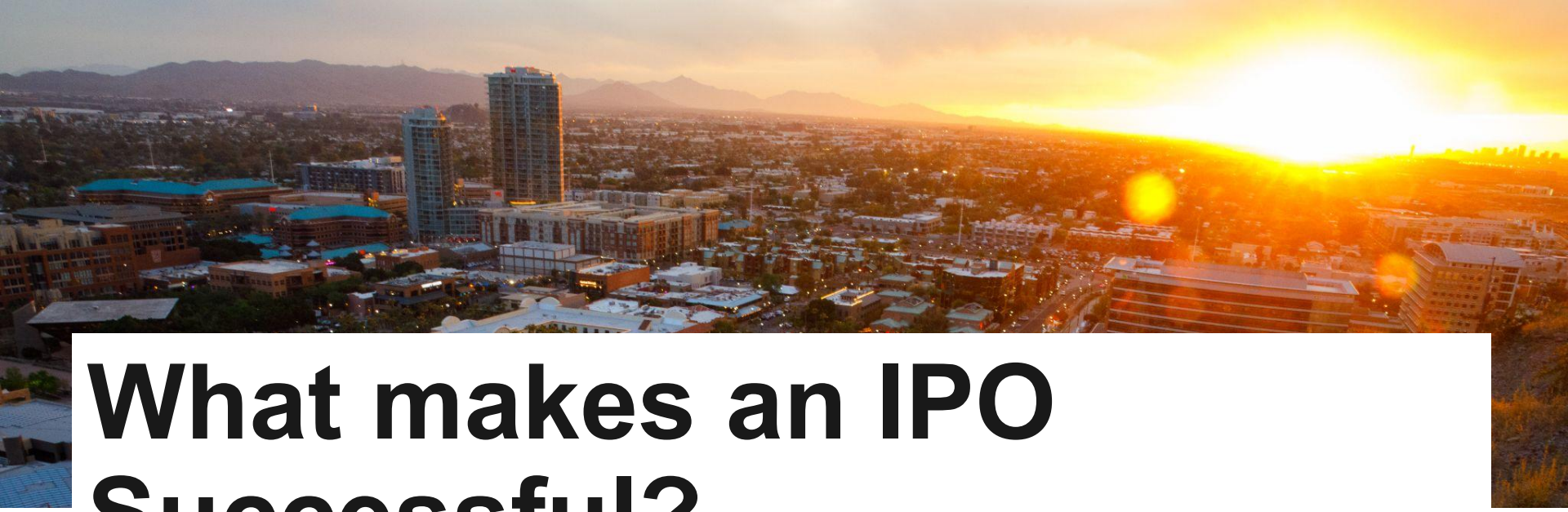


What makes an IPO Successful?

Data-driven analysis to identify key factors influencing IPO success, leveraging historical trends, financial metrics, and market conditions.

Presented By: Team Saturn

Arizona State University

Team Saturn

Presenting on Capstone Proposal

Julian Lin

Shraddha Pandey

Deepakshi Mathur

Gunner Tillemans

Ninad Shah

Business Problem

Identifying Key Factors and Predicting IPO Short and Long-Term Success

Analyzing IPO Success: A Multifaceted Approach to Post-IPO Performance

This study focuses on predicting post-IPO stock performance by leveraging historical SEC Filings data, market trends, and more. It incorporates analysis of past IPO trends, financial indicators, and timing strategies to forecast stock movement and optimize launch windows. We trained machine learning models to uncover patterns across financial and market variables, running simulations using real-world IPO data to evaluate predictive accuracy. And Simulated real world outcomes from our prediction models.

2. Data Sources and Preprocessing

Sources

- **SEC EDGAR Filings (2009–2024)**: Extracted pre-IPO financials from Forms S-1, 10-K, and 10-Q
- **SEC CIK to Ticker:** Two CIK crosswalks were found with current and historical tickers
- **EODHD Finance Data**: Pulled adjusted stock prices at IPO date, +6 months, +3 years.
- **FRED Macroeconomic Data**: Inflation rates and Federal Funds Rate at IPO time.
- **Yahoo Finance - S&P 500 Index**: Trend indicators using 30/50/200-day moving averages.
- **OpenCorporates:** We were granted access to extract companies incorporation date

Data Wrangling

CIK	ddate	Qtr	Tag	Value
20859	01/25/2022	1	Revenue	1,000,000
20859	04/25/2022	2	Assets	5,000,000
20859	08/25/2022	3	Expenses	750,000

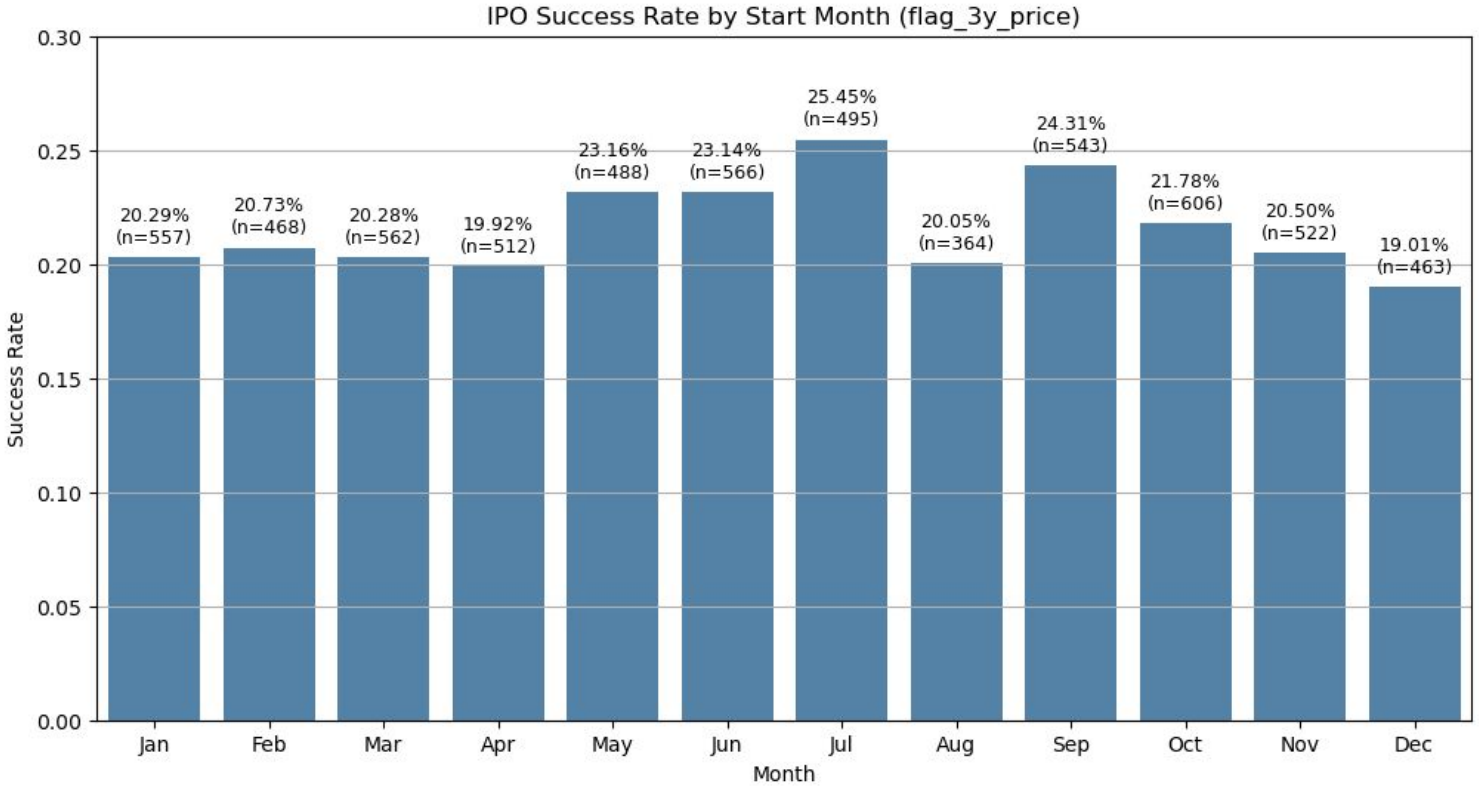
CIK	ddate	QRT	Revenue	Assets
20859	01/25/2022	1	1,000,000	5,000,000
20859	04/25/2022	2	800,000	5,000,000
20859	08/25/2022	3	600,000	5,000,000
20859	12/25/2022	4	1,200,000	4,000,000

Data Wrangling

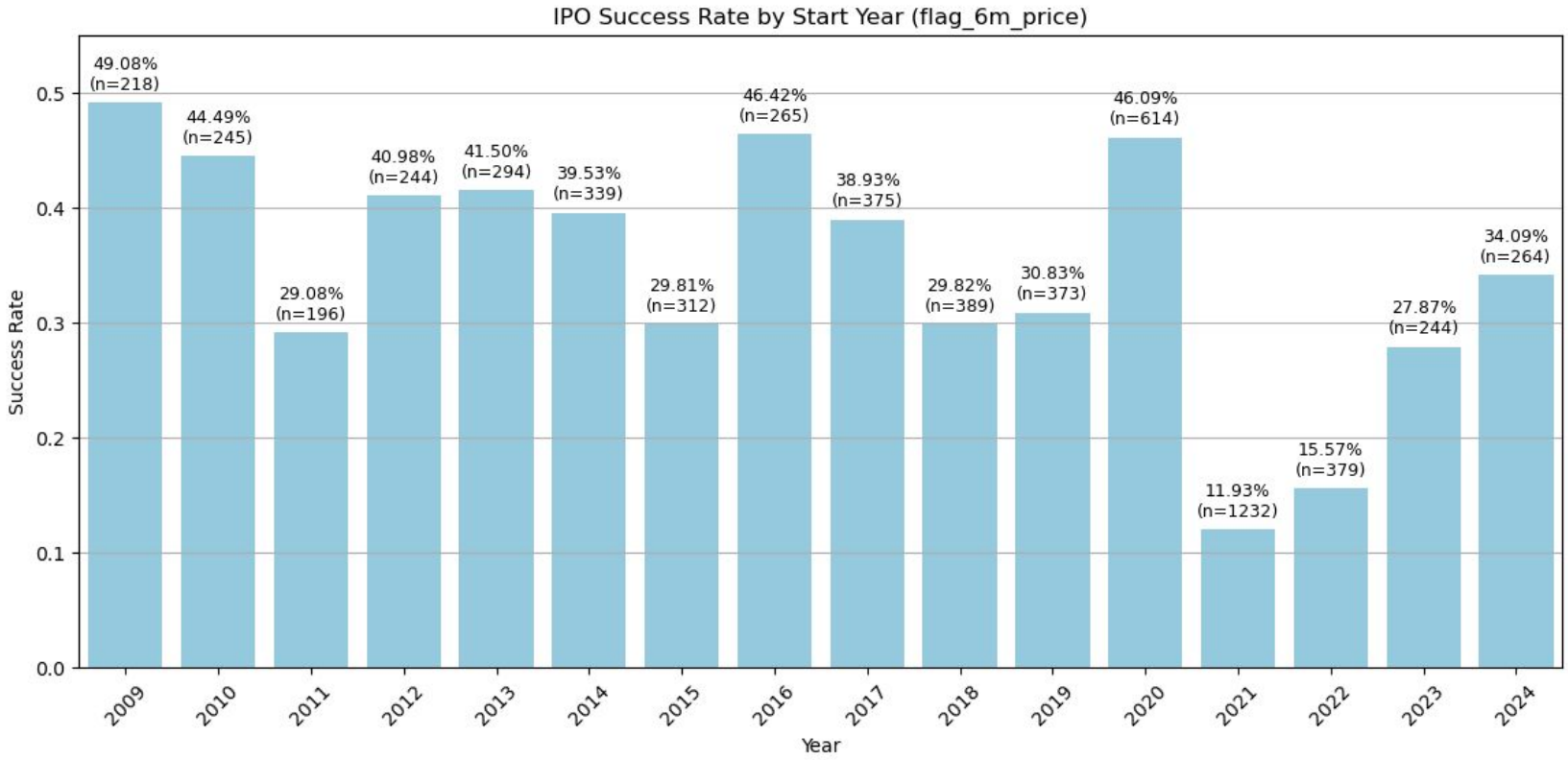
CIK	ddate	QRT	Revenue	Assets
20859	01/25/2022	1	1,000,000	5,000,000
20859	04/25/2022	2	800,000	5,000,000
20859	08/25/2022	3	600,000	5,000,000
20859	12/25/2022	4	1,200,000	4,000,000

CIK	ddate	Median Revenue Q1	Mean Revenue Q1	2Y ROC Asset Q1
20859	01/25/2022	500,000	1,000,000	2,247
58769	04/25/2022	57,000	500,000	5,000
20859	08/25/2022	60,000	57,000	7,000

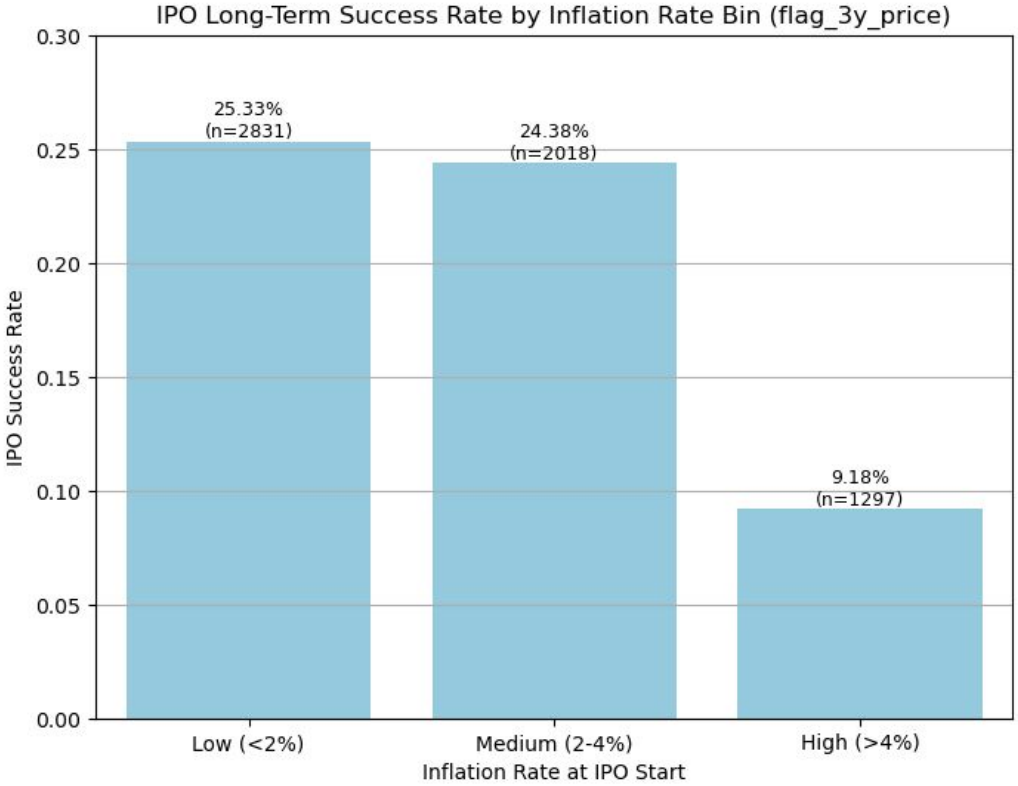
IPO Success Rate by Month

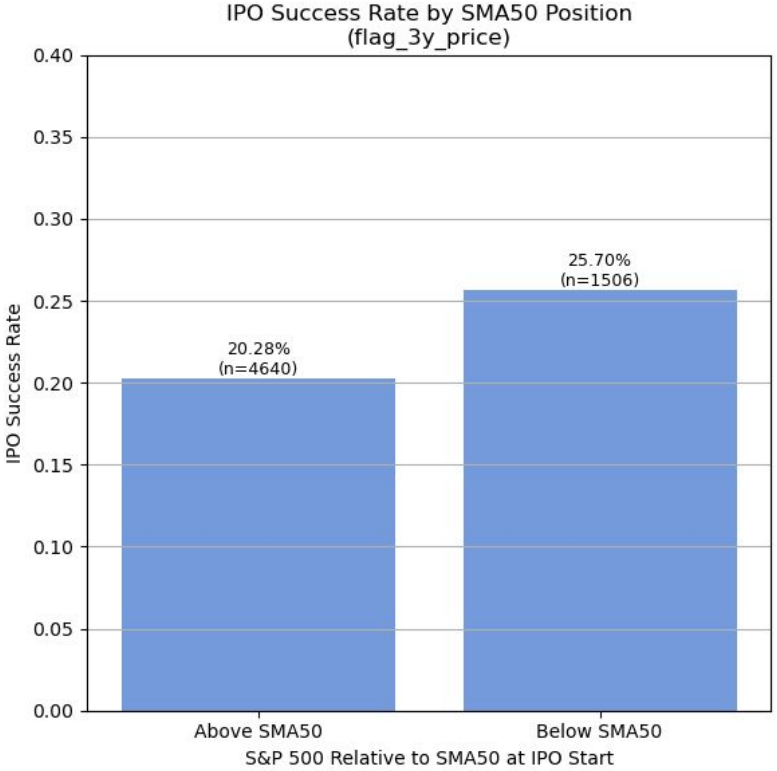


IPO Success Rate by Year

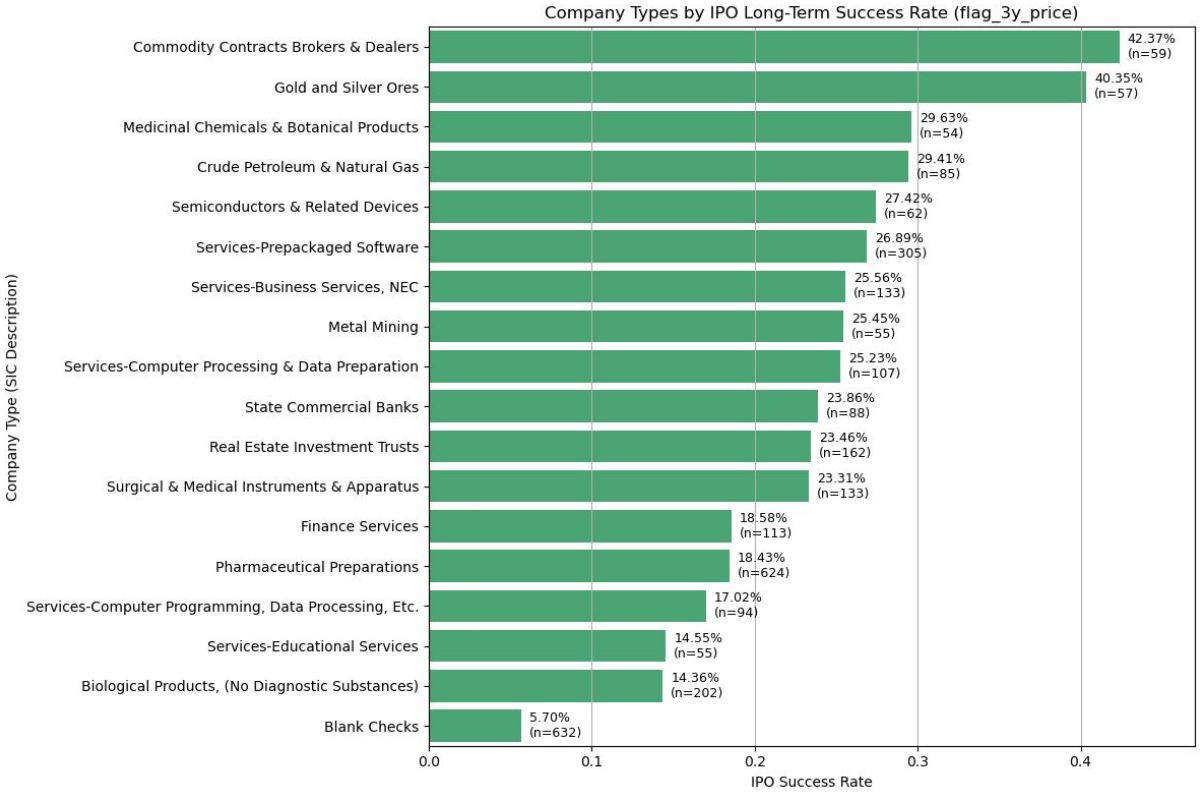


IPO Success Rate by Inflation Rate and S&P500

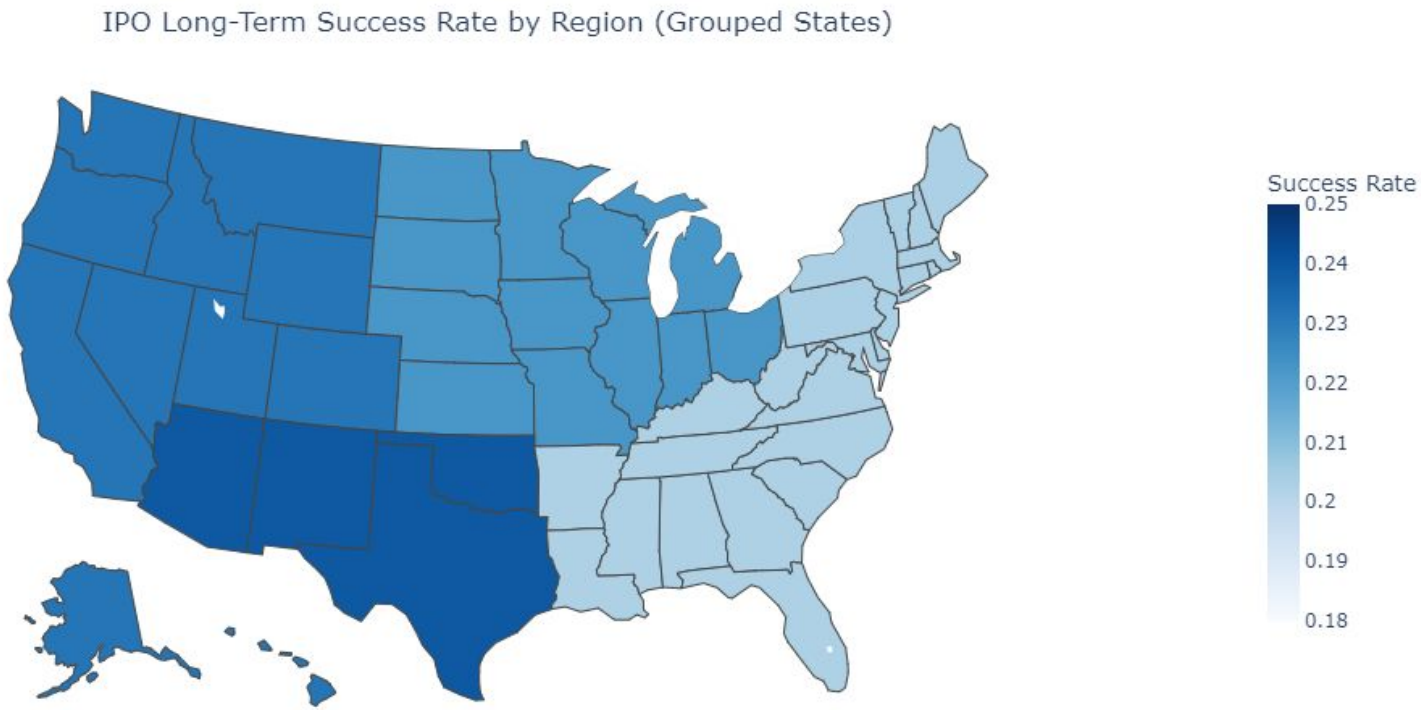




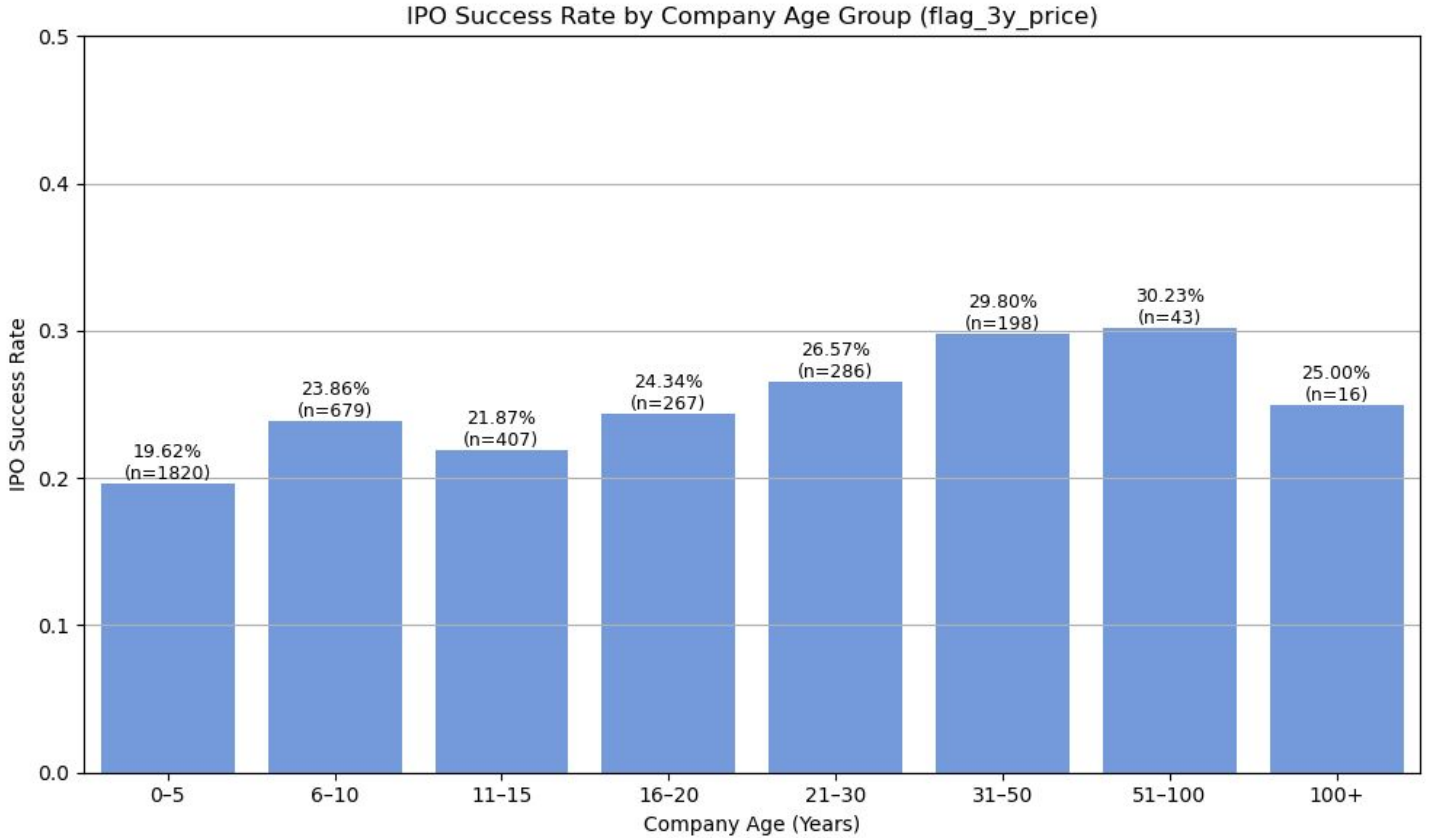
IPO Success Rate by Company Type



IPO Success Rate by US Region



IPO Success Rate by Company Age



Modeling and Evaluation

Model

- Deep Neural Network (PyTorch)
- 5-layer MLP, Dropout(0.3), ReLU activations
- Optimized using Adam, BCEWithLogitsLoss

Performance

- **Test Accuracy**: ~79%
- **Evaluation Metrics**: Precision, Recall, Confusion Matrix, AUC
- **Explainability**: SHAP summary & decision plots revealed key drivers

Tools and Libraries Used

`PyTorch`, `Pandas`, `yFinance`, `FRED`, `SHAP`, `Matplotlib`, `Optuna`

6. Modeling Approach

Model

Deep Neural Network (PyTorch)

Architecture

- 5 Fully Connected Layers
- Dropout Regularization (0.3)
- ReLU Activation
- Binary Cross Entropy Loss
- Adam Optimizer

Training

- 5,000 epochs
- Batch size: 256
- 80-20 train-test split
- Standardized input features

Results NN Model

- **Test Accuracy: ~69%** (Final Model)
- **Precision/Recall:**
  - Precision (Success): 0.81
  - Recall (Success): 0.77
- **Confusion Matrix:** Clear separation of successful vs failed IPOs.

```
=== NN on flag_3y_avg ===
Accuracy : 0.5949
Precision: 0.2905
Recall   : 0.7671
F1 Score : 0.4214
AUC      : 0.7056

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.55      0.69      1226
           1       0.29      0.77      0.42       292

    accuracy                           0.59      1518
   macro avg       0.60      0.66      0.55      1518
weighted avg       0.79      0.59      0.64      1518


########## Performance on flag_6m_avg ##########

=== NN on flag_6m_avg ===
Accuracy : 0.6877
Precision: 0.7500
Recall   : 0.0126
F1 Score : 0.0247
AUC      : 0.6777

Classification Report:
              precision    recall  f1-score   support

           0       0.69      1.00      0.81      1040
           1       0.75      0.01      0.02       478

    accuracy                           0.69      1518
   macro avg       0.72      0.51      0.42      1518
weighted avg       0.71      0.69      0.57      1518
```

Results LGBM Model

- **Test Accuracy: 74%** (Final Model)
- **Precision/Recall:**
  - Precision (Success): 0.81
  - Recall (Success): 0.77
- **Confusion Matrix:** Clear separation of successful vs failed IPOs.

```
########## LightGBM Performance on flag_3y_avg ##########

=== LGBM on flag_3y_avg ===
Accuracy : 0.7431
Precision: 0.3855
Recall   : 0.5651
F1 Score : 0.4583
AUC      : 0.7714

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.79      0.83      1226
           1       0.39      0.57      0.46       292

    accuracy                           0.74      1518
   macro avg       0.63      0.68      0.64      1518
weighted avg       0.79      0.74      0.76      1518
```

```
########## LightGBM Performance on flag_6m_avg ##########

=== LGBM on flag_6m_avg ===
Accuracy : 0.6825
Precision: 0.4969
Recall   : 0.6757
F1 Score : 0.5727
AUC      : 0.7509

Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.69      0.75      1040
           1       0.50      0.68      0.57       478

    accuracy                           0.68      1518
   macro avg       0.66      0.68      0.66      1518
weighted avg       0.72      0.68      0.69      1518
```

Results

- **Highest Win Rate:** 75.8**%** (Final Model)
- **Precision/Recall:**
  - Precision (Success): 0.81
  - Recall (Success): 0.77
- **Confusion Matrix:** Clear separation of successful vs failed IPOs.

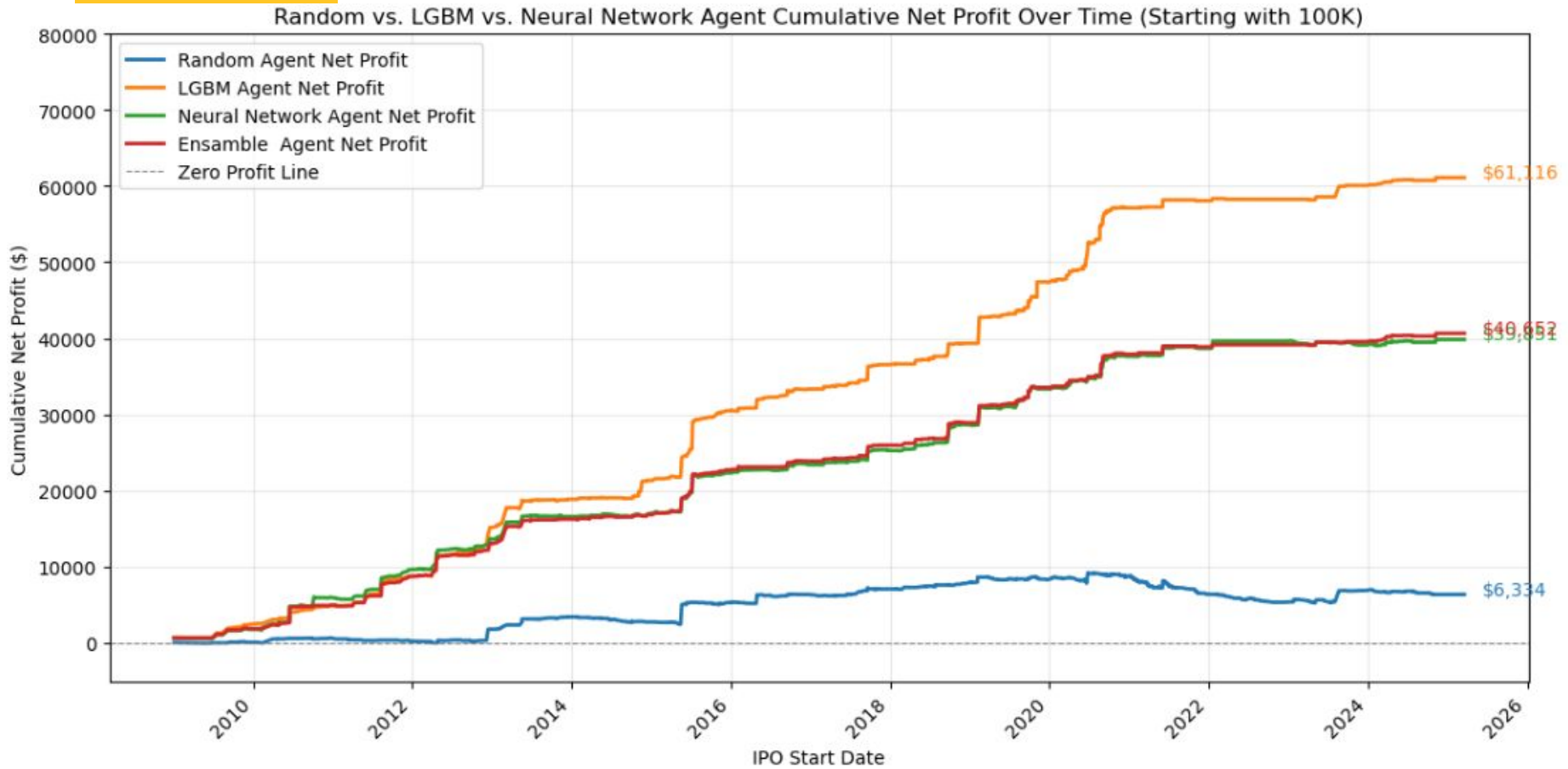
```
========== Simulation Summary ==========
Initial Capital per Strategy: $100,000.00
Investment per Signal:        $100.00
Investment for Both Signals:  $200.00
Prediction Threshold:         0.60
Total Opportunities in Test Set: 1420
Total Processed Opportunities:   1363
Total Skipped Opportunities:     57
```

Strategy	Final Capital	Total Profit
Random	$ 106,334.06	$ 6,334.06
LGBM	$ 161,116.00	$ 61,116.00
Neural Network	$ 139,850.54	$ 39,850.54
Ensemble	$ 140,652.26	$ 40,652.26

========== Extended Metrics ==========

Metric	Random	LGBM	Neural Network	Ensemble
Win Rate (%)	45.23	75.79	58.52	72.03
Total Trades	577	413	364	236
Average Win ($)	86.70	210.63	233.77	261.73
Average Loss ($)	-51.56	-48.11	-65.84	-58.22
Win/Loss Ratio	0.83	3.13	1.41	2.58
Max Drawdown ($)	3954.83	245.78	663.83	168.49
Avg Investment ($)	111.41	134.62	101.37	102.11
Sharpe Ratio	3.50	2.83	8.72	7.09

Results


Random vs. LGBM vs. Neural Network Agent Cumulative Net Profit Over Time (Starting with 100K)
Random Agent Net Profit
LGBM Agent Net Profit
Neural Network Agent Net Profit
Ensamble Agent Net Profit
Zero Profit Line
Cumulative Net Profit ($)
IPO Start Date
$61,116
$40,652
$39,891
$6,334
80000
70000
60000
50000
40000
30000
20000
10000
0
2010
2012
2014
2016
2018
2020
2022
2024
2026

Model Interpretation with SHAP

💡 What Drove IPO Success?

Top Positive Predictors

- High Net Margin
- Strong Revenue Growth
- Healthy Cash Ratio

Top Negative Predictors

- High Leverage (Debt-to-Equity)
- Low Asset Turnover Efficiency

Visual Insights Used

- SHAP Summary Plot – global feature importance
- SHAP Decision Plot – individual IPO explanations
- Correlation Heatmap – relationships among key drivers

Challenges Faced

Challenge	Mitigation
Large datasets (~71 GB raw)	Chunked processing, Drive mount
Missing fields in SEC data	Imputation, safe defaults
API limits on Yahoo Finance	Exponential backoff retries
Aligning IPO dates vs stock prices	Date matching ±2 day window
Black-box model interpretability	SHAP-based explanation

Conclusion

We developed a robust IPO success prediction model by integrating SEC financials, stock performance, and macroeconomic indicators.

The model delivers **actionable insights** for investors, VCs, and analysts, supporting better pre-IPO decision-making.

🚀 **Strategic Enhancements**

- Industry-specific model tuning
- NLP-based sentiment analysis of S-1 filings
- Inclusion of post-IPO real time deployment

1. Project Overview

IPO Success Prediction Using SEC EDGAR Filings and Macroeconomic indicators

Objective:

Build a machine learning model that predicts the **short-term success** (6 months stock growth) of companies after their Initial Public Offering (IPO), by analyzing historical SEC filings, macroeconomic conditions, and post-IPO performance.

3. Data Preprocessing - SEC

- SEC Filings: 3 tables num, sub, tag totaled to 71GB almost 200 Million rows
- Built a **CIK–Ticker mapping** from SEC JSON and TXT sources creating a unique list of IPO CIKs.
- Filtered SEC financials to **IPO companies only**.
- Used EODHD to pull adjusted stock prices at IPO start date, 6 months, and 3 years post-IPO.
- Integrated **macroeconomic indicators** from FRED:
    - Inflation rate (CPI) at IPO
    - Federal Funds Rate at IPO
- Enriched dataset with **S&P 500 trend indicators using Yahoo Finance API to find if** (price above/below SMA 30/50/200).

4. Feature Engineering

Key financial metrics engineered

- **Profitability** (Net Margin = Net Income / Revenue)
- **Leverage** (Debt to Equity = Total Liabilities / Equity)
- **Liquidity** (Cash Ratio = Cash / Liabilities)
- **Growth** (Revenue YoY, QoQ changes)
- **Efficiency** (Asset Turnover = Revenue / Assets)

Success Labelling

- **Short-Term Success**: ≥10% price gain in 6 months
- **Long-Term Success**: ≥50% gain in 3 years
- Binary classification targets created using post-IPO returns

5. Exploratory Data Analysis (EDA)

- Successful IPOs showed higher net margins and stronger liquidity.
- Revenue growth volatility was higher for failed IPOs.
- Macroeconomic indicators (high inflation, high rates) negatively correlated with IPO success.

Methodology

1. Data Collection

- **SEC EDGAR Filings (2009–2024)**: Extracted pre-IPO financials from Forms S-1, 10-K, and 10-Q.
- **Yahoo Finance**: Pulled adjusted stock prices at IPO date, +6 months, +3 years.
- **FRED Macroeconomic Data**: Inflation rates and Federal Funds Rate at IPO time.
- **S&P 500 Index**: Trend indicators using 30/50/200-day moving averages.

Latest Accomplishments (KPIs)

Data Collection & Cleaning:

- Extracted and cleaned over 15 years of SEC EDGAR data from 2009–2024 (Forms S-1, 10-K, 10-Q).
- Applied GAAP tag filtering: removed abstract/custom tags, retained standardized metrics.
- Joined tag, num, and sub tables to create a unified dataset with filing dates and financial values.

Feature Engineering:

- Derived core financial metrics:
  - Profitability: Net Margin = Net Income / Revenues
  - Leverage: Debt-to-Equity = Total Liabilities / Stockholders' Equity
  - Liquidity: Cash Ratio = Cash / Liabilities
  - Growth: YoY and QoQ Revenue and Cash Flow changes
  - Efficiency: Asset Turnover = Revenue / Total Assets

Labeling IPO Success:

- Defined "success" labels based on post-IPO performance:
  - Short-Term Success: ≥10% increase in 6 months
  - Long-Term Success: ≥50% increase in 3 years
- Matched IPO filing (S-1) data with market performance data from Yahoo Finance to apply these labels.

Exploratory Data Analysis (EDA):

- Visualized patterns in revenue, cash flow, and equity for successful vs underperforming IPOs.
- Identified initial correlations between financial ratios and IPO outcomes

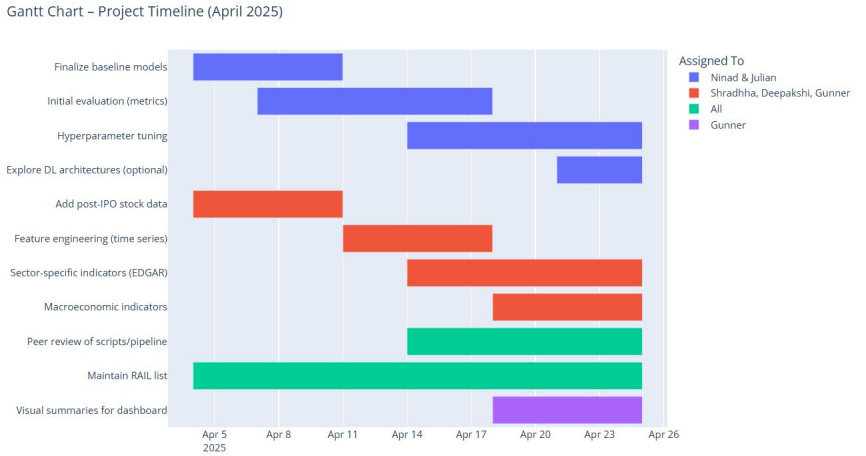
Next Major Tasks & Owners

- Model Development (Ninad & Julian)
  - Finalize baseline models (Random Forest, XGBoost) using currently cleaned features.
  - Conduct initial evaluation using classification metrics (accuracy, precision, recall, F1-score, ROC-AUC).
  - Start hyperparameter tuning using GridSearchCV or RandomizedSearchCV for optimal performance.
  - Begin exploring deep learning architectures (optional) based on current model performance.
- Data Engineering & Feature Expansion (Shraddha, Deepakshi, Gunner)
  - Augment financial data with post-IPO stock prices (Yahoo Finance, Bloomberg APIs).
  - Build additional features from time-series trends: revenue deltas, debt trends, net income change.
  - Extract sector-specific indicators from EDGAR metadata (industry codes, SIC).
  - Integrate macroeconomic indicators (e.g., inflation, Fed rates, unemployment) to enrich model context.
- Visualization Dashboards (Gunner)
  - Design and build interactive dashboards for visualization using Tableau on Financial feature distributions, Model performance metrics and Feature importance
- Team-Level Tasks
  - Conduct peer review of data cleaning scripts and modeling pipeline.
  - Maintain RAIL (Rolling Action Item List) to assign short weekly goals per member.
  - Prepare visual summaries of model performance for final dashboard.

Risks / Barriers / Obstacles

- Computational Constraints
  - Training advanced models (e.g., neural networks or large ensembles) may exceed local hardware capacity.
  - Mitigation: Use ASU's SOL supercomputer or university cloud resources for parallel processing and model training.
- Data Gaps & Incompleteness
  - Some IPOs may have missing key financial fields (e.g., EPS, Cash Flow), especially in earlier filings.
  - Mitigation: Implement imputation methods (mean/median fill, flagging), or develop logic to skip incomplete rows.
- Feature Imbalance & Noise
  - Certain features (e.g., outliers in revenue growth >1000%) may skew model behavior.
  - Mitigation: Apply scaling, clipping, and outlier detection as part of preprocessing pipeline.
- Model Interpretability
  - Ensemble models like XGBoost or black-box models may lack explainability, making it harder to interpret key drivers.
  - Mitigation: Use SHAP values and feature importance plots to explain predictions.
- Post-IPO Price Matching Difficulties
  - Aligning filing dates with corresponding market price dates may introduce alignment issues.
  - Mitigation: Use structured pipelines and date tolerance (±2 days) when fetching post-IPO returns.

Remaining Activities & How We'll Get There



Saturn Team - IPO Project Milestones

Timeline for entire project to ensure a developed and robust report by deadline