ML Algorithm Synthesizing Domain Knowledge for Fungal Spores Concentration Prediction

Md Asif Bin Syed[1], Azmine Toushik Wasi[2] & Imtiaz Ahmed[1]
[1]West Virginia University, [2]Shahjalal University of Science and Technology

Problem Formulation

Strict Quality Control Measures

The pulp-and-paper manufacturing sector requires strict quality control measures to ensure the end product is free from contaminants that can affect its usability for various applications.

Fungal Spore Concentration

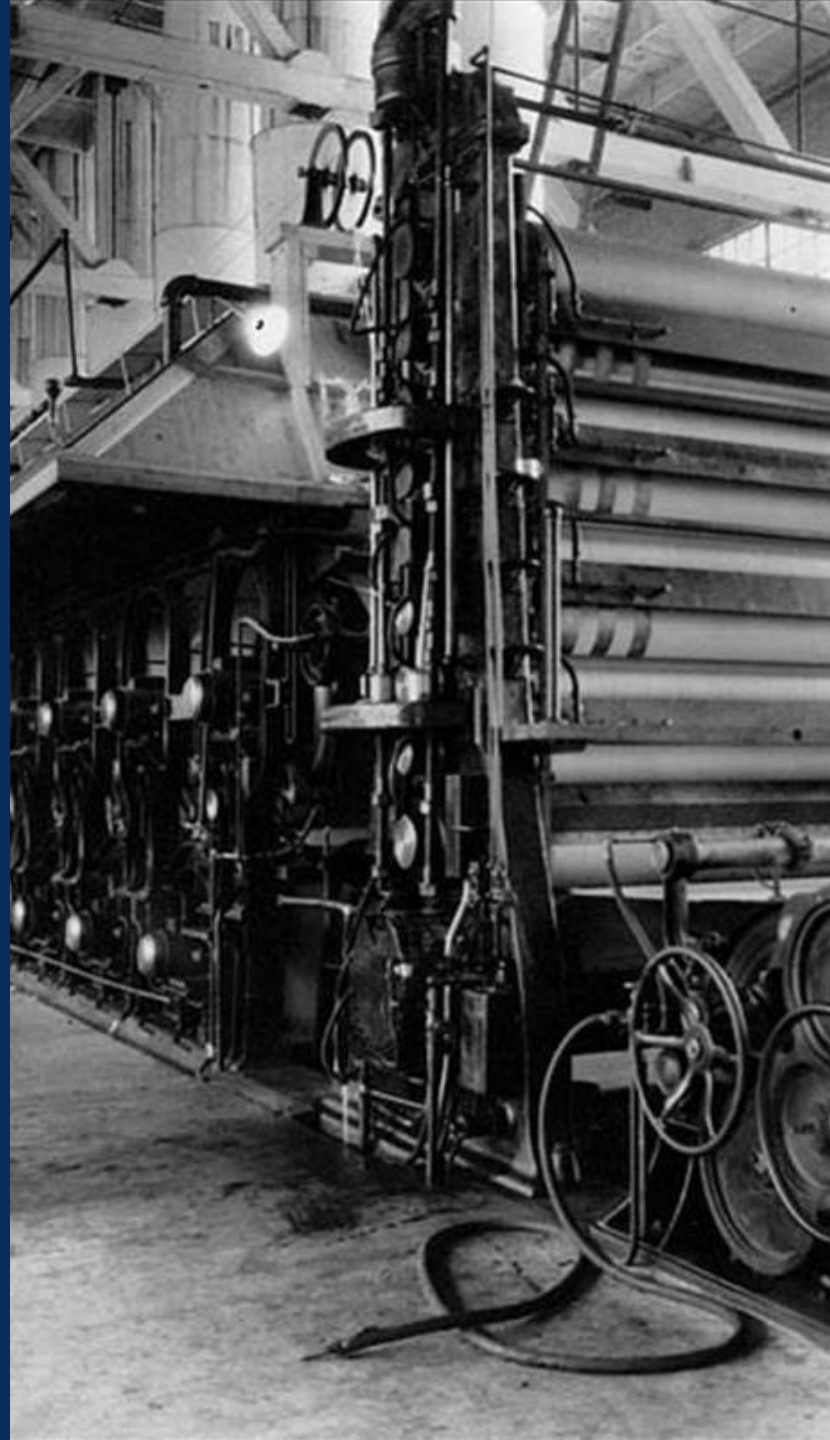
Fungal spore concentration is a critical quality metric in the industry and excessive amounts can severely affect the paper's usability.

Time-consuming Lab Tests

Current techniques for assessing fungal spore concentration involve time-consuming lab tests, delaying real-time control strategies.

Need For Precise Real-time Predictions

There is a need for precise real-time predictions of fungal spore concentration to maintain exceptional quality standards and enable effective real-time control strategies.



Timestamp	y_var	grade	x_var_1	x_var_2	x_var_3	x_var_4	...	x_var_113	...	x_var_119
2021-01-30 06:23:06	10	grade_1	34.30884	9.839399	64.52522	40.16026	...	42.46489	...	-
2021-01-30 10:09:38	25	grade_1	33.79903	9.861699	67.17016	32.47632	...	41.98848	...	-
2021-01-30 14:35:20	0	grade_1	33.68362	10.07202	66.39706	25.32029	...	45.81967	...	-
2021-01-30 19:46:07	5	grade_1	33.36954	9.795143	69.43247	27.25093	...	45.80038	...	-
2021-01-30 22:22:13	0	grade_1	34.64005	9.87202	68.75052	30.09163	...	45.80373	...	-
2021-01-31 06:12:39	0	grade_1	33.75125	9.645829	68.45326	25.97441	...	45.84367	...	-
2021-01-31 14:00:12	0	grade_1	33.3864	7.981672	69.65371	18.26637	...	45.83347	...	-
2021-01-31 18:20:22	5	grade_1	34.4358	9.841844	69.0898	29.73899	...	45.98977	...	-
2021-01-31 22:39:04	10	grade_1	33.20068	10.14254	69.5536	29.16336	...	45.86425	...	-
2021-02-01 02:12:04	0	grade_3	36.38562	9.908736	68.63308	30.34933	...	53.04411	...	-
...	...	...	...	...	...	...	...	...	...	...

Challenges

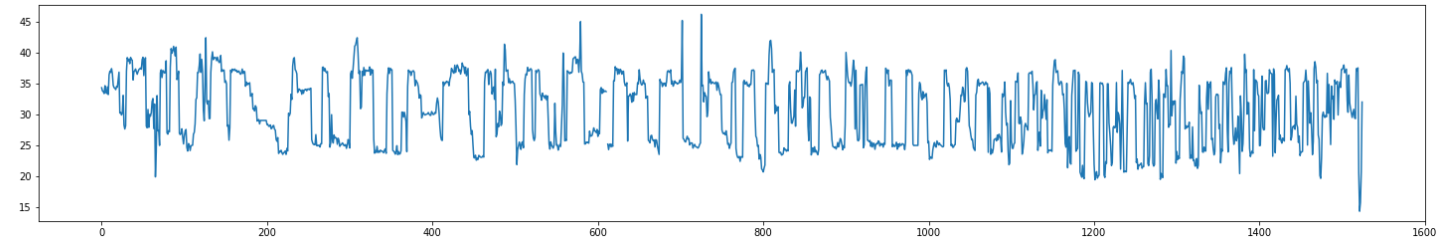
1. Dataset

We assessed the presence of missing values, identifying several variables with missing data. Feature selection with any methods were not working properly.

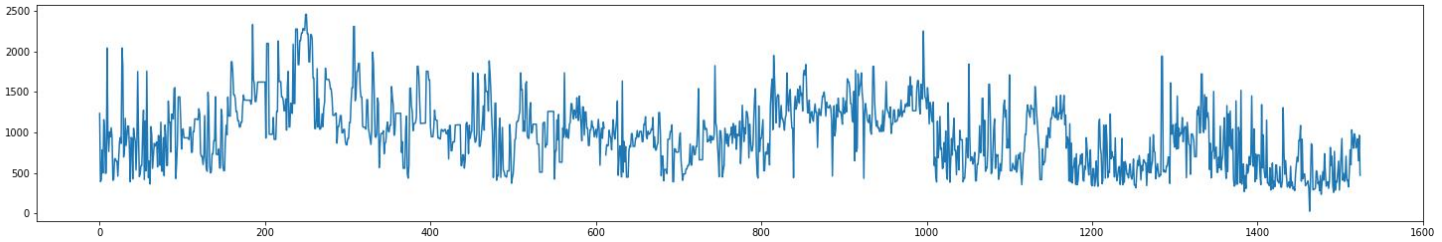
Challenges

Not Time Series

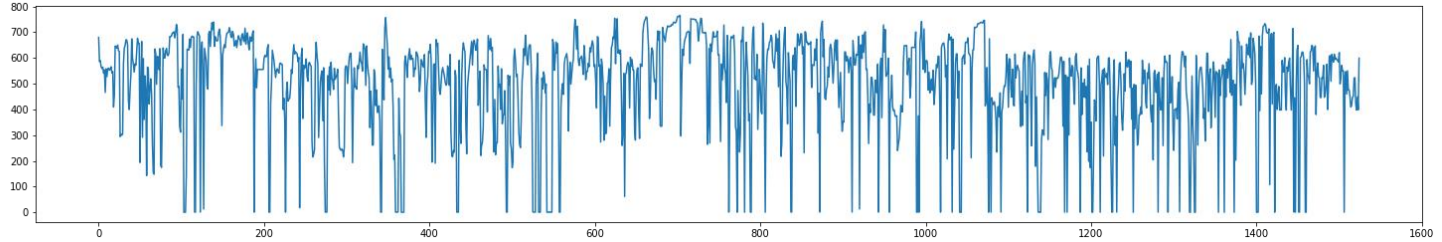
We investigated the potential dependency on time. By observing the provided figure, we attempted to discern any time dependency among the features.



x_var_1



x_var_70

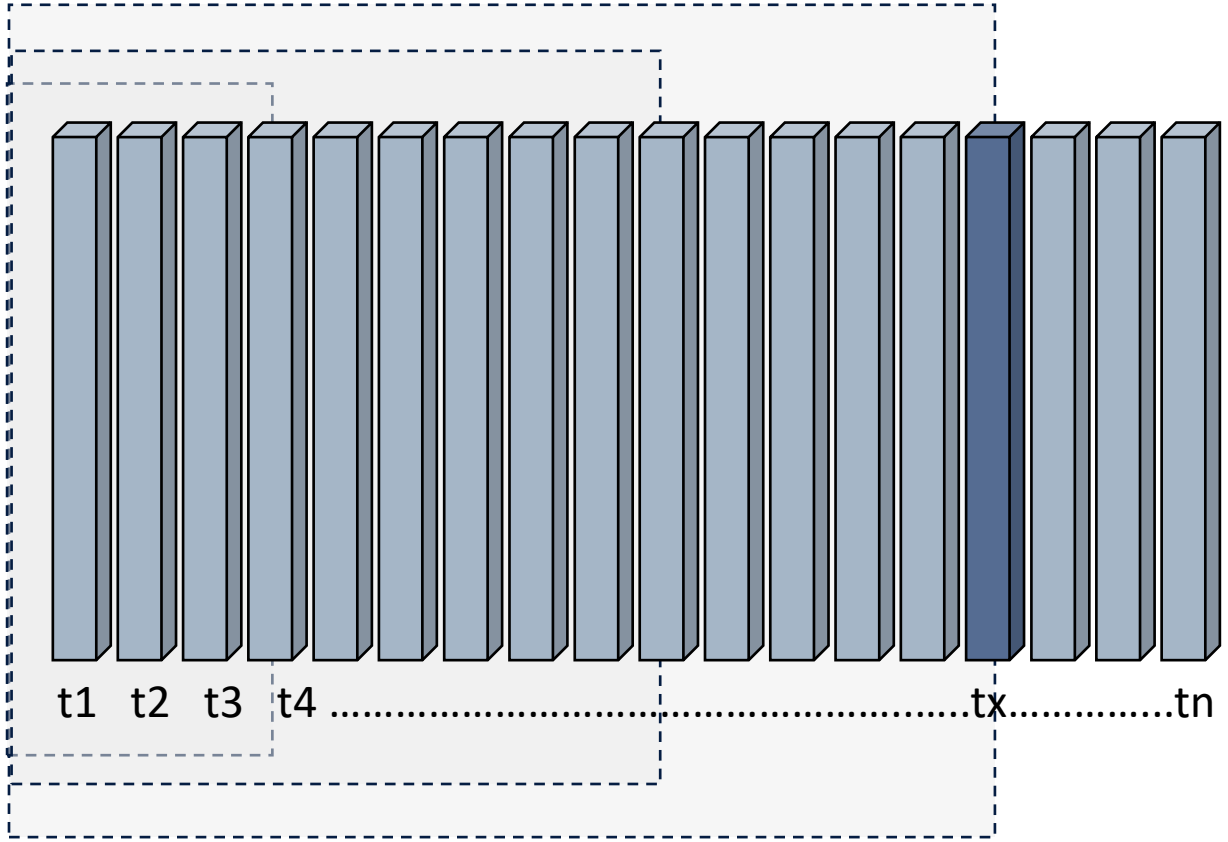


x_var_109

Challenges

Training Limitation

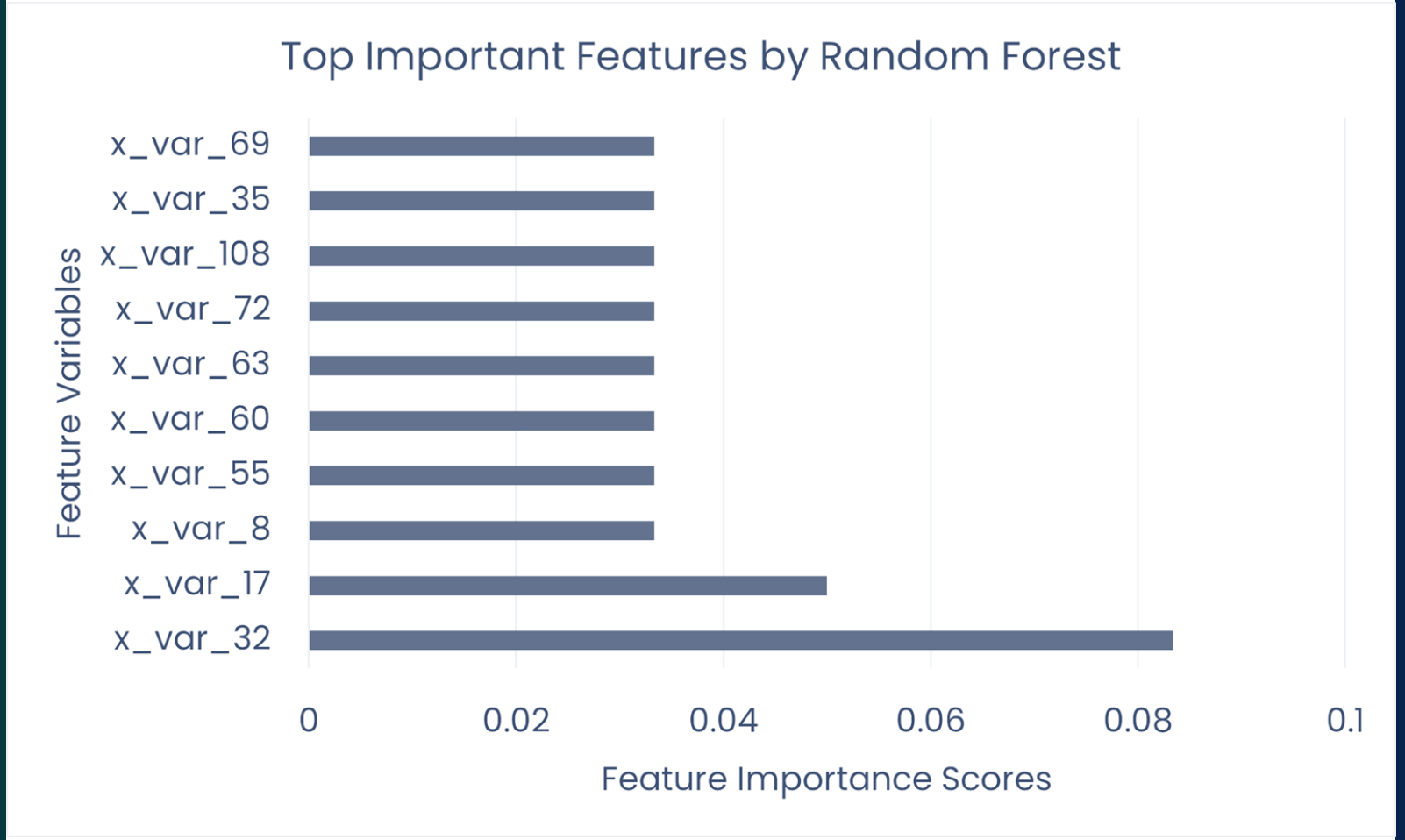
To predict the target value for a timestamp tx in the test set, we are not permitted to use trainset data with timestamp > tx.



Challenges

Feature Importance

Feature Importance was calculated by various methods but not found useful.



First Principle Thinking

Analyzing

Primary Predictions

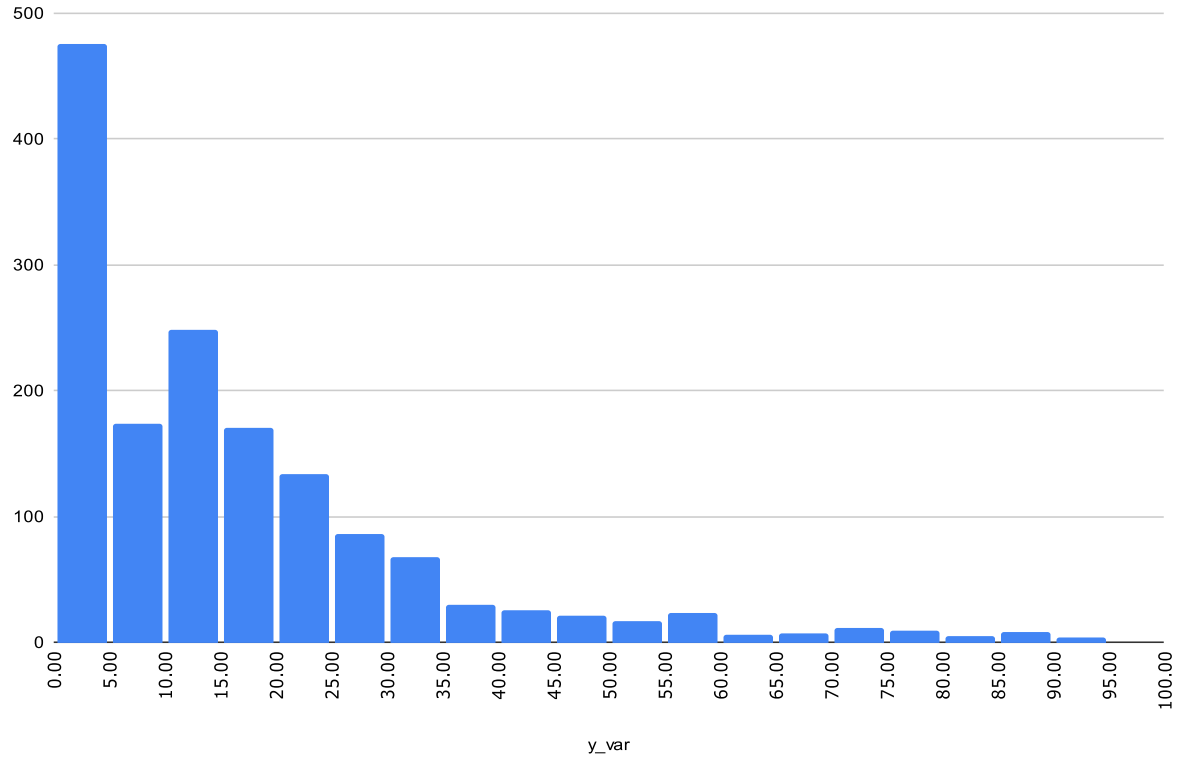
- ✓ Primary Predictions were unstable
- ✓ The values were problematic and can easily mislead evaluation metrics.

Timestamp	True Value	Primary Prediction
19-01-22 10:17:31	10	-5517182137000.93
08-11-22 2:32	0	-2.98
24-10-21 9:28	0	-2.11
05-12-21 11:28	0	-2.08
30-03-22 18:45	0	-1.40
17-02-22 7:46	0	-1.23
08-03-22 1:23	0	-1.19
22-12-21 10:56	0	-1.00
11-03-22 15:42	0	-0.92
07-11-21 1:32	0	-0.89

First Principle Thinking

Domain Expertise

- ✓ Target variable seems to be an integer which is divisible by 5.
- ✓ Our initial predictions are a lot close to multiples of 5, like we have 11, 11.5, 10.5 more than 12-12.5 which is a more uncertain stage.



Model

Ridge Regression

We found that Ridge Regression performed exceptionally well on our sensor dataset with a large number of input columns. This is likely because Ridge Regression's regularization technique helps to reduce overfitting and handle multicollinearity, leading to stable coefficient estimates and better predictive performance on new data.

01 Reduces Overfitting

Ridge Regression is a regularization technique that helps to reduce overfitting in ML models. It adds a penalty term to the cost function which reduces the magnitude of the coefficients, thus limiting the model's complexity and making it less prone to overfitting.

02 Handles Multicollinearity

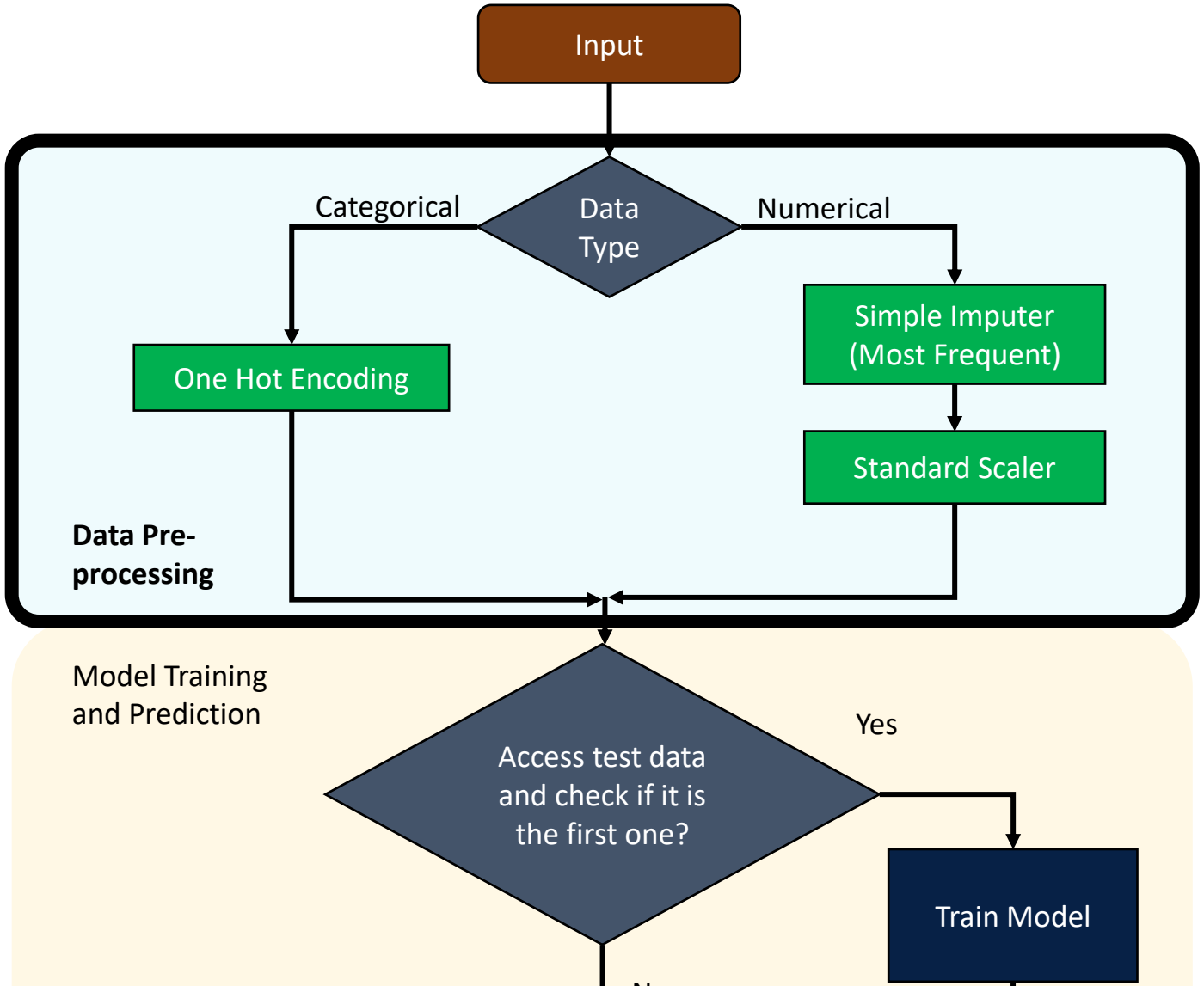
Ridge Regression performs well when there is multicollinearity between the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated with each other, which can lead to unstable and unreliable coefficient estimates in traditional linear regression. The penalty term in Ridge Regression helps to stabilize these estimates by reducing their variance.

03 Data Inconsistency Handling

Ridge Regression can provide a solution even when the data is inconsistent or when the number of samples is smaller than the number of features. This is because it introduces bias into the estimates, which can help to overcome problems caused by inconsistencies in the data.

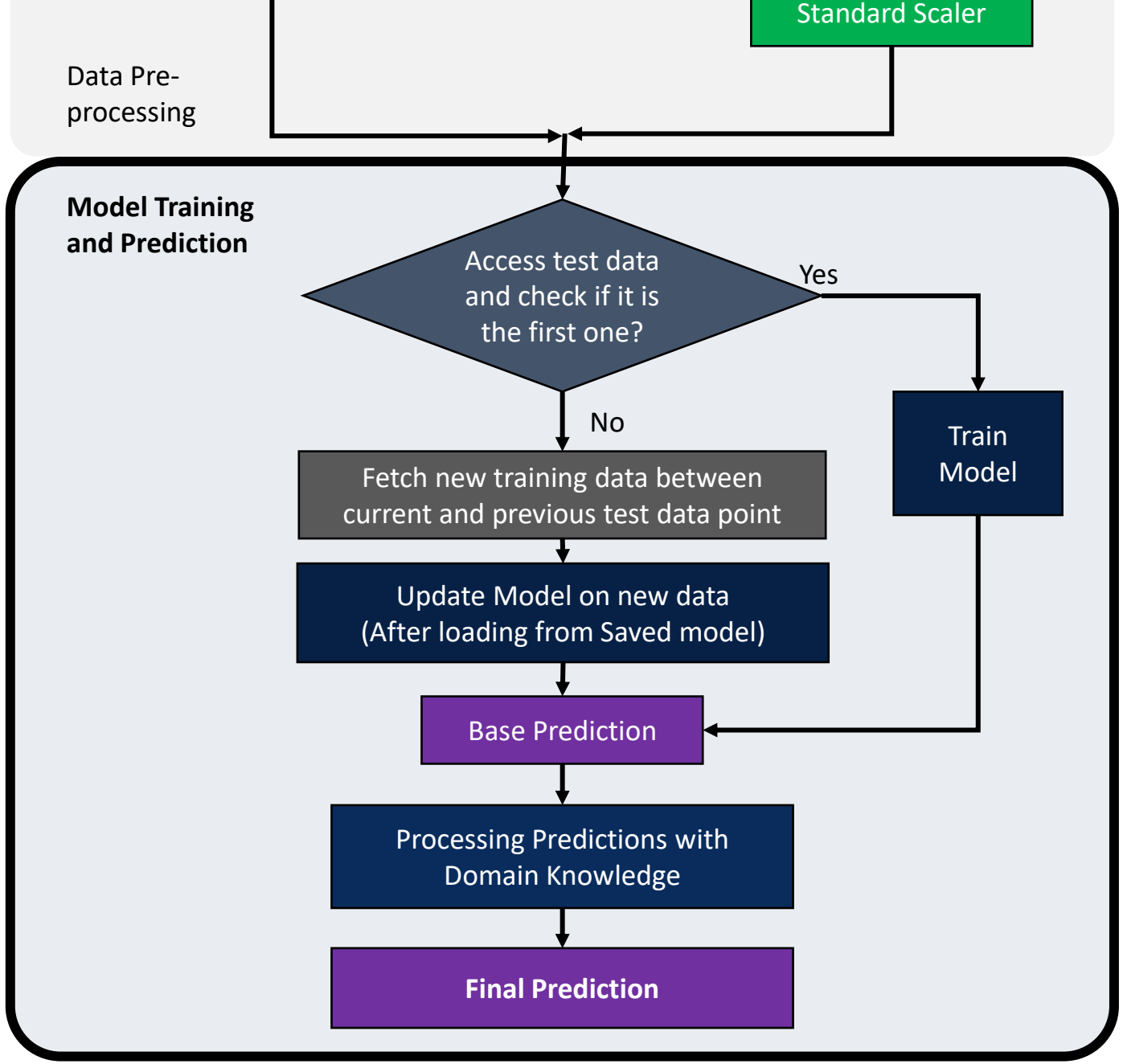
Model Architecture

Data Preprocessing



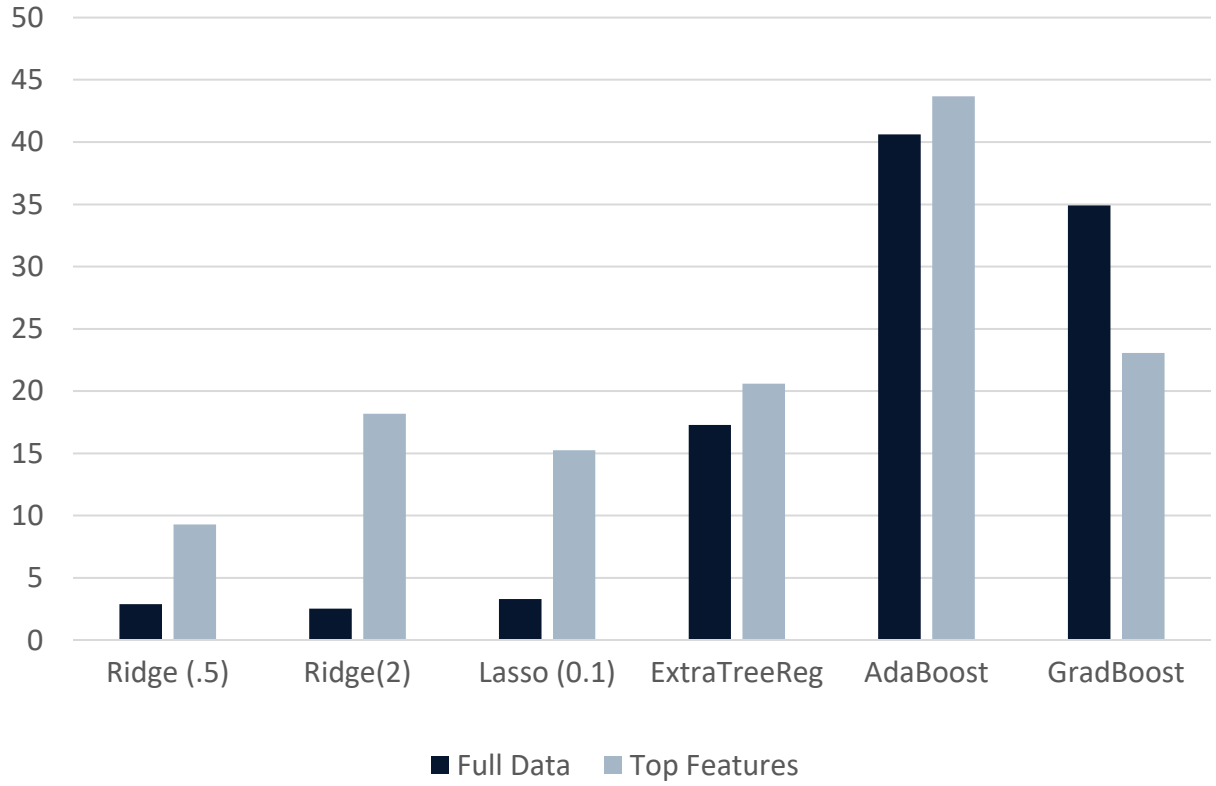
Model Architecture

Model Training

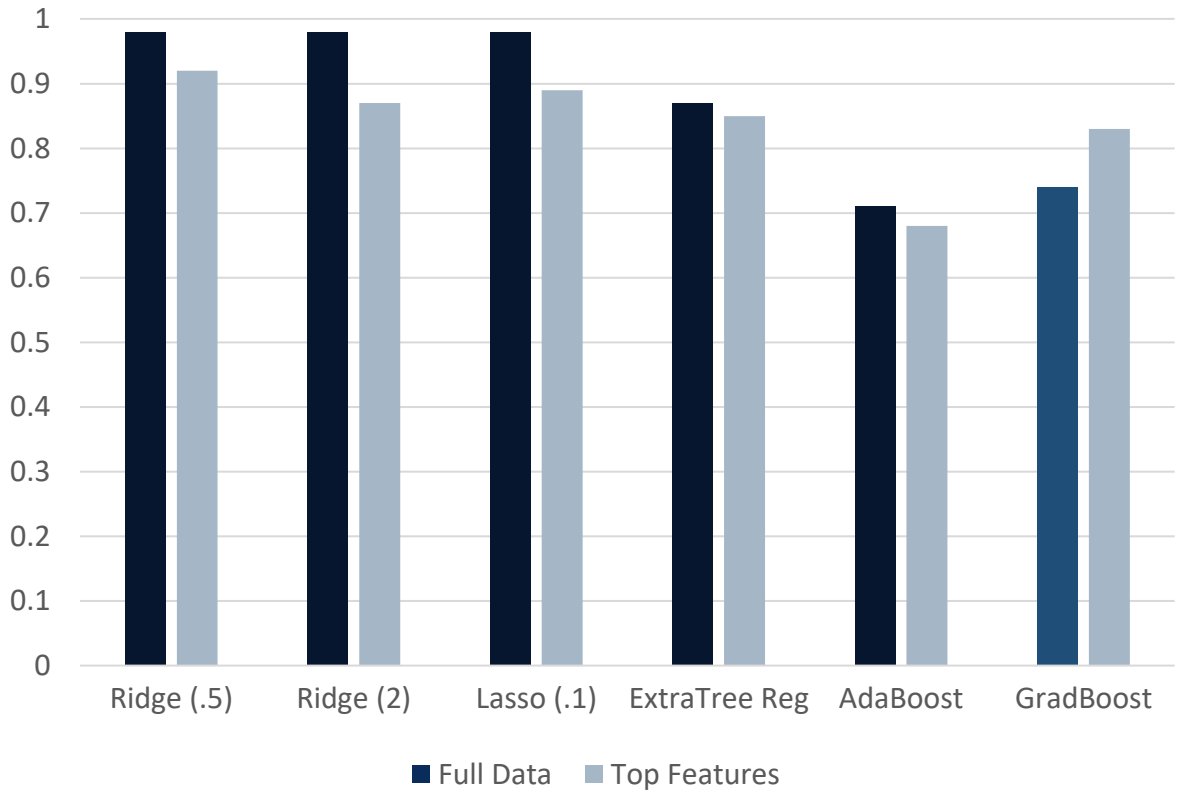


Scores

Mean Squared Error



R-square Score



Benefit

Fast, Efficient, Reliable

Ridge is a simple linear regression technique that does not require complicated neural network architectures or large amounts of training data. This simplicity leads to less complexity in terms of the model architecture and optimizations required during training, making it computationally efficient.

01 Low Memory Requirements

Our model requires much less memory than deep learning models because it does not have as many parameters to train. This low memory requirement makes it easier to integrate into embedded systems with limited memory resources.

02 Fast Training Time

Our model has a closed-form solution which results in a computationally efficient training process. This means that the training time for Ridge Regression is faster than deep learning techniques like recurrent neural networks (RNNs) or convolutional neural networks (CNNs).

03 Easy Deployment

Our model lightweight and easy to deploy on embedded systems because they do not require large processing power. This makes them ideal for implementing in resource-constrained devices such as embedded systems or Internet of Things (IoT) devices and sensors in the industry.

Thank You