Business Acquisition Evaluation

by Estella Liu

Abstract / Motivation:
In the dynamic realm of corporate growth, M&A activities are pivotal for accelerating expansion, accessing new markets, and acquiring innovation. Our project leverages AI to predict business acquisition likelihood, utilizing advanced classification techniques to identify and value potential targets effectively.

Background / Related Work:
M&A processes, traditionally reliant on financial expertise, are increasingly augmented by AI. Machine learning models, such as classification algorithms, have revolutionized target identification and valuation in the tech-driven marketplace.

Methodologies
Acquisition-Likelihood Model:
Purpose: Predict acquisition probability using industry, geography, and funding history data.
- Data Sources & Feature Engineering:
  - Compiled Crunchbase data into a unified SQL schema.
  - Engineered features: total funding, number of funding rounds, educational credentials.
- Core Technique:
  - Employed GradientBoostingClassifier via scikit-learn.
  - Imputed missing values and one-hot encoded categorical data.
  - Outcome: Probability score indicating M&A potential.

Discounted Cash Flow (DCF) Valuation Model:
Purpose: Estimate intrinsic value by forecasting and discounting future cash flows.
- Methodology:
  - Multi-phase growth projection, operating margin evolution, and WACC calculation.
  - Free cash flow forecast and terminal value estimation.
- Output: Cash-flow schedule, DCF value, and implied equity value per share.

Synergy Value Adjustment:
Purpose: Quantify incremental value from merged operations.
- Approach:
  - Identify synergy drivers and realization timeline.
  - Discount synergy cash flows and add to DCF value.
- Output: "With Synergies" price for strategic buyers.

Automated Preprocessing with DataFrameMapper & Custom Imputers
- Used a DataFrameMapper with ValueImputer, FundImputer, ParticipantsImputer, and OnceFittedLabelBinarizer to fill missing values, one-hot encode categorical fields, and standardize numeric columns in one pass. This ensures every new company record—regardless of missing degrees, funding rounds, or unseen categories—is transformed identically to the training data, enabling consistent acquisition-likelihood scoring.

Unified Gradient Boosting Pipeline for Prediction
- A Pipeline chains the preprocessing mapper with GradientBoostingClassifier so that calling .fit() simultaneously cleans the data and trains a powerful non-linear model. This design reduces manual steps, yields reliable probability estimates for each company's chance of acquisition, and provides built-in feature_importances_ for immediately identifying top predictors.

Model Interpretation via Feature Importances, Correlations & PDPs
- After training, plot the top-10 features driving acquisition probability (via feature_importances_), display a heatmap of their correlations with the actual acquired flag, and generate Partial Dependence Plots to show each feature's marginal effect. These visualizations demystify which traits (e.g., age, funding, degrees, industry) move the "acquired" needle, helping stakeholders understand and trust the model's recommendations.

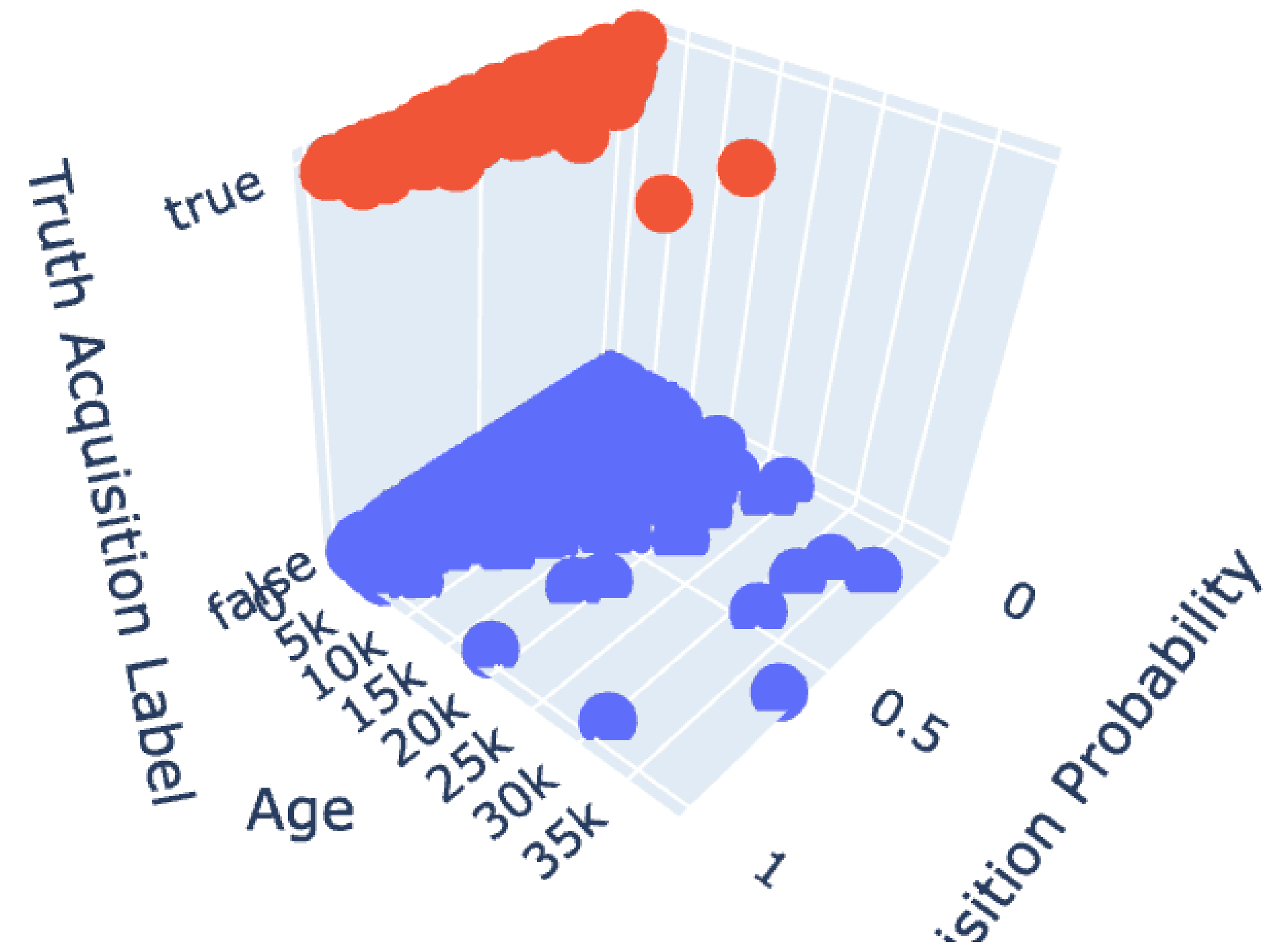


Figure 1: Predicted Acquisition Probability versus true acquisition label, plotted against age.

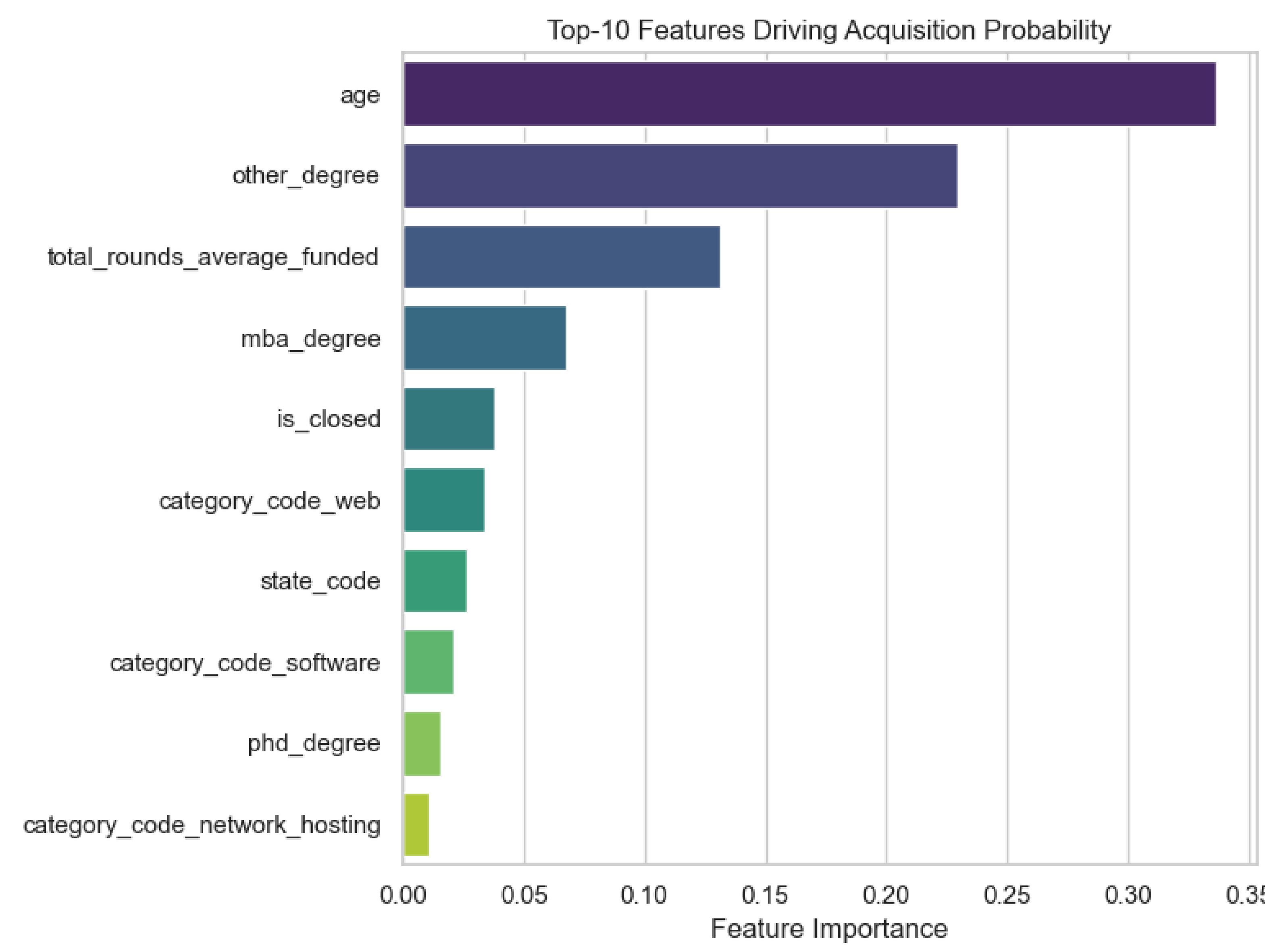


Figure 2: Top 10 most relevant features towards acquisition probability.

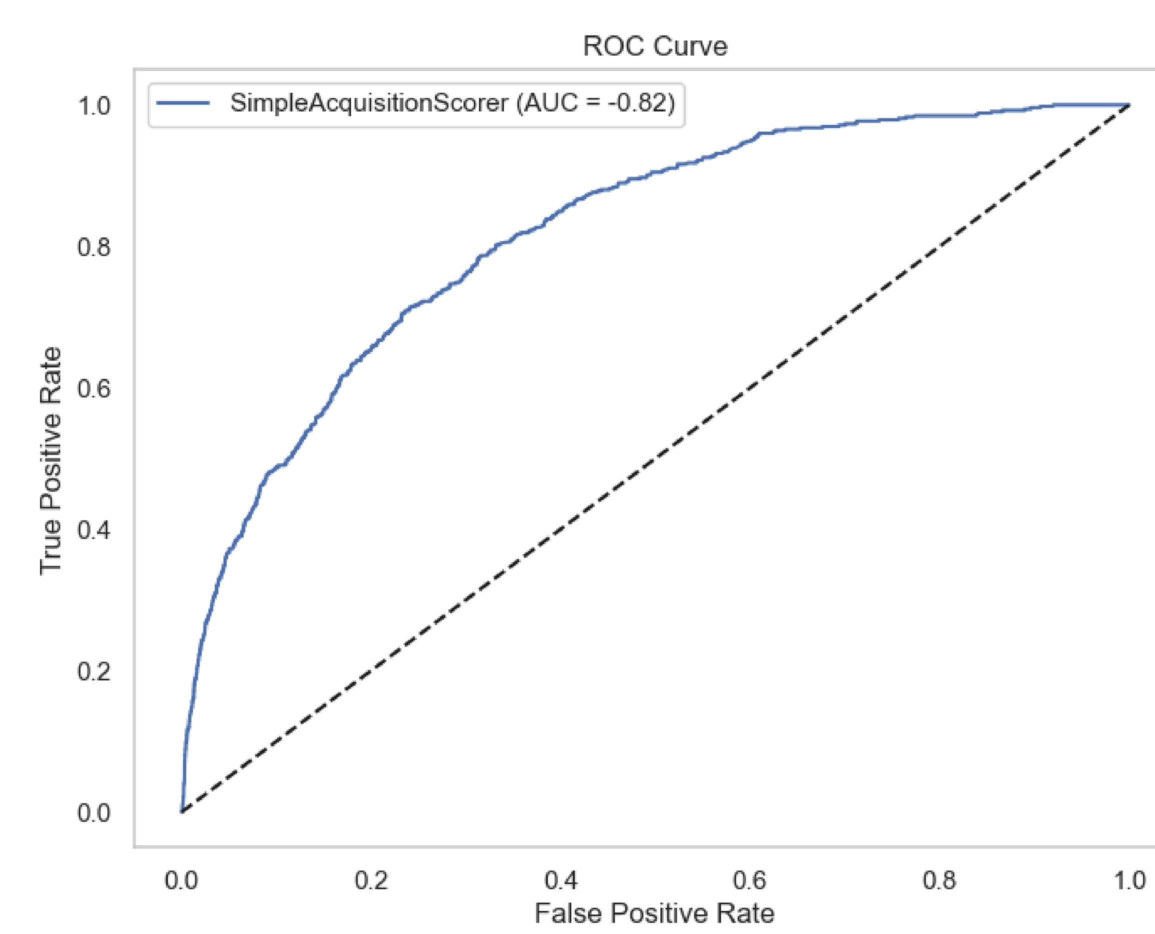


Figure 4: ROC Curve of the XGBoost scorer

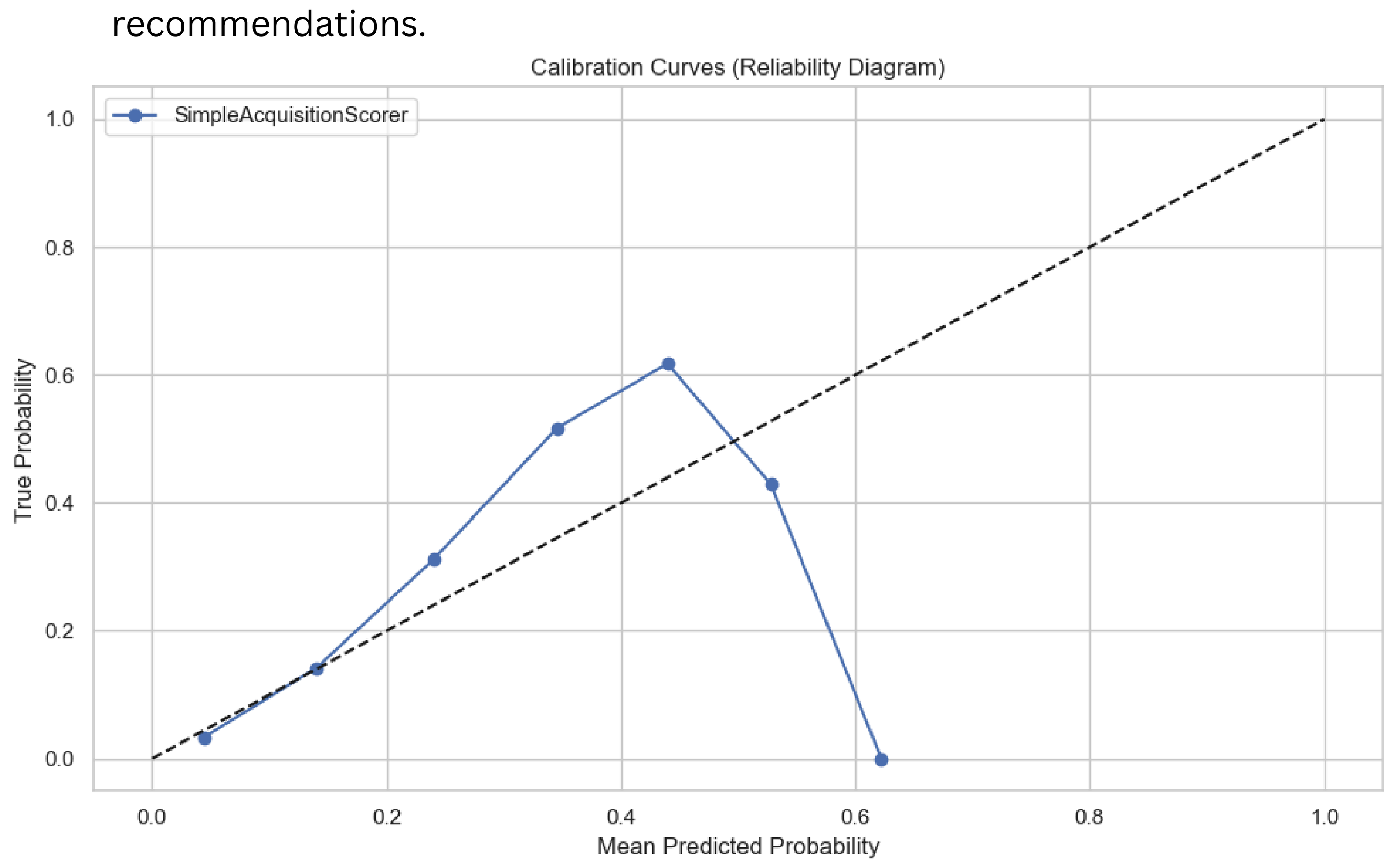


Figure 3: Expected Calibration curve of the SimpleAcquisitionScorer

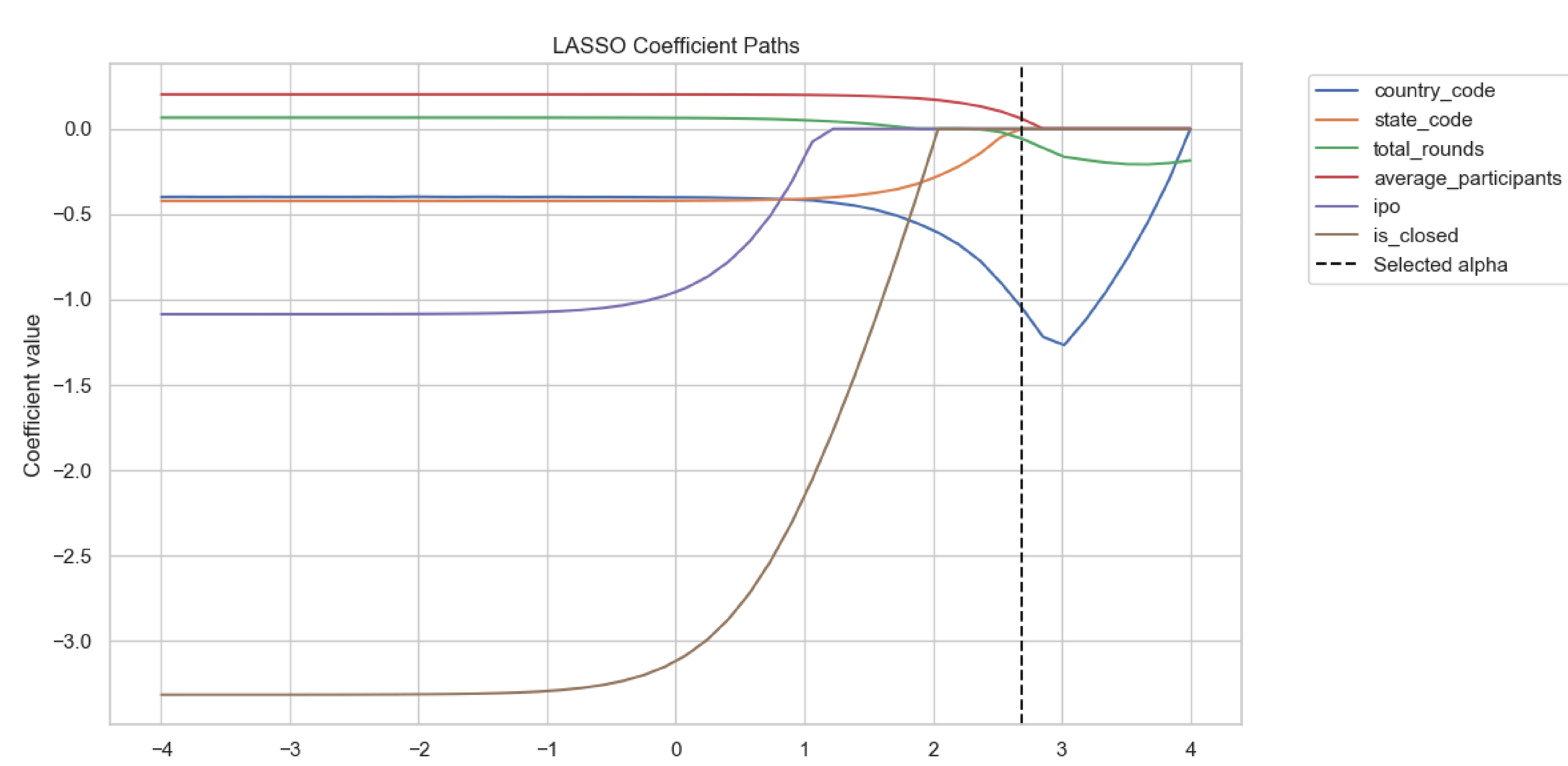
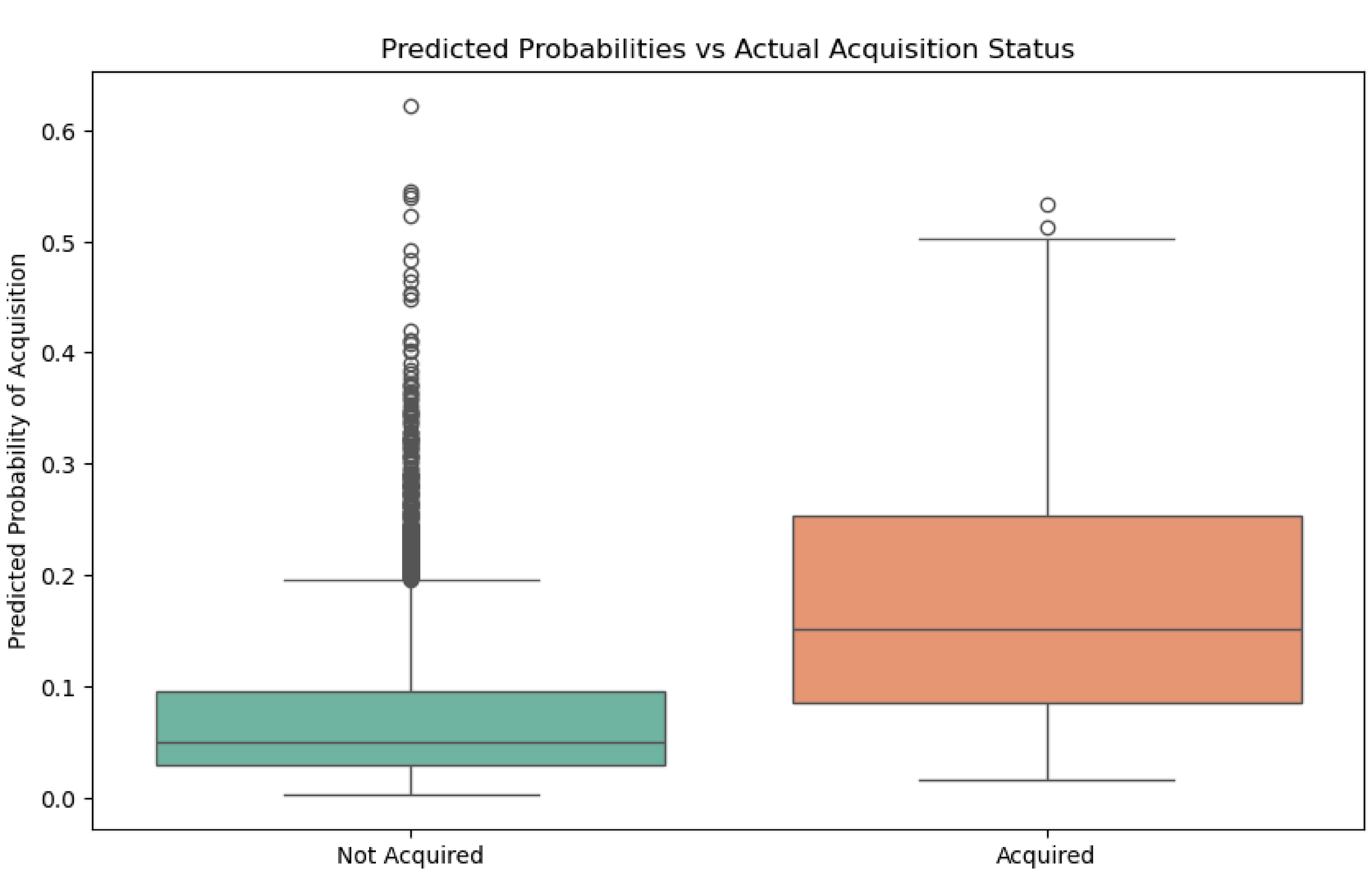


Figure 5: Plot of Lasso L2 coefficients for the model



Data & Analysis:
The data was compiled from Crunchbase, and contained row-column data on medium-sized business featrues vs acquisition rate. The data was preprocessed and parsed through the XGBoost model. It was found to have an ROC of **0.843**, a train accuracy of **0.919** and a test accuracy of **0.917.**

Results:
We found that our model performed quite well on **False** labeled values, or businessess that did not end up being acquisitied. However, while it did perform well at classifiying these, it also showed a trend of having high predicted probabilities for non-acquisitied businesses. This can be seen in the ECE curve in Figure 3 as well as the distribution of predicted probabilities in Figure 6.

Conclusions:
There is a high degree of correlation between features likle business age, degree received by founders, total rounds being funded, etc. However, there is a low rate of true positives showing that the model doesn't fit great to businesses that were acquisitied. This is also due to the lack of data available in ths subclass.