This report outlines the methodology and outcomes of predicting unknown system properties and addressing the key challenges of working with continuous data in stock value prediction.

Key Steps and Techniques Used:

Data Preprocessing: Separate missing data, scale feature values to remove bias, and identify and remove outliers using the 3 IQR criteria.

Feature Selection and Reduction: Employed pair-wise correlation and Principal Component Analysis (PCA) to reduce dimensionality and identify influential features.

Model Selection and Justification:

For Property 1, models like Multiple Linear Regression, Ridge Regression, and Random Forest were chosen for their robustness to outliers and complex relationships, with Random Forest displaying the lowest testing RMSE and highest R-squared value.

For Property 2, Lasso Regression, Gaussian Process Regression, and Random Forest were utilized, again with Random Forest performing best in terms of RMSE and R-squared.

For Property 3, Decision Tree, Logistic Regression, and Random Forest were employed, with Logistic Regression offering a balanced performance without overfitting.

Model Evaluation:

Accuracy, sensitivity, and specificity metrics were calculated, demonstrating the strengths and limitations of each model.

The Random Forest model exhibited high accuracy and robustness, suggesting its suitability for predicting continuous property values.

Predictions and Confidence:

The report discusses confidence in the predictions based on the comprehensive techniques used, which allowed for the assessment of both linear and non-linear relationships.

It is suggested that increased data availability could enhance model robustness and training diversity, thus improving confidence in predictions.

Handling Data Imbalance: The report mentions the use of techniques like SMOTE for balancing datasets when necessary.

Uncertainty Estimation: Strategies such as calculating confidence intervals, performing cross-validation, and leveraging model-specific uncertainty measures are recommended for assessing prediction uncertainty.

Solutions to Problem Statement Aspects:

Determined that none of the first 10 systems met the criteria for all three properties simultaneously, indicating potential data insufficiency or noise within these systems.

Justified model selections based on their mathematical underpinnings and relevance to the dataset's challenges.

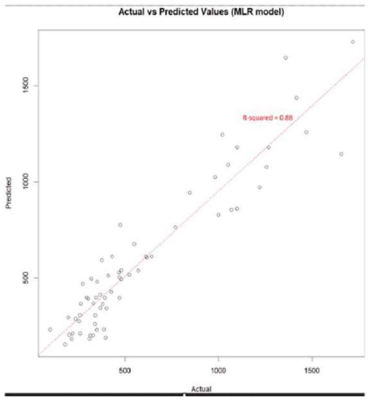
Addressed potential data imbalance and provided a framework for predicting missing continuous property values.

In conclusion, the report effectively addresses the problem statement by applying rigorous data science methodologies to predict system properties and stock values, while also considering the uncertainty and generalizability of the models used.

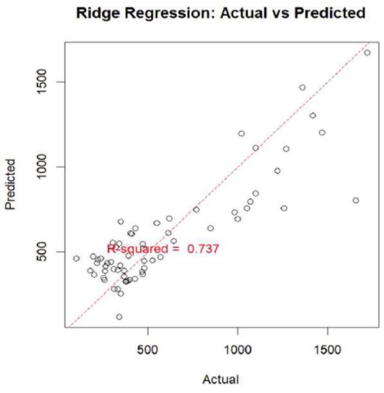
Plot of Actual vs Predicted value.

For Property 01:

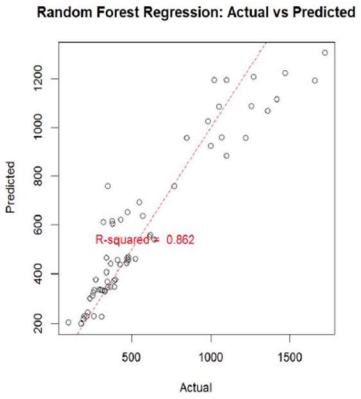
MLR



Ridge regression

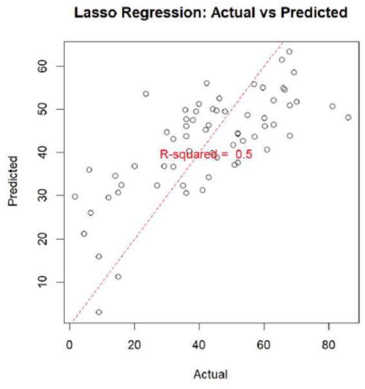


Random Forest

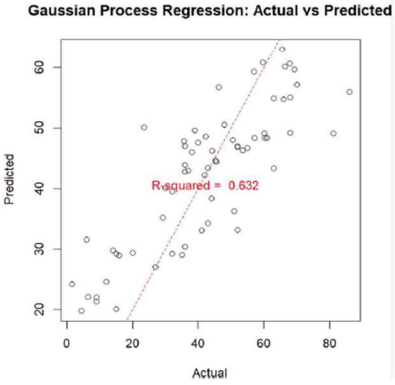


For Property 02:

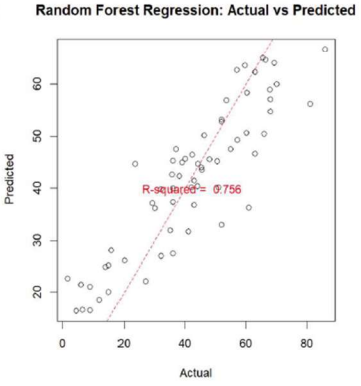
Lasso regression

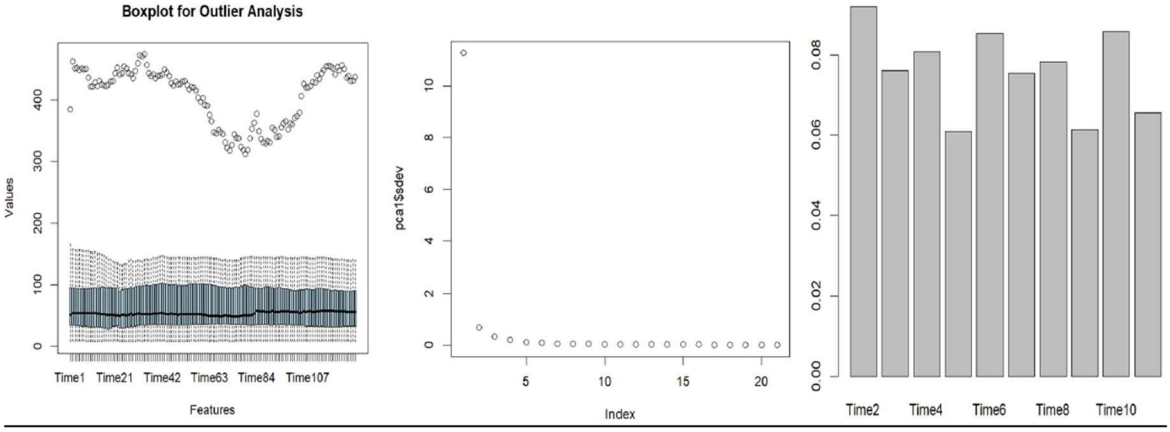


GPR



Random Forest





Model	Train RMSE	Test RMSE	R-squared
Lasso regression	18.17616	43.57231	0.4997178
PCR	15.16228	41.31031	0.6317625
Random forest	22.905548	48.48654	0.7557386

Predicted property	Lasso regression	PCR	Random forest
Company 1	41.4135	61.3048	51.69898
Company 2	61.4665	87.7221	98.4075
Company 3	55.7682	55.2096	66.0292
Company 4	66.9636	68.2637	74.8099