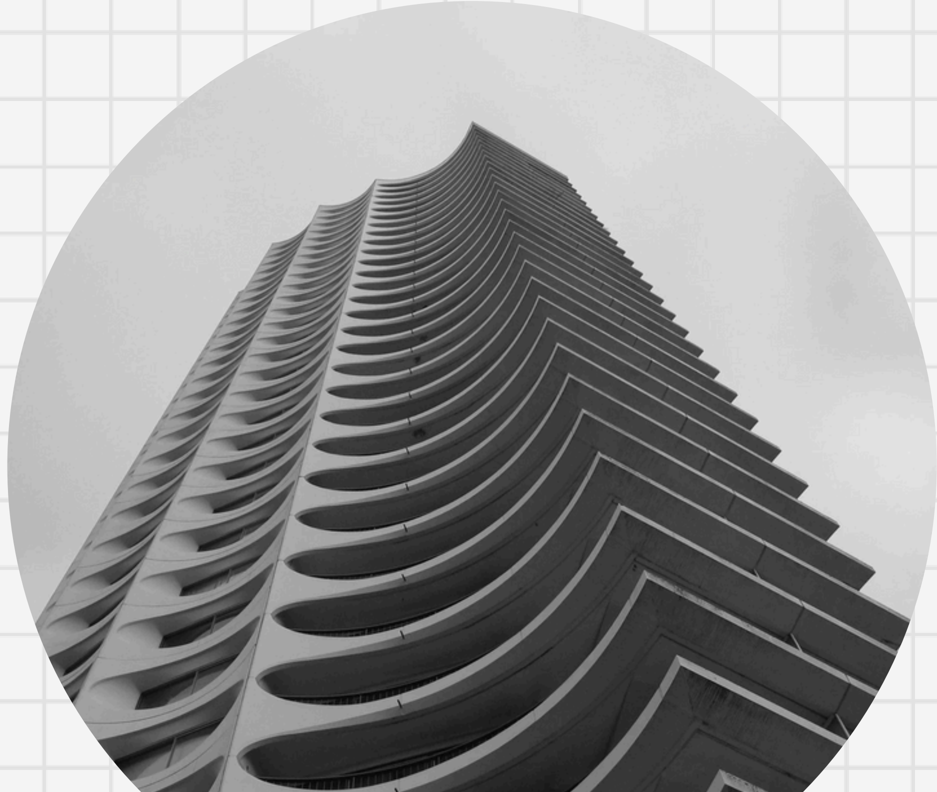
MLPR FINAL PRESENTATION

IPO BINARY CLASSIFICATION

MICHELE DAROOWALA, ESHANI PARULEKAR AND VARUN ARORA



Index

*"In 2024, India experienced a record-breaking year in its IPO market, with a total of **317** initial public offerings (IPOs) — which amounts to **more than one IPO being launched every day** on average."*

-Deva, P. (2024, December 23). 2024 Review: Indian IPO market shatters records as 317 issues raise ₹1.8 trillion. Mint.

Problem Statement:

Manually analyzing **lengthy IPO prospectuses** to predict listing gains is **time-consuming** and **overlooks market sentiment signals** critical to IPO performance.

What is the problem we are solving?

The Indian IPO market has become **increasingly active**, but predicting the success or failure (i.e., listing gain) of an IPO **remains complex.**

Traditionally, this is done by **manual analysis** of the Red Herring Prospectus(RHP).

What is a Red Herring Prospectus? (RHP)

RHPs are **comprehensive documents outlining key information** regarding companies risk factors, financials, management, etc.

RHPs are **long and difficult to read (400-600+ pages)**, making parsing and extracting relevant information from each extremely difficult.

Why is this important?

- Helps retail and institutional investors make **informed decisions**.
- **Aids financial advisors, analysts, and brokerage firms** to make their job more efficient and less time consuming.

Potential Applications:

1. **Automated** IPO recommendation engines.
2. **Risk assessment tools** for brokers and advisors.
3. **Data-driven dashboards** for fintech startups.

Impact:

Replacing manual, slow and inconsistent methods with **scalable, AI-powered analysis tools** tailored to Indian capital markets.

Literature Review

Traditional Methods:

Before the rise of AI/ML, IPO performance prediction was handled using econometric models with structured financial data. The following methods dominated:

Logistic Regression & Linear Regression

Commonly used to predict IPO underpricing and long-term performance.
Relied on variables like **ROE, EPS, issue size, and promoter holding.**

Ritter, J. R. (1991). The long-run performance of initial public offerings. The Journal of Finance, 46(1), 3–27. Ritter (1991) analyzed long-run IPO underperformance using traditional metrics.

Manual analysis of RHPs

Financial analysts **manually reviewed** Red Herring Prospectuses to assess company risks, competition, and purpose of funds.

Ding, Y. (2015). Investor attention and IPO performance: Evidence from prospectus textual analysis. Journal of Behavioral Finance, 16(3), 279–289. https://doi.org/10.1080/15427560.2015.1064933

Machine Learning Shift:

ML models emerged post-2010 as a response to the limitations of linear models in capturing nonlinearity and combining textual + numerical data.

SVM, Random Forest, XGBoost

These models handle high-dimensional data, nonlinear relationships, and class imbalance.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Textual + Sentiment Analysis using NLP

Used on RHPs and online forums (Reddit, Twitter) to model investor sentiment.

Singh, A., & Kalra, S. (2024). IPO forecasting using machine learning methodologies: A systematic review apropos financial markets in the digital era. Arthshastra Indian Journal of Economics & Research, 13(2), 23. https://doi.org/10.17010/aijer/2024/v13i2/173502

Aggarwal, S., Singh, M., & Gupta, A. (2021). A machine learning approach for IPO success prediction using investor sentiment and financial ratios. International Journal of Financial Studies, 9(4), 52. https://doi.org/10.3390/ijfs9040052

Investor Sentiment and Social Media Data

Liew & Wang (2015) analyzed IPO performance using Twitter sentiment and identified significant correlation with oversubscription and pricing.

Liew, J. K.-S., & Wang, T. (2015). Twitter-based investor sentiment and IPO performance. Journal of Behavioral Finance, 16(3), 232–245.

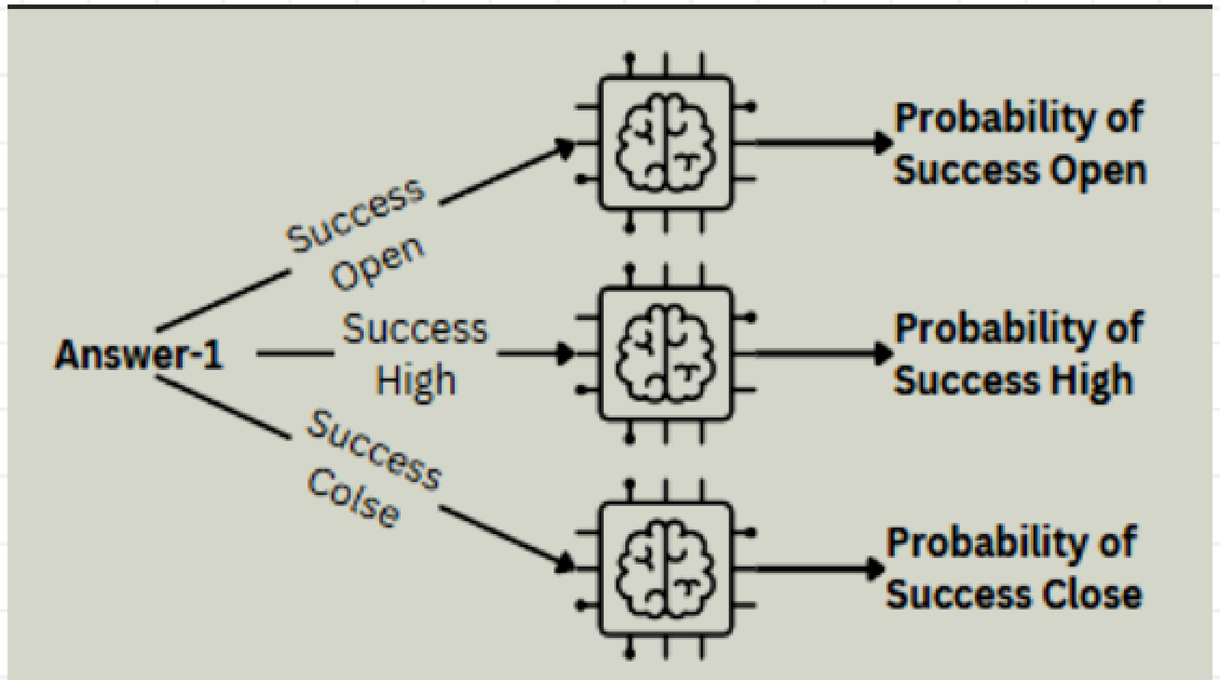
Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFinText system. ACM Transactions on Information Systems, 27(2), 1–19. https://doi.org/10.1145/1462198.1462204

Pre-existing Approaches:

			Main Board			SME		
Model	**Type**	**Input**	**AUC**	**F1 (0)**	**F1 (1)**	**AUC**	**F1 (0)**	**F1 (1)**
GLM	O	N-C	0.824	0.517	0.915	**0.698**	0.076	0.887
DRF	O	N-C	0.597	0.592	0.893	0.669	0.688	0.890
DL	O	N-C	**0.903**	**0.781**	**0.947**	0.679	0.003	0.887
XGB	O	N-C	0.773	0.233	0.893	0.670	0.077	0.887
GBM	O	N-C	0.712	0.597	0.893	0.606	**0.736**	**0.893**
Ens	O	N-C	0.824	0.278	0.902	**0.698**	0.027	0.887

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.

1. Each question and page of the prospectus was embedded using Nomic (Nussbaum et al., 2024).
2. Using only numeric and categorical (N-C) features (listed in Table LABEL:tab:ipo-variables), we trained five models using H2O AutoML.
3. Fine-tuned 26 DeBERTa-base models.



Dataset and Features Preprocessing

Considerations

1. Qualitative info from RHP
2. Financial market sentiment
3. Sentiment around IPO
4. Prospectus sentiment

Method of Collection

1. HDFC Securities
2. SEBI Website
3. Chittorgarh
4. Yahoo Finance API

Features, Dataset Details

1. Started with a list of ~400 IPO's, finally using ~150
2. Nifty Performance: 1 month, 3 month, 6 month before IPO listing data.
3. TFIDF & Sentiment weighted embeddings of risks, revenue growth, competitive landscape of IPO
4. Chittorgarh IPO recommendation
5. Issue price

Preprocessing Methodology

Qualitative Data

Custom RAG pipeline, TF-IDF, GLOVE embeddings of top 10 words, sentiment using NLTK SIA.

Quantitative Data

Standardization of Nifty Performance & Issue price.

Sentiment Data

One hot encoding of IPO recommendation.

ML Methodology

Base Model:

	precision	recall	f1-score	support
No	0.00	0.00	0.00	3
Yes	0.77	1.00	0.87	10
accuracy			0.77	13
macro avg	0.38	0.50	0.43	13
weighted avg	0.59	0.77	0.67	13

Qualitative information

- Unable to use LLM API’s.
- Failed to retrain Prospectus Roberta model.

Data

- Direct sentence embeddings of question answers of Prospectus Roberta dataset.

Attempt 2:

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.20	0.33	5
1	0.79	1.00	0.88	15
accuracy			0.80	20
macro avg	0.89	0.60	0.61	20
weighted avg	0.84	0.80	0.75	20

XG Boost

Classification Report:

	precision	recall	f1-score	support
0	0.50	0.33	0.40	3
1	0.89	0.94	0.91	17
accuracy			0.85	20
macro avg	0.69	0.64	0.66	20
weighted avg	0.83	0.85	0.84	20

Shallow Neural Net

- TF-IDF of answers in the Prospectus Roberta dataset.
- GLOVE embeddings (50 D) for top 10 words.

RAG Pipeline

- Attempted to use large context window models (Yi-34B-200K)
- Attempted Langchain with TinyLLama
- Finally, MiniLM-L6 with Mistral 7B
- Further steps - using model's with larger context windows and BERT embeddings

```
model_path = hf_hub_download(
    repo_id="TheBloke/Mistral-7B-Instruct-v0.1-GGUF",
    filename="mistral-7b-instruct-v0.1.Q4_K_M.gguf",
    local_dir=".",
    local_dir_use_symlinks=False
)
```

```
def create_vector_store(text):
    chunks = [text[i:i + CHUNK_SIZE] for i in range(0, len(text), CHUNK_SIZE)]
    embedder = SentenceTransformer("all-MiniLM-L6-v2")
    embeddings = embedder.encode(chunks, batch_size=32, show_progress_bar=False,
    index = faiss.IndexFlatL2(384)
    index.add(embeddings)
    return chunks, embedder, index
```

```
    prompt = f"""
You are an IPO expert evaluating investment opportunities. Use only the context below to ans

Context:
{context}

Question:
{query}

Answer (respond with a structured paragraph):
"""
    response = llm(prompt, max_tokens=768, temperature=0.3, top_p=0.9, repeat_penalty=1.1)
    return response["choices"][0]["text"].strip()
```

Performance Metrics and Deployability

Final Model

```python
model = Sequential([
    Dense(128, input_dim=X.shape[1], activation='relu'),
    BatchNormalization(),
    Dropout(0.3),

    Dense(64, activation='relu'),
    BatchNormalization(),
    Dropout(0.3),

    Dense(1, activation='sigmoid')  # Output layer for binary classifi
])
```

```
Most Common Lemmatized Words:

company: 937
issue: 395
share: 328
ipo: 308
price: 277
growth: 259
```

Custom corpus of prospectus stop words

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.67	0.73	6
1	0.92	0.96	0.94	24
accuracy			0.90	30

```python
def analyze_sentiment(text):
    if pd.isna(text):
        return 0

    sentiment_scores = sia.polarity_scores(text)
    return sentiment_scores['compound']
```

Sentiment & TF-IDF weighted GLOVE embeddings

Deployability & Future Scope

Not in place:
Information/Data Pipeline

Required:
GPU’s for larger context LLM’s

Thank You!