Before you begin...

- You'll be meeting virtually with your assigned student team on or around **Thursday, August 14th.**

- This template deck is what **we ask all Challenge Advisors to fill in and be ready to present during this first meeting.**

- Your assigned student team will have specific questions about the project related to their **Project Brief and Workplan assignment** (due to Break Through Tech on September 7, 2025). This Challenge Project Overview Template provides dedicated space to help you answer many of these questions (e.g., project milestones, preferred project workspaces).

Advice from former students: How to scope out and present your project

- Give us **structure, clear goals, and a timeline** to work toward at the outset of the project. Open-ended prompts can be hard for us since we're used to college course assignments.

- Provide **guidance on the initial steps** for us to take. Don't assume we know what to do, even if it's obvious to you! Some of us are new to Python and data science.

- Provide us with a **useable dataset** that (a) includes proper documentation and (b) doesn't require so much cleaning or pre-processing that it prevents us from getting started.

- **Help us anticipate potential challenges** during later phases of the project (and, when the time comes, help us tackle them by pointing us in the right direction).

- Suggest **resources** for us to better understand the problem space and possible approaches. What would **you** have found helpful as an undergrad?

BREAK
THROUGH
TECH

Using LLMs to Predict Stock Trends from Financial News

August 2024

We’re excited to be your Challenge Advisors!

Atena Sadeghi (Athena)
Enterprise data and AI
Executive consultant

[**Full Name** (pronouns)
Company Name
Job Title
Email Address]

Company overview

I am an independent consultant in the Field of Data and Ai. I provide end to end enterprise solution

- I started new role recently and I am working with Fedcap group which is a Nonprofit
- I am located in NYC and been in ML and AI field for past 10 years
- I also worked at Finance , Fintech, Media , Consulting and Tech companies

More about [Company name]

[XX%]

[Insert descriptive content (optional)]

[$XXM]

[Insert descriptive content (optional)]

[XX+]

[Insert descriptive content (optional)]

AI Studio Challenge Project Overview -1

CHALLENGE SUMMARY

In this project, you will build an NLP pipeline that analyzes public financial news headlines to predict the short-term movement (up or down) of a stock or market index, such as S&P 500 or TSLA. Using state-of-the-art language models (LLMs), you'll extract sentiment scores or generate text embeddings to represent news context numerically. This project addresses a key financial analytics challenge: transforming unstructured news data into actionable market insights. Your main objectives are to clean and align textual and financial data, apply LLM-based techniques for feature extraction, and train predictive models to forecast daily market trends.

Goal 1: Build a Clean, Aligned Dataset

1. 1- Collect and preprocess financial news headlines (e.g., from Kaggle or RSS feeds)
2. 2- Pull corresponding historical stock data using yfinance or similar APIs
3. 3- Align the news and price movement data at the daily level

   Success Metric: A structured dataset ready for modeling, with 90%+ date alignment completeness

AI Studio Challenge Project Overview - 2

Goal 2: Extract News Features with LLMs

1. Use models like FinBERT or other transformer-based LLMs to generate sentiment scores or embeddings from news headlines
2. Explore summarization techniques for multi-headline days

Success Metric: Feature extraction pipeline implemented with model interpretability visualizations (e.g., attention maps or SHAP)

Goal 3: Train and Evaluate a Predictive Model

1. 1- Train a supervised model (e.g., logistic regression, random forest, or neural net) to predict next-day price direction
2. 2- Tune model parameters and evaluate performance using Accuracy, ROC-AUC, and F1 Score

Success Metric: Model ROC-AUC > 0.70 on test set, with evidence of generalization and baseline comparisons

AI Studio Challenge Project Overview - 3

DESIRED OUTCOMES (December Presentation)

By December, we’d like the student team to present a well-documented machine learning pipeline that ingests financial headlines, applies LLMs for representation, and predicts market movement with measurable performance. The deliverables should include:

- A walkthrough of data sourcing, cleaning, and feature engineering steps
- Performance metrics of the final predictive model, with baseline comparisons
- Visualizations or dashboards showing the relationship between news sentiment and stock behavior
- Discussion of model limitations and suggestions for next steps or real-world deployment

Business context

Financial markets are highly sensitive to real-time news, headlines, and investor sentiment. Yet most firms still lack scalable tools to convert unstructured news into actionable market signals.

This Challenge Project focuses on solving a core business problem:

"How can we extract meaningful insights from financial news headlines and predict short-term stock movements?"

By combining stock data with NLP techniques and GenAI models like LLMs (e.g., FinBERT), the student team will attempt to forecast next-day stock direction based on daily news sentiment — a problem that touches both trading and risk functions in finance.

Suggested ML approach

Model / Approach	Why It’s Suitable
Logistic Regression	A strong baseline classifier for binary outcomes; interpretable and fast to train
Random Forest Classifier	Handles mixed feature types well; useful for small/medium datasets
XGBoost Classifier	High-performing for tabular data; provides built-in feature importance for explainability
FinBERT + Classifier	Use FinBERT (or another LLM) to generate sentiment scores or embeddings, then classify movement
Sentence-BERT + NN	Convert headlines to dense embeddings with \`sentence-transformers\`, then feed into shallow neural net
LSTM or BERT (optional)	For advanced teams: sequence models or fine-tuned BERT-based classifiers for headline-level prediction

Large Language Models

Some things to keep in mind if this is your first time working on a project involving LLMs:

Challenge Advisors:
Remove this slide if it is not applicable to your project. Feel free to modify the placeholder content.

- **Pre-trained Models**
  Start with pre-trained models (e.g., GPT, BERT).
- **Preprocessing**
  Perform text preprocessing (e.g., tokenization, normalization).
- **Fine-Tuning**
  Fine-tune the model on your specific dataset.
- **Evaluation**
  Use metrics like BLEU, ROUGE, and human evaluation.
- **Deployment**
  Plan for integration and scalability.

Data overview - 1

Financial News Headlines Dataset

- **Source**: https://www.kaggle.com/datasets/notlucasp/financial-news-headlines
- **Format**: CSV- Download the data from the above URL
- **Description**: Contains historical financial news headlines across multiple companies and sectors, with timestamps. Useful for extracting sentiment or embeddings using NLP models.
- **Potential Challenges**:
  - Timestamps may not perfectly align with market close times.
  - Some headlines may be repetitive or irrelevant.

Historical Stock Price Data

Source: https://aroussi.com/post/python-yahoo-finance

- **Format**: DataFrame (retrieved dynamically via code) https://github.com/ranaroussi/yfinance
- **Description**: Provides historical OHLCV (open, high, low, close, volume) data for any stock or index. Can be used to generate next-day movement labels.
- **Potential Challenges**:
  - Need to carefully handle market holidays/weekends
  - Volatility may be high; labeling should consider thresholds for movement

Data overview - 2

PREPROCESSING TASKS

- Clean and normalize headlines (e.g., lowercase, remove symbols, stopwords)
- Merge and align news data with daily stock closing prices
- Label data based on next-day stock return direction (e.g., `1` for up, `0` for down)
- Optional: Group multiple headlines per day into one summary using an LLM

Python libraries

Library	Purpose / Use Case
pandas	Data loading, cleaning, merging, and manipulation
numpy	Numerical computing, working with arrays and matrices
matplotlib / seaborn	Plotting and visualizing trends, stock prices, and evaluation metrics
scikit-learn	Building ML models, feature selection, cross-validation
xgboost	Advanced gradient boosting for high-performance modeling
tensorflow / keras	Training neural networks for deep learning
nltk / spaCy	NLP preprocessing: tokenization, stopwords, NER
transformers	Using pre-trained LLMs like FinBERT for embeddings or sentiment
sentence-transformers	Convert headlines to vector embeddings using LLMs
vaderSentiment	Fast rule-based sentiment analysis for finance text
yfinance	Pull historical stock data from Yahoo Finance
datetime	Handle and manipulate dates for aligning news and stock
shap (optional)	Explain model predictions using SHAP values
gradio / streamlit (optional)	Create an interactive app to showcase model predictions
langchain (advanced)	Build LLM-powered agents for context-aware modeling

Suggest tools and workspaces

[Enter a subtitle (optional)]

- [List your recommended free tools for:
  - Data Sharing and Collaboration (e.g., GitHub, Dropbox)
  - Coding & IDE (e.g., VS Code, PyCharm)
  - Project Management (e.g., GitHub Projects, Notion)
  - Other (e.g., Google Colab for free GPU resources)]
- [Explain how the students should use GitHub to collaborate]
- [Add any insights from your own professional life into methods or frameworks that you've found helpful – e.g., Agile, Kanban, task delegation]

Helpful resources

[Enter a subtitle (optional)]

- https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news
- Use ChatGPt codes for debugging

Project milestones and timeline

These are the monthly Milestones for your Challenge Project. You will need to complete several tasks for each of these.

MILESTONE 1

September

Data pipeline

MILESTONE 2

October

Models and performance

MILESTONE 3

November

Model improvement and report

Example Data preprocessing tasks

Task Description	Suggested Tools and Libraries
Handle missing values, outliers, and duplicates in both stock and news datasets.	Pandas, NumPy, Seaborn
Convert dates, normalize numerical values, encode categorical variables, and prepare text fields.	Pandas, Scikit-learn, NLTK, spaCy
Use dimensionality reduction techniques like PCA or feature selection to simplify the model input.	Scikit-learn, XGBoost (feature importance)
Generate more labeled examples through techniques like synonym replacement or backtranslation for NLP.	NLTK, Hugging Face Transformers, TextAttack
Visualize trends in sentiment, movement, and volume; plot class balance and correlation.	Matplotlib, Seaborn, Pandas

Examples of model performance tasks

Task	Description / Purpose	Tools / Libraries
Train/Test Split or Cross-Validation	Split data properly to avoid leakage. Use time-based split for financial/time-series data.	scikit-learn, pandas
Baseline Comparison	Compare against a simple baseline like majority class or previous day’s direction.	scikit-learn
Accuracy, Precision, Recall, F1 Score	Evaluate classification performance and handle class imbalance.	scikit-learn.metrics
ROC-AUC Curve	Visualize how well the model separates positive vs. negative classes.	scikit-learn, matplotlib
Confusion Matrix	Visualize TP, FP, TN, FN to interpret model errors.	scikit-learn, seaborn, matplotlib
Classification Report	Summarize precision, recall, and F1 score for each class.	scikit-learn
Feature Importance Plot	Visualize which features drive model predictions.	xgboost, shap, scikit-learn
SHAP Values (optional)	Explain individual predictions using feature contributions.	shap
Model Overfitting Check	Compare training vs. validation performance to detect overfitting.	matplotlib, pandas
Time-based Evaluation	Track model performance over time (e.g., weekly or monthly).	pandas, matplotlib
Error Analysis	Review misclassified samples to find patterns or edge cases.	pandas

MILESTONE #3 TASKS

{Optional}

- [**List a task** Include a brief task description]
- [**List a task** Include a brief task description]
- [**List a task** Include a brief task description]

How we’ll work together this semester

Category	Details
Check-in Meetings	Monthly check-ins aligned with milestones. Come prepared with updates, blockers, and demos. Notify in advance for any blockers.
Reporting	Bi-weekly written progress updates via Notion or GitHub Issues. Update task board and milestones regularly.
Communication	Preferred via email asadeghi@udel.edu or Break Through Tech Slack. Response within 72 hours. Tag for urgent issues.
Tools and Platforms	Use GitHub (version control), VS Code (IDE), Google Colab (GPU notebooks), Notion or GitHub Projects (task tracking).
Other Project Norms	Follow Agile-lite. Use mini-sprints and assign task ownership (e.g., different pipelines). Use Kanban boards for visual tracking.
Presentation Expectations	Divide roles for final slides: pipeline (data flow), data analysis (EDA, trends), model (architecture), evaluation (metrics, performance). Align with repo and tasks.

How to get started

Here’s what I suggest for your immediate next steps. I’ll follow up on your progress and help address any challenges in our next check-in meeting:

Review these slides and note down questions

I’ll email you a copy of this deck. Review it as a team and note down any questions you’d like to discuss in our next meeting.

Complete your “Project Brief and Workplan”

Continue working on your Project Brief and Workplan assignment, which is due next month. We’ll review it again in our next meeting.

[TBD next step]

[Add any additional next steps that you want to emphasize (e.g., access the dataset; set up dev/PM tools; do research)]

Questions?

What questions do you have?

Anything I can help clarify?

What are you most excited about?

Anything you’re unsure about?