A Model Evaluation on LightGBM vs. Catboost:

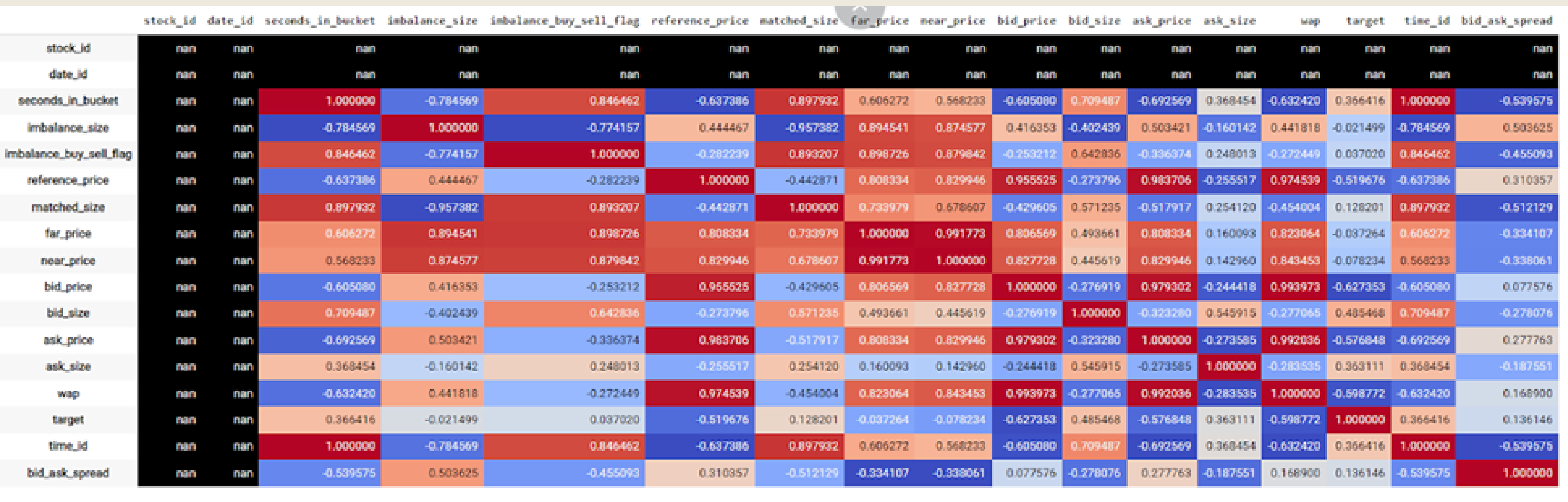
A Study on Nasdaq-listed Stock Closing Auction Price Prediction

JUNWEI SHEN & CHE ZHU

Introduction

- Main Objective: This study will utilize two cutting-edge gradient boosting models, LightGBM and Catboost, to predict the future price movements of stocks through a synthetic index composed of NASDAQ-listed stocks.
- Major Approach: We take a methodical approach encompassing data preprocessing, feature engineering, and model training and tuning, and we strive to minimize the mean absolute error (MAE).
- Responsible AI development: model comparison analysis of the two gradient-boosting models.

Fig 1: Nasdaq Stock Market



	stock_id	date_id	seconds_in_bucket	imbalance_size	imbalance_buy_sell_flag	reference_price	matched_size	far_price	near_price	bid_price	bid_size	ask_price	ask_size	wap	target	time_id	bid_ask_spread
stock_id	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
date_id	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
seconds_in_bucket	nan	nan	1.000000	-0.784569	0.846462	-0.637386	0.897932	0.606272	0.568233	-0.605080	0.709487	-0.692569	0.368454	-0.632420	0.366416	1.000000	-0.539575
imbalance_size	nan	nan	-0.784569	1.000000	-0.774157	0.444467	-0.957382	0.894541	0.874577	0.416353	-0.402439	0.503421	-0.160142	0.441818	-0.021499	-0.784569	0.503625
imbalance_buy_sell_flag	nan	nan	0.846462	-0.774157	1.000000	-0.282239	0.893207	0.898726	0.879842	-0.253212	0.642836	-0.336374	0.248013	-0.272449	0.037020	0.846462	-0.455093
reference_price	nan	nan	-0.637386	0.444467	-0.282239	1.000000	-0.442871	0.808334	0.829946	0.955525	-0.273796	0.983706	-0.255517	0.974539	-0.519676	-0.637386	0.310357
matched_size	nan	nan	0.897932	-0.957382	0.893207	-0.442871	1.000000	0.733979	0.678607	-0.429605	0.571235	-0.517917	0.254120	-0.454004	0.128201	0.897932	-0.512129
far_price	nan	nan	0.606272	0.894541	0.898726	0.808334	0.733979	1.000000	0.991773	0.806569	0.493661	0.808334	0.160093	0.823064	-0.037264	0.606272	-0.334107
near_price	nan	nan	0.568233	0.874577	0.879842	0.829946	0.678607	0.991773	1.000000	0.827728	0.445619	0.829946	0.142960	0.843453	-0.078234	0.568233	-0.338061
bid_price	nan	nan	-0.605080	0.416353	-0.253212	0.955525	-0.429605	0.806569	0.827728	1.000000	-0.276919	0.979302	-0.244418	0.993973	-0.627353	-0.605080	0.077576
bid_size	nan	nan	0.709487	-0.402439	0.642836	-0.273796	0.571235	0.493661	0.445619	-0.276919	1.000000	-0.323280	0.545915	-0.277065	0.485468	0.709487	-0.278076
ask_price	nan	nan	-0.692569	0.503421	-0.336374	0.983706	-0.517917	0.808334	0.829946	0.979302	-0.323280	1.000000	-0.273585	0.992036	-0.576848	-0.692569	0.277763
ask_size	nan	nan	0.368454	-0.160142	0.248013	-0.255517	0.254120	0.160093	0.142960	-0.244418	0.545915	-0.273585	1.000000	-0.283535	0.363111	0.368454	-0.187551
wap	nan	nan	-0.632420	0.441818	-0.272449	0.974539	-0.454004	0.823064	0.843453	0.993973	-0.277065	0.992036	-0.283535	1.000000	-0.598772	-0.632420	0.168900
target	nan	nan	0.366416	-0.021499	0.037020	-0.519676	0.128201	-0.037264	-0.078234	-0.627353	0.485468	-0.576848	0.363111	-0.598772	1.000000	0.366416	0.136146
time_id	nan	nan	1.000000	-0.784569	0.846462	-0.637386	0.897932	0.606272	0.568233	-0.605080	0.709487	-0.692569	0.368454	-0.632420	0.366416	1.000000	-0.539575
bid_ask_spread	nan	nan	-0.539575	0.503625	-0.455093	0.310357	-0.512129	-0.334107	-0.338061	0.077576	-0.278076	0.277763	-0.187551	0.168900	0.136146	-0.539575	1.000000

Fig 2: Correlation Heatmap of Features and Outcome: There are several correlations between different variables; besides relatively highly correlated with bid_price, ask_price, bid_size, and reference_price, the target variable also has a medium-level correlation with time_id

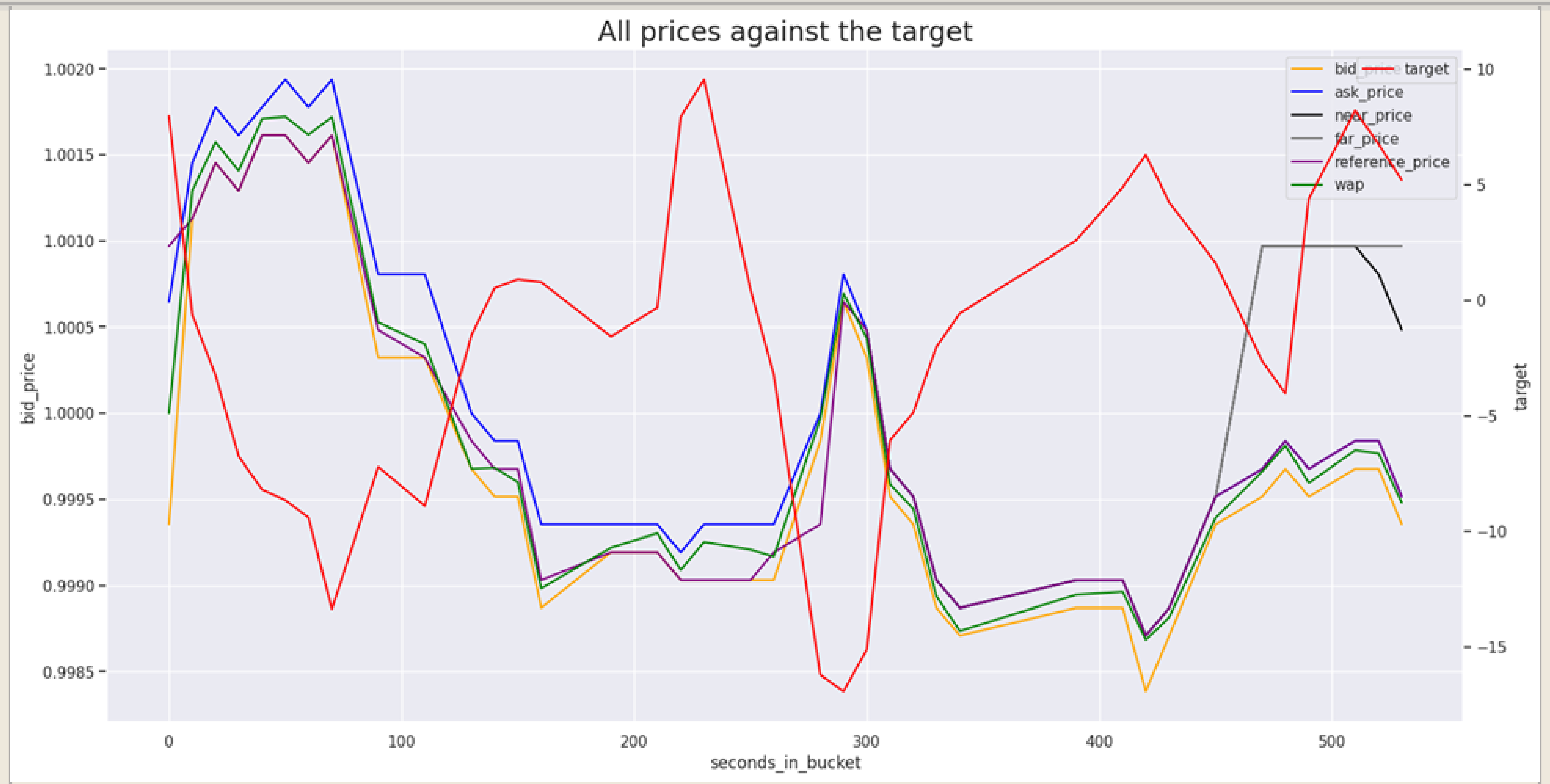


Fig 3: Different Price Measurements Change over Time: the target line shows a significant divergence from the price lines; the target line fluctuating in a different scale implies that its behavior or scale is different from that of the prices

Fig 4: LightGBM and Catboost Model Accuracy Results: Compared to a variance of the target (59.797), global minimum (-103.030), and global maximum (113.180), both models have relatively small MAE; The Catboost model slightly outperforms the LightGBM model with an MAE of 4.892

LightGBM Final MAE score	4.986
Catboost Final MAE score	4.892
Testing Set Target variance	59.797
Testing Set Target Range	(-103.030, 113.180)

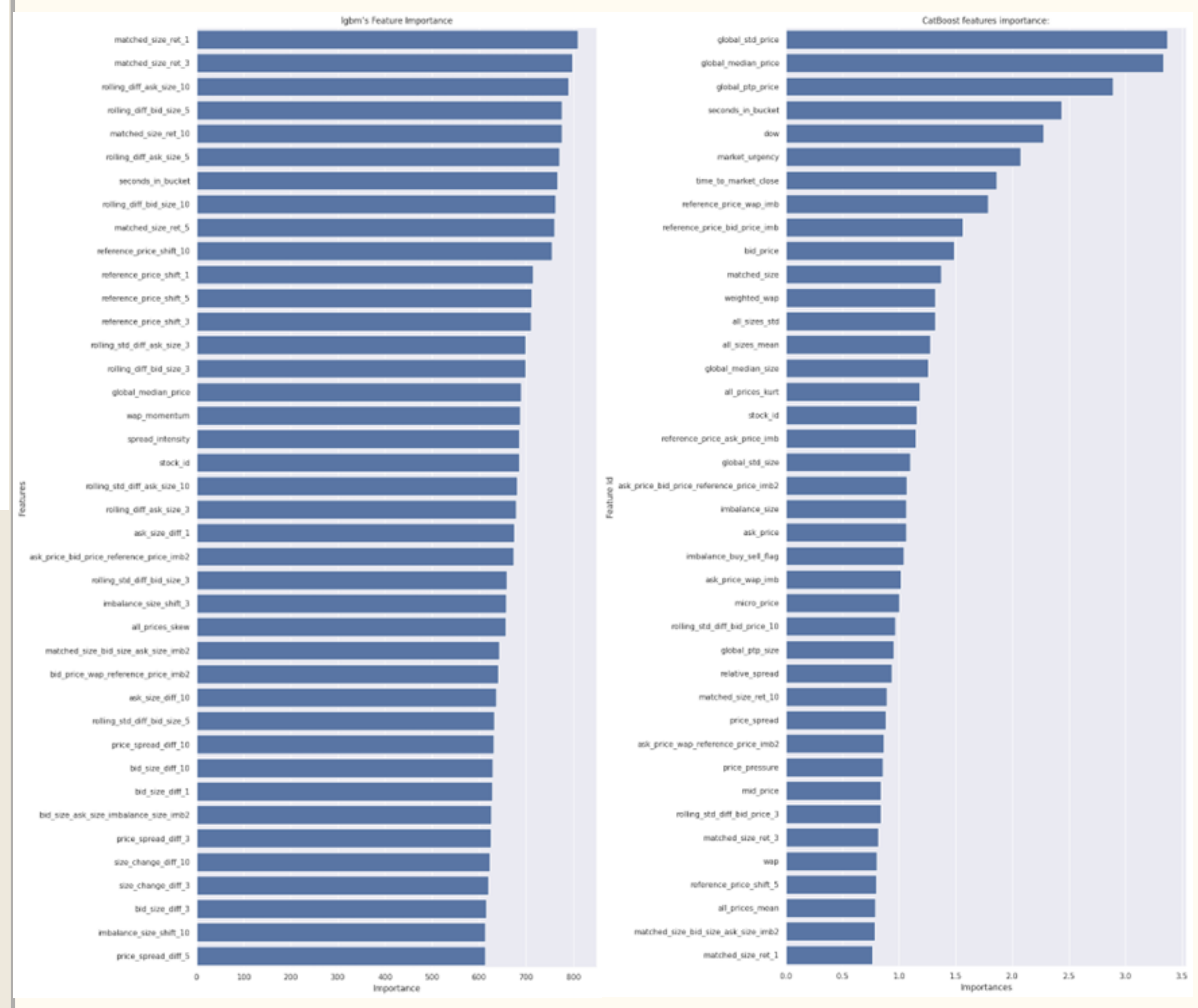


Fig 5: LightGBM (Left) and Catboost(Right) Feature Importance Result: Certain features are highly ranked in both results, implying that these features are strong predictors regardless of the model type. However, some features are given different significance levels in the ranking, this is the result of the algorithms' functionalitiess

Model Comparsion

CatBoost and LightGBM Overview:

- Part of the gradient boosting family, known for efficiency, speed, and performance.
- Distinct in optimization and handling of categorical data.

Performance Implications:

- CatBoost: Better for datasets with correlated features (e.g., auction data), due to ordered boosting and symmetric trees techniques.
- LightGBM: Offers similar accuracy levels with less computing time, ideal for complex ML tasks with large datasets.

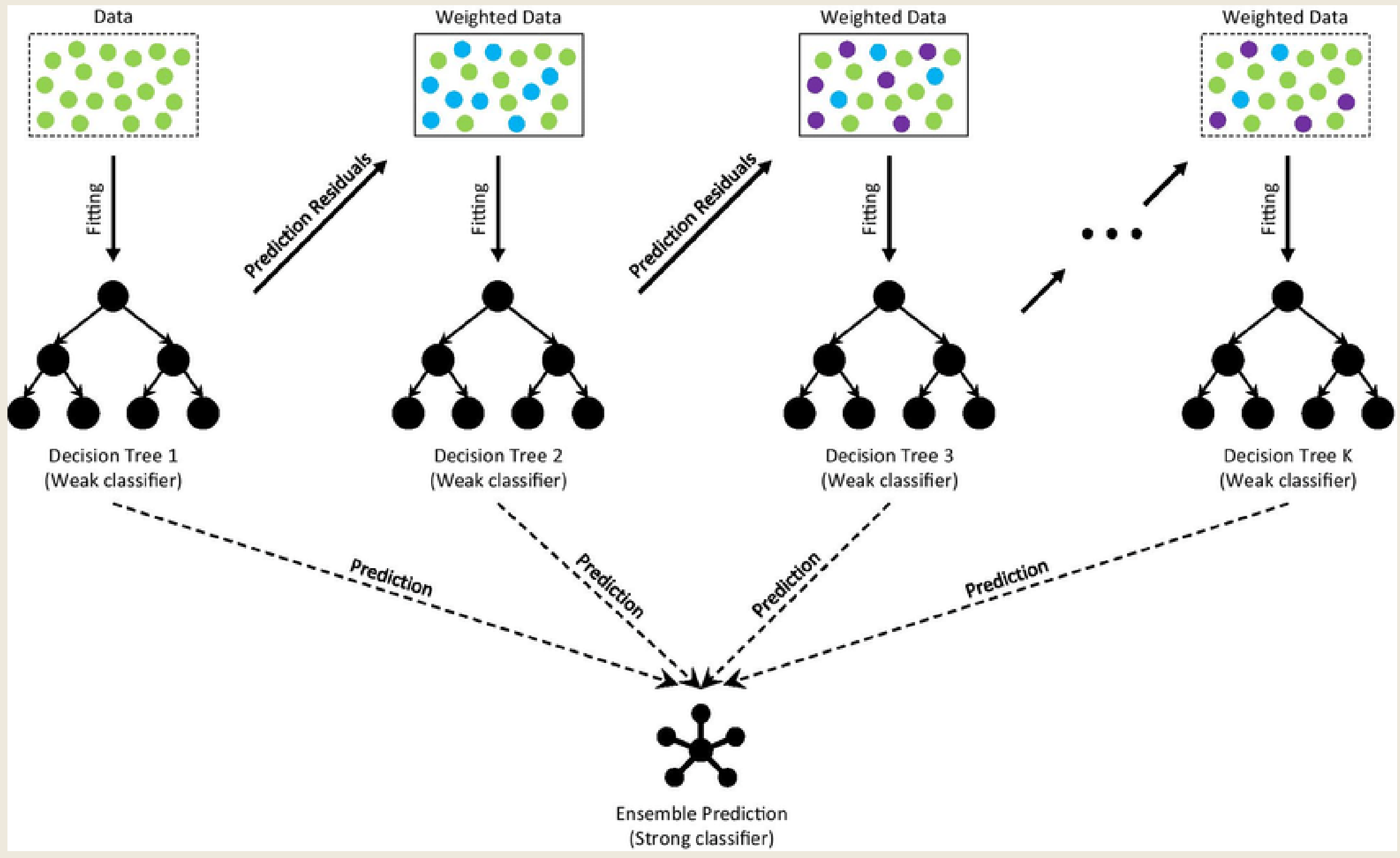


Fig 6: The architecture of Gradient Boosting Decision Tree

Bias Consideration

Stock market prediction models, influenced by human actions and major events, face challenges in adapting to rapid changes and biases. Despite high data transparency, selective data use and historical assumptions can mislead and reduce generalization. It's crucial to account for these factors to ensure fair and accurate market analysis