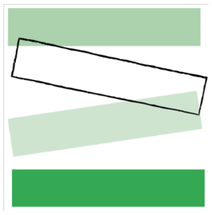
Data Collection + Evaluation



Chapter worksheet

Instructions

Use the exercises below as needed throughout your product's development.

Exercises

1. Get to know your data

Technical indicators like moving averages and historical stock prices, as well as trading volumes, are essential market data for MLOps stock prediction. External elements that can be quite important include industry performance and macroeconomic indicators such as inflation and interest rates. This data can be sourced from stock exchanges, public datasets, or financial APIs (like Yahoo Finance and Alpha Vantage). To train models, make sure the data is accurate, complete, and quality.

2. Speak with a domain expert [~1 hour]

Discussing important topics with a financial, stock forecasting, and MLOps expert is crucial. We first ask about the most pertinent data categories, market sentiment, and reputable data source recommendations. Examine which characteristics and technical indicators have the biggest effects on forecast accuracy and talk about how they incorporate technical and fundamental analysis into their models.

3. Data collection considerations matrix [~1 hour]

The purpose of data collection in MLOps for stock prediction is to assemble historical prices, trading volumes, Return difference, and external factors, as well as market sentiment, to efficiently support accurate, real-time stock movement predictions. A successful model is one that avoids overfitting, maintains accuracy under a range of market situations, generalizes well to new data, is simple to monitor, and can be quickly updated with fresh market insights

.4. Data Labelers + Task Design [~3 hours]

Ensure stock-related data labels and use Python modules such as Pandas for data manipulation, NumPy for numerical calculations, and Scikit-learn for preparation. Deep learning tools such as TensorFlow or PyTorch, as well as SpaCy for text data, are useful for sentiment analysis. Create a collaborative workspace utilizing Jupyter notebooks or, and provide explicit labelling guidelines and sample scripts. Implement Git version control for more effective collaboration and progress monitoring.

5. Write data disaster/diligence headlines [~1 hour]

1. *Data Integrity Crisis: When Inaccurate Data Disrupts Stock Predictions*
2. *The Risks of Unclean Data: How Poor Preprocessing Causes Model Failures*
3. *Real-Time Data Delays: A Hidden Risk in Stock Prediction Models*
4. *Overfitting Alert: When Too Much Data Causes a Disaster*
5. *Version Control Chaos: Preventing Data Mismatches in MLOps Pipelines*
6. *Bias in the Dataset: Identifying Hidden Prejudices Affecting Prediction*

1. Map user needs to data requirements

The first task your team must complete is to identify the type and scope of data needed to train an ML model that can meet your users' needs.

Use the template below for each unique user need your ML model will impact.

Example: building a recipe recommendation service that suggests new dishes to cook.

User needs & data needs	
Users	*Retail investors, financial analysts, institutional traders*
User action (core value prop)	*Make informed buy/sell/hold decisions based on stock prediction*
ML system output	*Predictions of stock price movements, volatility, or market trends*
ML system learning	*Patterns in stock market movements, correlations between features (e.g., trading volume, macroeconomic factors) and stock price changes*

Training dataset needed	*Historical stock prices, trading volume, technical indicators (moving averages, RSI)*
Key **features** needed in dataset	*Stock price (open, high, low, close, adjusted close)* *Return difference (SMB, HML, RMW, CMA)* *Volume of trades*
Key **labels** needed in dataset	*Price movement direction (up/down)* *Volatility prediction labels* *Buy/Sell/Hold signals* *User feedback on prediction accuracy*
Data **formatting**	*date (YYYY-MM-DD)* *open, close, high, low, volume (numeric values)*
Real world data considerations	*incorporate real-time data feeds*
Data source key user questions	*"How does the model make these predictions?"* *"Where does the data come from?"* *"How frequently is the data updated?"*

Synthesize your core data needs with the template below.

Our product/service uses:

- *Fama /French 5 Factors*
- *Yahoo Finance historical stock data*

to provide retail investors and financial analysts with stock predictions for informed trading decisions.

Critical labels for our data include:

- *Price movement direction (up/down)*
- *Volatility labels*
- *Buy/Sell/Hold signals*

We're aware of how the real world (e.g., time of year, changing trends) can impact the data used in our model.

To reflect the dynamism of the real world, we made sure our data includes:

- *Economic cycles and their impact on stock prices*
- *Seasonal effects like holiday spending trends*

2. Speak with a domain expert .

Once your team has the user-data needs template complete, identify **domain experts** who can give you feedback on your initial data hypotheses.

A domain expert is someone with a specialization in your ML model's subject area (not necessarily a ML expert) and can give you insights into the real-world implications of your data.

Questions for domain experts

1. What information is crucial in your field to predict stocks?
- What distinguishes useful from useless data for stock market forecasting and analysis?

2. How is information gathered in your field to forecast stocks?
- Is the main source of data the internet, specialized financial services like Bloomberg, or APIs like Yahoo Finance? Do you know of any trustworthy organizations or sources for gathering data?

3. What issues arise with the information utilized to forecast stocks?
- Do reporting inaccuracies, erroneous market sentiment assessments, or difficulties gathering real-time data pose recurring problems? How often do you come across these?

4. Do any temporal factors affect how data is collected?
- Do specific occasions or seasonal variables have a major impact on the gathering or interpretation of stock market data?

5. How simple or complex is it to reuse data for analysis of the stock market?
- What are the usual obstacles to repurposing old market data or correcting for events such as splits, dividends, and changes in the market?

6. When working with stock market data, what are the most important three aspects that consumers should know?
- Could you discuss some recommended practices or common mistakes that people should avoid when working with stock prediction data, particularly when creating machine learning models?

3. Data Collection Weighted Matrix[1]

Once your team knows what data will be required to train your model based on your answers to in the user + data needs template from exercise 1 and you've consulted with domain experts, you'll need to determine if you can get those data from:

- An existing dataset
- A new dataset

[1] This exercise is adopted from the weighted matrix exercise featured in Martin, Bella, and Bruce M. Hanington. **Universal Methods of Design**: 100 **Ways** to Research Complex Problems, Develop Innovative Ideas, and **Design** Effective Solutions. Beverly, MA: Rockport Publishers, 2012.

Use the weighted matrix below with your team to gain consensus on your data collection plan *(example matrix filled in below for a team with 6 people voting)*:

1. Have each team member vote for which dataset type is the best option for each row
    - The dataset criteria are suggested, you can change the criteria based on your team needs, but we strongly recommend always including 'fit for use case' and 'maintainability'
2. Multiply the number of votes for each option by the associated weight
3. Total the weighted number of votes per dataset option to give direction to your data collection plan

Dataset options → Data criteria ↓	Weight	**Existing dataset (no transformations)**	**Existing dataset (with transformations)**	**New dataset + Existing dataset**	**New dataset**
Fit for use case *Is this data appropriate for your users and use case?* *Consider PII and Protected Characteristics: in some regions it's illegal to use them to make certain predictions.* *Are there any risks of the dataset excluding certain user groups?* *Have you used the Facets tool or some other tool/technique to evaluate the dataset for bias?*	3	2	4	0	0
Legality / Compliance *What data standards are in place for compliance, licensing, documentation?* *See if you the dataset has a Data Card (or whether your team would need to create one)*	3	6	0	0	0
Maintainability *Does your team have a plan for maintaining the data post launch?* *How will data stay up to date over time?*	2	2	4	0	0
Data collection effort *How will the data be collected?* *How will your team ensure ethical data collection practices?*	2	1	5	0	0
Cost *What are the costs of choosing the most expedient data vs. the best data?*	1	1	5	0	0
Total	11	31	35	0	0

4. Data Labelers + Task Design

If your feature uses supervised learning and you are using a new dataset, you need to understand the people who will be teaching or evaluating your model, also known as "raters", (or “oracles”, “labelers”, or “analysts”).

Labelers can be:

- Employees at a labeling company
- Volunteers
- Your own team members
- Or a combination of all the above!

Use the questions below to get to understand potential mental model mismatches between your labelers vs. your users.

4.1 Who are your labelers?

- What are the perspectives or biases that labelers may be bringing to this task that could impact the quality of the labels?

Biases

Market sentiment bias: Labelers may have their own opinions on market trends, leading to biased labeling of stock news sentiment.
Recency bias: Labelers may give more weight to recent events, like a sudden market crash, even when long-term trends are more important.
Subjective interpretation: Especially with news sentiment data, interpretations may differ between labelers (e.g., labeling a news article as “neutral” vs. “positive”).

Contextual Knowledge Required:

Financial market knowledge to understand the impact of economic indicators and company news on stock prices.
Familiarity with sentiment analysis techniques (if manually labeling financial news).
Understanding of macroeconomic trends to label correctly based on historical events, such as recessions or market corrections.

- How will you compensate labelers fairly for their work?

Labelers could be compensated based on the complexity and critical nature of their work. Financial data labelers may require specialized compensation, whereas sentiment labeling for news could use an automated or crowd-sourced approach with appropriate rewards.

4.2 Task Instructions checklist

Help your labelers master a task by creating easy to use instructions.

DRAFT AND PILOT

☑ Draft instructions and budget time to get feedback from labelers on any aspects of the instructions that are unclear. *If you have already made instructions, don't worry! You can ask for feedback at any point.*

BITE-SIZE

☑ Break down instructions into manageable chunks by using bullets for steps, data items, or rules.

  □ In house labeling teams and 3rd party companies may have the benefit of doing in person/remote trainings, but that doesn't mean instructions shouldn't be broken down into easily referenceable chunks

EXAMPLES/IMAGES

☑ Add at least 3 positive, negative, and ambiguous examples to illustrate expectations.

□ If you are advertising a task on an open crowd platform, use images to capture worker interest in your task.

EXPLANATIONS

☑ Explain the overall goal of the effort to provide context and get labeler investment.

□

Explain criteria for acceptance, and clearly state what errors would trigger a rejection of the task. Allow for a feedback mechanism for labelers to flag ambiguous cases. ACCESSIBILITY

☑ Highlight if the task is fully accessible or requires specific abilities to complete.

Task: Labeling Market Events for Stock Price Prediction

Objective: The goal of this task is to label major market events that can influence stock prices. Your labeled data will be used to train a machine learning model that predicts future stock price mo

You are tasked with identifying and labeling major market events (e.g., earnings reports, mergers and acquisitions, government policies) that may influence stock prices.

Steps:

1. *Identify the event described in the article (e.g., earnings release, new government policy, product launch, merger).*
2. *Label the event based on its expected short-term impact (up to 3 months) and long-term impact (beyond 3 months) on the stock market.*
   - *Short-term:*
     - *Positive: The event is likely to cause an immediate increase in stock prices.*
     - *Negative: The event is likely to cause an immediate decrease in stock prices.*
     - *Neutral: The event is unlikely to cause immediate changes.*
   - *Long-term:*
     - *Positive: The event will likely boost the company's or the market's value over time.*
     - *Negative: The event could lead to long-term challenges for the company or market.*
     - *Neutral: The event won't likely have a lasting impact on stock prices.*
3. *Document event context: Provide a brief explanation in a comment for why you assigned the labels (e.g., "This earnings report shows strong revenue growth but rising costs that may impact long-term profitability.").*

Additional Instructions

- *Consistency: Ensure you label similar news articles and events consistently across the dataset.*
- *Use of external sources: You are encouraged to use trusted external sources (e.g., financial reports, market analysis) to verify the validity of the events or news articles.*

- *Time sensitivity: Make sure to account for the time period of the news or event. For example, news about an event from 6 months ago should be assessed based on the historical context.*
-

Examples:

1. *Positive Example:*
   "Tech Company A reports a 15% increase in quarterly profits driven by new product launches."
   - *Sentiment: Positive*
   - *Short-term impact: Positive*
   - *Long-term impact: Positive*
2. *Negative Example:*
   "Retail Company B's stock plunges after reporting a 10% drop in revenue due to supply chain issues."
   - *Sentiment: Negative*
   - *Short-term impact: Negative*
   - *Long-term impact: Neutral*
3. *Neutral Example:*
   "Economic data released today shows stable growth in the labor market."
   - *Sentiment: Neutral*
   - *Short-term impact: Neutral*
   - *Long-term impact: Neutral*

5. Submission

- *Submit your labeled dataset in the provided format (e.g., CSV/Excel).*
- *Ensure all flagged and ambiguous cases are documented with proper explanations.*

4.3 Task design and usability

In case you missed it - read the article First: Raters to understand how different types of labeling impact the design of labeling tools.

- □ **Do the task yourself!**
  - □ Catch and correct any usability issues prior to testing with labelers.
- □ **Observe people completing your task**
  - □ Can labelers complete key tasks quickly and without errors? *Yes*
  - □

Note: make it clear you are evaluating the task and not the individual's performance.

- ☐ **Plan for unsures**
  - ☐ Is your labeling UI forcing labelers to label prematurely or in error? *No*
  - ☐ How are you thinking about inter-rater reliability? *Inter-rater reliability will be managed through standardized guidelines, periodic reviews, and consensus mechanisms*
    - ☐ Will labelers be able to periodically indicate their level of confidence for a given task submission? (This technique can help reduce the need for multiple ratings)

      Yes, labelers can indicate their confidence using a rating scale, helping flag uncertain tasks for further review. This reduces the need for multiple ratings on high-confidence submissions
    - ☐ Can the data be labeled in more than one way?

      Yes, certain data points may have multiple valid interpretations, so we'll enable multi-class labeling and hierarchical labels. Labelers will also be allowed to add notes for ambiguous cases
- ☐ **Welcome feedback on your task/tool**
  - ☐ What incentives are there for labelers who speak up about discrepancies or interesting insights beyond the scope of the task?

    Labelers who identify discrepancies or provide valuable insights will be rewarded with performance bonuses or recognition through leaderboards
- ☐ **Provide feedback to labelers in a timely manner**
  - ☐ How will labelers know they are doing a good job and that their feedback is valued?

    Labelers will receive regular performance feedback through scorecards, highlighting accuracy, consistency, and quality. Positive contributions, such as valuable feedback or catching discrepancies will be acknowledged

Testing: Complete jobs on your own and see if labelers can correctly and consistently recognize and label price patterns, volatility spikes, or volume surges.

Premature Labeling: Watch out for early labeling of market movements, such as bearish or bullish patterns, by the user interface.

Inter-Rater Reliability: Adhere to uniform standards and conduct regular reviews when designating the direction of stock price movements using moving averages or candlestick patterns.

Confidence Rating: To cut down on rework, give labelers the option to rate their confidence in their ability to predict levels of support or resistance.

Multi-Labeling: Turn on multi-class labeling for technical indicators such as Bollinger Bands, MACD, and RSI amid unclear market situations.

Feedback and Incentives: Provide acknowledgment for identifying market irregularities or providing perceptive sentiment analysis, and use performance indicators such as accuracy in classifying uptrends or downtrends.

Additionally, you can use / modify the following questionnaire to evaluate the usability of your task:

Please evaluate the usability of the task you are working on.

	Agree	Disagree	Not applicable	Comments
1. The goal of the task is clear	○			
2. The task instructions are comprehensive	○			
3. The task instructions are easy to reference		○		
4. The task was easy to learn		○		

5. The steps to complete the task are in a logical sequence	◉			
6. The task shortcuts are useful	◉			
7. The task shortcuts are logical	◉			
8. It is easy to ask questions and get answers about the task		◉		
9. The time to complete the task is appropriate	◉			

5. Data disaster/diligence headlines

Write data disaster (and diligence) headlines to spot problematic data issues before they happen. Use these headlines to identify any data concerns to follow up with your engineering partners.

Data privacy	*Investors are upset to discover stock price prediction model uses personal financial data without explicit consent.*
	Stock price prediction model earns trust by only using publicly available stock data and anonymized sentiment data from financial news

Guiding questions

- How do you get access to the data? Do you have permission?

  Access comes from public datasets (stock prices, financial news, macroeconomic data) and the Fama/French 5 Factors data, all of which are freely available or licensed.

- What anonymization and/or aggregation techniques does your product use?

  Financial news sentiment is anonymized, and no personal user data is involved. Stock data and macroeconomic indicators are aggregated.

Data exclusion	*Stock price model criticized for lack of global market data, excluding non-U.S. investors.*
	Model praised for inclusive approach, incorporating global market data to serve international investors

Guiding questions

- What is the downstream, real-world effect of this model's performance?

  Investors using this model could make informed decisions across diverse markets if the dataset includes global data. Exclusion could lead to inaccurate predictions in foreign markets.

- What data is missing that would adversely impact certain user groups?

  Data on international stocks, currencies, and geopolitical events could be missing, impacting non-U.S. investors

Data ethics	*Public calls to boycott stock prediction service over poor compensation for financial news labelers.*
	Stock price model sets ethical standard by compensating financial analysts for accurate news labeling.

<table>
<tr><td colspan="2">Guiding questions
● Who are the humans involved collecting and/or labeling your data?
<em>Financial news sentiment data could be labeled by third-party financial experts or automated tools. If using human labelers, they should be adequately compensated</em>
● How are you compensating them for this critical work?
<em>Through incentives</em></td></tr>
<tr><td rowspan="2"><strong>Data transferability</strong></td><td><em>Stock price prediction tool abandoned after faulty use of historical social media data for market predictions.</em></td></tr>
<tr><td><em>Model praised for successfully incorporating well-vetted financial data sources, avoiding irrelevant datasets.</em></td></tr>
<tr><td colspan="2">Guiding questions
● What risks are present for using data not originally intended for your use case?
<em>Incorporating irrelevant or unverified data sources (like social media for financial predictions) could lead to faulty predictions</em></td></tr>
<tr><td rowspan="2"><strong>Data fragility</strong></td><td><em>Stock prediction model fails after disruptions in real-time market data feed.</em></td></tr>
<tr><td><em>Stock price prediction model remains robust by incorporating macroeconomic trends and accounting for major market events</em>.</td></tr>
<tr><td colspan="2">Guiding questions
● Does your data reflect the real world? e.g. for image-based systems does it include off center/blurry images?
<em>Yes, the model incorporates real-world events like market crashes, geopolitical events, and economic trends, ensuring robust predictions</em></td></tr>
</table>