# Imputing missing data using machine learning methods

Central Bank Business Survey Conference
October 29-30, 2024

Emil Mihaylov
Quantitative Analysis Specialist
Research Department

Federal Reserve Bank of Atlanta

# Introduction

- Machine learning can offer a viable alternative to current methods for imputation of missing data in business surveys.
- In the context of decision science, being able to rely on accurate data is crucial for making the best-informed decisions.
- ML models yield significantly superior predictive performance compared to current benchmarks
- Particularly, models that can capture complex relationships in the data, such as Random Forest and Neural Network work exceptionally well.

# About the SBU survey

- The Survey of Business Uncertainty (SBU), is a monthly business survey that elicits information from firms about employment, sales revenue, and prices and unit costs.
- The SBU survey is fielded to business executives from a broad range of industries and firm sizes.
- The panel consists of more than 2000 firms, which provide roughly 900 responses per month
- SBU’s special questions are instrumental for assessing business activity, economic conditions, and emerging issues.

# The issue with missing data in the SBU survey

- Core survey responses are often weighted using activity weights
- Results from special questions are typically weighted by firms' levels of current employment
- Every month, new respondents assigned to the sales or prices and costs questionnaires will not provide information about their level of employment
- The responses of those panelists will not have the corresponding employment weights
- To remediate this issue, we typically impute missing observations using employment data from the data providers, but those data are generally very stale.
- We also use a linear regression model to predict missing weights.

# What is machine learning?

- "Machine learning is a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data or other outcomes of interest." - "Deep Learning" by Goodfellow et al. (2016)
- ML models in general have more relaxed assumptions regarding data structures compared to traditional statistical models
- Many ML algorithms are very good at capturing non-linear relationships in the data
- Five supervised-learning regression models are compared in this exercise:
  - K-Nearest Networks (KNN)
  - Support Vector Regression (SVR)
  - Neural Networks
  - Ridge Regression
  - Random Forest

# Testing the performance of ML models - framework

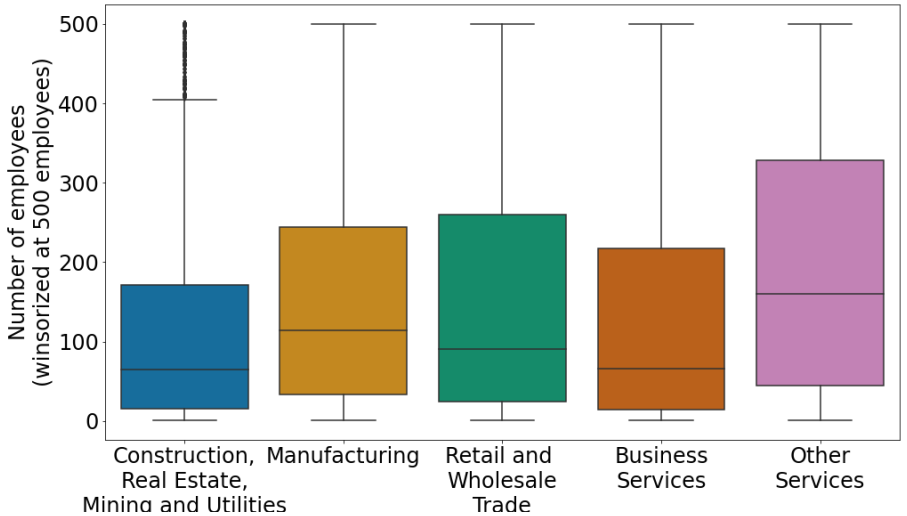
- Predictors: the models will utilize a set of firm characteristics (such as size, industry, location, and legacy employment data) and variables related to the date and duration of responses, as well as level of sales revenue.

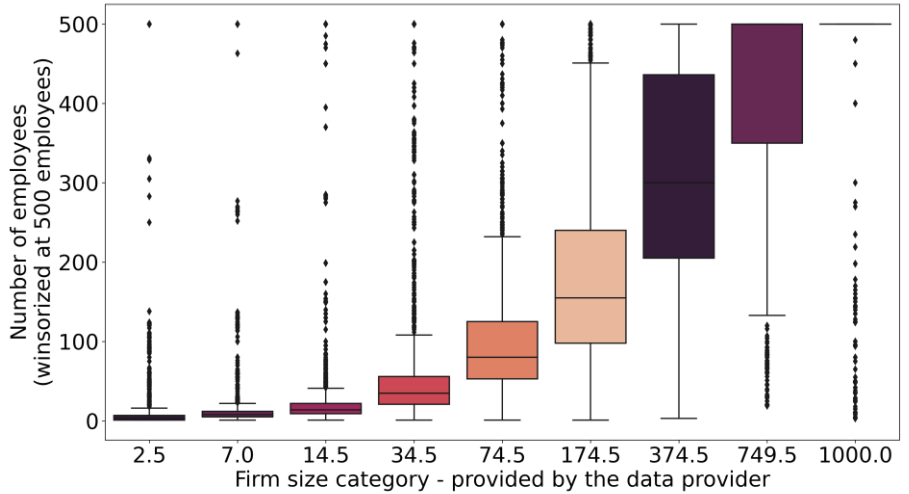
- All models are first fine-tuned on a subset of the data using five-fold cross validation

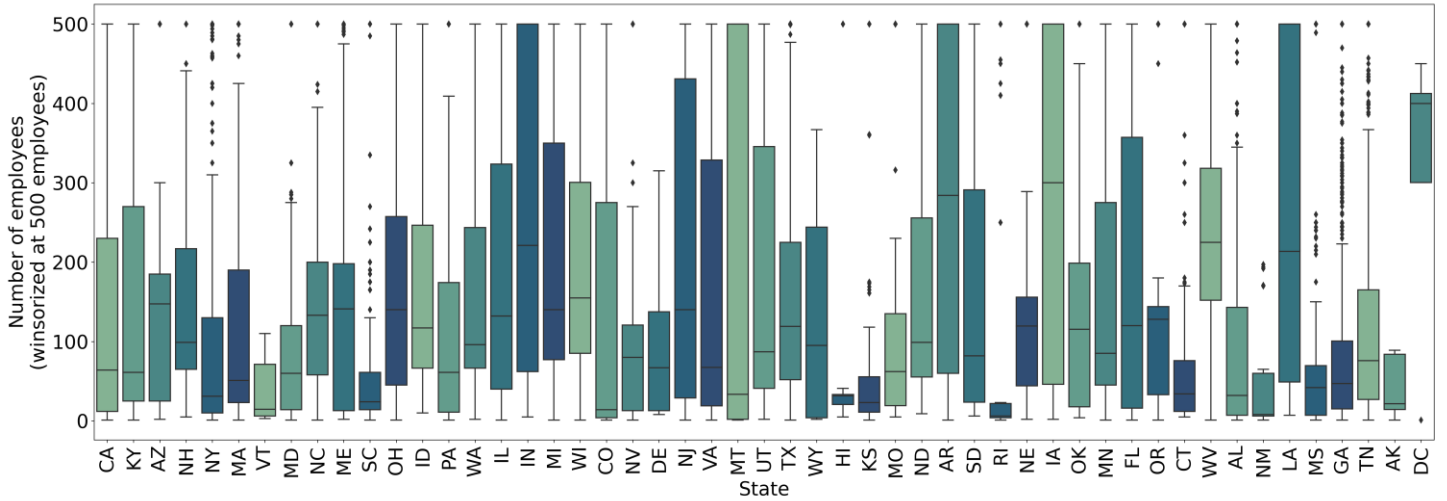
- Validation method : Monte Carlo Cross Validation – 20 iterations
  - Randomly split the data into 80/20 train/test splits
  - Train the models, make predictions using the test data, and calculate RMSEs

- Performance metric: Root mean squared error (RMSE) will be used to quantify the accuracy of responses.

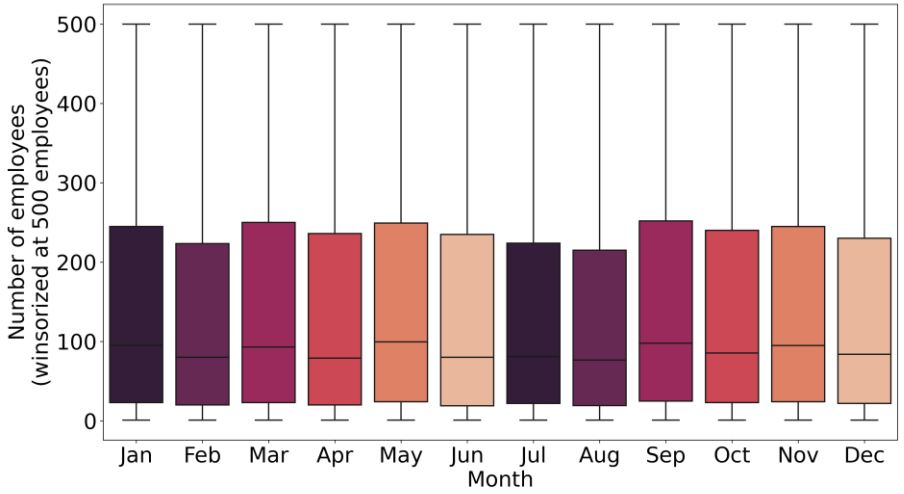
# Firm employment vary across sates and major industries

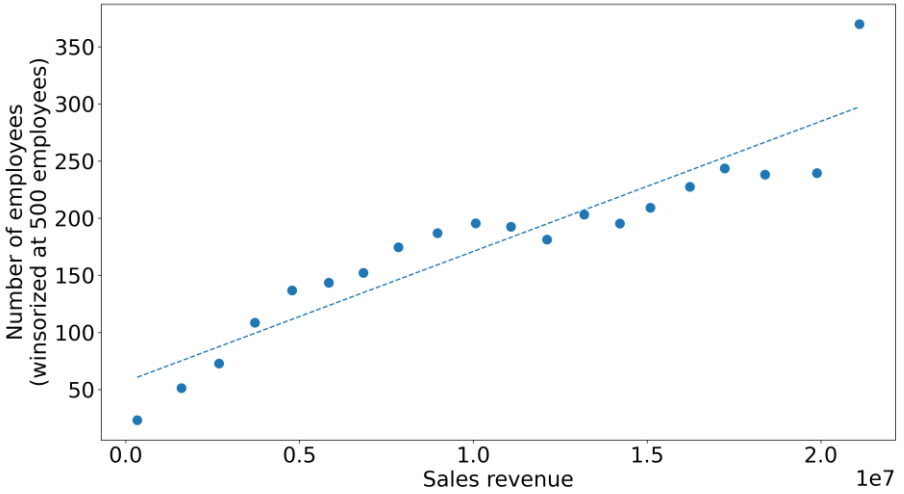


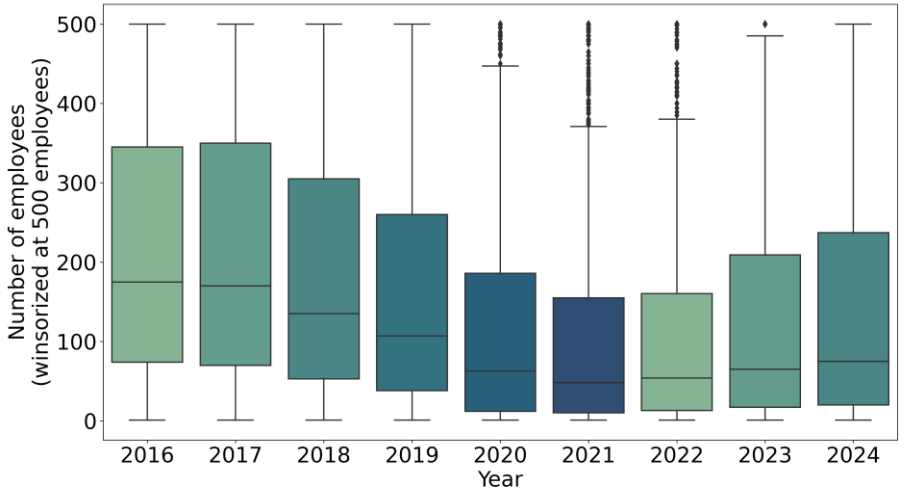



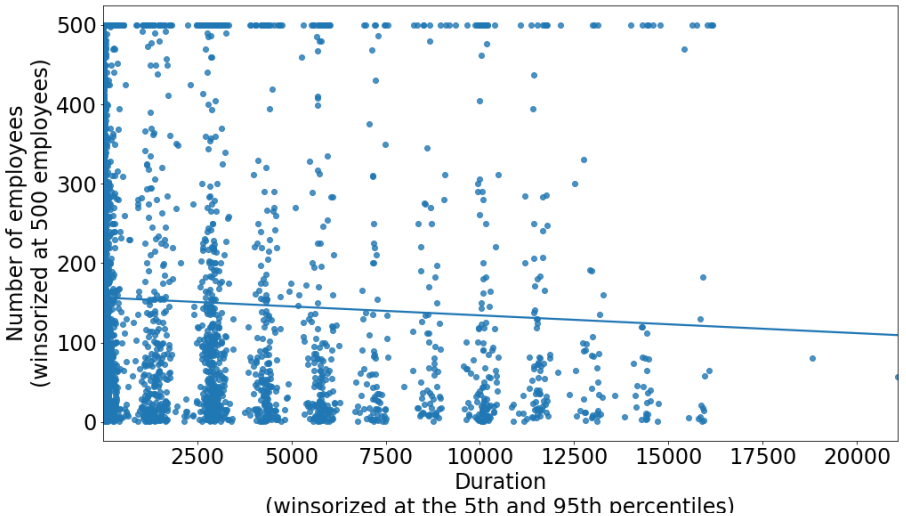



# Levels of sales revenue, time, and duration of responses also appear to be related to firm employment



# Machine Learning models enhance the imputation of missing data compared to the benchmark methods

## Results from Monte Carlo Cross-Validation exercise

| RMSE | *Emp. data from provider* | *SBU OLS Regression* | Ridge Regression | Support Vector Regression | KNN Regression | Neural Network Regression | Random Forest Regression |
|---|---|---|---|---|---|---|---|
| Average | *177.25* | *107.34* | | | | | |
| Standard Dev. | *3.67* | *2.60* | | | | | |
| Min | *168.0* | *102.0* | | | | | |
| 25% | *175.3* | *105.0* | | | | | |
| 50% | *177.2* | *108.3* | | | | | |
| 75% | *179.3* | *109.5* | | | | | |
| Max | *184.6* | *110.3* | | | | | |

- The benchmark methods are not so very accurate at imputing current employment levels

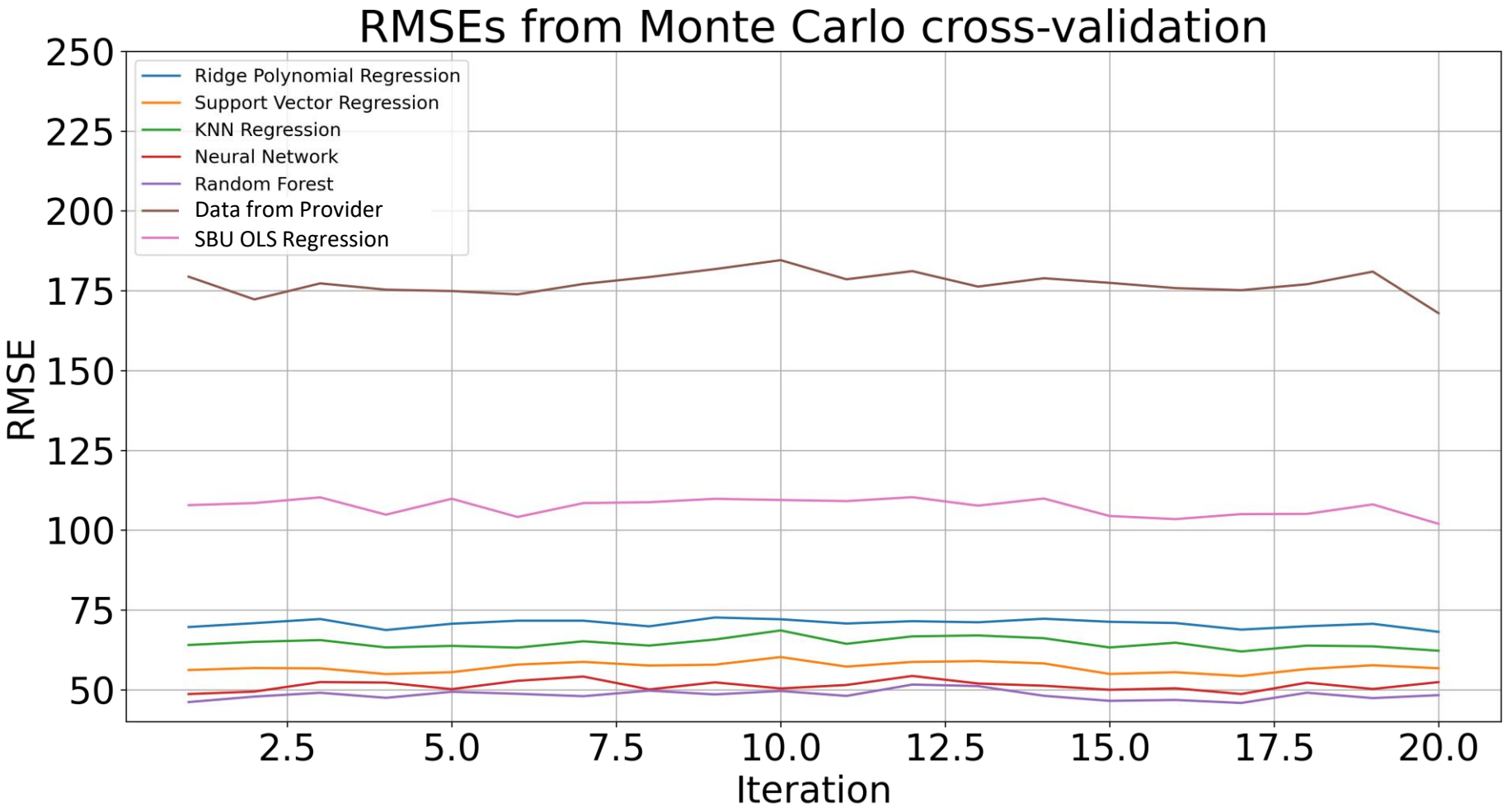
# Machine learning models enhance the imputation of missing data compared to the benchmark methods

## Results from Monte Carlo Cross-Validation exercise

| RMSE | *Emp. data from provider* | *SBU OLS Regression* | Ridge Regression | Support Vector Regression | KNN Regression | Neural Network Regression | Random Forest Regression |
|---|---|---|---|---|---|---|---|
| **Average** | *177.25* | *107.34* | 70.75 | 57.05 | 64.59 | 51.27 | 48.36 |
| **Standard Dev.** | *3.67* | *2.60* | 1.24 | 1.55 | 1.66 | 1.61 | 1.51 |
| **Min** | *168.0* | *102.0* | 68.1 | 54.3 | 62.0 | 48.6 | 45.8 |
| **25%** | *175.3* | *105.0* | 69.9 | 56.0 | 63.5 | 50.2 | 47.4 |
| **50%** | *177.2* | *108.3* | 70.9 | 57.0 | 64.2 | 51.4 | 48.2 |
| **75%** | *179.3* | *109.5* | 71.6 | 58.0 | 65.6 | 52.3 | 49.1 |
| **Max** | *184.6* | *110.3* | 72.6 | 60.2 | 68.6 | 54.3 | 51.6 |

- Machine learning models outperform the benchmarks by a long stretch
- Random Forest and Neural Networks are the most accurate models

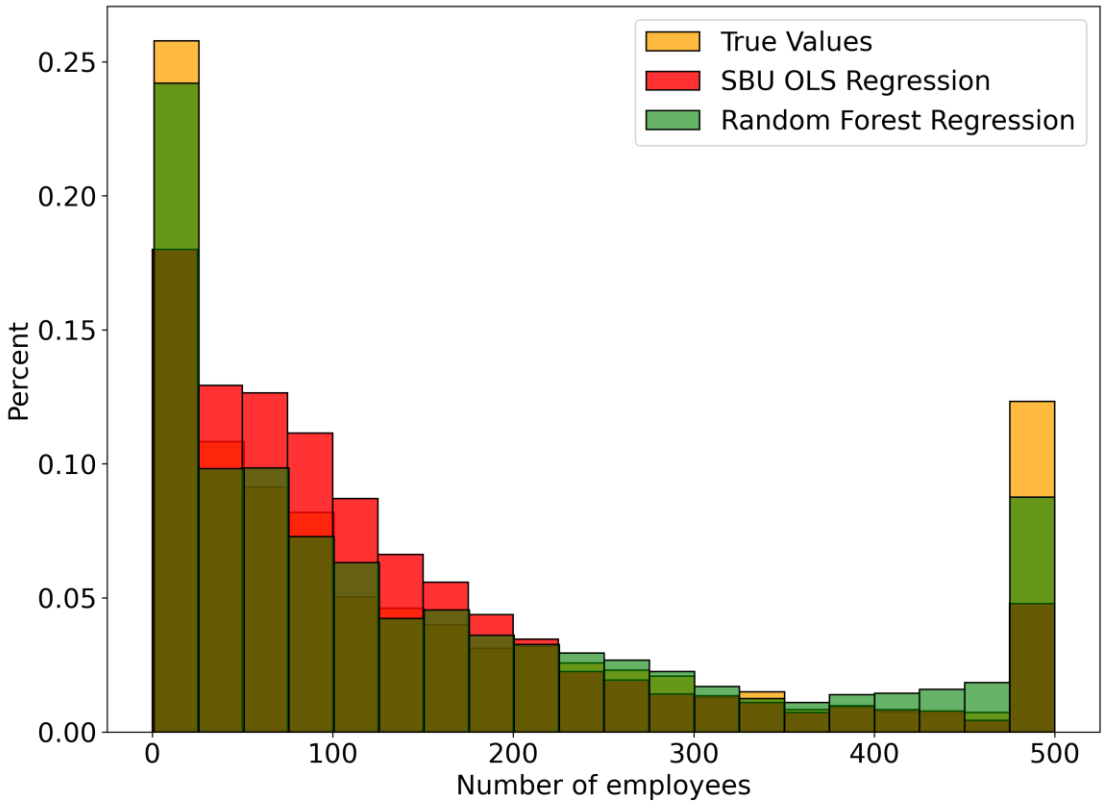
# Machine learning models enhance the imputation of missing data compared to the benchmark methods



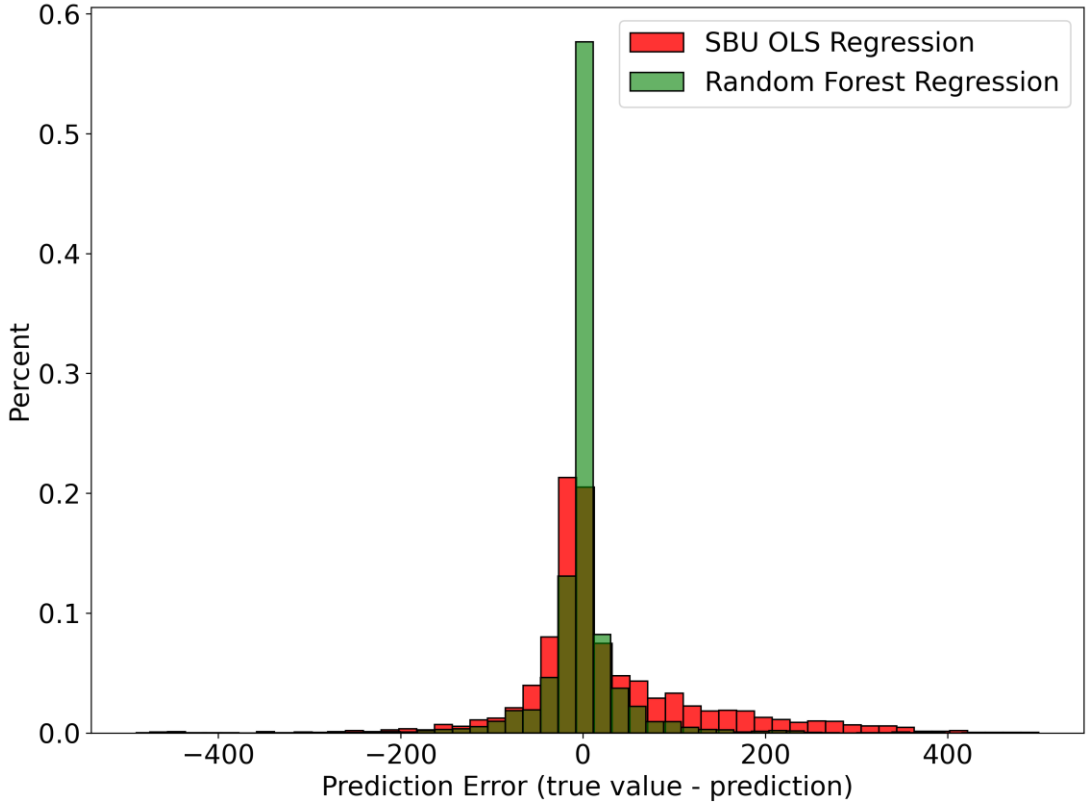

- Differences in model performance are highly significant

# The random forest model is much better at forecasting observations at the tails of the distribution

### Distribution of actual data and predictions




### Distribution of prediction errors



# **Sales revenue and number of employes** *(historical data provided by the data provider)* **are the most important determinants of full employment**

| Contributions of individual features to model predictions | Importance score |
|---|---|
| Sales revenue | 52.7% |
| Number of employees (from data provider) | 21.0% |
| State | 12.9% |
| Major sector | 3.2% |
| Year of response | 2.0% |
| Duration of response | 1.9% |
| Survey month | 1.1% |

# Summary

- Based on the conducted exercise ML models are significantly better at predicting missing employment data than the existing methods.
- Random Forest and Neural Networks were the two best performing ML models.
- Going forward, as we get more data and add more features to the dataset, these models will become even better at making predictions
- Missing data is a common problem in business surveys, affecting the accuracy and reliability of survey outcomes, and ML models present a viable solution to that problem.