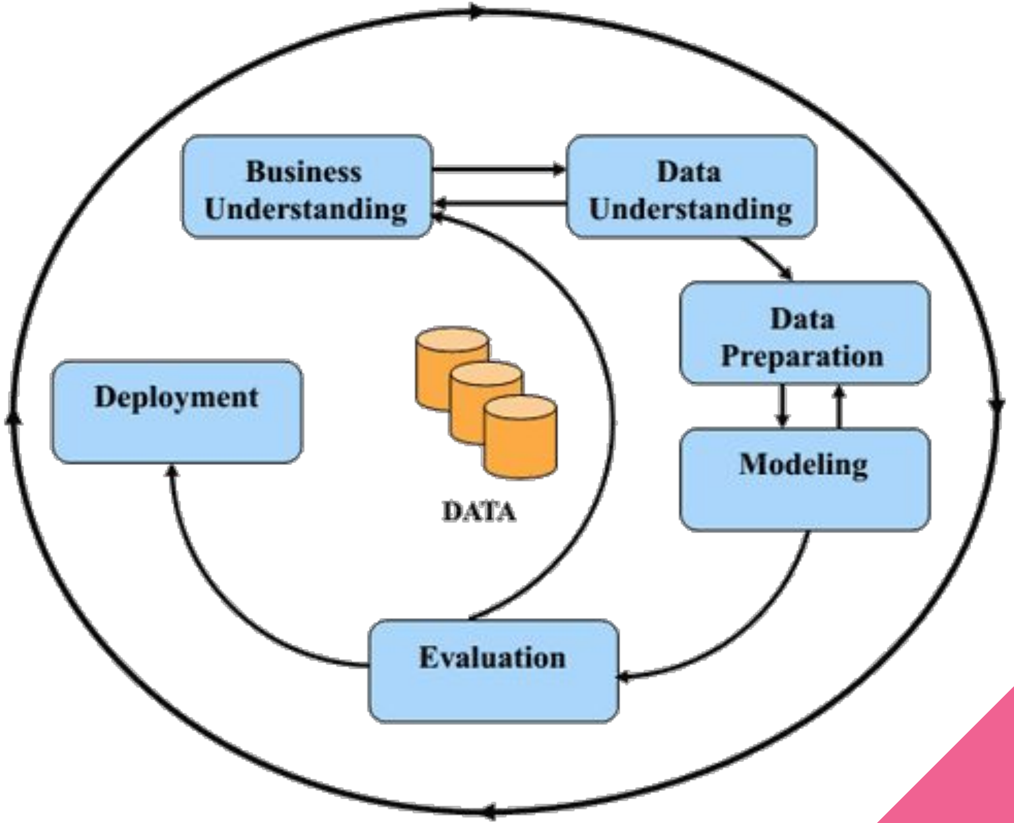
# Predicting Female Board Representation

Emmy Fortunato, Sean Jung,
Stephen Thomas, Qiyu Wang

# For this presentation, we will follow CRISP-DM



# Business Understanding

## We are motivated by answering the following business question:

**How can we use machine learning to predict the annual percentage of female representation on the board of directors of S&P 1500 American corporations?**


- Female board representation below 20 percent
- Gender-based diversity stimulates firm success

# Data Understanding

# We have used the following data sources.

wrds
The Global Standard for Business Research
Data sourced from Wharton Research Data Services (WRDS)

## ExecuComp: **Director Compensation**

- 107,908 observations and 107 columns
  - 2,162 unique companies
- Captures annual metrics regarding directors' compensation and personal characteristics
  - Age
  - Gender
  - Location
  - Company headquarters

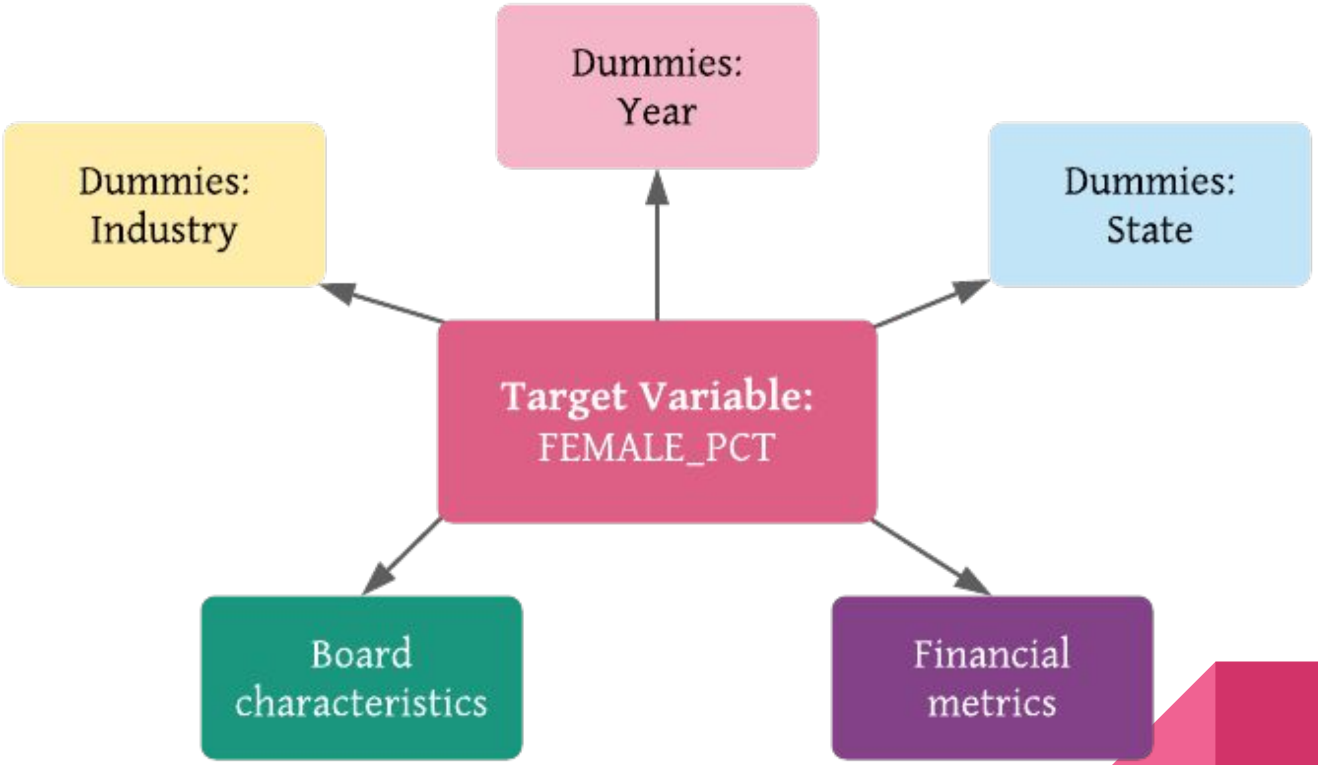
## Compustat: **Financial Ratios**

- 480,739 observations and 76 columns
- Firms' financial data:
  - Capitalization
  - Efficiency
  - Financial soundness
  - Solvency
  - Liquidity
  - Profitability
  - Valuation ratios

## BoardEx: **Organization Summary Analytics**

- 757,192 observations and 60 columns
- Annual board-related metrics:
  - Directors' tenure on their respective boards and in their respective companies,
  - Overall size of each director's network
  - Degree of nationality
  - Diversity of their board

# Our target variable and features are defined as:



# Data Preparation

# This is how we engineered the target variable

- Aggregated director-level instances **by company and year** to get the average board characteristics and company information on company-level
- Used the **financial metrics from the last month** of each year for each company; the **year-end records** present the snapshot/summary of performances of the year.
- Merged the three datasets based on common keys: ***'TICKER'*** and ***'YEAR'***
- **Normalization** since we used *k*NN and neural nets

# This is how we engineered the features

- **CEO_FEMALE** → Binary variable indicating whether the CEO is female or not
- **EXECDIR_COUNT** → Number of board members who served as executive director that year
- **EXEC_COUNT** → Number of board members
- **LENGTH_TERM** → The length of term CEO serve
- **FEMALE_PCT_LAST** → Last year's percentage of female representation for each company

# Modeling

# Modeling Processes

- **Target variable** → A numeric/fractional value, a percentage of female representation in each company for each year
- **Linear Regression** → Better speed of learning and comprehensibility
- **Ensemble Methods** → Better generalization performance
- **80(train)/20(test) split** → Evaluate the final performance
- **Cross-validation with five folds** → Optimize the hyperparameters of each model

# Results from Models

| Ordinary Least Squares Regressions | | |
|---|---|---|
| | OLS (statistical analysis) | OLS (machine learning) |
| R-squared | 0.722 | 0.708 |
| MAE | n/a | 0.044 |
| RMSE | n/a | 0.005 |

| Ridge, Lasso, Elastic Net Regresion | | | |
|---|---|---|---|
| | Ridge (L2) Regression | Lasso (L1) Regression | Elastic Net (L1 +L2) Regression |
| R-squared | 0.720 | 0.635 | 0.703 |
| MAE | 0.043 | 0.061 | 0.048 |
| RMSE | 0.005 | 0.005 | 0.005 |

# Results from Models

Ensemble Regressions

| | Stacking Regressor | AdaBoost Regressor | Bagging Regressor |
|---|---|---|---|
| **R-squared** | 0.856 | 0.702 | 0.713 |
| **MAE** | 0.029 | 0.047 | 0.048 |
| **RMSE** | 0.002 | 0.005 | 0.005 |

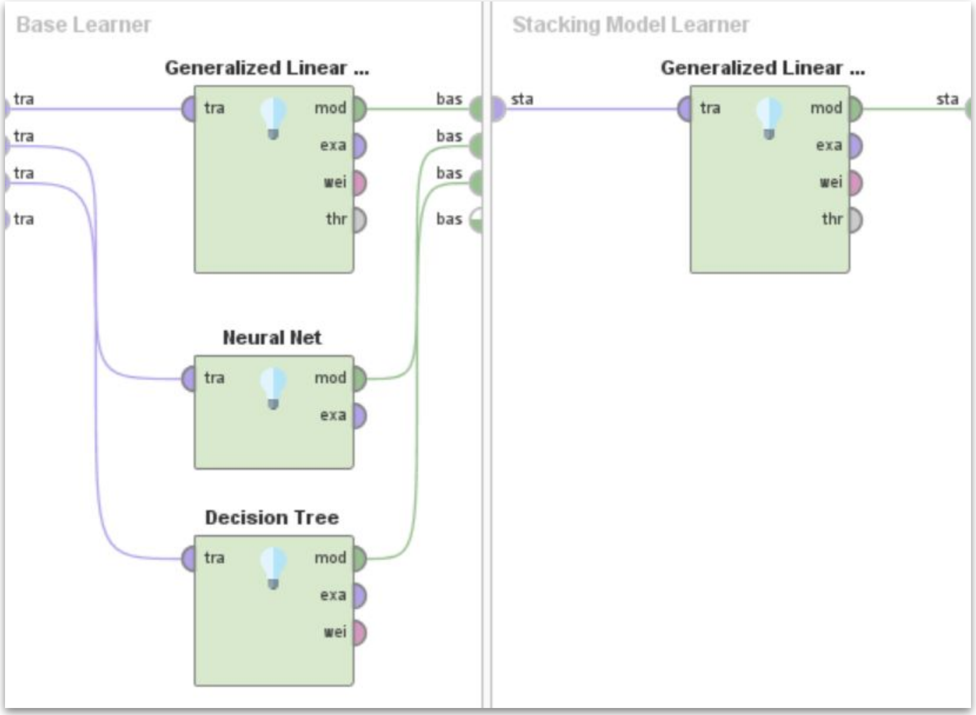
Tree-based Regressions

| | Decision Tree Regression | Random Forest Regressor | ExtraTrees Regressor |
|---|---|---|---|
| R-squared | 0.692 | 0.759 | 0.762 |
| MAE | 0.044 | 0.038 | 0.039 |
| RMSE | 0.005 | 0.004 | 0.004 |

| Gradient Boosted Tree Regression | Hist Gradient Boosting Regression |
|---|---|
| 0.692 | 0.732 |
| 0.043 | 0.042 |
| 0.005 | 0.005 |

# Evaluation and Deployment

# Our final model selected was a Stacking model.




**Performance Metrics**

$R^2$ = .856

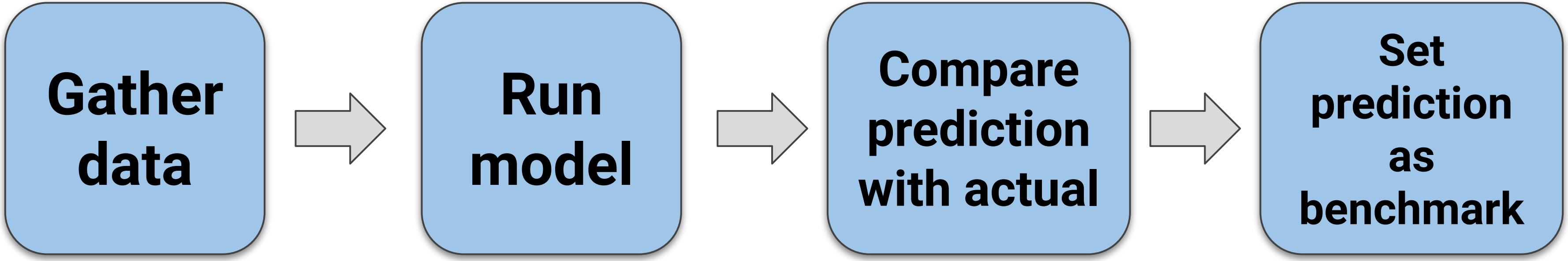
RMSE = .051

MAE = .030

# One deployment use case is understanding the impact of the most influential features.

| | Board Characteristics | Financial Ratios | Industries |
|---|---|---|---|
| **Negative** | • Average Director Age<br>• Average Director Salary | • High Profit<br>• Low Debt | • Construction<br>• Fuel<br>• Manufacturing<br>• Technology |
| **Positive** | • StDev Director Ages<br>• StDev Director Tenure<br>• Nationality Mix<br>• Network Size | • Low Invested Capital<br>• Low Assets Value | • Academia<br>• Consumer Goods<br>• Arts and Travel |

# Another deployment use case is using predictions to establish company gender diversity benchmarks.




*Data is a snapshot of the state of the company very end of the calendar year.

# Thank you!