ML Raport

AutoPrep

January 12, 2025

Abstract

This raport has been generated with AutoPrep.

Contents

1 Overview

1.1 System

System	Windows
Machine	AMD64
Processor	Intel64 Family 6 Model 140 Stepping 2, GenuineIntel
Architecture	64bit
Python Version	3.11.5
Physical Cores	4
Logical Cores	8
CPU Frequency (MHz)	2918.00
Total RAM (GB)	15.74
Available RAM (GB)	4.27
Total Disk Space (GB)	457.28
Free Disk Space (GB)	239.79

Table 1: System overview.

1.2 Dataset

Task detected for the dataset: regression.
Table 2 presents an overview of the dataset including the number of samples, features, and their types.

Number of samples	227
Number of features	9
Number of numerical features	9
Number of categorical features	0

Table 2: Dataset Summary.

Table 4 presents the distribution of missing values in the dataset.

feature	number of observations	fraction
P85	0	0.00
P75	0	0.00
RMT85	0	0.00
CS82	0	0.00
SS82	0	0.00
S82	0	0.00
ME84	0	0.00
REV84	0	0.00
REG	0	0.00

Table 3: Missing values distribution.

Table 5 presents the description of features in the dataset.

feature	type	dtype	space usage
P85	numerical	int64	3.6 kB
P75	numerical	int64	3.6 kB
RMT85	numerical	int64	3.6 kB
CS82	numerical	uint8	2.0 kB
SS82	numerical	uint8	2.0 kB
S82	numerical	uint8	2.0 kB
ME84	numerical	int64	3.6 kB
REV84	numerical	int64	3.6 kB
REG	numerical	uint8	2.0 kB

Table 4: Features dtypes description.

Table 6 presents the description of numerical features in the dataset.

index	count	mean	std	min	25%	50%	75%	max
P85	227.00	29.99	56.17	3.00	10.00	16.00	30.00	653.00
P75	227.00	29.52	57.77	4.00	10.00	15.00	28.00	671.00
RMT85	227.00	254.51	657.60	21.00	66.50	118.00	229.50	6720.00
CS82	227.00	9.18	4.98	1.00	6.00	8.00	11.00	34.00
SS82	227.00	21.95	7.23	8.00	17.00	21.00	27.00	46.00
S82	227.00	47.15	10.57	31.00	41.00	45.00	49.00	101.00
ME84	227.00	1842.41	4685.06	173.00	480.50	839.00	1580.50	47074.00
REV84	227.00	3048.31	5125.17	347.00	1134.50	1828.00	3174.00	59877.00
REG	227.00	4.33	2.08	1.00	2.00	4.00	6.00	8.00

Table 5: Numerical features description.

2 Eda

This part of the report provides basic insides to the data and the informations it holds..

2.1 Target variable and missing values

Figure 1 shows the distribution of the target variable.

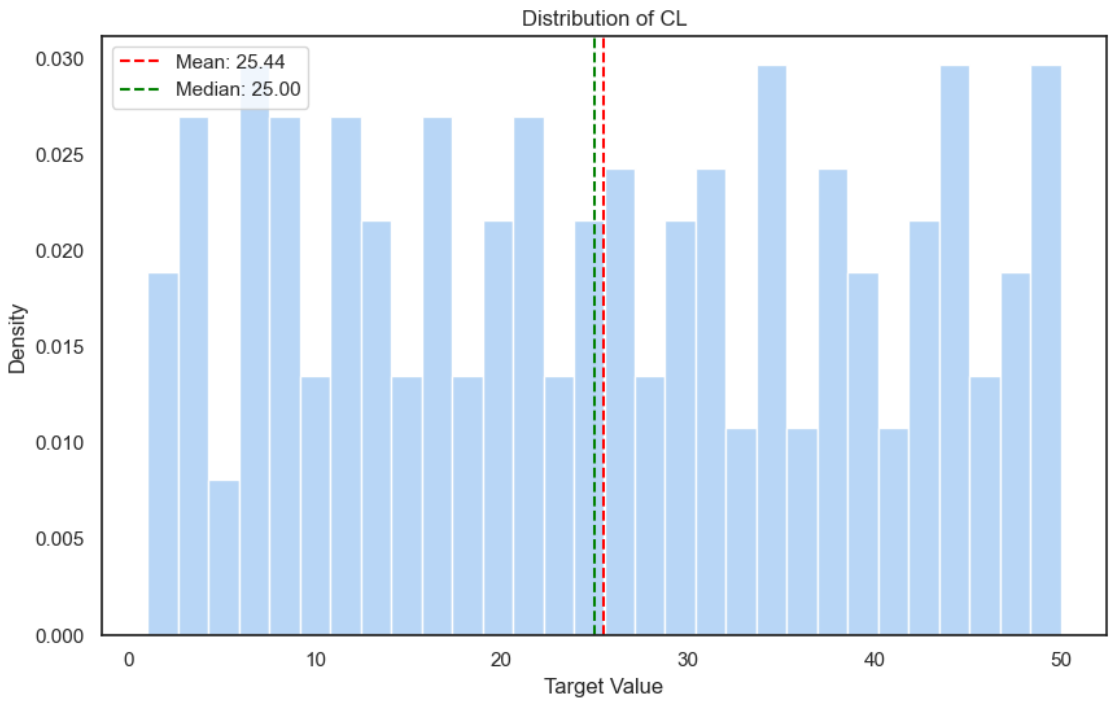

Figure 1: Target distribution.

2.2 EDA for numerical features

The distribution of numerical features is presented on histogram(s) below.

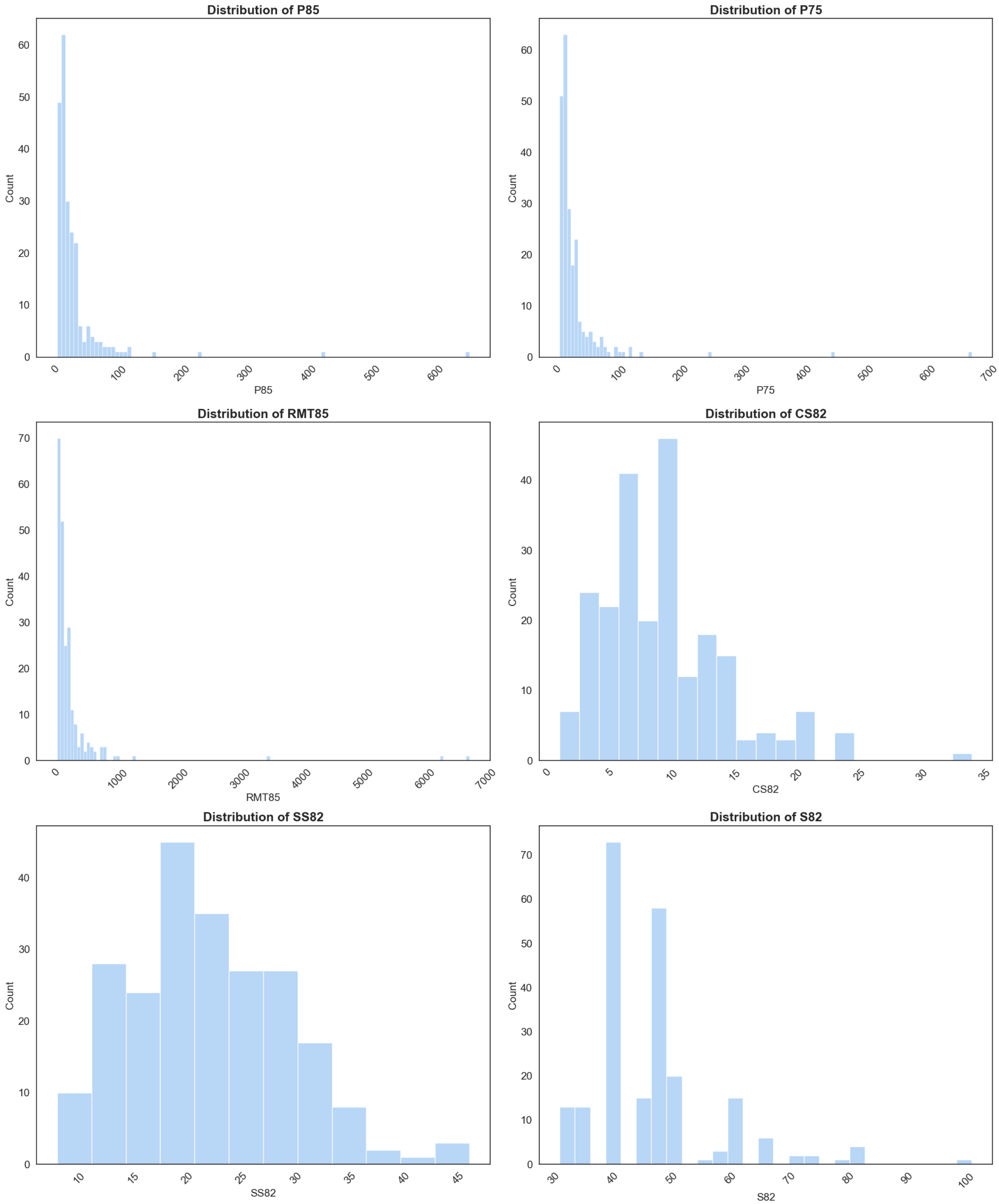

Figure 2: Numerical Features Distribution - Page 1

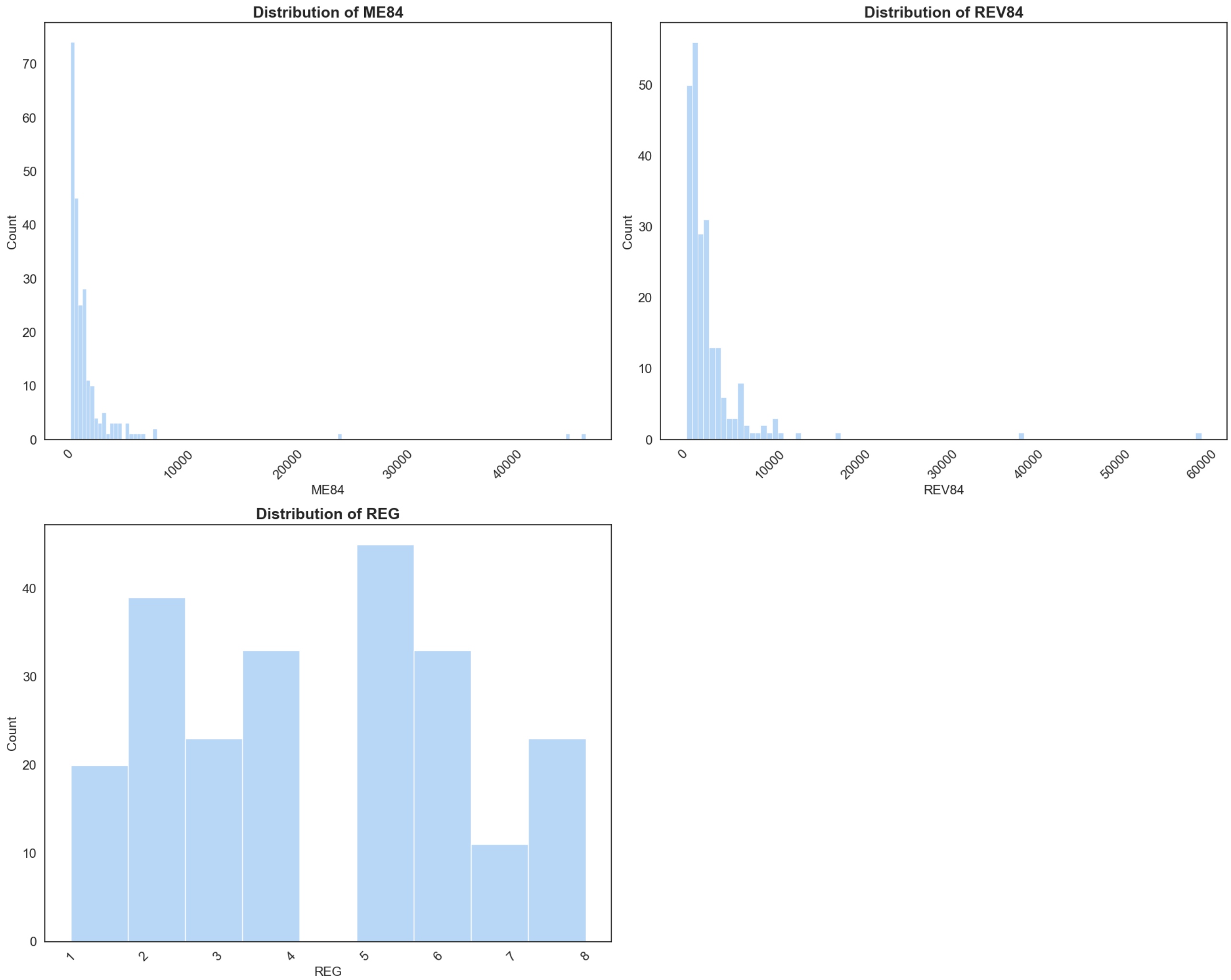

Figure 3: Numerical Features Distribution - Page 2

Figure 5 shows the correlation between features.

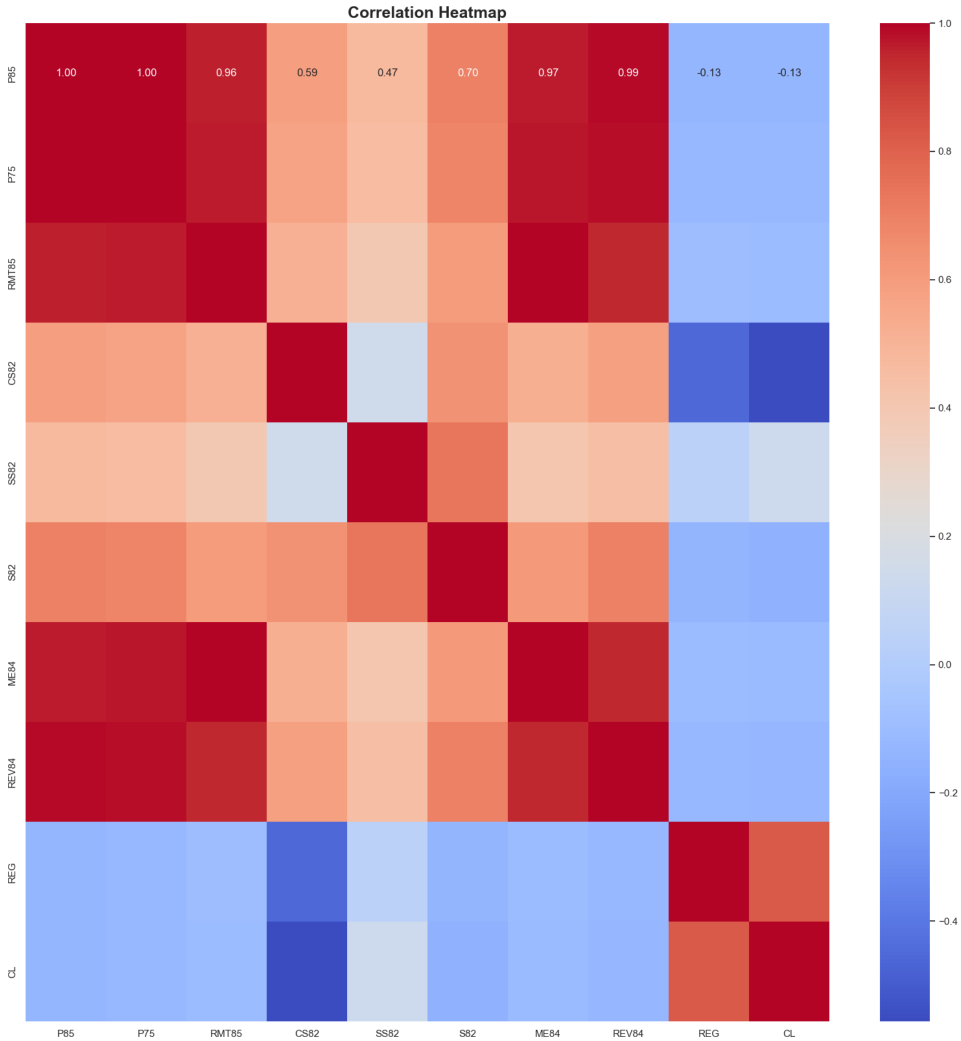

Figure 4: Correlation heatmap.

The boxplot of numerical features is presented on chart(s) below.

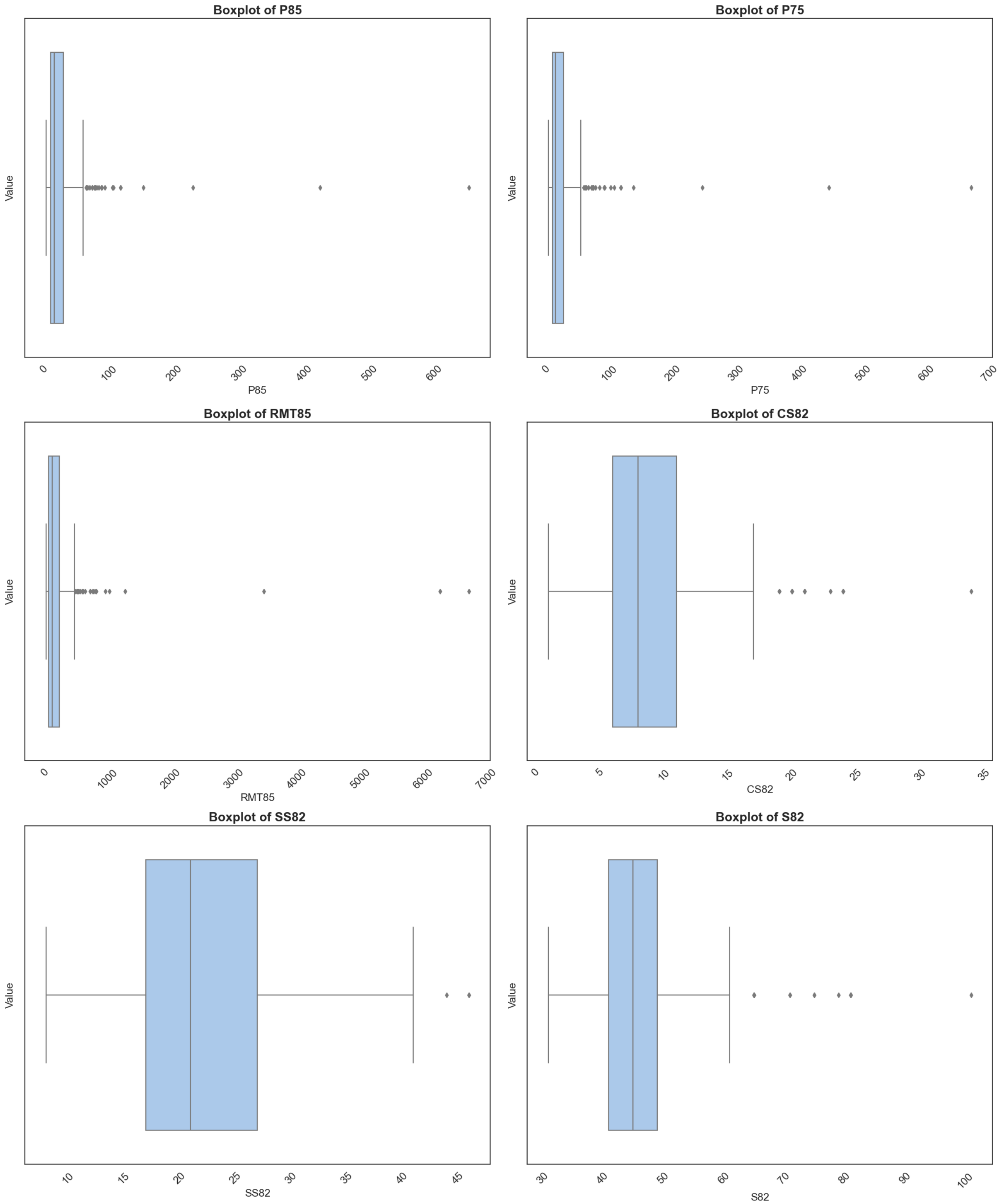

Figure 5: Boxplot page 1

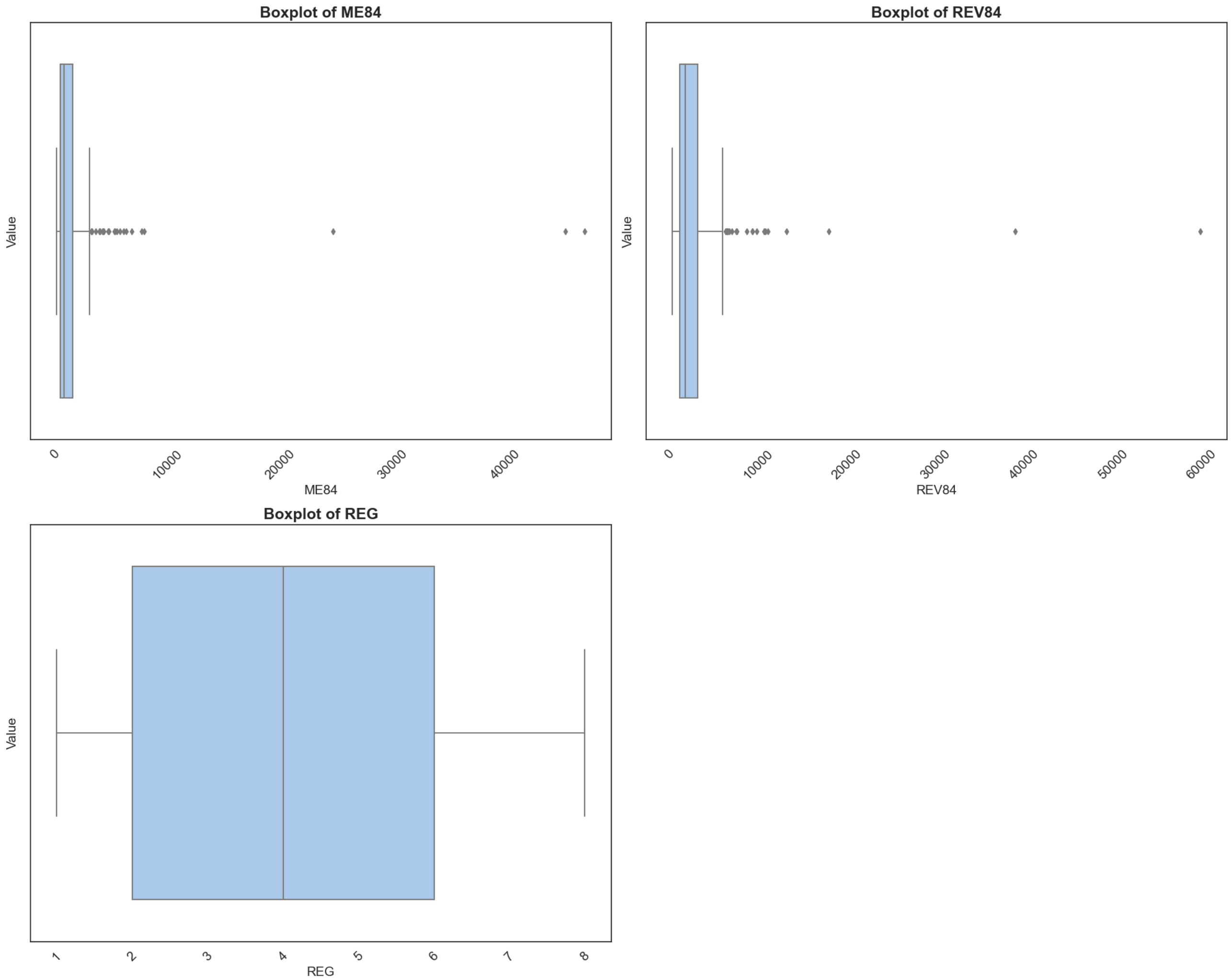

Figure 6: Boxplot page 2

3 Preprocessing

This part of the report presents the results of the preprocessing process. It contains required, as well as non required, steps listed below.
Required preprocessing steps

- Missing data imputation
- Removing columns with 100% unique categorical values
- Categorical features encoding
- Scaling
- Removing columns with 0 variance
- Detecting highly correlatd features

Additional preprocessing steps

- Feature selection methods : Correlation with the target or Random Forest feature importance
- Dimention reduction techniques: PCA, VIF, UMAP

Preprocessing process was configured to select up to 3 best unique preprocessing pipelines. Pipelines were scored based on a simple model. Tables below show detailed description of the best pipelines as well as all step combinations that were examined.

index	steps
0	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler
1	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, CorrelationSelector
2	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, FeatureImportanceClassSelector
3	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, FeatureImportanceRegressSelector
4	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, PCADimentionReducer
5	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, CorrelationSelector, PCADimentionReducer
6	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, FeatureImportanceClassSelector, PCADimentionReducer
7	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, FeatureImportanceRegressSelector, PCADimentionReducer
8	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, UMAPDimentionReducer
9	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, CorrelationSelector, UMAPDimentionReducer
10	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, FeatureImportanceClassSelector, UMAPDimentionReducer
11	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, FeatureImportanceRegressSelector, UMAPDimentionReducer
12	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, VIFDimentionReducer
13	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, CorrelationSelector, VIFDimentionReducer
14	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, FeatureImportanceClassSelector, VIFDimentionReducer
15	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler, FeatureImportanceRegressSelector, VIFDimentionReducer

Table 6: Pipelines steps overview.

index	file name	score	fit duration	score duration
0	preprocessing_pipeline_0.joblib	200.45	16 seconds	17 seconds
1	preprocessing_pipeline_1.joblib	192.80	a moment	a moment
2	preprocessing_pipeline_2.joblib	192.80	a moment	a moment

Table 7: Best preprocessing pipelines.

step	name	description	params
0	NAImputer	Imputes missing data.	{"numeric_imputer": "median", "categorical_imputer": "most_frequent"}
1	UniqueFilter	Removes categorical columns with 100% unique values. Dropped columns: []	{}
2	ColumnEncoder	Encodes categorical columns using OneHotEncoder (for columns with <5 unique values) or TolerantLabelEncoder (for columns with >=5 unique values). Encodes target variable using LabelEncoder if provided.	{}
3	VarianceFilter	Removes columns with zero variance. Dropped columns: []	{}
4	CorrelationFilter	Removes one column from pairs of columns correlated above correlation threshold: 0.8.	{}
5	ColumnScaler	Scales numerical columns using one of 3 scaling methods.	{"method": "standard"}
6	FeatureImportanceClassSelector	Selects the top 10.0% (rounded to whole number) of features most important according to Random Forest model for classification. Number of features that were selected: 0	{"k": 10.0}
7	UMAPDimentionReducer	Reduces the dimensionality of the data using UMAP.	{"n_components": null}

Table 8: Best pipeline No. 0: steps overview.

index	count	mean	std	min	25%	50%	75%	max
P85	227.00	-0.00	1.00	-0.48	-0.36	-0.25	0.00	11.12
CS82	227.00	0.00	1.00	-1.64	-0.64	-0.24	0.37	4.99
SS82	227.00	0.00	1.00	-1.93	-0.69	-0.13	0.70	3.33
S82	227.00	-0.00	1.00	-1.53	-0.58	-0.20	0.18	5.11
REG	227.00	-0.00	1.00	-1.60	-1.12	-0.16	0.80	1.77

Table 9: Best pipeline No. 0: output overview.

step	name	description	params
0	NAImputer	Imputes missing data.	{"numeric_imputer": "median", "categorical_imputer": "most_frequent"}
1	UniqueFilter	Removes categorical columns with 100% unique values. Dropped columns: []	{}
2	ColumnEncoder	Encodes categorical columns using OneHotEncoder (for columns with <5 unique values) or TolerantLabelEncoder (for columns with >=5 unique values). Encodes target variable using LabelEncoder if provided.	{}
3	VarianceFilter	Removes columns with zero variance. Dropped columns: []	{}
4	CorrelationFilter	Removes one column from pairs of columns correlated above correlation threshold: 0.8.	{}
5	ColumnScaler	Scales numerical columns using one of 3 scaling methods.	{"method": "minmax"}
6	FeatureImportanceClassSelector	Selects the top 10.0% (rounded to whole number) of features most important according to Random Forest model for classification. Number of features that were selected: 0	{"k": 10.0}
7	PCADimentionReducer	Combines PCA with automatic selection of the number of components to preserve 95% of the variance.	{"n_components": null}

Table 10: Best pipeline No. 1: steps overview.

index	count	mean	std	min	25%	50%	75%	max
P85	227.00	0.04	0.09	0.00	0.01	0.02	0.04	1.00
CS82	227.00	0.25	0.15	0.00	0.15	0.21	0.30	1.00
SS82	227.00	0.37	0.19	0.00	0.24	0.34	0.50	1.00
S82	227.00	0.23	0.15	0.00	0.14	0.20	0.26	1.00
REG	227.00	0.48	0.30	0.00	0.14	0.43	0.71	1.00

Table 11: Best pipeline No. 1: output overview.

step	name	description	params
0	NAImputer	Imputes missing data.	{"numeric_imputer": "median", "categorical_imputer": "most_frequent"}
1	UniqueFilter	Removes categorical columns with 100% unique values. Dropped columns: []	{}
2	ColumnEncoder	Encodes categorical columns using OneHotEncoder (for columns with <5 unique values) or TolerantLabelEncoder (for columns with >=5 unique values). Encodes target variable using LabelEncoder if provided.	{}
3	VarianceFilter	Removes columns with zero variance. Dropped columns: []	{}
4	CorrelationFilter	Removes one column from pairs of columns correlated above correlation threshold: 0.8.	{}
5	ColumnScaler	Scales numerical columns using one of 3 scaling methods.	{"method": "minmax"}
6	FeatureImportanceRegressSelector	Selects the top 10.0% (rounded to whole number) of features most important according to Random Forest model for regression. Number of features that were selected: 0	{"k": 10.0}
7	PCADimentionReducer	Combines PCA with automatic selection of the number of components to preserve 95% of the variance.	{"n_components": null}

Table 12: Best pipeline No. 2: steps overview.

index	count	mean	std	min	25%	50%	75%	max
P85	227.00	0.70	2.81	-0.65	-0.30	0.00	0.70	31.85
CS82	227.00	0.24	1.00	-1.40	-0.40	0.00	0.60	5.20
SS82	227.00	0.10	0.72	-1.30	-0.40	0.00	0.60	2.50
S82	227.00	0.27	1.32	-1.75	-0.50	0.00	0.50	7.00
REG	227.00	0.08	0.52	-0.75	-0.50	0.00	0.50	1.00

Table 13: Best pipeline No. 2: output overview.

You may also find all pipelines' runtime statistic in Table 16

Category	Value
Unique created pipelines	16
All created pipelines (after exploading each step params)	48
All pipelines fit time	29 seconds
All pipelines score time	37 seconds
scores_count	48.00
scores_mean	116.57
scores_std	72.67
scores_min	23.88
scores_25%	33.56
scores_50%	135.79
scores_75%	186.58
scores_max	200.45
Scoring function	<class 'str'>
Scoring model	RandomForestRegressor

Table 14: Preprocessing pipelines runtime statistics.

4 Modeling

4.1 Overview

This part of the report presents the results of the modeling process. There were 6 regression models trained for each of the best preprocessing pipelines.
The following models were used in the modeling process.

- BayesianRidge
- GradientBoostingRegressor
- KNeighborsRegressor
- LinearRegression
- LinearSVR
- RandomForestRegressor

4.2 Hyperparameter tuning

This section presents the results of hyperparameter tuning for each of the best 3 models using RandomizedSearchCV. Param grids used for each model are presented in the tables below.

Category	Value
max_iter	[300, 400, 500]
tol	[0.001, 0.0001, 1e-05]
alpha_1	[1e-06, 1e-07, 1e-08]
alpha_2	[1e-06, 1e-07, 1e-08]
lambda_1	[1e-06, 1e-07, 1e-08]
lambda_2	[1e-06, 1e-07, 1e-08]

Table 15: Param grid for model BayesianRidgeRegressor.

Category	Value
n_estimators	[100, 200, 300]
learning_rate	[0.1, 0.05, 0.02]
max_depth	[4, 6, 8]
min_samples_split	[2, 5, 10]
min_samples_leaf	[1, 2, 4]
subsample	[1.0, 0.5]
random_state	[42]

Table 16: Param grid for model GradientBoostingRegressor.

Category	Value
n_neighbors	[5, 10, 15]
weights	['uniform', 'distance']
algorithm	['auto', 'ball_tree', 'kd_tree', 'brute']
leaf_size	[30, 40, 50]
p	[1, 2]

Table 17: Param grid for model KNeighboursRegressor.

Category	Value
fit_intercept	[True, False]

Table 18: Param grid for model LinearRegression.

Category	Value
epsilon	[0.0, 0.1, 0.2, 0.5, 1.0]
C	[0.1, 1.0, 10.0, 100.0]
loss	['epsilon_insensitive', 'squared_epsilon_insensitive']
fit_intercept	[True, False]

Table 19: Param grid for model LinearSVR.

Category	Value
n_estimators	[100, 200, 300]
max_depth	[None, 5, 10, 15, 20]
min_samples_split	[2, 5, 10]
min_samples_leaf	[1, 2, 4]
max_features	['sqrt', 'log2', None]
bootstrap	[True, False]
random_state	[42]

Table 20: Param grid for model RandomForestRegressor.

Table 22 presents the best models and pipelines along with their hyperparameters, mean fit time, and test score.

Model	Pipeline	Best params	Mean fit time	Test score
BayesianRidge	final_pipeline_1.joblib	{"tol": 0.0001, "max_iter": 400, "lambda_2": 1e-07, "lambda_1": 1e-08, "alpha_2": 1e-07, "alpha_1": 1e-06}	a moment	234.22
BayesianRidge	final_pipeline_2.joblib	{"tol": 1e-05, "max_iter": 400, "lambda_2": 1e-06, "lambda_1": 1e-08, "alpha_2": 1e-06, "alpha_1": 1e-07}	a moment	234.25
BayesianRidge	final_pipeline_0.joblib	{"tol": 0.001, "max_iter": 300, "lambda_2": 1e-07, "lambda_1": 1e-08, "alpha_2": 1e-08, "alpha_1": 1e-07}	a moment	234.32

Table 21: Best models results

4.3 Interpretability

This section presents SHAP plots for the best model.

Figure 7: SHAP bar plot.

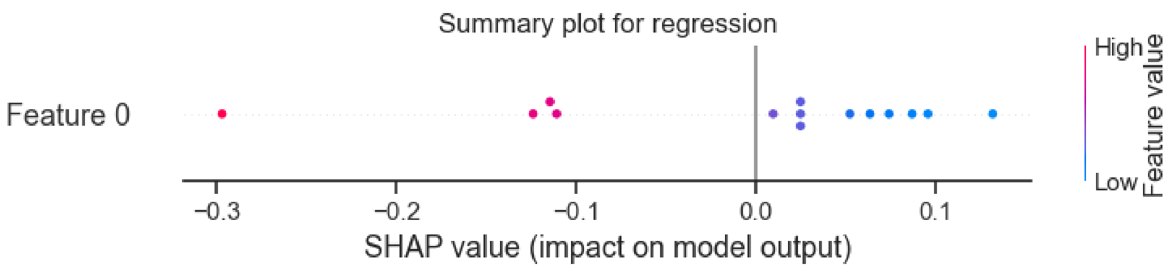

Figure 8: SHAP summary plot.

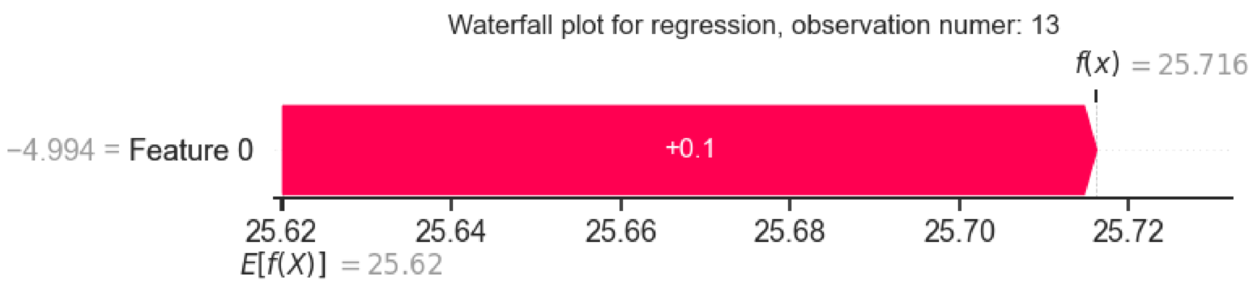

Figure 9: SHAP waterfall plot.