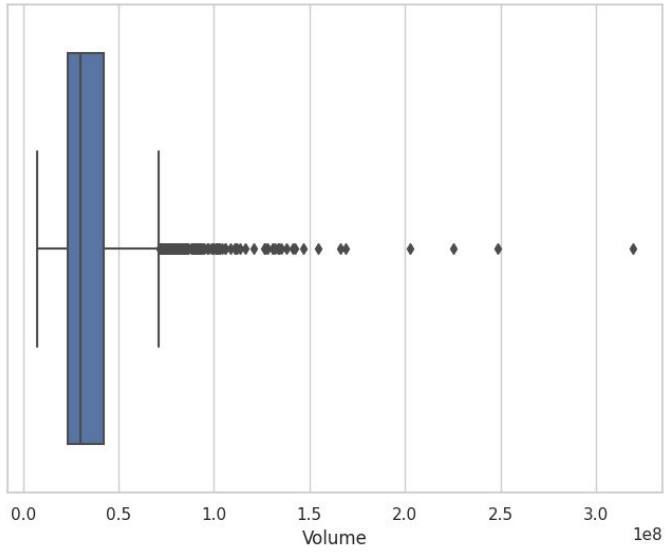
3220 rows × 14 columns

Dataset

We standardized the dataset to ensure uniformity and then applied K-means clustering and boxplot analysis to identify outliers. Among the 14 features, outliers were detected specifically in the 'Daily Return' and 'Volume' variables, indicating deviations from the general patterns observed in these two aspects of the data.







Methodology

- To ensure a realistic evaluation of our models on time series data, we've implemented a chronological train-test split with a **70:30** ratio.

- Our Next objective is to use a baseline **linear regression** model to predict Microsoft's (MSFT) next day's closing stock price based on historical data. We compute performance metrics, such as **Mean Squared Error (MSE)** and **R-squared (R2)**. All input features were modified so as to have the same scale.

- We then applied **Support Vector Regression (SVR)** to capture both linear and non-linear relationships and to enhance robustness against outliers in financial forecasting.

Methodology

- Our experiments on historical stock price data involved using SVR to model the relationship between various features and stock prices. Results showed SVR to be effective, particularly with a linear kernel.

- We also delved into **Lasso** and **Ridge** Regularization to prevent overfitting and enhance model generalization. Lasso introduced sparsity, emphasizing key features, while Ridge mitigated multicollinearity risks.

- Fine-tuning the hyperparameter **alpha** was achieved through Cross Validation techniques.

Methodology

- **Decision Tree Regressor** conducts a hyperparameter search using GridSearchCV, exploring various combinations of 'max_depth', 'min_samples_split', 'min_samples_leaf', and 'max_features'. The best-performing model is then selected and used to make predictions on the test set.

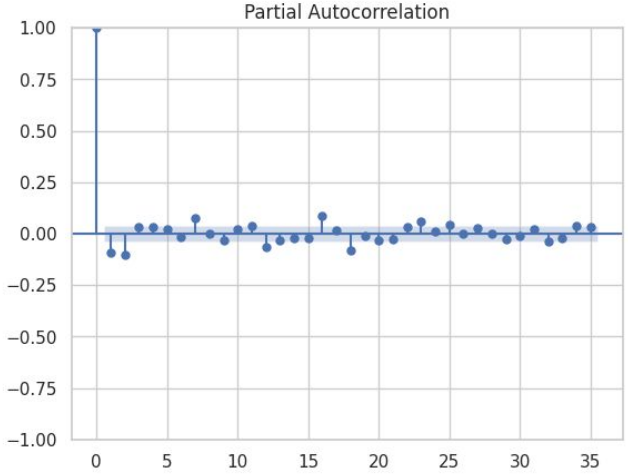
- **Random Forest** splits the data into training and testing sets, trains the model, and evaluates its accuracy, enhancing predictions through ensemble learning and parameter tuning.

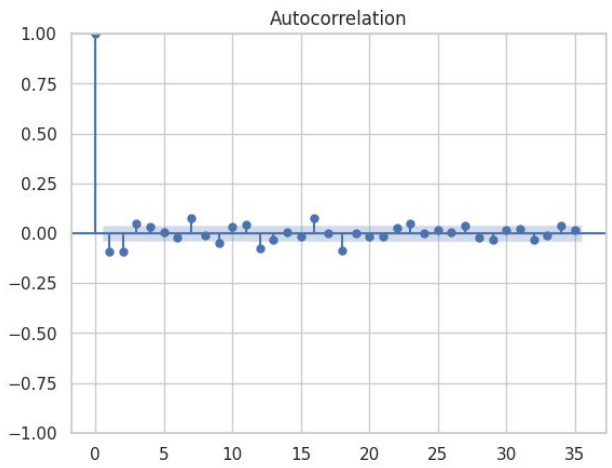
- **ANN** was applied to the dataset with 6 hidden layers using all of the features. It was observed that the network started to overfit at higher number of hidden layers.

- Different activation functions were explored and finally '**relu**' was used. **Dropout regularization** was also performed to prevent the dataset from overfitting.

Methodology

- Further, we explored the Autoregressive Integrated Moving Average (**ARIMA**) model, which is a technique applied to time series data. The three parameters p (order of the autoregressive term), d (order of the differencing) and q (order of the moving average term) were initialized to 0, 1 and 1 respectively after analyzing the Autocorrelation and Partial Autocorrelation graphs.





- Here, p refers to the the number of lagged values of the time series that are used to predict the current value. q refers to the error terms of the model that are lagged. d refers to the the degree of differencing, that is, the number of times we have to differentiate the time series to make it stationary.

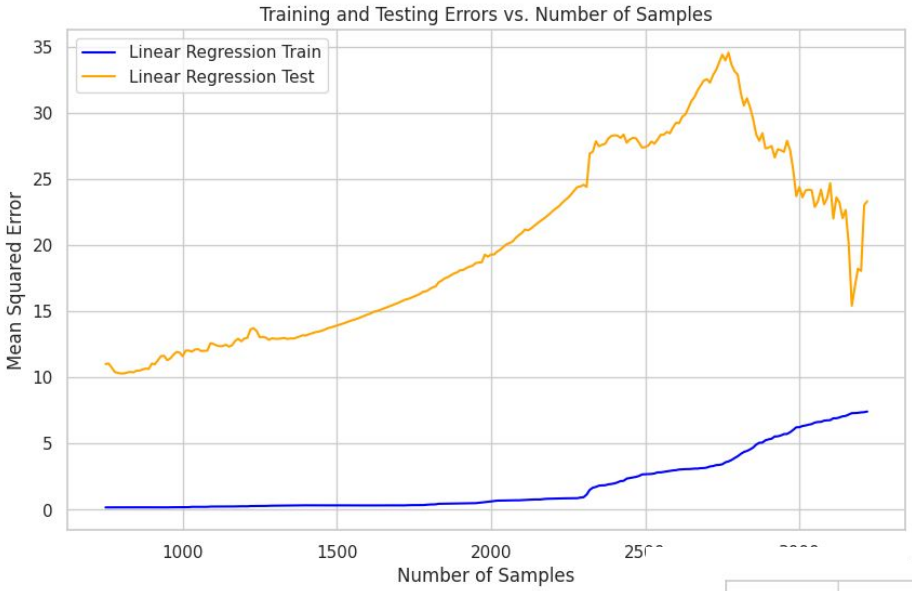
Result & Analysis

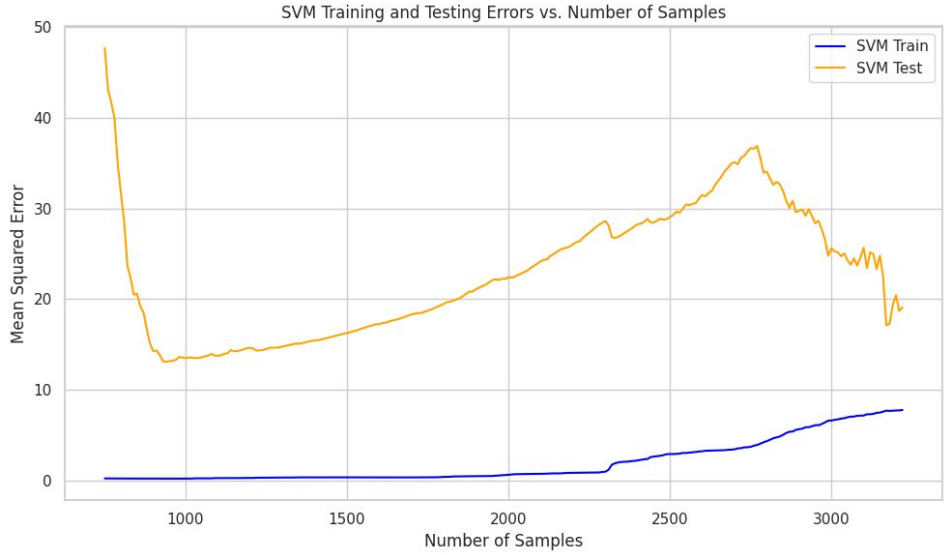
- **Linear Regression** performed quite well because of the data being heavily linearly related with an **MSE** of **23.63** and **R2 score** of **0.99**.

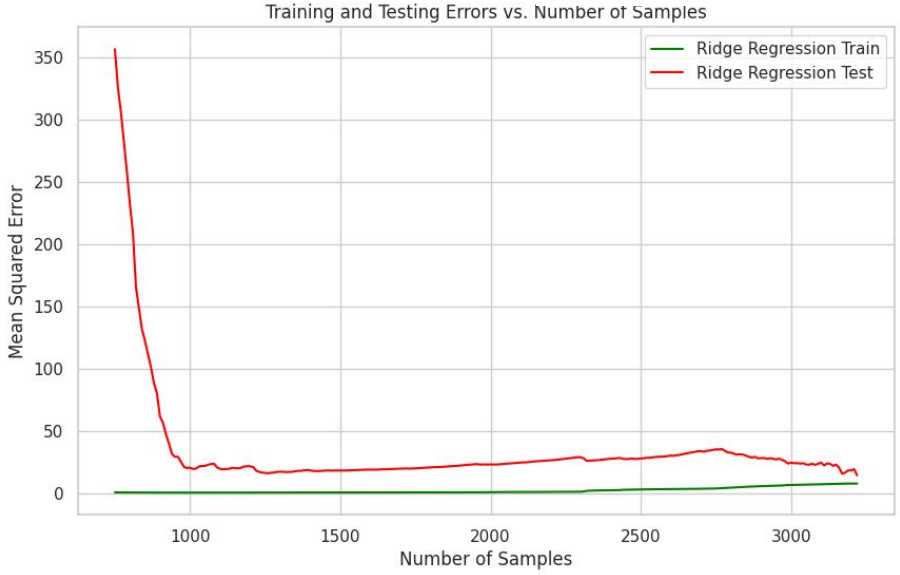
- **Support Vector Regression** also gave similar results with an **MSE** of **27.35** and **R2** score of **0.99**. Both the models showed extremely accurate predictions on the training set signifying overfitting and higher variance.

- Applying **Lasso** and **Ridge** Regularization with a hypertuned alpha led to improved observations of bias and variance but some amount of overfitting persisted.

Result & Analysis



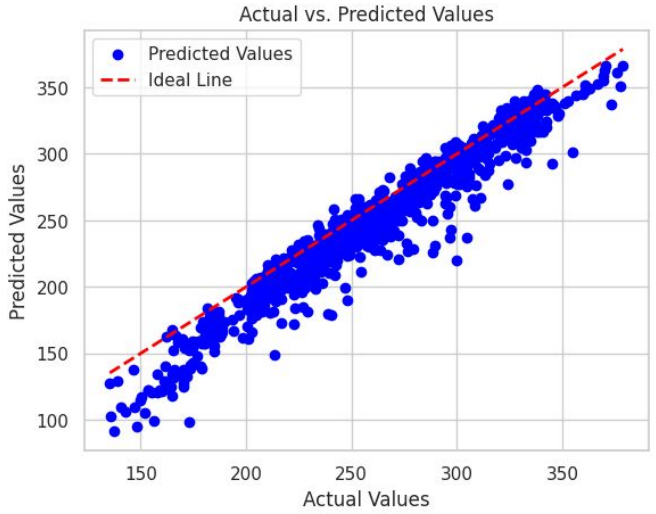


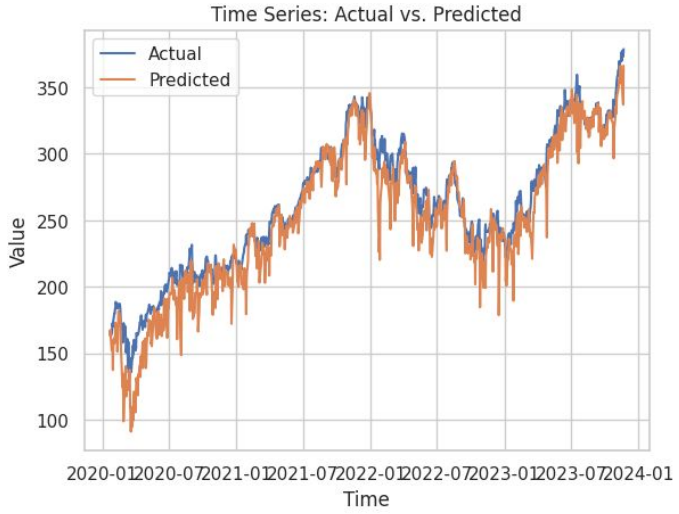


Results and Analysis

- Decision Tree and Random Forest did not perform to the expectations, having overfit extremely. Pruning and hyperparameter tuning using GridSearchCV was tried but it did not give promising results with the mean squared error being 12000 compared to a very low MSE for training data.

- ANNs, on the other hand, performed extremely well with a training MSE 34.78, testing MSE 308.5 and R2 Score 0.88. It was also observed that the model captured the trends of the stock fluctuations.

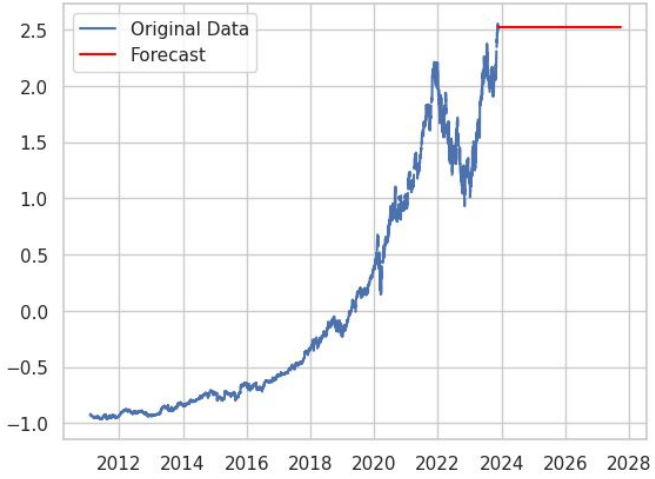




Results and Analysis

- The ARIMA Model was able to extensively learn the patterns in the stock data.

- But, it was observed unable to predict a good forecast on that testing data, the step forecast kept following the last forecast value (a horizontal line).



Contributions

Aryan Dhull	Data Collection, Exploratory Data Analysis, Data Preprocessing, K-Means, ANN, Report + Presentation
Deepanshu	Data Collection, Data Preprocessing, Linear Regression, SVR, Smoothing, Decision Tree, Report + Presentation
Pranav Aggarwal	Exploratory Data Analysis, K-Means, Naive-Bayes, Smoothing, Random Forest , Report + Presentation
Prerak Gupta	Data Collection, Linear Regression, SVR, ARIMA, Report + Presentation

References

- https://ieeexplore.ieee.org/document/7046047
- http://www.iaeng.org/publication/IMECS2009/IMECS2009_pp755-760.pdf
- https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/cit2.12067
- https://iopscience.iop.org/article/10.1088/1742-6596/1988/1/012041/meta
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7046047

QnA!