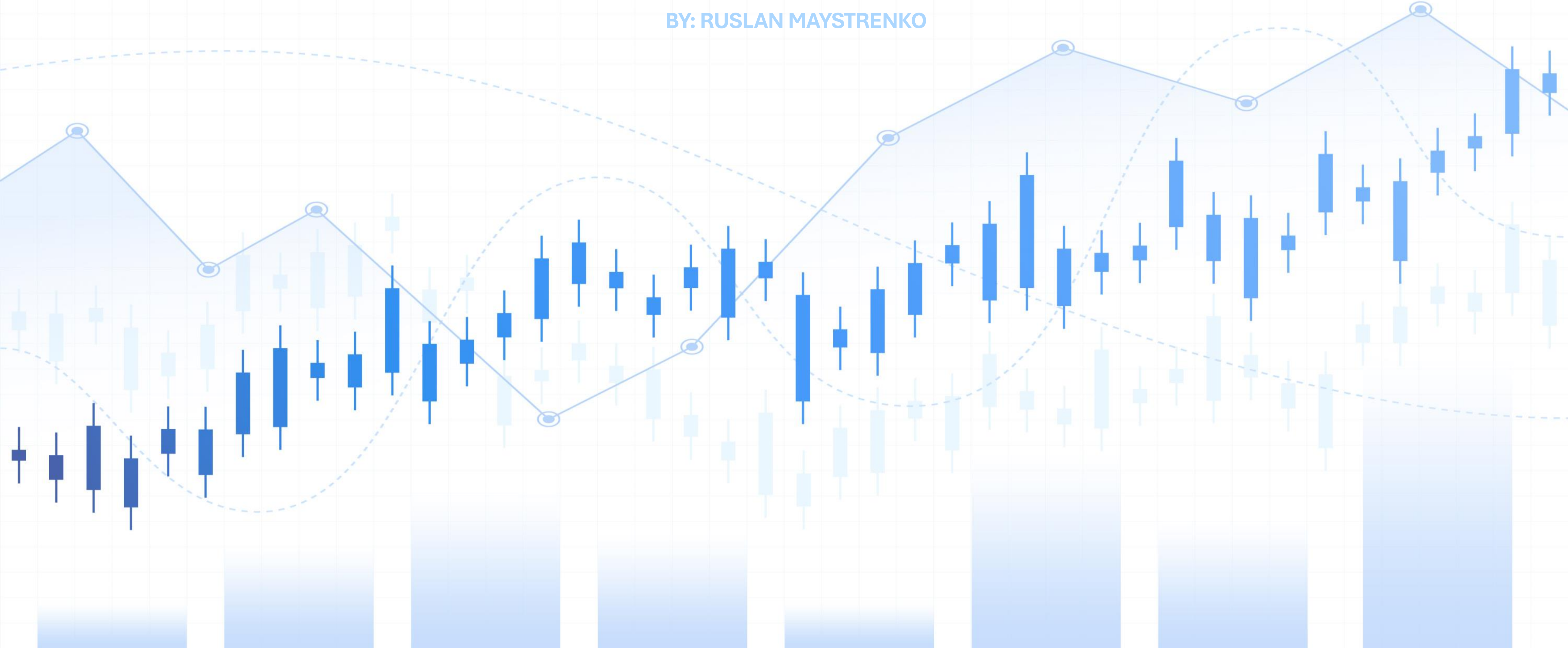
EMA Crossover Strategy Profitability Prediction

BY: RUSLAN MAYSTRENKO



Problem Statement

Problem

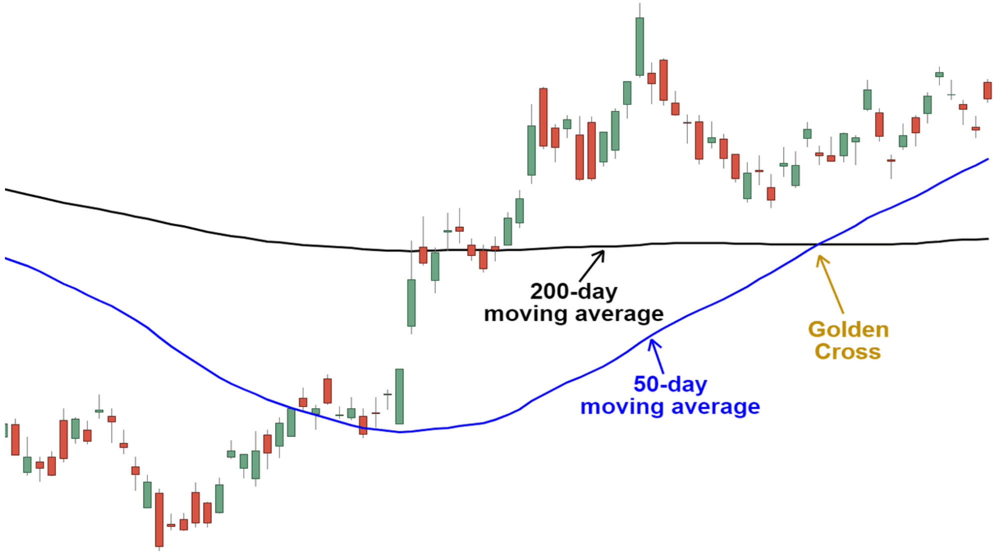
As EMA50 crosses over EMA200 a “Golden Cross” is formed which are widely used in technical analysis to generate buy signals, but not all result in profitable trades, leading to inconsistent returns and increased risk for traders

Objective

To develop a machine learning model to predict the success probability of EMA crossover events, labeling them as "good" (profitable, >3% return in 10 days) or "bad" (unprofitable), to enable better-informed trading decisions

Impact

Provides traders with a systematic, predictive tool to reduce risk and enhance profitability by filtering high-probability crossovers



Financial & Technical Foundations

Key Financial Terms

- **Ticker:** Unique identifier for stocks (e.g., AAPL, TSLA)
- **OHLCV:** Open, High, Low, Close prices, and Volume, foundational for price analysis and indicator calculations

Technical Indicators Overview

- **CCI (20-day):** Identifies overbought/oversold conditions by measuring price deviation from its mean
- **MACD:** Tracks momentum via the difference between 12-day and 26-day EMAs; the signal line (9-day EMA of MACD) and histogram (MACD - Signal) highlight trend shifts
- **RSI (14-day):** Momentum oscillator; values >70 (overbought) or <30 (oversold) indicate potential price corrections
- **EMA:** Weighted moving average emphasizing recent prices, used for trend identification (EMA50, EMA200)
- **ATR, HMA, Bollinger Bands:** Measure volatility, reduce lag in trend detection, and identify breakout points, respectively

These indicators provide the raw signals and features for predictive modeling, capturing momentum, volatility, and trend dynamics

Data Collection & Feature Engineering

Dataset Overview

- The data was sourced by first scraping S&P 500 ticker symbols from https://www.slickcharts.com/sp500
- Historical price data for these tickers, spanning January 1, 2010, to April 22, 2025, was then downloaded from Yahoo Finance resulting in a multi-indexed data frame for analysis containing:
    - Date, Ticker, Open, High, Low, Close, Volume

Feature Engineering

- Engineered 21 technical features to capture price dynamics:
    - **EMA Variants:** EMA10, EMA20, EMA50, EMA200 for trend analysis
    - **Momentum Indicators:** RSI_14, MACD, MACD_diff, CCI_20 for momentum and mean-reversion signals
    - **Volatility Indicators:** ATR_14, Bollinger Bands (BB_MID, BB_UPPER, BB_LOWER) for market volatility
    - **Custom Features:** ema_distance (EMA50 - EMA200), price_ema200_dist, and momentum_5 to quantify trend strength and price deviations
- Labeling:
    - **Crossover Detection:** Identified when EMA50 crosses above EMA200 using previous day EMA values (EMA50_prev, EMA200_prev)
    - **Target:** Binary label — 1 for "good" crossovers (10-day forward return >3%), 0 for "bad" (otherwise)

Date	Ticker	Open	High	Low	Close	Volume	EMA10	EMA20	EMA50	EMA200	momentum_5	ema_distance	ema50_slope	price_ema200_dist	EMA50_prev	EMA200_prev	RSI_14	MACD	MACD_signal	macd_diff	ATR_14	CCI_20	BB_MID	BB_STD	BB_UPPER	BB_LOWER	HMA_20
1/26/2012	CPAY	34	34	33.63	33.76	176600	33.117	32.21276	30.62883	29.71222	0.015643701	0.91661068	0.665551467	4.047775419	30.5010292	29.66890781	80.56	0.53	0.64671268	-0.116814	0.6121	103.42	31.997	1.427	34.85098	29.143016	34.12213
1/26/2012	JBL	20.19	20.2	19.49	19.68	3012200	19.54516	18.9702	18.12098	16.82132	-0.022894178	1.29966248	0.41480219	2.855124582	18.05749466	16.79246654	61.14	0.524	0.622240225	-0.09789	0.5051	82.09	18.701	1.082	20.8648	16.536979	20.20609
1/26/2012	ROK	60.99	61.6	58.92	59.12	1599200	60.51359	59.51335	57.09784	54.76525	-0.051375808	2.332590636	0.876207255	4.358502066	57.0151509	54.72120351	41.9	0.516	0.601076348	-0.084656	1.7595	25.974	59.038	2.6124	64.2623	53.812882	63.15396
1/26/2012	ATO	22.92	23.2	22.86	23.16	286400	22.8583	22.90328	23.04049	22.75223	0.026749463	0.28825561	-0.036668249	0.410577657	23.0354965	22.74808433	62.32	0.525	0.585873815	-0.06081	0.3233	26.972	22.942	0.3066	23.55505	22.328513	22.51603
1/26/2012	MPWR	14.34	14.9	13.88	14.38	158000	14.21265	13.72461	12.63871	12.0007	-0.002985178	0.638005396	0.444093754	2.376478486	12.56774988	11.97668588	60.2	0.648	0.598279652	0.049623	0.5061	80.553	13.662	0.721	15.1042	12.220196	14.46962
1/26/2012	DHI	12.63	12.8	12.19	12.26	9943400	12.14897	11.82197	11.06594	10.0312	0.004982369	1.034746961	0.265555056	2.231013134	11.01711653	10.00865014	58.2	0.739	0.626454559	0.1127	0.4578	93.344	11.776	0.5645	12.90488	10.646897	12.62955
1/26/2012	MAS	9.157	9.41	8.804	8.84	8567547	8.795358	8.434746	7.672877	7.428978	-0.046692909	0.243898687	0.269762284	1.410688855	7.625253064	7.414722045	57.09	0.783	0.657779334	0.125299	0.3732	82.748	8.4411	0.5813	9.603779	7.2783947	9.324375
1/26/2012	CAG	13.69	13.8	13.55	13.6	3368242	13.57535	13.45957	13.14125	12.46873	-0.005150718	0.672512689	0.103857151	1.13296031	13.12245366	12.45728505	58.44	0.899	0.73744549	0.161658	0.1473	85.678	13.475	0.1515	13.77844	13.172426	13.70475

Modelling Approach

Model Selection:

- Focused on **tree-based ensemble models**: LightGBM, Decision Tree, Random Forest, XGBoost, CatBoost
- **Excluded** SVM and Logistic Regression due to their poor performance on tabular financial data with non-linear relationships

Evaluation Focus:

- **Prioritized recall for class 1** ("good" crossovers) to maximize capture of profitable trades, as missing opportunities (false negatives) directly reduces returns

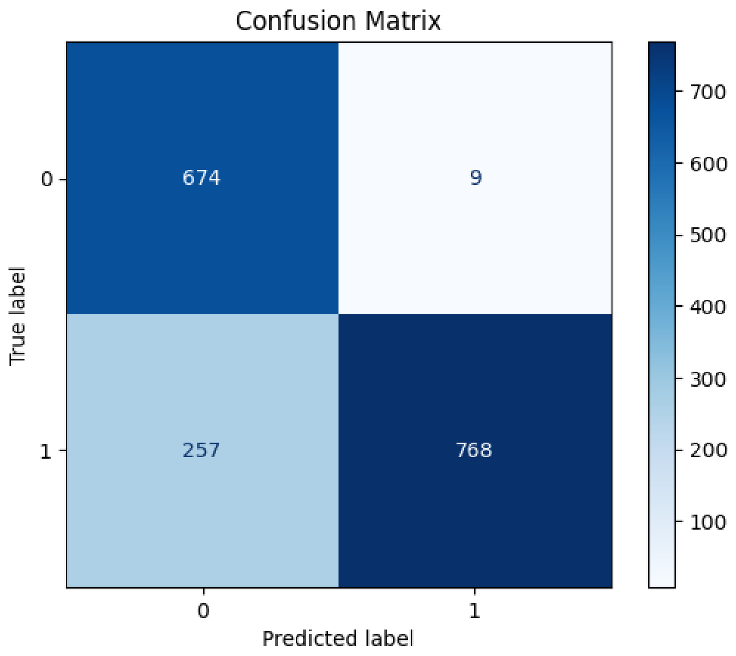
Modelling Process:

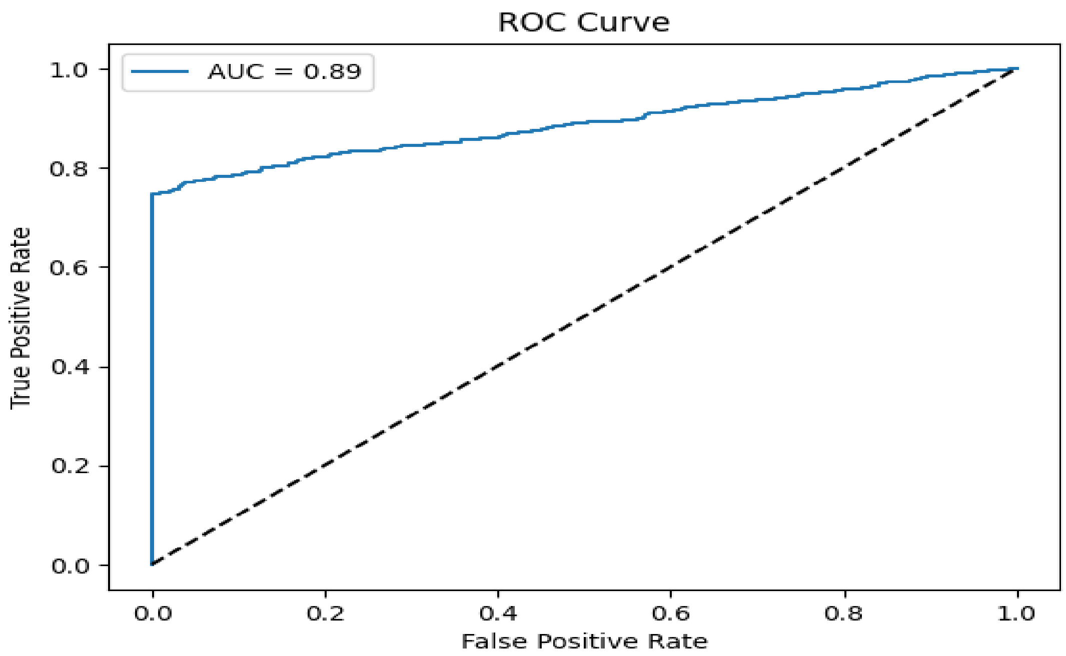
- Compared classification reports of untuned classifiers and hyperparameter tuned classifiers
- Hyperparameter tuning was done using `RandomizedSearchCV` with the scoring attribute set to 'f1'
- After evaluating all metrics, **the untuned LightGBM model remains the champion with a higher recall for class 1 at 0.78**

Champion Model: LightGBM

	Class 0	Class 1
Precision	*0.74*	*0.94*
Recall	*0.93*	*0.78*
F1-Score	*0.82*	*0.85*
Support	*683*	*1025*
Accuracy	*0.84*	
Macro Avg Precision	*0.84*	
Macro Avg Recall	*0.85*	
Macro Avg F1-Score	*0.84*	
Weighted Avg Precision	*0.86*	
Weighted Avg Recall	*0.84*	
Weighted Avg F1-Score	*0.84*	

LightGBM Analysis

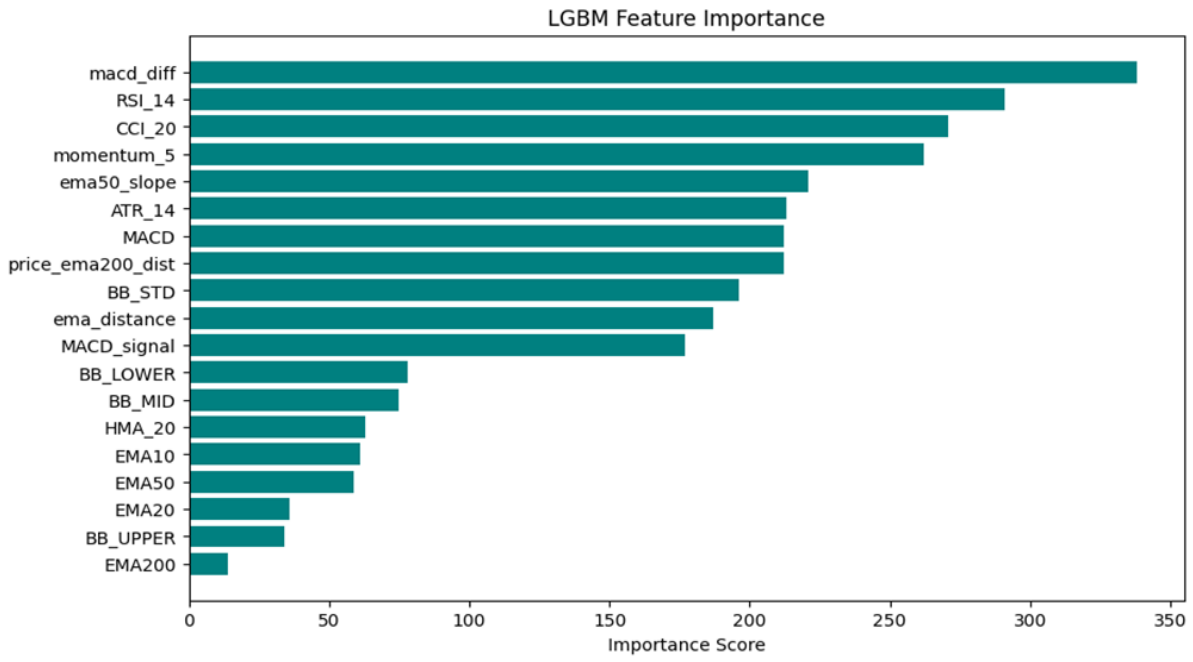


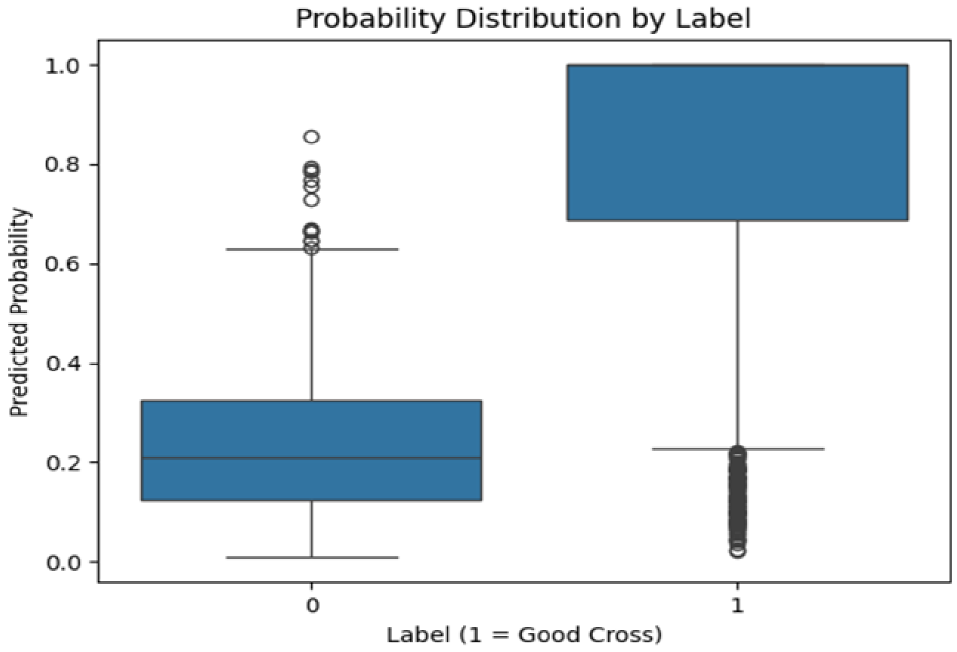


This matrix shows that the model correctly classified 674 of the class 0 (underperforming) cases and 768 of the class 1 (outperforming) cases, with a moderate number of false positives 9 and false negatives 257

An AUC score of 0.89 indicates strong separability between the two classes, showing that the model is capable of distinguishing good trading signals from poor ones.

LightGBM Analysis





This feature importance shows indicators like `macd_diff`, `RSI_14`, and `CCI_20` were the most influential in determining the class label, highlighting which technical indicators had the strongest predictive value.

This boxplot shows class 1 examples have significantly higher predicted probabilities compared to class 0, confirming that the model is correctly assigning higher confidence to truly positive cases while still maintaining a reasonable spread.

Simulated Backtest

With our selected model we now want to test our predictions in a simulated backtest, here's how it will flow:

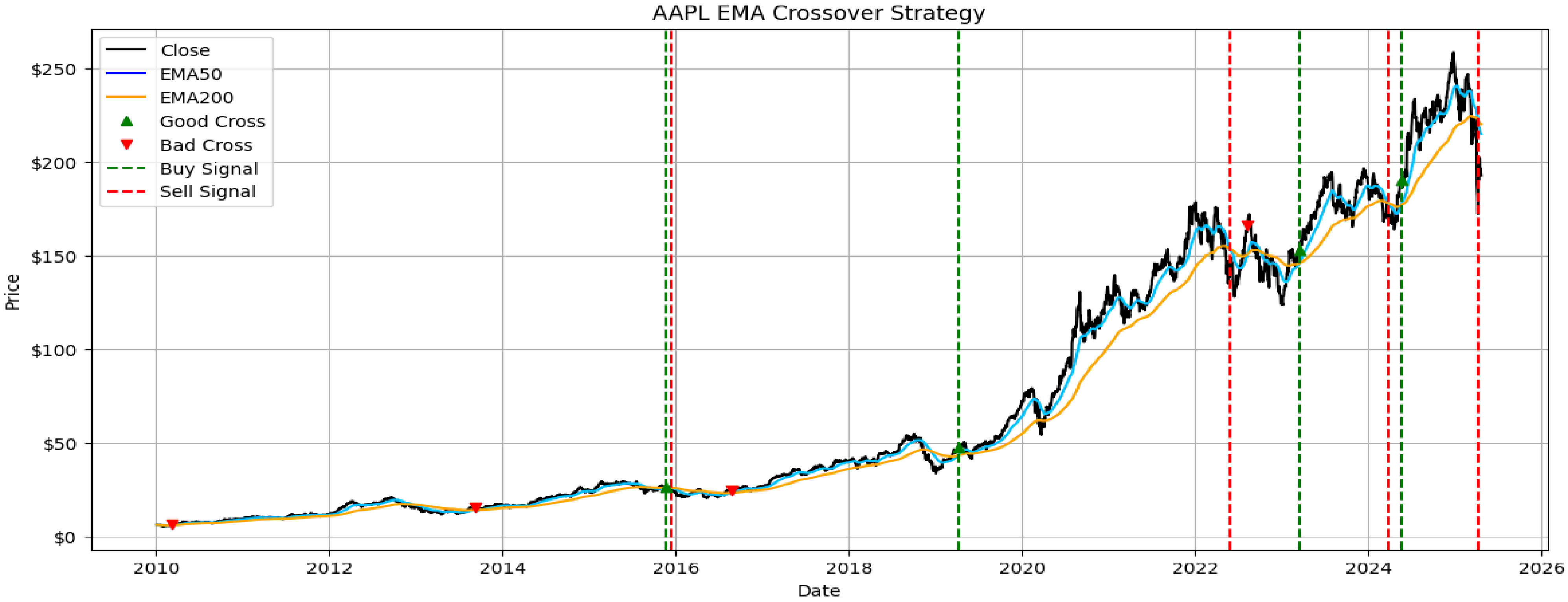
- The backtest will be run on one stock, it will not be a portfolio backtest
- Initial cash is set to $10,000
- A buy signal will be generated depending if the LGBM model predicted a good cross (class 1)
- A sell signal will be generated once EMA50 goes equal or below EMA200 (EMA50 <= EMA200)
- For every buy signal, purchase as much stock with the cash available (all in)
- For every sell signal, sell all stock

Simulation Objective

- The goal of this simulation is to see if the model is able to learn the **patterns** that make a profitable EMA crossover
- The primary objective is **not** to maximize cumulative return; however, refining sell signal diagnostics and integrating short-term strategies could significantly enhance overall profitability

We will be looking at the performance of AAPL, TSLA, and META stock

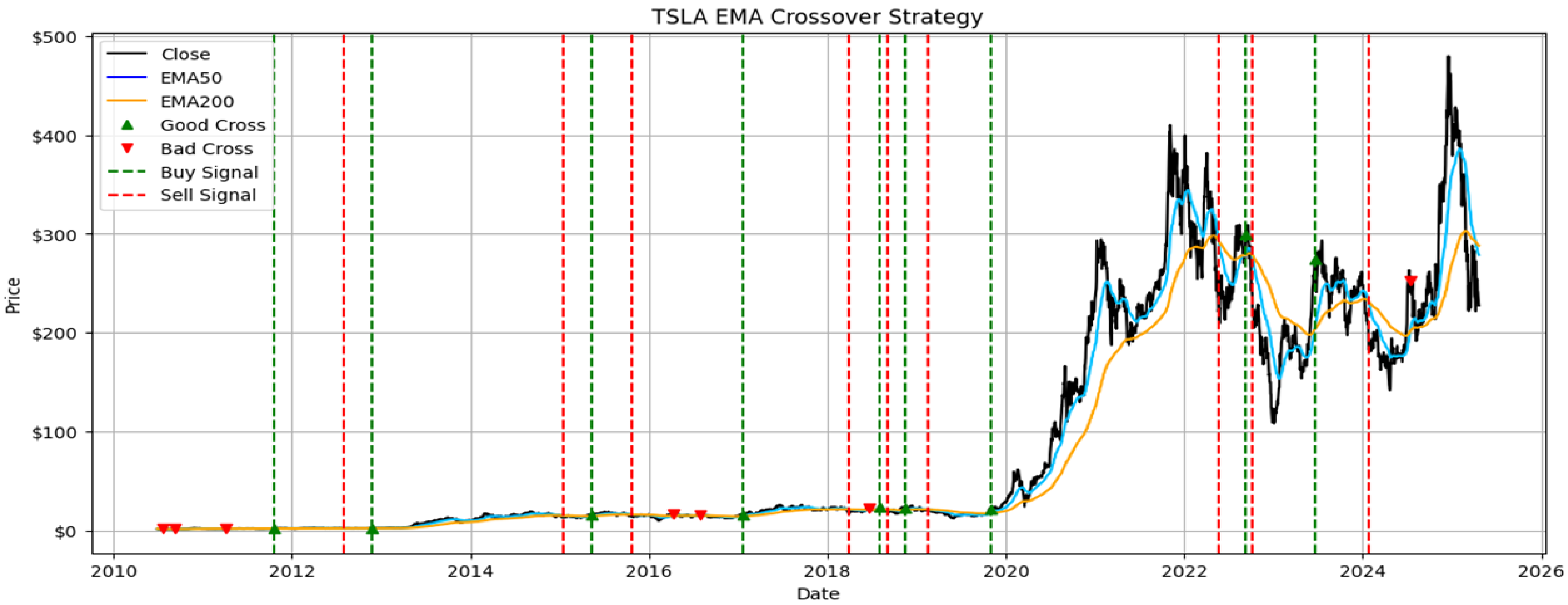
AAPL




```
Final Portfolio Value: $28,453.67
Cumulative Return: 184.54%
```

For AAPL the model performed fairly well accurately predicting the huge trends in 2019, 2023, and 2024 but making a small error in around late 2016. The cumulative return percentage would definitely be higher if our sell signal wasn't as lagging

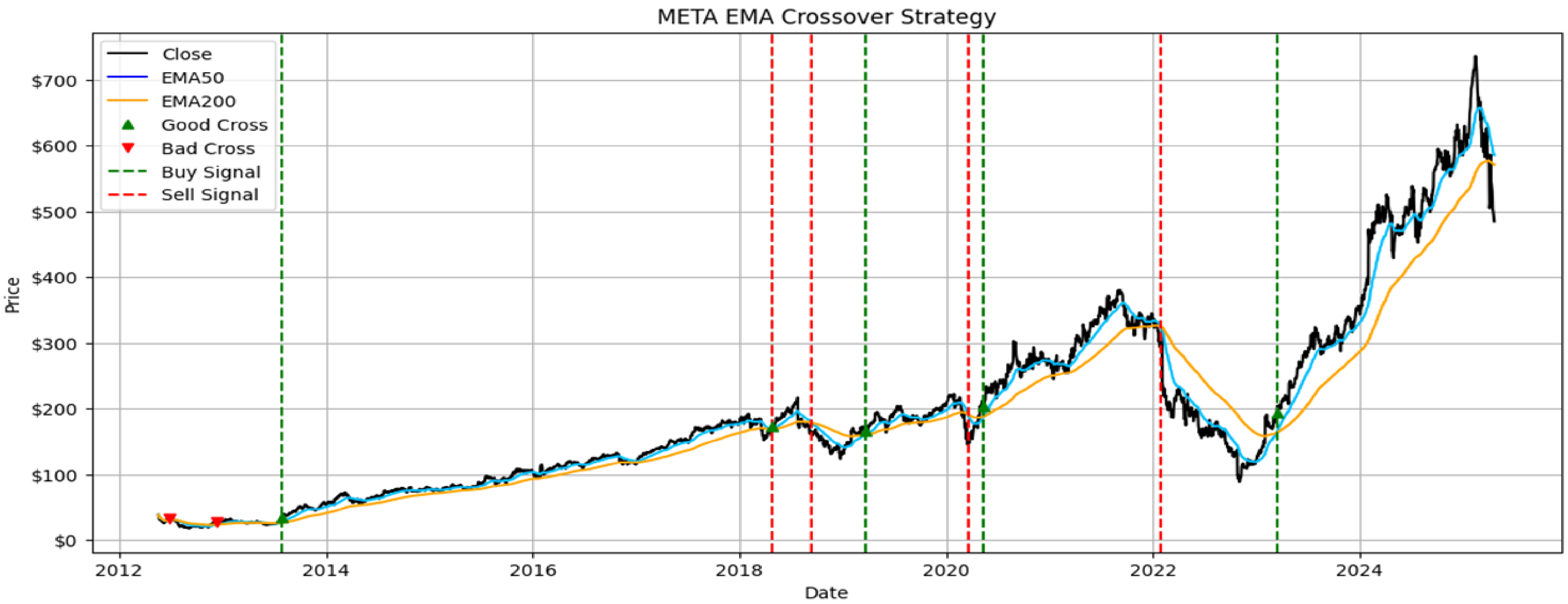
TSLA



Final Portfolio Value: $211,599.71
Cumulative Return: 2016.00%

For TSLA our model made more mistakes but accurately predicted the huge leap in late 2019. The big mistake that stands out is the bad cross in 2024, although this could be due to the stocks high volatility which the model seems not to perform well in

META




```
Final Portfolio Value: $142,120.58
Cumulative Return: 1321.21%
```

For META the detection of good crosses appears to be as intended accurately capturing profitable crosses. In 2018 we see a good cross and then a sell signal almost immediately afterwards this is probably due to EMA50 and EMA200 being extremely close to one another – more in-depth research would have to be done to be certain.

Limitations & Recommendations

Limitations

- **Sell signals lag** potentially missing optimal exit points
- Model **struggles with high-volatility** stocks, leading to prediction errors
- **Crossovers are infrequent**, limiting short-term trading opportunities.

Recommendations

- **Refine sell signals** to capture earlier exits
- Incorporate **macroeconomic or industry-specific data** to better handle volatility
- **Combine with short-term strategies** to increase trade frequency and cumulative returns
- Deploy as a **live trading tool** or **stock screener** to scan and predict crossover probabilities in real time.

Summary & Strategic Value

Summary

- This project explored a data driven approach to identifying **profitable** EMA crossover events using **machine learning.** By engineering features from common technical indicators and training a LightGBM classifier, we were able to **isolate high probability "good" crossovers** across S&P 500 stocks
- This serves as a **strong foundation** for building machine learning **enhanced technical trading systems**, transforming a basic EMA crossover into a more robust and informed decision engine

Strategic Value

- Transforms a basic EMA crossover into a **data-driven decision engine**
- **Empowers traders** to focus on high-probability trades, reducing risk and enhancing returns
- Provides a **scalable framework** for live trading and stock screening applications

Next Steps

Explore tuning for volatility, integrate short-term strategies, and deploy in real-time trading systems