PURDUE UNIVERSITY | Krannert School of Management

Feature Engineering for Sparse Demand Prediction

Hsiao Yu Hsu, Robyn Campbell, Stefanie Walsh, Zinnia Arshad, Matthew A. Lanham

Purdue University, Krannert School of Management

hsu230@purdue.edu; campb170@purdue.edu; walsh88@purdue.edu; zarshad@purdue.edu; lanhamm@purdue.edu

ABSTRACT

This study provides feature engineering recommendations for analytics practitioners on how to improve demand forecasts for sparsely demanded specialized products. There is no large-scale study to date that provides an in-depth empirical investigation of feature engineering approaches and their associated predictive gains when trying to predict sparse demand–which is one of the most challenging prediction problem classes one can encounter in practice. In collaboration with a large national auto parts retailer, we develop predictive models of demand for 47k+ products where 26k of them have less than five units sold in a year In our study is we will run thousands of feature engineering experiments to identify where we see cross-validated predictive gains for a set of common predictive modeling algorithms.

INTRODUCTION

- Feature Engineering can significantly decrease data cleaning time as well as allow Data Scientists to engage with more appealing parts of their job, like building models.

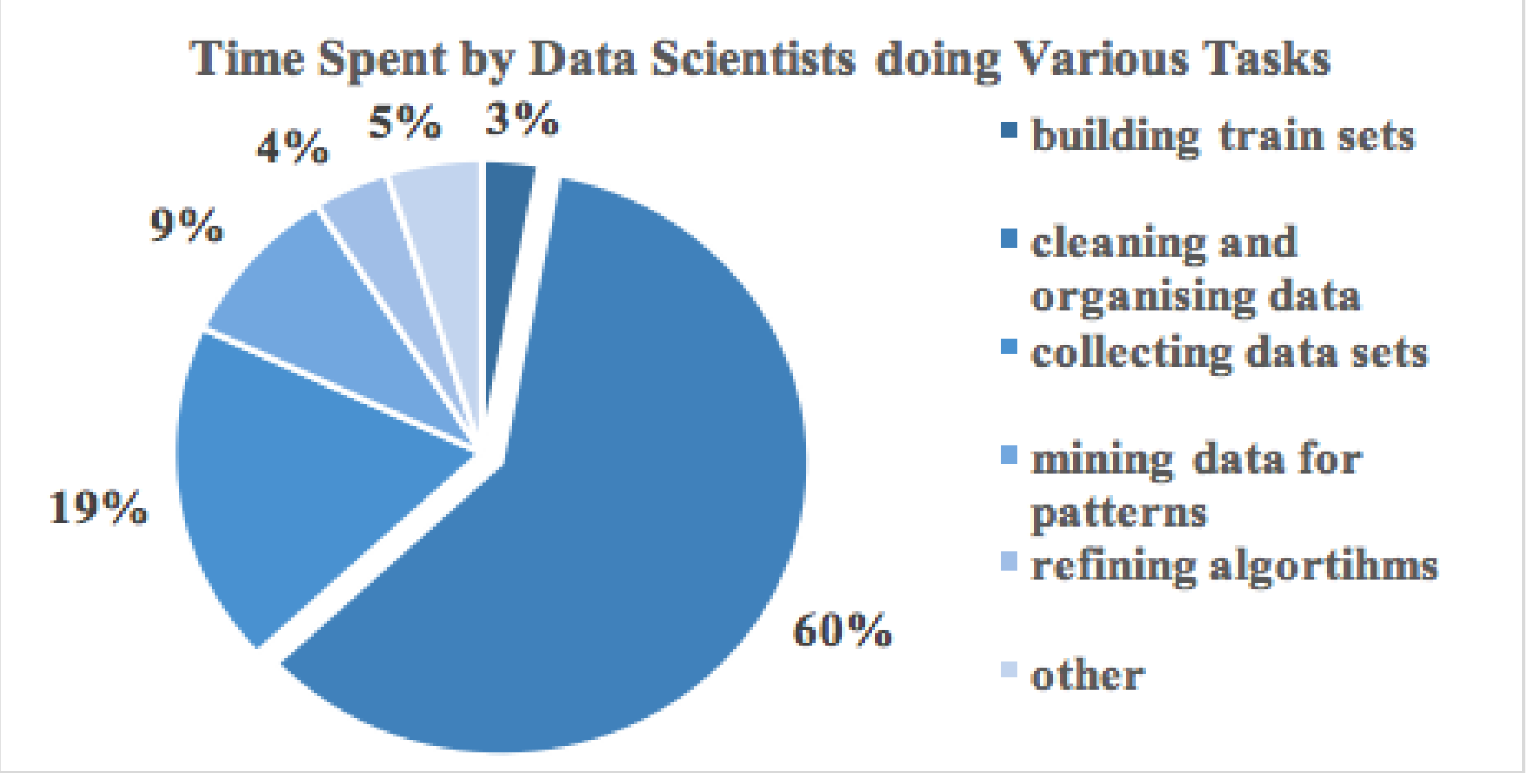


Fig 1. Time Spent on Each Data Science Activity

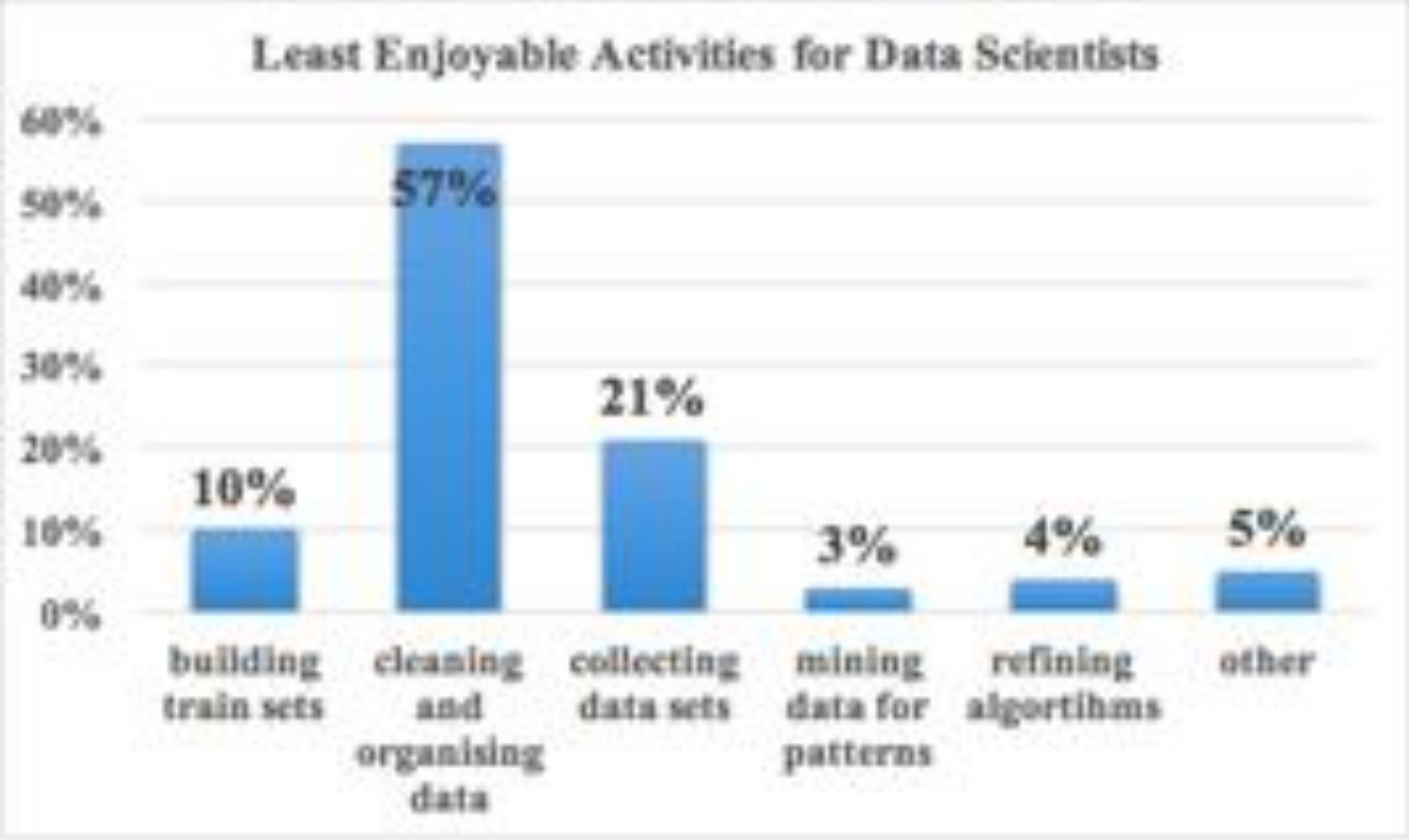


Fig 2. Least Enjoyable Data Science Activities Rated By Data Scientists

RESEARCH QUESTIONS

- Which feature engineering techniques work the best?
- How does the impact of these techniques change with type of model?
- How does the impact of feature engineering change with data granularity/amount of data/sparse vs. continuous data?

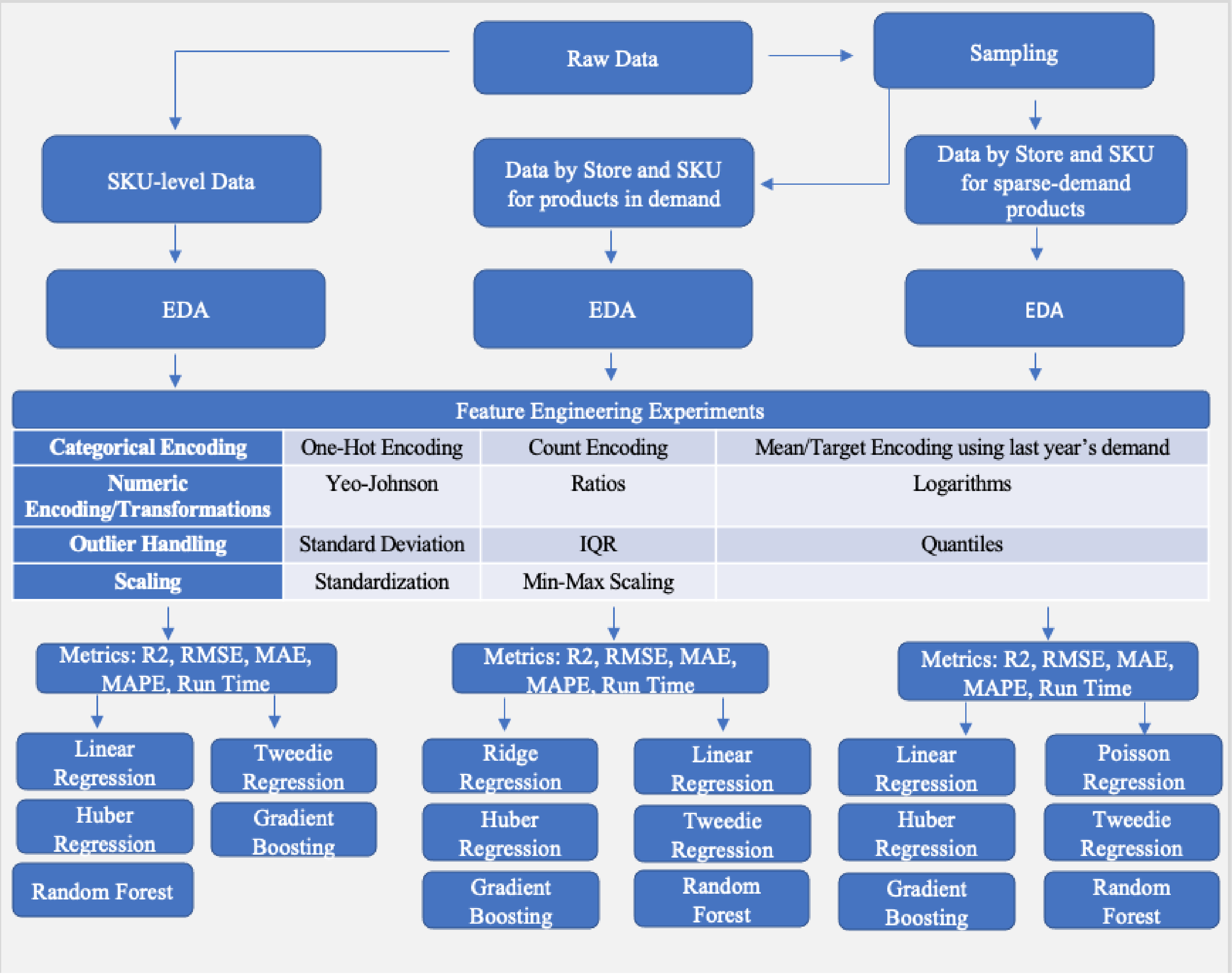
ACKNOWLEDGEMENTS

We would like to thank the Data Science Team at the national auto-parts retailer for guidance and support on this project.

CONCLUSIONS

- One-hot encoding consistently took longer to run and process than count encoding or mean encoding for equal results
- Running more complex models like gradient boosting or random forest significantly increased run time
- Regression models are significantly more likely to have significant ratios than gradient boosting or random forest models. Thus, making it best numeric method to increase Linear Regression accuracy
- Scaling and outlier removal did not impact SKU-Level data
- When using alternative regression models, utilize Z-score scaling to boost R2

METHODOLOGY


Raw Data → Sampling

SKU-level Data → EDA

Data by Store and SKU for products in demand → EDA

Data by Store and SKU for sparse-demand products → EDA

Feature Engineering Experiments			
Categorical Encoding	One-Hot Encoding	Count Encoding	Mean/Target Encoding using last year's demand
Numeric Encoding/Transformations	Yeo-Johnson	Ratios	Logarithms
Outlier Handling	Standard Deviation	IQR	Quantiles
Scaling	Standardization	Min-Max Scaling	

Metrics: R2, RMSE, MAE, MAPE, Run Time → Linear Regression, Tweedie Regression, Huber Regression, Gradient Boosting, Random Forest

Metrics: R2, RMSE, MAE, MAPE, Run Time → Ridge Regression, Linear Regression, Huber Regression, Tweedie Regression, Gradient Boosting, Random Forest

Metrics: R2, RMSE, MAE, MAPE, Run Time → Linear Regression, Poisson Regression, Huber Regression, Tweedie Regression, Gradient Boosting, Random Forest

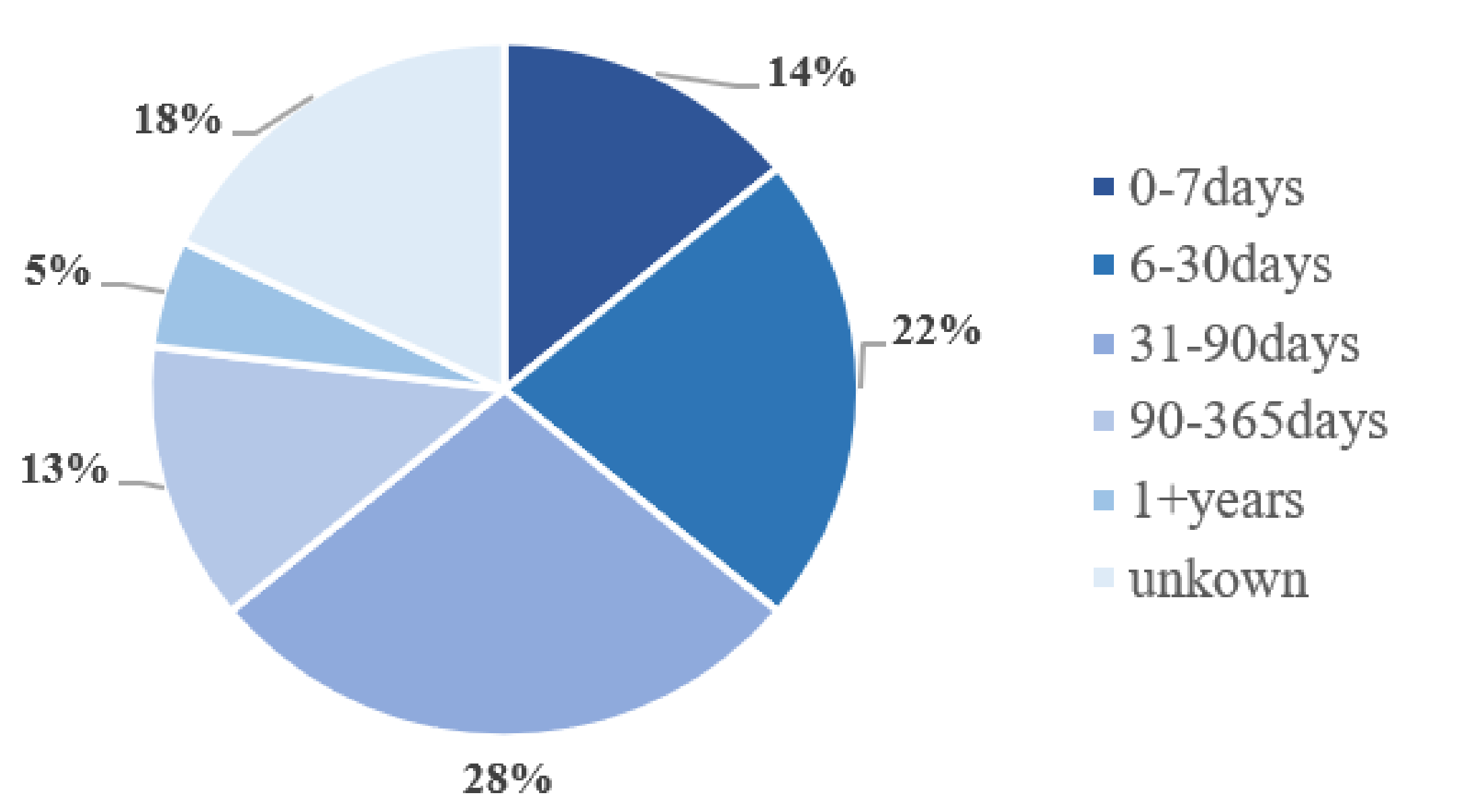
Fig 3. Study Design

EXPECTED BUSINESS IMPACT

From our analysis and experiments, there are serveral benefits and business impacts that we can contribute to our client

- Decrease the running time for demand forecast using feature engineering algorithm by 20% which identify the cost-saving opportunity
- Develop dynamic experiments to make modeling more flexibly on different circumstances

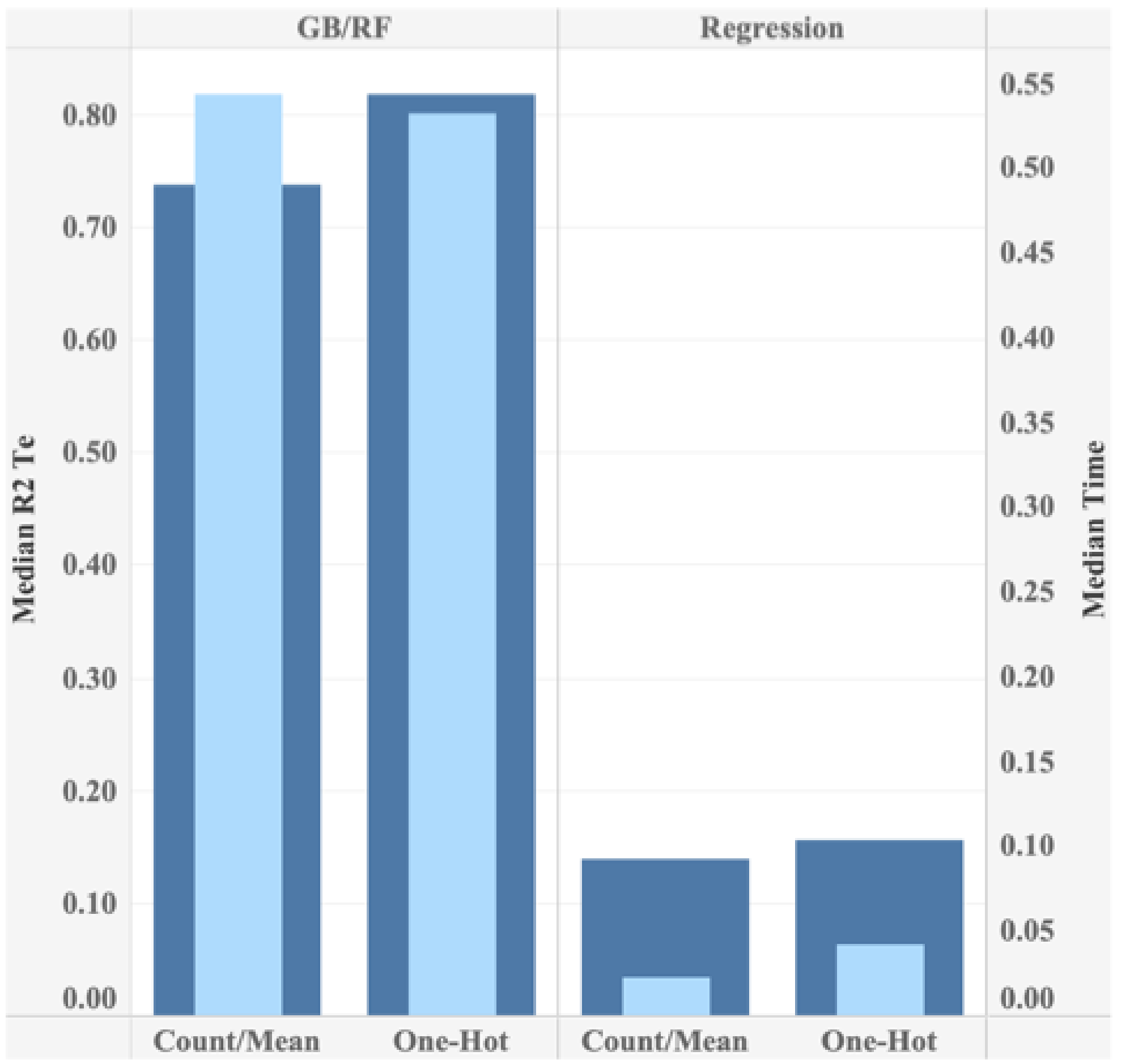
Deployment Time for Machine Learning Model



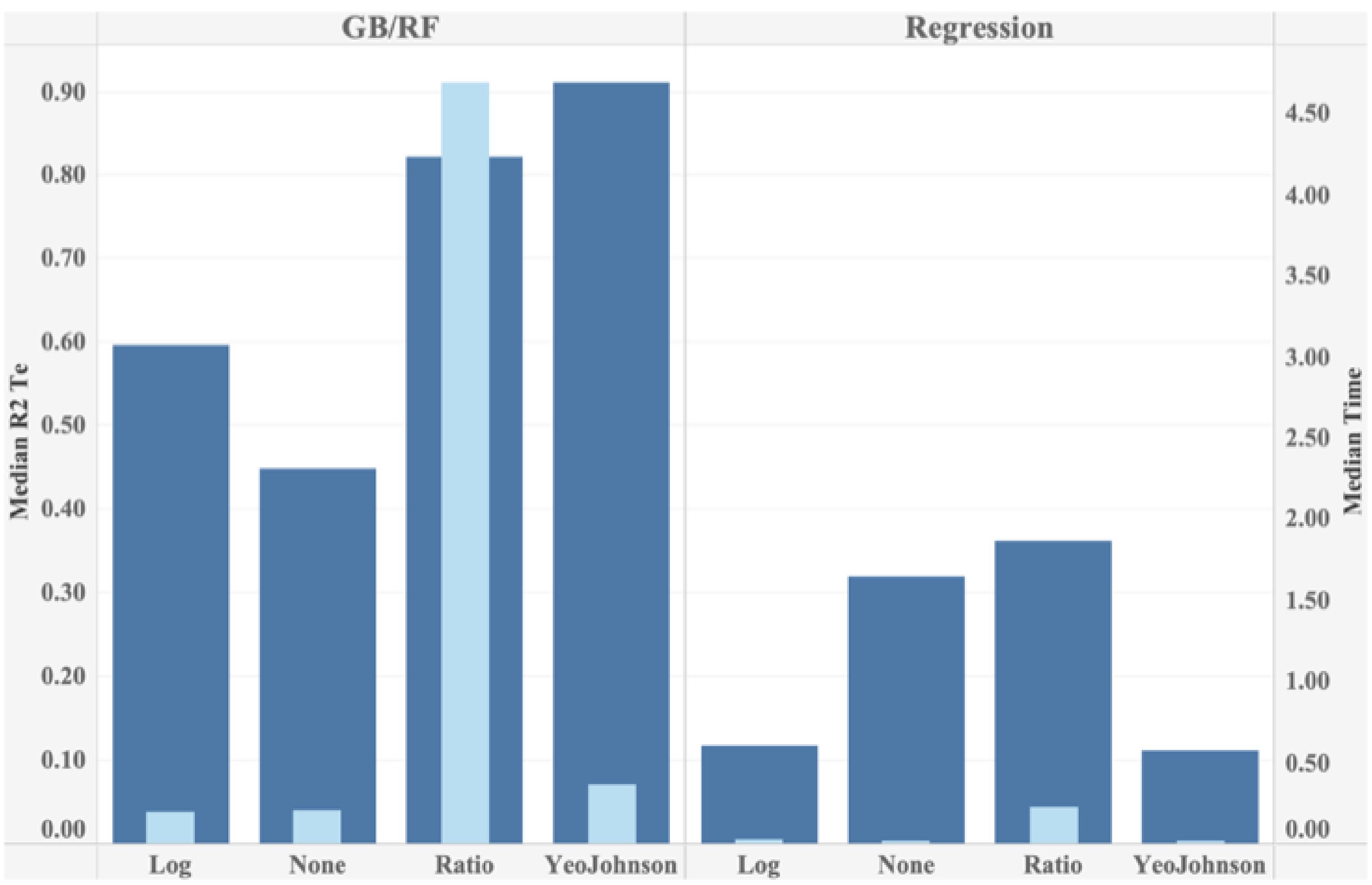
STATISTICAL RESULTS

Median R2 Te
Median Time

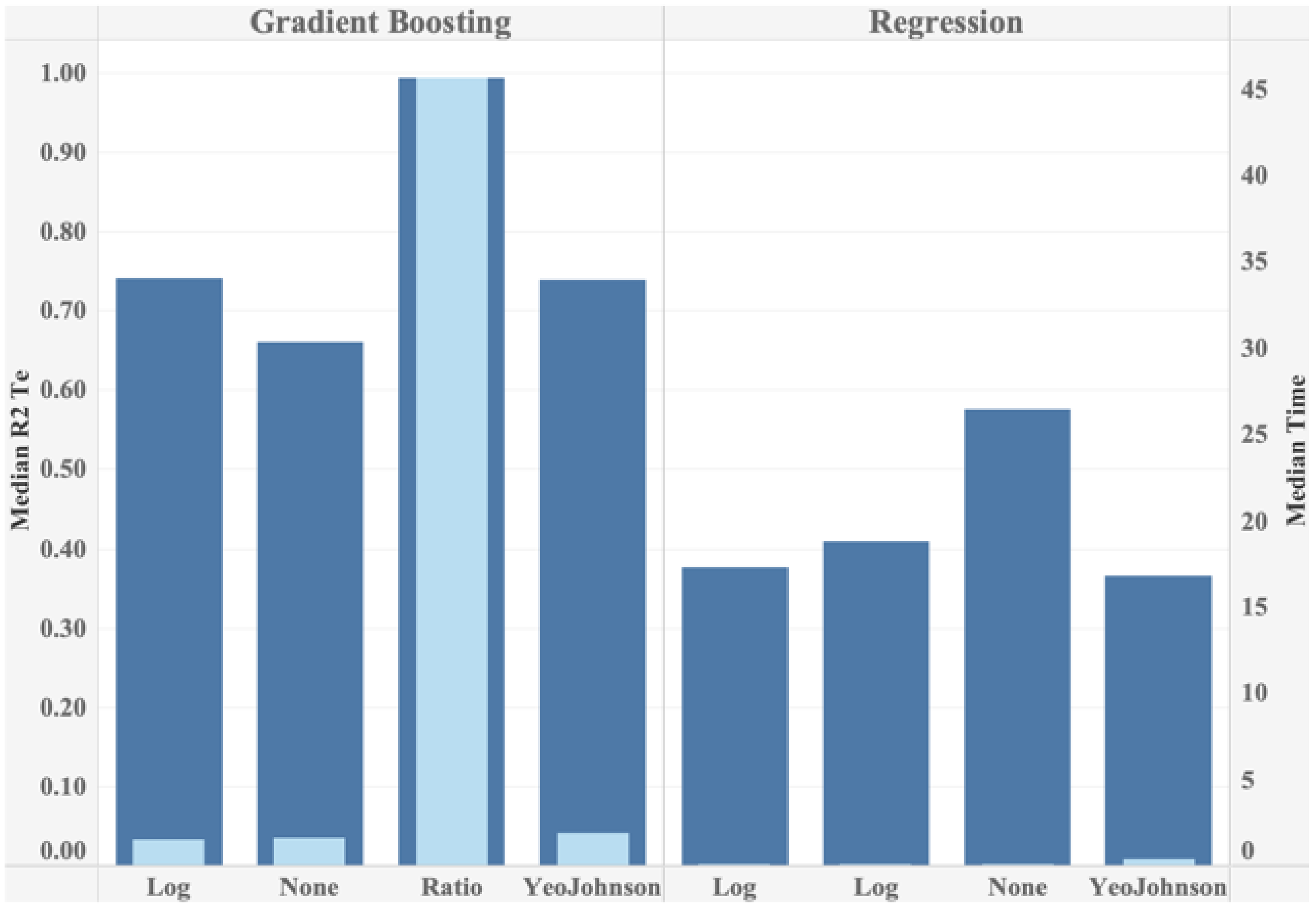
Categorical Encoding Results for National Level Data By Model



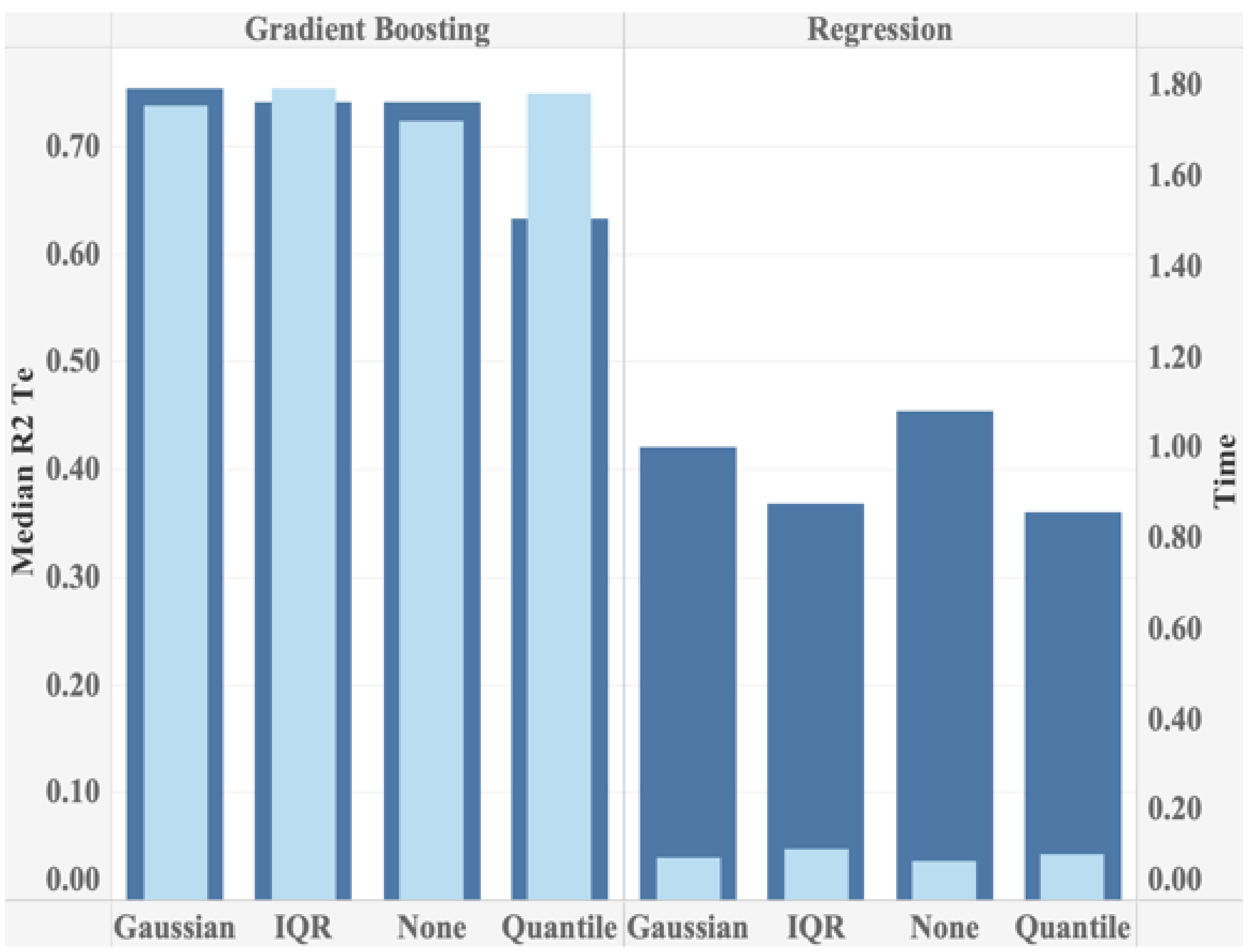
Numeric Transformation Results for National Level Data By Model



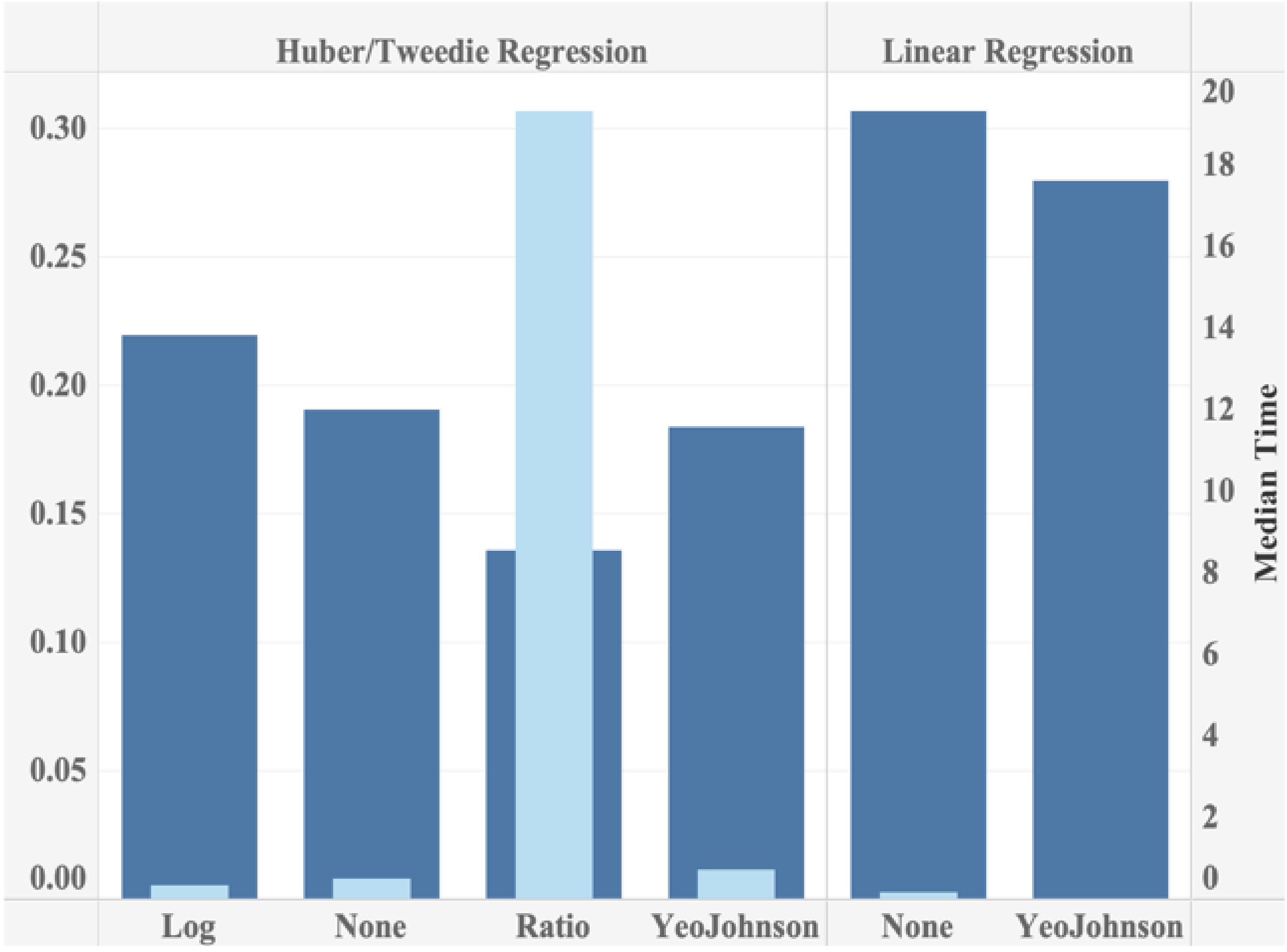
Numeric Transformation Results for Store-SKU Level Continuous Demand Data By Model



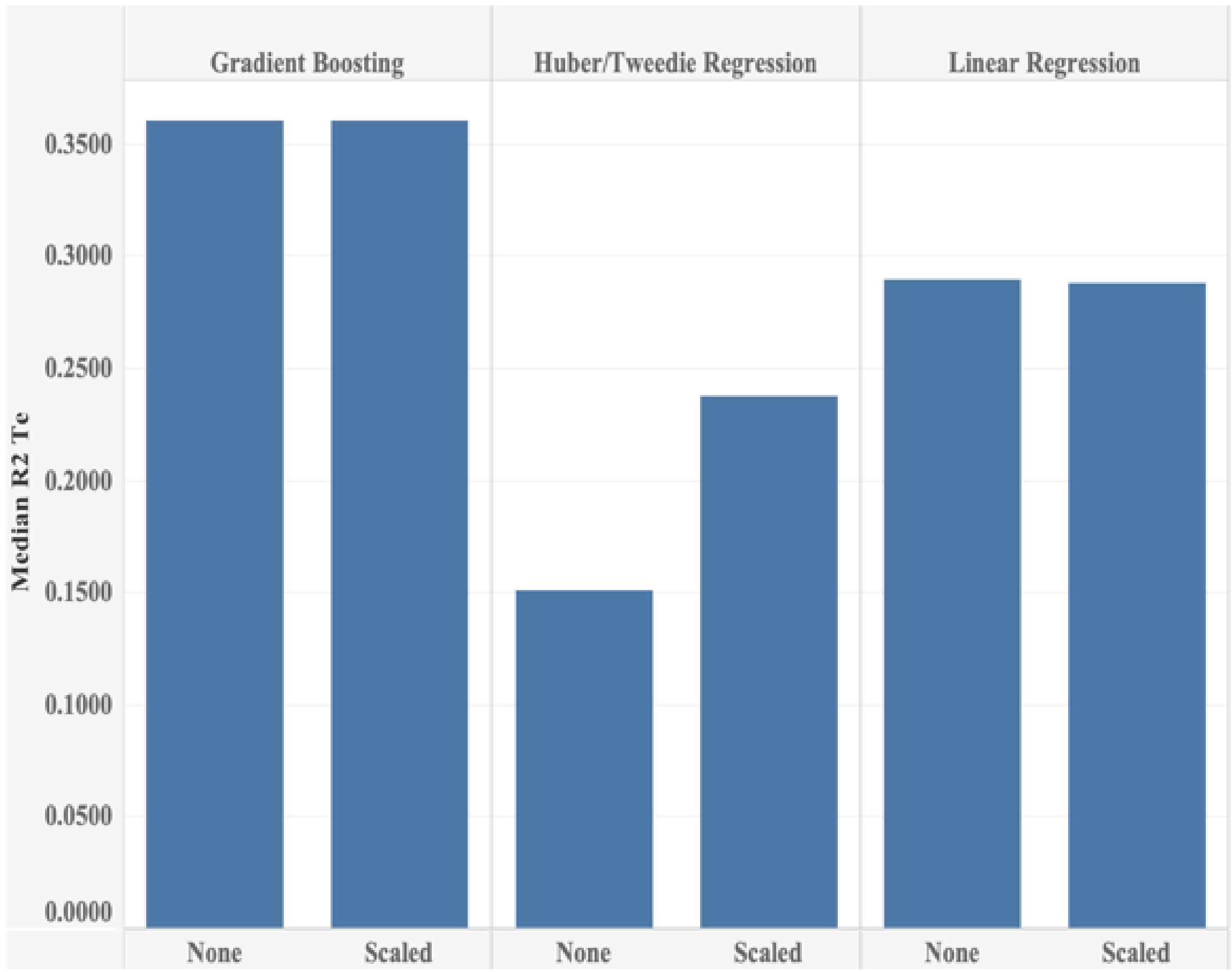
Outlier Removal Results for Store-SKU Level Continuous Demand Data By Model



Numeric Transformation Results By Model for Store-SKU Level Data With Sparse Demand



Outlier Removal and Scaling Results By Model for Store-SKU Level Data With Sparse Demand


Gradient Boosting
Huber/Tweedie Regression
Linear Regression
Median R2 Te
None
Scaled
None
Scaled
None
Scaled