MACHINE LEARNING TECHNIQUES TO PREDICT STOCK PRICES AND REGIME CHANGES IN FINANCIAL MARKETS

Saloni D Jaitly, Jie Chen, Dominique Morris

Supervised by: Prof. Jörg Osterrieder, Prof. Ali Hirsa

IEOR
Industrial Engineering
& Operations Research
COLUMBIA | ENGINEERING

zh
aw

INTRODUCTION

THE PREDICTION OF STOCK PRICES IN FINANCIAL MARKETS IS CRUCIAL FOR INFORMED DECISION-MAKING, RISK MANAGEMENT, AND MARKET ANALYSIS. ACCURATE FORECASTS ENABLE INVESTORS TO OPTIMIZE STRATEGIES AND ALLOCATE RESOURCES EFFECTIVELY, WHILE ALSO SERVING AS ECONOMIC INDICATORS AND SUPPORTING ALGORITHMIC TRADING. THIS PROJECT AIMS TO LEVERAGE MACHINE LEARNING TECHNIQUES TO PREDICT STOCK PRICES AND IDENTIFY REGIME CHANGES, HARNESSING THE ADAPTABILITY AND PATTERN RECOGNITION CAPABILITIES OF THESE MODELS. PREVIOUS WORK ACHIEVED ACCURACIES RANGING FROM 0.50 TO 0.52, MOTIVATING OUR EXPLORATION OF MACHINE LEARNING'S POTENTIAL IN FINANCIAL MARKETS. HOWEVER, CHALLENGES EXIST, INCLUDING MARKET COMPLEXITIES AND VOLATILITY. ASSUMPTIONS ARE MADE REGARDING DATA STATIONARITY, INDEPENDENCE OF OBSERVATIONS, MARKET EFFICIENCY, AND FEATURE RELEVANCE. THROUGH THIS PROJECT, WE SEEK TO ENHANCE ACCURACY, CAPTURE NON-LINEAR DYNAMICS, AND SUPPORT DECISION-MAKING IN FINANCIAL MARKETS USING MACHINE LEARNING.

DATA

THE DATA FOR THIS PROJECT IS OBTAINED FROM VARIOUS SOURCES, SUCH AS FINANCIAL DATA PROVIDERS, STOCK EXCHANGES, DATA VENDORS, AND APIS. IT UNDERGOES A THOROUGH ACQUISITION PROCESS, INCLUDING DATA PREPROCESSING, FEATURE EXTRACTION, AND CLEANING. THE HISTORICAL STOCK DATA COLLECTED SERVES AS THE FOUNDATION FOR ANALYSIS AND PREDICTION. PREPROCESSING AND CLEANING PROCEDURES ENSURE CORRECT FORMATTING, HANDLING OF MISSING VALUES, AND EXTRACTION OF RELEVANT FEATURES FOR FUTURES TRADING AT A 5-MINUTE INTERVAL. THESE PROCEDURES INVOLVE FORMATTING DATE AND TIME COLUMNS, REMOVING UNNECESSARY COLUMNS, FILTERING OUT INVALID OBSERVATIONS, RESAMPLING THE DATA, AND EXTRACTING SIGNIFICANT FEATURES. ADDITIONALLY, USER INPUT IS INCORPORATED TO INCLUDE FUTURE DATA, ENSURING UP-TO-DATE INFORMATION FOR ACCURATE PREDICTIONS

METHODOLOGY

1) DATA COLLECTION AND PREPROCESSING: WE COLLECT RELEVANT DATA FROM VARIOUS SOURCES AND CLEAN IT TO ENSURE ACCURACY AND CONSISTENCY. THIS INCLUDES HANDLING MISSING DATA, NORMALIZING VARIABLES, AND REMOVING OUTLIERS.

2) FEATURE EXTRACTION: WE EXTRACT MEANINGFUL FEATURES FROM THE PREPROCESSED DATA THAT CAPTURE IMPORTANT PATTERNS AND RELATIONSHIPS, SUCH AS MOVING AVERAGES AND VOLATILITY MEASURES.

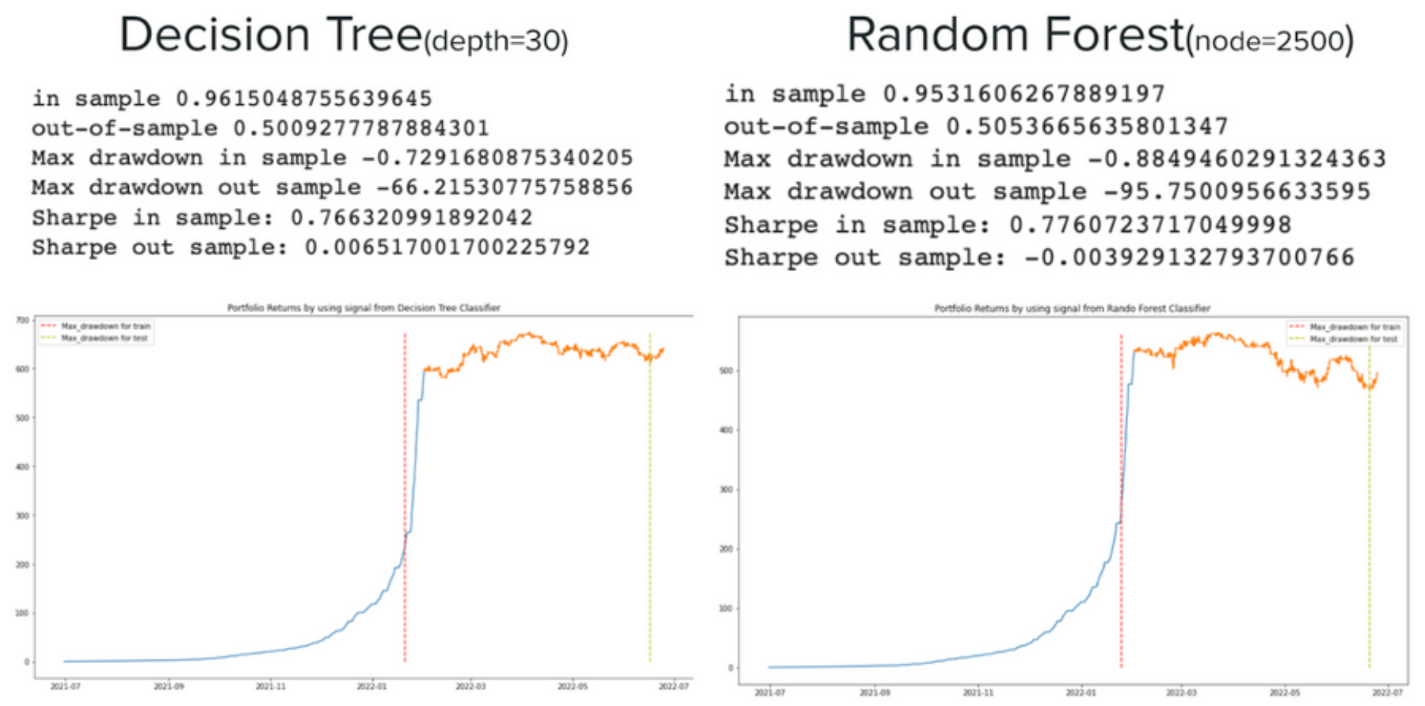
3) MODEL SELECTION: WE SELECT APPROPRIATE MODELS FOR STOCK PRICE PREDICTION AND REGIME DETECTION, CONSIDERING OPTIONS LIKE LINEAR REGRESSION, RANDOM FORESTS, AND GEOMETRIC BROWNIAN MOTION (GBM) FOR PRICE PREDICTION, AND CLUSTERING ALGORITHMS AND GAUSSIAN MIXTURE MODELS (GMM) FOR REGIME DETECTION.

4) EVALUATION: WE EVALUATE THE PERFORMANCE AND INSIGHTS GAINED FROM THE MODELS TO UNDERSTAND STOCK PRICE MOVEMENTS AND REGIME CHANGES. THIS HELPS IN DECISION-MAKING AND INVESTMENT STRATEGIES. ACCURACY AND PREDICTIVE POWER OF THE MODELS ARE ASSESSED.

DATA COLLECTION AND ANALYSIS TECHNIQUES INCLUDE COLLECTING HISTORICAL STOCK DATA, MARKET INDICATORS, AND FUTURES DATA. WE APPLY DATA CLEANING TECHNIQUES TO ENSURE INTEGRITY AND USE TIME-SERIES AND STATISTICAL ANALYSIS TO IDENTIFY TRENDS, PATTERNS, AND CORRELATIONS. MACHINE LEARNING MODELS, ESPECIALLY UNSUPERVISED TECHNIQUES LIKE GMM AND CLUSTERING ALGORITHMS, ARE USED FOR STOCK PRICE PREDICTION AND REGIME DETECTION. THESE MODELS IDENTIFY CHANGES IN VOLATILITY, TRENDS, OR MARKET STATES, AIDING IN EVALUATING MARKET CONDITIONS. CUSTOM FUNCTIONS ARE IMPLEMENTED, SUCH AS 'VOLATILITY_REGIME_ANALYSIS()' FOR REGIME ANALYSIS, A TRAIN-TEST SPLIT FUNCTION, 'RUN_CLASSIFIER' FOR TRAINING AND EVALUATING CLASSIFIERS, AND 'PLOT_FEATURE_IMPORTANCE' FOR VISUALIZING FEATURE IMPORTANCE.

FINDINGS AND OBSERVATIONS

1) ACCURACY:
PREVIOUS DECISION TREE AND RANDOM FOREST MODELS ACHIEVED AN IN-SAMPLE ACCURACY OF 0.7, WHICH INCREASED TO 0.97 BY ADJUSTING LEAF NODES AND MAX DEPTH. HOWEVER, OUT-OF-SAMPLE ACCURACY REMAINED AROUND 0.5, INDICATING POOR GENERALIZATION TO UNSEEN DATA.



Decision Tree(depth=30)

```
in sample 0.9615048755639645
out-of-sample 0.5009277787884301
Max drawdown in sample -0.7291680875340205
Max drawdown out sample -66.21530775758856
Sharpe in sample: 0.766320991892042
Sharpe out sample: 0.006517001700225792
```

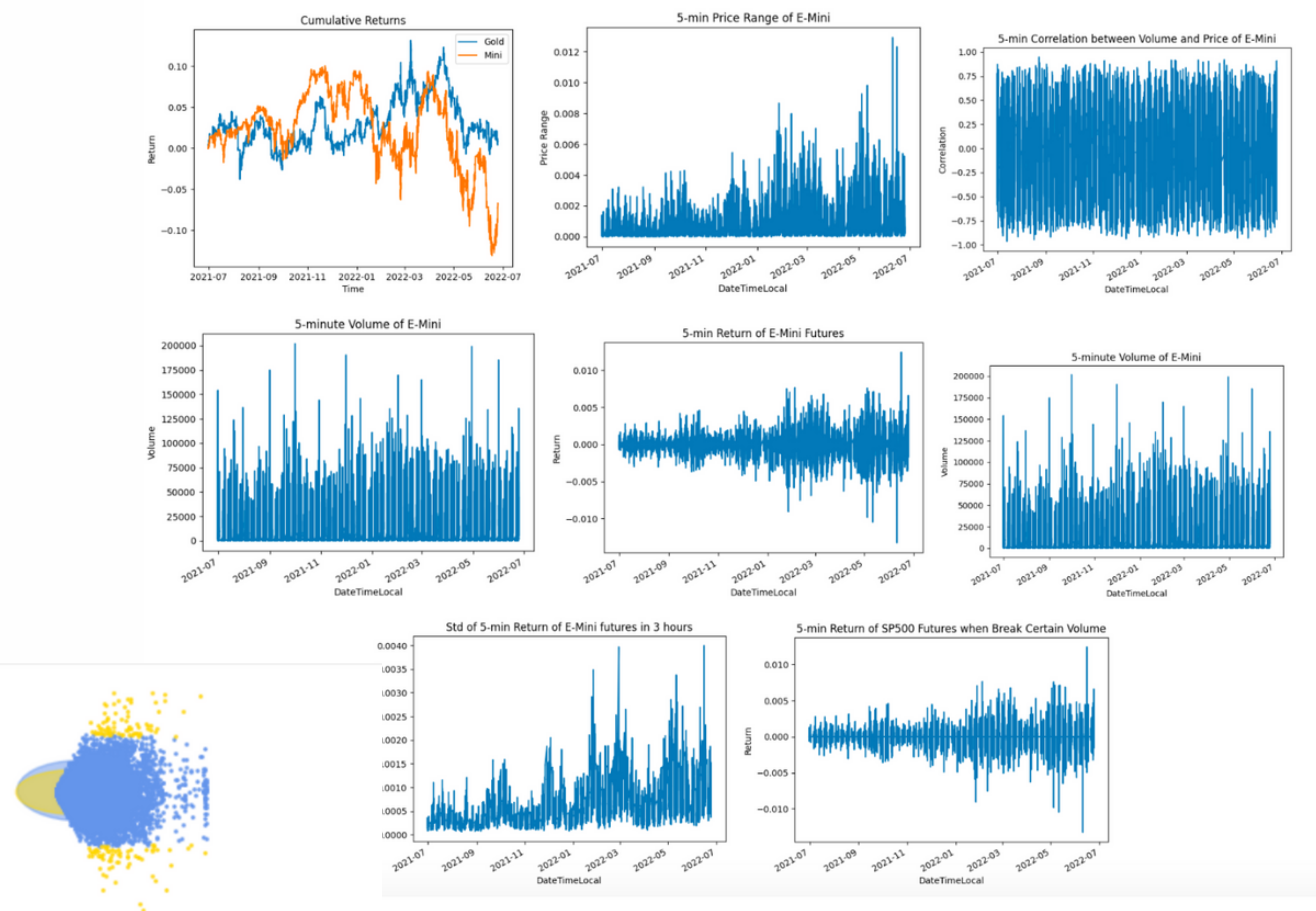
Random Forest(node=2500)

```
in sample 0.9531606267889197
out-of-sample 0.5053665635801347
Max drawdown in sample -0.8849460291324363
Max drawdown out sample -95.7500956633595
Sharpe in sample: 0.7760723717049998
Sharpe out sample: -0.003929132793700766
```

Portfolio Returns by using signal from Decision Tree Classifier
Max_drawdown for train
Max_drawdown for test
Portfolio Returns by using signal from Rando Forest Classifier

2)SIMPLIFIED CLASSIFICATION MODELS:
SIMPLIFIED MODELS PROVIDED DETAILED OUTPUTS, AIDING IN UNDERSTANDING MODEL PREDICTIONS. FEATURE IMPORTANCE PLOTS OFFERED INSIGHTS INTO THE SIGNIFICANCE OF DIFFERENT FEATURES.

3) VOLATILITY REGIME ANALYSIS:
THE "VOLATILITY_REGIME_ANALYSIS()" FUNCTION ANALYZED FUTURES DATA BY EXAMINING THE STANDARD DEVIATION OF 5-MINUTE RETURNS FOR S&P500 FUTURES OVER A 3-HOUR PERIOD. QUANTILES (E.G., 33RD AND 67TH PERCENTILES) WERE USED TO CATEGORIZE VOLATILITY REGIMES, PROVIDING INSIGHTS INTO MARKET DYNAMICS. THIS INFORMATION CAN GUIDE ADJUSTMENTS IN TRADING OR INVESTMENT STRATEGIES BASED ON DIFFERENT VOLATILITY REGIMES.



Gaussian Mixture Model for Regime Classification

Fig C: Time Series plots for feature-futures-data-5min

CONCLUSION

THE DECISION TREE AND RANDOM FOREST MODELS ACHIEVED AN IN-SAMPLE ACCURACY OF 0.7, WHICH WAS IMPROVED TO 0.97 BY INCREASING LEAF NODES AND MAX DEPTH. HOWEVER, THE OUT-OF-SAMPLE ACCURACY REMAINED AROUND 0.5, INDICATING POOR GENERALIZATION TO UNSEEN DATA. CATEGORIZING AND ANALYZING VOLATILITY REGIMES PROVIDED VALUABLE INSIGHTS INTO MARKET DYNAMICS, ENABLING THE ADJUSTMENT OF TRADING OR INVESTMENT STRATEGIES.QUANTILE-BASED REGIMES AND MORE ADVANCED CLASSIFICATION METHODS WERE EXPLORED BUT A GAUSSIAN MIXTURE MODEL WAS ADOPTED AS A MORE PROMISING APPROACH IN PREDICTING DIRECTIONAL MOVEMENTS. CHALLENGES WERE FACED IN ADDRESSING OVERFITTING, IMPLEMENTING MANUAL INPUT OF DATA, IMPROVING OUTPUT PROCESSING, SIMPLIFYING DECISION TREE AND RANDOM FOREST MODELS.

ACKNOWLEDGMENTS

WE WOULD LIKE TO EXPRESS OUR SINCERE GRATITUDE AND APPRECIATION TO THE FOLLOWING INDIVIDUALS AND INSTITUTIONS FOR THEIR INVALUABLE SUPPORT AND CONTRIBUTION TO THIS PROJECT.
FIRST AND FOREMOST, WE EXTEND OUR GRATITUDE TO COST (COOPERATION IN SCIENCE AND TECHNOLOGY) AND COLUMBIA UNIVERSITY FOR THEIR GENEROUS SUPPORT, WHICH MADE THIS RESEARCH POSSIBLE.
WE WOULD LIKE TO EXTEND SPECIAL THANKS TO PROFESSOR HIRSA, PROFESSOR OSTERRIEDER, PROFESSOR HADZI MISHEVA, AND TEACHING ASSISTANT MIAO WANG FOR THEIR GUIDANCE, EXPERTISE, AND CONTINUOUS SUPPORT THROUGHOUT THE DURATION OF THIS PROJECT. THEIR VALUABLE INSIGHTS AND ADVICE WERE INSTRUMENTAL IN SHAPING OUR RESEARCH AND ENHANCING ITS QUALITY. WE WOULD ALSO LIKE TO ACKNOWLEDGE THE SIGNIFICANT CONTRIBUTIONS OF WASSIM BOUTABRATINE, RUBEN ILLOUZ, AND MARC AL HAJ. THEIR PRIOR WORK IN THIS RESEARCH PROJECT LAID A STRONG FOUNDATION AND PROVIDED VALUABLE INSIGHTS THAT PAVED THE WAY FOR OUR CURRENT STUDY.

FUTURE STEPS

\IN ORDER TO IMPROVE THE METHODOLOGY, SEVERAL RECOMMENDATIONS ARE PROPOSED. FIRSTLY, THE DATASET SHOULD BE SPLIT INTO TRAINING AND TESTING SETS USING A ROLLING OR EXPANDING WINDOW TECHNIQUE. THIS ALLOWS FOR EVALUATING THE MODEL'S PERFORMANCE ON UNSEEN DATA THAT CLOSELY RESEMBLES REAL-TIME SCENARIOS, ENABLING ASSESSMENT OF ITS ABILITY TO ADAPT TO CHANGING MARKET CONDITIONS.
ADDITIONALLY, FEATURE ENGINEERING AND SELECTION SHOULD BE EXPLORED FURTHER. CREATING NEW FEATURES OR TRANSFORMING EXISTING ONES CAN CAPTURE RELEVANT INFORMATION AND ENHANCE THE MODEL'S PREDICTIVE POWER. FEATURE SELECTION TECHNIQUES, SUCH AS RECURSIVE FEATURE ELIMINATION OR L1 REGULARIZATION, CAN HELP IDENTIFY THE MOST INFORMATIVE FEATURES FOR STOCK PRICE PREDICTION AND REGIME ANALYSIS.
ENSEMBLE METHODS, SUCH AS RANDOM FOREST, GRADIENT BOOSTING, OR MODEL STACKING, SHOULD BE CONSIDERED. THESE METHODS COMBINE MULTIPLE MODELS TO LEVERAGE THEIR INDIVIDUAL STRENGTHS AND IMPROVE OVERALL PERFORMANCE. STACKING, IN PARTICULAR, INVOLVES TRAINING MULTIPLE MODELS AND COMBINING THEIR PREDICTIONS USING ANOTHER MODEL AS A META-LEARNER, ENABLING BETTER CAPTURE OF COMPLEX PATTERNS IN STOCK PRICES AND REGIME CHANGES.
INTEGRATING EXTERNAL DATA SOURCES, SUCH AS MACROECONOMIC INDICATORS, FINANCIAL NEWS SENTIMENT, OR SOCIAL MEDIA SENTIMENT, CAN PROVIDE ADDITIONAL INSIGHTS. BY INCORPORATING THESE DATA SOURCES INTO THE MODELING PROCESS, THE MODEL'S ABILITY TO CAPTURE DYNAMIC MARKET CONDITIONS AND IMPROVE PREDICTIVE ACCURACY CAN BE ENHANCED.
IN SUMMARY, THE PROPOSED IMPROVEMENTS INCLUDE USING A ROLLING OR EXPANDING WINDOW TECHNIQUE FOR DATASET SPLITTING, EXPLORING ADVANCED FEATURE ENGINEERING AND SELECTION TECHNIQUES, CONSIDERING ENSEMBLE METHODS AND MODEL STACKING, AND INCORPORATING EXTERNAL DATA SOURCES AND SENTIMENT ANALYSIS. THESE ENHANCEMENTS AIM TO ENHANCE THE MODEL'S PERFORMANCE AND ABILITY TO CAPTURE REAL-TIME MARKET DYNAMICS.