Predicting Market Capitalization

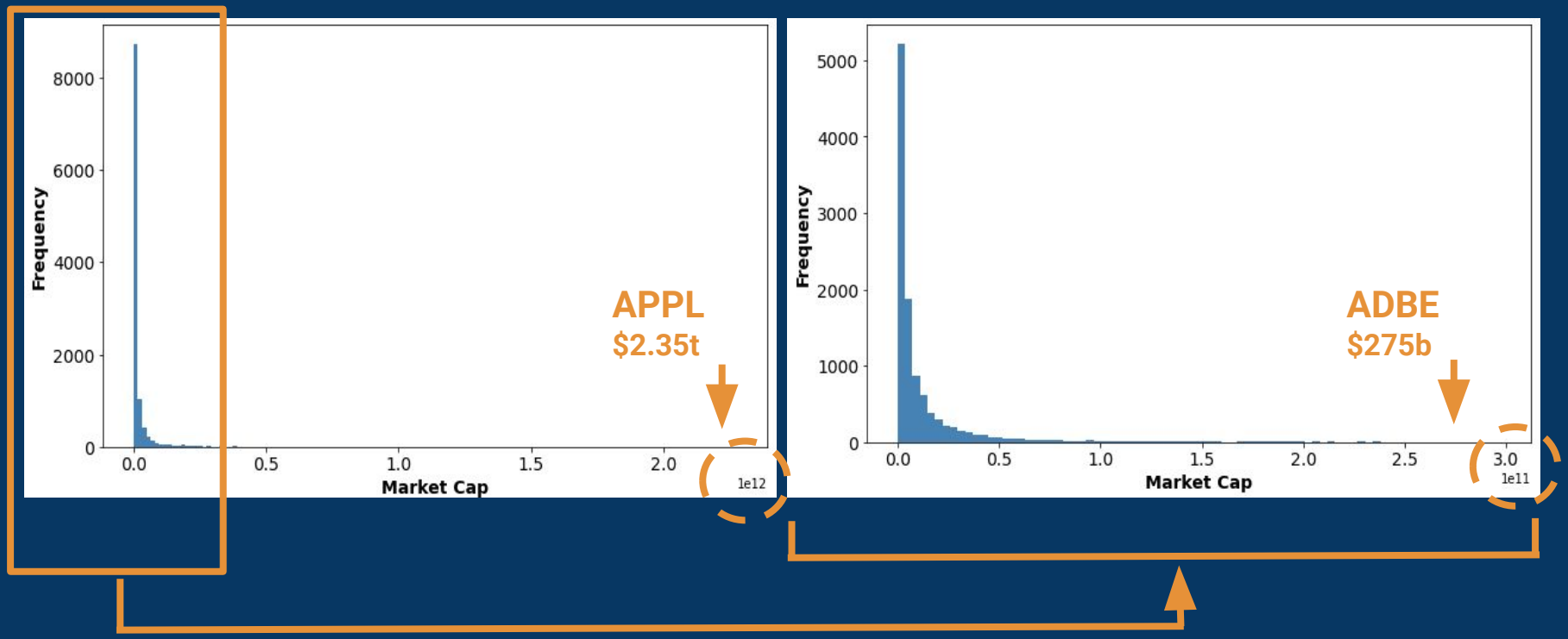
Market Capitalization

1) [Market Cap] = [# outstanding shares] x [price per share]
2) Core metric: values firm equity by public market
3) Project Goal: predict market cap of US-listed, publicly-traded companies using data from company financial statements

Market Capitalization Data

Target (Response) Variable

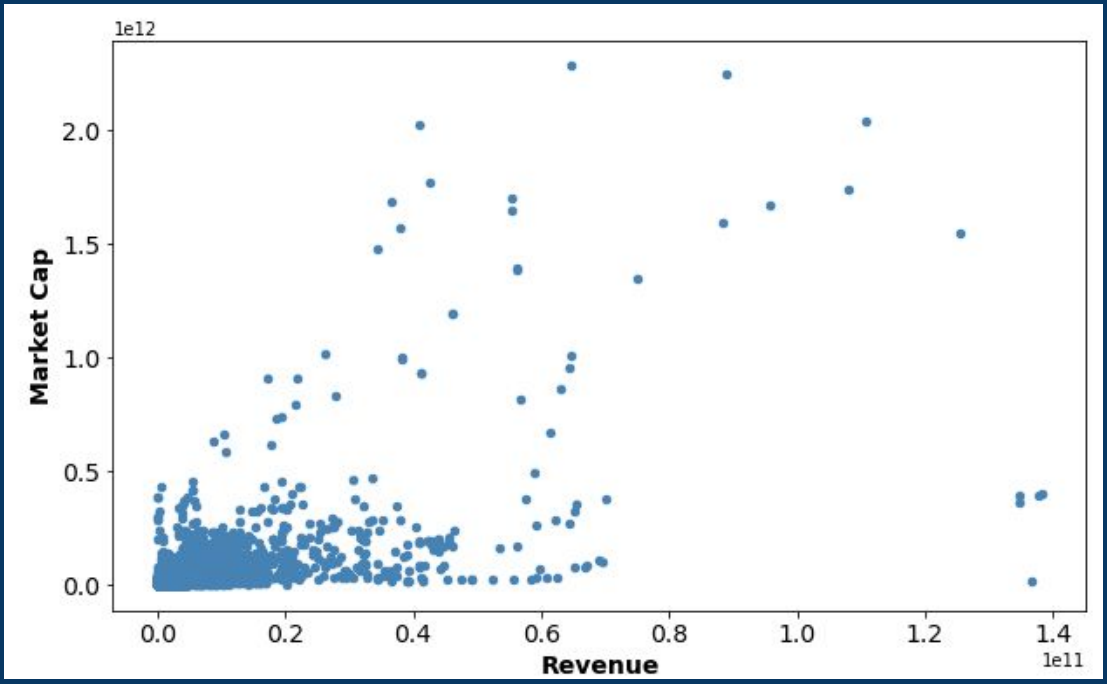
- **Market Capitalization** (USD; public, US-listed companies)
- Equals [# outstanding shares] x [USD price per share]
- Source: https://www.alphavantage.co/documentation
- >$1b to $2.35t (APPL) Market Cap
- Long-tailed distribution



Market Capitalization Data

Feature (Predictor) Variables

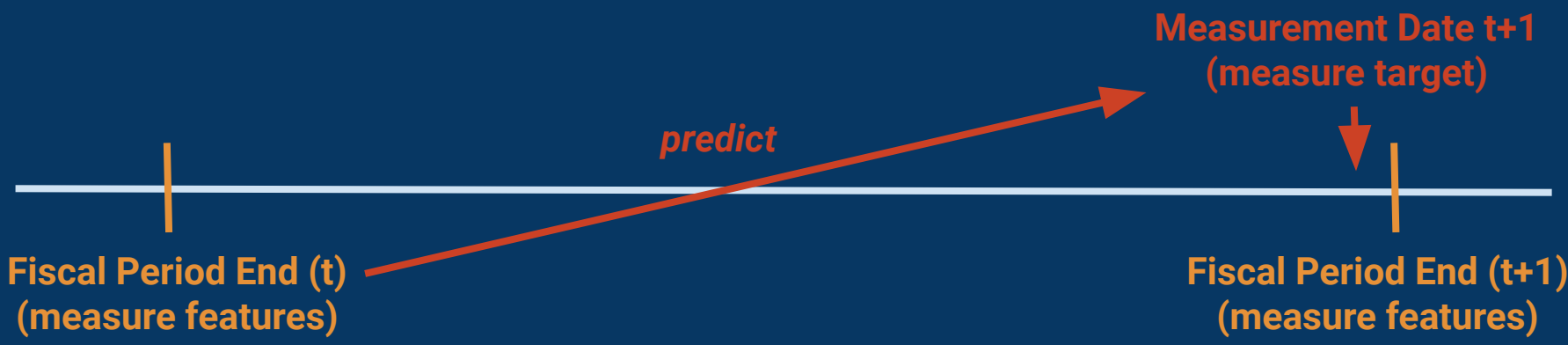
- Metrics: quarterly financial statements
- Source: https://www.alphavantage.co/documentation
- 23 different features considered
- Revenue, Gross Profit, Net Income, Operating Cash Flow, Share Issuance/Repurchase, Dividend Payout



Machine Learning Goal

Prediction Scenario

- Financial metrics measured with respect to fiscal reporting periods (Q1, Q2, etc)
- Predict next-quarter market cap given financial metrics known in the current (and prior) fiscal quarter(s)



Machine Learning Goal: predict market capitalization just prior to the next fiscal reporting date using feature measurements from the current fiscal reporting date

Exploratory Analysis: In-Scope Sample

Defining In-Scope Records

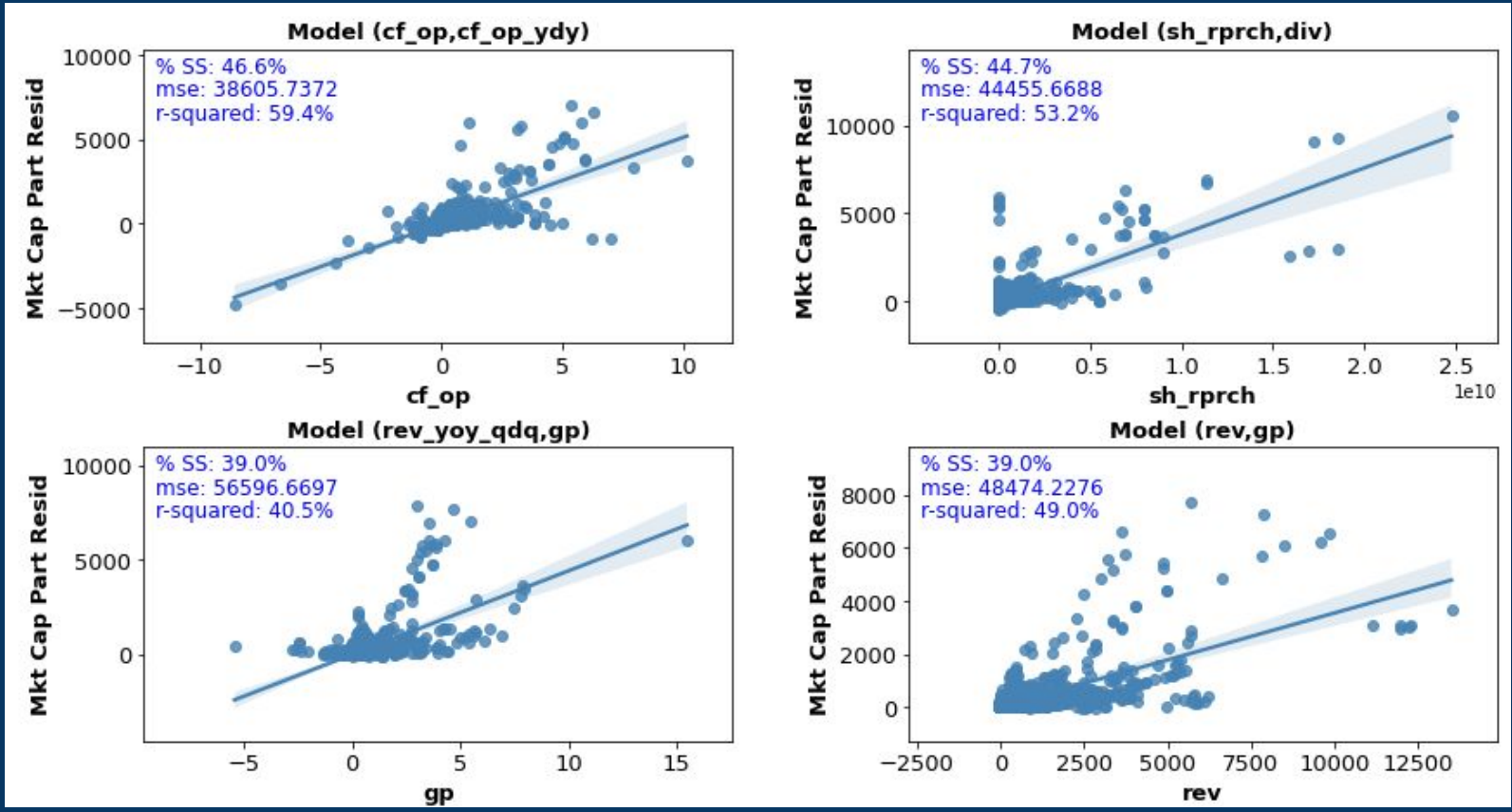
- Firms reporting in USD currency
- Records with missing (NaN) values for core metrics excluded
- Median quarterly revenue of $5m or more, others excluded
- Inconsistent, nonlogical, extreme values excluded

Final Result Set

- 2,122 distinct stock tickers
- 9,723 observations (~5 fiscal quarters per stock ticker)
- 1,692 stock tickers (80%) have 5 observations (5 fiscal quarters)
- Market Cap ranges from $1.006b to $2.358t as of Sep 2021
- Fiscal reporting dates range from 2020-01-31 to 2021-06-04
- Multiple API calls to AlphaVantage with up to 30,000 records per batch (Balance Sheet, Income Statement, Cash Flow Statement)

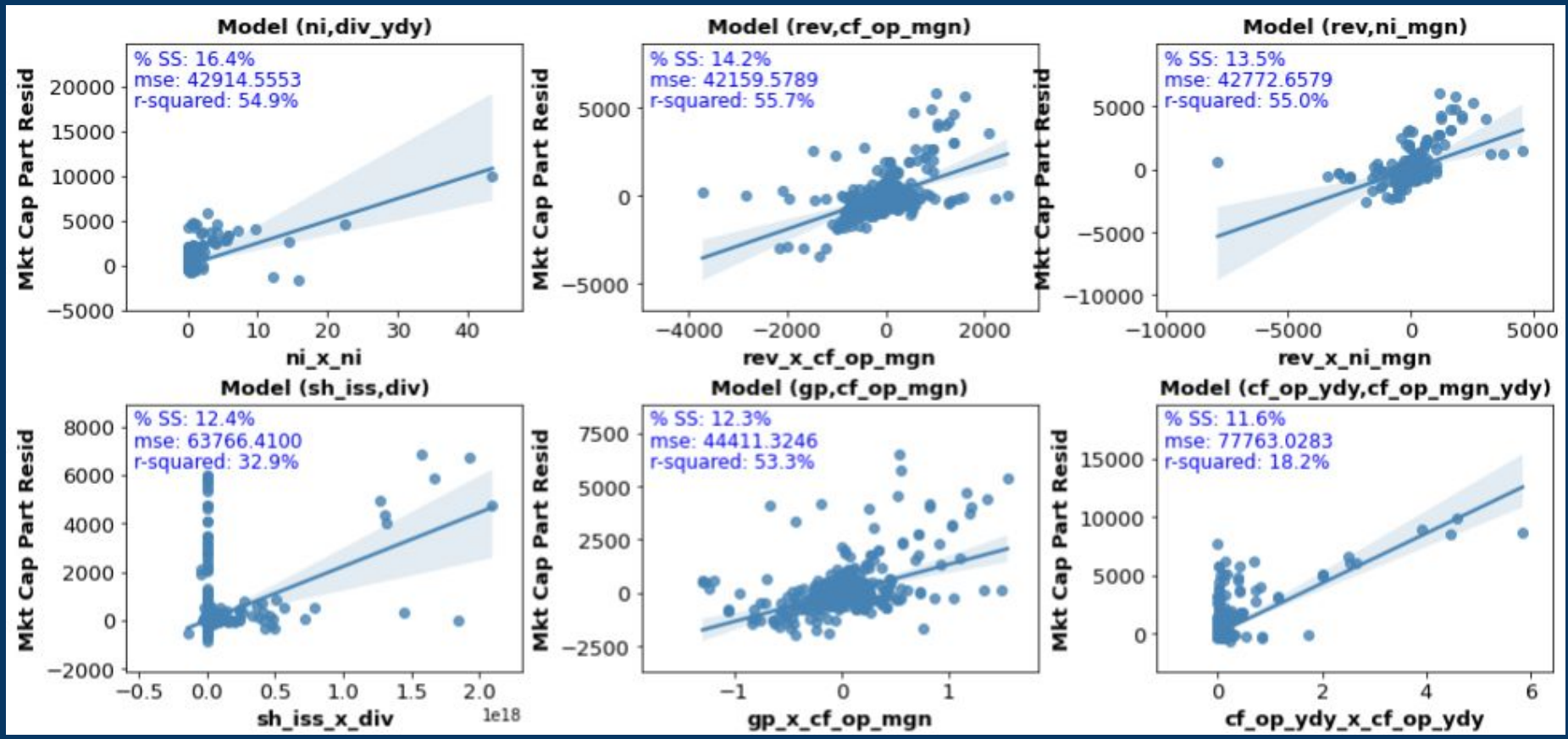
Exploratory Analysis: 1st Order Effects

- All Feature Pairs Linear Regressions (RobustScaler)
- Partial Residual Plots - Top 1st-Order Effects
  - Operating Cash Flow, Share Repurchase, Gross Profit, Revenue, Net Income
  - Top 5 1st-Order effects explain 33% to 44% of variation



Exploratory Analysis: 2nd Order Effects

- Pairwise Linear Regressions (RobustScaler)
- Partial Residual Plots - Top 2nd-Order Effects
  - Net Inc x Net Inc, Rev x Op Cash Flow, Rev x Net Inc Mgn, Sh Iss x Div
  - Top 18 2nd-order effects explain 5% to 16% of variation



Regression Models: 1st Iteration

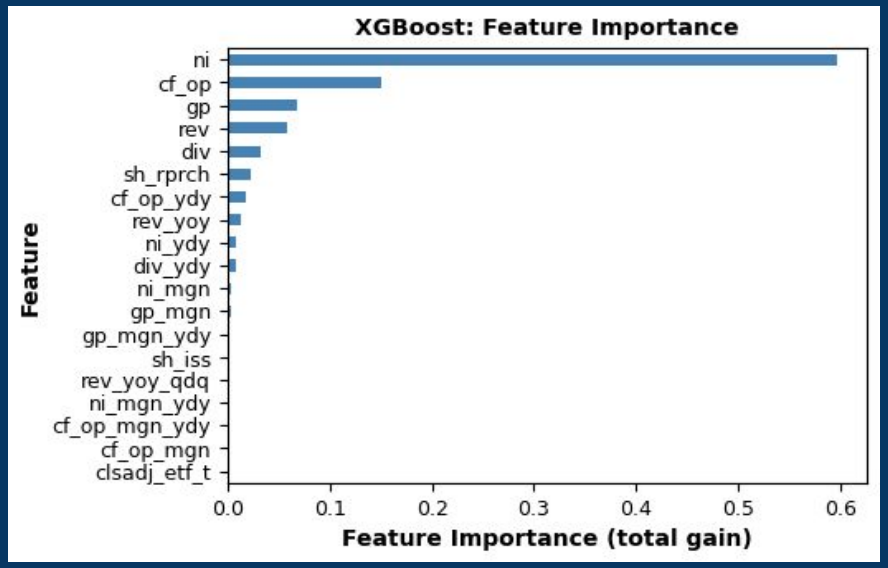
Alternative Models

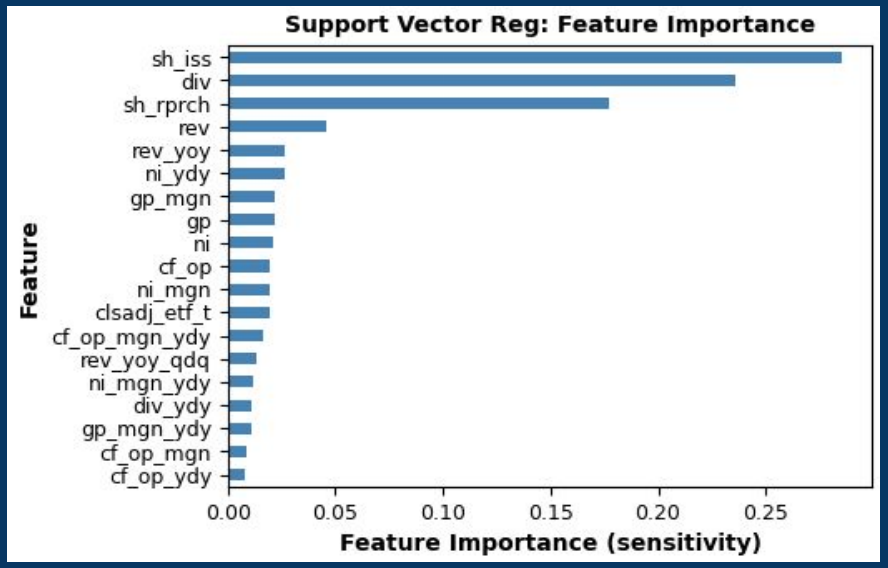
1) Linear Lasso Regression (sci-kit learn)
    a) RobustScaler
    b) Cross validation, 4-fold, MAE
        i) k-best features, alpha L1 regularization
2) Boosted Decision Tree (xgboost)
    a) RobustScaler
    b) Cross validation, 4-fold, MAE
        i) max_depth, learning_rate, reg_alpha
3) Support Vector Regression (sci-kit learn)
    a) QuantileScaler
    b) Cross validation, 4-fold, MAE
        i) max_depth, learning_rate, reg_alpha

Evaluation

1) Feature Importance
2) Absolute Percent Error (APE) on test data (80/20)

Regression 1: Feature Importance





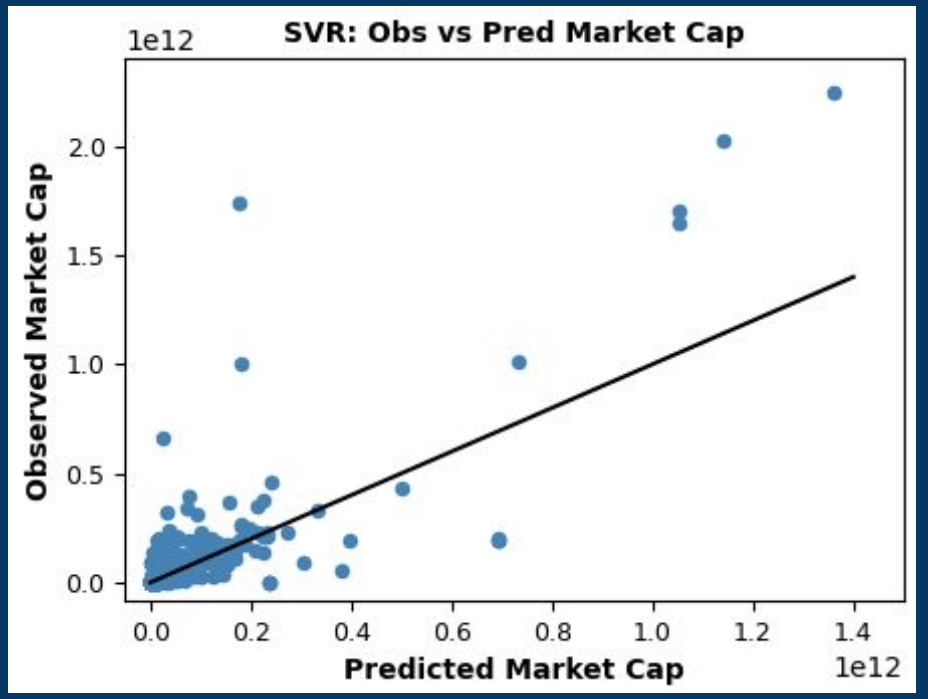
Top Features

- Operating Cash Flow
- Share Repurchase
- Net Income
- Gross Profit
- Dividend

Regression 1: Test Metrics

Model	r-squared	APE-mean	APE-median
Linear Regression	70%	235%	60%
Boosted Tree	83%	218%	62%
Support Vector Regression	64%	129%	41%

- High r-squared
- High APE: mean vs median
- Best: Support Vector Regression



Regression Models: 2nd Iteration

Feature Engineering: 1st Iteration

1) Long-tailed distributions
2) Scale units: USD amount vs percent (%)
3) RobustScaler vs QuantileScaler
4) Mean vs median APE

Feature Engineering: 2nd Iteration

Goal: Better handle distributional properties

1) Add 4 indicator features (gp, div, sh rep, sh iss)
2) SQRT transform all USD amounts
3) Regress all USD amounts on Revenue, use residuals
4) Scale all variables to 0 mean, 1 std dev

Regression Models: 2nd Iteration

Alternative Models

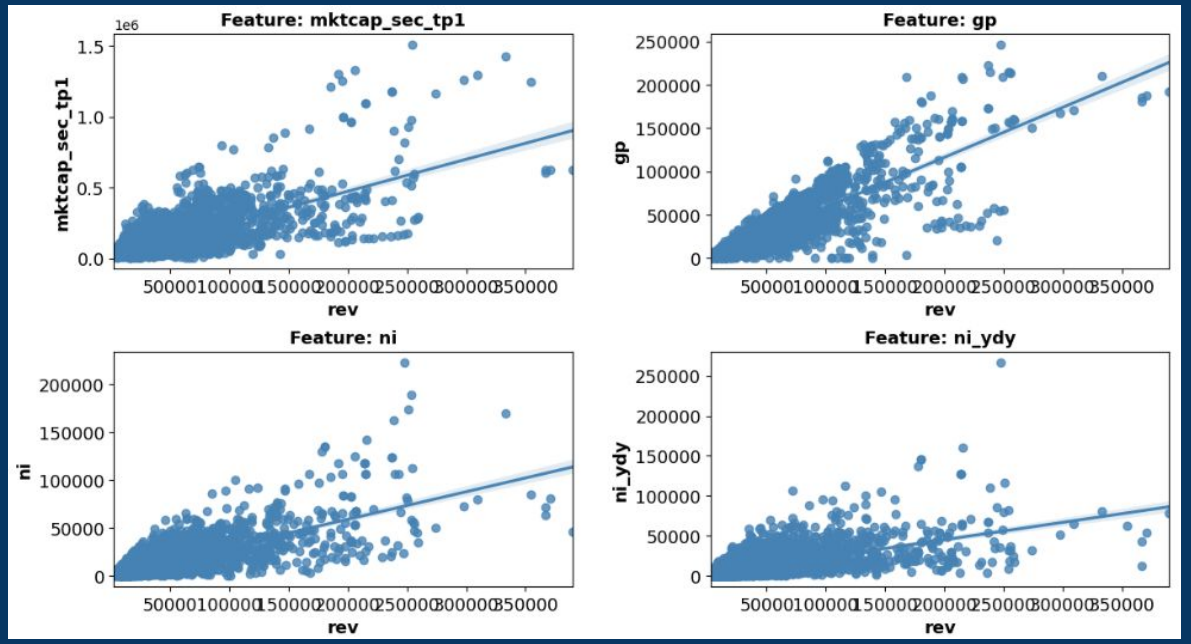
1) **Linear Lasso Regression** (sci-kit learn)
2) **Boosted Decision Tree** (xgboost)
3) **Support Vector Regression** (sci-kit learn)
4) **Neural Network** (keras, tensorflow)
    a) Linear + Nonlinear layers, vary L1 regularization

Evaluation

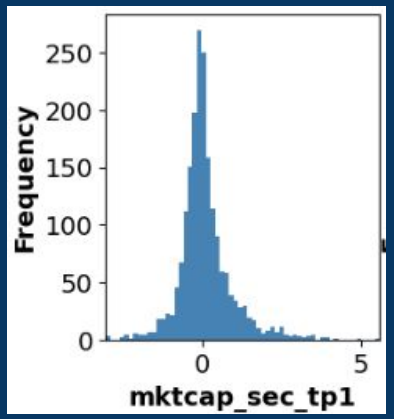
1) Feature Importance
2) Absolute Percent Error (APE) on test data (80/20)

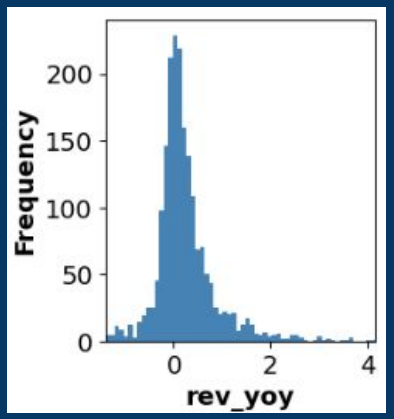
Regression 2: Feature Engineering

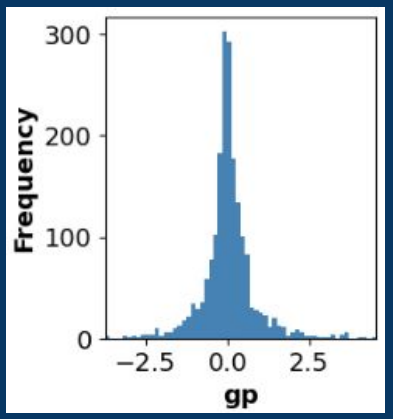
- SQRT transform
- Regression on revenue



- 0 mean, 1 std dev
- Transformed variables



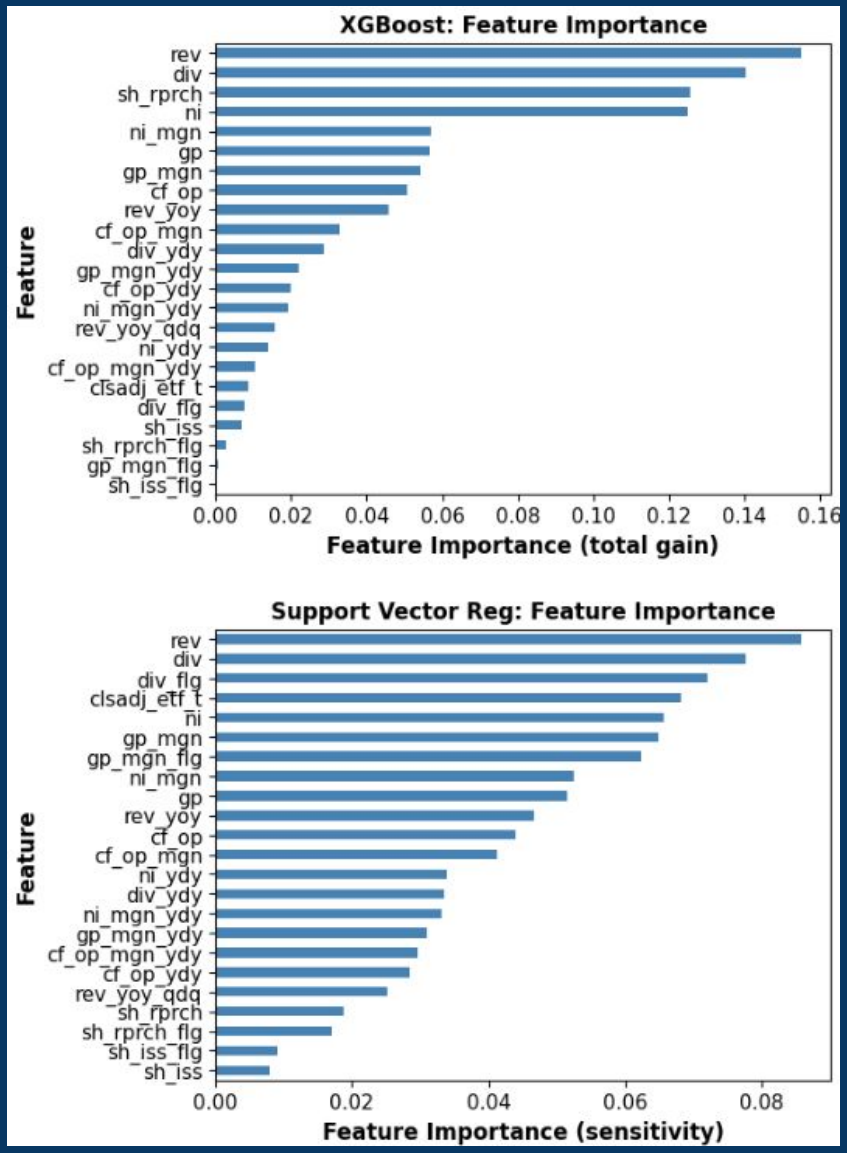




Regression 2: Feature Importance

Feature Importance

- More evenness
- Revenue still top feature
- Top features
  - Revenue
  - Dividend
  - Net Income
  - Share Repurchase
  - Gross Profit



Regression 2: Test Metrics

Model	r-squared	APE-mean	APE-median
Linear Regression	67%	176%	47%
Boosted Decision Tree	89%	151%	40%
Support Vector Regression	36%	150%	43%
Neural Network 1	12%	237%	46%
Neural Network 2	83%	193%	47%

- High r-squared
- High APE: mean vs median
- NN models no improvement vs linear model
- Best: Boosted Decision Tree

Recommendations & Next Steps

Recommendations

- Boosted decision tree for one-quarter-ahead forecast
- Top features
  - revenue
  - dividend amount
  - share repurchase amount
  - net income
- Segment the data (e.g., sector, firm size)

Next Steps

- Exploratory model, residual analysis: feature effects
- Mixture models based on data segmentation
- Enhance neural network model