Rasha Alharthi
Nov 18 · 9 min read · Listen

Analyzing the Saudi Stock Exchange (Tadawul) with Machine Learning

SAUDI EXCHANGE BRAND LOGO

Content

1. Introduction

There has been much written about the possibilities of data science; it is a combination of various tools, algorithms, and machine learning principles to discover hidden patterns in raw data, and it is being used in many businesses to help them save money and increase revenue, including stock markets, where it provided an in-depth understanding of the stock market and financial statistics, as well as forecasting future events, which is a critical input into many types of planning and decision-making.

Saudi's 2030 Vision and Tadawul:

As part of achieving Saudi Arabia Vision 2030, Saudi Arabia is working to make its financial market the primary regional market and one of the 10 most major markets in the world through the **Financial Sector Development Program**, which focuses on increasing the size, depth, and development of Saudi capital markets so that they become advanced markets and attract local and foreign investment, allowing them to play a crucial role in developing the national economy and expanding its sources of income.

Saudi Tadawul Group which was established in March 2021, works closely with all interested parties and investors to develop the market by increasing market liquidity, local investor confidence, access to the market, efficiency, and level of transparency, and attracting foreign investment to open up to

various segments locally, regionally, and globally for what “Tadawul” represents in terms of volume and liquidity in the international Saudi financial market.

Business problem:

In this project, we analyze and study the “Tadawul” dataset from various aspects to determine the factors that affect the local trading market and use them to predict future behavior that will help Tadawul in making decisions.

Project goal:

We have two different goals for analyzing the dataset, first, the regression model predicts the close price, and the other one is the classification model to predict the change class if it’s a good change, a bad change, or stable.

> *Regression model target:*

The Closing Price helps the investor understand the market sentiment of the stocks over time. It is the most accurate matrix to determine the valuation of stock until the market resumes trading the next day.

> *Classification model target:*

The target here is the change category, the change term refers to the difference between a stock’s closing price on a trading day and its closing price on the previous trading day. The change is termed a positive change when the initial value is lower than the final value, and negative if the end value is lower than the initial value.

2. Data Review

This is the data of Saudi stock market companies from 2001–12–31 until 2020–04–16. It was collected from Saudi Stock Exchange (Tadawul) website.

The dataset link: https://www.kaggle.com/datasets/salwaalzahrani/saudi-stock-exchange-tadawul

> *The data contains 593819 rows and 14 columns.*

Meaning of each column (each row in the database represents the price of a specific stock at a specific date):

> *Symbol (int64): reference number of the company*
>
> *Name (object): Name of the company*
>
> *Trading_name (object): The trading name of the company*
>
> *Sectoer (object): The sector in which the company operates*
>
> *Date (object): The date of the stock price*
>
> *Open (float64): The opening price*
>
> *High (float64): The highest price of the stock on that day*
>
> *Low (float64): The lowest price of the stock on that day*
>
> *Close (float64): The closing price*
>
> *Change (float64): The change in price from the last day*
>
> *Perc_Change (float64): The percentage of the change*

> *Volume_traded (float64): The volume of the trades for the day*
>
> *Value_traded (float64): The value of the trades for the day*
>
> *No_trades (float64): The number of trades for the day*

3. Data Preprocessing

Here we will not include all the preprocessing steps, we will mention only the important step that we used in the following sections.

We added a new column from the per_change column in the dataset, which is the Change_category (the classification target).

We create the classes from the per_change as below:

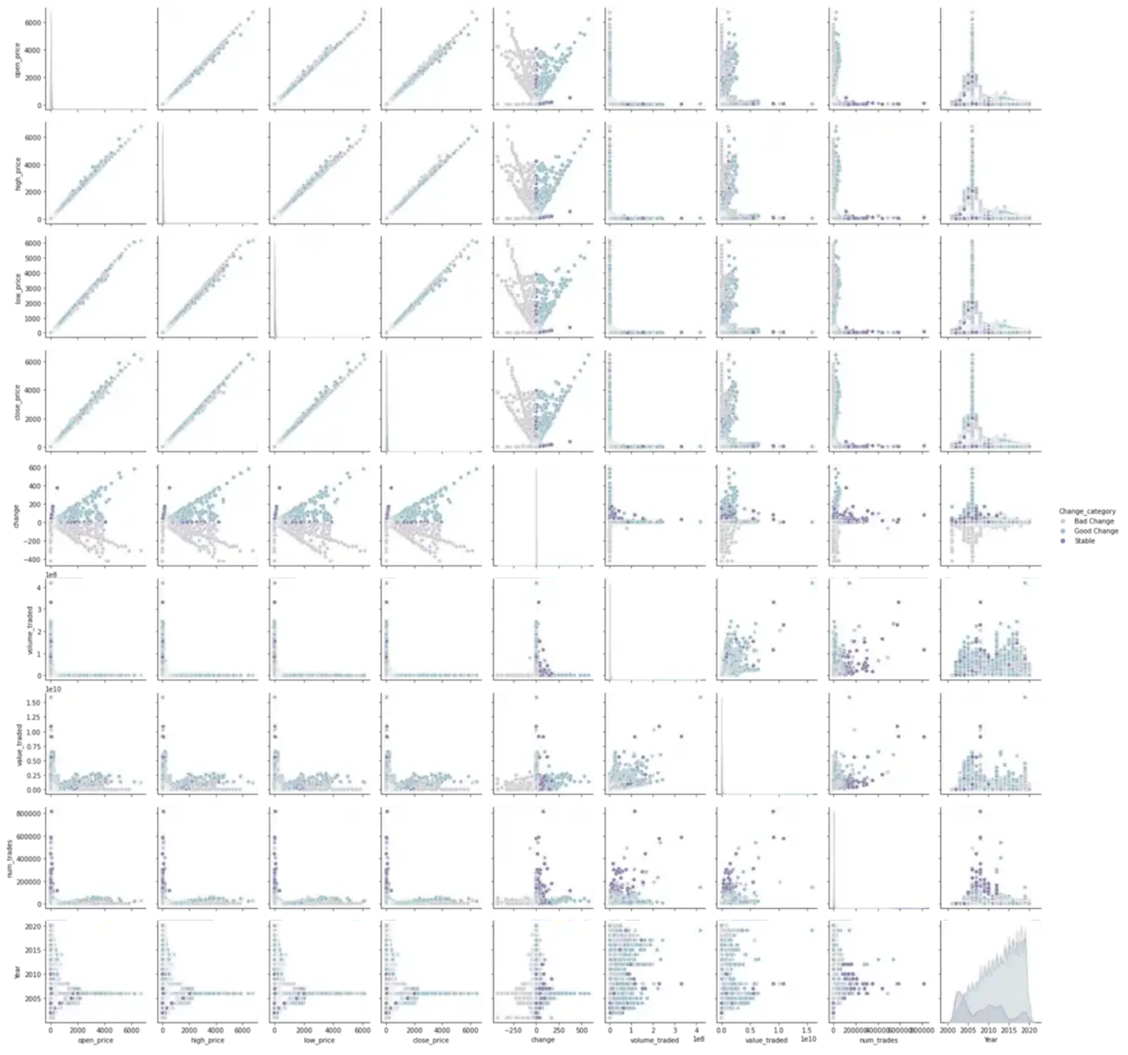
- **Good change:** if the per_change > 0
- **Bad Change:** if the per_change < 0
- **Stable:** if the per_change = 0

4. Data Exploration

Plot 1 (Pair plot): This plot shows multiple insights as follows:

- Trade distribution Rises each year but in 2020 it is broken down
- We can see one peak in open price, high price, low price, and close_price which means the data has no varied regions, And the Volume data is right-skewed.
- The highest price in open_price, high_price, low_price, close_price is more than 6000
- The lowest price in open_price, high_price, low_price, close_price is 0

- We can see all the relationships on daily returns between all the features. A simple glance shows an interesting correlation between close price and open, high and low price daily returns.
- The highest change (Good Change) is almost 600
- The worst change (Bad Change) is -400
- We notice that the good and bad changes are almost similar.

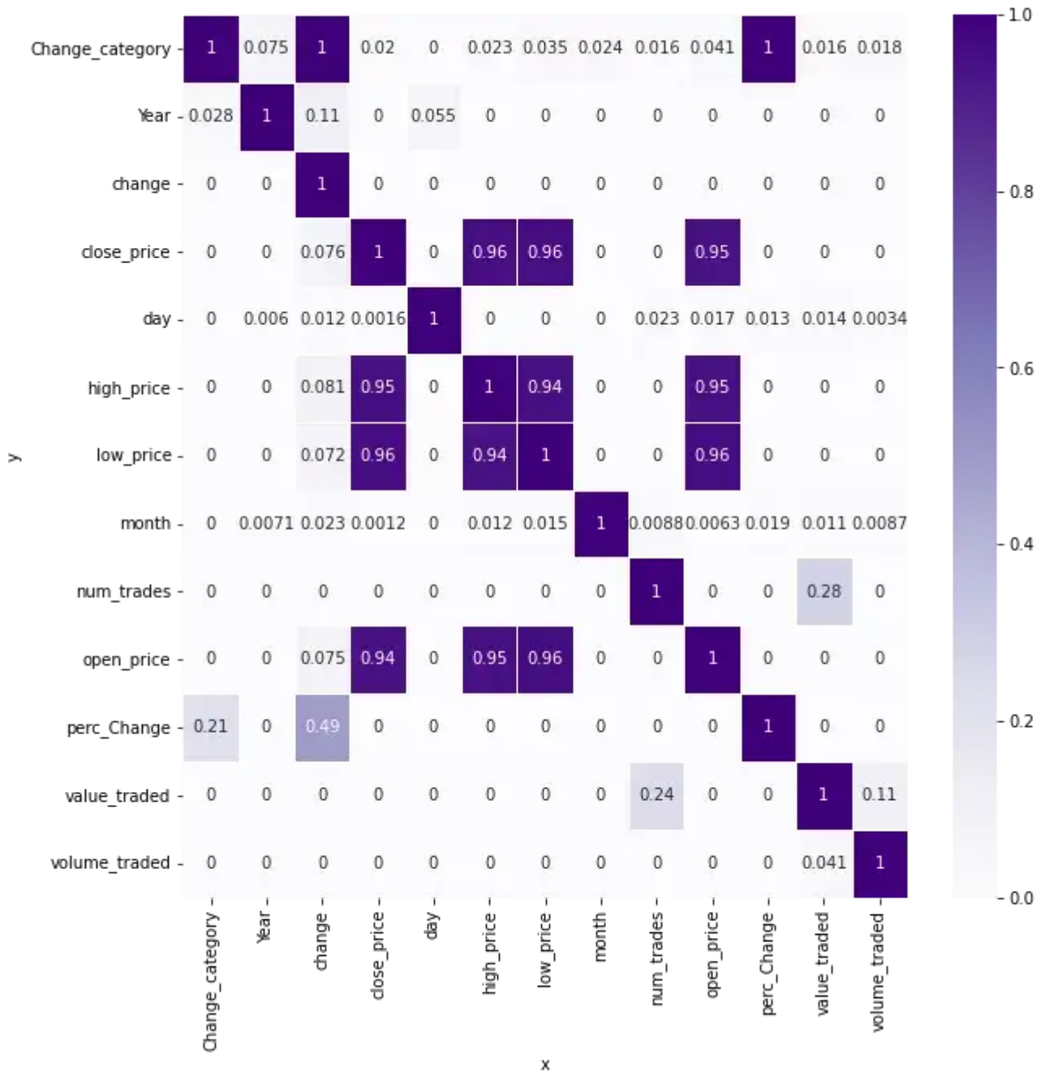


Pair plot

Plot 2 (Predictive Power Score matrix plot): The Predictive Power Score (PPS) is an asymmetric, data-type-agnostic score that can detect linear or non-linear

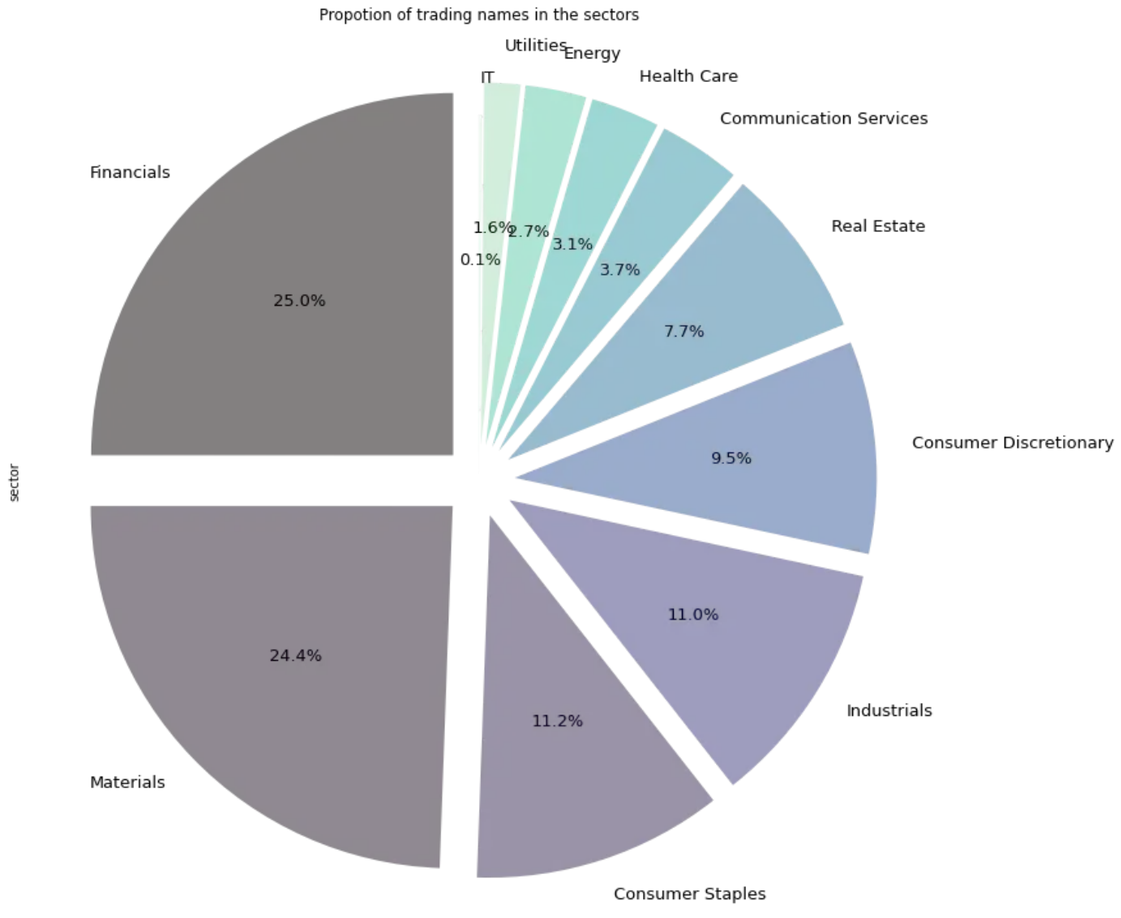
relationships between two columns.

This plot shows high relationships between open_price, high_price, low_price, with our regression target (clos_price)



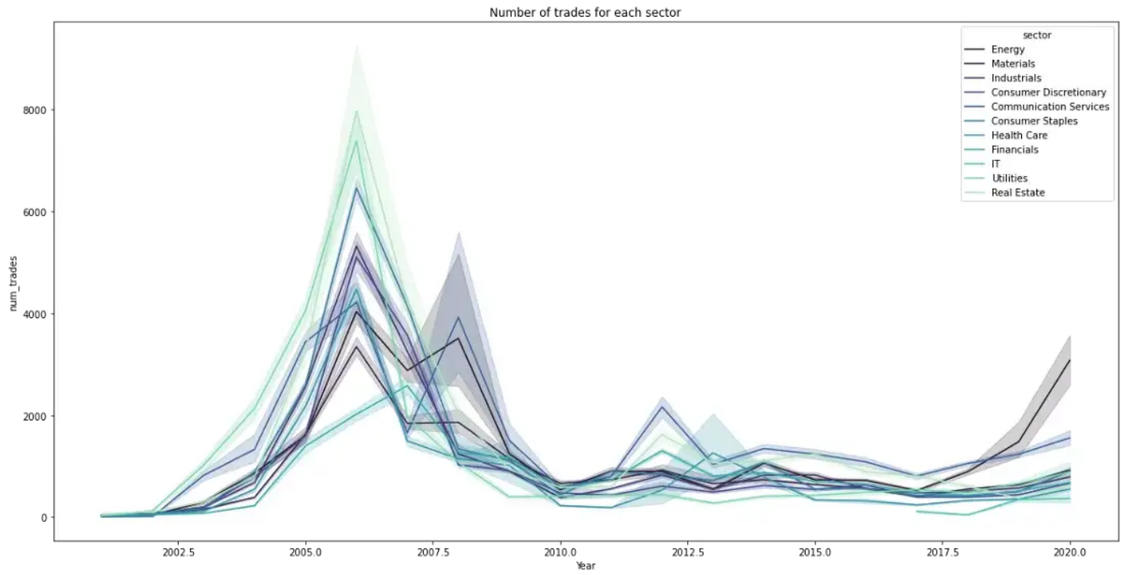
PPS matrix plot

Plot 3 (Pie plot): This plot shows that most of the trading companies are from the financial and materials sectors.



The proportion of trading names in the sectors

Plot 4 (Line plot): This plot shows the number of trades for each sector in each year. As shown that the number of trades got increased in 2006, especially in the Real Estate and Utilities sectors.



The number of trades for each sector in each year

Plot 5 (Box Plot): These two plots show the Close and Open prices in the largest sector (Financial sector):

close_price and open_price in Bupa Arabia for cooperative insurance

- the higher median in 2020, then in 2019 lastly 2018
- the highest variation in 2019, then 2018 lastly 2020
- no outliers

close_price and open_price in the company for cooperative insurance

- the highest median in 2020, then in 2019 lastly 2018
- the highest variation in 2019 then 2018 lastly 2020
- 2020 has outliers in the open and close prices

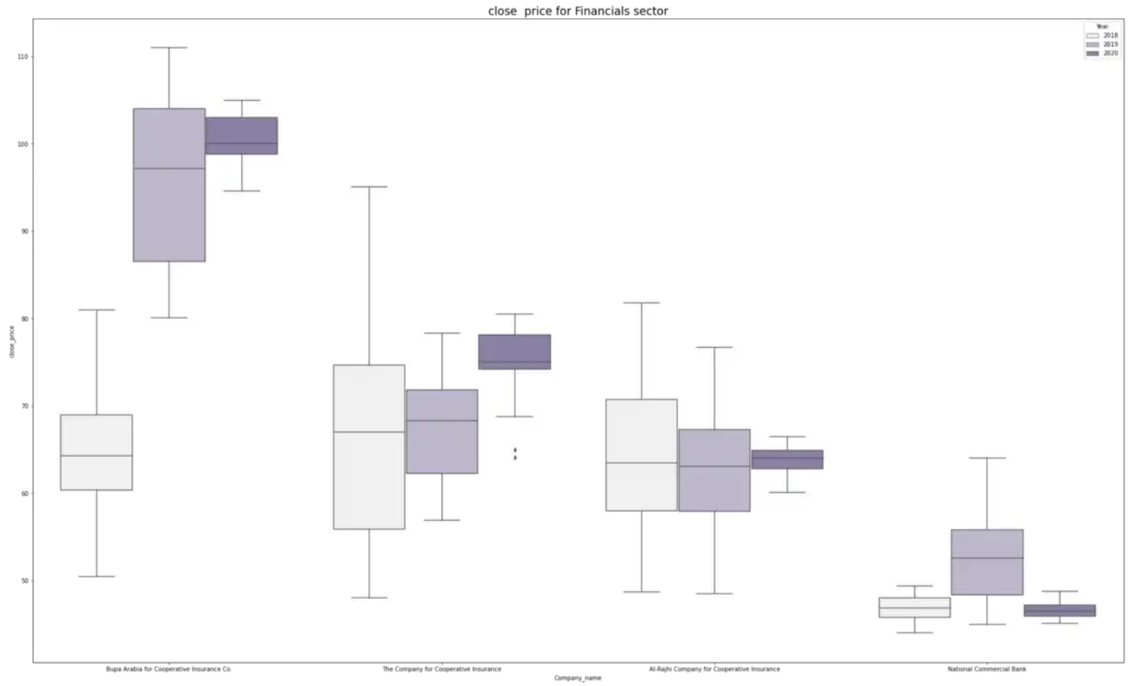
close_price and open_price in alrajhi

- the highest median in 2020, then 2018 lastly 2019

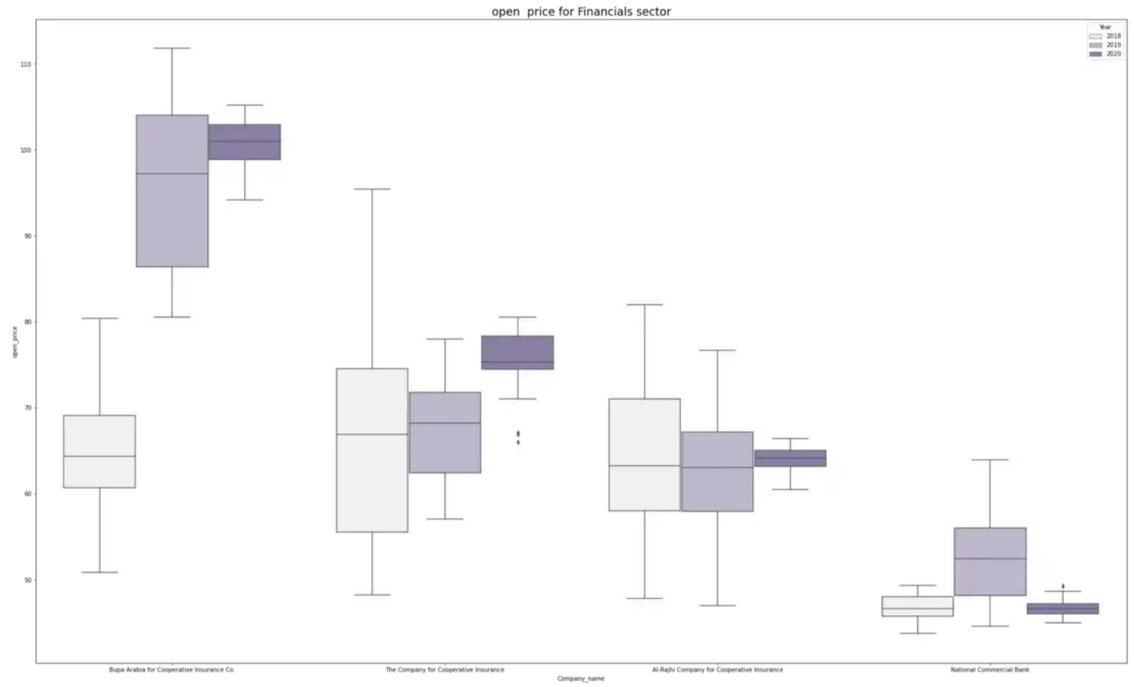
- the highest variation in 2018, then 2019 lastly 2020
- no outliers

close_price and open_price in national commercial bank

- the highest median in 2019, then 2018 lastly 2020
- the highest variation in 2019, then 2018 lastly 2020
- there is an outlier in the open price in 2020



Close price for the Financials sector



Open price for the Financials sector

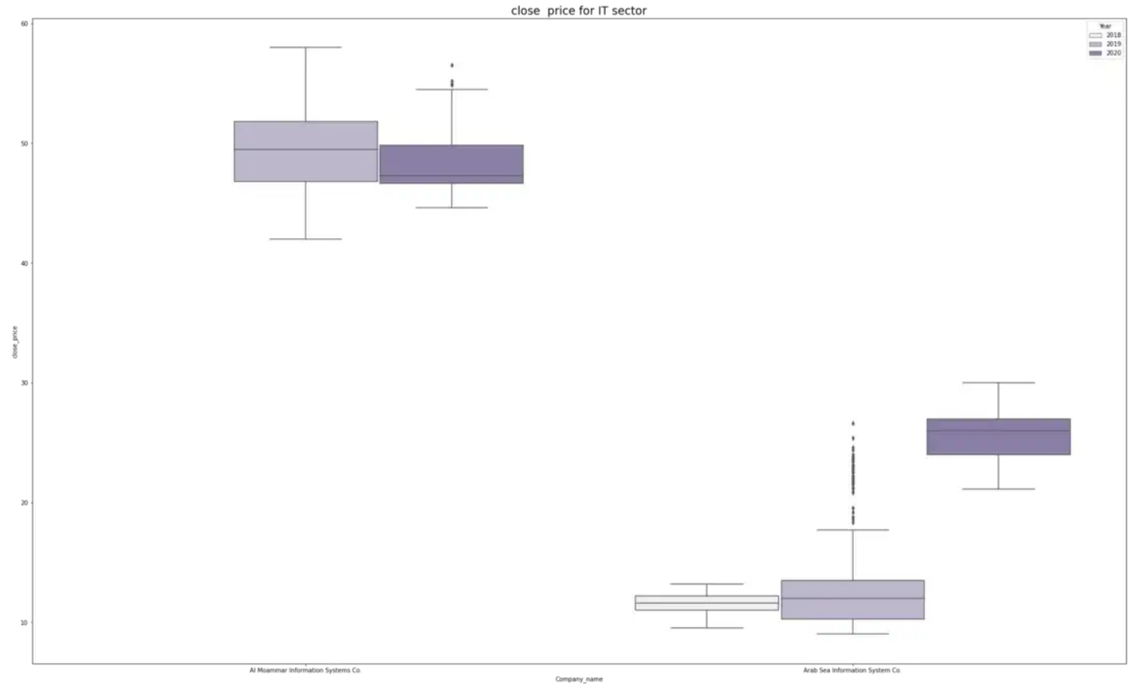
Plot 5 (Box Plot) continued: These two plots show the Close and Open prices in the largest sector Information Technology (IT) sector:

close_price and open_price in 2019 al Moammar information systems

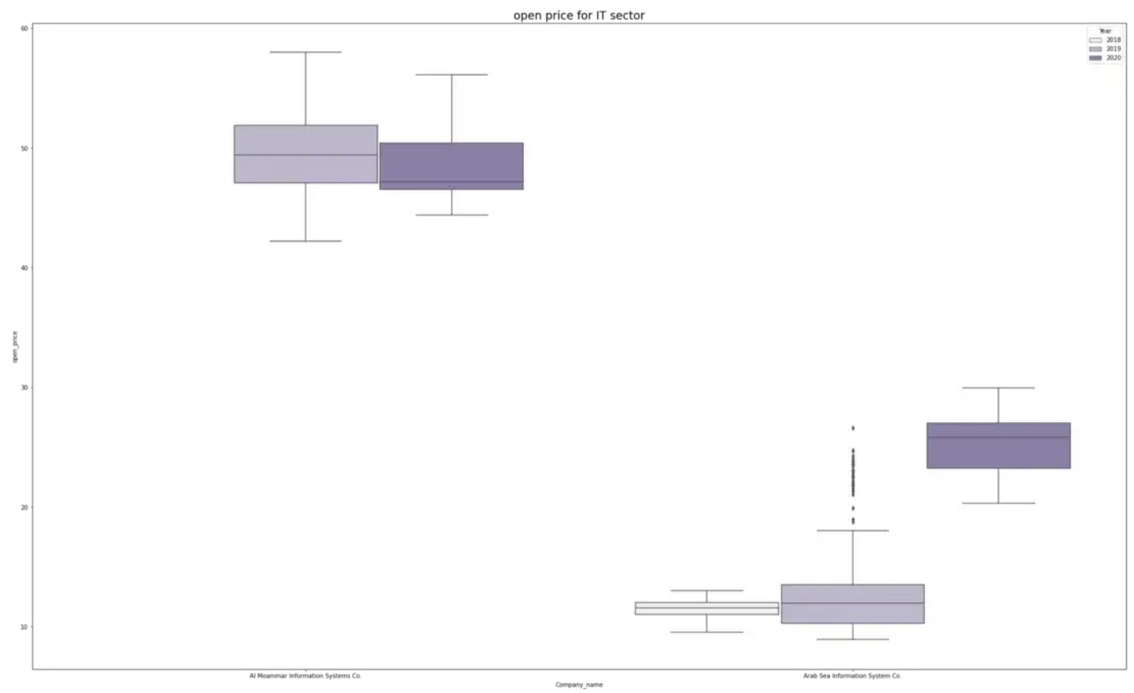
- the highest median in 2019, then 2020
- the highest variation in 2019, then 2020
- close price in 2020 has outliers
- non-existent in 2018

close_price and open_price in Arab sea information systems

- the highest median in 2020, then 2019 lastly 2018
- the highest variation in 2020, then 2019 lastly 2018
- there are outliers in 2019 in the open and close prices



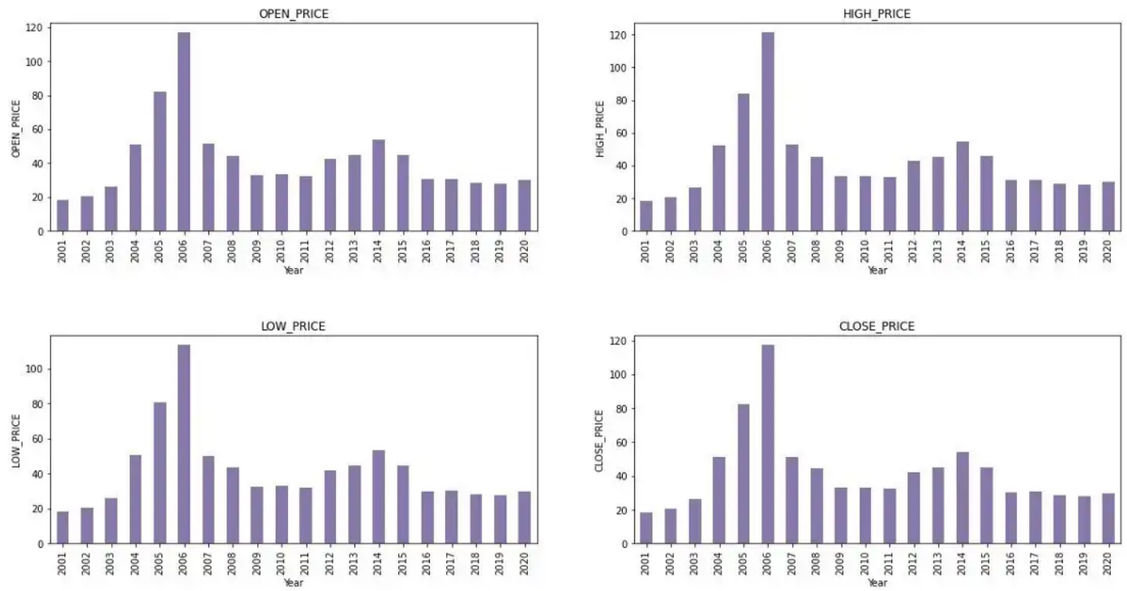
Close price for the IT sector



Open price for the IT sector

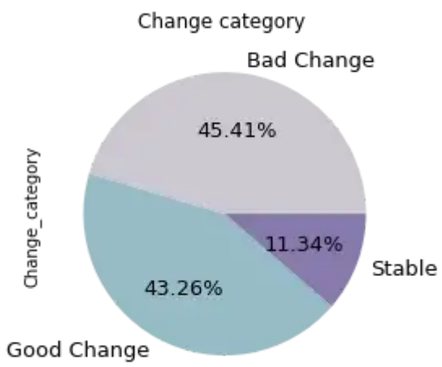
Plot 6 (Bar plot):

- this plot shows that the stock prices increase from 2001–2006
- the stock prices fell in 2007.
- the stock prices stable from 2016–2020



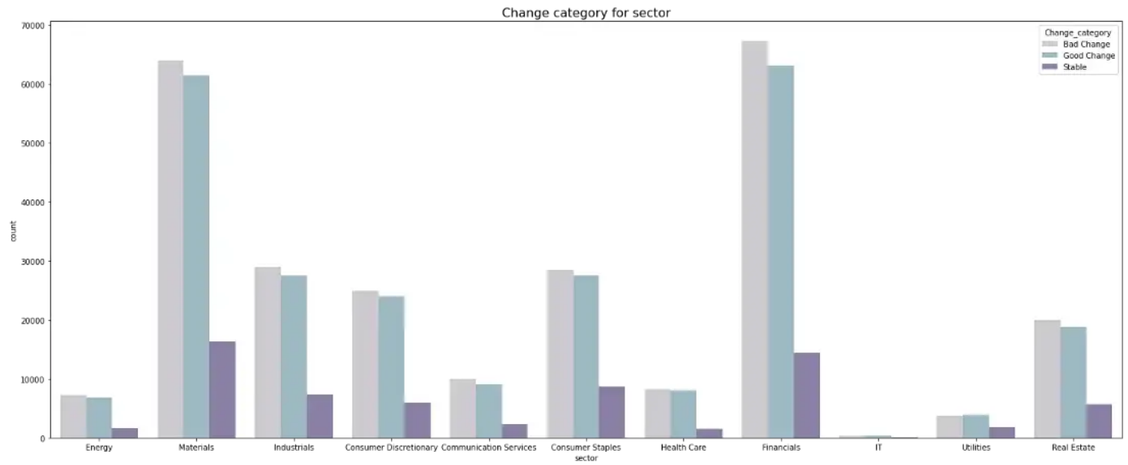
Bar plot

Plot 7 (Pie plot): This plot shows that the Bad Changes class has the highest proportion of the overall changes (45.41%), while the Stable class has the lowest proportion (11.34%). Nonetheless, the Good change class is extremely close to the Bad change class, with a proportion of (43.26%).



The distribution of change categories in the dataset

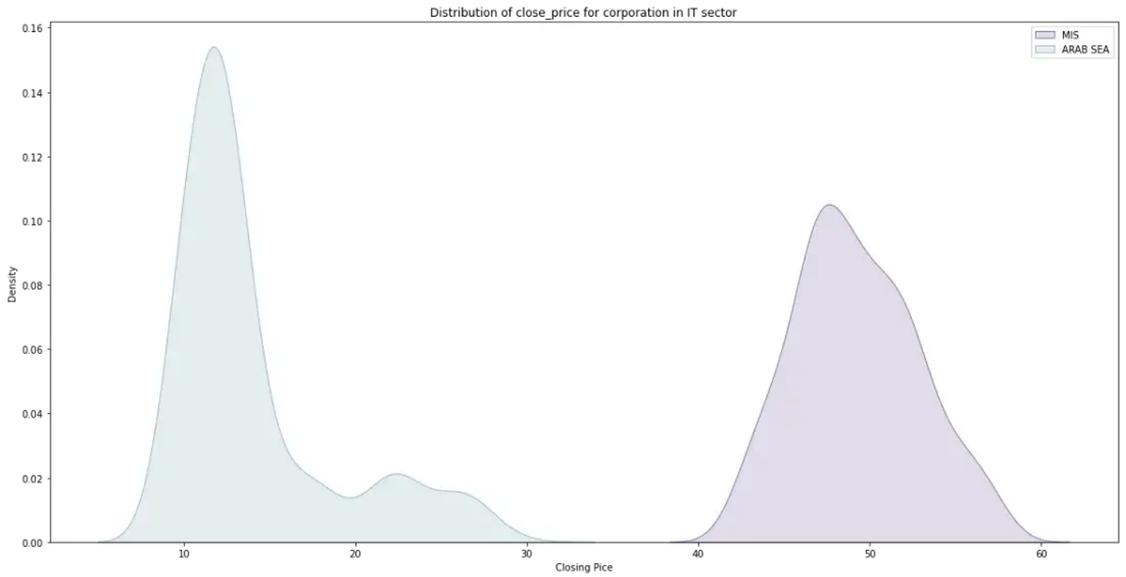
Plot 8 (Count plot): This plot shows that all sectors are close in terms of good or bad changes, while the Stable class is the lowest across all sectors.



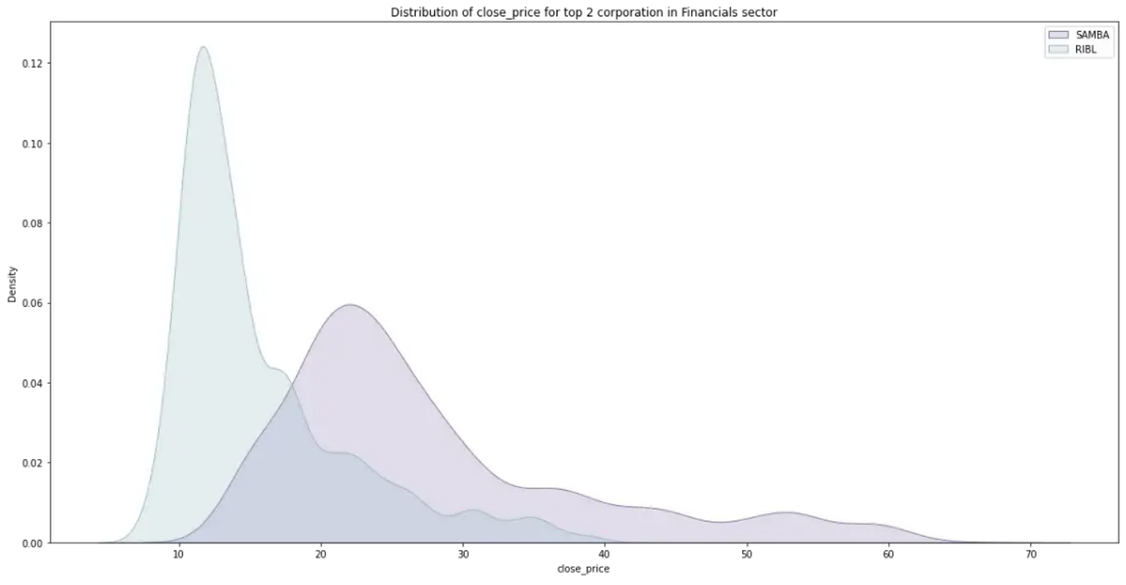
The number of each change category in each sector

Plot 9 (KDE plot): These two plots show us the closing price density for the largest sector companies (only the best 2 companies in terms of the closing price) and the lowest sector companies.

- IT: MIS has the highest density when the close price is almost 13, where ARAB SEA has the highest density when the close price is almost 46.
- Financial: RIBL has the highest density when the close price is almost 12, whereas SAMBA has the highest density when the close price is almost 23.

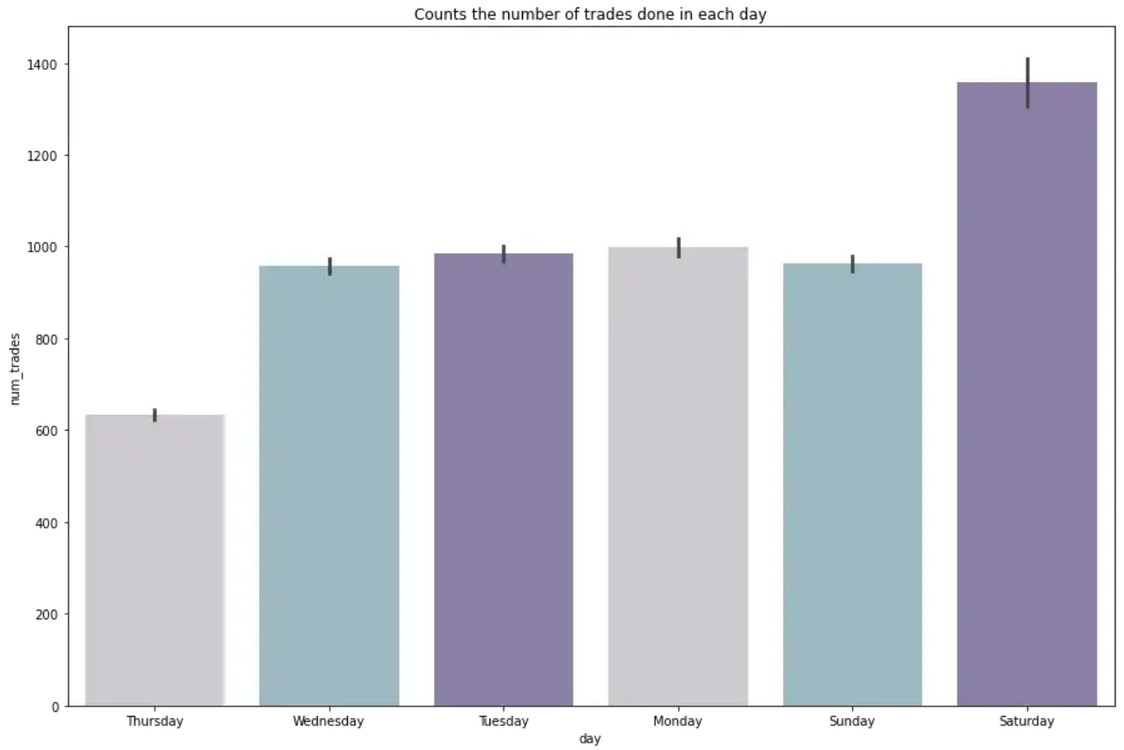


Distribution of close prices for corporations in the IT sector



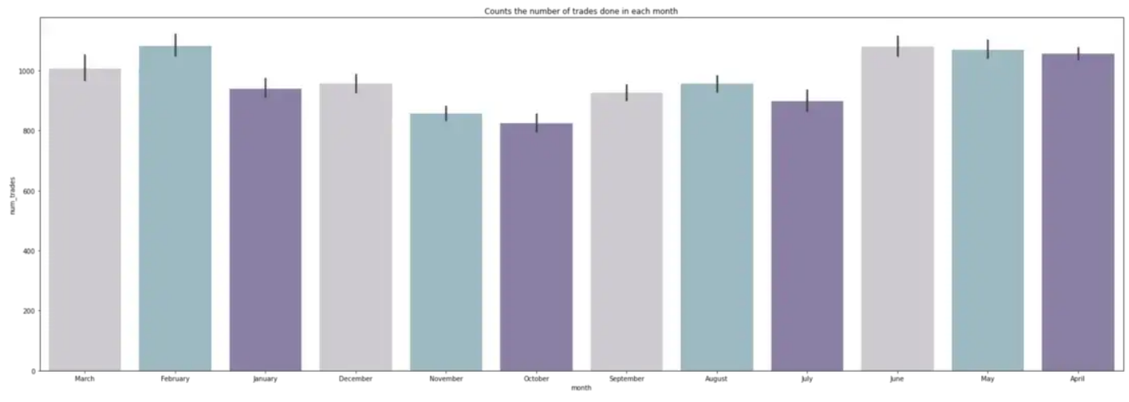
Distribution of close price for top 2 corporations in the Financials sector

Plot 10 (Bar plot): This plot shows the number of trades on weekdays. We notice that the number of trades increases at the start of the investment market on Saturday, while it decreases at the end of the week on Thursday.



Counts the number of trades done each day

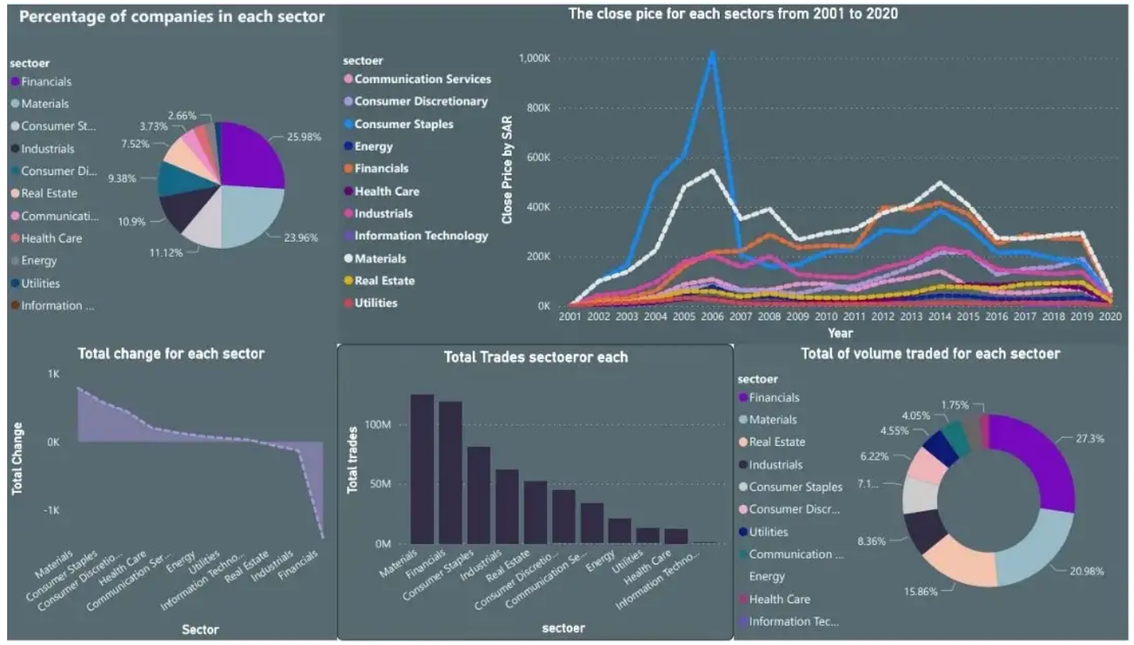
Plot 10 (Bar plot) continued: This plot shows the number of trades in the months. We notice that the number of trades is close in the first and last months of the year, while it varies throughout the rest of the year.



Counts the number of trades done each month

5. Dashboard

We created this dashboard to highlight the important insights of our dataset so that people can easily understand what the project is about.



Saudi Stock Exchange (Tadawul) Dashboard

6. Our Approach: Building Machine Learning Models

> *First: Building Regression Models (predict the close price)*

We choose the regression models to predict the close price because it is a continuous value (float). First, we create the baseline model using the DummyRegressor model, then we trained many regression models based on the open_price, low_price, and the change, then compare the results using the cost functions for the regression models in order to choose the best one with the highest R2-score and the least possible error. Finally, we tuned the best model using the Random_Search method, and this is the final result as is shown in the table below:

	Model	R2	MSE	MAE	RMSE
0	Baseline model	-0.000013	7217.246782	28.401061	84.954381
1	Linear Regression	0.999606	2.844279	0.426416	1.686499
2	Random Forest	0.986577	96.875029	5.931222	9.842511
3	KNN	0.999515	3.501239	0.301051	1.871160
4	GBR	0.999215	5.665061	0.671754	2.380139
5	XGB	0.999213	5.676292	0.396874	2.382497

Comparison of all the regression model metrics

We found that the best model is Linear Regression Model because it gives the highest R2-score with a value of 0.9996 and the lowest error in the Root Mean Squared Error (RMSE) with a value of 1.6.

> *Second: Building Classification Models (predict the change category)*

We choose the classification models to predict the change category because it is a categorical value (object). First, we create the baseline model using the DummyClassifier model, then we trained many classification models based on the low_price, high_price, open_price, and close_price, then compare the results using the classification reports in order to choose the best one. Finally, we tuned the best model using the Random_Search method, and this is the final result as is shown in the table below:

	Model	Accuracy	Recall	precision	F1 score
0	Baseline model	0.454031	0.206144	0.454031	0.283548
1	Logistic Regression	0.758729	0.741616	0.758729	0.713941
2	Random Forest	0.750956	0.758492	0.768822	0.760486
3	KNN	0.756320	0.744912	0.756320	0.744185
4	GBC	0.693741	0.694983	0.693741	0.666824

Comparison of all the classification model metrics

We found that the best model is the Random Forest Model because it gives the highest F1-score metric in the classification reports with a value of 76%. The reason we look at the F1-score metric is that we had unbalanced classes.

7. Results

In this project, we started with some usual cleaning and preprocessing, and then we created a new column to help us predict the change class using the classification model. Finally, we attempted to predict the close price using many regression models, and we only chose the best one and tried to optimize its results, and we found that the best result we got was from the Linear regression with an R2-score of 0.9996 and RMSE of 1.6. We also attempted to predict the change category by testing several classification models and optimizing the best one; the best result we obtained was from the Random Forest model with an F1-score of 76%.

8. Future Work

- Collect more recent and relevant data.
- Try other machine learning models.

- Handling the unbalanced classes in the change category and evaluating the model.
- Try another target.
- Use the Time series model; Since we have time series data.