Project Name

Technical Report

Table of Contents

[Business Problem 3](#_Toc42527016)

[Further Clarifications 3](#_Toc42527017)

[Assumptions 3](#_Toc42527018)

[Data Description 4](#_Toc42527019)

[Data Source 4](#_Toc42527020)

[Data Extraction 4](#_Toc42527021)

[Features Analysis 5](#_Toc42527022)

[Outliers 5](#_Toc42527023)

[Missing Values 5](#_Toc42527024)

[Target Variable 5](#_Toc42527025)

[Preliminary Analysis 6](#_Toc42527026)

[Metric Definition 6](#_Toc42527027)

[Baseline estimation 6](#_Toc42527028)

[Data-Driven Expectations 6](#_Toc42527029)

[Machine Learning Exploration 7](#_Toc42527030)

[Model 1 Evaluation 7](#_Toc42527031)

[Data Preparation 7](#_Toc42527032)

[Results 7](#_Toc42527033)

[Model 2 Evaluation 7](#_Toc42527034)

[Data Preparation 7](#_Toc42527035)

[Results 7](#_Toc42527036)

[Model N… 7](#_Toc42527037)

[Model Comparison 7](#_Toc42527038)

[Model Tuning 7](#_Toc42527039)

[Results 8](#_Toc42527040)

[Interpretation of Output 8](#_Toc42527041)

[Future Developments 8](#_Toc42527042)

Change Log

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Author | Version | Notes |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Business Problem

\*\*\* Description of Business Problem as per project agreement or contract

## Further Clarifications

\*\*\* Notes on subsequent meetings, requests, evolution of demand during the project

## Assumptions

\*\*\* A list of assumption made on the details of the project due to lack of prior information or knowledge or due to the necessity of imposing constraint

# Data Description

\*\*\* High-level description of the data to be used and from which domain. Size of the dataset, number of rows and columns (and so problem’s dimensionality) used in the final solution to be delivered (a mention to the original set of data taken into account is a plus).

## Data Source

\*\*\* Description of the data source: e.g. SQL Server DB available in a specific location, reachable at a certain URL, etc… Describe what the source system is there for, why is it keeping data there?

## Data Extraction

\*\*\* Explain the query used to extract the data, it may bring a lot of implicit semantics in the data as well as some transformations with it… In case of OSISoft Systems for instance, it would valuable to describe the frequency resolution of the data tags in use, different time-based aggregation carry different semantics and information as well as it is important to keep track of it in case of changes in such kind of resolution.

# Features Analysis

\*\*\* For each of the columns in the dataset used in the delivered Solution provide: name, data type, range of values, if there are missing values, mean, 5-value-summary (min, 1st quartile, median, 3rd quartile, max), (univariate) outliers

## Outliers

\*\*\* Describe how outliers were detected into data and how the project is supposed to deal with them (remove them from the dataset, replace them with the median value, leave them as they are, etc…)

## Missing Values

\*\*\* Describe if the dataset contains missing values or incomplete data and how to deal with those

## Target Variable

\*\*\*Provide the same level of information used for the feature columns in the dataset in order to properly describe the target variable

# Preliminary Analysis

\*\*\* Perform cross-correlation, check for auto-correlation, etc… according to the problem. Include all the possible explanatory diagrams with comments

\*\*\* Discuss if the problem is well-represented by the available features or not. E.g. if few features are highly correlated with the target variable you could think about a Feature Selection or dimensionality reduction. If all the features are highly correlated, maybe the dimensionality is too low and you need more columns…

## Metric Definition

\*\*\* According to the problem define 3 metrics that will describe the solutions to the problem from slightly different angles (e.g. precision-recall curve, F1-Score, accuracy), explaining it. Then pick one to be used for model comparison in the next steps.

## Baseline estimation

\*\*\* Perform a statistical analysis to derive the baseline of the problem. E.g. there are 100 rows, 60 are positive and 40 are negative, the baseline is 60%

\*\*\* Define the learning baseline using naïve methods like 1NN or Naïve Bayes. They should perform better than the analysed distribution of data, and will provide a concrete learning-oriented baseline to derive expectations from.

## Data-Driven Expectations

\*\*\* Define a solid expectation from the ML process in terms of the metric chosen and based on the above estimated baselines

# Machine Learning Exploration

\*\*\* Given the problem, the data and the analysis performed until now, one or more Machine Learning approaches could have been used to build up the final Solution and try to perform better than the baseline. Explain reasons why and point of strength and weakness each of the ML approaches used in the Solution.

## Model 1 Evaluation

### Data Preparation

\*\*\* Describe if the dataset require some modification due to the biases of the ML algorithm

### Results

\*\*\* Present results of running a 10-fold cross-validation!

## Model 2 Evaluation

### Data Preparation

\*\*\* Describe if the dataset require some modification due to the biases of the ML algorithm

### Results

\*\*\* Present results of running a 10-fold cross-validation!

## Model N…

\*\*\* As above…

## Model Ensembling

\*\*\* If more than a single ML model is used in the Solution describe how they are thought to be used together and how they cooperate to build up the final Solution. Different options are available: ensemble models, model pipelines, voting, etc…

## Model Tuning

\*\*\* Explain how to tune hyper-parameters

\*\*\* Consider feature engineering, etc…

# Results

\*\*\* Summarize final scores and performances achieved by the chosen approach in the context of a fully-fledged experimental setup (10-fold cross validation, large dataset, etc…)

## Interpretation of Output

\*\*\* Describe in which format and formalism the model provides scores in output and how to interpret the results in the context of the initial Business Problem Definition.

## Future Developments

\*\*\* Highlight limitations and possible lines of improvement for future project iterations, if any.