TÉCNICO LISBOA

DS 2025

Data Science Project

Team nr: 16	**Student 1:**	Antero Morgado	**IST nr:**	1119213
	Student 2:	David Ferreira	**IST nr:**	1107077
	Student 3:	José Fernandes	**IST nr:**	1103727
	Student 4:	Olha Buts	**IST nr:**	1116276

The present document presents a template for the Data Science Project report. It specifies the mandatory format and suggests the structure to follow. All text with grey background shall be replaced with the analysis made over the datasets. Put your charts in the *images* folder, and set the name of the file in the *includegraphics* command, after uncommenting it.

CLASSIFICATION

1 DATA PROFILING

May be used to describe any useful observation about the data, and that was used in the current project. An example is the use of any domain knowledge to process the data or evaluate the results. **Shall not exceed 200 characters.**

Data Dimensionality

Shall contain all relevant information and charts respecting to the data dimensionality perspective, such as the number of records and number of dimensions, and their impact on the following analysis. **Shall not exceed 500 characters.**

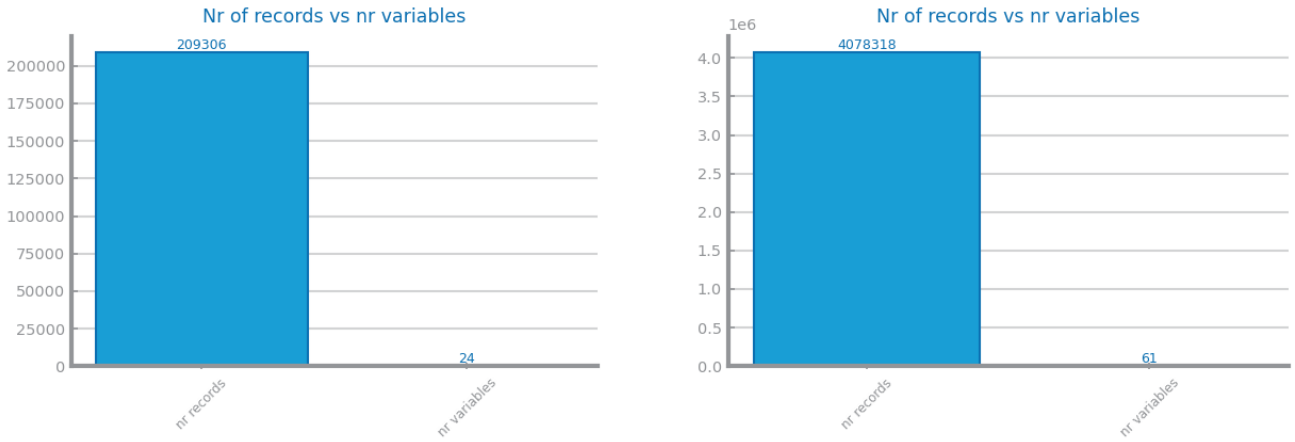


Figure 1: Nr Records x Nr variables for dataset 1 (left) and dataset 2 (right)

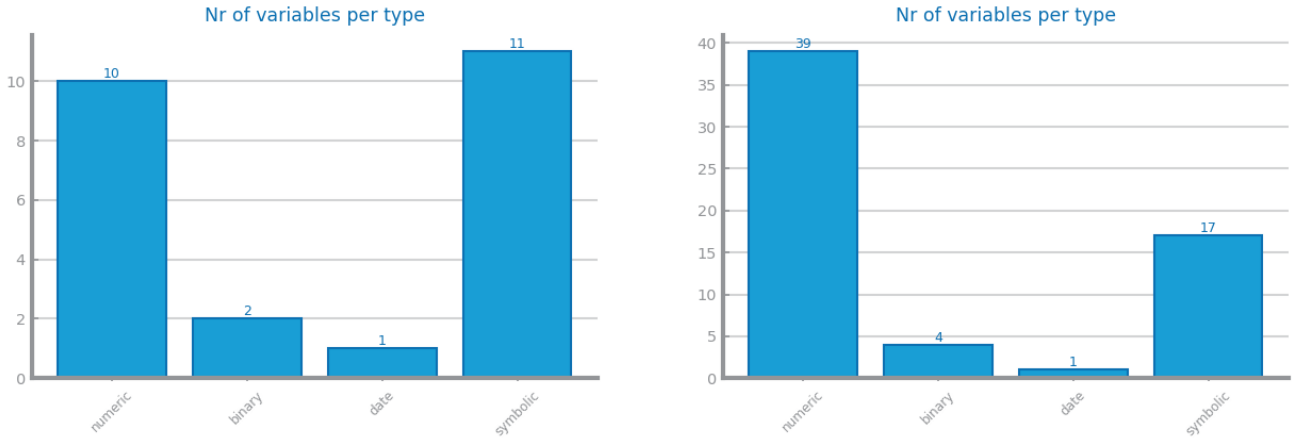


Figure 2: Nr variables per type for dataset 1 (left) and dataset 2 (right)

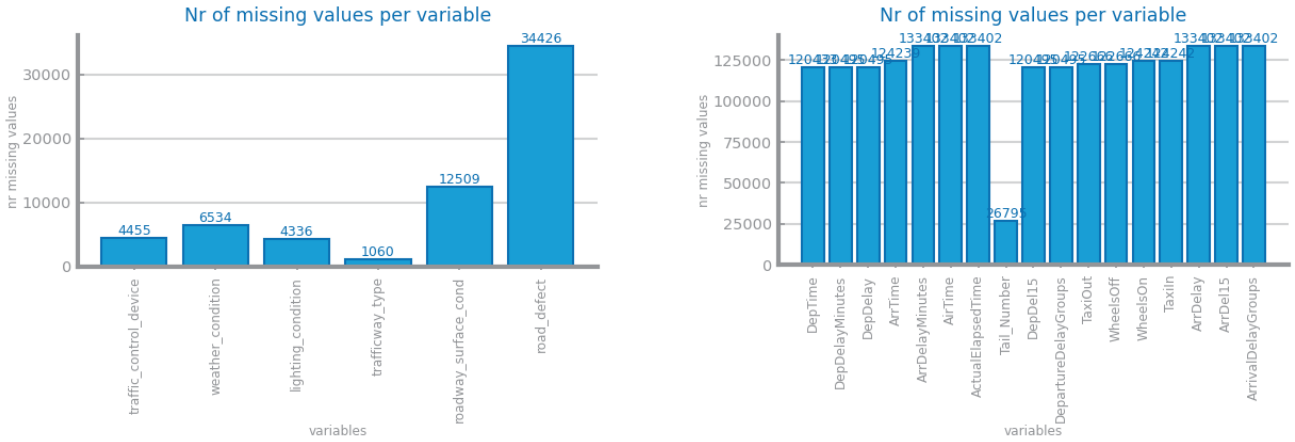


Figure 3: Nr missing values for dataset 1 (left) and dataset 2 (right)

Data Distribution

Shall contain all relevant information and charts respecting to the data distribution perspective, such as each variable distribution, type, domain and range. May be used to describe any useful observation about the data, and that was used in the current project. **Shall not exceed 500 characters.**

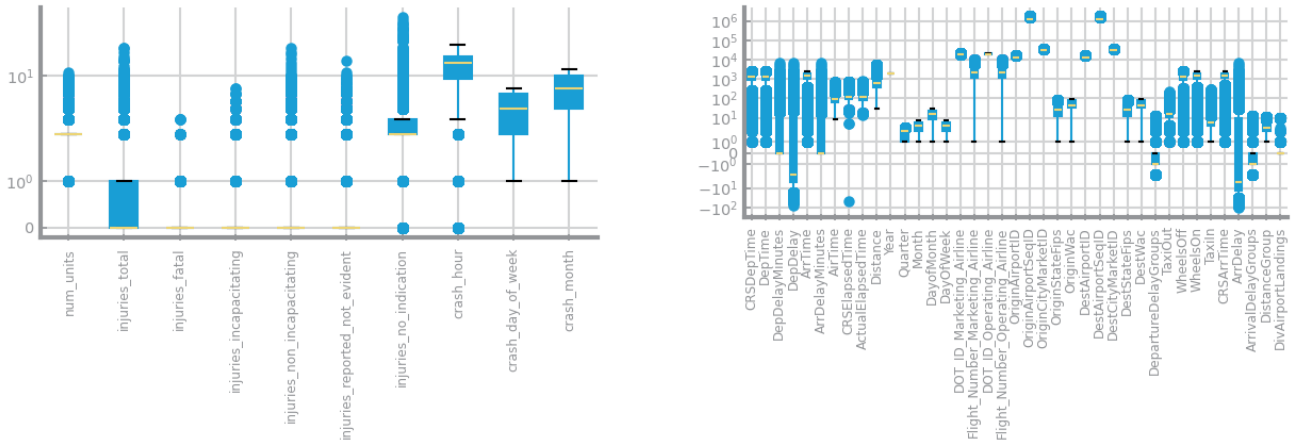


Figure 4: Global boxplots dataset 1 (left) and dataset 2 (right)

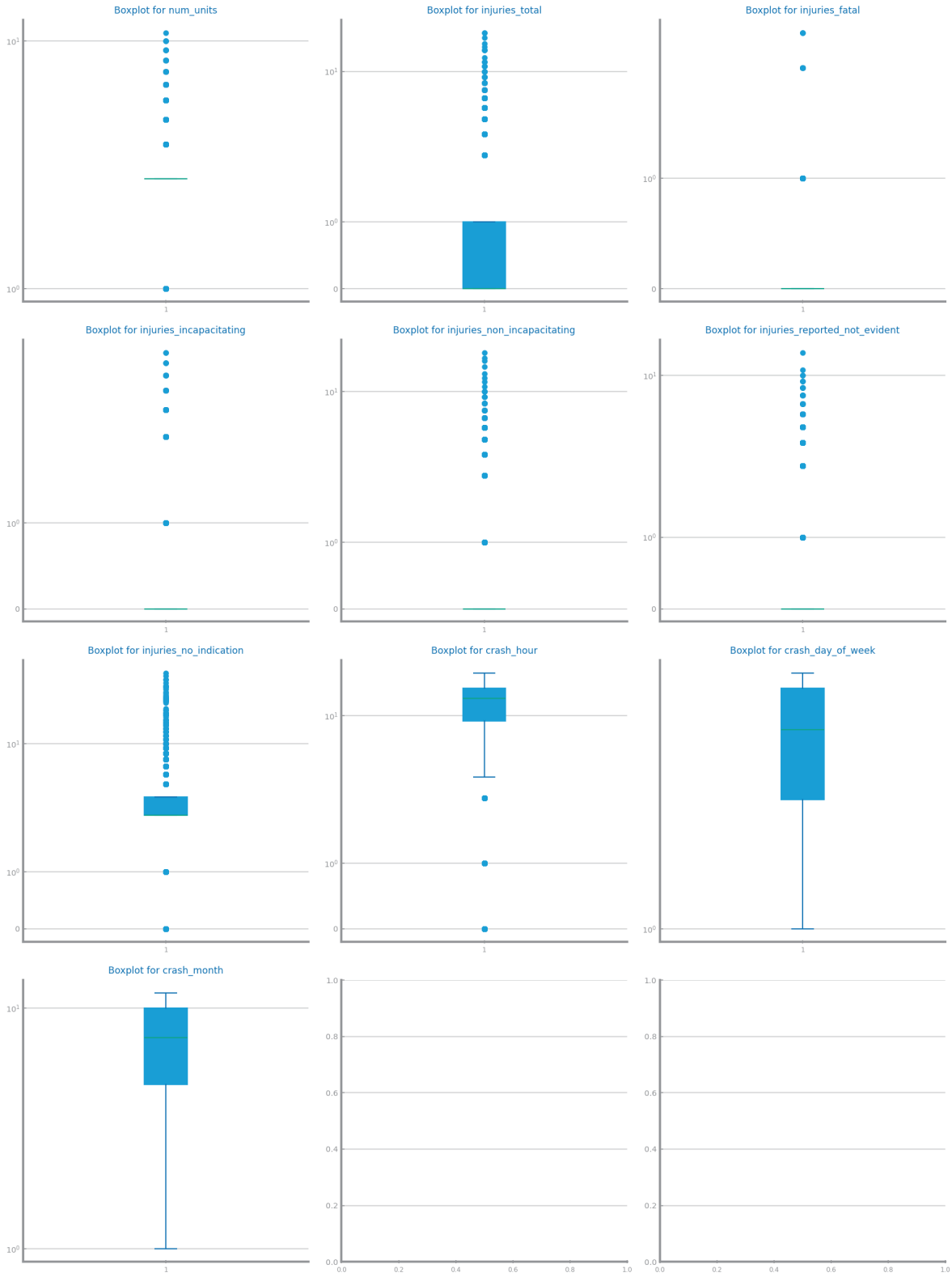


Figure 5: Single variables boxplots for dataset 1

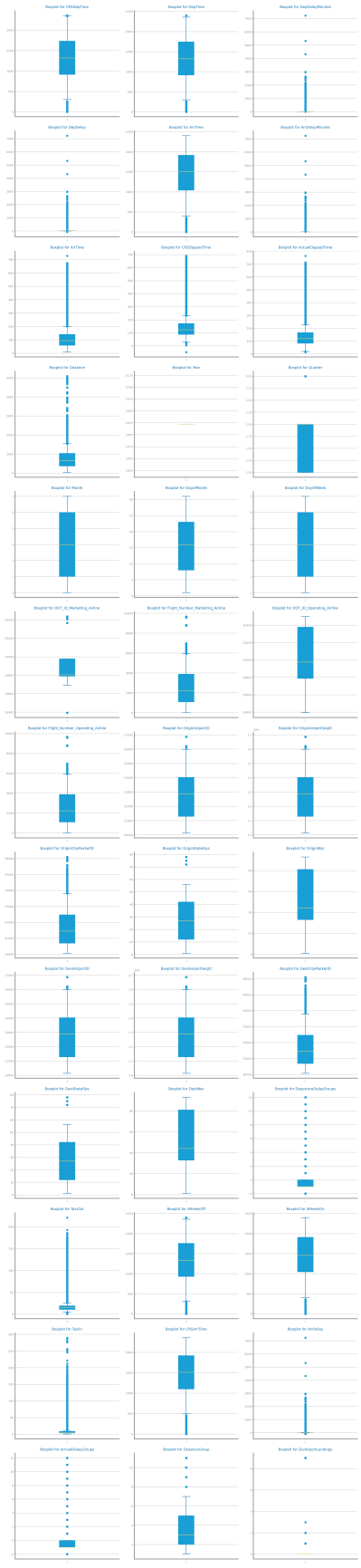


Figure 6: Single variables boxplots for dataset 2

Figure 7: Histograms for dataset 1

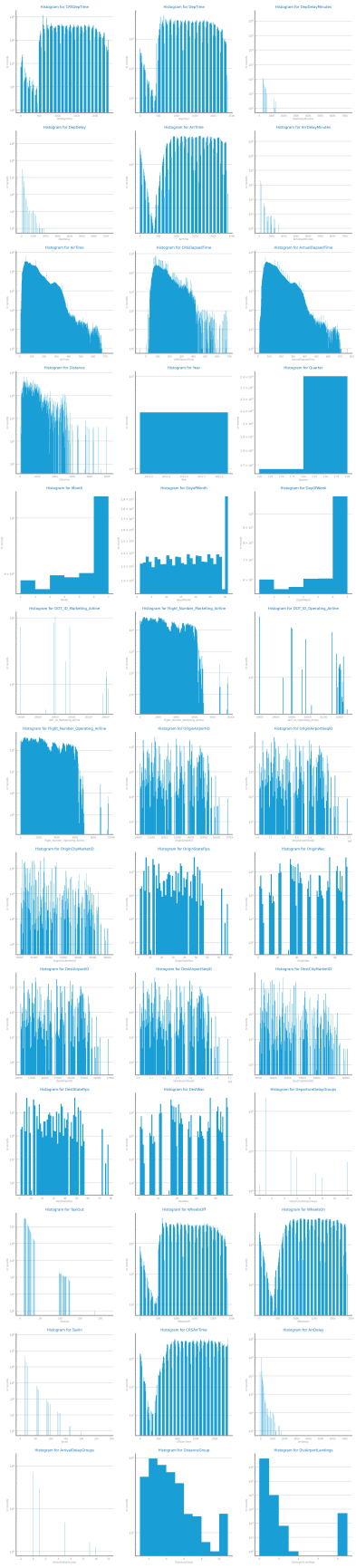


Figure 8: Histograms for dataset 2

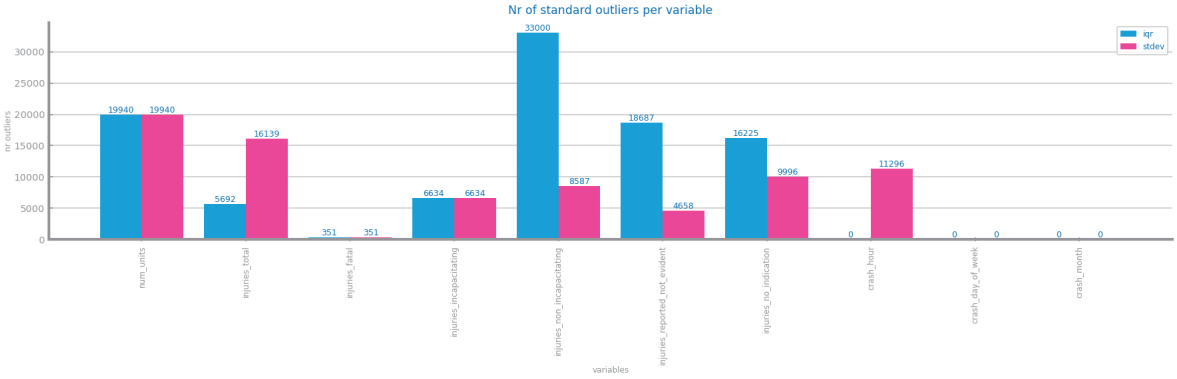


Figure 9: Outliers study dataset 1

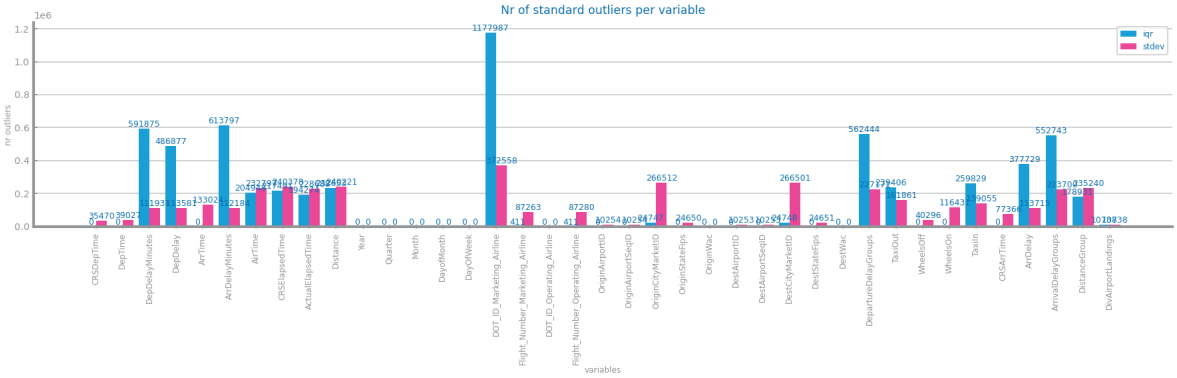


Figure 10: Outliers study dataset 2

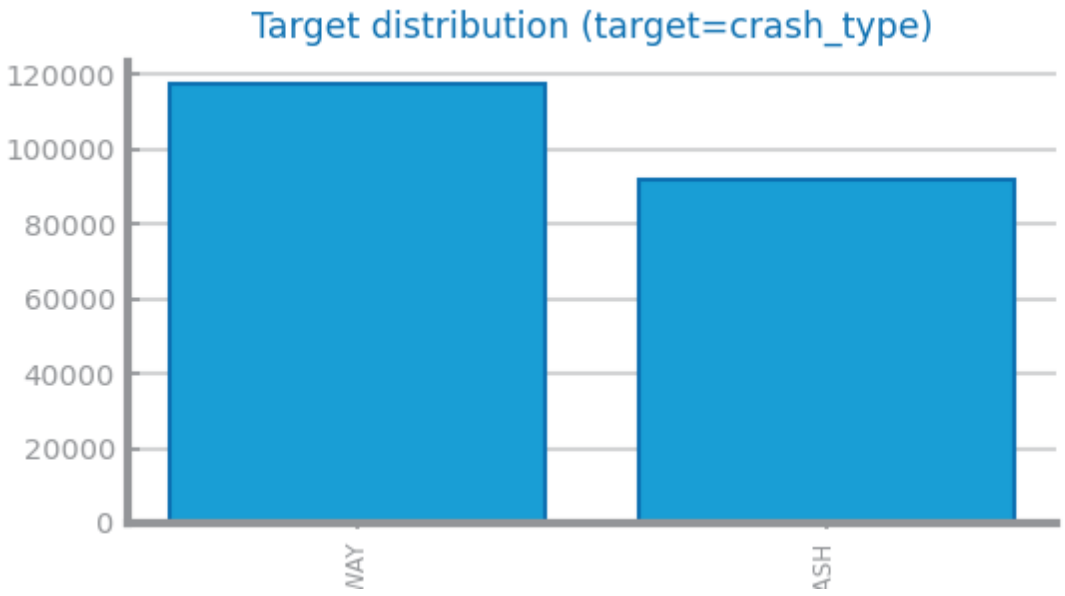


Figure 11: Class distribution for dataset 1

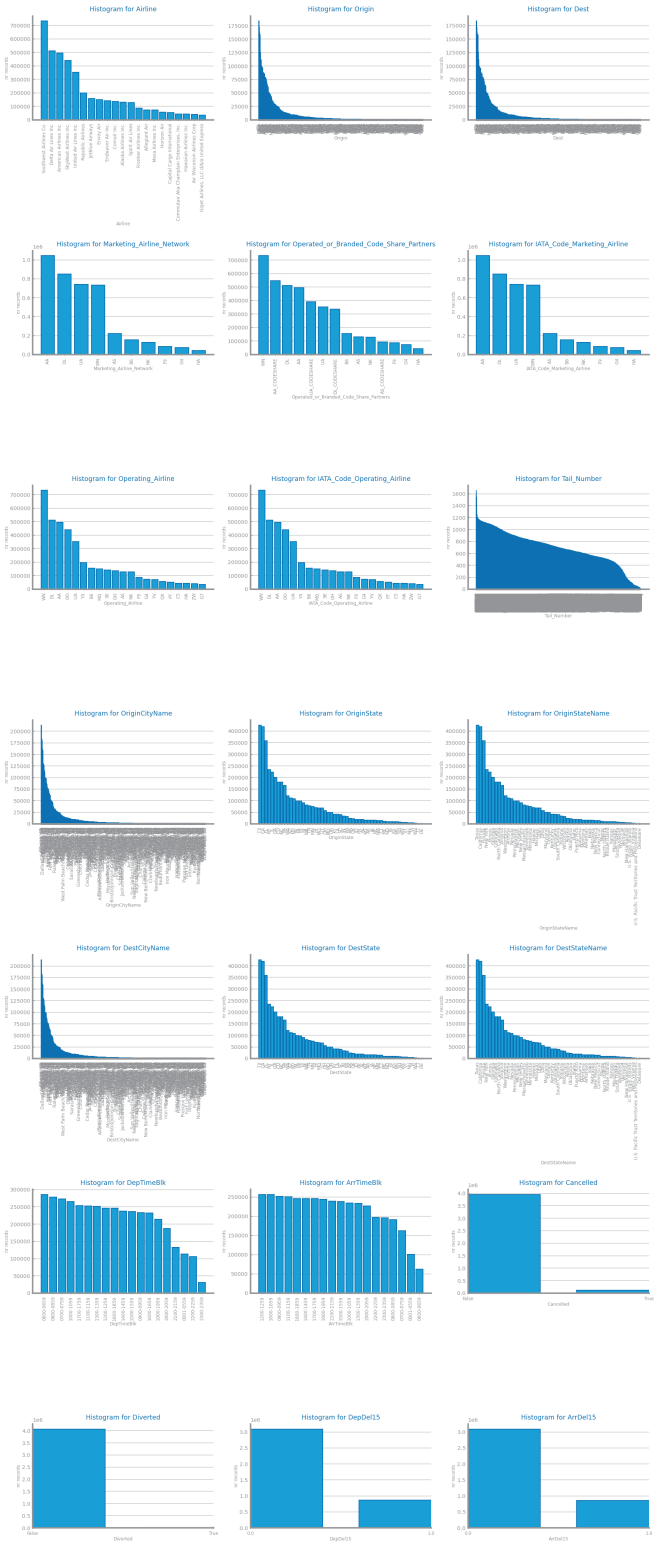


Figure 12: Class distribution for dataset 2

Data Granularity

Shall contain all relevant information and charts respecting to the data granularity perspective, such as the impact of different granularities considered for each variable. May present additional taxonomies if needed. **Shall not exceed 500 characters.**

Figure 13: Granularity analysis for dataset 1

Figure 14: Granularity analysis for dataset 2

Data Sparsity

Shall contain all relevant information and charts respecting to the data sparsity perspective, such as domain coverage and correlation among variables. **Shall not exceed 500 characters.**

Figure 15: Sparsity analysis for dataset 1

Figure 16: Sparsity analysis for dataset 2 - View on Google Drive

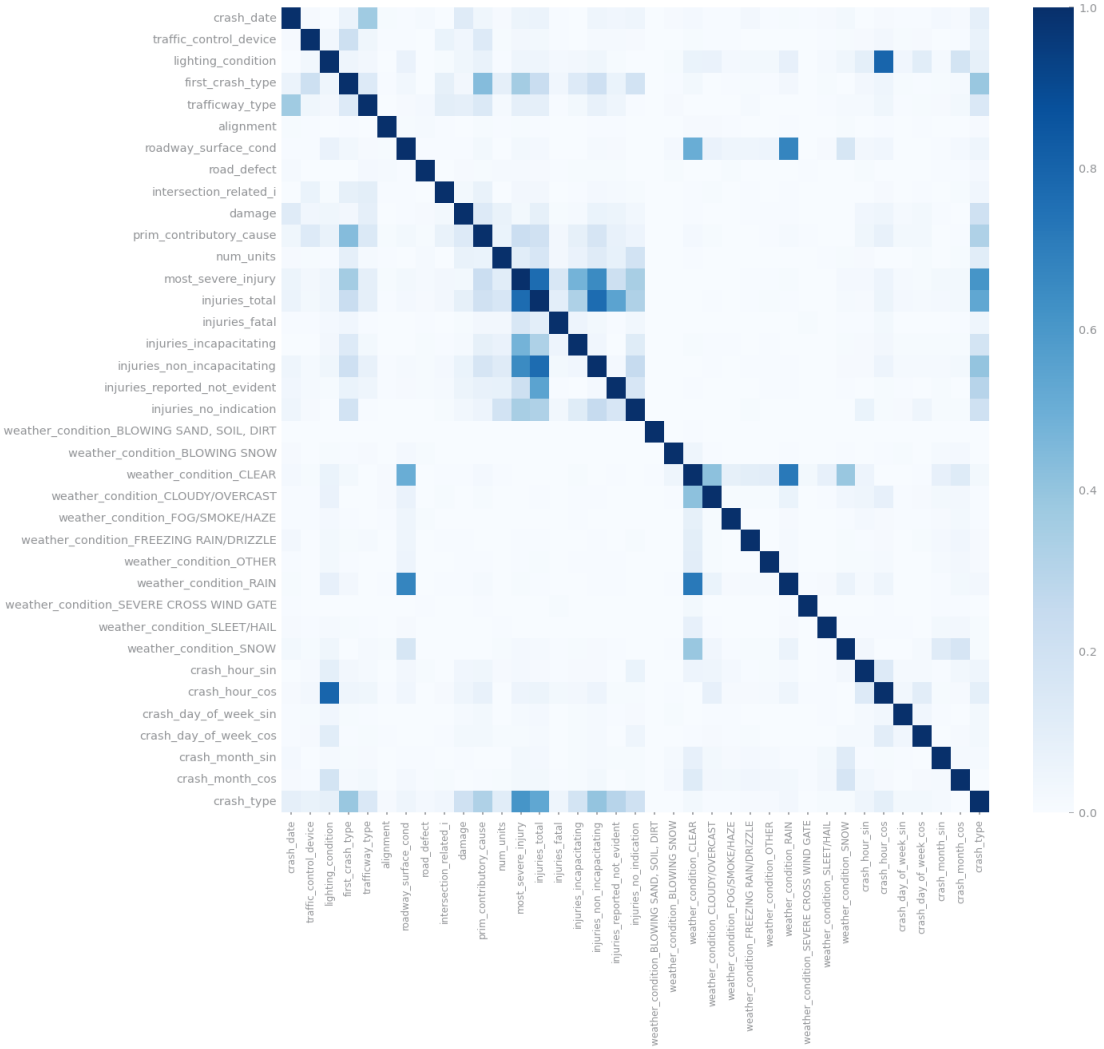


Figure 17: Correlation analysis for dataset 1

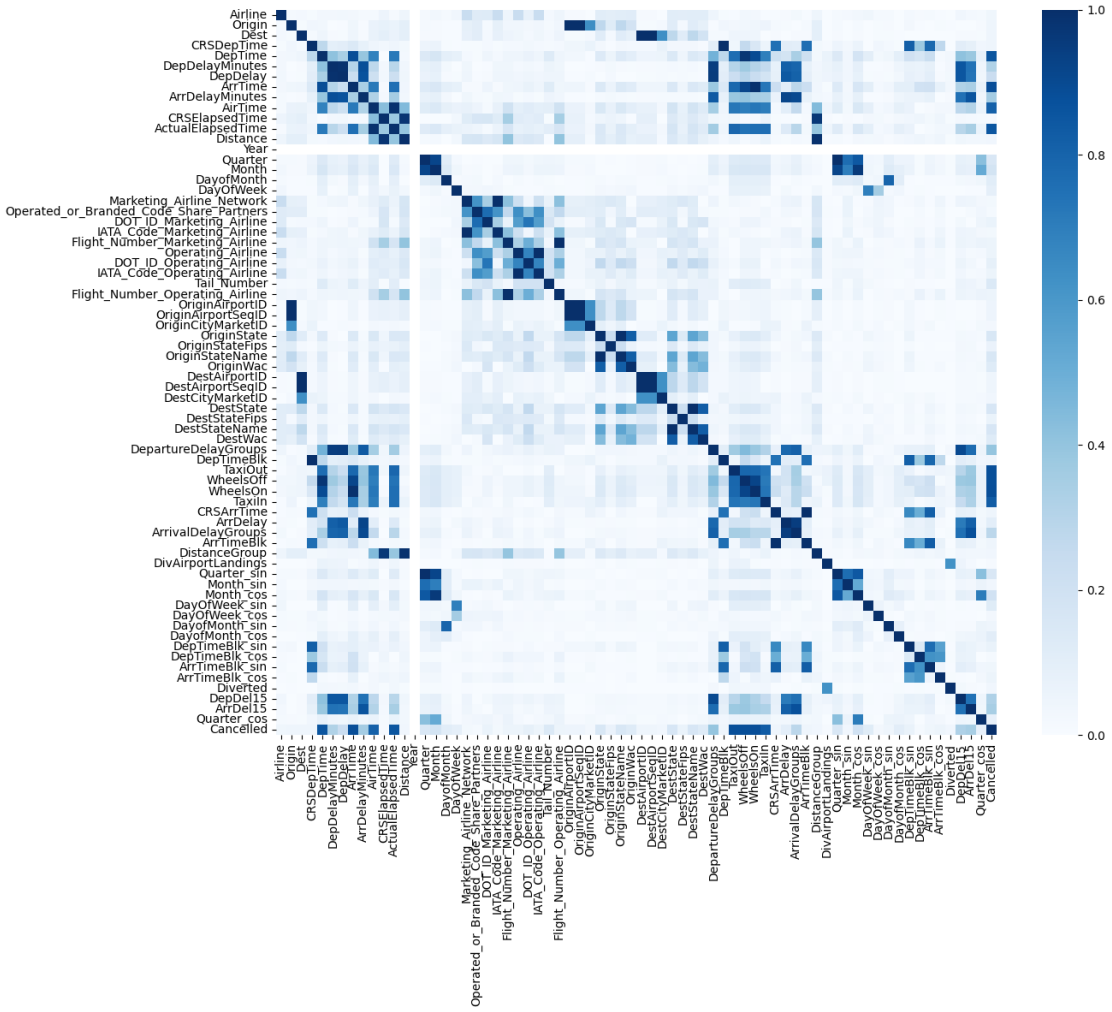


Figure 18: Correlation analysis for dataset 2

2 DATA PREPARATION

Variables Encoding

Shall contain all relevant information respecting to the transformation of variables. The list of variables under each one of the transformations, shall be presented. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 500 characters *for each dataset*.**

Missing Value Imputation

Shall contain all relevant information and charts respecting to missing values imputation, such as the choices made and the impact of the different approaches on modelling results. Shall also clearly reveal the approach selected to proceed with the processing. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 500 characters.**

Figure 19: Missing values imputation results with different approaches for dataset 1

Figure 20: Missing values imputation results with different approaches for dataset 2

Outliers Treatment

Shall contain all relevant information and charts respecting to outliers imputation, such as the choices made and the impact of the different approaches on modelling results. Shall also clearly reveal the approach selected to proceed with the processing. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 500 characters.**

Figure 21: Outliers imputation results with different approaches for dataset 1

Figure 22: Outliers imputation results with different approaches for dataset 2

Scaling

Shall contain all relevant information and charts respecting to scaling transformation, such as the choices made and the impact of the different approaches on modelling results. Shall also clearly reveal the approach selected to proceed with the processing. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 200 characters.**

Figure 23: Scaling results with different approaches for dataset 1

Figure 24: Scaling results with different approaches for dataset 2

Balancing

Shall contain all relevant information and charts respecting to balancing transformation, such as the choices made and the impact of the different approaches on modelling results. Shall also clearly reveal the approach selected to proceed with the processing. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 500 characters.**

Figure 25: Balancing results with different approaches for dataset 1

Figure 26: Balancing results with different approaches for dataset 2

Feature Selection

Shall contain all relevant information and charts respecting to feature selection based on filtering out redundant (based on correlation) and relevant (based on variation) variables. The different choices and their impact on the modelling results shall be presented and explained. Should also clearly reveal the approach selected to proceed with the processing. All explanations shall be based on data characteristics. **Shall not exceed 500 characters.**

Figure 27: Feature selection of redundant variables results with different parameters for dataset 1

Figure 28: Feature selection of redundant variables results with different parameters for dataset 2

Figure 29: Feature selection of relevant variables results with different parameters for dataset 1 (variance study)

Figure 30: Feature selection of relevant variables results with different parameters for dataset 2 (variance study)

Feature Extraction (optional)

Shall contain all relevant information and charts respecting to feature extraction, in particular PCA. The different choices and their impact on the modelling results shall be presented and explained. **Shall not exceed 200 characters.**

Figure 31: Principal components analysis and feature extraction results for dataset 1

Figure 32: Principal components analysis and feature extraction results for dataset 2

Additional Feature Generation (if done)

Shall contain all relevant information and charts respecting to feature generation. The different choices and their impact on the modelling results shall be presented and explained. Shall summarise all variables generated and the formula used to derive them (in a table). **Shall not exceed 200 characters.**

Figure 33: Feature generation results for dataset 1

Figure 34: Feature generation results for dataset 2

3 MODELS' EVALUATION

Shall be used to point out any important decision taken during the training, including training strategy and evaluation measures used. **Shall not exceed 500 characters.**

Naïve Bayes

Shall be used to present the results achieved with each one of Naïve Bayes implementations, comparing and proposing explanations for them. If any of the implementations is not used, a justification for it shall be presented. Shall be used to present the evaluation of the best model achieved. **Shall not exceed 300 characters.**

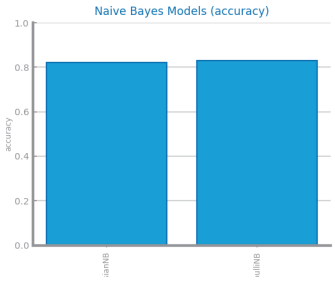


Figure 35: Naïve Bayes alternatives comparison for dataset 1

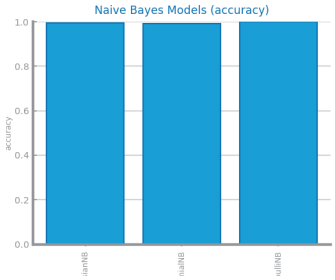


Figure 36: Naïve Bayes alternative comparison for dataset 2

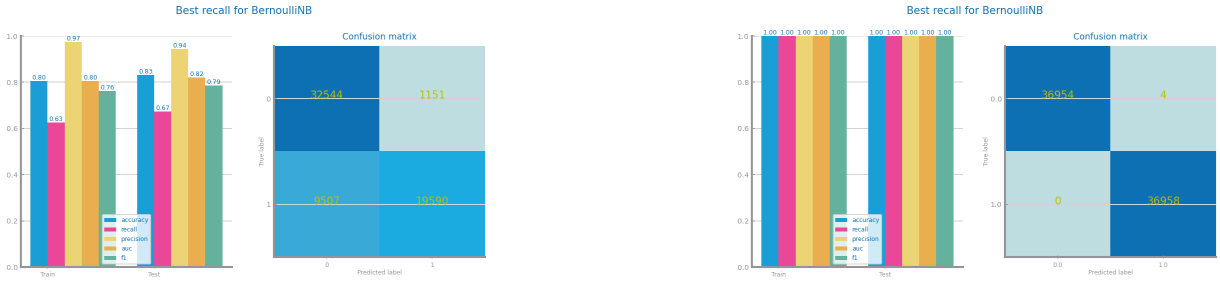


Figure 37: Naïve Bayes best model results for dataset 1 (left) and dataset 2 (right)

KNN

Shall be used to present the results achieved through different similarity measures and KNN parameterisations. The results shall be compared and explanations for them shall be presented. The justification for the chosen similarity measures shall be presented. Shall be used to address the overfitting phenomenon, studying the conditions under which models face it. Shall be used to present the evaluation of the best model achieved. **Shall not exceed 500 characters.**

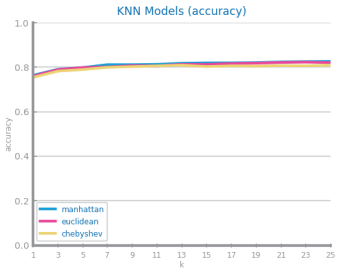


Figure 38: KNN different parameterisations comparison for dataset 1

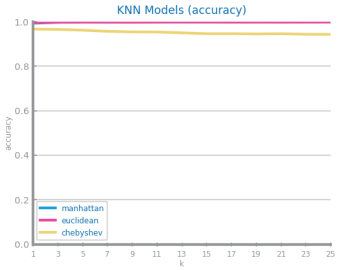


Figure 39: KNN different parameterisations comparison for dataset 2

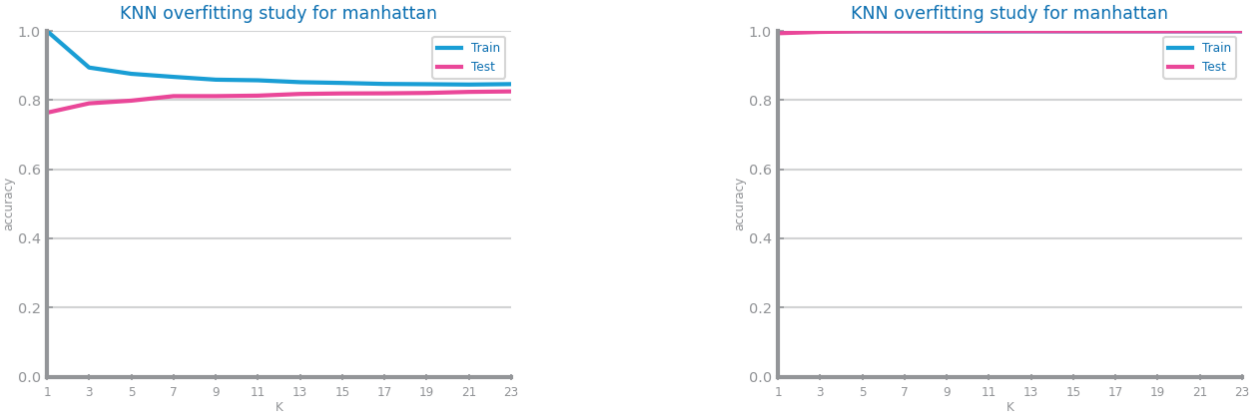


Figure 40: KNN overfitting analysis for dataset 1 (left) and dataset 2 (right)

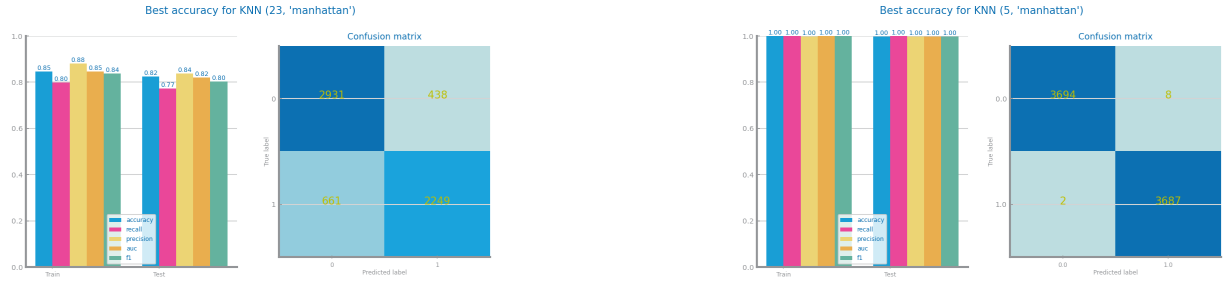


Figure 41: KNN best model results for dataset 1 (left) and dataset 2 (right)

Logistic Regression

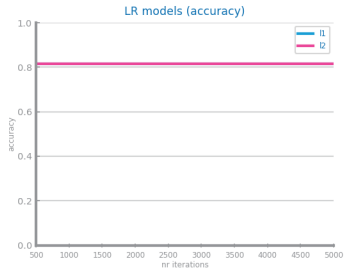


Figure 42: Logistic Regression different parameterisations comparison for dataset 1

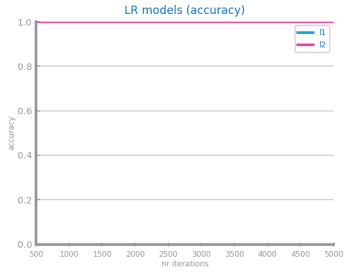


Figure 43: Logistic Regression different parameterisations comparison for dataset 2

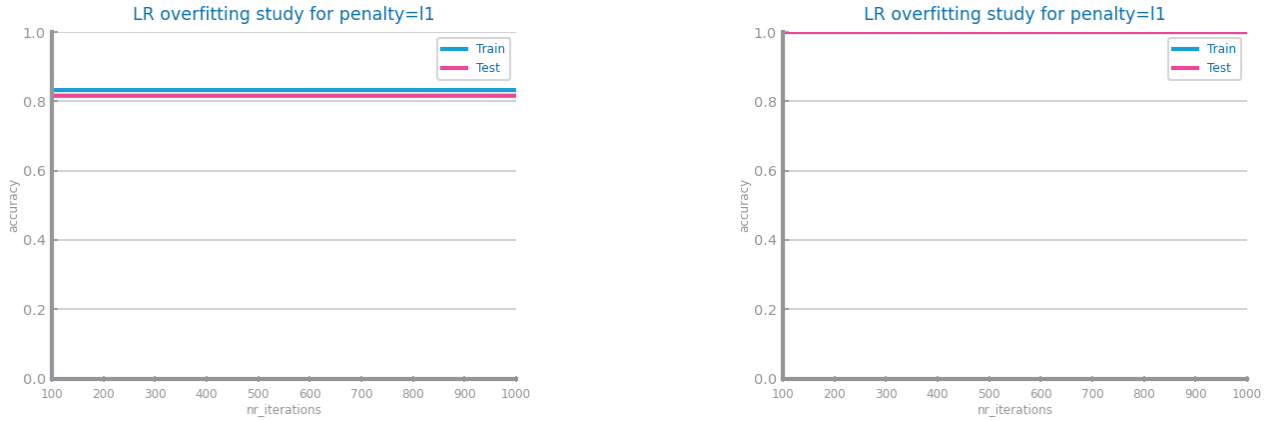


Figure 44: Logistic Regression overfitting analysis for dataset 1 (left) and dataset 2 (right)

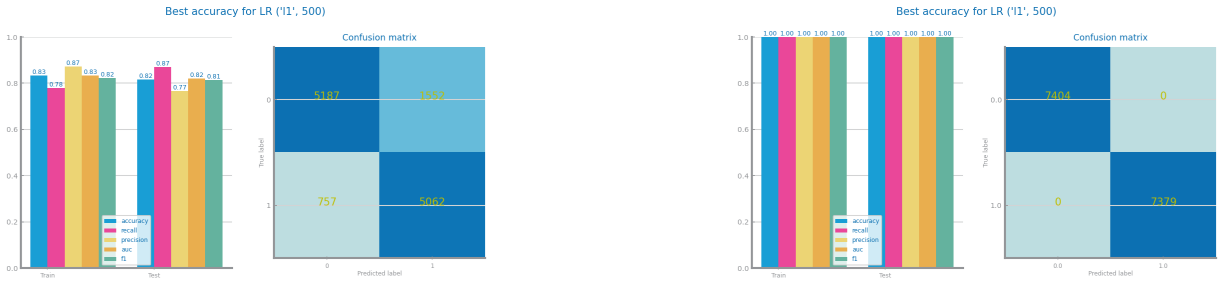


Figure 45: Logistic Regression best model results for dataset 1 (left) and dataset 2 (right)

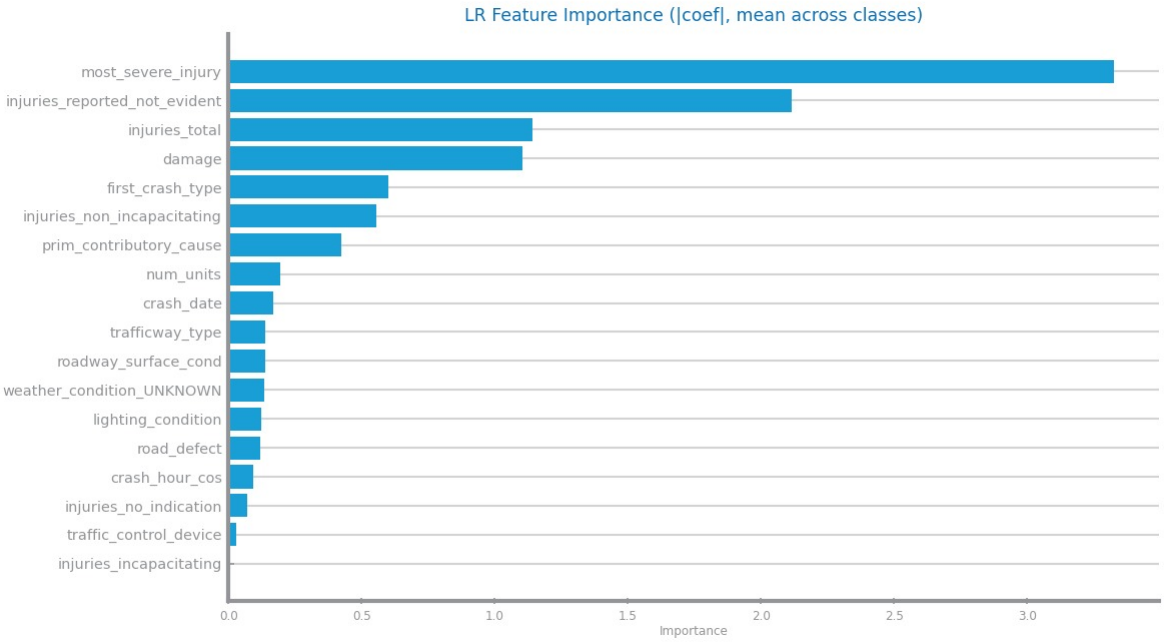


Figure 46: Logistic Regression feature importance for dataset 1

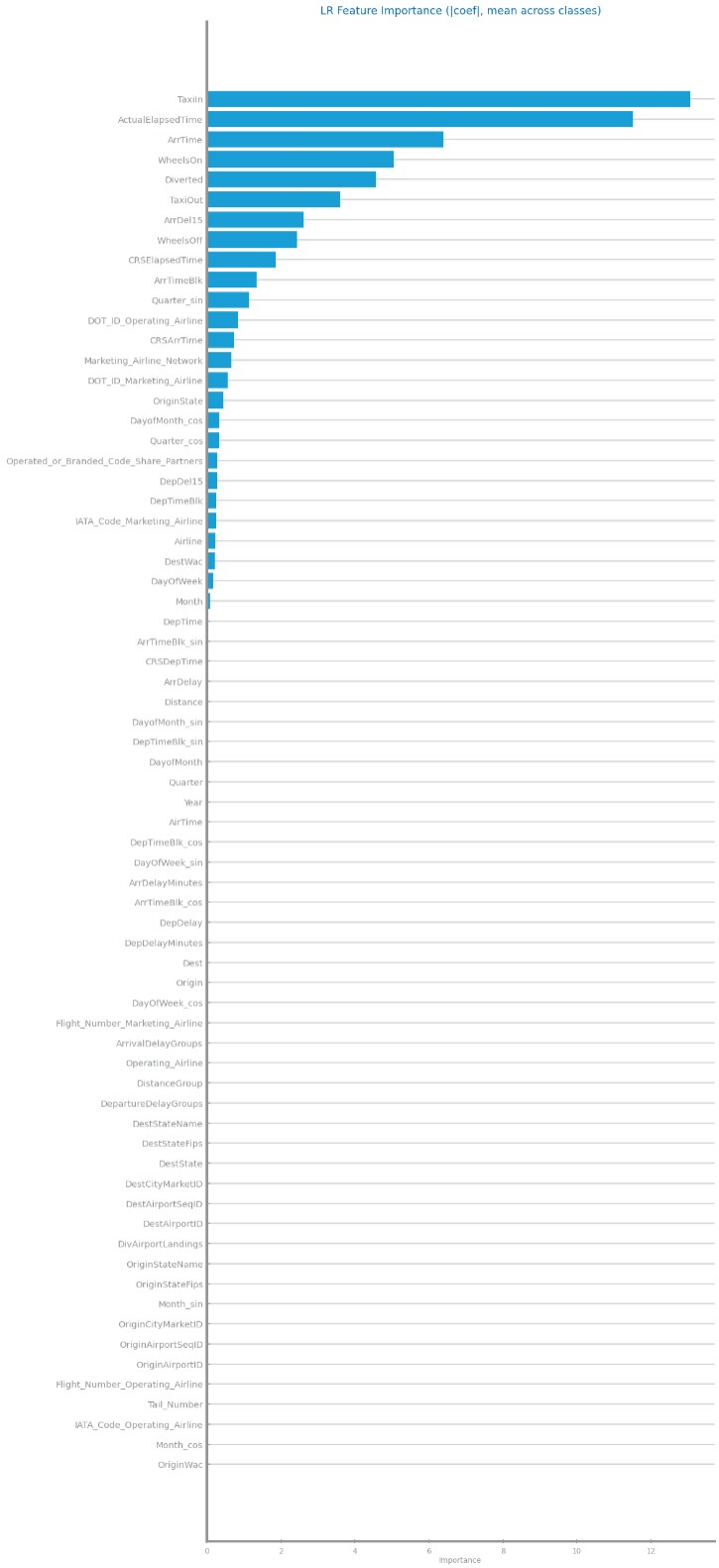


Figure 47: Logistic Regression feature importance for dataset 2

Decision Trees

Shall be used to present the results achieved through different parameterisations for the train of decision trees. The results shall be compared and explanations for them shall be presented. Shall be used to address the overfitting phenomenon, studying the conditions under which models face it. Shall be used to present the evaluation of the best model achieved. Shall be used to present the best tree achieved and its succinct description. **Shall not exceed 500 characters.**

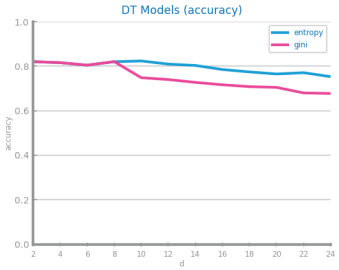


Figure 48: Decision Trees different parameterisations comparison for dataset 1

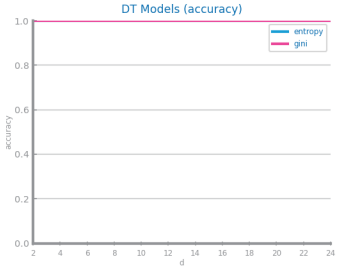


Figure 49: Decision Trees different parameterisations comparison for dataset 2

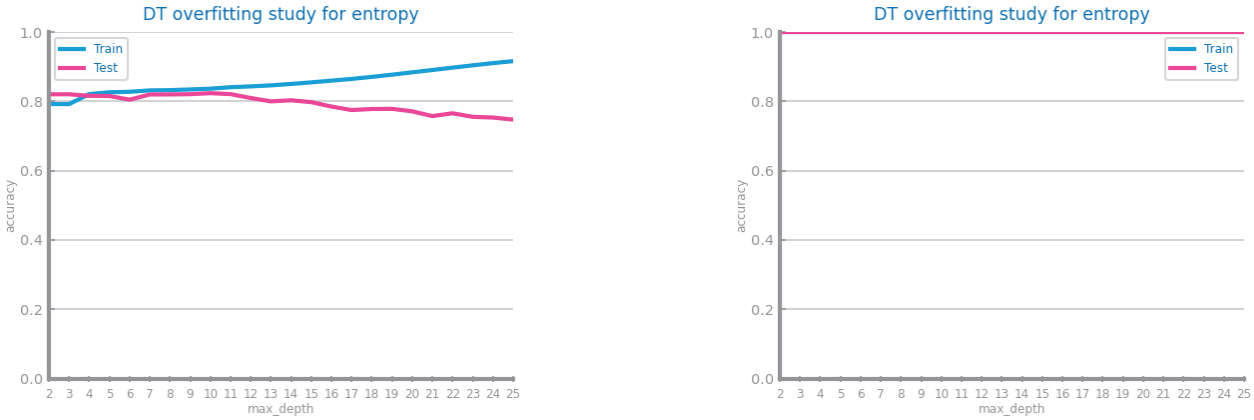


Figure 50: Decision Trees overfitting analysis for dataset 1 (left) and dataset 2 (right)

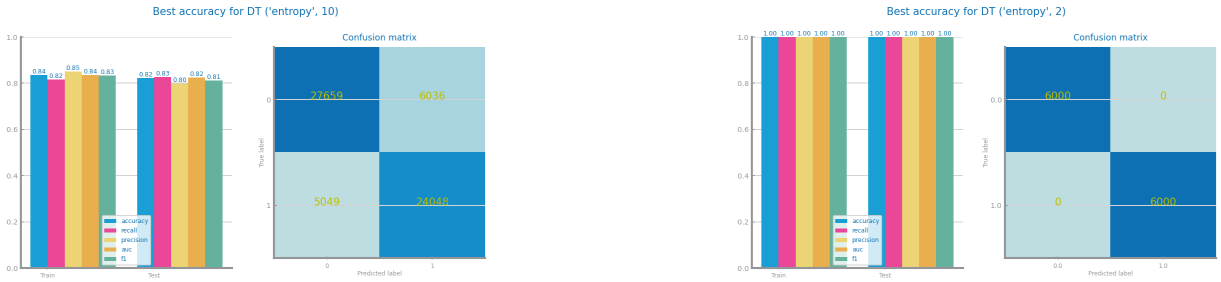


Figure 51: Decision trees best model results for dataset 1 (left) and dataset 2 (right)

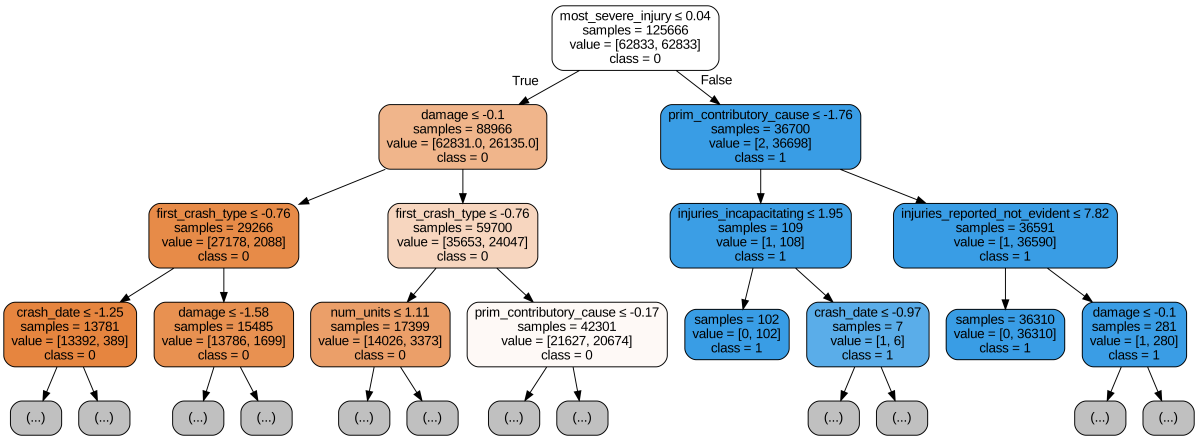


Figure 52: Best tree for dataset 1

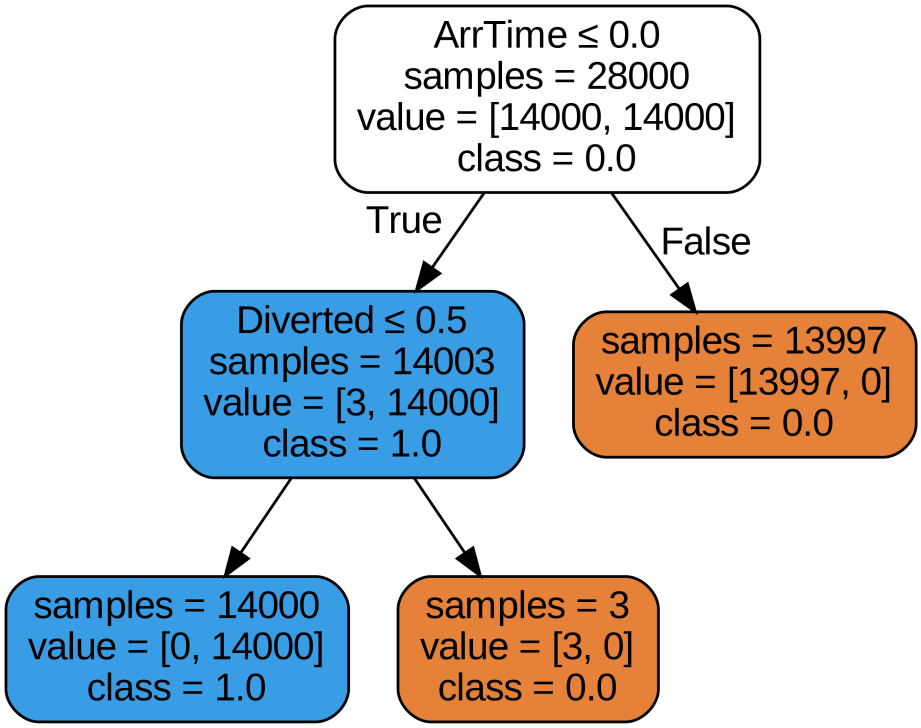


Figure 53: Best tree for dataset 2

Random Forests

Shall be used to present the results achieved through different parameterisations for the train of random forests. The results shall be compared and explanations for them shall be presented. Shall be used to address the overfitting phenomenon, studying the conditions under which models face it. Shall be used to present the evaluation of the best model achieved. May be used to present the most important variables in the model. **Shall not exceed 500 characters.**

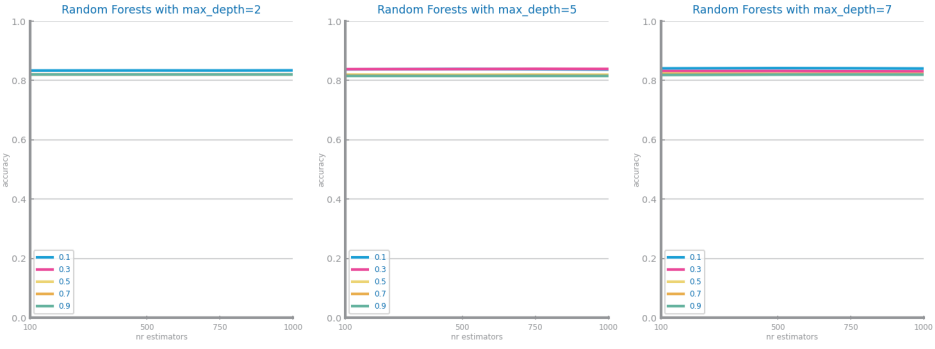


Figure 54: Random Forests different parameterisations comparison for dataset 1

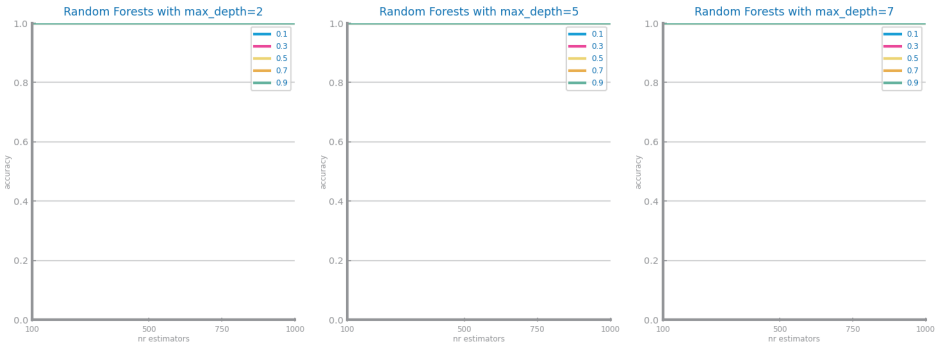


Figure 55: Random Forests different parameterisations comparison for dataset 2

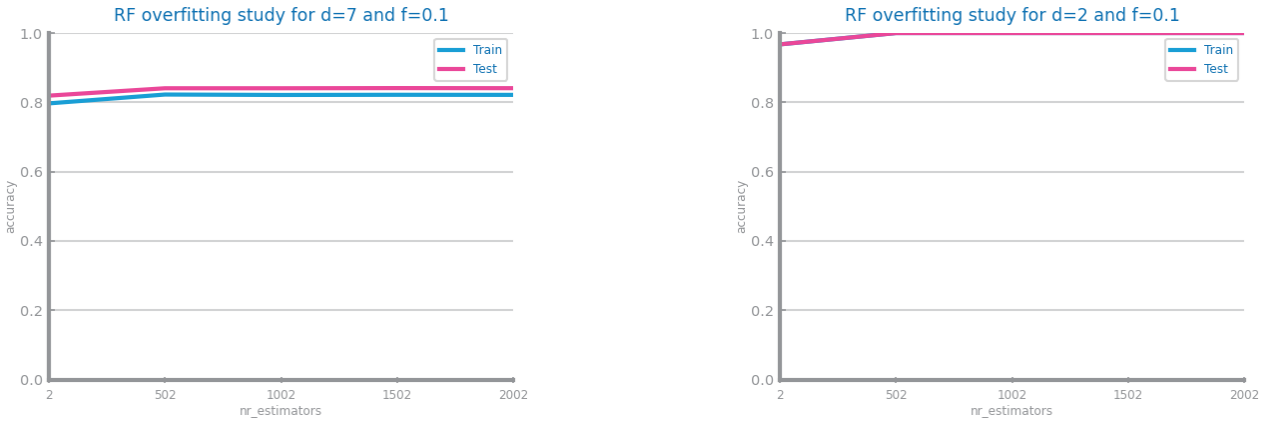


Figure 56: Random Forests overfitting analysis for dataset 1 (left) and dataset 2 (right)

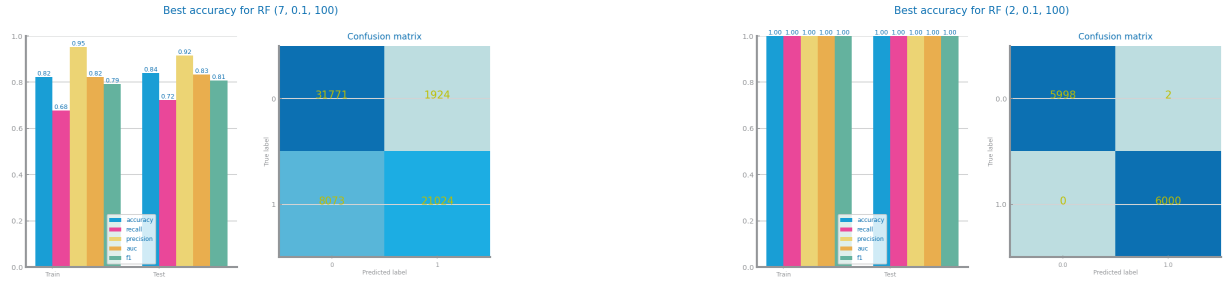


Figure 57: Random Forests best model results for dataset 1 (left) and dataset 2 (right)

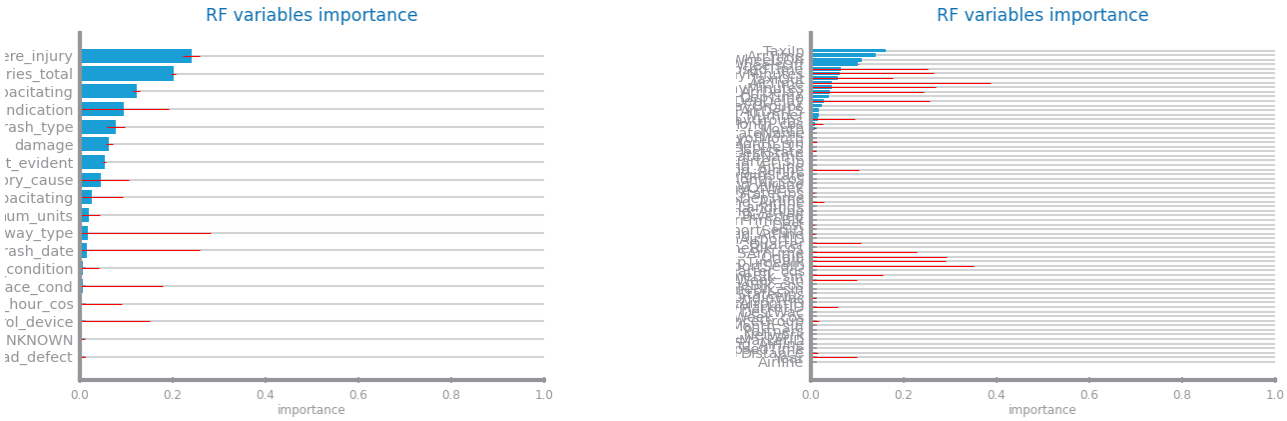


Figure 58: Random Forests variables importance for dataset 1 (left) and dataset 2 (right)

Gradient Boosting

Shall be used to present the results achieved through different parameterisations for the train of gradient boosting. The results shall be compared and explanations for them shall be presented. Shall be used to address the overfitting phenomenon, studying the conditions under which models face it. Shall be used to present the evaluation of the best model achieved. May be used to present the most important variables in the model. **Shall not exceed 500 characters.**

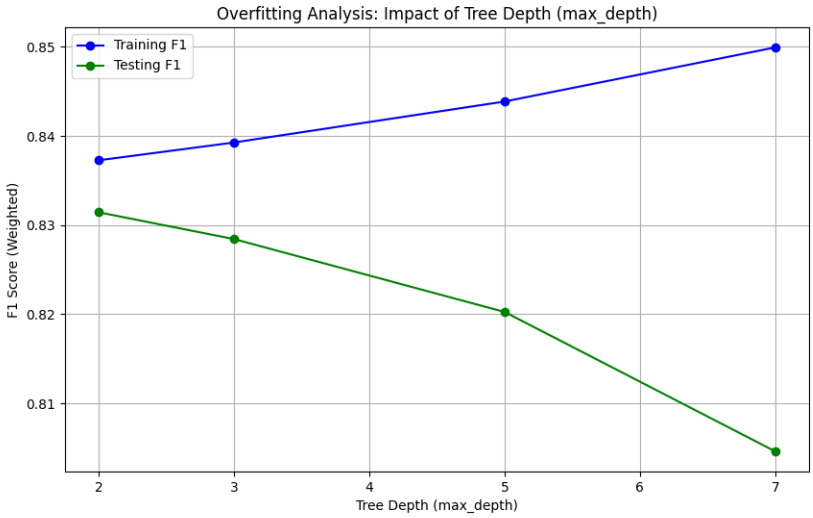


Figure 59: Gradient boosting different parameterisations comparison for dataset 1

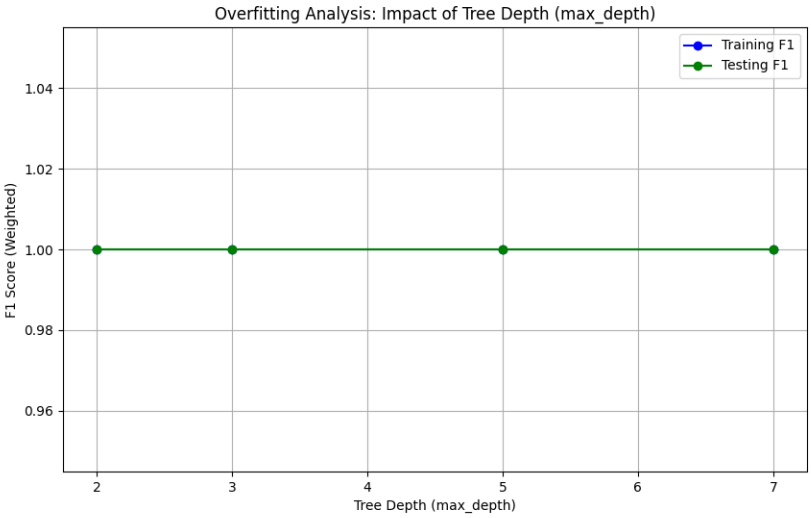


Figure 60: Gradient boosting different parameterisations comparison for dataset 2

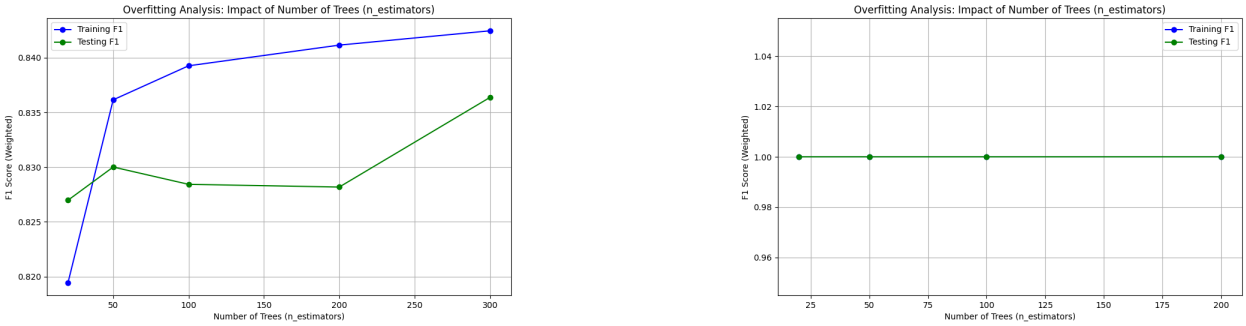


Figure 61: Gradient boosting overfitting analysis for dataset 1 (left) and dataset 2 (right)

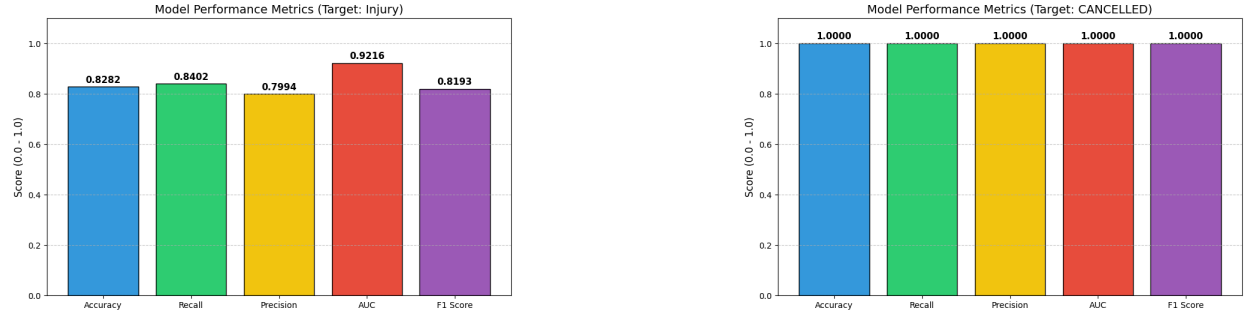


Figure 62: Gradient boosting best model results for dataset 1 (left) and dataset 2 (right)

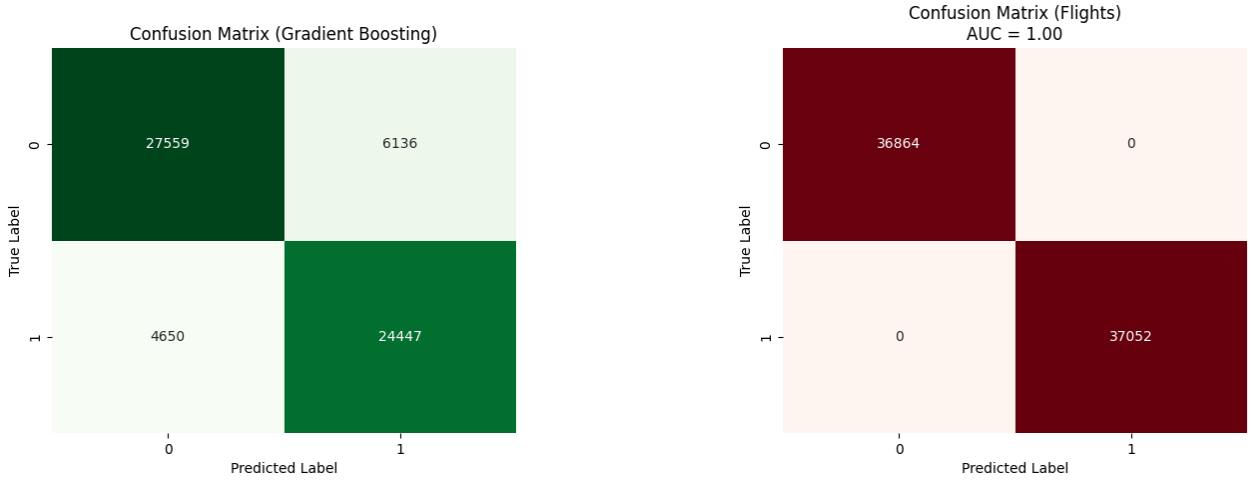


Figure 63: Gradient boosting variables importance for dataset 1 (left) and dataset 2 (right)

Multi-Layer Perceptrons

Shall be used to present the results achieved through different parameterisations for the train of MLPs. The results shall be compared and explanations for them shall be presented. Shall be used to address the overfitting phenomenon, studying the conditions under which models face it. Shall be used to present the evaluation of the best model achieved. **Shall not exceed 500 characters.**

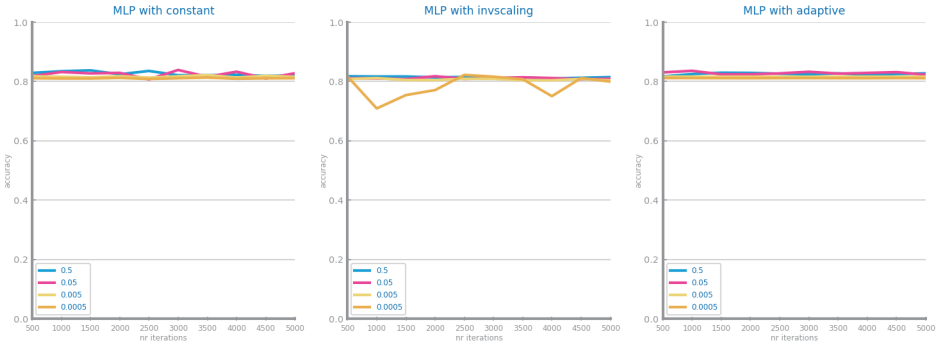


Figure 64: MLP different parameterisations comparison for dataset 1

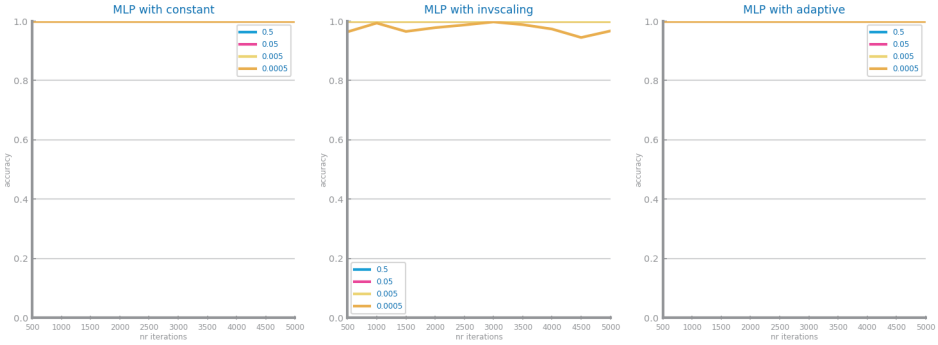


Figure 65: MLP different parameterisations comparison for dataset 2

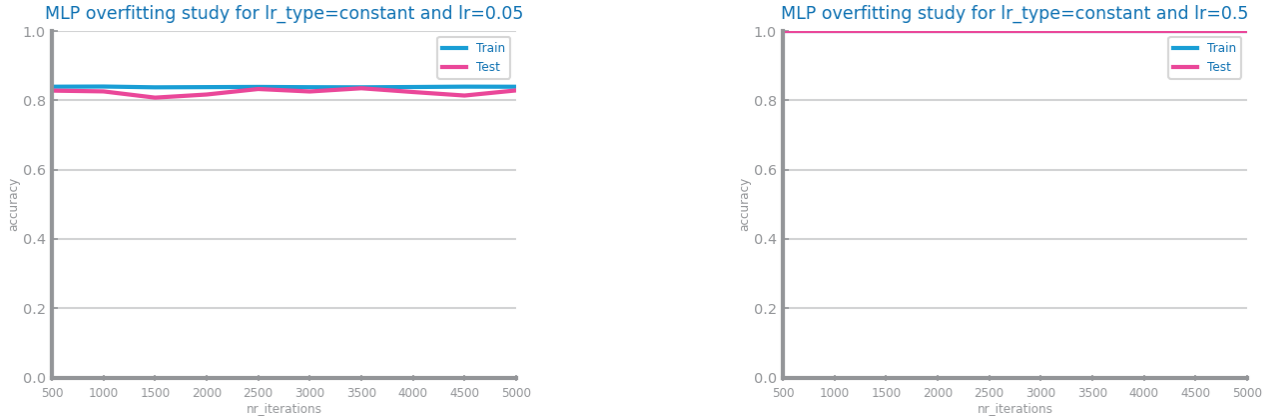


Figure 66: MLP overfitting analysis for dataset 1 (left) and dataset 2 (right)

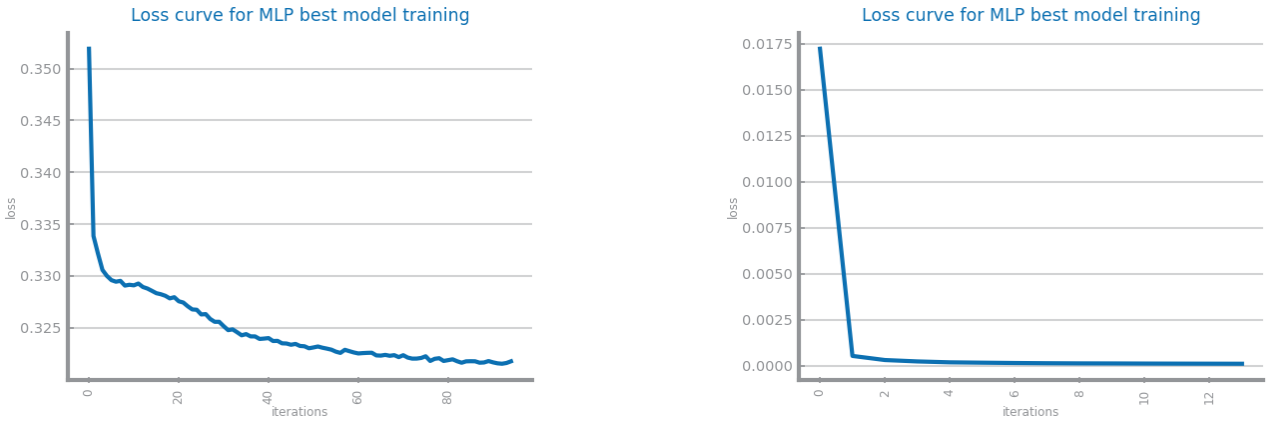


Figure 67: Loss curve analysis for dataset 1 (left) and dataset 2 (right)

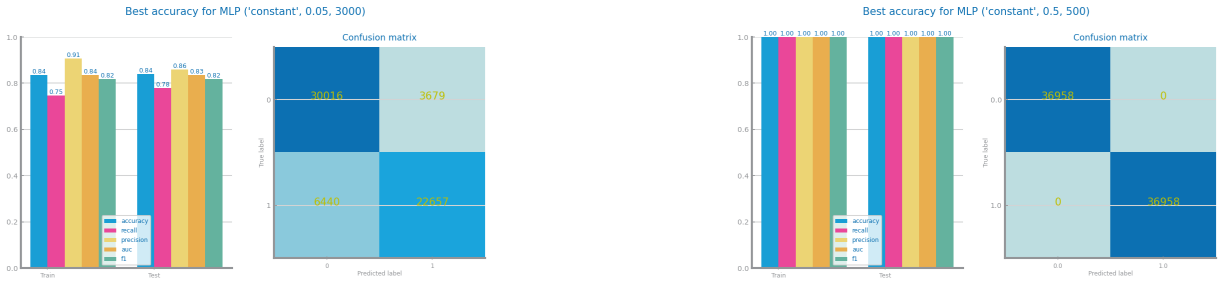


Figure 68: MLP best model results for dataset 1 (left) and dataset 2 (right)

4 CRITICAL ANALYSIS

Shall be used to present a summary of the results achieved with the different modelling techniques, and the impact of the different preparation tasks on their performance. A cross-analysis of the different models may also be presented, identifying the most relevant variables common to all of them (when possible) and the relation among the patterns identified within the different classifiers. A critical assessment of the best models shall be presented, clearly stating if the models seem to be good enough for the problem at hand. **Additional charts may be presented here. Shall not exceed 2000 characters.**

TIME SERIES ANALYSIS

5 DATA PROFILING

Data Dimensionality and Granularity

May be used to identify the most atomic granularity and two other different granularities to consider. **Shall not exceed 500 characters.**

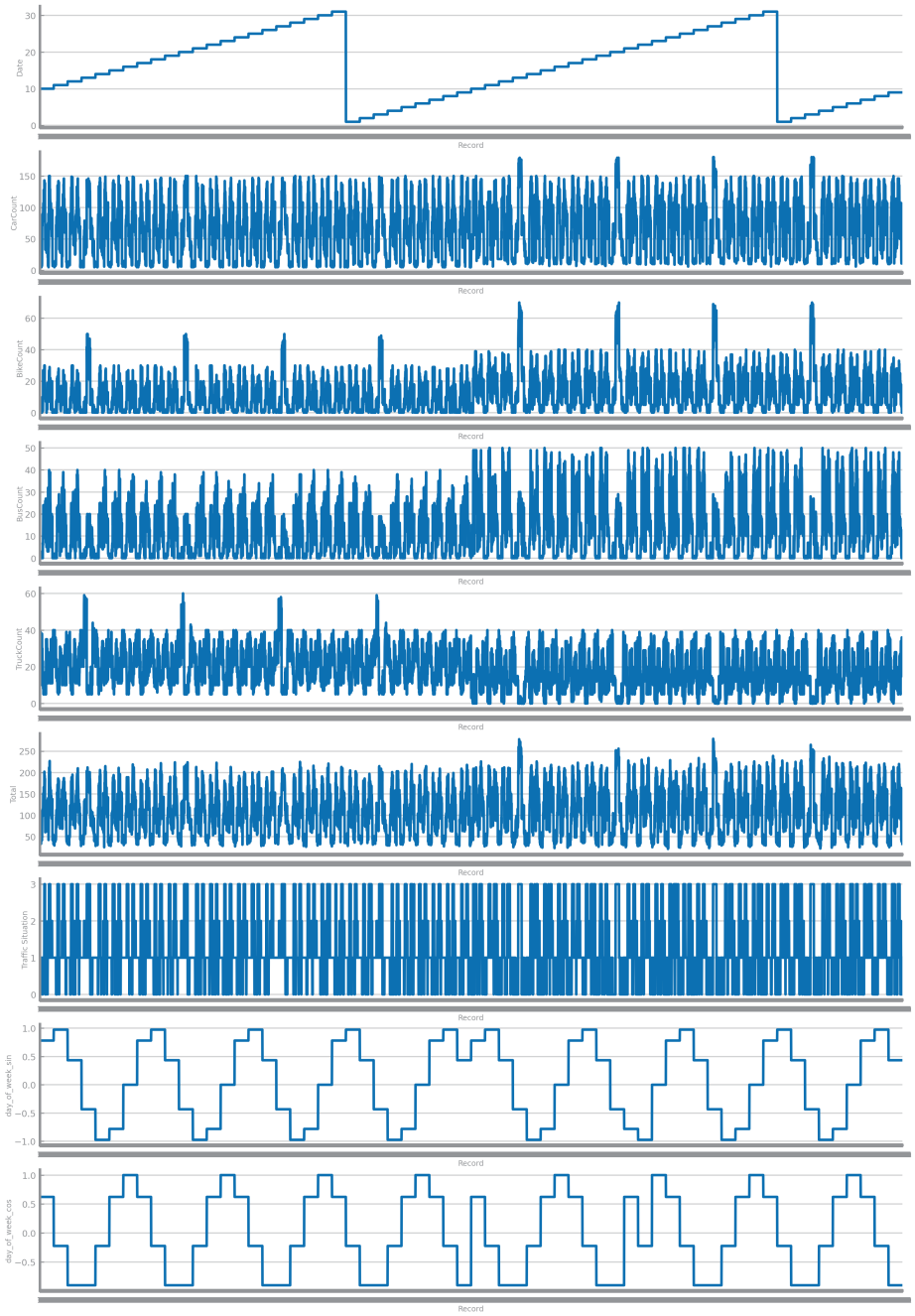


Figure 69: Time series 1 at the most granular detail

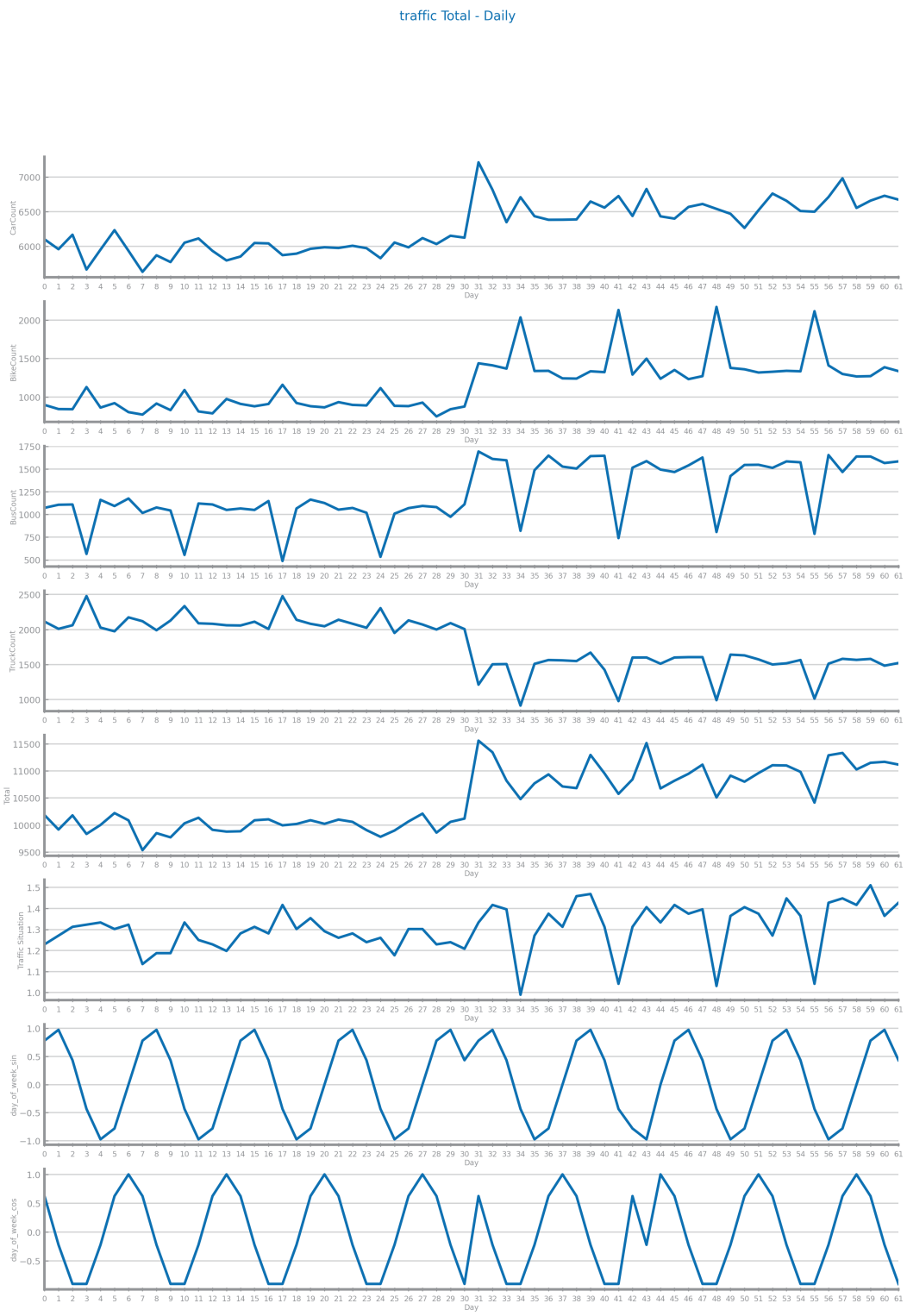


Figure 70: Time series 1 at the second chosen granularity

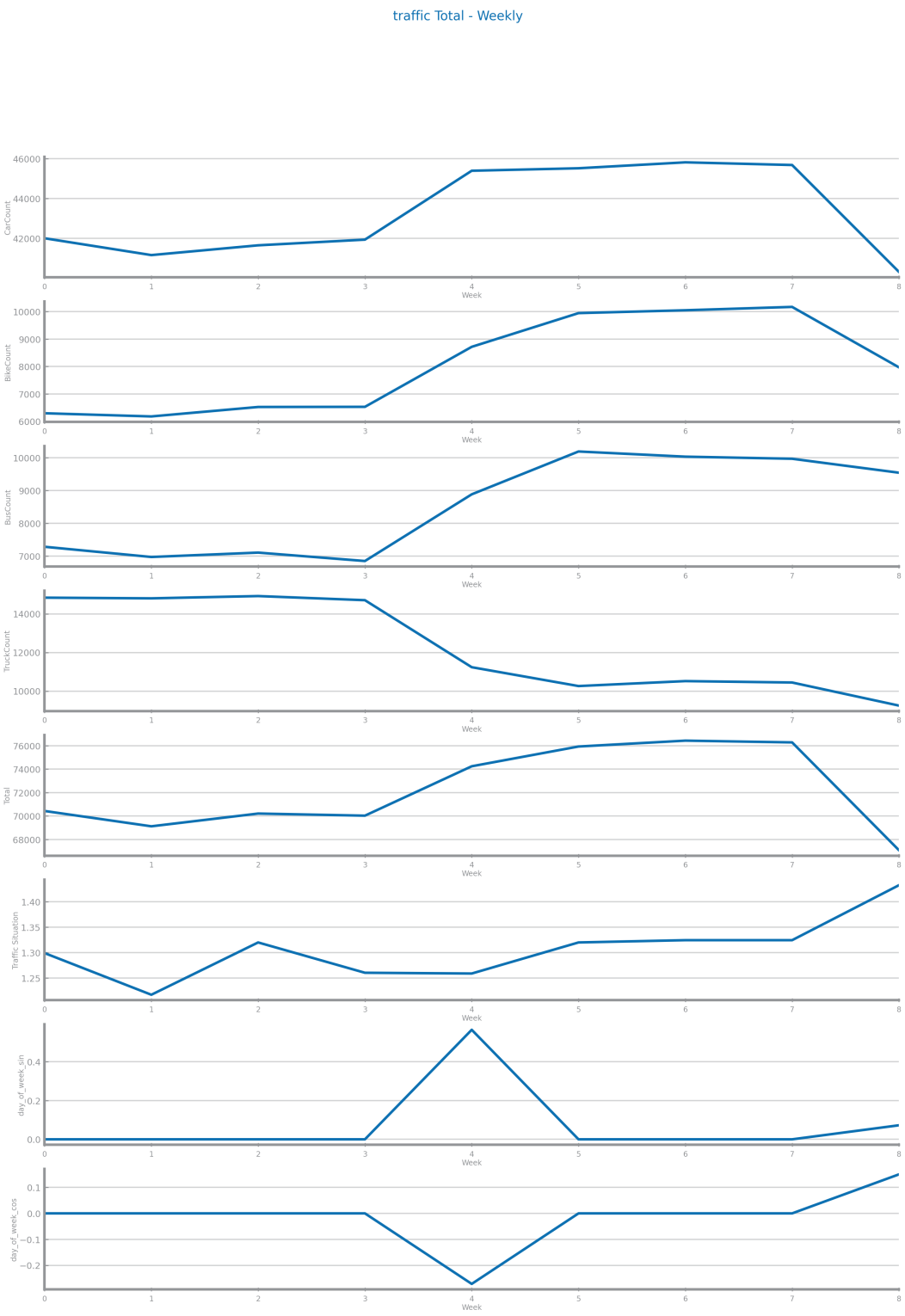


Figure 71: Time series 1 at the third chosen granularity

Data Distribution

Shall be used to perform the data analysis at those three different granularities, concerning the series distribution. **Shall not exceed 500 characters.**

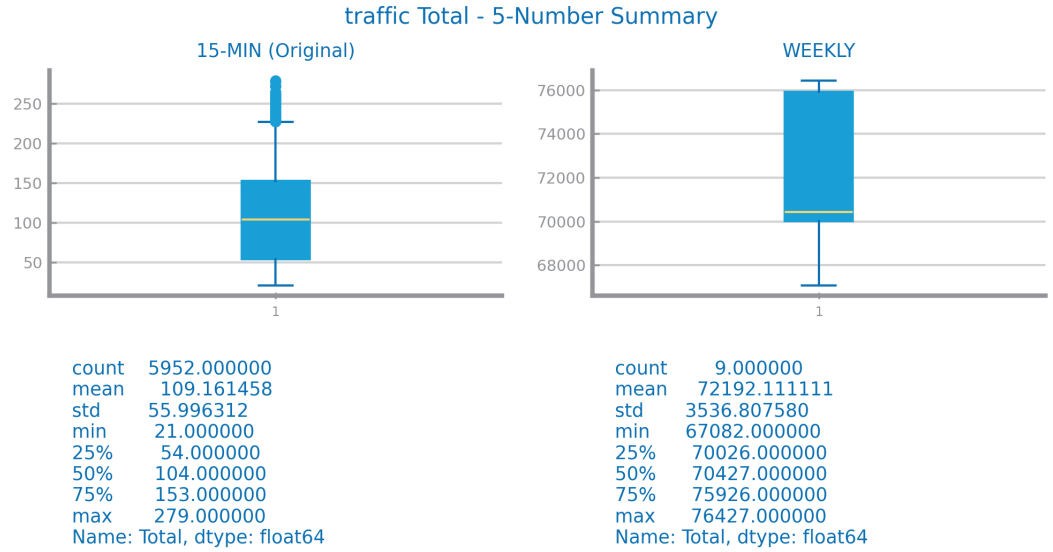


Figure 72: Boxplot(s) for time series 1

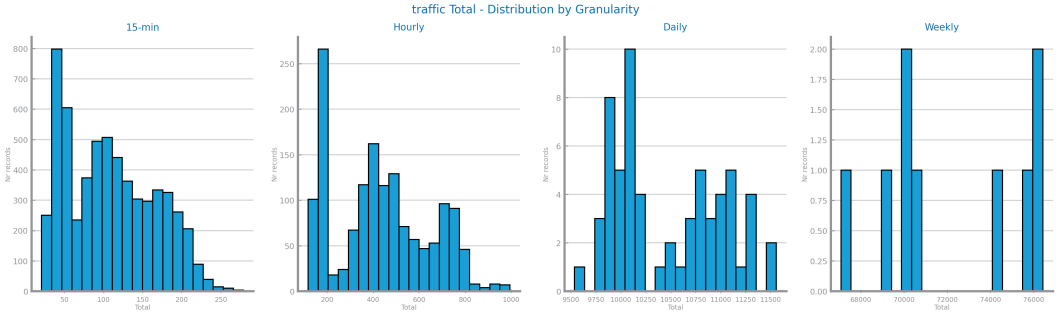


Figure 73: Histogram(s) for time series 1

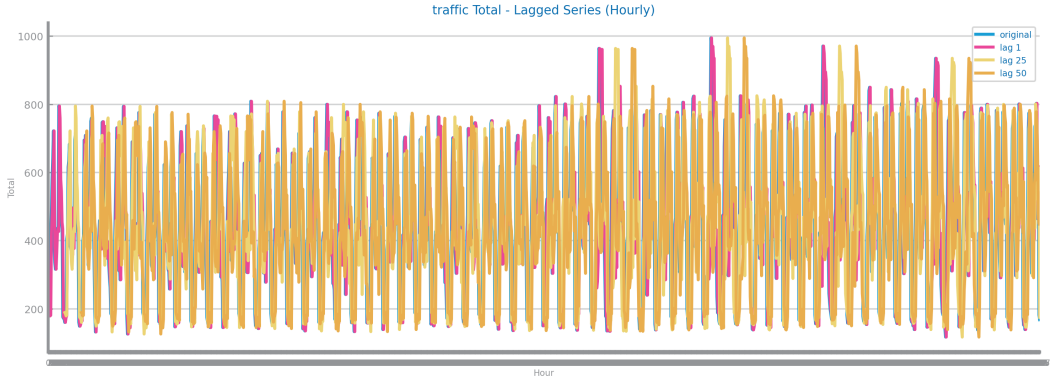


Figure 74: Autocorrelation lag-plots for original time series 1

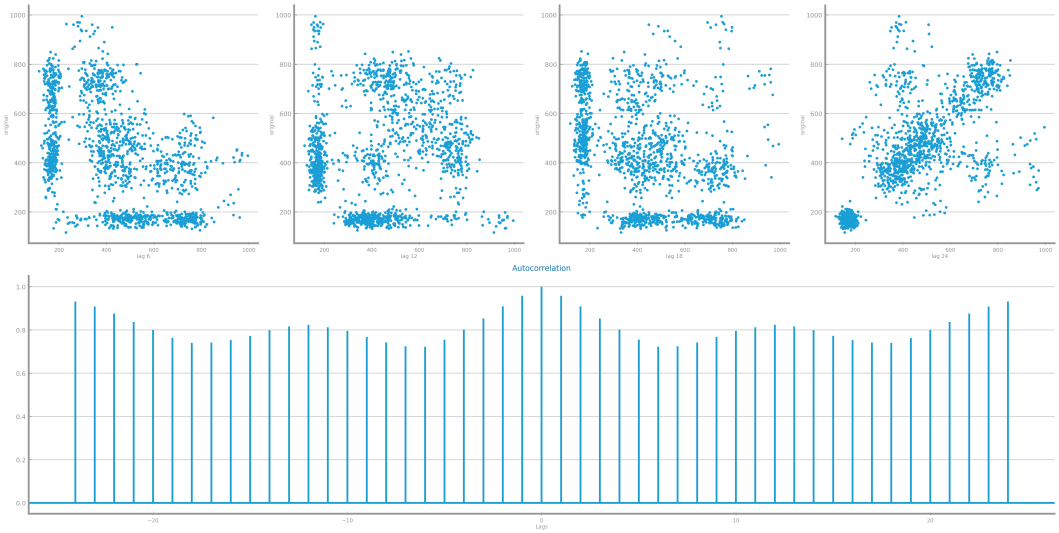


Figure 75: Autocorrelation correlogram for original time series 1

Data Stationarity

Shall be used to perform the data analysis at those three different granularities, concerning the series stationarity. **Shall not exceed 300 characters.**

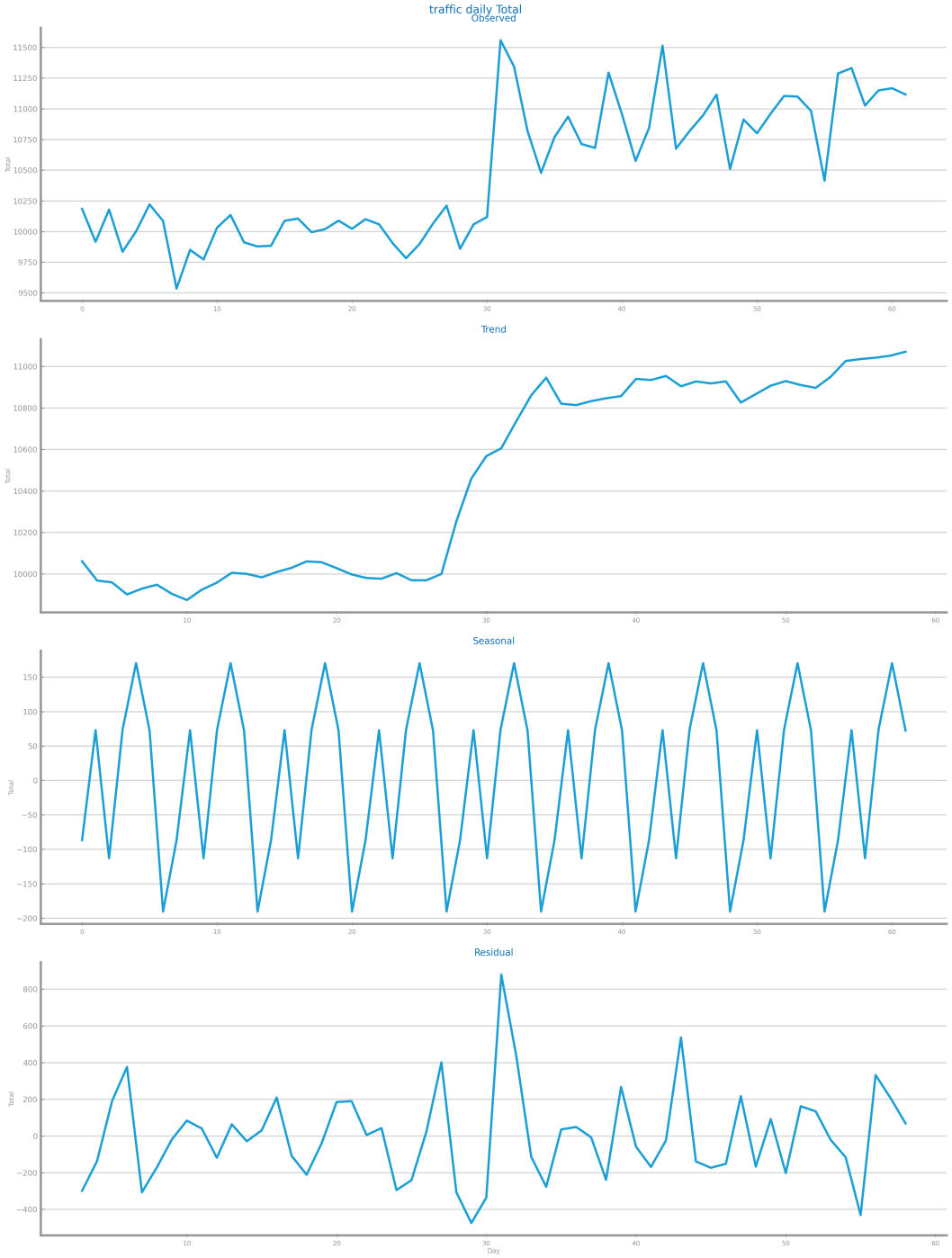


Figure 76: Components study for time series 1

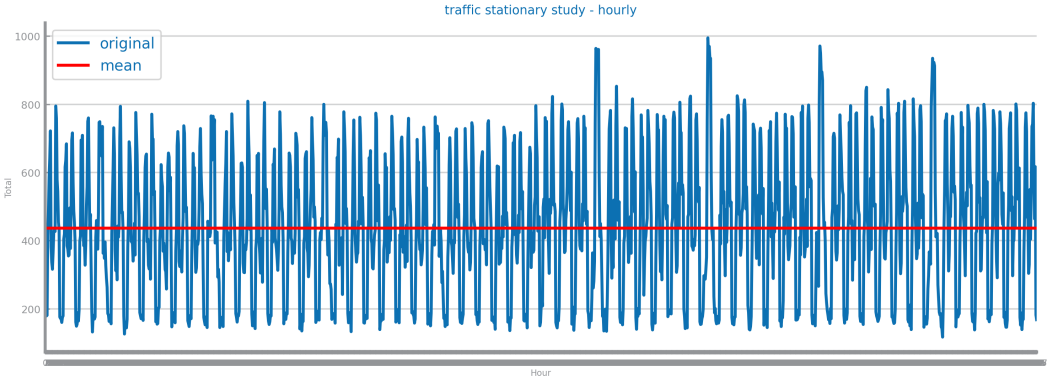


Figure 77: Stationarity study for time series 1

Augmented Dickey-Fuller Test Results:

Original (15-min):

- ADF Statistic: -14.441
- p-value: 0.000
- Critical Values: 1%: -3.431, 5%: -2.862, 10%: -2.567
- **The series IS stationary**

Hourly:

- ADF Statistic: -8.903
- p-value: 0.000
- Critical Values: 1%: -3.435, 5%: -2.864, 10%: -2.568
- **The series IS stationary**

Daily:

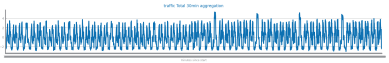
- ADF Statistic: -0.826
- p-value: 0.811
- Critical Values: 1%: -3.548, 5%: -2.913, 10%: -2.594
- **The series IS NOT stationary**

6 DATA TRANSFORMATION

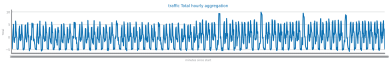
Aggregation

Shall describe the results of applying three different aggregations over both datasets, and identifying the granularity chosen to proceed. **Shall not exceed 300 characters.**

(a) No Aggregation (original after scaling)



(b) 30min Aggregation



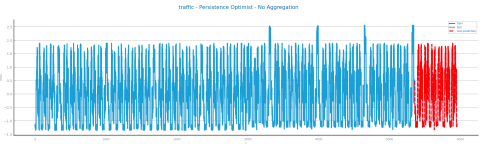
(c) Hourly Aggregation

(d) Daily Aggregation

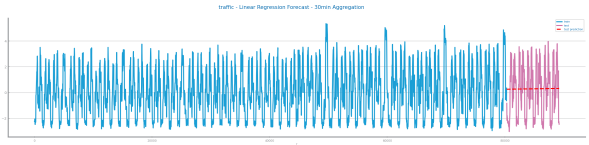
(e) Weekly Aggregation

Figure 78: Time series plots after different levels of aggregation

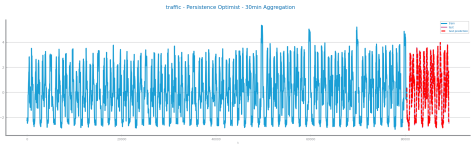
(a) Linear Regression – No Aggregation



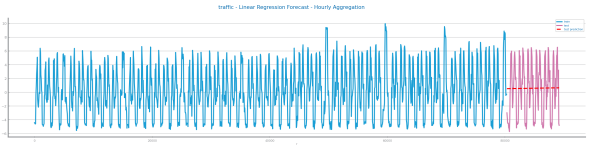
(b) Persistence Optim – No Aggregation



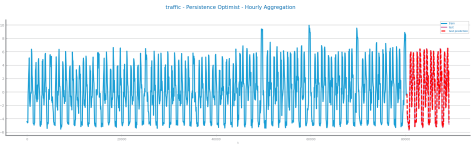
(c) Linear Regression – 30min Aggregation



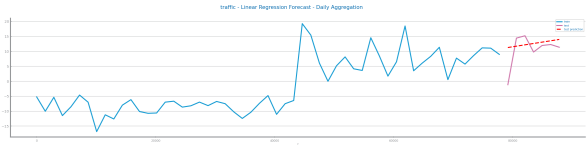
(d) Persistence Optim – 30min Aggregation



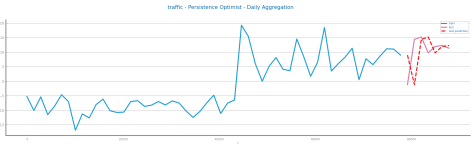
(e) Linear Regression – Hourly Aggregation



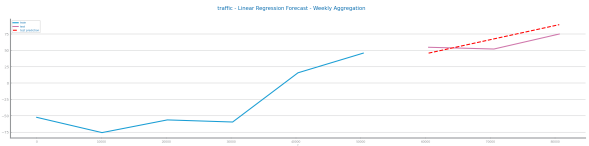
(f) Persistence Optim – Hourly Aggregation



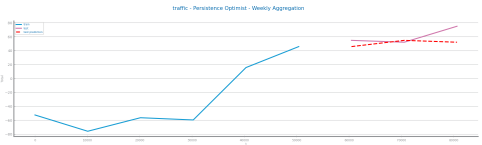
(g) Linear Regression – Daily Aggregation



(h) Persistence Optim – Daily Aggregation

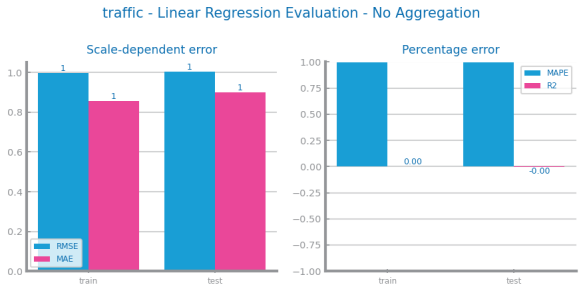


(i) Linear Regression – Weekly Aggregation

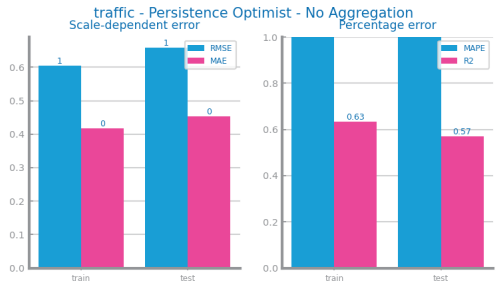


(j) Persistence Optim – Weekly Aggregation

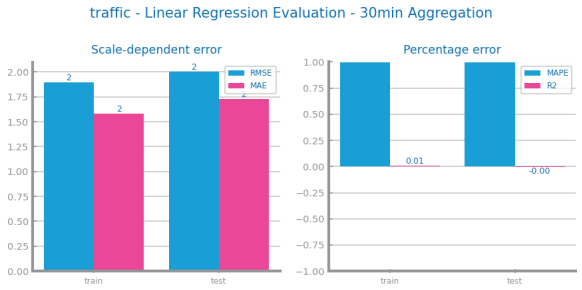
Figure 79: Forecasting plots for Linear Regression and Persistence Optim after different aggregations



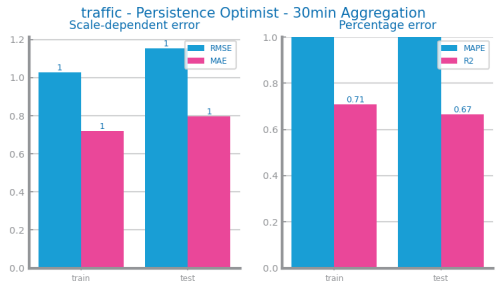
(a) Linear Regression Evaluation – No Aggregation



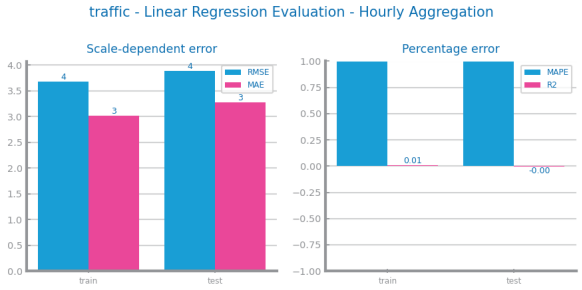
(b) Persistence Optim Evaluation – No Aggregation



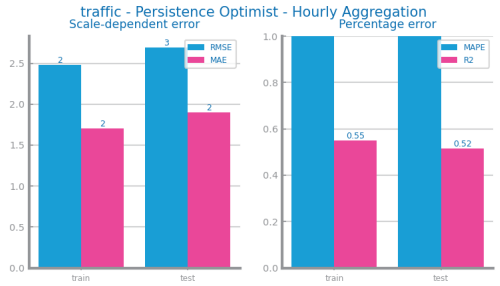
(c) Linear Regression Evaluation – 30min Aggregation



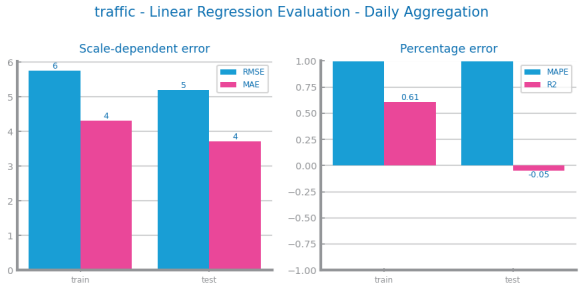
(d) Persistence Optim Evaluation – 30min Aggregation



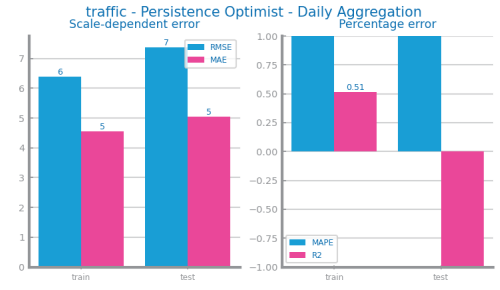
(e) Linear Regression Evaluation – Hourly Aggregation



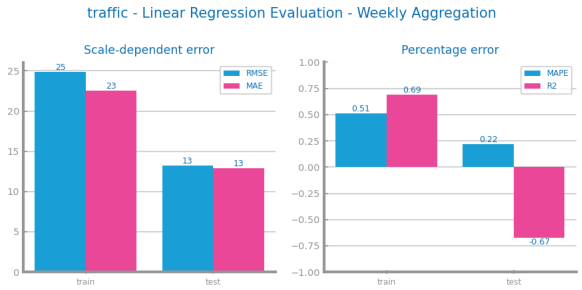
(f) Persistence Optim Evaluation – Hourly Aggregation



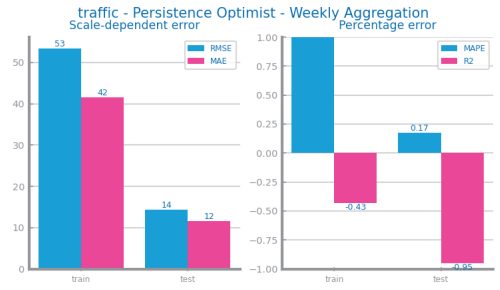
(g) Linear Regression Evaluation – Daily Aggregation



(h) Persistence Optim Evaluation – Daily Aggregation



(i) Linear Regression Evaluation – Weekly Aggregation



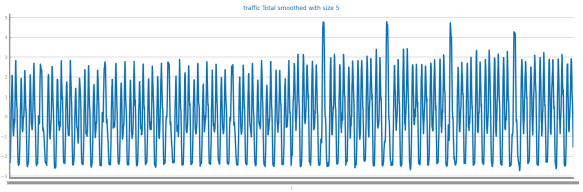
(j) Persistence Optim Evaluation – Weekly Aggregation

Figure 80: Evaluation results for Linear Regression and Persistence Optim after different aggregations

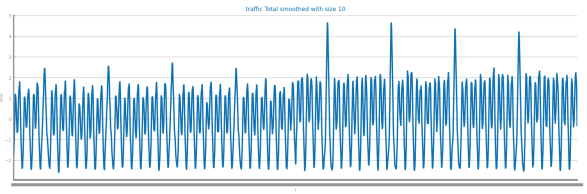
Smoothing

Shall describe the results of applying smoothing transformations over both datasets, and identifying the best result to proceed. **Shall not exceed 300 characters.**

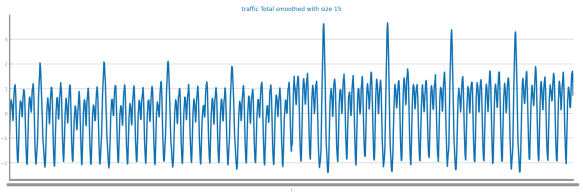
(a) Smoothing Size 3



(b) Smoothing Size 5

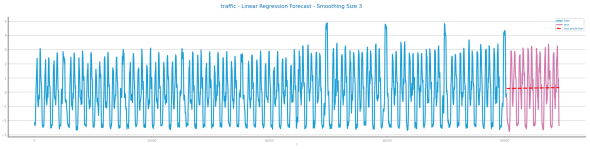


(c) Smoothing Size 10



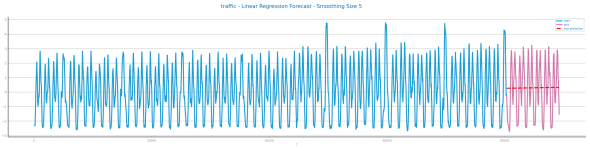
(d) Smoothing Size 15

Figure 81: Time series plots after applying moving average smoothing with different window sizes



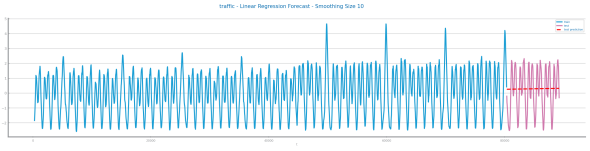
(a) Linear Regression – Size 3

(b) Persistence Optim – Size 3

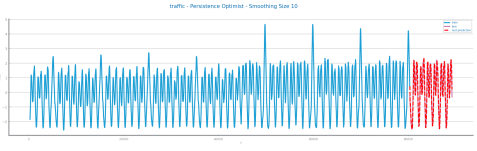


(c) Linear Regression – Size 5

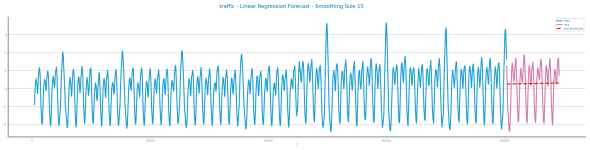
(d) Persistence Optim – Size 5



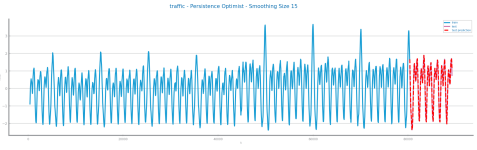
(e) Linear Regression – Size 10



(f) Persistence Optim – Size 10

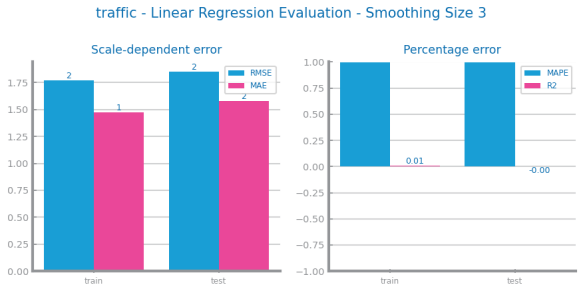


(g) Linear Regression – Size 15

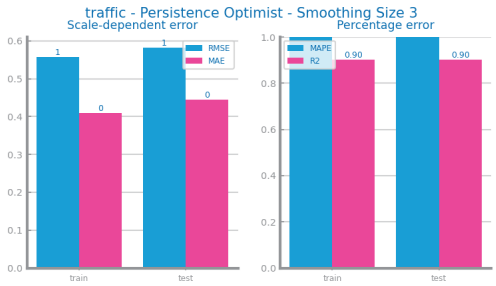


(h) Persistence Optim – Size 15

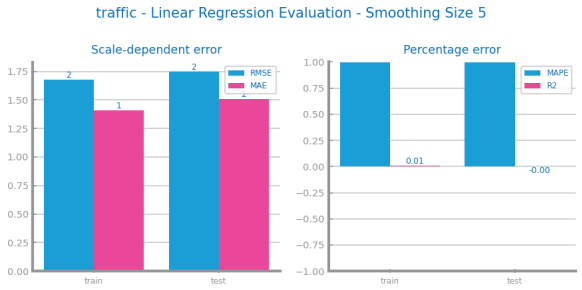
Figure 82: Forecasting plots for Linear Regression and Persistence Optim after different smoothing window sizes



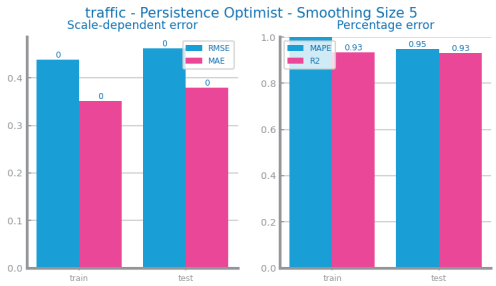
(a) Linear Regression Evaluation – Size 3



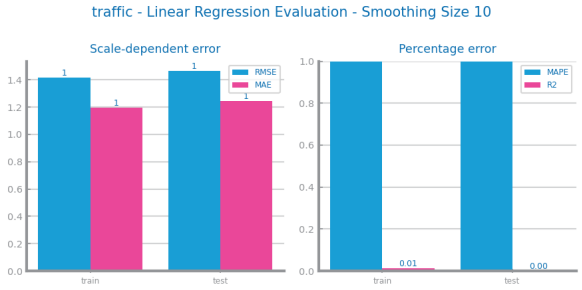
(b) Persistence Optim Evaluation – Size 3



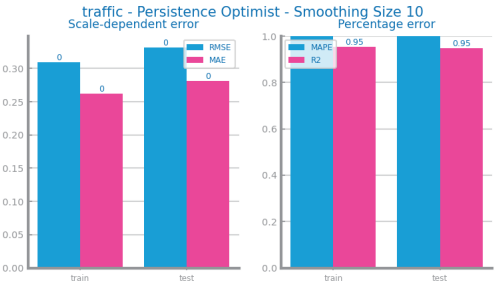
(c) Linear Regression Evaluation – Size 5



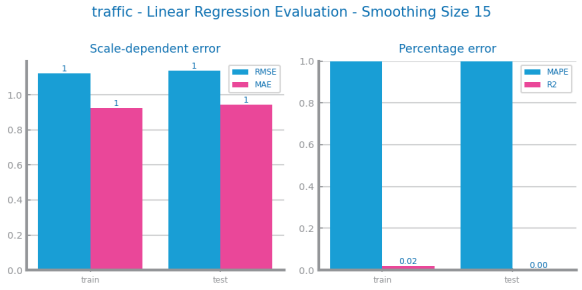
(d) Persistence Optim Evaluation – Size 5



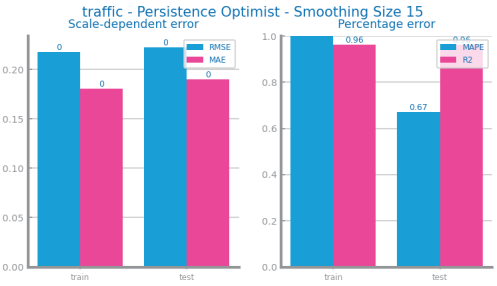
(e) Linear Regression Evaluation – Size 10



(f) Persistence Optim Evaluation – Size 10



(g) Linear Regression Evaluation – Size 15



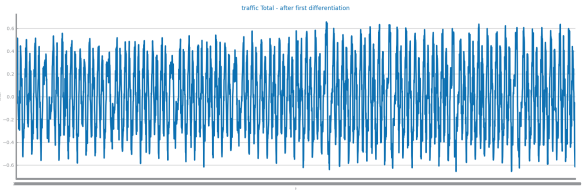
(h) Persistence Optim Evaluation – Size 15

Figure 83: Evaluation results for Linear Regression and Persistence Optim after different smoothing window sizes

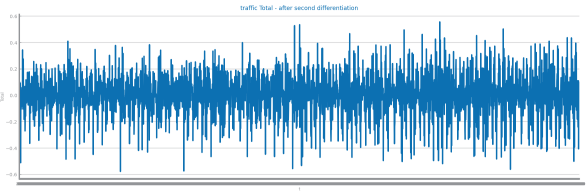
Differentiation

Shall describe the results of applying two consecutive differentiation of both datasets, and identifying the best result to proceed. **Shall not exceed 300 characters.**

(a) No Differentiation

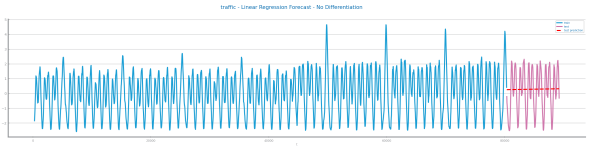


(b) First Differentiation

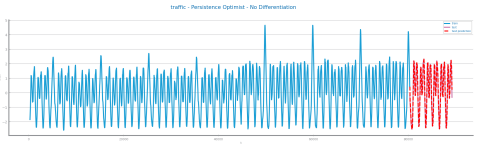


(c) Second Differentiation

Figure 84: Time series plots after applying zero, first, and second differentiation

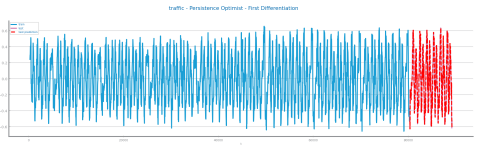


(a) Linear Regression – No Differentiation

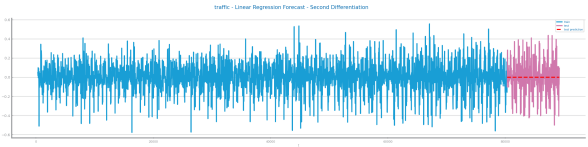


(b) Persistence Optim – No Differentiation

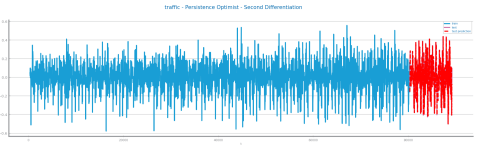
(c) Linear Regression – First Differentiation



(d) Persistence Optim – First Differentiation

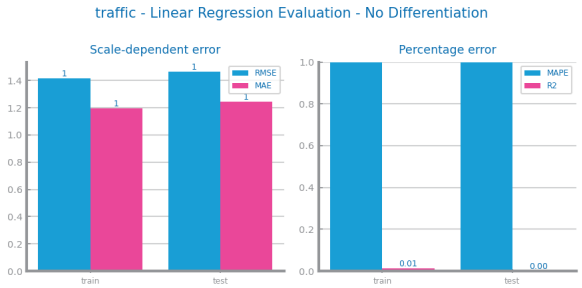


(e) Linear Regression – Second Differentiation

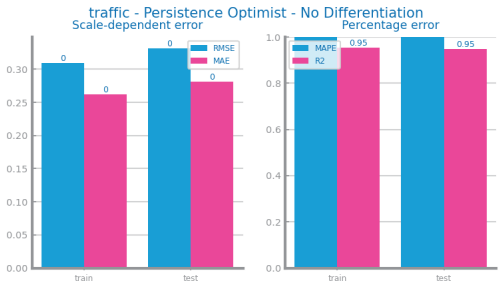


(f) Persistence Optim – Second Differentiation

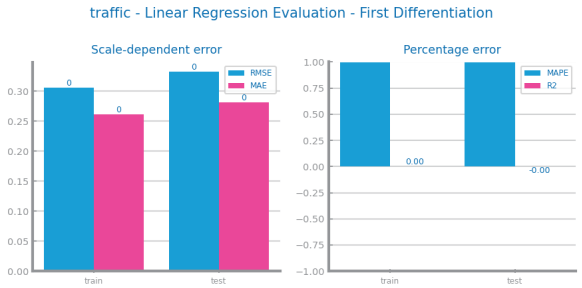
Figure 85: Forecasting plots for Linear Regression and Persistence Optim after different levels of differentiation



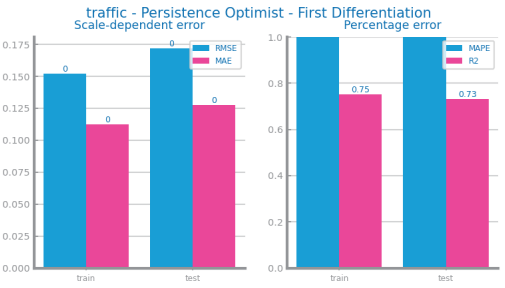
(a) Linear Regression Evaluation – No Differentiation



(b) Persistence Optim Evaluation – No Differentiation

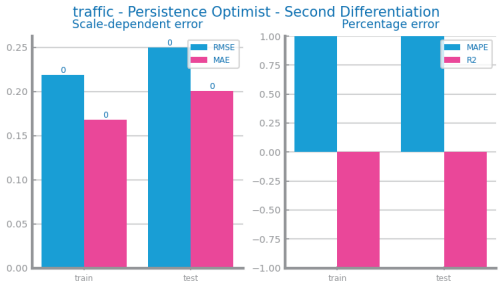


(c) Linear Regression Evaluation – First Differentiation



(d) Persistence Optim Evaluation – First Differentiation

(e) Linear Regression Evaluation – Second Differentiation

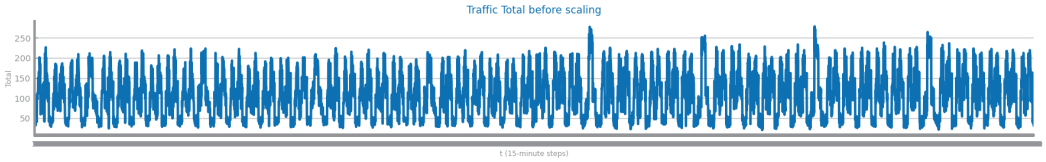


(f) Persistence Optim Evaluation – Second Differentiation

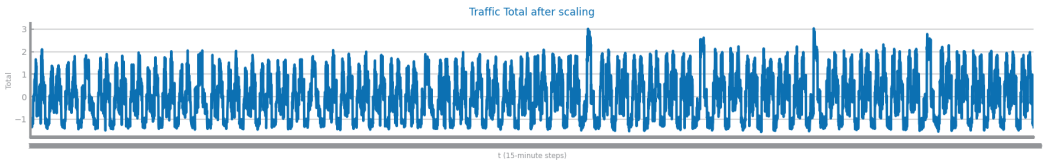
Figure 86: Evaluation results for Linear Regression and Persistence Optim after different levels of differentiation

Scaling

Shall describe the results of applying other transformations over both datasets, and identifying the best result to proceed. **Shall not exceed 500 characters.**



(a) Before scaling



(b) After StandardScaler()

Figure 87: Effect of StandardScaler on the original 15-minute time series

7 MODELS' EVALUATION

Shall be used to summarise the transformations done over the original time series. **Shall not exceed 500 characters.**

Exponential Smoothing Model

Exponential Smoothing was applied with varying smoothing parameter alpha. The hyperparameter study identified the optimal alpha that maximises R^2 on the validation set. The best model achieves solid performance on the test set, with predictions closely following the actual traffic patterns after the selected data transformations. **(312 characters)**

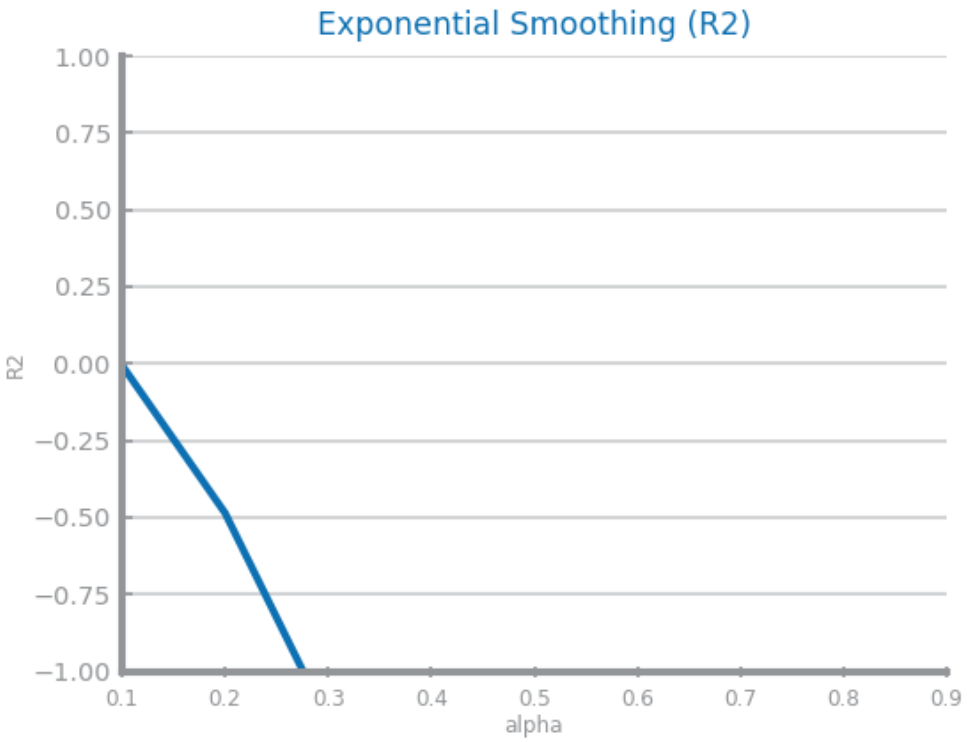


Figure 88: Hyperparameter study: R^2 as a function of alpha for Exponential Smoothing

Figure 89: Forecasting plots obtained with the best Exponential Smoothing model (predictions in red vs actual test data in pink)

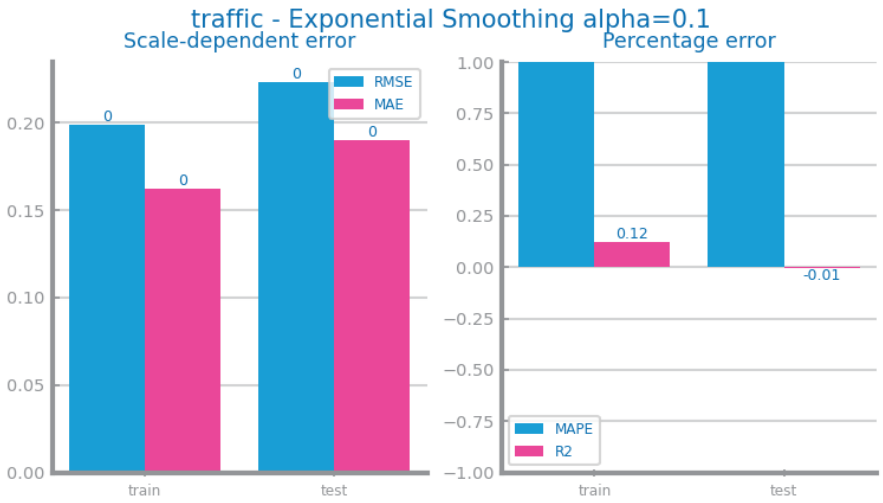


Figure 90: Performance metrics of the best Exponential Smoothing model (RMSE, MAE, MAPE, R^2)

Multi-layer Perceptrons Model

A Multi-layer Perceptron (MLP) was trained with different hidden layer configurations. The hyperparameter search showed convergence of R^2 across architectures. The selected model provides competitive forecasting accuracy on the transformed traffic time series. **(278 characters)**

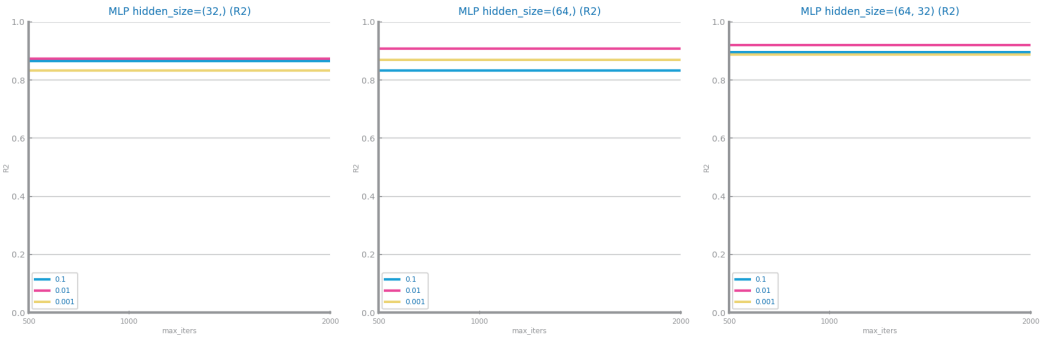


Figure 91: Hyperparameter study: R^2 convergence for different MLP hidden layer configurations

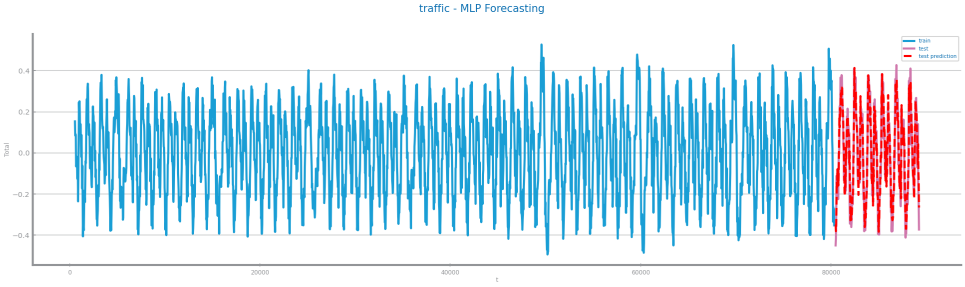


Figure 92: Forecasting plots obtained with the best MLP model (predictions in red vs actual test data in pink)

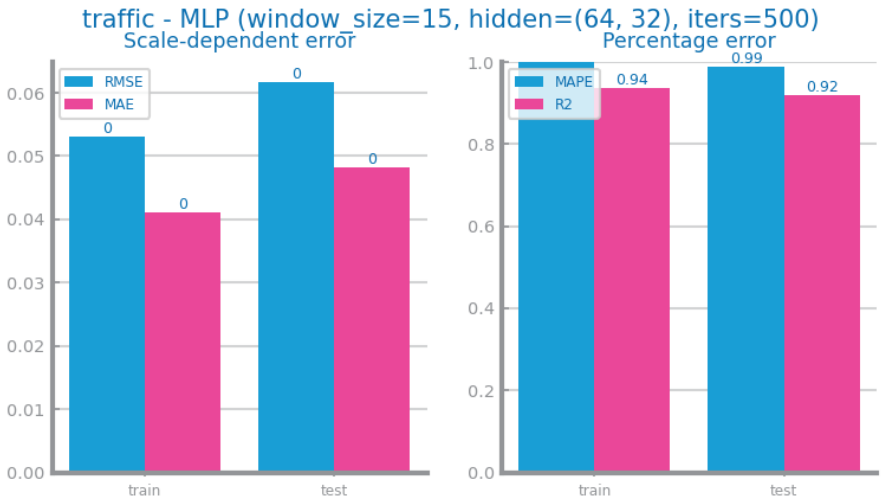


Figure 93: Performance metrics of the best MLP model (RMSE, MAE, MAPE, R^2)

ARIMA Model

ARIMA models were evaluated through a grid search over orders (p,d,q). The best univariate configuration (p=5, d=1, q=7) was identified using validation performance. The model captures the stationary patterns effectively after differentiation, delivering reliable short-term forecasts on the traffic series. **(298 characters)**

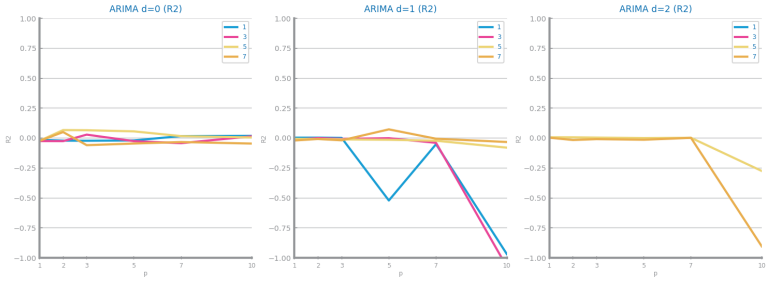


Figure 94: Hyperparameter study: best ARIMA configuration (p=5, d=1, q=7) – univariate

Figure 95: Forecasting plots obtained with the best ARIMA model (predictions in red vs actual test data in pink) – univariate

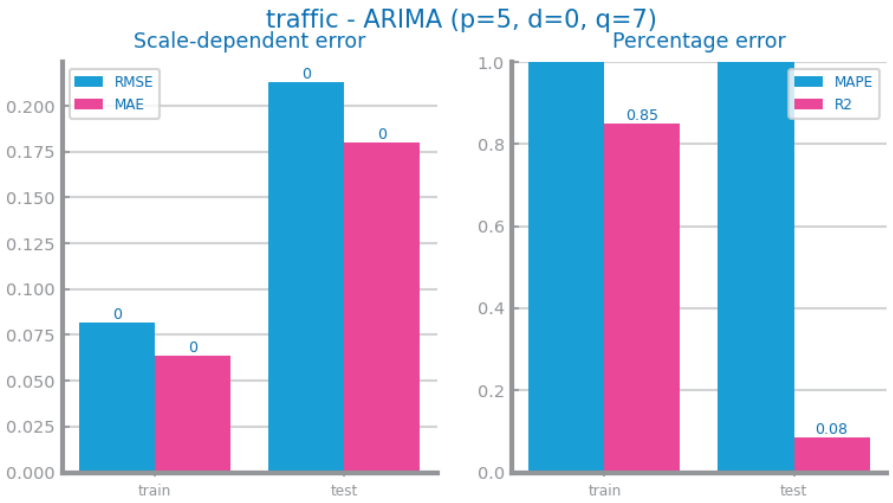


Figure 96: Performance metrics of the best ARIMA model (RMSE, MAE, MAPE, R^2) – univariate

LSTMs Model

Long Short-Term Memory networks were tuned over sequence length, hidden units, and training epochs. The optimal configuration (sequence length=4, hidden=25, epochs=2100) was selected based on validation R^2. The deep learning approach excels at capturing non-linear dependencies in the transformed traffic data. **(326 characters)**

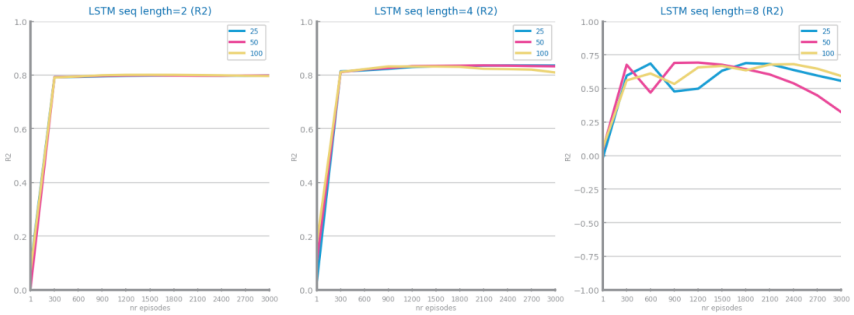


Figure 97: Hyperparameter study: best LSTM configuration (sequence length=4, hidden=25, epochs=2100) – univariate

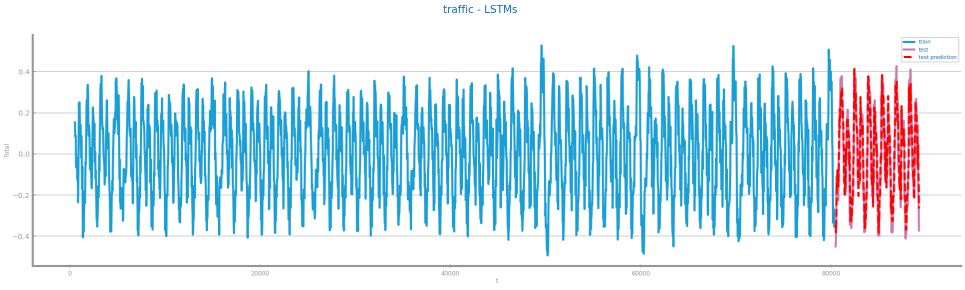


Figure 98: Forecasting plots obtained with the best LSTM model (predictions in red vs actual test data in pink) – univariate

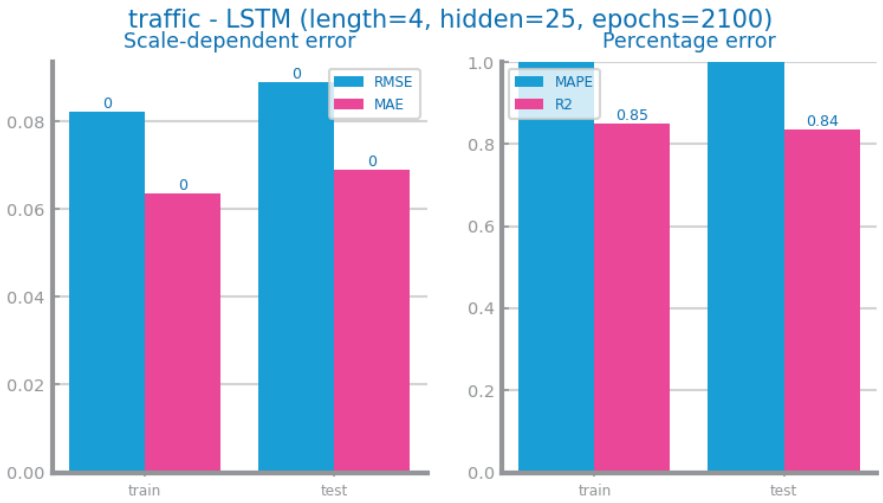


Figure 99: Performance metrics of the best LSTM model (RMSE, MAE, MAPE, R^2) – univariate

8 CRITICAL ANALYSIS

Shall be used to present a summary of the results achieved with the different forecasting techniques, and the impact of the different preparation tasks on their performance. A critical assessment of the best models shall be presented, clearly stating if the models seem to be good enough for the problem at hand. Additional charts may be presented here. **Shall not exceed 2000 characters.**