TÉCNICO LISBOA

DS 2022

Data Science Project

Team nr: Insert here	**Student 1:** Insert here	**IST nr:** Insert here
	Student 2: Insert here	**IST nr:** Insert here
	Student 3: Insert here	**IST nr:** Insert here

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 200 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 variable boxplots for dataset 1

Figure 6: Single variable boxplots s for dataset 2

Figure 7: Histograms for dataset 1

Figure 8: Histograms 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 200 characters.**

Figure 9: Granularity analysis for dataset 1

Figure 10: 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 300 characters.**

Figure 11: Sparsity analysis for dataset 1

Figure 12: Sparsity analysis for dataset 2

Figure 13: Correlation analysis for dataset 1

Figure 14: Correlation analysis for dataset 2

2 DATA PREPARATION

Variables Encoding

Shall contain all relevant information respecting to the transformation of variables, including *dummification*. 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.**

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 200 characters.**

Figure 15: Missing values imputation results with different approaches for dataset 1

Figure 16: 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 200 characters.**

Figure 17: Outliers imputation results with different approaches for dataset 1

Figure 18: 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 19: Scaling results with different approaches for dataset 1

Figure 20: 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 200 characters.**

Figure 21: Balancing results with different approaches for dataset 1

Figure 22: 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 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 200 characters.**

Figure 23: Feature selection of redundant variables results with different parameters for dataset 1

Figure 24: Feature selection of redundant variables results with different parameters for dataset 2

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 25: Principal components analysis and feature extraction results for dataset 1

Figure 26: Principal components analysis and feature extraction results for dataset 2

Feature Generation (optional)

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 summarize all variables generated and the formula used to derive them (in a table). **Shall not exceed 300 characters.**

Figure 27: Feature generation results for dataset 1

Figure 28: 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 300 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 29: Naïve Bayes alternatives comparison for dataset 1

Figure 30: Naïve Bayes alternative comparison for dataset 2

Figure 31: 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 parameterizations. 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 400 characters.**

Figure 32: KNN different parameterizations comparison for dataset 1

Figure 33: KNN different parameterizations comparison for dataset 2

Figure 34: KNN overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 35: KNN best model results for dataset 1 (left) and dataset 2 (right)

Decision Trees

Shall be used to present the results achieved through different parameterizations 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 36: Decision Trees different parameterizations comparison for dataset 1

Figure 37: Decision Trees different parameterizations comparison for dataset 2

Figure 38: Decision Trees overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 39: Decision trees best model results for dataset 1 (left) and dataset 2 (right)

Figure 40: Best tree for dataset 1

Figure 41: Best tree for dataset 2

Random Forests

Shall be used to present the results achieved through different parameterizations 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 42: Random Forests different parameterizations comparison for dataset 1

Figure 43: Random Forests different parameterizations comparison for dataset 2

Figure 44: Random Forests overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 45: Random Forests best model results for dataset 1 (left) and dataset 2 (right)

Figure 46: Random Forests variables importance for dataset 1 (left) and dataset 2 (right)

Gradient Boosting

Shall be used to present the results achieved through different parameterizations 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 47: Gradient boosting different parameterizations comparison for dataset 1

Figure 48: Gradient boosting different parameterizations comparison for dataset 2

Figure 49: Gradient boosting overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 50: Gradient boosting best model results for dataset 1 (left) and dataset 2 (right)

Figure 51: 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 parameterizations 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. In particular by analysing the *loss_curve_* available at the end of each train. Shall be used to present the evaluation of the best model achieved. **Shall not exceed 500 characters.**

Figure 52: MLP different parameterizations comparison for dataset 1

Figure 53: MLP different parameterizations comparison for dataset 2

Figure 54: MLP overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 55: Loss curves analysis for dataset 1 (left) and dataset 2 (right)

Figure 56: 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 modeling 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 Granularity

May be used to identify the most atomic granularity and two other different granularities to consider. **Shall not exceed 300 characters.**

Figure 57: Time series 1 at the most granular detail

Figure 58: Time series 1 at the second chosen granularity

Figure 59: Time series 1 at the third chosen granularity

Figure 60: Time series 2 at the most granular detail

Figure 61: Time series 2 at the second chosen granularity

Figure 62: Time series 2 at the third chosen granularity

Data Distribution and Stationarity

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

Figure 63: Boxplot(s) for time series 1

Figure 64: Boxplot(s) for time series 2

Figure 65: Histogram(s) for time series 1

Figure 66: Histogram(s) for time series 2

Figure 67: Stationarity study for time series 1

Figure 68: Stationarity study for time series 2

6 DATA TRANSFORMATION

Aggregation

Shall describe the results of applying the persistence model over the three different aggregations over both datasets, and identifying the granularity chosen to proceed. **Shall not exceed 200 characters.**

Figure 69: Forecasting plots after different aggregations on time series 1

Figure 70: Forecasting results after different aggregations on time series 1

Figure 71: Forecasting plots after different aggregations on time series 2

Figure 72: Forecasting results after different aggregations on time series 2

Smoothing

Shall describe the results of applying the persistence model over different smoothing transformations over both datasets, and identifying the best result to proceed. **Shall not exceed 200 characters.**

Figure 73: Forecasting plots after different smoothing parameterizations on time series 1

Figure 74: Forecasting results after different smoothing parameterizations on time series 1

Figure 75: Forecasting plots after different smoothing parameterizations on time series 2

Figure 76: Forecasting results after different smoothing parameterizations on time series 2

Differentiation

Shall describe the results of applying the persistence model over two consecutive differentiation of both datasets, and identifying the best result to proceed. **Shall not exceed 200 characters.**

Figure 77: Forecasting plots after first and second differentiation of time series 1

Figure 78: Forecasting results after first and second differentiation of time series 1

Figure 79: Forecasting plots after first and second differentiation of time series 2

Figure 80: Forecasting results after first and second differentiation of time series 2

7 MODELS' EVALUATION

Shall be used to summarize the transformations done over the original time series. **Shall not exceed 200 characters.**

Simple Average Model

Shall be used to present the results achieved through the simple average model. **Shall not exceed 200 characters.**

Figure 81: Forecasting plots obtained with Simple Average model over time series 1

Figure 82: Forecasting plots obtained with Simple Average model over time series 2

Persistence Model

Shall be used to present the results achieved through the persistence model. **Shall not exceed 200 characters.**

Figure 83: Forecasting plots obtained with Persistence model over time series 1

Figure 84: Forecasting plots obtained with Persistence model over time series 2

Rolling Mean Model

Shall be used to present the results achieved through the rolling mean forecasting algorithms. **Shall not exceed 500 characters.**

Figure 85: Forecasting study over different parameterizations of the rolling mean algorithm over time series 1

Figure 86: Forecasting plots obtained with the best parameterization of rolling mean algorithm, over time series 1

Figure 87: Forecasting results obtained with the best parameterization of rolling mean algorithm, over time series 1

Figure 88: Forecasting study over different parameterizations of the rolling mean algorithm over time series 2

Figure 89: Forecasting plots obtained with the best parameterization of rolling mean algorithm, over time series 2

Figure 90: Forecasting results obtained with the best parameterization of rolling mean algorithm, over time series 2

ARIMA Model

Shall be used to present the results achieved through the ARIMA forecasting algorithms. **Shall not exceed 500 characters.**

Figure 91: Forecasting study over different parameterizations of the ARIMA algorithm over time series 1

Figure 92: Forecasting plots obtained with the best parameterization of ARIMA algorithm, over time series 1

Figure 93: Forecasting results obtained with the best parameterization of ARIMA algorithm, over time series 1

Figure 94: Forecasting study over different parameterizations of the ARIMA algorithm over time series 2

Figure 95: Forecasting plots obtained with the best parameterization of ARIMA algorithm, over time series 2

Figure 96: Forecasting results obtained with the best parameterization of ARIMA algorithm, over time series 2

LSTMs Model

Shall be used to present the results achieved through LSTMs. **Shall not exceed 500 characters.**

Figure 97: Forecasting study over different parameterizations of LSTMs over time series 1

Figure 98: Forecasting plots obtained with the best parameterization of LSTMs, over time series 1

Figure 99: Forecasting results obtained with the best parameterization of LSTMs, over time series 1

Figure 100: Forecasting study over different parameterizations of the LSTMs over time series 2

Figure 101: Forecasting plots obtained with the best parameterization of LSTMs, over time series 2

Figure 102: Forecasting results obtained with the best parameterization of LSTMs, over time series 2

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.**