MINI PROJECT CANVAS

Title (preliminary): Improve investment returns with maching learning method

Group members: Jie Wang, Yunjing Ma, Qianhui Zhong

Workshop # : 3

MOTIVATION

Which is the target group of our mini-project? Who is the end-user?

What are their objectives? What needs do we need to address with our work?

How will they benefit from this proposed solution?

The target group is all the investors who are interested in the stock market. Not only the individual investor but also the institutional investor can be our mini project’s end-user.

The objective is, through the prediction results of stock prices, to analyze the current state and future trends of the stock market, optimize existing investment portfolios, and thereby achieve higher investment returns. To fulfill their needs, we need to explore the advanced method of machine learning to build the model as accurately as possible.

By using our proposed solution, on the one hand, they can increase investment returns, on the other hand, they can minimize investment losses as much as possible.

In our mini project, we will use Tesla as an example, but our solution can also apply in other companies.

DATA COLLECTION

Which data sources are we planning to use?

Mention database tables, API methods, websites to scrape, etc.

Which is the data management plan?

First, we use iFnd API to fetch some original data, like some categories of stock price and some technical indexes. IFind API can be seen as an extension of EXCEL, we can use some formulas and parameters to fetch the data we want.

Second, we get the public interest through Google Trends index (https://trends.google.com/trends) by using the key words of “Tesla stock”.

Lastly, we use date as the key to join two parts of data, after that, we create our project’s original dataset.

PREPROCESSING

What are the goals of the preprocessing pipeline?

Give some examples of data preprocessing steps.

What are some possible data cleaning/wrangling methods you’re planning to use?

What are some possible data transformations that could be useful?

Any feature engineering necessary?

The goals of a data preprocessing pipeline are focused on transforming raw data into a structured, clean and usable format for further analysis and modeling.

Preprocessing steps:
(1) Data integration: Our data comes from the iFind API and Google Trends. The time spans of the data from each source are different, after comparison, we ultimately selected daily data from 1/1/2021 to 30/9/ 2024.
(2) Feature engineering: Since stock trading only occurs on weekdays, indicators such as opening price and closing price will periodically show zero values. To handle these zero values, we need to add a new feature, "Is trading day or not,", to filter out trading day data. This category feature has two labels: “Yes” or “No”.

EXPLORATORY DATA ANALYSIS (EDA)

Look at the data!

What steps are you planning to take towards exploring and understanding better the data you have?

What properties would be meaningful to summarize/visualize in this step?

First, we look at the data to gain a general understanding of them, like use df.head() to get the basic structure of the dataset, use df.describe() to get the basic statistical information, use df.info () to get the data type.

Second, we visualize the data we collected; detailed information of this part will be shown in the “Visualizations”.

Third, we will do the correlation analysis to check the correlation, and, in this part, we may reduce the dimensionality of the variables based on the correlation analysis.

(Something related data processing will be shown in “Preprocessing”)

Through this step, we can get a better understanding of the data, and after this step (including preprocessing and visualization), we can confirm the final useful variables that as our model’s input.

VISUALIZATIONS

List any meaningful visualizations you are planning to produce that will be useful to the end user?

Are you planning to produce any interactive visualizations?

If so, which types of interactivity might be useful to the end user?

For the indexes of stock price, we will plot the line charts to observe the long-term trend and short-term fluctuation.

For the technical indexes, line charts are also suitable to observe these time series characteristics.

For the correlation analysis, scatter plot and correlation heatmap may be useful. We’ll use these visualizations to observe correlation strength, correlation direction, relationship patterns (Detect whether there are any variable pairs that are significantly correlated) between variables.

In our case, we are not planning to produce interactive visualizations.

LEARNING TASK 🐭

(focus on problem definition)

Define the problem setting.

Is this supervised / unsupervised / other…?

Classification / regression / other…?

What are we planning to learn? E.g. What is the target variable / learning outcome?

What variables are we using as input?

After we processed the data and EDA, we can get the most relevant variables, then we will use machine learning method to explore how these variables

LEARNING APPROACH ▯

(focus on solution implementation)

Which ML/statistical methods seem more relevant for the defined problem setting and why?

Which evaluation metrics could be relevant?

Is any special treatment relevant regarding how we choose to split the data or how we cross-validate?

First, we need to know the characteristics of our data: All our data is chronological. Stock prices have temporal dependency: future stock prices are influenced by past price movements. The stock price is affected by many variables, and it has a long-term trend and short-term fluctuation.

When handling the new category feature “Is trading day or not”, we convert it to numerical values: “Yes” to 1, “No” to 0.

Some transformations like data normalization will be helpful, we’ll scale numerical features to a specific range like 0 to 1 to the index of public interest, ensuring consistency across variables and transform features to standardize.

During our process of data preprocessing, feature engineering is necessary, we will add a new feature “Is trading day or not”

COMMUNICATION OF RESULTS 📢

Which type of deliverable will benefit most the end-user? Do we choose to write a blog post, create a website, an app, or other..?

How do we communicate best our results to the predefined target group?

Short description of your interface/workflow (if applicable).

We will write a technical report. Besides, we will communicate the results with a blog post. To make it easier for our end-users, we will attach the methods of how to use our model in our deliverable. And we will publish our project on GitHub.

DATA PRIVACY AND ETHICAL CONSIDERATIONS 🔐

(if applicable)

Are there any fairness constraints that apply to our proposed pipeline?

Is there a need to ask for consent during the data collection process?

Is there a need for data pseudonymization/anonymization?

Any other privacy considerations that come to mind?

We are planning to only use public data that can be obtained by API or downloading.

influence the stock price. It is supervised and regression. We have 13 variables as our original variables, they are opening price, closing price, highest price, lowest price, trading volume, percentage change, MA (Moving Average), MACD (Moving Average Convergence Divergence), KDJ (Stochastic Oscillator), RSI (Relative Strength Index), BOLL (Bollinger Bands), Is trading day or not. And the dimensionality of variables may be reduced through prior process, and then the final variables will as our model's input.	Stock price is non-stationary data. Then the following methods seem more relevant: 1. Multiple linear regression: Because the stock price may be affected by all the other variables. 2. LSTM (Long Short-Term Memory): (1) Learns patterns based on the chronological order. (2) LSTM can retain long-term dependencies. (3) Can capture long-term trends and short-term fluctuation of stock prices. (4) Can handle multivariate time series data. (5) Suitable for handling non-stationary data, where the mean and variance change over time. (6) Can be used for multi-step forecasting. (7) The relation between stock price and other factors may be complex, so linear models may not be sufficient to describe the relationship between them. Evaluation indicators： 1. Mean Squared Error (MSE) 2. Root Mean Squared Error (RMSE) 3. Mean Absolute Percentage Error (MAPE) 4. Mean Absolute Percentage Error (MAPE) 5. coefficient of determination (R^2) 80-20 Train-Test Split (the data is split in chronological order to ensure that the training set is always earlier in time than the validation or test set, preventing data leakage. ）			

ADDED VALUE

Is there a possibility for added value from the data we're planning to use?

What is the added value?

How are predictions turned into added value for the end-user?

By predicting stock price changes, it helps investors identify the right time to buy and sell stocks, thereby increasing investment returns. Stock forecasts help investors seize possible rising trends or avoid potential falling risks.

LEGEND

WEEK 1: Data collection/preprocessing

WEEK 2: EDA & visualizations

WEEKS 3-4: Machine/deep learning

WEEK 5: Fairness & data privacy