Data Science Canvas

Project:

Team:

Problem Statement

Business Case & Value Added

Which business case should be analyzed and what added value does it generate?

The case is predicting each trading day's high and low for NIFTY 500 stocks and generating early buy/sell signals. Added value: improves intraday decision-making, reduces trader bias, and enables realistic profit estimation through systematic backtesting including costs and slippage.

Data Landscape

Which data is required for this and which is already available? Which additional data has to be collected?

Available: minute-level OHLCV data for all NIFTY 500 stocks.
Required: engineered indicators (returns, ATR, rolling highs/lows, VWAP).

Model Selection

Which analysis methods can be considered on the basis of the specific data landscape and the business case?

Minute-level OHLCV data with highly non-linear and noisy patterns requires models that can capture complex dependencies.
Thus, tree-based gradient boosting models XGBoost, LightGBM, CatBoost , neural network, Stacking Ensemble ensembled together with probabilities[0.3, 0.2, 0.2, 0.2, 0.1] were used for BUY/SELL classification, whiletwo separated sequential3 layered (128 → 64 → 32) LSTM models were used for predicting daily high and low prices. These models balance accuracy, speed, and the ability to generalize across noisy intraday data.

Model Requirements

Which model requirements must be complied with in order to obtain a valid model?

Model must train using only past data (strict time-based split) and avoid look-ahead when generating indicators. Predictions must be early enough for trading and robust under different market regimes.

Software & Libraries

Which software should be used? Is there already a standard solution? Which libraries are used?

Python-based workflow using pandas, NumPy, scikit-learn, XGBoost/LightGBM, and statsmodels. Matplotlib/seaborn/plotly for visualization and Dash for reporting.

Skills

What skills are needed to provide the data and model development?

Time-series preprocessing, indicator engineering, ML modeling with tree ensembles and simple neural nets, backtesting logic, and performance evaluation including transaction-cost modeling.

Execution & Evaluation

Model Evaluation

Which indicators require quality control and validation and how should they be interpreted? Is real-time monitoring necessary?

Classification signals must be evaluated using **Accuracy, F2 Score, and AUC**, ensuring the model correctly detects directional movement patterns.
For high/low regression, **R^2 and RMSE** must be monitored to verify the quality and stability of numerical predictions.
Cross-validation is needed to ensure generalization across stocks and days.

Data Storytelling

What requirements does the target group have for the presentation of the results and how do I effectively communicate this data?

Stakeholders need clear, visual summaries of intraday trends, cluster behavior, model accuracy, and prediction errors.
Charts comparing actual vs predicted highs/lows, F2-score bars, ROC curves, and cluster-wise EDA help them quickly evaluate reliability.

Data Collection & Preparation

Data Selection & Cleansing

Which of the available data is relevant? Do the data have to be cleaned up?

Use regular-hours minute data only.
Clean duplicates, erroneous ticks, missing minutes, and negative or zero-price anomalies.
Align timestamps to uniform intervals and normalise inconsistent formatting.

Data Integration

In which system should the data from different sources be migrated?

Combine all stock files into a unified columnar dataset (parquet).
Integrate indicators and additional market features in a feature store ensuring strict time alignment.

Data Collection

How and with which methods should additionally required data be collected? What properties has this data to fulfil?

Data is downloaded from the Kaggle Algo Trading dataset (2015–2025) containing minute OHLCV per stock.
All data must have synchronized timestamps, consistent trading-session boundaries, and maintain minute-level granularity without gaps.

Explorative Data Analysis

Are there outliers or structures to be considered? Creation of descriptive key figures for the first assessment of the data.

Check for missing minutes, extreme price spikes, zero-volume entries, and duplicated ticks.

Adopted from: https://github.com/tomalytics/datasciencecanvas