Group 18

IPO QUANTITATIVE ANALYSIS AND LISTING PREDICTOR

AKSHITHA, DIVY GUPTA, MOKSHITH, NAGAVENI

INDEX

1.Problem Statement

2.Literature Survey

3.Dataset and Features Preprocessing

4.ML Methodology

5.Performance Metrics &
Deployability of ML Solution

PROBLEM STATEMENT

"To predict IPO listing gains in the Indian stock market by analysing firm fundamentals together with investor demand and market sentiment"

RHP (Red Herring Prospectus):

RHP is an official document released by a company before launching its IPO that provides important information to investors.

GMP (Grey Market Premium) :

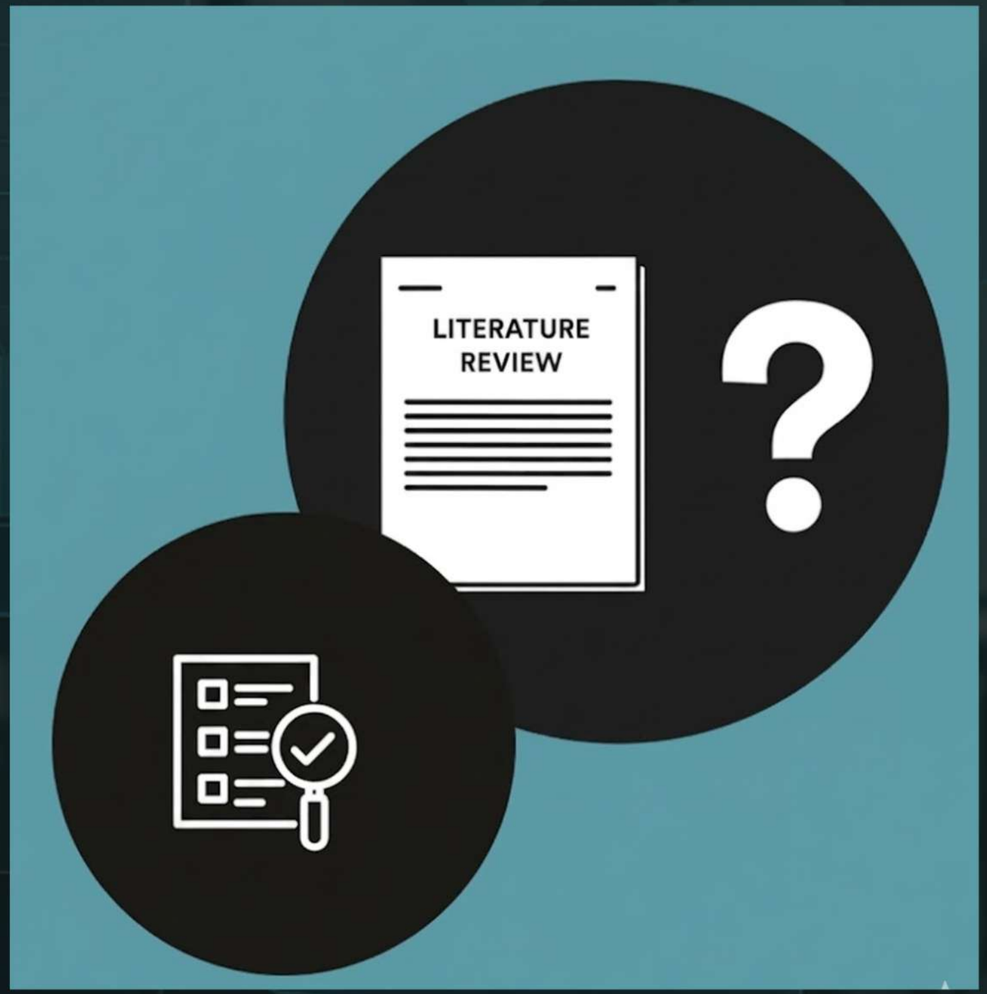
GMP refers to the unofficial price at which IPO shares trade in the grey market before the stock is listed on the exchange.

WHY IT MATTERS?

- To improve market efficiency , capital allocation
- Retail investors often rely on hype rather than analysis
- Reduces market mispricing
- Generally IPO markets involve high uncertainty - if prospectus language can predict performance, it reduces uncertainty and improves decision making(words influences the investor behaviour)

Group 18

LITERATURE SURVEY



PHASE 1 AND PHASE 2

PHASE	AUTHOR & YEAR	PAPER CITATIONS	KEY FINDINGS	KEY VALUE
PHASE 1 Underpricing (1990s – Early 2000s)	Ritter (1991)	Ritter, J. R. (1991). The long-run performance of initial public offerings. Journal of Finance, 46(1), 3–27.	IPOs give strong 1st-day returns but underperform long-run vs seasoned firms	Established listing-day prediction is worthwhile
	Ritter & Welch (2002)	Ritter, J. R., & Welch, I. (2002). A review of IPO activity, pricing, and allocations. Journal of Finance, 57(4), 1795–1828.	Standard pricing models (DCF, comparables) cannot explain IPO underpricing	Avg 1st-day return = 18.8% (US, 1980–2001)
PHASE 2 Behavioural & Sentiment (2000 – 2015)	Loughran & Ritter (2004)	Loughran, T., & Ritter, J. R. (2004). Why has IPO underpricing changed over time? Financial Management, 33(3), 5–37.	Psychological biases (anchoring, prospect theory) cause issuers to accept below-market prices	Rational pricing cannot explain underpricing
	Cornelli, Goldreich & Ljungqvist (2006)	Cornelli, F., Goldreich, D., & Ljungqvist, A. (2006). Investor sentiment and pre-IPO markets. Journal of Finance, 61(3), 1187–1216.	Grey market prices predict first-day returns high GMP = investor overexcitement, not fundamental value	GMP contradicts the Efficient Market Hypothes
	Kim & Ritter (1999)	Kim, M., & Ritter, J. R. (1999). ValuingIPOs. Jou rnal of Financial Economics, 53(3), 409–437.	Financial ratios from the prospectus (P/E, P/Sales, P/Book) explain cross-sectional IPO pricing	Validated RHP fundam -entals as predictors
	Shetty et al. (2023)	Shetty, C., Vinish, P., Aluru, S., Pinto, P., & Hawaldar, I. T. (2023). IPO subscription dynamics: A comprehensive inquiry into the Indian stock market. Investment Management and Financial Innovations, 20(4), 400–415.	QIB subscription alone is the strongest predictor of Listing gains in India	QIB explains ~48% of list ing -gain variance
	Lowry (2003)	Lowry, M. (2003). Why does IPO volume fluctuate so much? Journal of Financial Economics, 67(1), 3–40.	IPO volume fluctuates with investor sentiment and adverse selection costs	Justifies IPO_Year & market-timing features in model

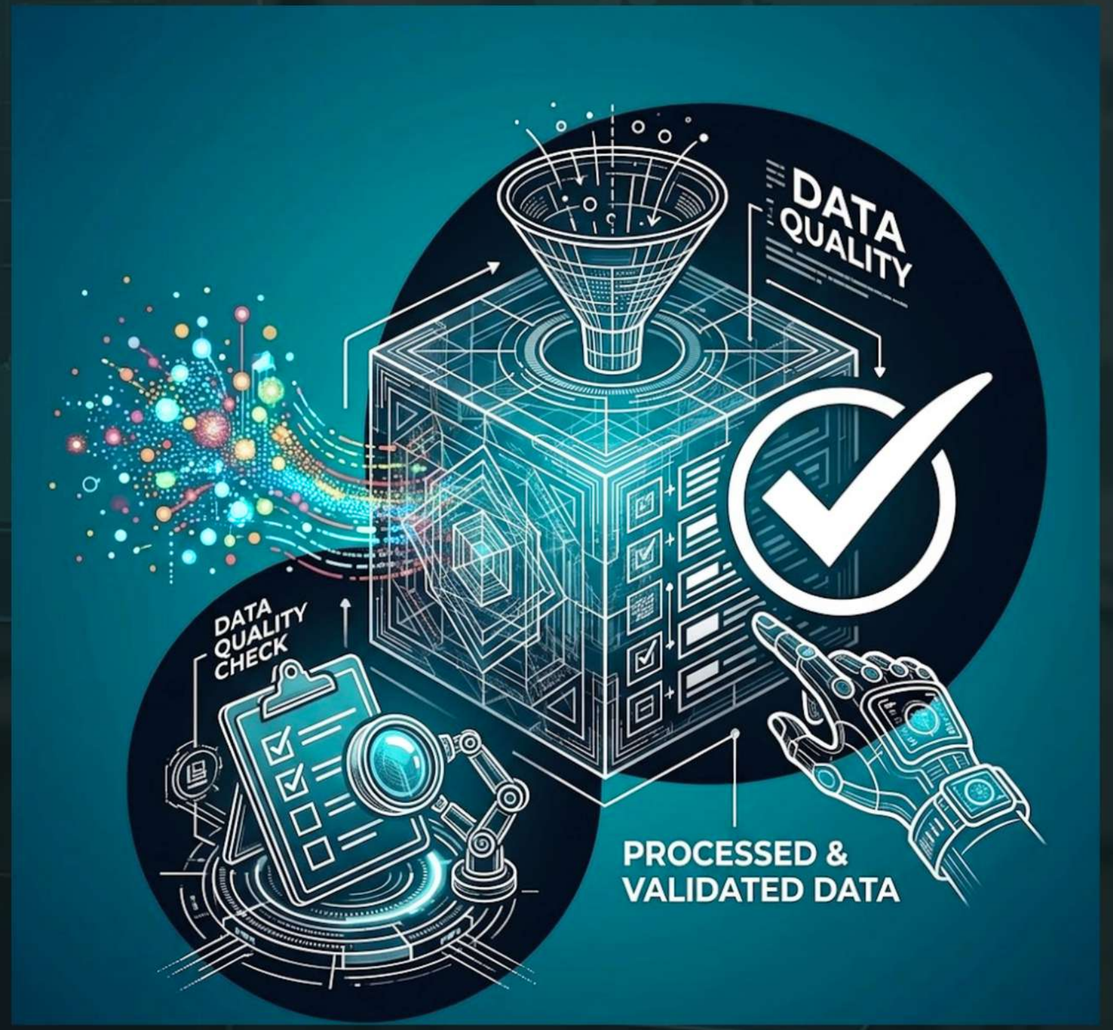
PHASE 3

PHASE	AUTHOR & YEAR	PAPER CITATIONS	KEY FINDINGS	KEY VALUE
PHASE 3 Machine Learning Takes Over (2015 – Present)	Gu, Kelly & Xiu (2020)	Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. Review of Financial Studies, 33(5), 2223–2273.	Tree models & neural nets capture non-linear interactions OLS misses entirely	+6.2% predictive R^2 over OLS 0.4% Out of Sampe R^2
	Baba & Sevil (2020)	Baba, B., & Sevil, G. (2020). Predicting IPO initial returns using random forest. Borsa İstanbul Review, 20(1), 13–23.	Random Forest outperforms OLS on 1st-day return prediction	RF RMSE 0.148 vs OLS RMSE 0.187
	Katsafados et al. (2023)	Katsafados, A. G., Leledakis, G. N., Pyrgiotakis, E. G., Androutsopoulos, I., Chalkidis, I., & Fergadiotis, E. (2023). Textual information and IPO underpricing: A machine learning approach. Journal of Financial Data Science, 5(2), 100–135.	Prospectus text + financials outperforms either alone; NLP adds measurable signal	Best accuracy 73.6% Text adds +6.1pp
	Alahmadi & Yilmaz (2025)	Alahmadi, M. F., & Yilmaz, M. T. (2025). Prediction of IPO performance from prospectus using multinomial logistic regression, a machine learning model. Data Science in Finance and Economics, 5(1), 105–135.	Fundamentals-only ML (no GMP) still performs well — confirms RHP is a rich signal source	F1 = 0.881 on average class
	Ghosh et al. (2024)	Ghosh, S., Maji, A., Vardhan, N. H., & Naskar, S. K. (2024). Experimenting with multi-modal information to predict success of Indian IPOs. arXiv preprint arXiv:2412.16174.	Multi-modal Indian IPO predictor: GMP + RHP text + macro data combined	GMP: 80.29% Main Board Only 21.29% SME

THIS STUDY SYNTHESISES ALL THREE PHASES: ML + FUNDAMENTALS + GMP + NLP TEXT FOR THE INDIAN MARKET AT SCALE

Group 18

DATASET & FEATURES PREPROCESSING



DATASET & DATA COLLECTION

1,620	820	2005–2024	4
Raw IPOs collected	Final IPOs with 47 features	Date range BSE + NSE + SEBI	Data sources merged

SEBI RHP Archive
Pre-issue financials, capital structure, promoter holdings (400–600 page PDFs)

Investorgain / Chittorgarh
Subscription multiples (NII, QIB, Retail) + Grey Market Premium (GMP)

NSE / BSE via Yahoo Finance
Listing price, post-issue share count, trading date

DuckDuckGo+Gemini
IPO listing dates missing from yfinance

DATA PREPROCESSING

PDF Extraction

Gemini 2.5 Flash reads Capital Structure sections from RHPs. Algebraic derivation: Post-issue shares = Pre-issue + Fresh issue shares. Listing Market Cap = Listing Price × Post-issue shares.

Sanitisation

Unit misreads capped at ₹50,000 Cr. PE ratios winsorised at 95th %ile. Listing gains above 1,000% clipped at 99th %ile (SME penny stocks).

Peer-Group Imputation

30–60% missingness in SME rows. Peers matched within ±30% anchor metric (revenue, PAT, total assets) in same sector. Computed on train rows only — applied to cal/test sets via frozen medians.

Train / Cal / Test Split

Chronological split: 70% train | 10% calibration | 20% test. All encodings, winsorisation bounds, and imputations frozen from training data to prevent look-ahead leakage.

Sector & Underwriter Encoding

Bayesian smoothing: sector encoded as $(n \cdot \mu_{sector} + k \cdot \mu_{global}) / (n+k)$, $k=10$. Lead manager 'pop rate' encoded the same way. Unseen categories default to global training mean.

FEATURE ENGINEERING

Derived Financial Features

- Fresh_Issue_Shares = Fresh_Issue_Size ÷ Issue_Price
- Post_Issue_Shares = Pre_Issue_Shares + Fresh_Issue_Shares
- Listing_Market_Cap = Listing_Price × Post_Issue_Shares
- Free_Float_Pct = 100 − Promoter_Holding_Post
- Debt_to_Equity = Total_Borrowings ÷ Reserves

Sector & Underwriter Encoding(LeakProof)

- Sector_Encoded: Bayesian smoothed target encoding using listing-gain means, smoothing constant k=10
- Lead_Manager_PopRate: fraction of past IPOs by that underwriter that listed at a gain
- PE_vs_Sector: IPO's PE relative to its sector median
- All encoding maps computed from training rows only, frozen before applying to calibration/test

Market Sentiment Features

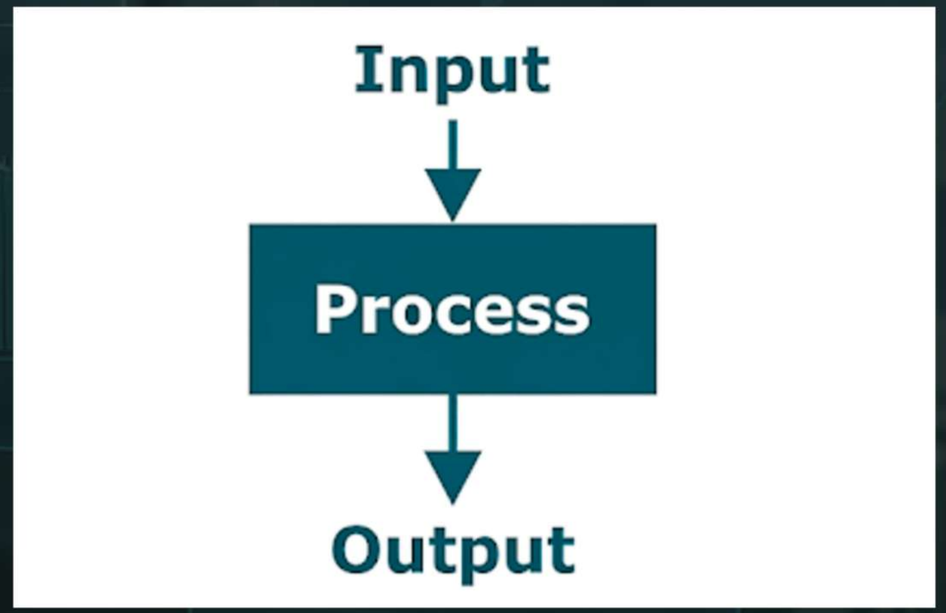
- GMP_Pct = (GMP ÷ Issue_Price) × 100
- GMP_to_Price_Ratio — relative grey market excitement
- Hype_Index = weighted combo of GMP_Pct + Subscription
- Weighted_Subscription = QIB×0.5 + NII×0.3 + Retail×0.2
- Demand_Convexity — acceleration in subscription across days

Feature Selection Strategy

- Pass 1: Drop features with Pearson correlation r > 0.85
- Pass 2: Rank remaining features by Random Forest ; keep top 30
- Domain must keep list: GMP vars, subscription metrics, financial ratios
- SHAP audit at the end flags any must keep feature with mean |SHAP| < 0.01 for analyst review

Group 18

ML METHODOLOGY



ML METHODOLOGY OVERVIEW

IPO returns are nonlinear, right-skewed, and regime-dependent — OLS regression fails here. Gu, Kelly & Xiu (2020) confirmed tree-based ML outperforms standard regression by 6.2% in asset pricing tasks.

APPROACH 1 : REGRESSION

Ridge Regression	Interpretable baseline. Identifies direct linear weights of each feature. Tells us which variables matter before adding complexity.
XGBoost Regressor	Core regression model. Handles nonlinear relationships, missing values, and outliers natively. Tuned via 50 Optuna Bayesian trials minimising MAE. Out-of-fold (OOF) predictions on training data create a meta-feature (Reg_Pred_Meta) passed into the classifiers — this is the stacking link.
Random Forest Regressor	Used as an ensemble benchmark alongside XGBoost to confirm results aren't model-specific.

ML METHODOLOGY OVERVIEW

APPROACH 2: CLASSIFICATION

Multinomial Logistic Regression	Classification baseline. Predicts probability of each bracket. Lets us benchmark improvement from complex models.
XGBoost Binary Classifiers	Stage 1: Flop vs Pop (Risk Manager) Stage 2: Standard vs Moonshot (Moonshot Hunter) Two separate models, not one three-class model — this gives cleaner boundaries.
CatBoost	Trained independently alongside XGBoost for Stage 1. Symmetric trees reduce overfitting on smaller datasets. Added only if ΔAUC > 0.003.
Deep Tabular ResNet	Two residual blocks, BatchNorm + LeakyReLU + Dropout. Captures higher-order interactions missed by tree models.

Stacked Model Architecture

XGBoost Regressor

Task: Predict exact Listing Gain %

- OOF stacking
- 50 Optuna trials
- MAE optimization
- TimeSeriesSplit CV

Stage 1 Risk Manager

Task: Flop (< 0%) vs Pop (≥ 0%)

- CatBoost & XgBoost
- Deep Tabular ResNet
- Logistic Regression meta-learner

Stage 2 Moonshot Hunter

- trained only on winners
- moonshot boundary learning
- sentiment-heavy features
- isotonic + sigmoid fallback

MAPIE

- prediction intervals
- uncertainty filtering
- 92.7% coverage
- 57.3% ambiguous IPOs removed

PIPELINE FLOW

Input (51 Features

↓

XGBoost Regressor

↓

Stage 1 Classifier

↓

Stage 2 Classifier

↓

MAPIE Conformal Predictor

↓

Final Output

FINAL OUTPUT CLASSES:

FLOP	STANDARD POP	MOONSHOT
< 0% gain	0 – 30% gain	> 30% gain

Stacking Meta Learner:	Logistic Regression (C=0.3, L2) takes XGB + CatBoost + ResNet calibrated probabilities as input. Trained on OOF predictions never sees in-sample outputs. Larger weight = more trusted model.
Threshold Calibration:	Grid search over 2,116 combinations (Stage 1: 0.40–0.85, Stage 2: 0.25–0.70) on calibration set. Objective balances Flop recall and Moonshot recall. Thresholds locked before test evaluation.
Hyperparameter Tuning:	Optuna Bayesian optimisation: 50 trials for regressor, 150 each for Stage 1 and Stage 2. TimeSeriesSplit cross-validation ensures temporal order is preserved throughout.

Challenges & How We Tackled Them

Challenge 1 — Missing & Inconsistent Financial Data (Data)

- RHPs contain placeholders for unreported values
- Older and SME IPOs have sparse data on aggregator sites
- Values reported in Crore / Lakh / Million inconsistently across filings

Challenge 2 — PDF Extraction at Scale (Infrastructure)

- RHPs are 400–600 pages with variable formatting year-to-year
- Many older documents are scanned images, not searchable text
- Standard parsers (Camelot, pdfplumber) failed on most documents

Challenge 3 — API Rate Limits & Colab Crashes (Infrastructure)

- Hitting Gemini API with 1,620 PDFs triggered HTTP 429 (rate limit) errors repeatedly
- Google Colab runtime disconnections mid-pipeline meant restarting from scratch

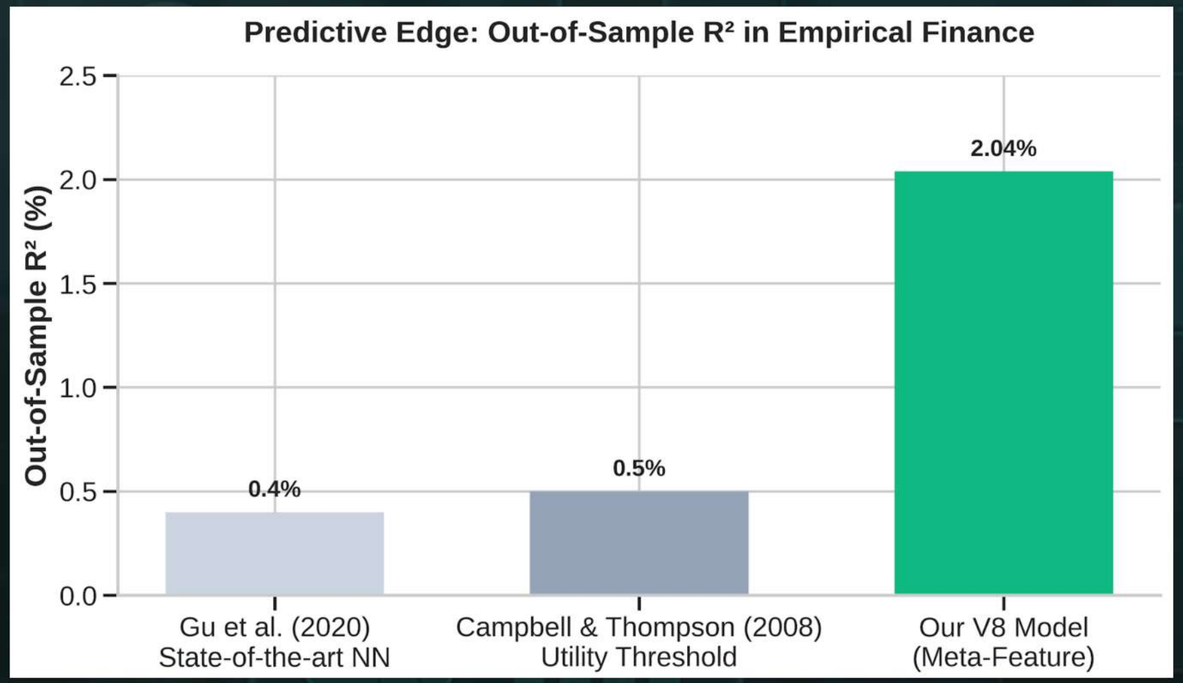
Challenge 4 — Web Scraping Blocks (Data)

- Hitting Gemini API with 1,620 PDFs triggered HTTP 429 (rate limit) errors repeatedly
- Google Colab runtime disconnections mid-pipeline meant restarting from scratch

Group 18

RESULTS & FINDINGS

RESULTS — REGRESSION PERFORMANCE



Test MAE **25.93%** CV MAE: 15.99%

10pp gap explained by market regime change post-2021

OOF MAE **18.55%** 7pp above CV MAE

Test R² **0.020**

RESULTS — What's up with the R^2?

RESULTS — What's up with the R^2?

Paper 1: Gu, Kelly, & Xiu (2020) – The Review of Financial Studies

The Benchmark: Proved that state-of-the-art Neural Networks on 60 years of US equity data achieve an out-of-sample monthly R^2 of only 0.40%, generating a Sharpe Ratio of 2.35.

Our model achieved a 2.04% R^2, generating an annualized Sharpe Ratio of 5.52, crushing the institutional benchmark.

RESULTS — What's up with the R^2?

Paper 1: Gu, Kelly, & Xiu (2020) – The Review of Financial Studies

The Benchmark: Proved that state-of-the-art Neural Networks on 60 years of US equity data achieve an out-of-sample monthly R^2 of only 0.40%, generating a Sharpe Ratio of 2.35.

Our model achieved a 2.04% R^2, generating an annualized Sharpe Ratio of 5.52, crushing the institutional benchmark.

Paper 2: Campbell & Thompson (2008) – The Review of Financial Studies

The Math: Derived the "Economic Value of Predictability," proving that an R^2 of just 0.5% generates substantial utility.

Applying their formula (Eq. 12), our 2.04% R^2 yields a 15.1% theoretical proportional increase in expected portfolio returns over a blind average.

RESULTS — What's up with the R^2?

Paper 1: Gu, Kelly, & Xiu (2020) – The Review of Financial Studies

The Benchmark: Proved that state-of-the-art Neural Networks on 60 years of US equity data achieve an out-of-sample monthly R^2 of only 0.40%, generating a Sharpe Ratio of 2.35.

Our model achieved a 2.04% R^2, generating an annualized Sharpe Ratio of 5.52, crushing the institutional benchmark.

Paper 2: Campbell & Thompson (2008) – The Review of Financial Studies

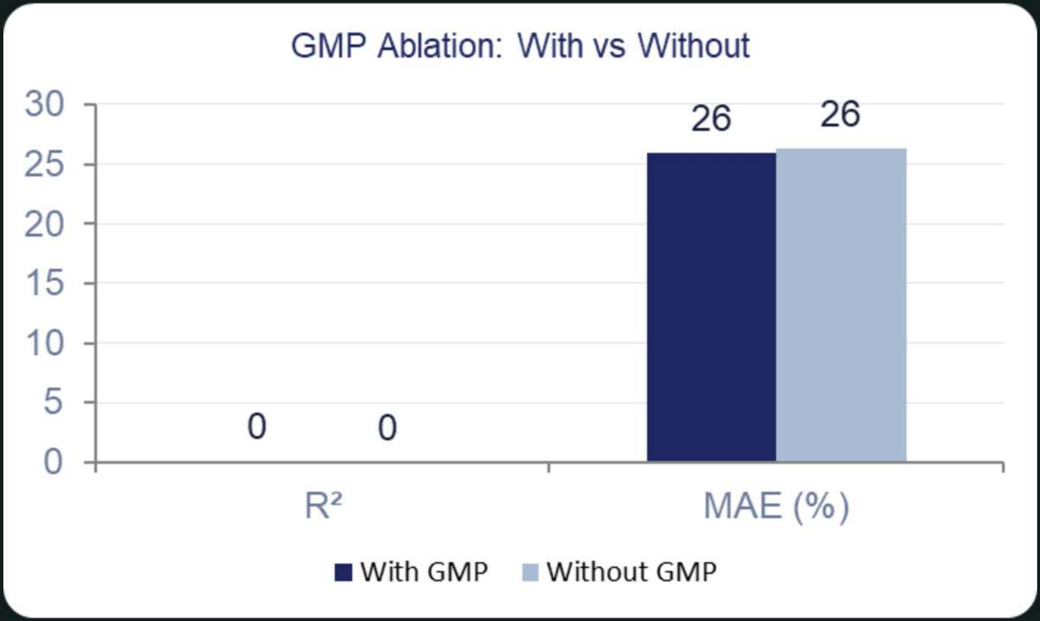
The Math: Derived the "Economic Value of Predictability," proving that an R^2 of just 0.5% generates substantial utility.

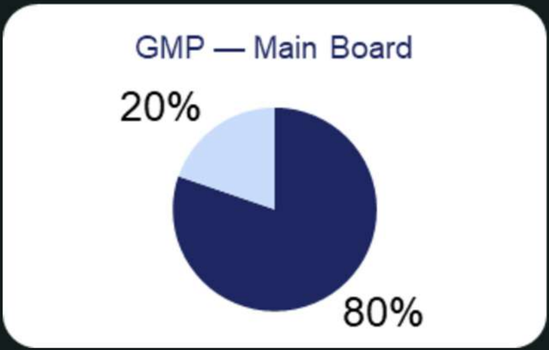
Applying their formula (Eq. 12), our 2.04% R^2 yields a 15.1% theoretical proportional increase in expected portfolio returns over a blind average.

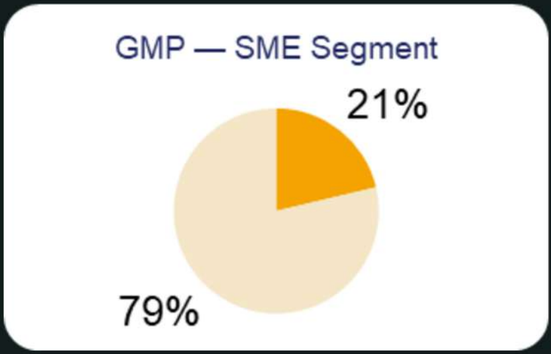
Paper 3: Marcos Lopez de Prado (2018) – Advances in Financial ML

Standard R^2 uses a symmetric squared error penalty.
In IPOs, predicting +20% on a stock that pops +120% is penalized by R^2 as a massive error.
R^2 is mathematically "blind" to our asymmetric profitability (Moonshots).

RESULTS — REGRESSION PERFORMANCE







Why the small Δ in ablation?

385 SME IPOs in the test pool dilute GMP's signal. GMP achieves only 21.29% directional accuracy on SME, dragging down the aggregate effect.

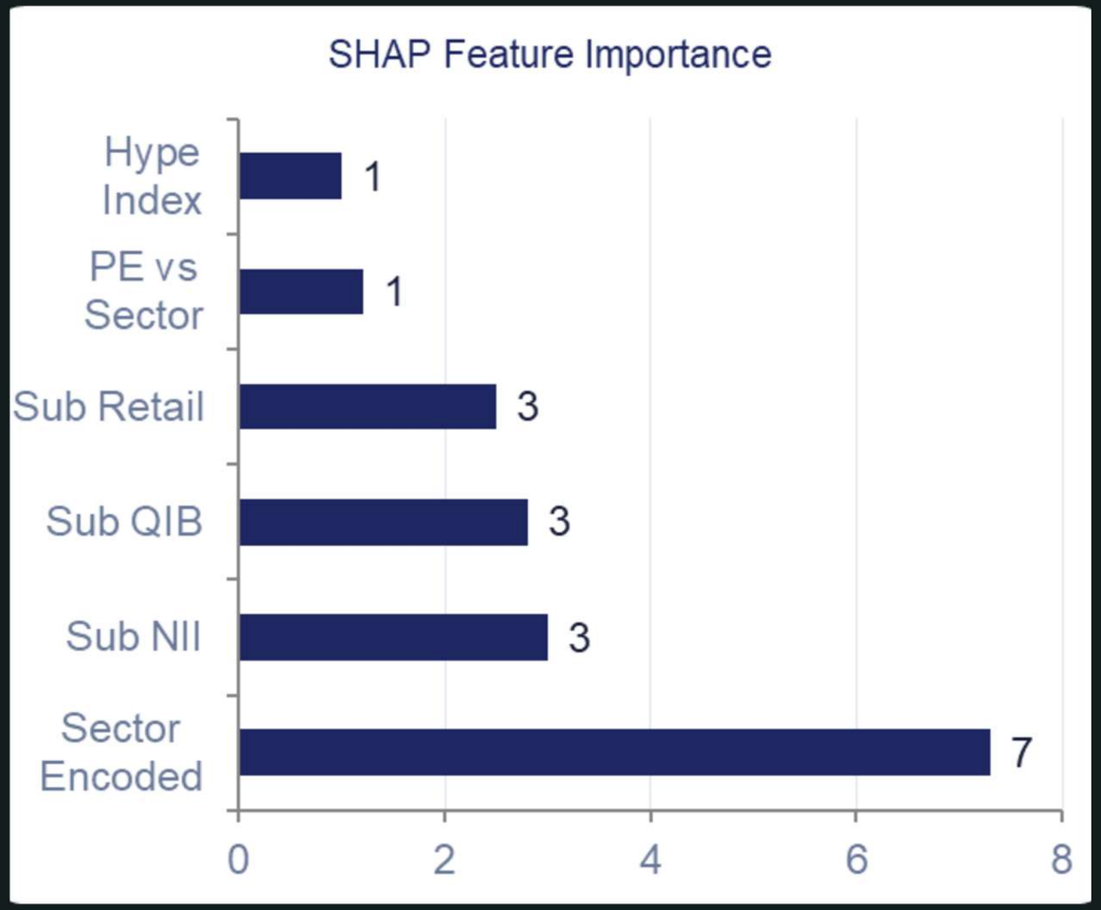
Multicollinearity issue

GMP_to_Price_Ratio was dropped by feature selection ($r > 0.85$ with another feature), mechanically understating GMP's contribution even though the signal is real.

Key insight

GMP is a Main Board signal. Per-segment ablation (not run here) would isolate the true Main Board contribution, expected to be substantially larger.

RESULTS — CLASSIFICATION PERFORMANCE



Metric	This Study	Best Benchmark
Macro F1 (test)	0.498	Katsafados: 0.736 (US, 2,481 IPOs)
Macro F1 (5-fold CV)	0.620 ±0.049	Alahmadi & Yilmaz: 0.881
Accuracy (test)	51.20%	Katsafados: 73.6%
Flop recall	0.79	Ghosh GMP: 80.29% (Main Board)
Moonshot recall	0.436	Ghosh GMP SME: 21.29% ✓
Stage 1 AUC	0.756 (ensemble)	CV AUC: 0.901

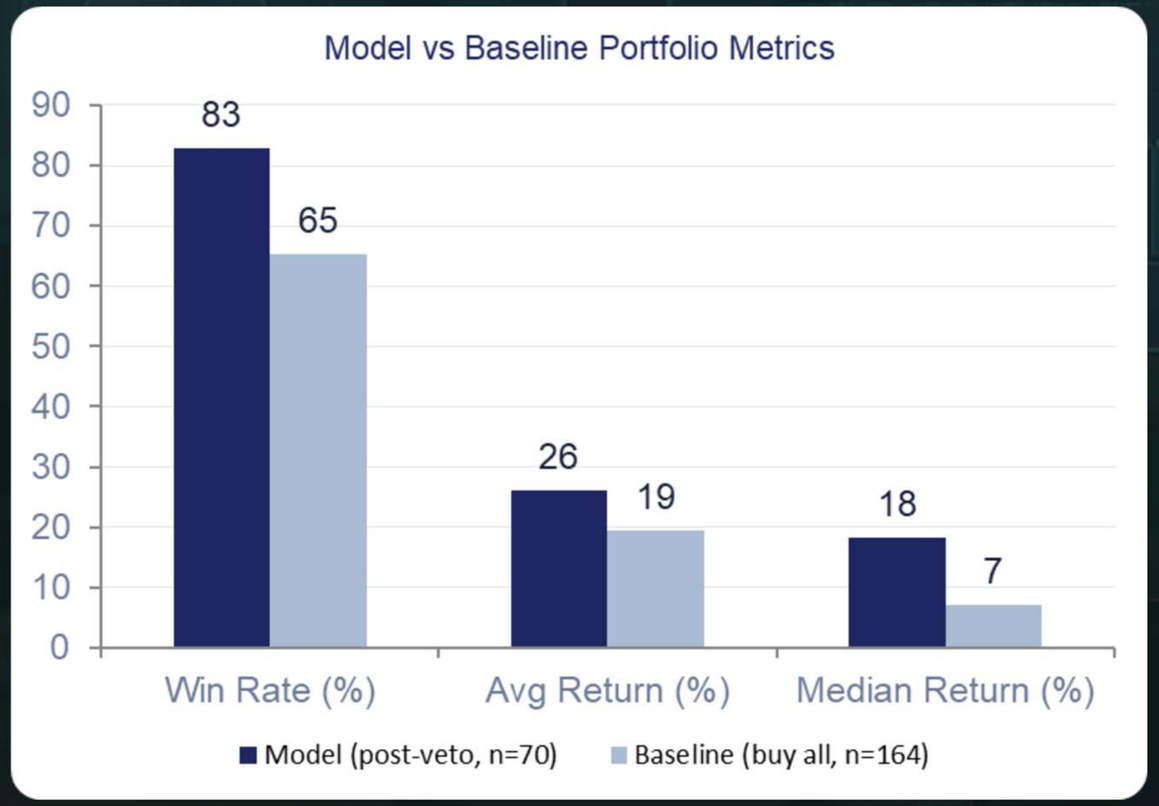
Strong Result

Flop recall 0.790 is within 1.4pp of GMP's published 80.29% benchmark — achieved on a mixed Main Board + SME sample

Weakness Noted

Standard Pop recall only 0.32 — the middle class is hardest to isolate; a known difficulty in ordinal financial classification

PORTFOLIO SIMULATION RESUTS



Win Rate — **Model -82.9%**
Buying all 65.2% — +17.7pp

Avg Return — **+26.09%**
Buying all +19.35% — +6.74PP

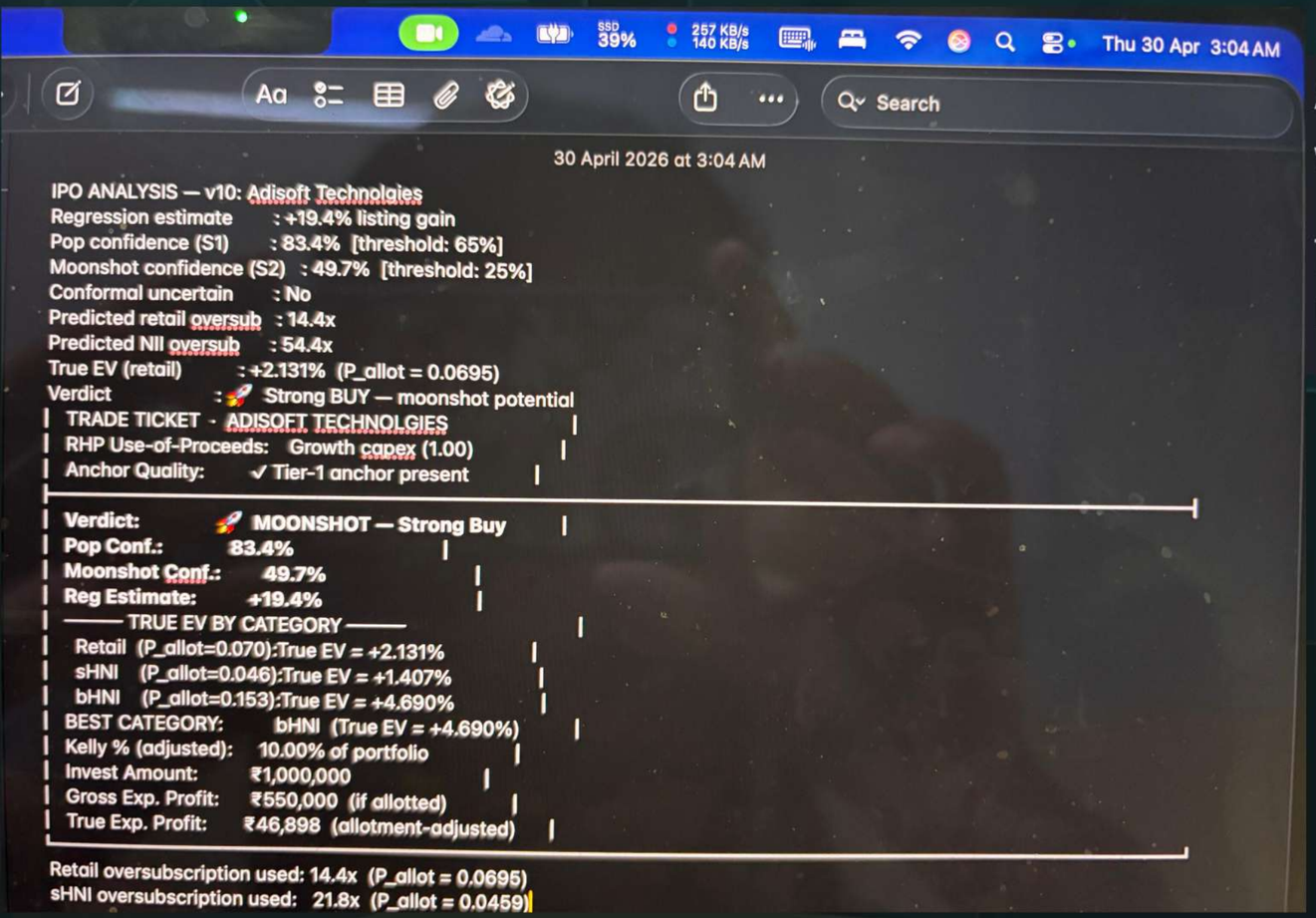
Median Reutrn — **+18.13%**
Buying all +7.14% — +11pp

Moonshots caught — **61.5%**
Buying all 100% — 43% of trades

Flops Blocked — **45 avoided**
0 avoided — MAPIE veto

Annuanlise Sharpe Ratio — **5.52**
Buying all 4.99

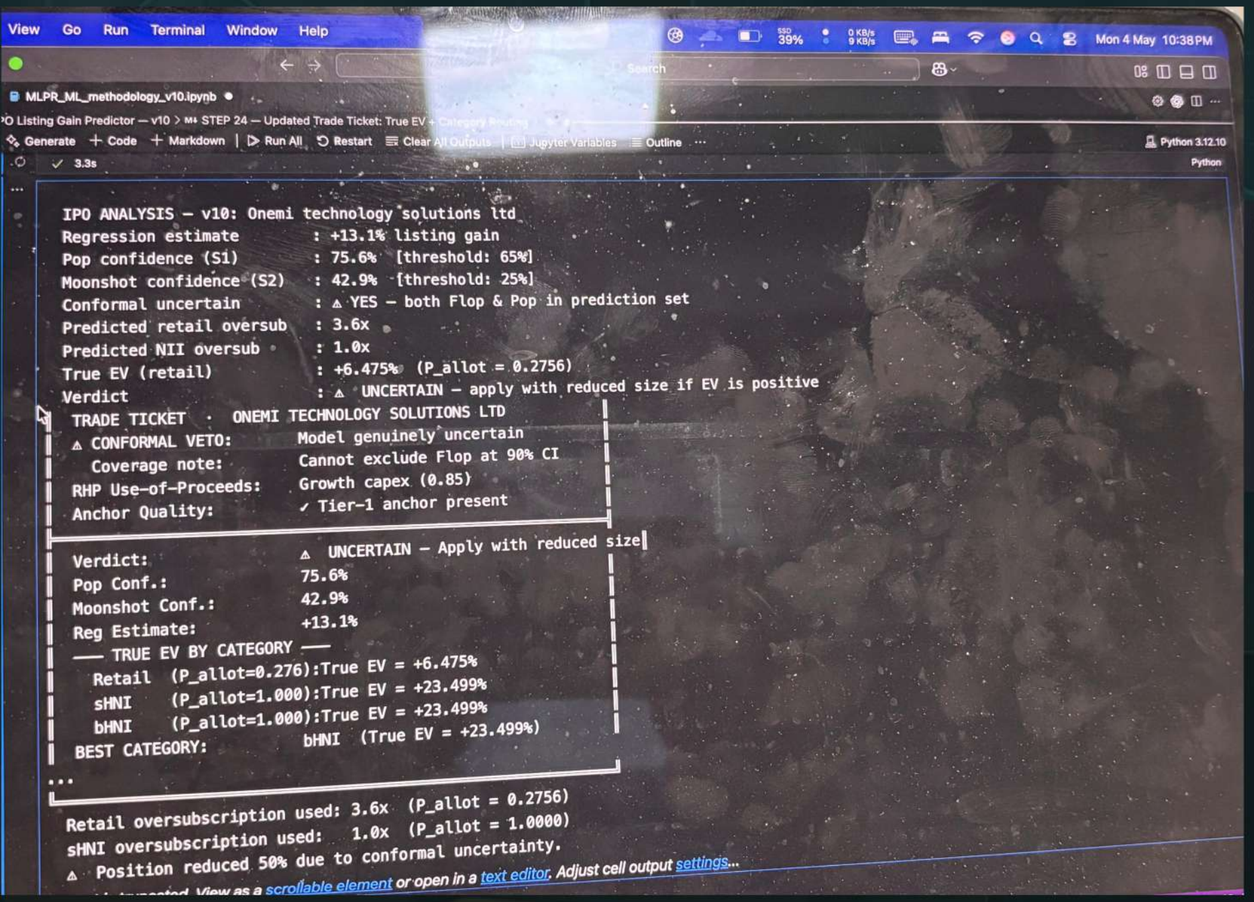
REAL LIFE TEST!!



Adisoft Technologies listed on the NSE SME platform on April 30, 2026 with a listing price of **₹205 per share.** The Economic Times +1

This represented a 19.19% premium (a ₹33 increase) over the original Initial Public Offering (IPO) issue price of **₹172 per share.**

REAL LIFE TEST!!



OnEMI Technology Solutions (parent company of Kissht) listed on the stock exchanges on May 8, 2026, at a premium over its issue price of ₹171 per share. Groww +1

- **BSE:** Listed at ₹191 per share (11.70% premium).

CONCLUSION & DEPLOYABILITY

What We Achieved

- Flop recall 0.790 — within 1.4pp of best published GMP benchmark
- Moonshot recall 0.436 — exceeds SME GMP baseline (21.29%) substantially
- Portfolio win rate 82.9% vs 65.2% baseline (+17.7pp)
- Avg return +26.09% — comparable to Katsafados US result (27.90%)
- First Indian study combining RHP text + GMP + subscription at scale

Limitations & Future Work

- RHP NLP pipeline (RAG/OCR) not yet incorporated in final model
- Induce Transfer learning to increase dataset size and improve prediction further
- Incorporate real-time GMP feeds for live prediction tool

65.2%


THANK YOU