Deliu Maria, December 7, 2023

Management of Information, Department of System Engineering, Technical University of Moldova

**Forecasting Data Storage Size in Banking Systems:**

**A Comparative Analysis of Predictive Models**

**Abstract**

This study investigates the dynamics within a banking system by analysing a comprehensive dataset including system characteristics, operations, and storage behaviours. Using a sum of statistical techniques and machine learning models, the research aims to find correlations, patterns, and predictive insights within the system. The analysis includes a detailed exploration of the impact of variables such as system types, system stages, number of accounts, number of transactions per day and others on system size. Through meticulous data analysis using diverse regression models, including Multiple Linear Regression, Ridge Regression, Lasso Regression, Random Forest and others, this study explores the predictive capabilities of each model. The investigation also entails feature selection methods like Filtering, Wrapper, Tree-based, and Regularization techniques to ascertain the most influential features for accurate predictions. Additionally, the study investigates multivariate data analysis, showcasing the relationships between multiple variables using heatmaps and pair plots. Results indicate that Multiple Linear Regression, Ridge, and Lasso Regression models exhibit superior predictive performance. While Random Forest Regression demonstrates high accuracy metrics, its performance significantly falters when predicting future data. The comparative analysis of R2 scores highlights the efficiency of MLR, Ridge, and Lasso models, emphasizing their robustness and reliability.

This study underscores the importance of feature selection and model assessment in understanding banking system behaviour, offering insights valuable for decision-making in financial institutions.

## Introduction

In an era defined by escalating reliance on digital infrastructure, the banking sector stands at the forefront of technological evolution. The exponential growth of data within banking systems has prompted a necessity to understand the relationships between system attributes and operational outcomes [1]. This study delves into the complexities of banking system dynamics, seeking to find the deep relationships between various system parameters and the resulting impact on system sizes. The enormous volume and diversity of data generated within banking systems have set the need for advanced analytical methodologies to find valuable insights. The main problem lies in guessing how much data space will be needed in the future. This guessing game often leads to spending too much on storage and sometimes makes it hard to find what we need quickly. Also, the equipment we already have might not be used as well as it could be, which costs extra and might slow things down. This research is set to unravel this complex ecosystem through an extensive analysis of diverse system attributes and behaviours. Central to this research is an in-depth examination of variables such as backup size, system stages, system types, number of accounts and transactions, each playing a serious role in shaping system architectures.

The study adopts a step-by-step approach, using a spectrum of statistical analyses and machine learning models to discern underlying patterns, dependencies, and predictive capabilities within the banking system. By employing a suite of regression models including Multiple Linear Regression, Ridge, Lasso, Elastic-Net and others, this research aims to ascertain the predictive prowess of each model and evaluate them in forecasting system sizes.

Furthermore, the study employs rigorous feature selection techniques, exploring diverse methods such as Filtering, Wrapper, Tree-based, and Regularization methods to choose the most influential attributes that drive system sizes. The research also includes multivariate data analysis, utilizing visual tools like heatmaps and pair plots to unravel relationships between various system parameters. The culmination of this investigation not only sheds light on the inherent complexities of banking system behaviours but also presents pragmatic insights crucial for informed decision-making within the company. Understanding the relationships between system attributes and their impact on system sizes is imperative for optimizing operational efficiencies and leading to noticeable growth in modern banking infrastructures.

**Data and Methods**

The methodology adopted for forecasting data storage size in banking systems uses a systematic approach, with all the critical phases instrumental in understanding and predicting storage needs. It involves a series of sequential steps, including data collection, preprocessing, exploratory data analysis (EDA) [2], feature selection, and model selection.

After collecting the data from the banking systems, an extensive data preprocessing stage involved cleaning and feature engineering to enhance the dataset’s suitability for analysis.

An Exploratory Data Analysis (EDA) followed, utilizing various analytical techniques to derive insights into dataset characteristics and patterns [3]. Additionally, a feature selection process identified key variables essential for forecasting data storage size within banking systems. Finally, the model selection phase involved the application and comparison of predictive models to forecast storage requirements, considering their performance metrics.

The subsequent sections will delve deeper into each phase, detailing the specifics, processes, and outcomes of each stage, contributing to the comprehensive understanding and prediction of data storage size in banking systems.

1. **Data Description**

Exploratory data analysis (EDA) starts by importing essential libraries for data manipulation and visualization, tools such as pandas [4], numpy [5], matplotlib.pyplot [6], and seaborn [7]. Two datasets were used for analysis [Anexa A, Datasets]. These datasets, "History" and "Systems," provide an extensive array of information crucial for analysing the storage patterns, system characteristics, and operational attributes within banking systems. Next thing is to pre-process the data for further analysis.

1. **Data Preprocessing**

During data pre-processing [8], I focused on data cleaning and feature engineering to optimize the datasets for further analysis.

### *Data Cleaning*

The first step involved data cleaning procedures to ensure data integrity and relevance for analysis. This included identification and removal of unnecessary columns. In the "History" dataset, the 'ID', 'SYSTEM\_id', and 'DATE' columns were seemed irrelevant for further analysis and finally removed. Similarly, in the "Systems" dataset, the 'ID' column was eliminated to streamline the dataset. Also, data cleaning included duplicate and empty values check. These checks for duplicates and empty fields were conducted across the datasets. Fortunately, no duplicates or empty rows were identified among the records. This data cleaning process ensured that the datasets were free from redundant information, irrelevant columns, duplicate records, or empty values, facilitating a more focused and meaningful analysis.

### *Feature Engineering*

Following data cleaning, the next step involved the creation of new features derived from existing data attributes. The newly engineered features include:

'Active\_All': A ratio representing the number of active accounts relative to the total number of accounts.

'Backup\_Efficiency': A ratio representing the size of the backup in relation to the system's size.

'Kept\_size': The residual amount of data remaining after transmission for storage.

### *Defining Variables*

After feature engineering, variables were categorized into numerical and categorical categories for convenience in further analysis:

Numeric variables (Historical data):

['SIZE', 'LOAD\_TPD', 'ACCOUNTS\_ALL', 'ACCOUNTS\_ACTIVE', 'Non\_kept\_size', 'Backup\_size', 'LongOps\_min', 'Active\_All', 'Backup\_Efficiency', 'Kept\_size']

Categorical variables (System data):

['Name', 'Stage', 'Description', 'Type', 'Data Keep, years', 'Backup retention, month', 'Depreciation period, years', 'Data retention, years']

### *Data Standardization and Normalization:*

For numerical data, standardization and normalization techniques were applied to bring variables to a consistent scale:

MinMax Scaling: Scaled values to a specified range (typically 0 to 1) while preserving relative proportions.

Standardization: Centered the data by the mean and scaled by the standard deviation, resulting in a mean of 0 and a standard deviation of 1.

Additionally, for categorical data, One-hot Encoding was implemented to convert categorical variables into binary (dummy) variables, making their utilization easier in numerical-based models without imposing false hierarchy or significance among categories.

1. **Exploratory Data Analysis (EDA)**

Conducting a detailed Exploratory Data Analysis (EDA) was vital to easier understand the dataset's characteristics and distributions. The EDA started by calculating descriptive statistics for each numerical variable and generating visualizations. Histograms were utilized to visualize the distribution of numerical variables, while scatterplots were examined for potential correlations among variables. Additionally, pie charts and bar charts were employed for categorical variables to depict their distributions. Statistics provided a comprehensive overview of the dataset's characteristics (see [Annex A, Figure 2]). The univariate data analysis involved the visualization of numerical variables to explore their distributions and symmetry. Each variable was individually assessed using graphical representations generated from the dataset. The visualization methods used were histograms and box plots, providing insights into the distributions and skewness of each variable. Bivariate data analysis involved using barplots and a pairplot to show the relationships between different aspects. Multivariate data analysis, heatmap, was used to assess correlations between variables.

1. **Feature Selection**

Feature selection plays an important role in enhancing modeling accuracy by leaving only the most important variables and excluding the others. This process aids in fortifying model interpretability, reducing overfitting risks, and refining model generalization. Various filtering methods were employed to extract informative variables for modeling. Techniques such as Heatmap for visualizing correlations, Mutual Information [10] for evaluating dependencies, SelectKBest [11] for selecting top variables, and ANOVA F-test [12] for assessing variable significance were utilized. From wrapper methods, Recursive Feature Elimination (RFE) was used to identify crucial variables by progressively eliminating less significant ones [13]. Random Forest [14], tree-based method, assesses feature importance based on their contribution to reducing uncertainty within each tree. Also regularization methos where applied. Their penalty system was used to identify important features.

1. **Model Selection**

Selecting a suitable model after selecting variables is essential to obtain optimal results. A suitable model must match the nature of the data, taking into account its structure and features. This may involve selecting a model that can perform well on the selected features and has the best generalization ability on new data. It is important to consider the size and quality of the data, as well as the context and purpose of the problem, in order to select the most appropriate model for the banking system. Different models were tested and used to try and predict future banking systems size. All the models were tested both on all features and on selected ones. Also, Shapley values were used, which are a game theory technique used in machine learning to determine the contribution of each feature to a model's prediction. They measure the importance of features by identifying their influence on changes in model predictions for each specific observation. Multiple Linear Regression (MLR) [19] is a statistical model used to analyse the relationship between a dependent variable and multiple independent variables. This model is one of the basic methods of regression analysis and is used to predict or identify relationships in data. Ridge Regression is a regression technique that augments conventional linear regression with L2 regularization. It is specifically designed to model the relationship between a dependent variable and predictors while mitigating overfitting concerns, typically associated with multicollinear data. Lasso Regression, a linear regression technique employing L1 regularization, minimizes the loss function and reduces feature coefficient weights. Elastic-Net Regression combines the features of Lasso (L1 regularization) and Ridge (L2 regularization) methods, effectively blending their advantages. It addresses both feature selection, akin to Lasso, and control over coefficient scaling, similar to Ridge, aiding in tackling multicollinearity and creating more resilient, accurate models. Random Forest is an ensemble machine learning model that leverages multiple decision trees. It builds several decision trees during training and amalgamates them to yield more accurate predictions. KNeighborsRegressor [20] is a nearest neighbour regression model that utilizes the proximity of the "k" nearest neighbours to forecast values for a target variable in new data. Employed primarily for regression problems, it predicts values based on the average of the target variable’s nearest neighbours. Gradient Boosting Regressor [21] is an ensemble algorithm that sequentially constructs a predictive model, enhancing it with each iteration. It utilizes gradient descent to reduce the error of previous models, combining weak models into a stronger one capable of highly accurate predictions in regression problems.

**Results**

The analysis defined an overall evaluation of various regression models applied to predict the size of the banking system based on all and selected features. These models were assessed using different metrics to gauge their performance and effectiveness in predicting the system's size accurately. During EDA, I found out that the distribution of systems by type and stage revealed intriguing patterns (see [**Error! Reference source not found.**]): Graphs showcased an almost even distribution among the system types. Notably, Test and Backup systems each represented 11.1% of the total, while Development, Backup, Production, and Archive systems occupied 16.7% each. However, the distribution among system types differed slightly. DWH accounted for 11.1%, OLTP for 27.8%, Mixed for 27.8%, and Bkp for 33.3%. This variation in distribution is attributable to the prevalence of OLTP systems in real-time transaction processing and the importance of data backup, as reflected by the Bkp systems.

*Univariate data analysis* gave the following insights:

System size (see [Annex A, Figure 3]): Demonstrated an even distribution without anomalous values, displaying skewness close to zero, indicating a near-normal distribution.

Load (see [Annex A, Figure 4]): Showed uneven distribution and positive skewness, with the majority of values within the 10-25 million range, validating prior statistics.

Total and active accounts (see [Annex A, Figure 5] and [Annex A, Figure 6]): Presented a relatively uniform distribution for both total and active accounts, with a slight right tail evident in the asymmetry.

Data transferred to storage and remaining data (see [Annex A, Figure 7] and [Annex A, Figure 8]): Depicted an evenly distributed pattern for remaining data and a slightly right-skewed distribution for transferred data.

Backup size (see [Annex A, Figure 9]): Exploring the backup size distribution reveals a left-skewed pattern (asymmetry -0.4), indicating a slight shift towards smaller values in most instances.

Duration of complex operations (see [Annex A, Figure 10]): Observing the duration of complex operations shows a noticeable concentration around 450 units. Prior to system optimization, these operations consistently took around 400-450 minutes. After optimization, the duration decreased to approximately 100-250 minutes, resulting in a left-skewed distribution (asymmetry -0.5).

Stages and Types (see [Annex A, Figure 11]): Analysing categorical data shows nearly uniform distribution across system stages. However, fewer systems are evident in the testing and reserve stages. Regarding system types, only two data warehouses are seen, while the remaining types exhibit an even distribution.

Data storage and backup (see [Annex A, Figure 12]): The analysis indicates that a majority of systems store data for less than a year, with few exceptions primarily within archives and business analytics systems, storing data for 20 years. However, backup distribution appears relatively uniform, evenly distributed between backups for a month and less than a month.

*Bivariate data analysis* showed the following:

Size comparison by system Stage and Type (see [Annex A, Figure 13]): Archives and backups occupy the most space within banking systems, logically due to storing substantial information for extended periods. In opposite, test, stage and standby systems utilize the least space. Backup systems consistently utilize more space, while OLTP online transaction systems occupy the least.

Data storage Size and Duration and Depreciation relationship (see [Annex A, Figure 14]): Systems storing data for 2 and 7 years occupy more space compared to those storing data for shorter (less than a year) or longer periods (20 years). Systems with data becoming obsolete in 7 years occupy more space than those with depreciation in 0 and 5 years.

Pairplot (see [Annex A, Figure 15]): The pairplot visualizes relationships among variables, highlighting a linear and positive correlation between size, accounts (active and all), load, and stored size.

Utilizing a heatmap [9] (see [Annex A, Figure 16]) for multivariate analysis unveils strong associations among variables presented in matrix form. It demonstrates substantial correlations among variables. Notably, system size shows positive correlations with the number of accounts (both total and active), load, and backup size. Oppositely, a negative correlation exists between size and backup efficiency along with operations. This indicates that positive alterations in independent variables positively influence the dependent variable in the former case, while inversely affecting it in the latter.

During feature selection, following results were obtained:

Correlation (see [Annex A, Figure 16]): Correlation analysis was employed to identify variables strongly impacting the dependent variable. While this method aids in variable selection, caution is necessary to address biases and multicollinearity concerns. Notably, the system's remaining size exhibits dependency on factors such as system size, accounts (both general and active), and system load. While correlation can be used with categorical data (see [Annex A, Figure 17]), it is not commonly used in feature selection because there are other, better methods.

Mutual Information: Employing Mutual Information showed the degree of dependence between variables. It highlighted the relationship between remaining size (2.51), accounts (active 2.38 and total 2.43), transferred size (2.20), and system load (1.78), mirroring the findings observed in correlation analysis. Also, for the categorical variables, type seems to influence size the most (OLTP type 0.57, Mixed 0.32, Backup 0.22 and DWH 0.17).

SelectKBest: Utilizing SelectKBest, the five most impactful variables were determined, showcasing a pattern consistent with previous selections. Variables like accounts, load, remaining memory, and backup size emerged as predominant contributors to the model. For categorical variables it was archive stage, backup type, OLTP type, depreciation period of 7 years and data retention of 7 years.

ANOVA F-test: ANOVA F-test assessed categorical traits' impact on the target variable, paralleling SelectKBest findings and showing the significance of selected categorical features.

RFE (wrapper meyjod) detected varying sets of important variables compared to previous methods, emphasizing the total number of accounts, transferred size, backup size, backup efficiency, and remaining size. RFE for categorical data highlighted key features such as the archive stage, storage types (mixed and OLTP), and data retention for a 2-year period.

Random forest method reiterated the significance of features previously identified, emphasizing total (0.27), and active (0.18) accounts, system load (0.23), and remaining size (0.15). The consistency across methods reaffirms the reliability of these features as key contributors to the model.

Random Forest for categorical data: A similar trend persists, underscoring the importance of the archive stage (0.55) and OLTP type (0.18), although the significance of other categorical variables is very close to 0.

Lasso (L1 Regularization) [15]: Lasso's penalty-induced feature selection eliminated two variables—ratio of active accounts to the total and backup efficiency. The RMSE (24.45) snd standard deviation of RMSE (9.48) remained relatively low across this analysis, signifying acceptable model performance.

Lasso for categorical data: Alarmingly, Lasso discarded all features except for system stages and types, presenting concerns regarding the suitability of this dataset for forecasting.

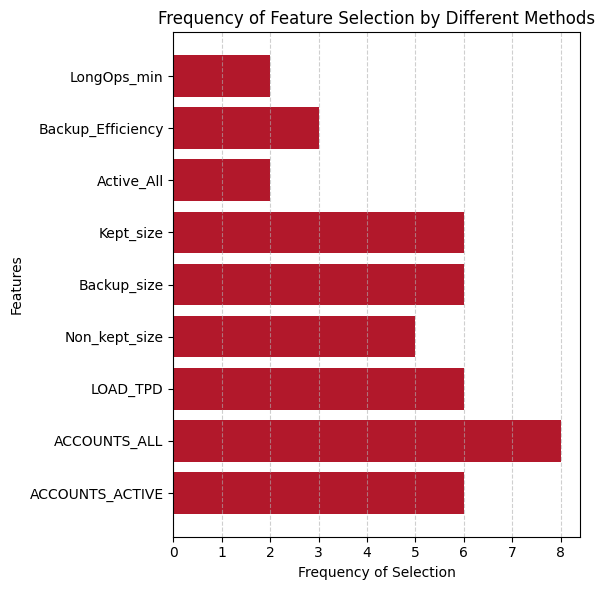
Ridge (L2 Regularization) [16]: Ridge regularization demonstrated a significantly low RMSE (0.017), highlighting the importance of remaining size and transferred size as critical variables for the model.

Ridge for categorical data: Similar to Lasso outcomes, Ridge faced challenges with the dataset, hinting at potential limitations in its application for predictive modeling. RMSE was very high (6817.49).

Elastic Net (Combining L1 and L2 Regularization) [17]: Elastic Net's extensive regularization parameter nearly nullified most coefficients, indicating minimal reliance on most features except for accounts and system load.

Elastic Net for categorical data: The RMSE (6817.48) pattern repeats, which is very bad, this method emphasized the significance of the archive stage, 7-year periods, and backup types, but still this method cannot be relied upon with such high error.

These results show that data with information about systems is unlikely to be suitable for forecasting, but still was checked further. As for the remaining data, a graph was made with the frequency of their selection by methods.



***Figure 1.***Frequency of feature selection

It is noticeable that the number of accounts, total and active, the load on the system and the size of the remaining space and backup were the most frequently selected variables. Variance Inflation Factor (VIF) [18] was also taken into account, which, when using all the characteristics, showed very high parameters with the highest Kept\_size equal to 21340.01.

During Model selection, different models were tested with the following results:

*Multiple Linear Regression (MLR)*

All features: The model's performance, evaluating all features, highlights multicollinearity issues. While an R-squared value of 1.0 might seem promising, it signifies model overfitting, rendering this approach impractical for further consideration.

Selected features (see [Annex A, Figure 18]): Summary of coefficients, root mean square error, and R-squared score obtained from the selected features: The model performs significantly better with the selected features. Achieving an R-squared value of 0.99983 on training set and 0.99985 on test set indicates strong performance with a minimal error (RMSE on training set 15.061 and on testing set 16.116). Comparison of actual values and values predicted by the model shows that the values are almost identical. Model error estimate shows that the distribution of errors is very symmetrical, close to normal, which very well shows the absence of bias and the good performance of the model. From the residual plots the randomness of the errors is clearly visible, another good indicator of a good model.

It can be seen that for this model (see [Annex A, Figure 19]), the largest contribution was made by the size of the backup, the number of users and the transferred size. You can also notice that the higher the value of these variables, the higher the value of the predicted size.

To test the model, 10 new values of independent variables were taken and the model, based on them, predicted the future size of the system for the next 10 months (see [Annex A, Figure 20]), the predicted size looks very realistic, which indicates that the model is working well.

Categorical data: As deducted earlier, data on systems is most likely not suitable for forecasting. But for verification, MLR and selected features were taken. Although MLR demonstrated high R2 scores during training (0.936), its performance on categorical data for forecasting was inadequate (RMSE during training 1020.62 and testing 3978.23). Despite strong training performance, the model struggled to generalize to new data, suggesting potential noise or limited training data.

*Ridge Regression Model*

All Features (see [Annex A, Figure 21]): The Ridge Regression model demonstrates high R2 scores (0.9992 and 0.99694) on both training and testing datasets. While the RMSE values (32.8 and 73) indicate acceptable size prediction, the model shows more errors on the test data than during training. Comparison of predicted and actual values shows that the model's deviations are slightly more pronounced compared to the linear model but still closely align with real values. Model error estimation indicates minor skewness to the left, but errors are otherwise uniformly distributed. The comparison of actual and predicted values appears promising. Residual plots show that the residuals are marginally concentrated in the positive direction of actual values but maintain a presence on both sides. Shapley values show that the largest contributions to this model were made by backup size, efficiency, and user count. Higher values of these variables positively correlate with predicted size, except for backup efficiency, which exhibits the opposite effect.

Selected Features (see [Annex A, Figure 22]): On the selected features, this model showcased superior performance with an R2 score close to 1 and impressive RMSE values (1.686 and 1.724), signifying strong training outcomes. Comparison of predicted and actual values shows that predicted values closely match actual values, displaying high accuracy. Model error estimation shows a uniform distribution of errors and highlights the model's strong performance. Residual plots showcase scattered errors and indicate a lack of bias in the model, which is a result of successful training. Shapley values (see [Annex A, Figure 23]) show similar to previous observations, that all variables positively impact the model, with higher attribute values leading to increased predictions. To test the model, 10 new values of independent variables were taken and the model, based on them, predicted the future size of the system for the next 10 months (see [Annex A, Figure 24]), the predicted size looks very realistic, which indicates that the model is working well.

*Lasso Regression Model*

All Features (see [Annex A, Figure 25]): Comparatively, this model outperformed the Ridge Regression model in all aspects. The R2 scores are very close to 1, and the RMSE values (8.595 and 9.56) are strong indicators of successful training. Comparison of predicted and actual values results indicate nearly identical values between predictions and actual data. Model error estimation indicates that the errors are uniformly distributed, with predicted values closely aligned with the actual values. Residual plots show that while slightly more values are seen on the positive side, deviations remain small. Shapley values similar to other models, indicate that most features positively influence the predicted values, except for backup efficiency.

Selected Features (see [Annex A, Figure 26]): Lasso performed less effectively with its selected features compared to using all of them. A marginal decrease in R2 and an increase in RMSE (14.785 and 15.348) were observed, indicating slightly less accurate predictions. Comparison of predicted and actual values results depict a close match between predicted and actual values, maintaining high accuracy. Model error estimation shows that the errors are distributed nearly evenly, with predicted values lying close to the line, showcasing very small errors but slightly more visible than the previous model. Residual plots showcase scattered errors and indicate a lack of bias in the model, signalling successful training outcomes. Future size prediction shows that the model accurately predicts future system sizes based on 10 new values of independent variables, indicating efficient functionality.

*Elastic-Net Regression Model*

All Features (see [Annex A, Figure 27]): Comparatively, this model exhibited lower R2 scores (0.988 and 0.957) and considerably higher RMSE values (127.4 and 273.2) compared to previous models. This might suggest poorer learning performance and less generalization ability to the broader population. Comparison of predicted and actual values shows that deviations from actual values are notably more pronounced in this model compared to previous ones, indicating less favourable performance. Model error estimation indicates the error distribution exhibits a distinct tail, with noticeable deviations. Residual plots showcase errors that are concentrated above the actual values, accompanied by a pronounced tail, observed in both training and testing data.

Selected Features (see [**Error! Reference source not found.**]): On the selected features, the model marginally outperformed its performance with all features but still lags behind other models, as indicated by higher RMSE values (94.671 and 92.745). Comparison of predicted and actual values shows that improvements are evident in comparison to the model using all features, but deviations remain apparent. Model error estimation shows that though a tail on the left is observed in error distribution, it's less pronounced than the previous instance. Residual plots show that errors are still gathered above the actual values, but the spread is less pronounced. Future size prediction showcased model generating plausible predictions based on 10 new independent variables, indicating satisfactory performance.

*Random Forest Model*

All Features (see [Annex A, Figure 29]): While the model displayed slightly inferior performance on testing data (RMSE 99.20) than training data (RMSE 46.63), the high R2 (on training 0.998 and on testing 0.994) suggests decent model performance. Comparison of predicted and actual values shows that the distinction between the predicted and actual values is almost undetectable. Model error estimation demonstrates a left-tail, similar to other models, with predicted variable values nearly aligned with slight deviations. Residual plots show that most errors are in proximity to the line, signifying favourable accuracy, albeit with an anomaly displaying a negative deviation of over -200. Shapley values show that this model assigns more weight to features compared to previous models, with a slightly altered significance order, distinctly varying from the prior models.

Selected Features (see [Annex A, Figure 30]): The model performed better on the selected features, with R2 nearing 1 and reduced RMSE values (38.61 and 49.648). Feature selection evidently enhanced the model's performance. Comparison of predicted and actual values demonstrates that predicted values closely mirror actual values, displaying minimal differences. Model error estimation shows a uniform distribution of errors signifies robust performance, with most errors falling between -50 and 50, indicating satisfactory accuracy. Residual plots show that errors are scattered chaotically and close to zero, suggesting a low or negligible bias in this model. Shapley values show that selected variables contribute almost uniformly and positively to the model, indicating higher feature values correspond to higher prediction values. Future size prediction showcased model generating unplausible predictions based on 10 new independent variables, indicating disastrous performance. The values are all just a straight horizontal line.

*KNeighborsRegressor Model*

All Features (see [Annex A, Figure 31]): Compared to other models, this model's performance is relatively inferior, as indicated by a slightly lower R2 (0.989 and 0.976) and slightly worse RMSE values (118.93 and 204.91). Comparison of predicted and actual values demonstrates that noticeable differences exist between the actual and predicted values, indicating poorer performance compared to other models. Model error estimation shows that though the error distribution is uniform, deviations from the line are more pronounced than in other models, suggesting less accurate predictions. Residual plots showcase the error distribution that appears chaotic, which is favourable. However, high errors contribute to a less favourable outcome.

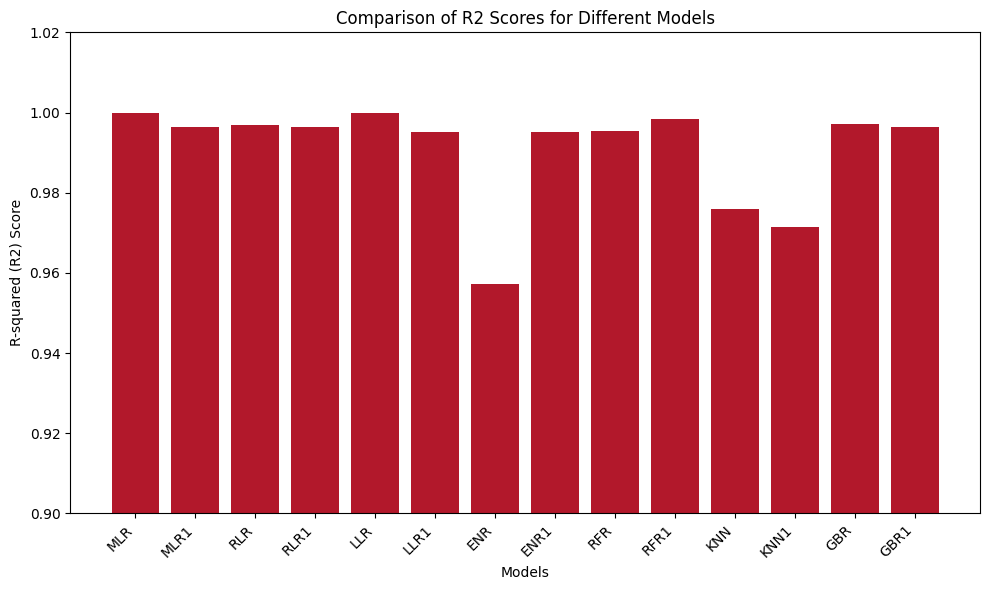
Selected Features (see [Annex A, Figure 32]): The R2 assessment (0.983 and 0.971) suggests worse coping with selected features compared to using all features, and the RMSE values (149.36 and 223.61) indicate less accurate predictions. Comparison of predicted and actual values shows that significant differences between predicted and actual values are noticeable, indicating worse performance in comparison to other tested models. Model error estimation shows that despite a uniform error distribution, the model's accuracy remains low, evident from the considerable deviations from the line. Residual plots that the errors are evenly distributed around zero, indicating minimal bias influence. However, these errors range significantly from -400 to 400, portraying lower accuracy in predictions.

*Gradient Boosting Regressor Model*

All Features (see [Annex A, Figure 33]): The model performed notably better on the training data than on the testing data with R2 in both cases being close to 1 (0.9999 and 0.996) but RMSE on training set 0.088 and on testing set 73.71. This disparity indicates potential bias and noise due to features, where the model understands the training data well but struggles to generalize to the entire population. Comparison of predicted and actual values shows that the differences between predicted and actual values are minor and hardly noticeable. Model error estimation demonstrates uniform error distribution without any deviation from the line, indicating consistent predictions. Residual plots show that while there's a slight trend in errors on the training data, the errors overall are chaotic. Training set errors are minimal, but the model's performance is comparatively poorer on the testing set.

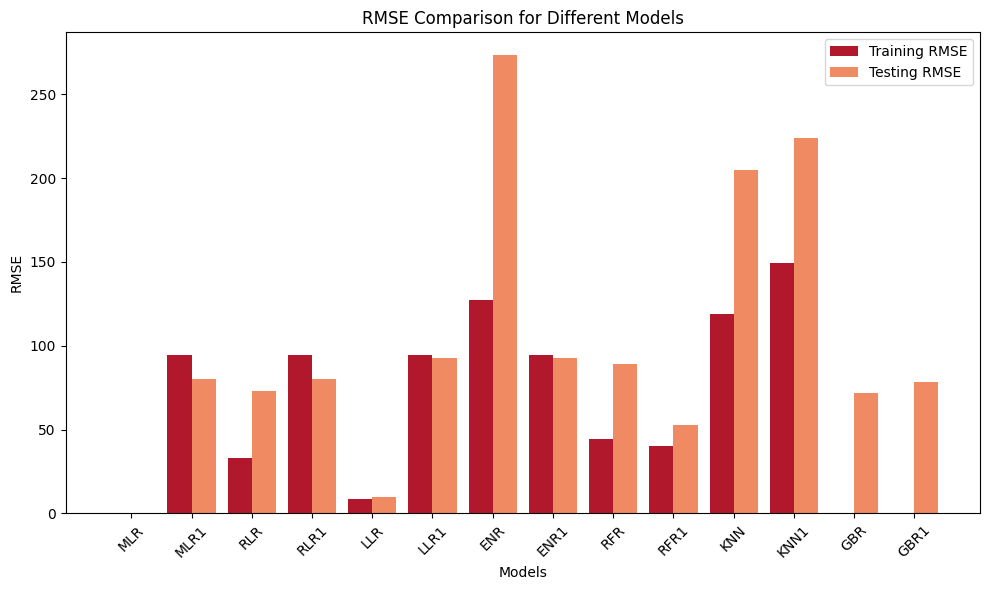
Selected Features (see [Annex A, Figure 34]): Similar to the performance on all features, the model excels during training (R2 0.9999 and RMSE 0.094) but exhibits lower performance during testing (R2 0.996 and RMSE 82.62). Comparison of predicted and actual values shows that minor differences are observed between the predicted and actual values. Model error estimation demonstrates normal error distribution, and errors are minimal, indicating robust performance. Residual plots showcase a discernible trend in the error spread, implying bias in the model, which results in low error rates but may affect generalizability.

The evaluation of the models revealed varying performance across different algorithms. Two critical metrics, the R2 score and the Root Mean Square Error (RMSE), were used to assess the models' predictive accuracy.



***Figure 2*** R2 comparison between models

Among these models, Multiple Linear Regression (MLR) on selected features, Ridge Regression (RLR), Random Forest (RFR) and Lasso Regression (LLR) showcased the highest R2 scores, indicating their superior ability to explain the variance in the system's size. Notably, MLR on selected features attained a perfect R2 score, which was influenced by overfitting of the model due to multicollinearity.



***Figure 3*** RMSE comparison between models

The RMSE evaluation graph [**Error! Reference source not found.**]) highlights the performance of models on both training and testing sets. Overall, Multiple Linear Regression (MLR) using selected features, Ridge Regression (RLR), and Lasso Regression (LLR) emerged as the most robust models. These models demonstrated consistent performance, achieving high R2 scores and exhibiting minimal error rates in prediction. While Random Forest Regression displayed impressive metrics initially, its inability to predict future data accurately raises concerns about its applicability in forecasting the banking system's size effectively. The consistent performance of MLR on selected features, Ridge, and Lasso models underscores their reliability and potential suitability for predicting the banking system's size.

**Conclusion**

The investigation into various regression models aimed to predict the size of the banking system based on a range of features has yielded valuable insights into the ability of different models. *The key findings are:*

**System Type:** Throughout the EDA and during feature selection, it was evident, that the type of the system plays a significant role in its size. For this study, a production system was used for prediction, but we can be sure, after the analysis, that if Archive system or Backup system were used, the size would be significantly larger. **Model Performance:** Through overall evaluation, it became evident that Multiple Linear Regression, Ridge Regression, and Lasso Regression emerged as the top-performing models. These models showed exceptional predictive power, demonstrating high accuracy in estimating the size of the banking system.

**Random Forest Challenges:** While Random Forest Regression initially displayed promising metrics, its disastrous struggle in accurately forecasting future data diminishes its reliability for predictive purposes in this context. The disparity between training and future data predictions raises concerns regarding its practical use for future estimations. **Importance of Feature Selection:** The models that excelled—Multiple Linear Regression, Ridge, and Lasso underscored the significance of thoughtful feature selection. These models showcased that focusing on relevant features significantly enhances predictive accuracy.

In conclusion, the evaluation of regression models for predicting the banking system's size emphasized the importance of model selection, feature relevance, and predictive accuracy. Leveraging robust models such as MLR on selected features, Ridge, and Lasso can significantly improve forecasting accuracy, providing invaluable insights for future planning and decision-making in the banking domain.

**Annex A**

**Datasets**

1. "History" Dataset

This dataset includes records spanning 50 months referring to a specific production system. It includes the following columns:

ID: Identification number for each observation.

DATE: Date of the observation.

SYSTEM\_id: Identifier for the respective system.

SIZE: System size measured in gigabytes.

LOAD\_TPD: Number of transactions processed per day.

ACCOUNTS\_ALL: Total count of accounts.

ACCOUNTS\_ACTIVE: Count of active accounts.

Non\_kept\_size: Quantity of data transmitted for storage.

Backup\_size: Size of the system's backup.

LongOps\_min: Duration of lengthy operations in minutes.

2. "Systems" Dataset

This dataset overall information regarding various banking systems, containing the following attributes:

ID: Identification number for each observation.

NAME: Name of the system.

Stage: Production stage of the system.

Description: Brief description detailing the system.

Type: Classification of system type.

Size, Gb: Current system size measured in gigabytes.

Data Keep, years: Duration of data storage availability online in years.

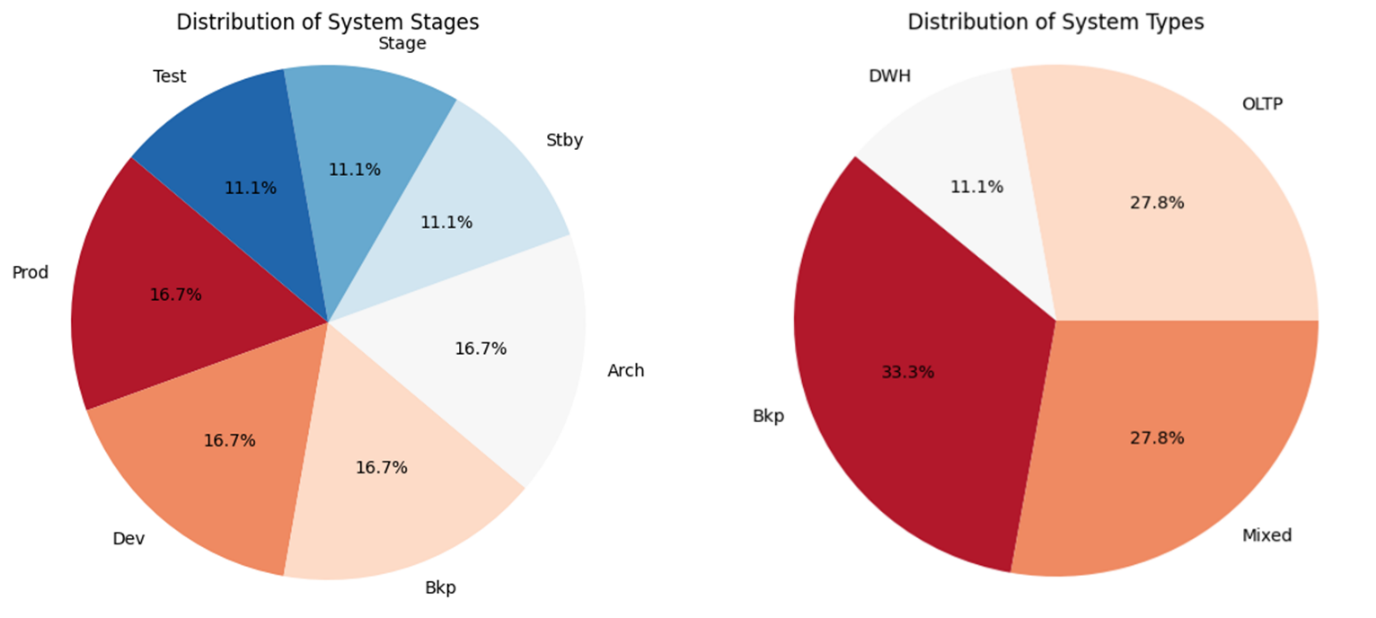
Backup retention, month: Retention period for system backups in months.

Depreciation period, years: Duration for data aging in years.

Data retention, years: Timeframe for data storage in years.

**Figure 1**

Title: Distribution of Systems by Stage

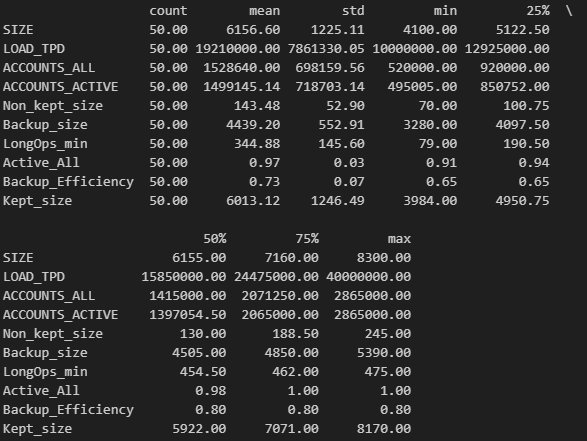


Annex A, Figure 1

Description: This Pie plot showcased an almost even distribution among the system types. Notably, Test and Backup systems each represented 11.1% of the total, while Development, Backup, Production, and Archive systems occupied 16.7% each. The distribution among system types differs. DWH accounted for 11.1%, OLTP for 27.8%, Mixed for 27.8%, and Bkp for 33.3%. This variation in distribution is attributable to the prevalence of OLTP systems in real-time transaction processing and the importance of data backup, as reflected by the Bkp systems.

**Figure 2**

Title: Descriptive statistics



Annex A, Figure 2

Description: System Size (SIZE): Average system size hovered around 6155 GB, ranging from 4100 to 8300 GB, with most systems falling between 5122 to 7160 GB.

Number of Transactions per Day (LOAD\_TPD): Average daily transactions were around 19,210,000, ranging from 10,000,000 to 40,000,000, with the majority processing between 12,925,000 and 24,475,000 transactions.

Number of Accounts (ACCOUNTS\_ALL and ACCOUNTS\_ACTIVE): Systems averaged about 1,528,640 total accounts, slightly fewer being active at approximately 1,499,145. The most system held between 920,000 to 2,071,250 accounts.

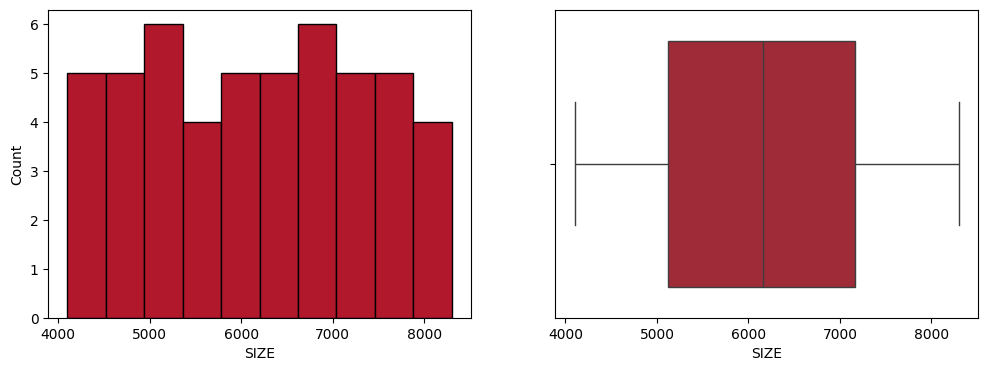
Backup Size (Backup\_size): Average backup size equated to about 4439.20 GB, with most systems falling between 4097 to 4850 GB.

Active Accounts Percentage (Active\_All): The average percentage of active accounts was about 97%, signifying high user engagement in banking systems.

Backup Efficiency: Average backup efficiency stood at approximately 80%, highlighting a substantial ratio of backup size to system size.

**Figure 3**

Title: System Size Distribution

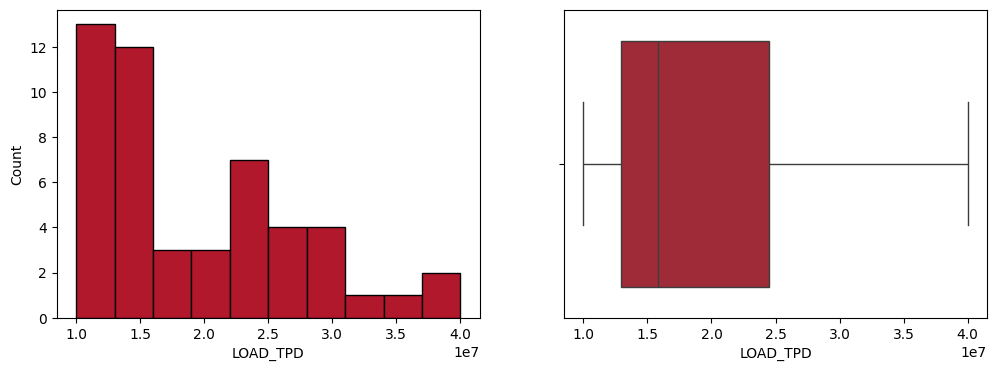


Annex A, Figure 3

Description: Both histogram and boxplot demonstrated an even distribution of size without anomalous values, with skewness close to zero, indicating a near-normal distribution

**Figure 4**

Title: Load Distribution

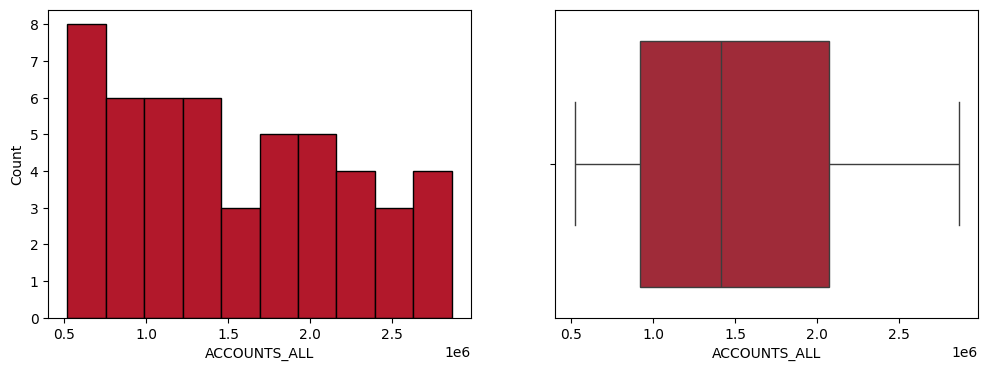


Annex A, Figure 4

Description: Both histogram and boxplot showed uneven distribution and positive skewness, with the majority of values within the 10-25 million range, validating prior statistics.

**Figure 5**

Title: Total Accounts Distribution

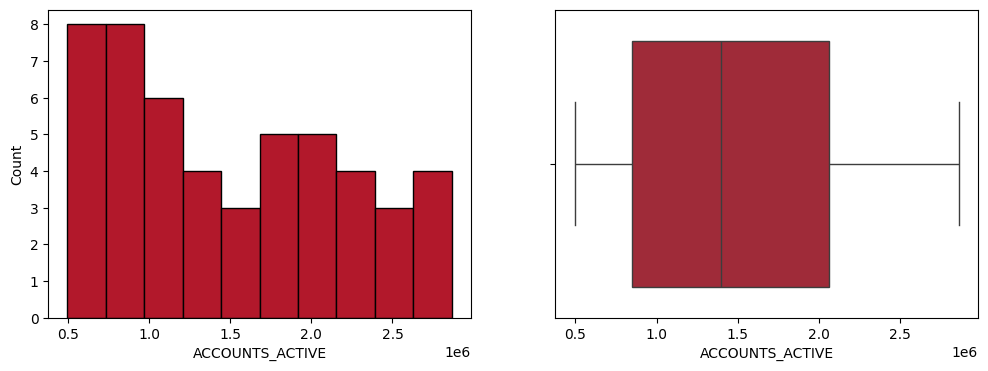
****

Annex A, Figure 5

Description: Both graphs presented a relatively uniform distribution for total accounts, with a slight right tail evident in the asymmetry.

**Figure 6**

Title: Active Accounts Distribution

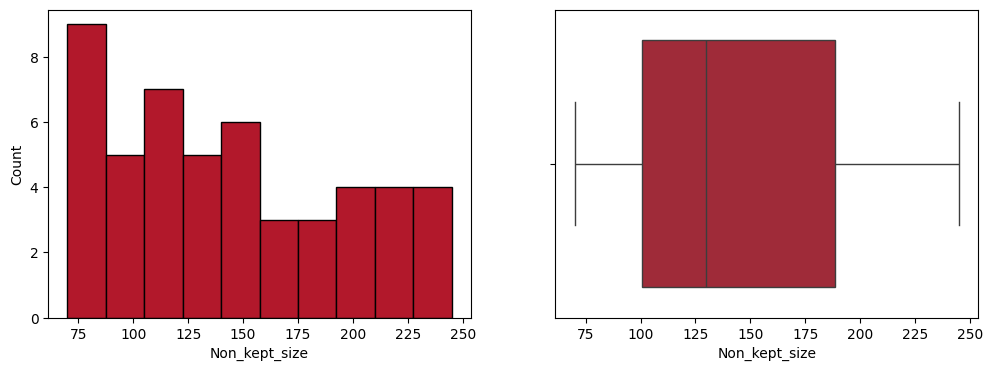


Annex A, Figure 6

Description: Both graphs presented a relatively uniform distribution for active accounts, with a slight right tail evident in the asymmetry.

**Figure 7**

Title: Transferred Data Distribution

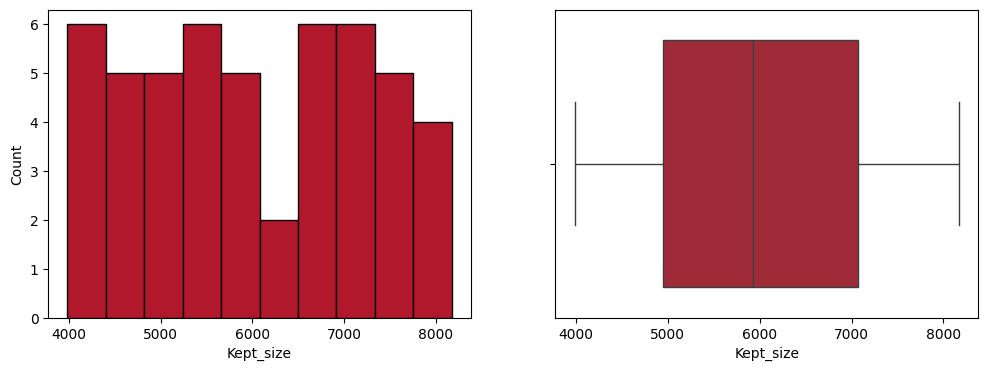
****

Annex A, Figure 7

Description: Graphs depicted a slightly right-skewed distribution for transferred data.

**Figure 8**

Title: Remained Data Distribution

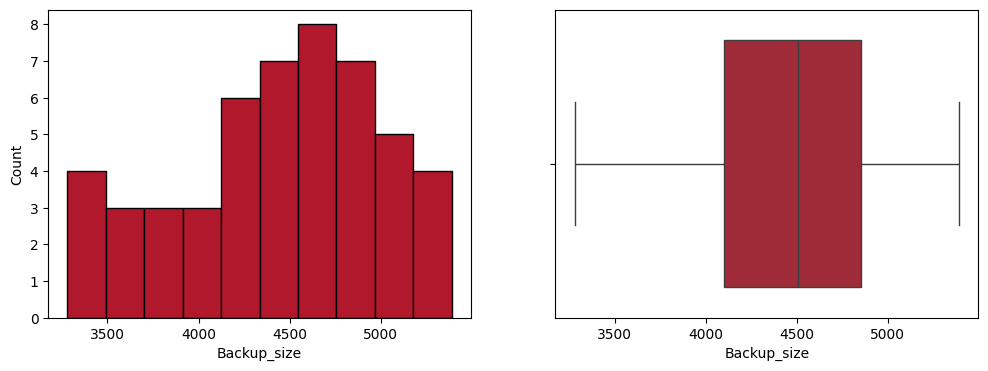
****

Annex A, Figure 8

Description: Both graphs showing an evenly distributed pattern for remaining data

**Figure 9**

Title: Backup Size Distribution

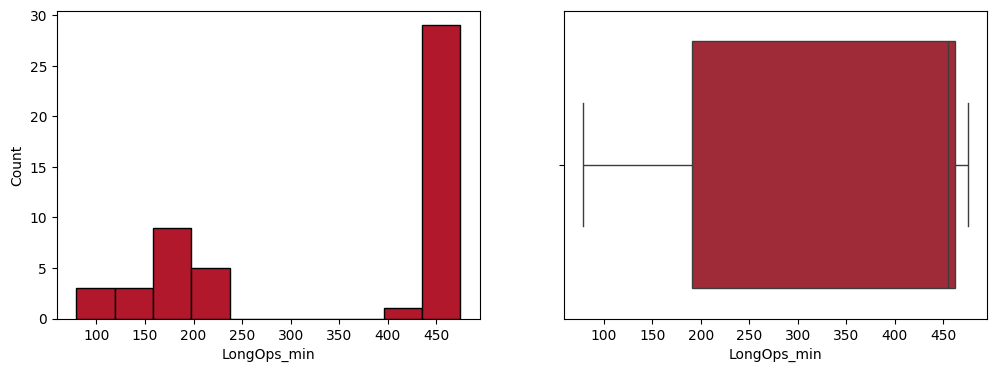


Annex A, Figure 9

Description: Exploring the backup size distribution reveals a left-skewed pattern (asymmetry -0.4), indicating a slight shift towards bigger values in most instances.

**Figure 10**

Title: Duration of Complex Operations Distribution

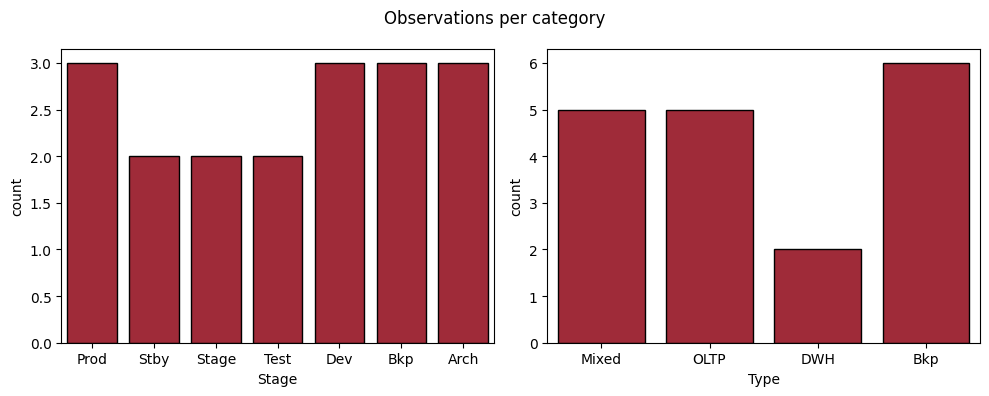


Annex A, Figure 10

Description: Observing the graphs for duration of complex operations shows a noticeable concentration around 450 units. Prior to system optimization, these operations consistently took around 400-450 minutes. After optimization, the duration decreased to approximately 100-250 minutes, resulting in a left-skewed distribution (asymmetry -0.5).

**Figure 11**

Title: System Stages and Types Distribution

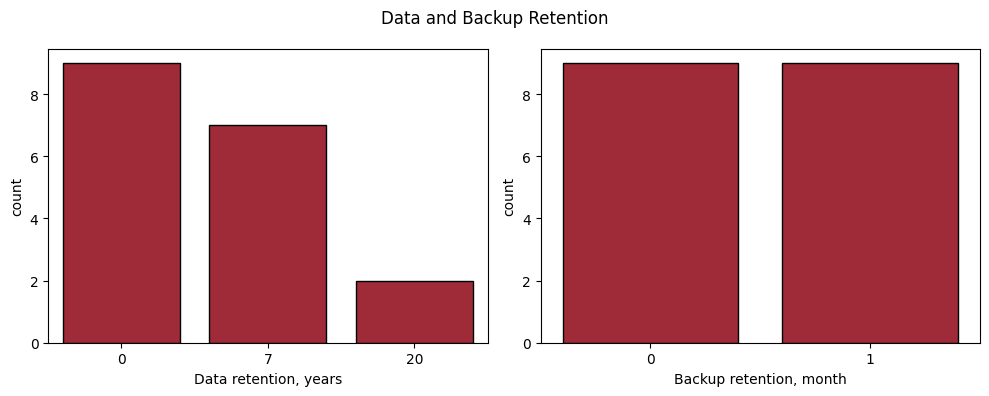


Annex A, Figure 11

Description: Analysing categorical data shows nearly uniform distribution across system stages. However, fewer systems are evident in the testing and reserve stages. Regarding system types, only two data warehouses are seen, while the remaining types exhibit an even distribution.

**Figure 12**

Title: Data and Backup Retention Distribution

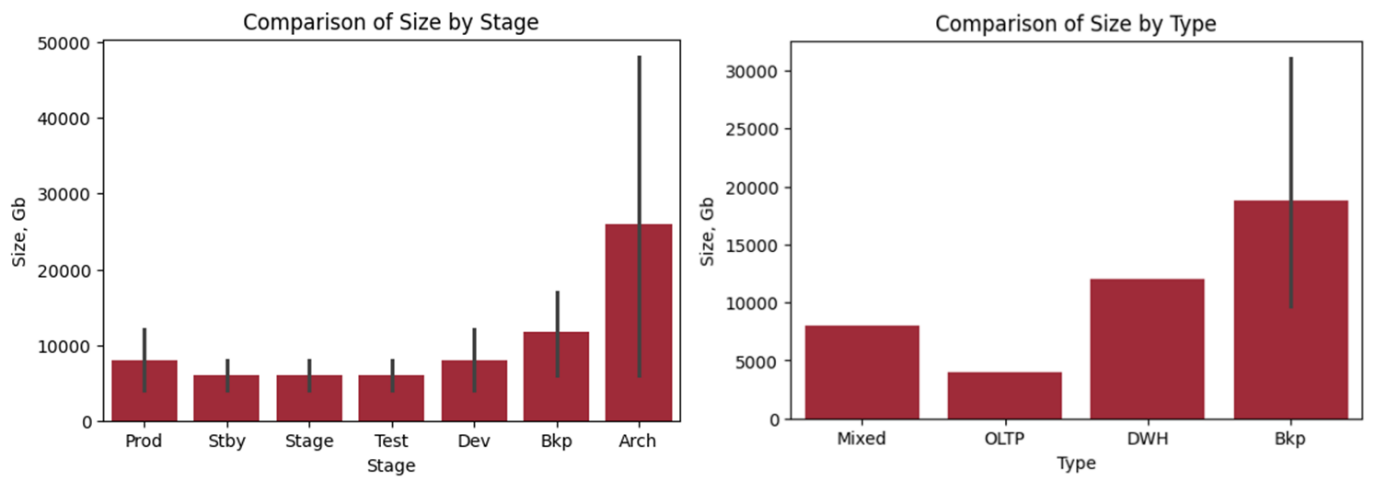


Annex A, Figure 12

Description: The analysis indicates that a majority of systems store data for less than a year, with few exceptions primarily within archives and business analytics systems, storing data for 20 years. However, backup distribution appears relatively uniform, evenly distributed between backups for a month and less than a month.

**Figure 13**

Title: Comparison of Size by System Stage and Type

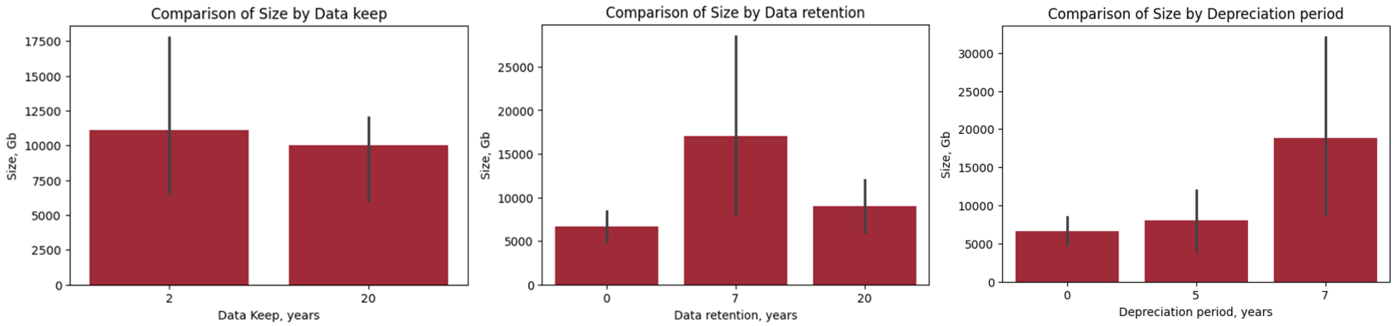


Annex A, Figure 13

Description: Archives and backups occupy the most space within banking systems, logically due to storing substantial information for extended periods. In opposite, test, stage and standby systems utilize the least space. Backup systems consistently utilize more space, while OLTP online transaction systems occupy the least.

**Figure 14**

Title: Comparison of Size by data keep and data retention and depreciation period

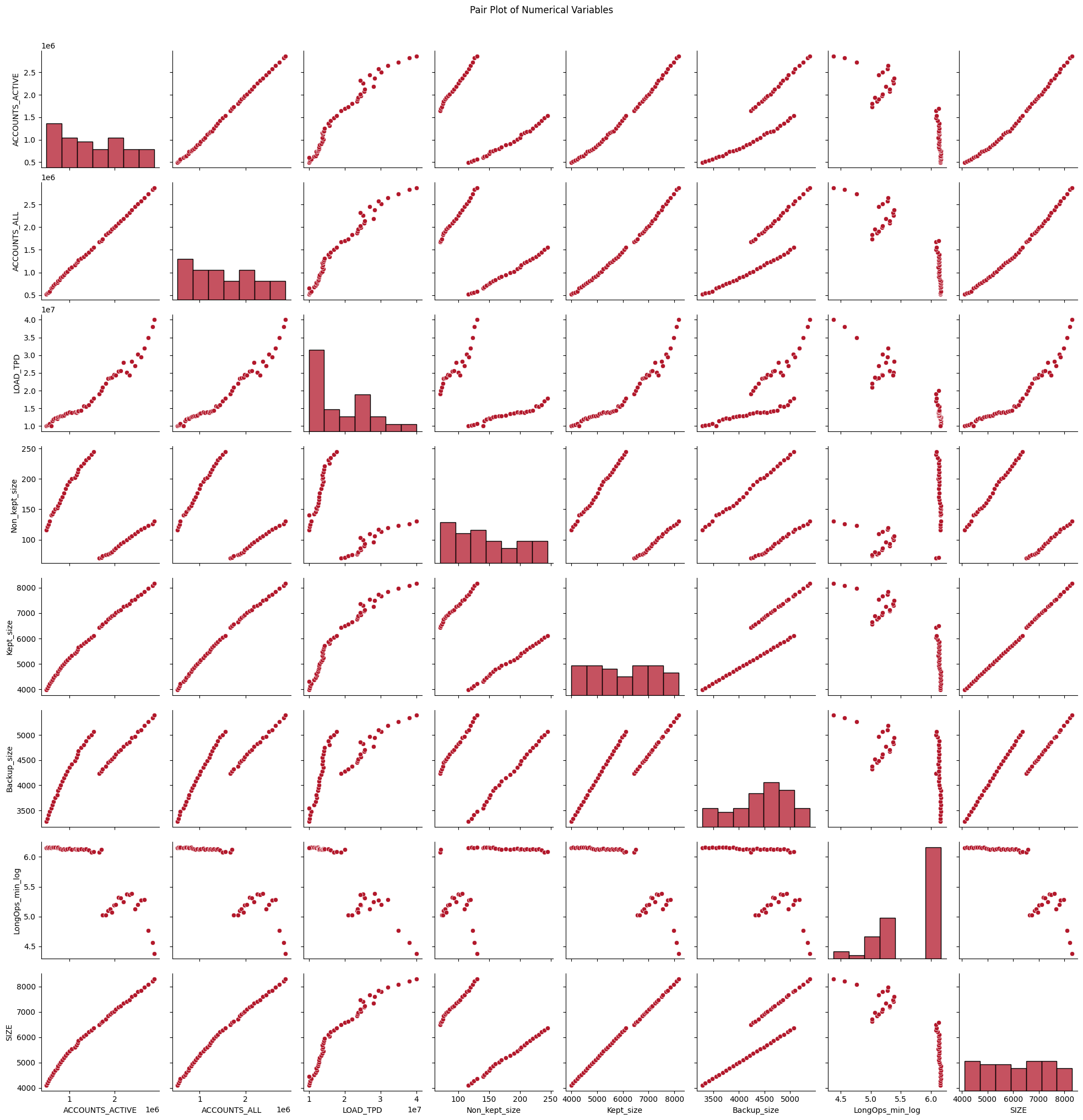


Annex A, Figure 14

Description: Systems storing data for 2 years occupy more space compared to those storing data for longer periods (20 years). Systems retaining data for 7 years occupy more space compared to those storing data for shorter (less than a year) or longer periods (20 years). Systems with data becoming obsolete in 7 years occupy more space than those with depreciation in 0 and 5 years.

**Figure 15**

Title: Pair Plot of Numerical Variables

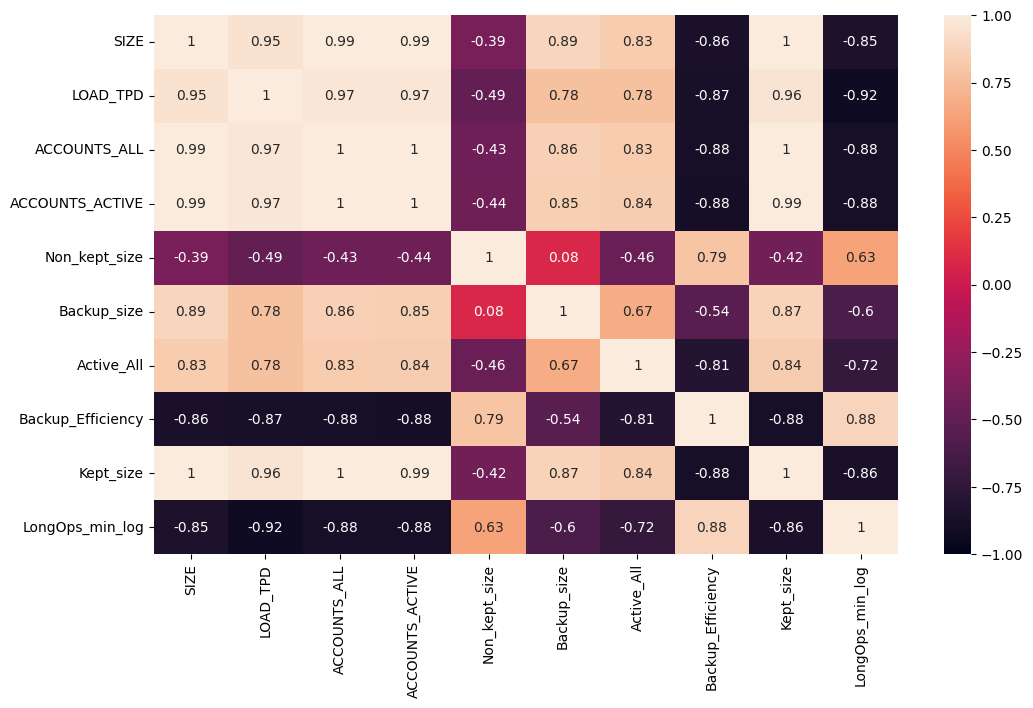
****

Annex A, Figure 15

Description: The pairplot visualizes relationships among variables, highlighting a linear and positive correlation between size, accounts (active and all), load, and stored size.

**Figure 16**

Title: Heatmap for Multivariate Analysis



Annex A, Figure 16

Description: Heatmap demonstrates substantial correlations among variables. Notably, system size shows positive correlations with the number of accounts (both total and active), load, and backup size. Oppositely, a negative correlation exists between size and backup efficiency along with operations. This indicates that positive alterations in independent variables positively influence the dependent variable in the former case, while inversely affecting it in the latter.

**Figure 17**

Title: Correlation for Categorical Data

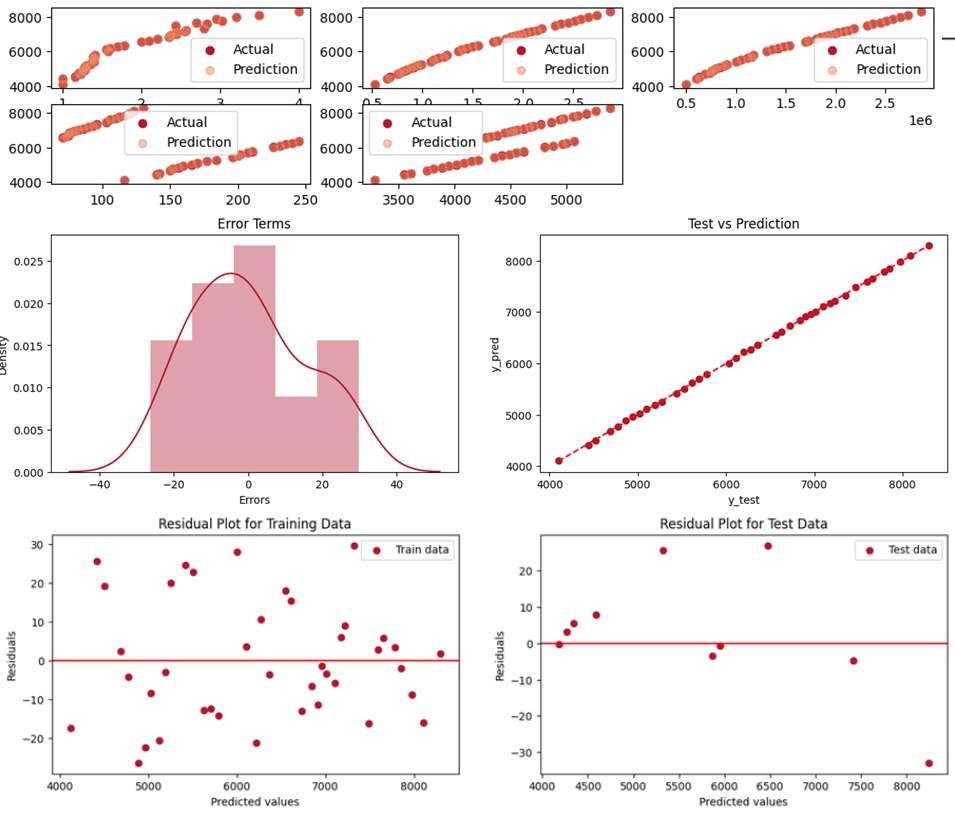


Annex A, Figure 17

Description: While correlation can be used with categorical, it is not commonly used in feature selection because there are other, better methods. Here we can see strong positive relationship of size with Archive stage and Backup Type, also with depreciation and retention period of 7 years. Also, negative relationship can be seen with the OLTP system type, as well as with data retention and depreciation of under a year and backup retention of under a month.

**Figure 18**

Title: MLR model evaluation

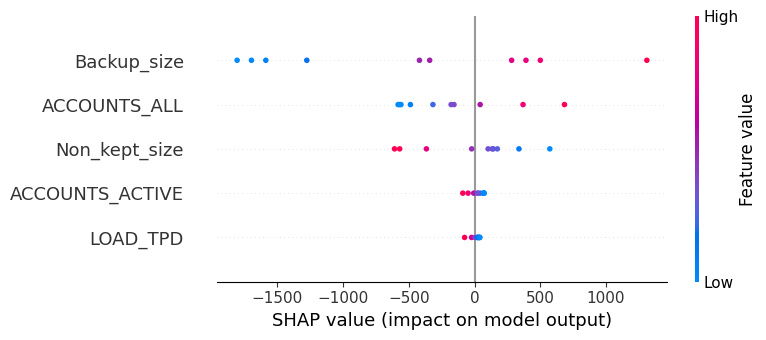


Annex A, Figure 18

Description: Comparison of actual values and values predicted by the shows that the values are almost identical. Model error estimate shows that the distribution of errors is very symmetrical, close to normal, which very well shows the absence of bias and the good performance of the model. From the residual plots the randomness of the errors is clearly visible, another good indicator of a good model.

**Figure 19**

Title: Shapley Values for the MLR model (selected features)

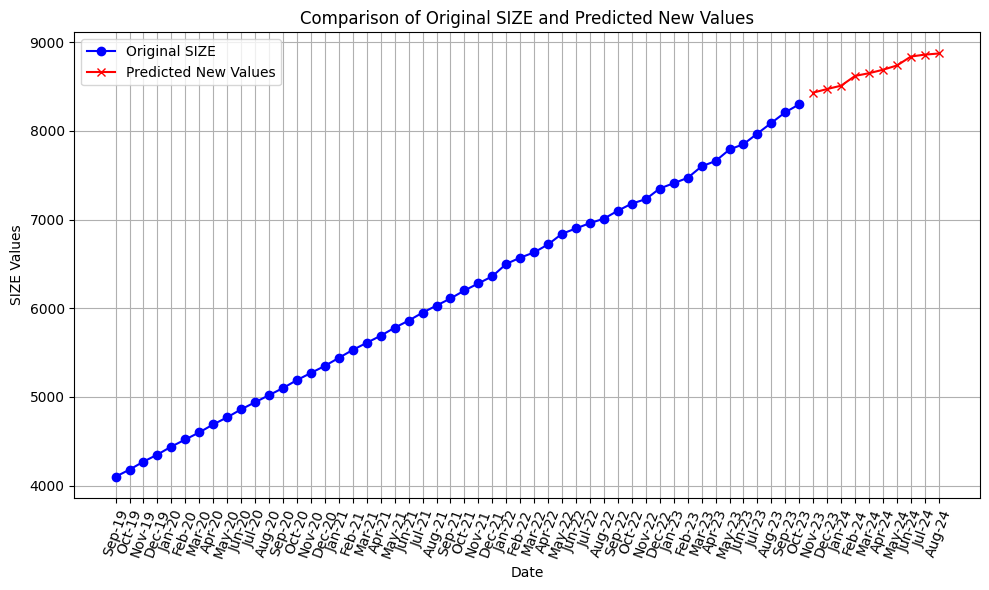


Annex A, Figure 19

Description: It can be seen that for this model the largest contribution was made by the size of the backup, the number of users and the transferred size. You can also notice that the higher the value of these variables, the higher the value of the predicted size.

**Figure 20**

Title: Prediction of Size by MLR model

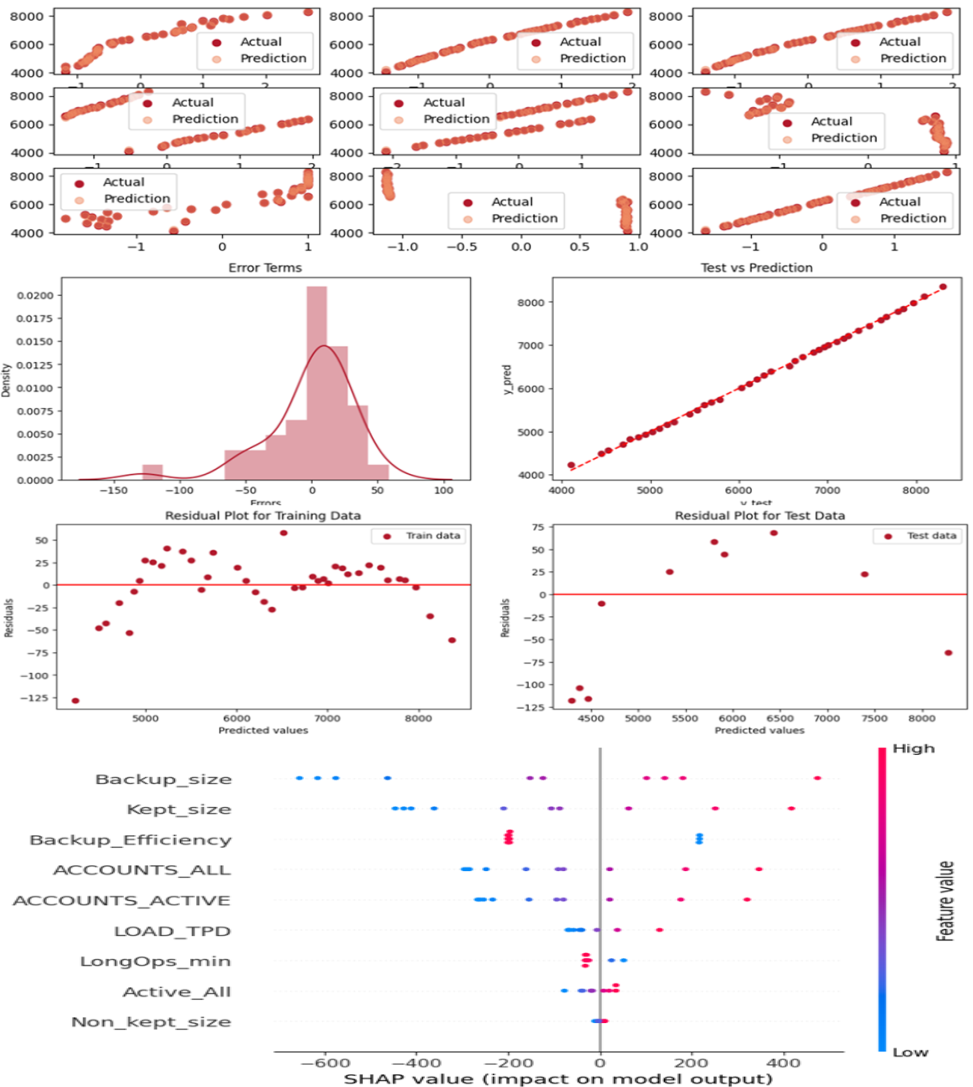


Annex A, Figure 20

Description: To test the model, 10 new values of independent variables were taken and the model, based on them, predicted the future size of the system for the next 10 months, and the predicted size looks very realistic, which indicates that the model is working well.

**Figure 21**

Title: RLR model evaluation (all features)

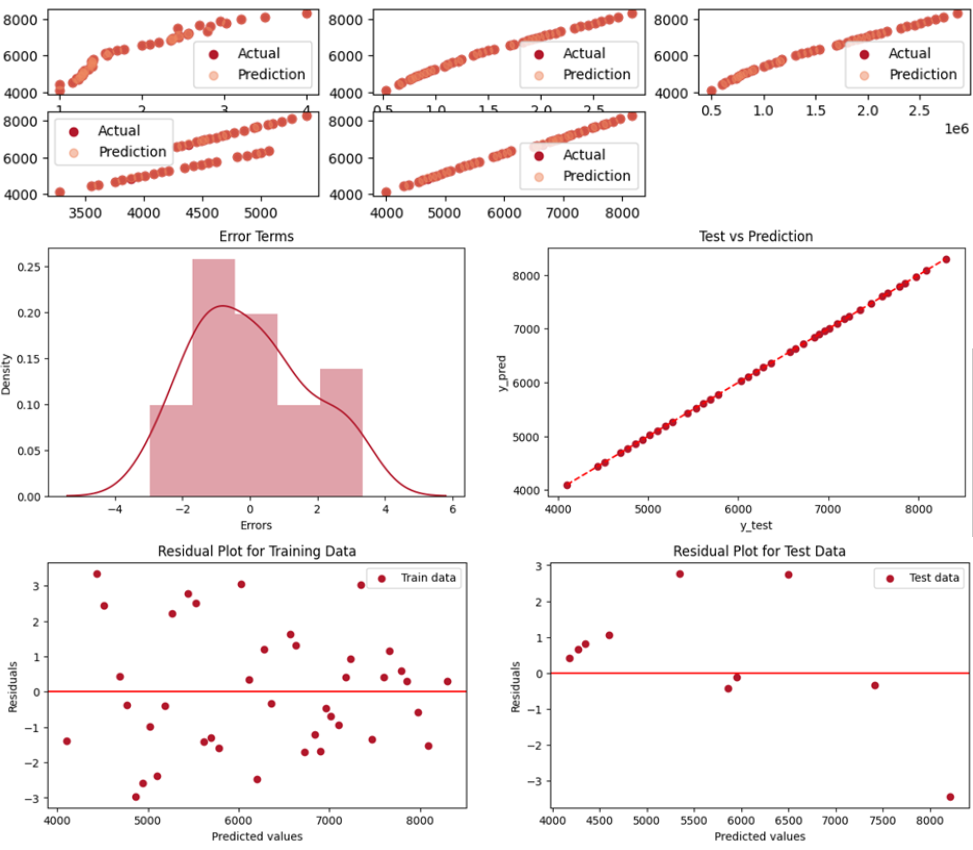


Annex A, Figure 21

Description: Comparison of predicted and actual values shows that the model's deviations are slightly more pronounced compared to the linear model but still closely align with real values. Model error estimation indicates minor skewness to the left, but errors are otherwise uniformly distributed. Residual plots show that the residuals are marginally concentrated in the positive direction of actual values but maintain a presence on both sides. Shapley values show that the largest contributions to this model were made by backup size, efficiency, and user count.

**Figure 22**

Title: RLR model evaluation (selected features)

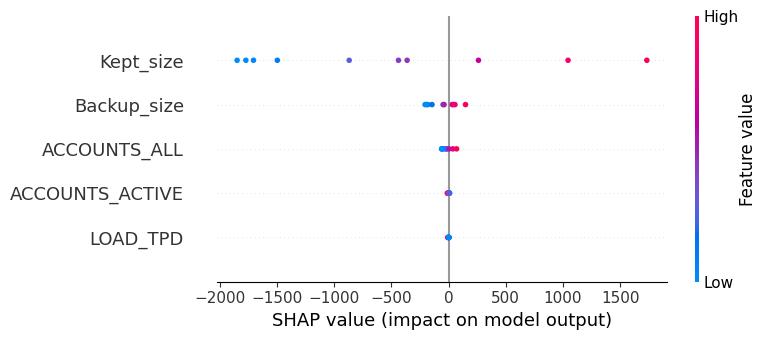


Annex A, Figure 22

Description: Comparison of predicted and actual values shows that predicted values closely match actual values, displaying high accuracy. Model error estimation shows a uniform distribution of errors and highlights the model's strong performance. Residual plots showcase scattered errors and indicate a lack of bias in the model, which is a result of successful training.

**Figure 23**

Title: Shapley Values for the RLR model (selected features)

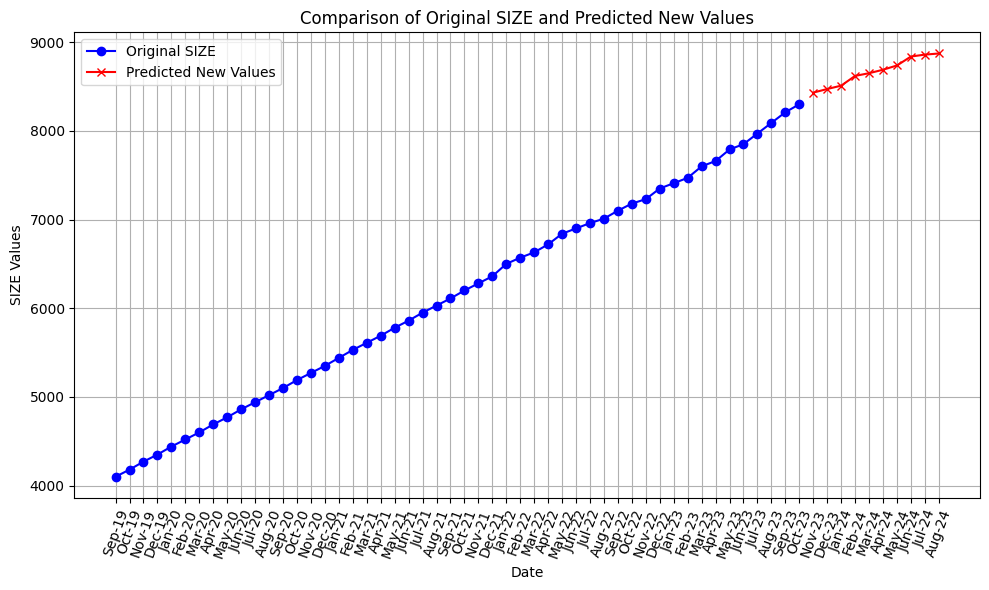


Annex A, Figure 23

Description: Shapley values show similar to previous observations, that all variables positively impact the model, with higher attribute values leading to increased predictions.

**Figure 24**

Title: Prediction of Size by RLR model

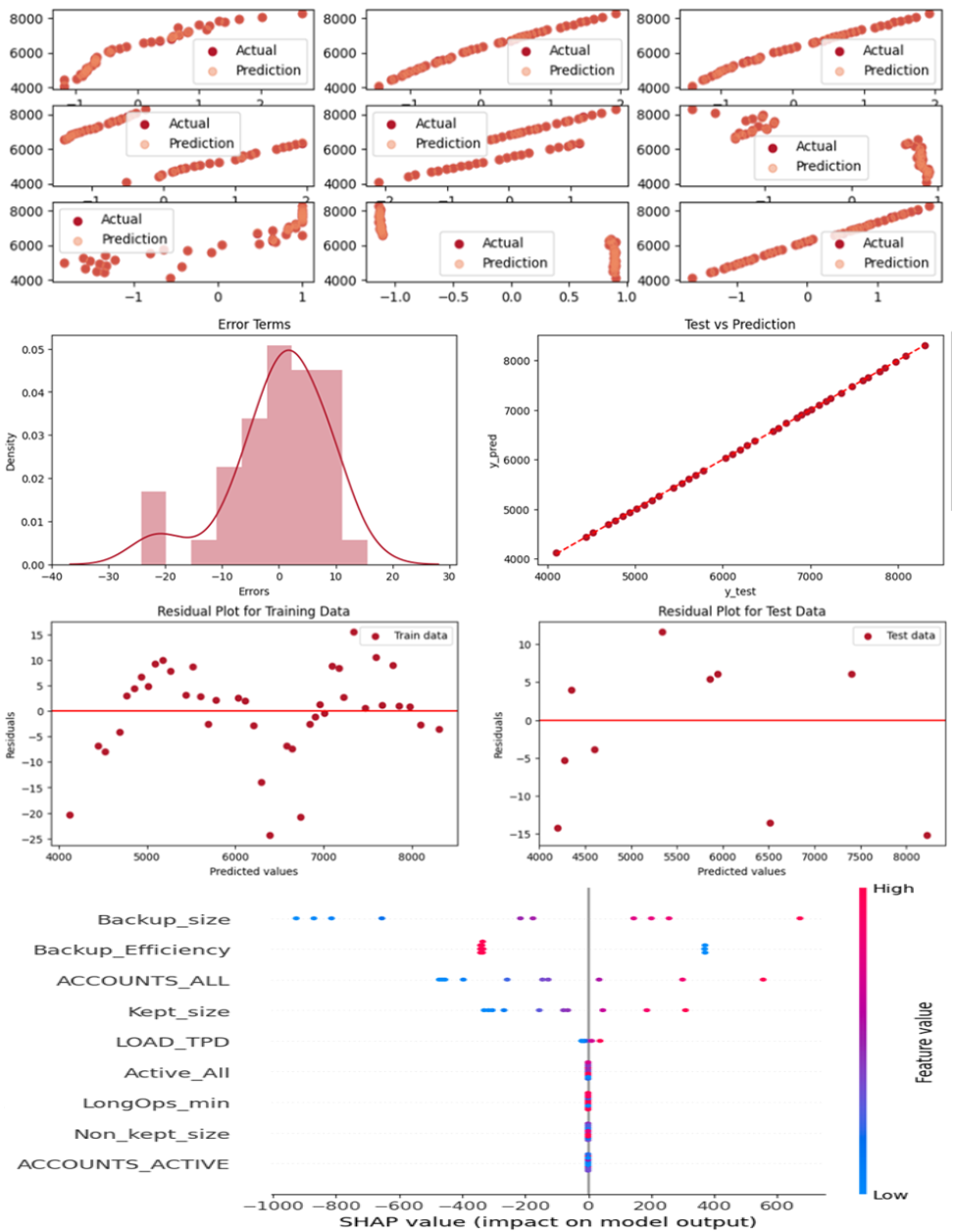


Annex A, Figure 24

Description: The graph shows that the predicted size looks very realistic, which indicates that the model is working well.

**Figure 25**

Title: LLR model evaluation (all features)

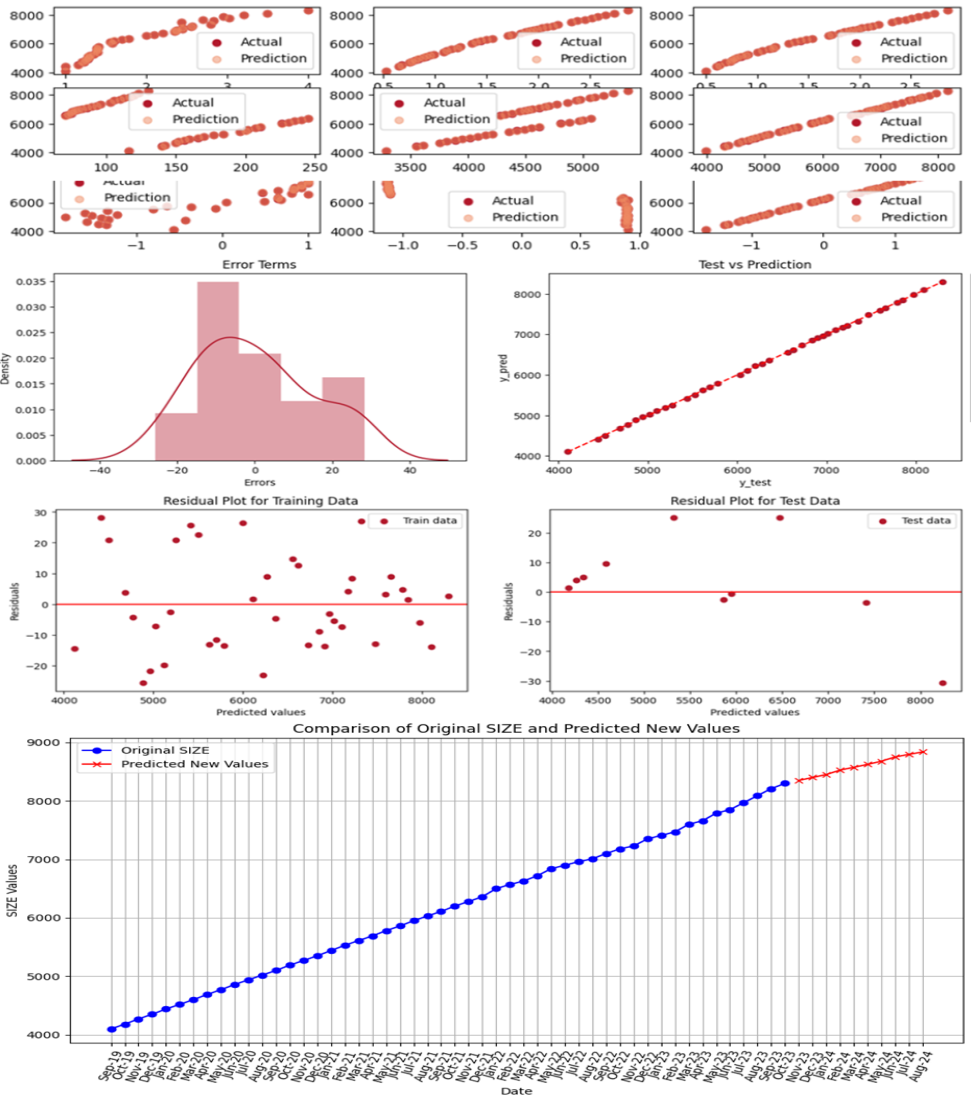


Annex A, Figure 25

Description: Comparison of predicted and actual values results indicate nearly identical values between predictions and actual data. Model error estimation indicates that the errors are uniformly distributed, with predicted values closely aligned with the actual values. Residual plots show that while slightly more values are seen on the positive side, deviations remain small.Shapley values similar to other models, indicate that most features positively influence the predicted values, except for backup efficiency.

**Figure 26**

Title: LLR model evaluation (selected features)

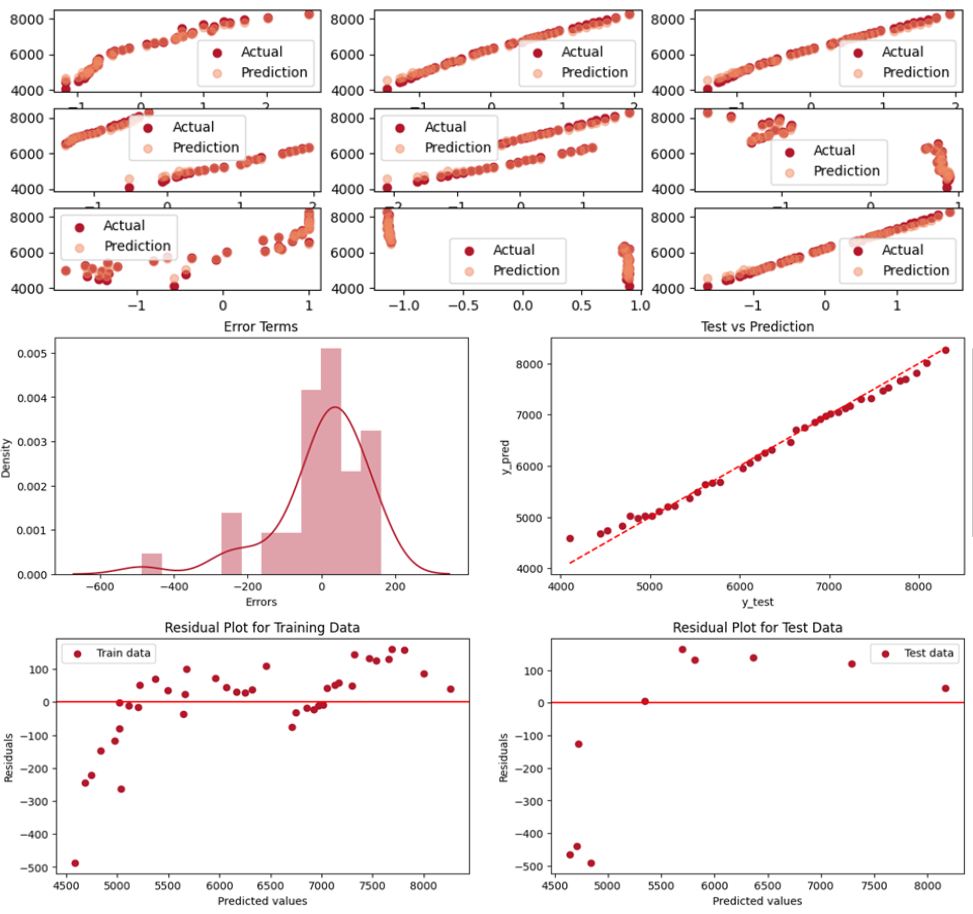


Annex A, Figure 26

Description: Comparison of predicted and actual values results depict a close match between predicted and actual values, maintaining high accuracy. Model error estimation shows that the errors are distributed nearly evenly, with predicted values lying close to the line, showcasing very small errors but slightly more visible than the previous model. Residual plots showcase scattered errors and indicate a lack of bias in the model, signalling successful training outcomes.Future size prediction shows that the model accurately predicts future system sizes based on 10 new values of independent variables, indicating efficient functionality.

**Figure 27**

Title: ENR model evaluation (all features)

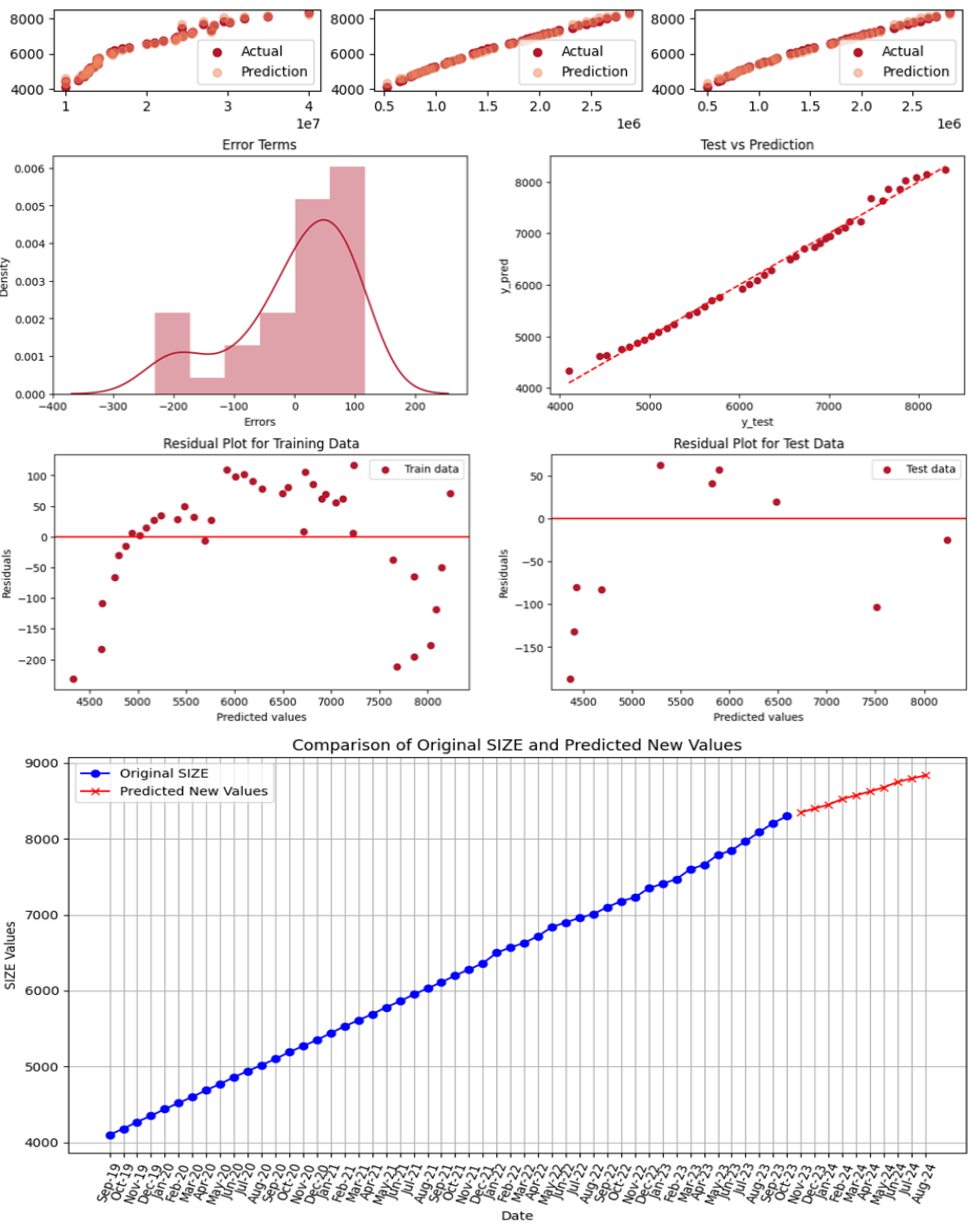


Annex A, Figure 27

Description: Comparison of predicted and actual values shows that deviations from actual values are notably more pronounced in this model compared to previous ones, indicating less favourable performance. Model error estimation indicates the error distribution exhibits a distinct tail, with noticeable deviations. Residual plots showcase errors that are concentrated above the actual values, accompanied by a pronounced tail, observed in both training and testing data.

**Figure 28**

Title: ENR model evaluation (selected features)

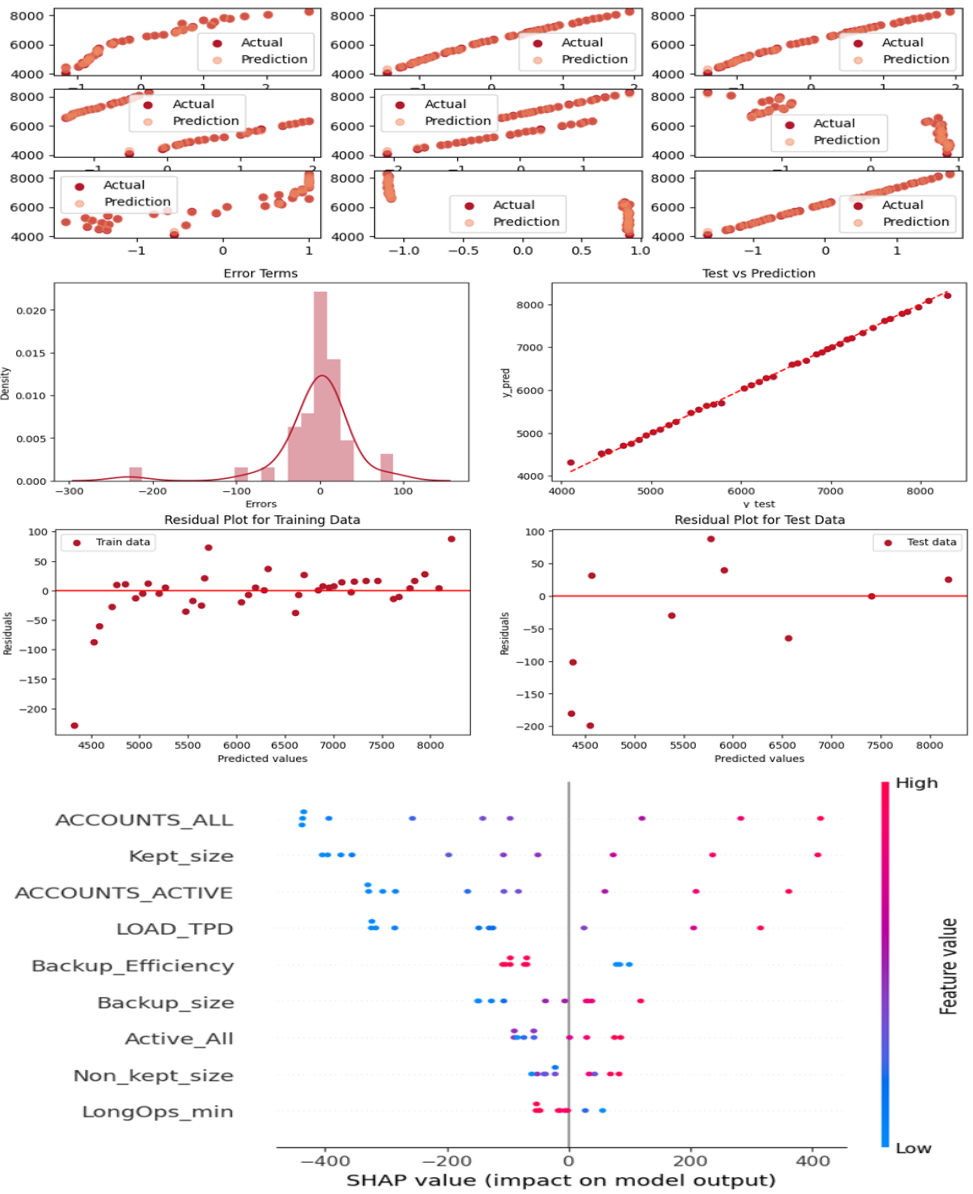


Annex A, Figure 28

Description: Comparison of predicted and actual values shows that improvements are evident in comparison to the model using all features, but deviations remain apparent. Model error estimation shows that though a tail on the left is observed in error distribution, it's less pronounced than the previous instance. Residual plots show that errors are still gathered above the actual values, but the spread is less pronounced. Future size prediction showcased model generating plausible predictions based on 10 new independent variables, indicating satisfactory performance.

**Figure 29**

Title: RFR model evaluation (all features)

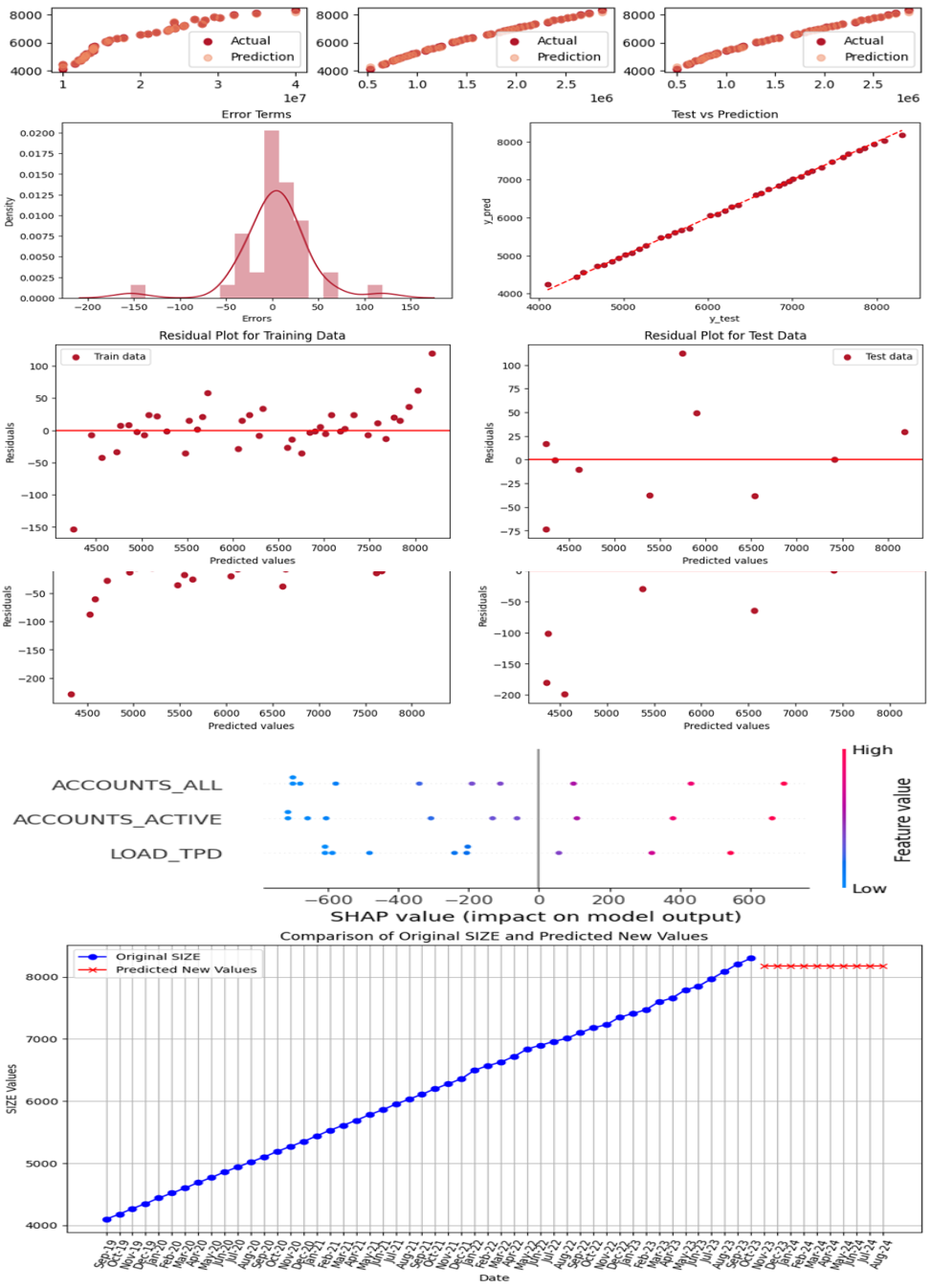


Annex A, Figure 29

Description: Comparison of predicted and actual values shows that the distinction between the predicted and actual values is almost undetectable. Model error estimation demonstrates a left-tail, similar to other models, with predicted variable values nearly aligned with slight deviations. Residual plots show that most errors are in proximity to the line, signifying favourable accuracy, albeit with an anomaly displaying a negative deviation of over -200. Shapley values show that this model assigns more weight to features compared to previous models, with a slightly altered significance order, distinctly varying from the prior models.

**Figure 30**

Title: RFR model evaluation (selected features)



Annex A, Figure 30

Description: Comparison of predicted and actual values demonstrates that predicted values closely mirror actual values, displaying minimal differences. Model error estimation shows a uniform distribution of errors signifies robust performance, with most errors falling between -50 and 50. Errors are scattered chaotically and close to zero, suggesting a low or negligible bias in this model. Shapley values show that selected variables contribute almost uniformly and positively to the model. Future size prediction showcased model generating unplausible predictions based on 10 new independent variables, indicating disastrous performance.

**Figure 31**

Title: KNN model evaluation (all features)

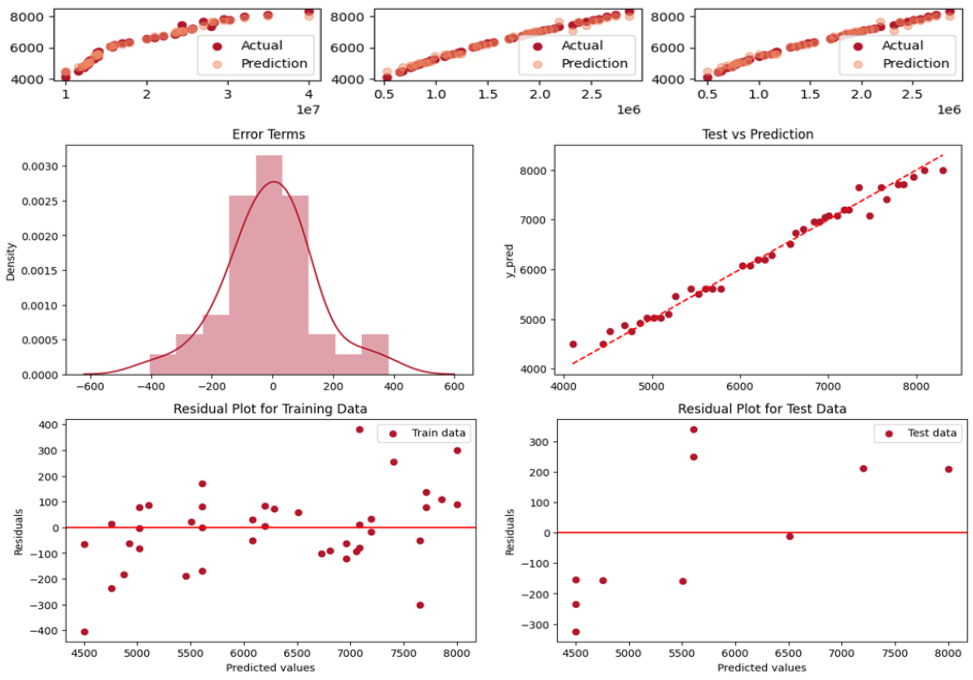


Annex A, Figure 31

Description: Comparison of predicted and actual values demonstrates that noticeable differences exist between the actual and predicted values, indicating poorer performance compared to other models. Model error estimation shows that though the error distribution is uniform, deviations from the line are more pronounced than in other models, suggesting less accurate predictions. Residual plots showcase the error distribution that appears chaotic, which is favourable. However, high errors contribute to a less favourable outcome.

**Figure 32**

Title: KNN model evaluation (selected features)

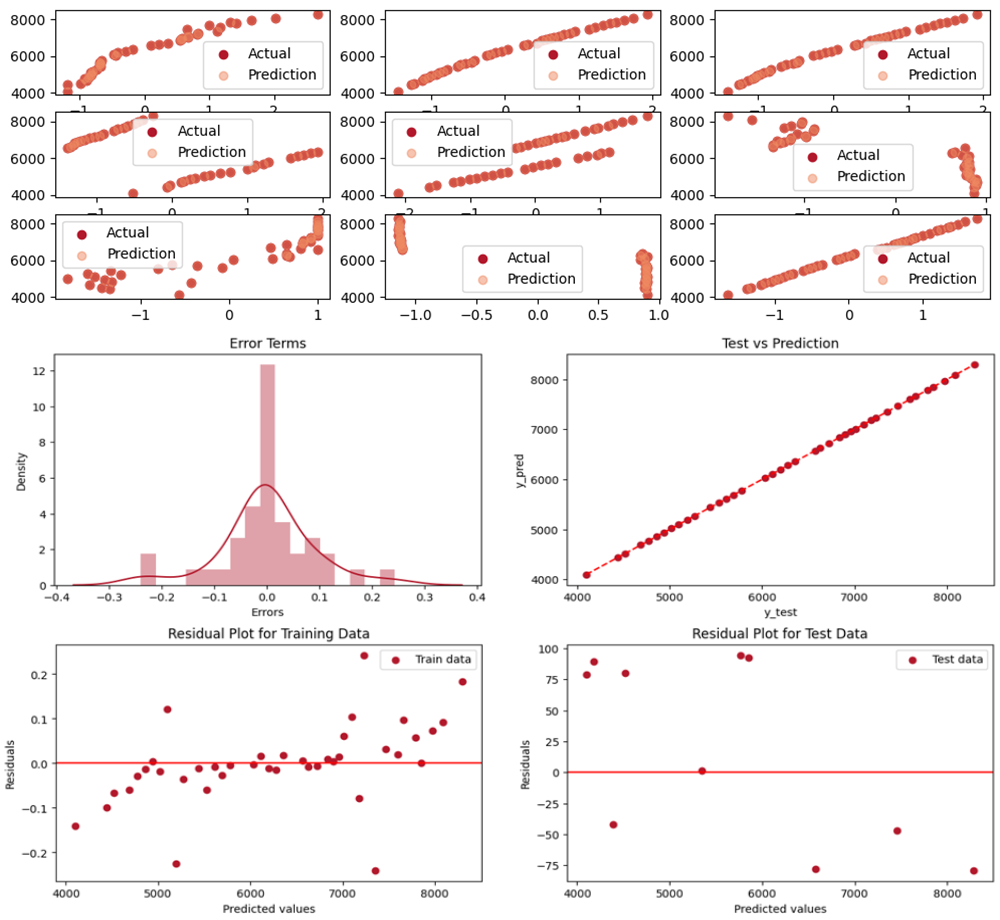


Annex A, Figure 32

Description: Comparison of predicted and actual values shows that significant differences between predicted and actual values are noticeable, indicating worse performance in comparison to other tested models. Model error estimation shows that despite a uniform error distribution, the model's accuracy remains low, evident from the considerable deviations from the line. Residual plots demonstrate that the errors are evenly distributed around zero, indicating minimal bias influence. However, these errors range significantly from -400 to 400, portraying lower accuracy in predictions.

**Figure 33**

Title: GBR model evaluation (all features)

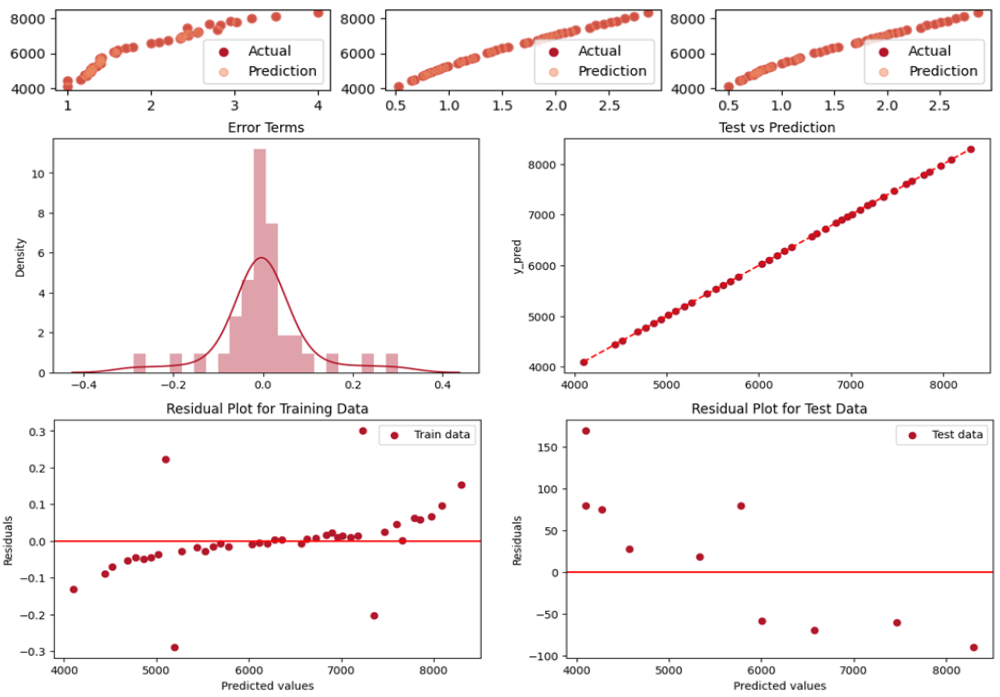


Annex A, Figure 33

Description: Comparison of predicted and actual values shows that the differences between predicted and actual values are minor and hardly noticeable. Model error estimation demonstrates uniform error distribution without any deviation from the line, indicating consistent predictions. Residual plots show that while there's a slight trend in errors on the training data, the errors overall are chaotic. Training set errors are minimal, but the model's performance is comparatively poorer on the testing set.

**Figure 34**

Title: GBR model evaluation (selected features)



Annex A, Figure 34

Description: Comparison of predicted and actual values shows that minor differences are observed between the predicted and actual values. Model error estimation demonstrates normal error distribution, and errors are minimal, indicating robust performance. Residual plots showcase a discernible trend in the error spread, implying bias in the model, which results in low error rates but may affect generalizability.

Bibliography

[1] ‘Addressing The Growing Storage Needs For Financial Services | StoneFly’. Accessed: Dec. 13, 2023. [Online]. Available: https://stonefly.com/blog/addressing-growing-storage-needs-financial-services/

[2] O. US EPA, ‘Exploratory Data Analysis’. Accessed: Dec. 13, 2023. [Online]. Available: https://www.epa.gov/caddis-vol4/exploratory-data-analysis

[3] ‘7-Steps Predictive Modeling Process’. Accessed: Dec. 13, 2023. [Online]. Available: https://rstudio-pubs-static.s3.amazonaws.com/223423\_8ca6fccca1e44939be3f85ecbfa9598f.html

[4] ‘pandas - Python Data Analysis Library’. Accessed: Dec. 13, 2023. [Online]. Available: https://pandas.pydata.org/about/

[5] ‘NumPy - About Us’. Accessed: Dec. 13, 2023. [Online]. Available: https://numpy.org/about/

[6] ‘History — Matplotlib 3.8.2 documentation’. Accessed: Dec. 13, 2023. [Online]. Available: https://matplotlib.org/stable/users/project/history.html

[7] M. Waskom, ‘seaborn: statistical data visualization’, *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.

[8] ‘6.3. Preprocessing data’, scikit-learn. Accessed: Dec. 13, 2023. [Online]. Available: https://scikit-learn/stable/modules/preprocessing.html

[9] ‘seaborn.heatmap — seaborn 0.13.0 documentation’. Accessed: Dec. 13, 2023. [Online]. Available: https://seaborn.pydata.org/generated/seaborn.heatmap.html

[10] A. Kraskov, H. Stögbauer, and P. Grassberger, ‘Estimating mutual information’, *Phys. Rev. E*, vol. 69, no. 6, p. 066138, Jun. 2004, doi: 10.1103/PhysRevE.69.066138.

[11] A. AK, ‘Clever Cuts: Uncovering the Power of SelectKBest for Feature Selection in Machine Learning’, Medium. Accessed: Dec. 13, 2023. [Online]. Available: https://medium.com/@abelkuriakose/clever-cuts-uncovering-the-power-of-selectkbest-for-feature-selection-in-machine-learning-c8d20d75c82f

[12] A. F. Siegel, ‘Chapter 15 - ANOVA: Testing for Differences Among Many Samples and Much More’, in *Practical Business Statistics (Seventh Edition)*, A. F. Siegel, Ed., Academic Press, 2016, pp. 469–492. doi: 10.1016/B978-0-12-804250-2.00015-8.

[13] J. Brownlee, ‘Recursive Feature Elimination (RFE) for Feature Selection in Python’, MachineLearningMastery.com. Accessed: Dec. 13, 2023. [Online]. Available: https://machinelearningmastery.com/rfe-feature-selection-in-python/

[14] T. K. Ho, ‘Random decision forests’, in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Aug. 1995, pp. 278–282 vol.1. doi: 10.1109/ICDAR.1995.598994.

[15] R. Tibshirani, ‘Regression Shrinkage and Selection via the Lasso’, *J. R. Stat. Soc. Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[16] A. E. Hoerl and R. W. Kennard, ‘Ridge Regression: Biased Estimation for Nonorthogonal Problems’, *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970, doi: 10.2307/1267351.

[17] ‘Elastic Net Regularization’, OpenGenus IQ: Computing Expertise & Legacy. Accessed: Dec. 13, 2023. [Online]. Available: https://iq.opengenus.org/elastic-net-regularization/

[18] ‘Variance Inflation Factor (VIF)’, Investopedia. Accessed: Dec. 13, 2023. [Online]. Available: https://www.investopedia.com/terms/v/variance-inflation-factor.asp

[19] J. D. Jobson, ‘Multiple Linear Regression’, in *Applied Multivariate Data Analysis: Regression and Experimental Design*, J. D. Jobson, Ed., in Springer Texts in Statistics. , New York, NY: Springer, 1991, pp. 219–398. doi: 10.1007/978-1-4612-0955-3\_4.

[20] ‘Nearest Neighbors regression’, scikit-learn. Accessed: Dec. 13, 2023. [Online]. Available: https://scikit-learn/stable/auto\_examples/neighbors/plot\_regression.html

[21] L. Mason, J. Baxter, P. Bartlett, and M. Frean, ‘Boosting Algorithms as Gradient Descent’, in *Advances in Neural Information Processing Systems*, MIT Press, 1999. Accessed: Dec. 13, 2023. [Online]. Available: https://proceedings.neurips.cc/paper\_files/paper/1999/hash/96a93ba89a5b5c6c226e49b88973f46e-Abstract.html