MACHINE LEARNING 2 – PROJECT PROPOSAL APPROVAL DOCUMENT

About Me

My name is **Kamil Kashif**, student ID **428629**, and I am completing the ML2 project **individually**. I chose to work alone because this project will serve as the foundation for my future research (my upcoming Master's thesis) and possibly a potential publication to a journal ideally with you ☺ helping later with my PhD applications. Working individually allows me to fully control the methodology, create a consistent research pipeline, and reuse the models for my personal algorithmic trading experiments.

Introduction

This project examines the application of modern machine learning techniques to cryptocurrency price forecasting, using ETH–USD and BTC–USD market data obtained from the Binance API. The study integrates both regression and classification components into a single, coherent framework, enabling the prediction of future prices as well as directional market movements. These predictions are subsequently used to construct and evaluate an algorithmic trading strategy.

Dataset Description and Data Source

The project uses two high-liquidity cryptocurrency datasets: **ETH–USD** and **BTC–USD**. Both datasets consist of historical trade-level data retrieved directly from the **Binance Exchange API**, ensuring accurate, high-resolution market information. The data covers the period from **1 January 2024 to 30 September 2025**, providing a sufficiently long and volatile timeframe for training, validating, and testing predictive models. Data will be collected at the **aggregated trade (aggTrades)** level, which includes timestamp, price, quantity, taker side, and trade notional value. These raw trades are subsequently transformed into standardized features such as minute-level OHLCV bars and microstructure-based indicators. For the OHLCV data you may find this link: LINK_TO_GD_LINK. The microstructure data is not there yet as its being downloaded and is expected to be completed by the end of the next week, also that data is huge as its contain around a million rows for each day. All data is obtained programmatically using REST API.

Feature Engineering

After collecting the raw Binance trades, the data is transformed into a structured time-series dataset combining OHLCV information and microstructure signals. The following data is computed:

OHLCV FEATURES	MICROSTRUCTURE FEATURES
Open, High, Low, Close, Volume	Taker Buy Volume (quote-weighted)
Realized Volatility	Taker Sell Volume (quote-weighted)
Log Returns	Cumulative Volume Delta (CVD)
Moving Averages	Cumulative CVD
Rolling Volume	Taker Buy Ratio (TBR)

Note: This document was formatted and written with assistance from an AI model (GPT-5.1).
The ideas and project design are entirely my own.

Machine Learning Tasks (Objective, Hyperparameter Tunning, Trading Strategy)

Regression Task: The regression component aims to predict the future closing price at time t + sequence_length. This continuous forecast is used to generate long-only trading signals based on expected future price changes.

Classification Task: The classification component predicts the direction of future market movement, using log-returns mapped into three classes through volatility-adjusted thresholds: 1 (buy), 0 (Hold/Neutral), -1 (Sell).

I would like to employ the following models: Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and as a baseline use Random Forest to check if advanced models are always viable. Furthermore, I would also like to test and Ensembling approach to check for improved stability and robustness. For regression, take the weighted average of predictions for each model used and for classification, take the majority voting across all models. All excluding the baseline model Random Forest.

The full evaluation relies on **Walk-Forward Optimization (WFO),** a methodology used in financial modeling to avoid look-ahead bias. Each window is divided into Train, Validation, and Test. And, finally an Out-of-Time data. For each window, models are retrained and re-optimized, ensuring that all assessments reflect realistic predictive performance under evolving market conditions. This procedure provides a more reliable estimate of generalization than a single static train-test split.

Instead of traditional grid search or Keras Tuner, the project uses a modern MCP based optimization library – Optuna. The link is here: **LINK**. Furthermore, the tuning process will maximize Information Ratio 2 (IR2), a performance-oriented metric composed of Annualized Returns Compounded, Annualized Standard Deviation, and Maximum Drawdown. This objective is selected because conventional error metrics (e.g., MSE, accuracy) do not align with trading performance. IR2 directly evaluates the stability, profitability, and drawdown characteristics of the model generated strategy, making it more appropriate for financial forecasting.

The project implements a long-only algorithmic trading strategy derived from the regression and classification model outputs. For regression, trades are executed when the predicted future price exceeds or falls below the current price by a volatility-adjusted threshold, resulting in either entering a long position, exiting to neutrality, or staying out of the market. For classification, model predictions directly map to trading actions: buy (1), neutral (0), or exit/avoid long positions (–1). Short selling is intentionally excluded to reflect practical constraints and the intended real-world application of the system. All strategies are benchmarked against a standard Buy & Hold approach to assess the value of the machine learning framework.

Summary

This project develops a unified machine learning framework for forecasting cryptocurrency prices and generating long-only trading signals using deep learning, ensemble methods, and walk-forward optimization. The approach is fully reproducible, grounded in rigorous methodology, and designed to support future academic research and practical algorithmic trading applications.

Note: This document was formatted and written with assistance from an AI model (GPT-5.1).
The ideas and project design are entirely my own.