FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Performance Management Analytics For The Automotive Industry: An Empirical Study

João Carlos Dias Correia Pinto

WORKING VERSION

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Carlos Manuel Milheiro de Oliveira Pinto Soares

Co-supervisor: Filipe David Maia Ferreira

August 3, 2016

Resumo

A indústria automóvel está altamente dependente de informação e empresas como a *Volkswagen Autoeuropa* estão permanentemente a recolher dados na forma de métricas para monitorizar e controlar o estado da empresa. No entanto, essas métricas são referentes ao passado, o que impede uma abordagem estratégica proativa. Este projeto usa *Machine Learning, Data Mining, Estatística* e *Inteligência Artificial* para prever o valor que um indicador de eficiência irá assumir no dia seguinte. Foi realizado um estudo empírico, comparando alguns algoritmos: *Random Forest, Partial Least Squares, Artificial Neural Networks, M5, Support Vector Machines* e *k-Nearest Neighbors* foram testados. Foram também comparados treze conjuntos diferentes de variáveis preditivas.

Abstract

The automotive industry heavily relies on information and companies such as *Volkswagen Autoeuropa* are permanently acquiring data in the form of metrics to monitor and control the state of the company. However, those metrics are referred to the past, preventing a proactive strategic approach. This project uses Machine Learning, Data Mining, Statistics and Artificial Intelligence to predict the value that a major efficiency indicator will assume one day ahead. We carried out an empirical study, comparing several algorithms: *Random Forest, Partial Least Squares, M5, Artificial Neural Network, Support Vector Machines* and *K Nearest Neighbors* were tested. Thirteen different sets of predictive variables were tested.

Acknowledgement

The completion of this project would not be possible without the help, assistance and support of many people whose names may not all be enumerated. However, their contribution was not forgotten and a deep appreciation remains untouchable.

A special thanks is directed to my supervisors Carlos Soares and Filipe Ferreira for all the support and insight given to me during the development of this project.

My friends also played a decisive role, backing me and giving me strength to never give up.

All my family supported me until the end of my degree. A special word of appreciation must be directed to my uncle António and my aunt Conceição. I can't thank enough to my parents Carlos and Maria João for every single moment they stood by me and helped me, being more than just parents and making me who I am today. Unfortunately, my grandmother could not see this day come, but will always have a special place in my heart.

João Correia Pinto

"If you're going to try,
go all the way.
There is no other feeling like
that.
You will be alone with the gods,
and the nights will flame with fire.
You will ride life straight to
perfect laughter, it's
the only good fight there is."

Henry Bukowski

Contents

List of Figures

List of Tables

Abbreviations and Symbols

ANN	Artificial Neural Networks
BSC	Balanced Scorecard
CRISP-DM	Cross Industry Process for Data Mining
DB	Database
DM	Data Mining
EDA	Exploratory Data Analysis
ETL	Extract, Transform and Load
HPU	Hours Per Unit
IDE	Integrated Development Environment
k-NN	k-Nearest Neighbors
KPI	Key Performance Indicators
MAE	Mean Absolute Error
ML	Machine Learning
PCA	Principal Component Analysis
PLS	Partial Least Squares
PME	Performance Measurement Engine
PMS	Performance Management System
RF	Random Forest
RMSE	Root Mean Square Error
SBU	Strategic Business Units
SVM	Support Vector Machines

Chapter 1

Introduction

The business world and more specifically the automotive industry is on permanent change needing to accelerate their strategic performance. In order to be competitive, companies have to access information to be able to act under different conditions. In this paradigm, the automotive industry is permanently collecting data about the business that is used to calculate high level performance indicators. Those are organized on Performance Management Systems (PMS) and help on the decision making process, giving information to monitor and control the state of the company.

On the other hand, a critical misalignment can be observed between strategic and operational layers. Enterprises apply reactive optimization and improvement methods (contrarily to predictive ones) based on feedback information, strictly related with past actions. That happens due to the lack of knowledge and support tools to forecast future manufacturing systems' behaviour, basing their planning processes in oversimplified approaches.

Due to this fact, the necessity of predicting future values for key performance indicators (KPI) becomes more and more clear. This feature grants a company the capacity of acting in anticipation. It closes the gap between strategic plan and operational execution, opening a perspective that allows the strategic layer to control the operational one with anticipation instead of acting after events occur. Predictive KPIs are obtained recurring to Data Mining (DM), Machine Learning (ML), Statistics and Artificial Intelligence.

A predictive performance indicator model using data from the automotive industry was previously created ("Performance Management System Analytics for the Automotive Industry") [2]. It was a preliminary work that was able to develop a predictive model by testing and comparing five different algorithms. However, the author recognized that the project should have continuity and be improved. The plant was only considered as a whole without any domain separation. The data set should be larger and with more external variables to make the model more reliable. A more diverse set of algorithms should also be considered. A phase of deployment should also be implemented, integrating the developed model with a performance measurement engine (PME) tool [2].

1.1 Objectives

The main objective of this project is testing and comparing different algorithms for the prediction of the *Hours per Unit* KPI with a data set provided by *Volkswagen Autoeuropa*. It is divided into two main phases:

The first one is based on the project "Performance Management System Analytics for the Automotive Industry" [2]. It consolidates that approach, testing and analysing more algorithms with more complete data sets to improve the results. Those data sets contemplate one entire year of observations instead of four months. The observations are also separated by car models (instead of only analysing the plant as a whole).

The second phase focuses on a feature engineering problem. With the same information that existed before, a new dataset is built in order to get one that is better suited for the desired outputs. It demands a thorough analysis before the attempts and extensive testing to verify if the accuracy increases at each experience.

Chapter 2

Literature Review

In this chapter, the basis of the project is presented as well as the previous projects related to this one. The concepts of *Performance Management System* (section 2.1), *Performance Measurement* (section 2.2), *Key Performance Indicator* (section 2.3), *Machine Learning* (section 2.4) and *Data Mining* (section 2.5) are explained. The intuition behind the algorithms used in the project is also given in section 2.8. Two projects that served as a basis to this one are summarized in section 2.9. This project deals with a time-series problem, treating it as a regression problem.

2.1 Performance Management System

A Performance Management System (PMS) is a permanent communication process that works as an assistant to accomplish strategic objectives of the organization, including the clarification of expectations, the identification and establishment of objectives, feedback and results verification [11]. It makes it possible to examine and discuss the individual contribution for the organization development [12] creating a consistent relationship between strategy, planning, implementation and controlling [13]. The clarification of expectations and objectives increases the team motivation and company's efficiency and growth, allowing a better organization performance [14] [15].

2.2 Performance Measurement Engine

Measuring performance is a necessary part of any PMS. It consists in measuring the system indicators to obtain the necessary quantitative data to understand if the critical objectives are being obtained in terms of efficiency and effectiveness. It provides methods and tools for measuring, monitoring and managing processes.

A part of the performance measurement system at *Volkswagen Autoeuropa* is a performance measurement engine (PME). It is an interface software responsible for the communication between the measurement system and the user.

The difference between a Performance Measurement System and a Performance Management System is that the first one does exclusively measurements while a Performance Management System chooses a particular course of action and studies how those actions relate to other departments and the overall achievement of company strategy [2][16][17].

2.2.1 Historical Development of Performance Measurement

There were three main stages for the development of the Performance Measurement that exists nowadays. Those are described below:

- **Stage 1** *(1850-1925):* Focused on the development of costs and management accounting that evolved from the older accounting systems that were revealed as insufficient due to the evolution of businesses into multiple plants and multi-division firms. The new systems were able to compare the performance between multiple sites and to monitor fluctuations in demand and production.

- **Stage 2** *(1974-1992):* The multi-dimensional performance measurement became necessary as well as the budgetary planning and control. In spite of it, the critics pointed that these plans lacked strategic focus and encouraged short-term thinking. Because of it, multi-dimensional frameworks were developed. The mind-set also changed during this time from “goods-producing” to “customer satisfying” and an emphasis was placed on non-financial performance measures.

- **Stage 3** *(1992-2000):* Development of strategy maps, business models and cause-effect diagrams. These newly developed models, an evolution on the balanced scorecard, translated the concept of "leading and lagging indicators" into a visual representation where each element of performance is linked to another. Organizations developed systems to statistically correlate the main drivers for success, using historic performance measurement data [2][18].

2.3 Key Performance Indicator (KPI)

The Key Performance Indicators are quantitative values that help measuring the progress of the company relative to its own objectives [19]. Those measurements allow the achievement of information related to the main performance dimensions, allowing its permanent monitoring and the immediate acknowledgement of the behavior of the organization.

Therefore, when defining a KPI, it is critical to properly specify the architecture of the system to be controlled, through the definition of the data structuring and grouping, as well as its outcomes when applicable.

To properly define a KPI, it is also important to specify its metric, which can be represented by a hierarchical tree, and target values.

The KPIs may be divided between KPI_0 and high level KPIs (KPI_{1+}). The KPI_0 are the ones that are directly extracted from the performance measurement engine. They are then used to calculate high level KPIs by mathematical operations defined previously.

The target KPI to be predicted in this project is Hours Per Unit (HPU).

2.3.1 The historical evolution of KPIs

The evolution of the KPIs have depended of the evolution of the business itself, which dynamics tend to accelerate over the time.

The development of KPIs have passed through 3 stages (waves) as it is described below:

- **First Stage**: Focused on what had happened historically to individual product lines and strategic business units (SBUs). The information was not centralized or structured as a unit. The Operational Managers forced to calculate, track and react to KPIs that were completely retrospective.

- **Second Stage**: During this phase the focus changed from a siloed product/SBU mind-set to an enterprise perspective, being the data integrated into data warehouses. The focus started to incorporate the customer, process and learning perspectives rather than just the financial aspects. The measurements did not only target the past results but also the present health of the company.

- **Third Stage:** This approach is not fully implemented on most companies and is the main focus of this project. It consists on changing from a reactive paradigm (on which the company tried to react to past data) to a proactive paradigm (where the data is predicted and the company tries to react to the predictions the best way possible). [2][1]

The historical evolution of KPIs is summarized in fig. 2.1.

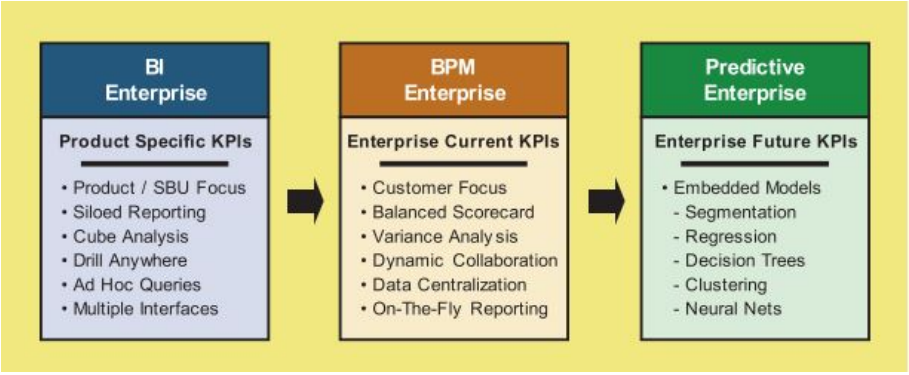


Figure 2.1: The Three stages of KPI Development [1]

2.4 Machine Learning

Machine Learning is a field of computer science. Its definition is not consensual. According to Arthur Samuel it is "Field of study that gives computers the ability to learn without being explicitly programmed" [20]. Another definition by Tom Mitchel for a well-posed learning problem is: "A computer program is said to learn from experience E (observed examples) with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E." [21]. In other words, the machine has the hability to learn a given task by experience and the more experience it has, the better its performance on accomplishing that task.

2.5 Data Mining

Data Mining is the process of analyzing a great quantity of data from different perspectives and summarizing it to find interesting padrons to obtain relevant information. It is an interdisciplinary subject that involves Database Systems, Data Wharehouse, Statistic, Machine Learning and Computing [22][23].

There are several types of DM algorithms being each one used according to the context [13]. More than one algorithm can be used on the same data set depending on the situation and more than one algorithm might be used for the same situation producing different results. A few examples of algorithms that might be used on KPI predictions are following presented:

- **Classification Algorithms:** predict one or more discrete variables, based on other attributes on the dataset

- **Regression Algorithms:** predict one or more continuous variables, based on other attributes on the dataset

2.5.1 Supervised Learning

The learning process is based on example pairs: an input and the respective output. A supervised learning algorithm is able to determine the value of a new output based on different input values.

2.5.2 Unsupervised Learning

In this type of learning, there are inputs but they are not associated to any output (they are not labeled). The learning algorithm as the function of finding some structure in the data (for example, dividing it on clusters).

2.5.3 Data Mining Tasks

2.5.3.1 Clustering

Clustering consists on grouping the similar data in the same group so that the data can have the greatest similitude degree inside a cluster and the lowest possible similitude degree outside that one. It makes it possible to analyse the data collectively instead of individually [23][24].

2.5.3.2 Association

This task consists on a method based on the discovery of interest relations and variables on big databases, describing and associating rules (strong rules) discovered on databases using different interest measures. [24]

2.5.4 Data Mining Methodology

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a structured process model used to implement DM projects creating a more accessible management.

It has four different levels of abstraction: [3][25]

1. **Phases:** A data mining project is broken down into six different phases. They are following explained and summarized in fig. 2.2.

   (a) **Business Understanding:** focused on understanding the objectives and requirements of the project from a business point-of-view and on converting that knowledge on the definition of a DM problem and on a preliminary plan.

   (b) **Data Understanding:** this phase collects the initial data and then performs activities that allow the familiarization with the data including the identification of the problems related to its quality and the acquirement of new perspectives about the data.

   (c) **Data Preparation:** necessary activities for the final data model building that serves as an input for the modeling tool. It changes the data set in order to avoid misleading results and obtain better results

   (d) **Modeling:** several modeling techniques are selected and applied and their parameters are calibrated to apropriate values.

   (e) **Evaluation:** the model is analyzed to assure that the project objectives are clear. Therefore, the process shall be reviewed and a decision must be made about how to use the DM results.

   (f) **Deployment:** several plans are developed to monitor and maintain the previously created model as well as to present it in a way that allows the end user to use it.

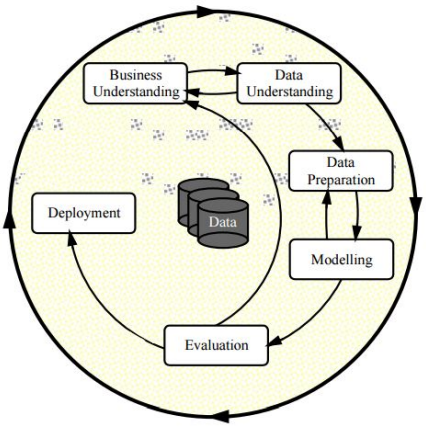


Figure 2.2: Phases in the DM process [2]

2. **Generic Tasks:** constitute each phase. They are called this way because they should be so general that it is possible to cover every DM possibility. They should be complete (in a way that all the DM process and applications are covered) and stable (to make the model adapta ble to non-predicted developments).

3. **Specialized Tasks:** describe how generic tasks action should be taken on specific situations.

4. **Process Instances:** compilation of action decisions and results of a DM application implementation.

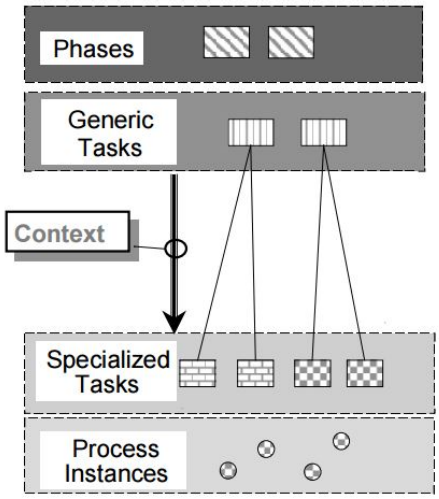


Figure 2.3: Four Level Breakdown of the CRISP-DM Methodology for Data Mining [3]

2.5.5 Data Mining Evaluation

To evaluate data mining models, the data set must be divided into a train and a test set. The first one is used to fit the parameters and learn. The model can not be tested with the training set to prove its capacity to generalize. For that reason, a test set that has not played any part in the formation of the classifier is necessary. Both sets come from the the same data set in order to minimize the effects of data discrepancies.

During the train phase, the training set may be divided into a train and a validation set. The validation set is used by the model to know how well it has performed on the training task and to know which is the best path to follow.

Once the evaluation is complete, all the data can be used to form the final classifier.

Usually, the larger the training data the better the classifier. The larger the test data the more accurate the error estimate. Due to this, it is important to find a good trade-off between the sizes of the train and the test set [26][27].

2.6 Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a statistical method for data description. Application of EDA techniques largely determines the types of other techniques which a data analyst can use to examine a given set of data. It includes measures of central tendency, dispersion measures and distribution measures [28]. A robust statistic is one that is not much affected by extreme values. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population [29]. In the following formulas, X_i is a variable, θ^2 is the variance, n is the number of observations, x_i is an observation, $\bar{X}_i$ is the mean value of a variable and s is the standard deviation of the sample.

2.6.1 Measures of central tendency

Measures of central tendency are the ones that try to describe a set of data by identifying the central position within that set of data. Examples of measures of central tendency are:

- **Mean:** It is not a robust value.
- **Median:** used to observe the symmetry of a distribution or the existence of outliers. It is more robust than the mean.
- **Trimmed Mean:** Discards the examples on the extremes of an ordered sequence. It is necessary to define the percentage of examples to eliminate on each extreme (percentile). It adds robustness to the mean.
- **Quartiles:** divides the data into four parts:

  1st Quartile (Q1): Value for which 25% of the values are inferior to this one.

  2nd Quartile (Q2): Median

3rd Quartile (Q3): Value for which 75% of the values are inferior to this one.

- **Box Plots:** Graphical description of the data dividing it on quartiles as showed on fig. 2.4

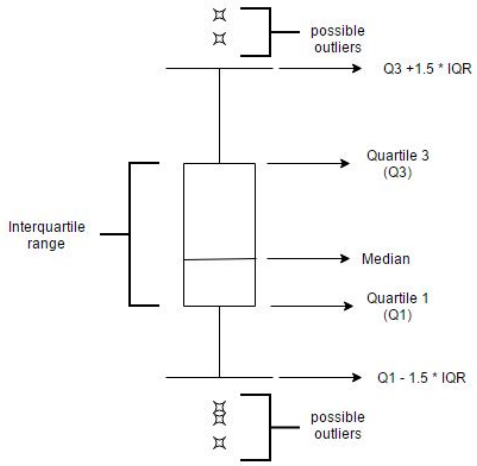


Figure 2.4: Box plot model

2.6.2 Measures of Dispersion

Measures of dispersion are the ones that describe a set of data by showing how spread that data is. Examples of measures of dispersion are range, variance, standard deviation, average absolute deviation and interquartille range.

- **Range:** difference between the maximum and minimum value
- **Variance:** measures how spread out the numbers are from the mean.

$$\theta^2(X_i) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X}_i) \tag{2.1}$$

- **Standard Deviation:** square root of the variance.
- **Average Absolute Deviation:** average of the absolute deviations from a central point.

$$AAD(X_i) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{X}_i| \tag{2.2}$$

- **Interquartile Range:** difference between the upper and lower quartiles

2.6.3 Distribution Measures

Measures of distribution are the ones that describe a set of data by showing the shape that it takes including its symmetry and flatness.

2.6.3.1 Moment

In statistics, the moment measures the shape of a data set.

$$moment_k(X_i) = \frac{\sum_{i=1}^{n}(x_i - \bar{X}_i)^k}{n-1} \tag{2.3}$$

If k =1, it is the first central moment

If k=2, it is variance, the 2nd central moment

If k=3, it is skewness, the 3rd central moment

If k=4, it is kurtosis, the 4th central moment

The first two moments are location and dispersion measures, respectively.

The 3rd and 4th moments are distribution measures, showing how the values are distributed.

2.6.3.2 Skewness

Skewness measures the symmetry of the distribution around the mean. It is divided by the standard deviation to make it independent from the scale.

$$skewness(X_i) = \frac{moment_3(X_i)}{s^3} = \frac{\sum_{i=1}^{n}(x_i - \bar{X}_i)^3}{(n-1)s^3} \tag{2.4}$$

If skewness = 0 -> normal distribution

If skewness > 0 -> distribution concentrated on the left side

If skewness < 0 -> distribution concentrated on the right side

2.6.3.3 Kurtosis

Kurtosis measures the flatness of the distribution function.

$$kurtosis(X_i) = \frac{moment_4(X_i)}{s^4} = \frac{\sum_{i=1}^{n}(x_i - \bar{X}_i)^4}{(n-1)s^4} \tag{2.5}$$

2.7 Regression

Regression is a statistical technique used in DM to find the relationships between variables, searching for a model that best characterizes a given data sample. It estimates the quantitative effect of the causal (dependent) variables upon the variables that they influence (independent variables).

Mathematically, the regression predicts a value for the dependent variable Y through the formula:

$$Y = F(x, \theta) + e \tag{2.6}$$

Being Y a set of dependent variables to predict $(y_1, y_2, \ldots, y_n)$, $F(x, \theta)$ the function that defines the sample depending on the set of independent variables $X(x_1, x_2, \ldots, x_n)$ and the set of parameters $\theta(\theta_1, \theta_2, \ldots, \theta_n)$ and being the error (e) associated to the prediction process. A model of a

regression problem with a single predictor is represented on fig. 2.5, where the points represent the predictor's values and the line represents the achieved prediction.

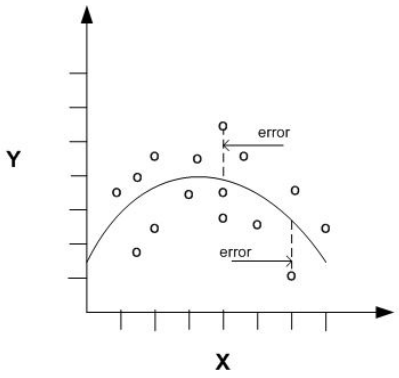


Figure 2.5: Non-linear regression with a single predictor [4]

2.7.1 Evaluate the Regression Quality

Two of the most used methods to obtain the resulting error when a regression is calculated are the *Root Mean Square Value* (RMSE) and the *Mean Absolute Error* (MAE).

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{2.7}$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n} | y_j - \hat{y}_j | \tag{2.8}$$

Being n the number of samples, y_j the predicted value for the index j and $\hat{y}_j$ the predicted value for the index j [30] [31].

2.8 Algorithms

2.8.1 Decision Trees

Decision tree is a method used both for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

As established in fig. 2.6 there is a **root node**, **branches** and **leaf nodes**. In the branches, the conditions on the predictor variables are settled and a path is chosen between the root node and the leaf node through a *if-then* rule. Each node leads to other nodes through the accomplishment of more conditions until it reaches the leaf node.

On decision tree learning the most basic algorithm (ID3) [32][33] consists on a top-down approach that starts by finding the best attribute to be used as the test at the root node of the tree.

A descendant of the root attribute is then created for each possible value of this attribute and the training examples are sorted to the appropriate descendent node [21]. The whole process is then repeated at each descendent node. To improve the predictive accuracy of the constructed tree, *pruning* is used to reduce its overfitting to the data. It consists on going up to the leaf of the tree and make an error estimation. If the error is small, the tree is kept and if not, the tree is discarded. In fig. 2.6 a scheme of a decision tree may be observed, where the input leads to the division into three different trees to reach the final output. [28]

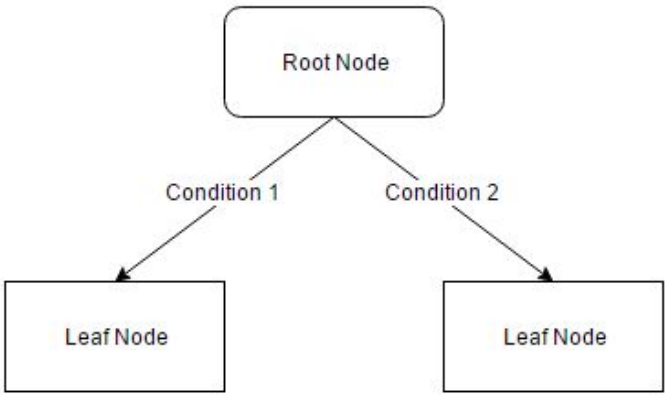


Figure 2.6: Decision tree model

2.8.2 Random Forest (RF)

Random Forest is an ensemble of B trees. To produce an output for an object, B outputs are created (one for each tree). If it is a classification algorithm, each tree votes and the most voted classifier is chosen. If instead the algorithm is a regression one, an average of the set of the output values is calculated to get the final output. A scheme of a random forest may be observed on fig. 2.7. [34]

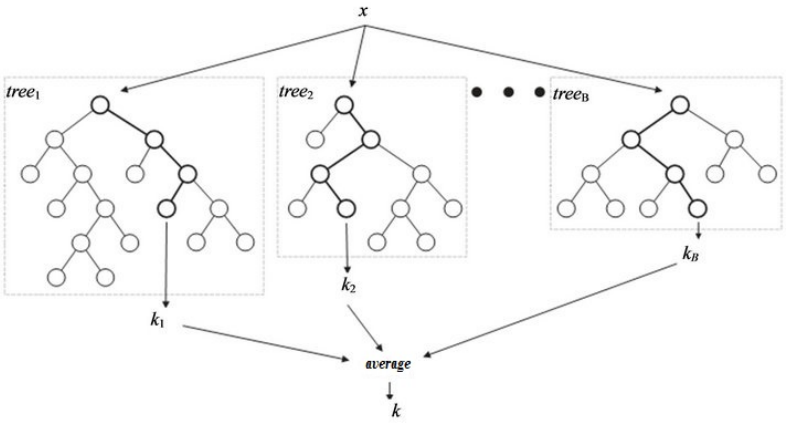


Figure 2.7: Random Forest model [5]

2.8.3 M5

M5 is a modified regression tree algorithm used for regression. It passes through three main steps to make a prediction [35] [32]:

2.8.3.1 Building

The tree is built using a splitting criterion. It calculates the reduction of error that comes from each possible outcome of a following leaf. The error formula is stated in the following equation:

$$\Delta e = \sigma(T) - \sum_{n=1} \frac{|T_i|}{|T|} \times \sigma(T_i) \tag{2.9}$$

Where Δe is the error difference, σ is the standard deviation, T is the set of example that reaches the node and T_i is the resulted set from splitting the node according to the selected attribute.

The splitting process ceases when all the instances that reach a node vary less than 5% of the standard deviation of the original instance or when only a few instances remain.

2.8.3.2 Pruning

Pruning is the second step of the M5 model. The process is already described in section 2.8.1. M5 calculates the error by multiplying the residual of the error by $(n+\nu)/(n-\nu)$, where n is the number of training cases and ν is the number of parameters in the model. It is on to do not underestimate the error. If the estimated error is lower at the parent, the leaf node may be dropped.

2.8.3.3 Smoothing

The tree is smoothed to avoid sharp discontinuities at the leaves of the tree after the pruning process. It consists on adjusting the values along the root to the leaf of a given predicted value.

2.8.4 Partial Least Squares (PLS)

Partial least squares is a linear regression technique. It decomposes the predictors' matrix and the target matrix into principal components that maximize the variance of the data using *Principal Components Analysis* (PCA). The variables that were correlated before are replaced by a usually smaller set of uncorrelated variables carrying almost the same information. That way, the data is projected into a smaller dimensional subspace, retaining most of the information. The regression is then made between the principal components of each of the variables [36] [37].

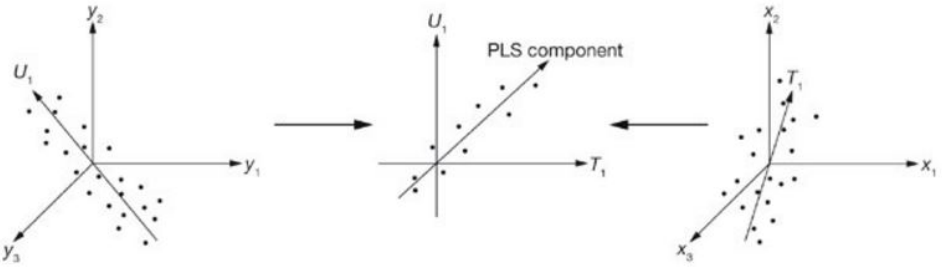


Figure 2.8: Partial Least Squares model: x_1, x_2 and x_3 are predictors, y_1, y_2 and y_3 are target variables and T_1 and U_1 are the respective principal components [6]

2.8.5 K-Nearest Neighbors (k-NN)

K-nearest neighbors is an algorithm that may be used for classification and regression. It starts by looking for the closest trained values relative to the test point. The number of trained points considered is the value of k (a parameter of the algorithm).

If it is a classification problem, each one of the considered trained points votes with its own classification. The most voted one will classify the test point.

In a regression, the process depends on the cost function. If the objective is to minimize the quadratic error, the mean value of the values of the trained objects is calculated and equals the value of the prediction for the test point. If the objective is to minimize the absolute deviation, the median is used instead of the mean. [28]

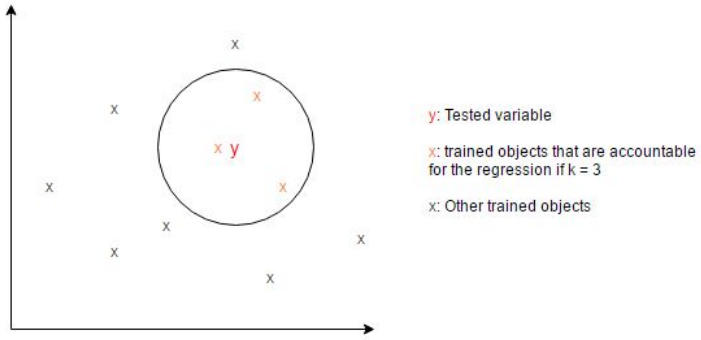
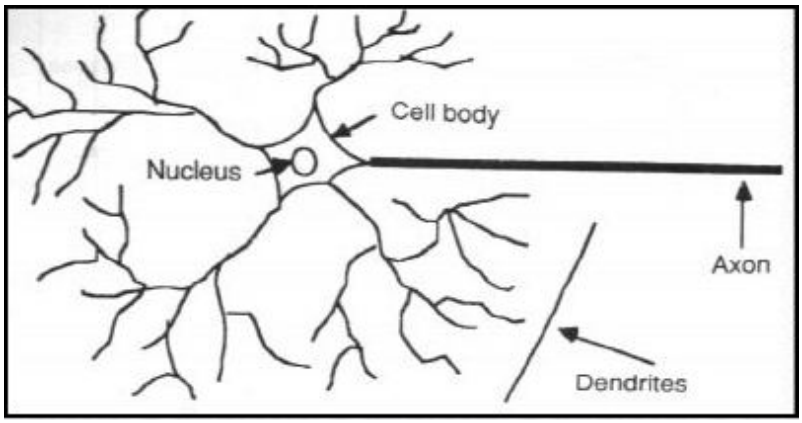


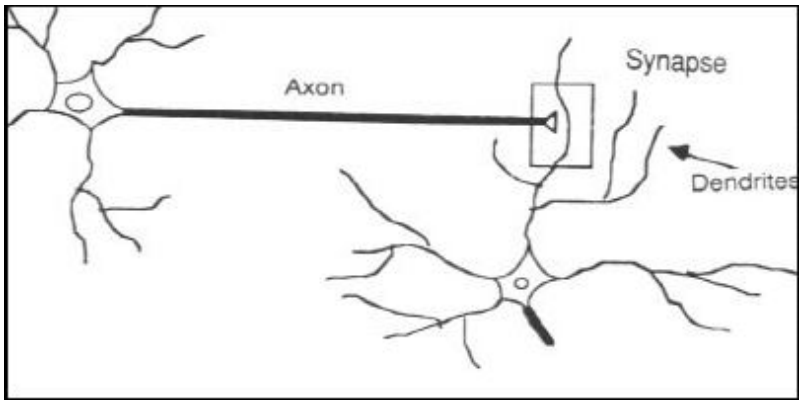
Figure 2.9: k-NN model [6]

2.8.6 Artificial Neural Networks (ANN)

ANNs functioning is based on the way the human brain works. On a human neuron, the dendrite receives the signal and sends electrical signals through axons to other neurons. This process is called synapse.



(a) Components of a neuron [7]



(b) Functioning of a synapse

Figure 2.10: Models of the components (2.10a) and the functioning of a synapse (2.10b) [7]

ANNs try to simulate the neural activity deducing the essential functions of neurons and their interconnections.

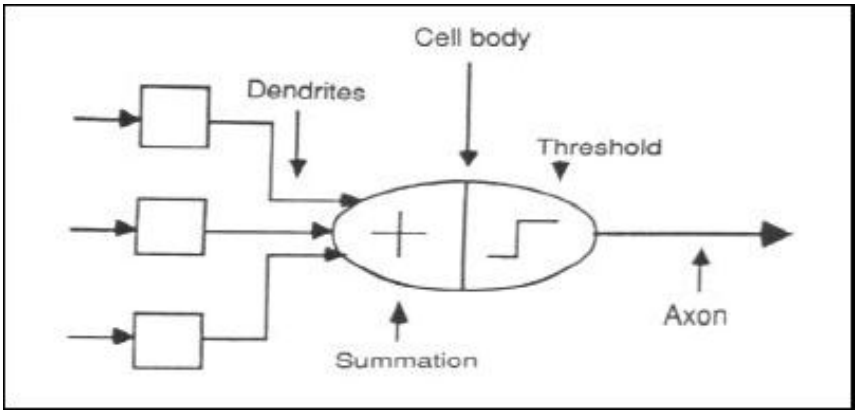


Figure 2.11: A machine neuron model [7]

Most of the functions of ANNs are on the field of pattern recognition. They do not follow a linear path unlike most computational systems. Instead, information is processed collectively and

in parallel through a network of nodes.

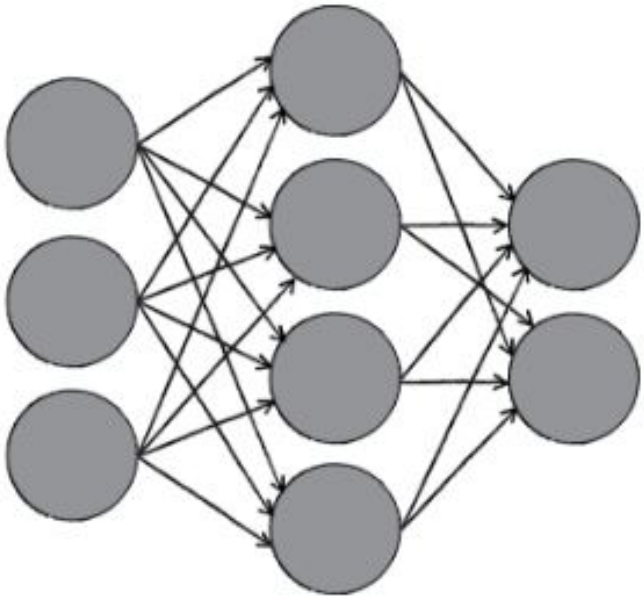


Figure 2.12: Model of neural connections [8]

The nodes change its internal information structure based on the information flowing through them. Learning is done by adjusting weights (a number associated to each connection that controls the signal between two neurons). If the ANN generates the desired output, there is no need to adjust the weights. On the other hand, if the error is significant, the weights are adjusted. [7][8]

2.8.7 Support Vector Machines (SVM)

This section is divided between rigid margin SVMs (2.8.7.1) and SVM applied to non-linear functions (2.8.7.2).

2.8.7.1 Rigid Margin SVMs

Consider the data representation in fig. 2.13.

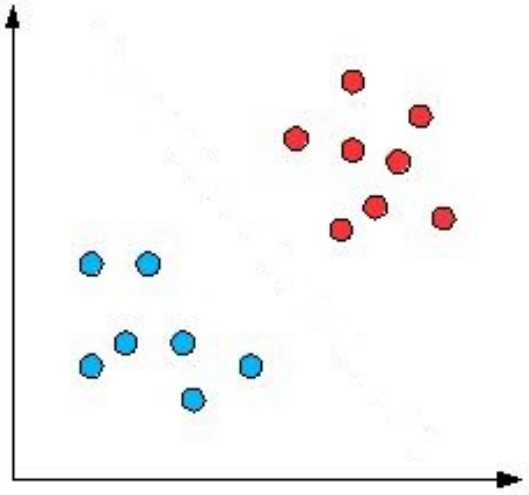


Figure 2.13: Data set representation. The red and blue dots belong to different classes.

If it is intended a division with a straight line between the two classes that allows a classification of a new object, it would be intuitively drawn in the middle of the two classes. That happens because it would create a safety space between the two classes if the line is as far as possible from both classes, minimizing the risk of failure as shown on fig. 2.14.

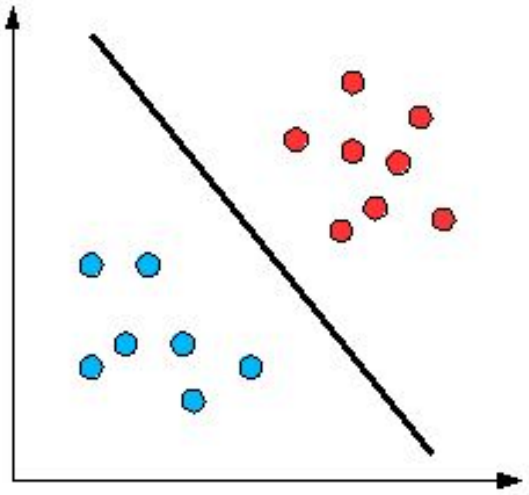


Figure 2.14: Data set representation with a division straight line.

2.8.7.2 SVMs applied to non-linear functions

Non-linear functions may not be separated by straight lines. A technique called mapping is used. The data is mapped into higher dimensions where it exhibits linear patterns and a linear model may be applies in the new input space. In other words, the feature representation is changed.

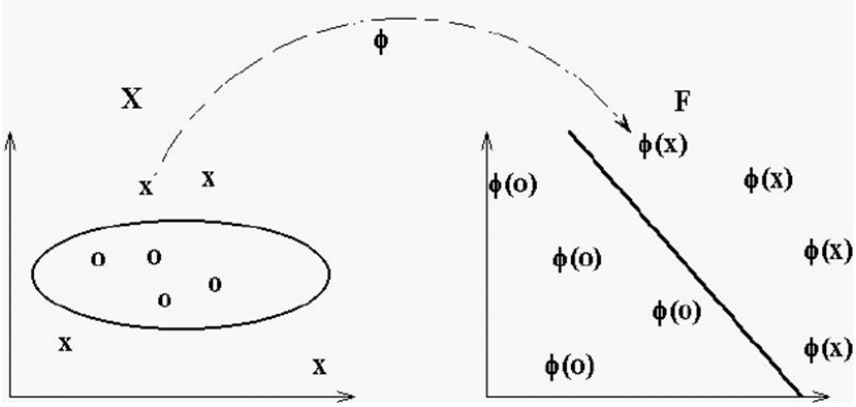


Figure 2.15: Mapping technique. [9]

2.9 Related Projects

The main basis for this project is *Performance Management Analytics for the Automotive Industry* by Oliveira [2]. It consisted on the creation of a predictive model for the indicator *units per hour* based on the data collected on the company *Autoeuropa – Automóveis, Lda.* and used DM and ML techniques. After understanding the business and data and preparing the data, 5 different algorithms were used to train the data set. The problem was elaborated considering the time-series as a regression. All of them were tested and the error was tested using RMSE and MAE to select the best model.

Proactive Supply Chain Performance Management with Predictive Analytics by Nenad Stefanovic [13] describes how to create a predictive model for a Supply Chain using predictive KPIs to measure its performance. It started by explaining the importance of a proactive thinking correlating Business Intelligence and Key Performance Indicators with the needs of the supply chain. Later on, two approaches for building KPI prediction models were used: data mining dimensions and prediction tables. Data mining dimensions is a dimension with a special father-son relationship on the data defined through DM application instead of user defined. Prediction Tables are usually used to measure a group. Data mining predictions are performed within ETL (extract, transform and load) process. ETL package pulls data from the data source, performs a prediction task and loads the results into a prediction table.

Chapter 3

Case Study

In this chapter the case study is explained following the CRISP-DM methodology phases (section 2.5.4). In section 3.1 the main aspects of the business are explained. In section 3.2, the problem is formulated mathematically. The process to collect the data that is used to predict is explained in section 3.3. The main topics of data analysis are stated in section 3.4.

3.1 Business Understanding: Volkswagen Autoeuropa

The unprocessed data was provided by *Volkswagen Autoeuropa, Lda.*. It is an automotive production company located in Palmela, Portugal. It started activity in 1995, representing an investment of 1970 million euros with a production area of 1 100 000 square meters. In 2015 it produced three *Volkswagen* models: *Scirocco*, *EOS* and *MPV*.

On such a complex industry, information as a high impact on performance. For that reason *Volkswagen Autoeuropa* is permanently acquiring KPIs, including *Hours per Unit* (HPU).

However, while those KPIs are past related, a proactive strategic thinking is not happening. The change of paradigm is of crucial importance for the company.

Besides all of that, the business strategy of *Volkswagen Autoeuropa* is *Make to Order* (MTO), which means that the factory produces after receiving the orders [38].

3.2 Problem Formulation

Let X be a matrix of n predictors and $X_1, X_2, ..., X_n$ be each one of the predictors. Let X_{it} be one observation of each predictor X_i and X_{iT} its latest. Let Y be the target variable, Y_t an observation and Y_T its last observation. The objective is to predict Y_{T+1} based on X_{ij} and Y_j ($j \leq T$).

3.3 Data Collection

The data is being permanently collected at *Volkswagen Autoeuropa* plant. It is provided on *Excel* data files: *raw data*. There are three types of raw data files used to calculate the target indicator:

- **Structure:** classifies the cost centers. It indicates which cost centers and employees are and are not accountable for the Harbour calculation.

- **Status:** classifies each employee indicating his cost center, position, function and working schedule.

- **Hours:** indicates the number of hours that each employee worked and on which task.

- **Train:** states the number of hours spent on training by each employee and the type of training they received.

Administratorname	Costc	Orgun	Objectname	NameofE	NameofE	HA(	Excpe	Excreas	HStru	RCla	CSSDetail	RClassAre
BA Executive Direct.	4310000	50000078	Executive Director Business Area	Indirect	Manager	n	1	eds	not	not	-	not
BA Executive Direct.	4310000	50000080	Executive Directors Support Team	Indirect	Manager	n	1	eds	not	not	-	not
BA Executive Direct.	4310000	50000080	Executive Directors Support Team	Indirect	Technician	n	1	eds	not	not	-	not
BA Executive Direct.	4310000	50108352	Communication, Sustainability & CO Image	Indirect	Technician	n	1	eds	not	not	-	not
BA Prod Mgt&Planning	4310030	50002126	Evaluation Analysis Coordination	Indirect	Technician	y	0	not	ss	acss	Engineering current Product	bpt
BA Prod Mgt&Planning	4310030	50002226	Technical Changes Coordination	Indirect	Technician	y	0	not	ss	acss	Engineering current Product	bpt

(a) Example of the first six lines of a *Structure* data file

Administratornam	Costct	Acronym	Orgur	OrganizationalUnit	Persn	Complete name	NameofEG	NameofES	Hiring Dat
BA Plant Manager	4310000	AG	50000078	Plant Manager	30088	ANTONIO AFONSO MELO PIRES	Manager	Indirect	01/10/1992
BA Plant Manager	4310000	AGU	50000080	Plant Manager Team	30172	MARIA CARLA BARBOSA SOARES	Technician	Indirect	02/12/1992
BA Plant Manager	4310000	AGU	50000080	Plant Manager Team	31504	RUTE ISABEL FERREIRA COELHO	Technician	Indirect	20/06/1994
BA Plant Manager	4310000	AGU	50000080	Plant Manager Team	41209	KATHARINA EGGER	Technician	Indirect	20/10/2014
BA Assembly	4310020	AGFMP1C1	50066206	Zone B - URQ 1A	33783	PAULO JORGE S RODRIGUES GONCALVES	Operator	Direct	18/09/1995
BA Finance	4310020	AFR	50002425	Accounting	30303	ELSA MARIA OLIVEIRA CARETO	Manager	Indirect	15/03/1993

(b) Example of the first six lines of a *Status* data file

EEsubgrp	EEgroup	Period	TmType	Time type descript.	nHours
Direct	Operator	201504	10	Attendance	163.00
Direct	Operator	201504	7512	Training Onsite	11.00
Direct	Operator	201504	10	Attendance	35.00
Direct	Operator	201504	10	Attendance	75.00
Direct	Operator	201504	10	Attendance	131.00
Direct	Operator	201504	7512	Training Onsite	43.00

(c) Example of the first six lines of a *Hours* data file

Persl	FullName	CostCtr	B.EventText	EmployeeGroup	Employee Sub-Group	RealHours
31268	JOAO LUIS ALVES ABRANTES	4316100	ALEMÃO - A1	Technician	Indirect	0.50
34838	MARGARIDA PAULA SANTOS CALEIRA	4310190	ALEMÃO - A1	Technician	Indirect	3.00
41049	JOAO PEDRO DELGADO SOARES	4310090	ALEMÃO - A1	Manager	Indirect	3.00
31668	CARMEN SOFIA GOMES EMILIANO	4310160	ALEMÃO - A1	Technician	Indirect	4.50
30339	JOSE ANTONIO COSTA CORREIA	4316100	ALEMÃO - A1	Technician	Indirect	6.50
30721	MANUEL SOUSA GUERRA	4316100	ALEMÃO - A1	Technician	Indirect	6.50

(d) Example of the first six lines of a *Train* data file

Figure 3.1: Examples of the first six lines of each of the raw data files types. From the top to the bottom, the first one is an example of a *Structure* data file, following *Status*, *Hours* and *Train* data files examples

However, raw data can not be analysed in this form. It enters on an ***Extract, Transform and Load* (ETL)** software. It groups and transforms all the data into KPI_0 (KPIs that are not calculated using other indicators), storing it on a Mongo database (DB). The data stored on MongoDB may be accessed and explored.

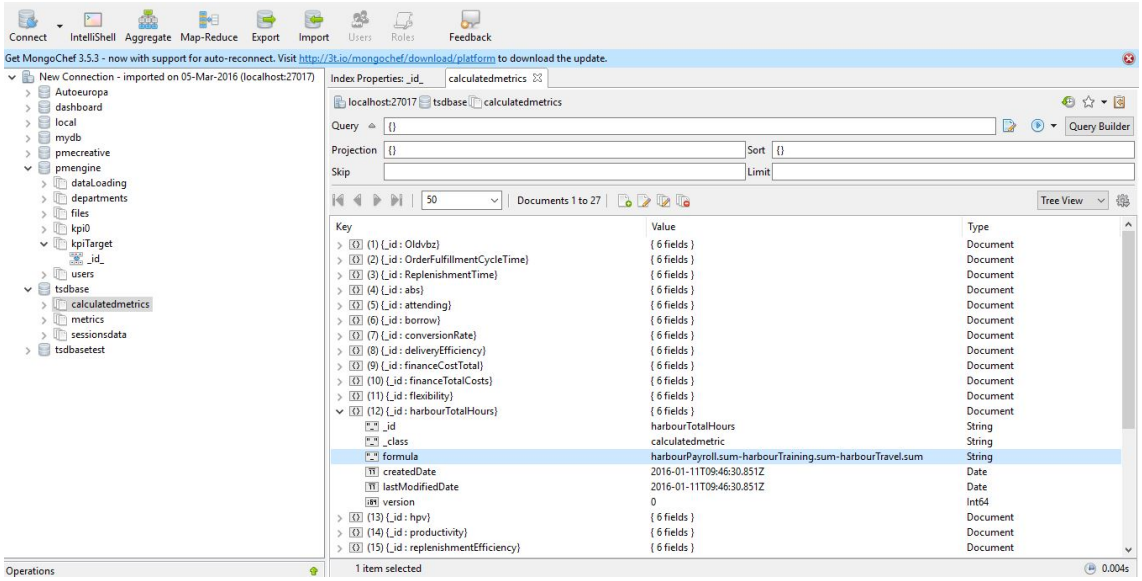
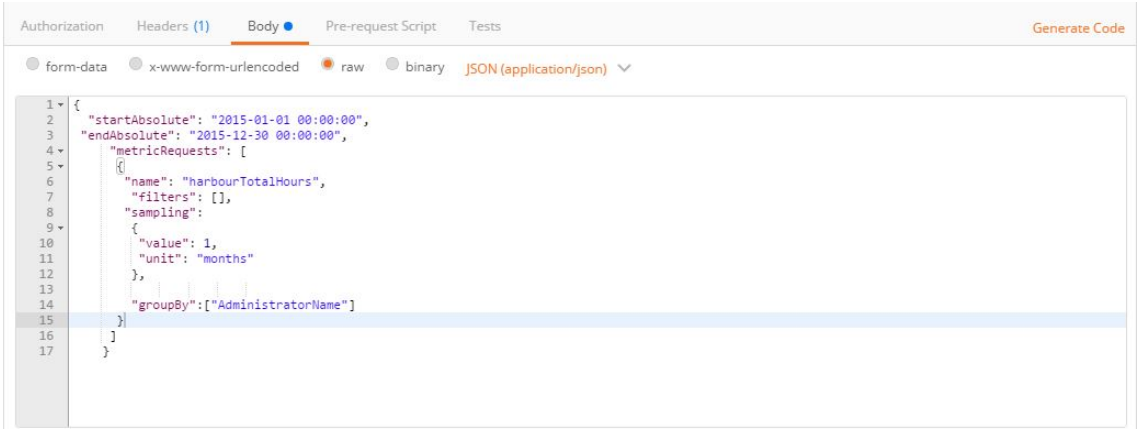
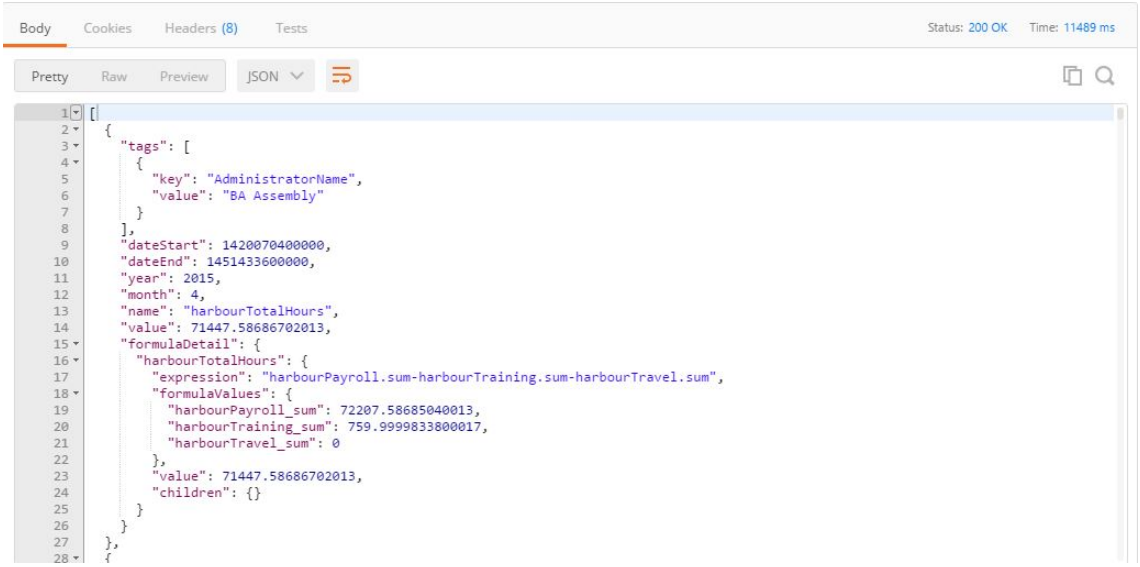


Figure 3.2: Inside *MongoDB* using the software *MongoChef*

The data stored in MongoDB is handled by **TSDBase**. It establishes a client-server connection with MongoDB. TSDBase performs the necessary mathematical operations to acquire KPI_{1+} (KPIs calculated based on other KPI values) for a given period of time. To send requests, the user has to use *JQuery* requests. The interface software used to send requests was *Postman*.



(a) Example of *JQuery* request sent via *Postman*



(b) Example of *JQuery* response received via *Postman*

Figure 3.3: Example of *JQuery* request and response using the software *Postman*

Statistics are then compiled on the performance measurement engine (PME) used by *Volkswagen Autoeuropa*. The user may access it to see statistics, analysis and reports.

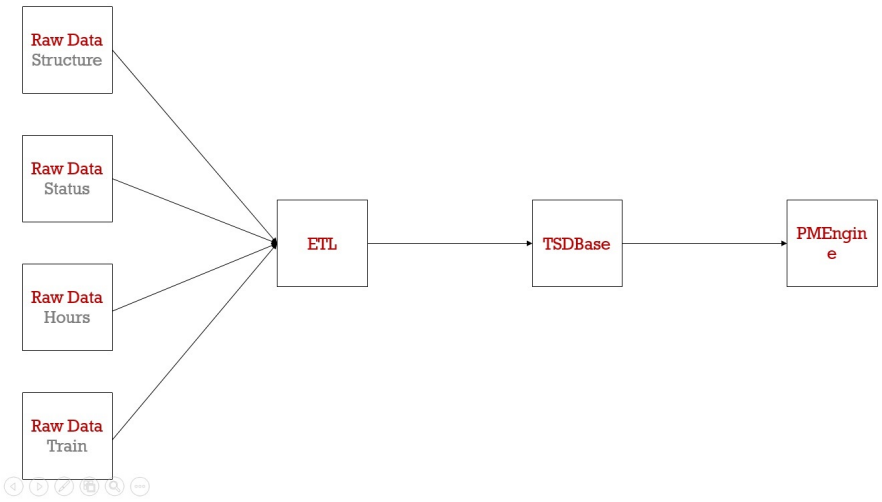


Figure 3.4: Data extraction model

3.4 Data Understanding

3.4.1 Harbour Variable

The main objective is to calculate the KPI *Hours per Unit*. It states how many hours are necessary to produce one unit:

$$HPU = \frac{Number\,of\,working\,hours}{Volumes} \tag{3.1}$$

However, there are several ways of calculating the number of working hours, which creates different types of HPU indicators. One of those is *Harbour*. In *Harbour*, the number of working hours equals the number of hours that are paid in total (*Payroll*), excluding the hours spent by employees on business travels (*Travel*) and training (*Training*). The hierarchical tree that describes the dependences of the *Harbour* indicator is shown in fig. 3.5.

$$Harbour = \frac{Payroll - Travel - Training}{Volumes} \tag{3.2}$$

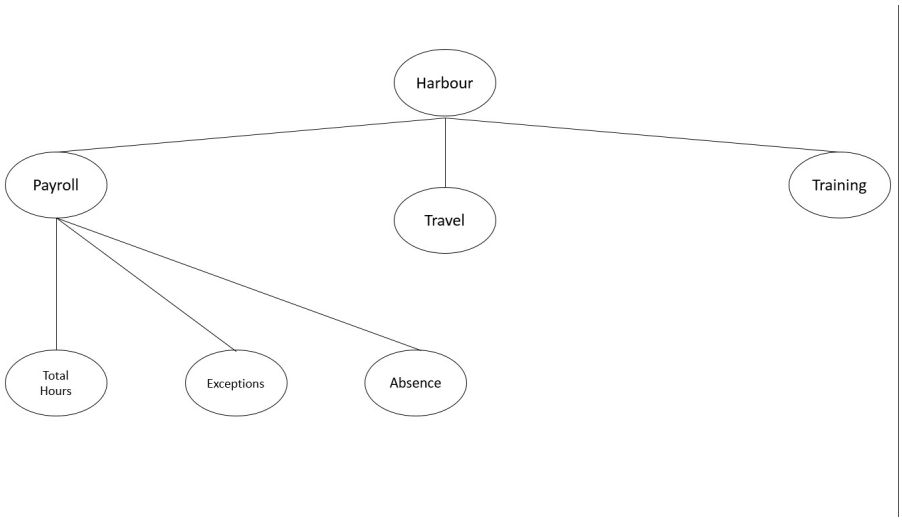


Figure 3.5: Hierarchical tree of the *Harbour* KPI

3.4.2 Plant Data Set

Given that the data set was referred to the total production on the plant of the industry on an entire year, there were three hundred and sixty-five observations. The days that production did not occur (weekends, holidays, etc.) were excluded, with two hundred and fifty-four observations remaining. The original files contained six variables, as shown on fig. 3.6:

- Payroll
- Travel
- Training
- Hours (= Payroll - Travel - Training)
- Volumes
- Harbour

Another variable was created. In each observation it assumes the value of Harbour in the next one. That variable is called *HarbourNextDay* and it is the target one.

	Payroll	Travel	Training	Hours	volumes	Harbour	HarbourNextDay
1	18420.03	36	80	18304	448	40.85764	41.31414
2	18423.26	13	108	18301	443	41.31414	41.94159
3	18451.53	43	91	18317	437	41.94159	40.62398
4	18471.63	56	93	18322	451	40.62398	39.83277
5	18425.59	86	98	18241	458	39.83277	39.90662
6	18378.11	65	92	18221	457	39.90662	41.05094

Figure 3.6: First six lines of the plant data frame

3.4.3 Plant Data Set with individual model features

Another provided data file contained the plant data set taking into account the domain separation by car model produced. There are three different models in the data set (EOS, Scirocco and MPV) and the sum of the hours and volumes of three of them is equal to the value on the previous data set. The new data set maintained the first one plus fifteen more variables (*Payroll, Travel, Training, Hours* and *Volumes* for each of the models).

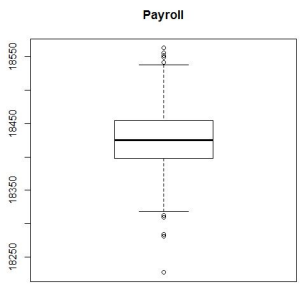
eosPayroll	eosTravel	eosTraining	eosHours	eosVolumes	mpvPayroll	mpvTravel	mpvTraining	mpvHours	mpvVolumes	sciPayroll	sciTravel	sciTraining	sciHours	sciVolumes
1842.003	4	8	1830	45	11052.02	22	48	10983	269	5526.009	11	24	5491	134
1842.326	1	11	1830	44	11053.96	8	65	10981	266	5526.978	4	33	5490	133
1845.153	4	9	1832	44	11070.92	26	55	10990	262	5535.458	13	27	5495	131
1847.163	6	9	1832	45	11082.98	34	56	10993	271	5541.488	17	28	5497	135
1842.559	9	10	1824	46	11055.35	51	59	10945	275	5527.676	26	30	5472	137
1837.811	7	9	1822	46	11026.87	39	55	10932	274	5513.433	20	28	5466	137

Figure 3.7: First six lines of the added variables to the plant data set

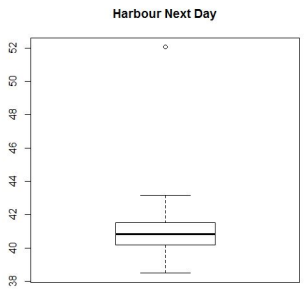
3.4.4 Exploratory Data Analysis

Many statistical analyses were conducted. In this section the most relevant ones are described.

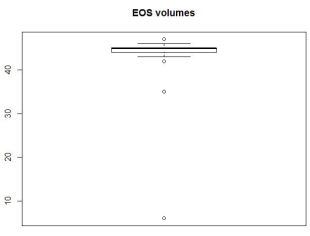
In figure 3.8 the box plots that expose the most relevant information are shown. The points outside of the boxes are possible outliers. It is important to identify and analyse them so that they can be identified as outliers or not. Only the most relevant ones are presented here.



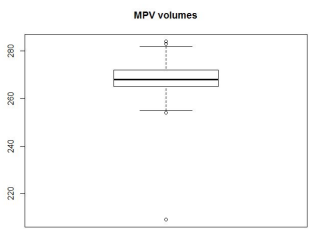
(a) Box plot of *Payroll* of the plant



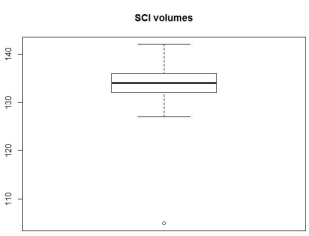
(b) Box plot of *HarbourNext-Day* of plant



(c) Box plot of *volumes* of EOS model



(d) Box plot of *volumes* of MPV model



(e) Box plot of *volumes* of SCI model

Figure 3.8: Most relevant box plots from the data set

The first box plot has many possible outliers. Looking at the dependence KPI tree (fig. 3.5), it becomes clear that *Payroll* is dependent on three other variables. It is enough to vary one of those to vary *Payroll* which justifies the variation of this variable.

Looking at the second box plot, a value is clearly out of the other values range. However, it was not far enough to be treated as an outlier and it was only considered an atypical day of production. It is justifiable by the date when it happened (last day of production of the year, right before holidays and Christmas). This value does not appear in the *Harbour* column because it was the last day of production and the last row of the *Harbour* variable was deleted so that missing values would not appear. It is used for all the models and predictions.

The third, fourth and fifth box plots are unusually strange because of the number of possible outliers that are presented. However, this is normal due to the business strategy (MTO) (section 3.1) applied at *Volkswagen Autoeuropa*. If the number of cars produced are dependent of the number of orders, then it is normal that in some days the production of a specific model is very high or low. In spite of this, there is a value that stills curious: only six units of the EOS model produced on the eleventh of August. This one, after analysis, was considered an outlier. The value of the sum of the production of the three models is different from the value of the whole plant and therefore the observation was deleted from the data set.

		Plant	EOS	MPV	SCI
MEAN	**Payroll**	18426	1842.56	11055.35	5527.68
	Travel	51.33	5.15	30.77	15.42
	Training	100.62	10.04	60.41	30.17
	Hours	18274.07	1827.33	10964.19	5482.05
	Volumes	447.69	44.56	268.35	134.19
	Harbour	40.84	-	-	-
	HarbourNextDay	40.88	-	-	-
STANDARD DEVIATION	**Payroll**	48.72	4.90	29.44	14.72
	Travel	18.85	1.93	11.25	5.66
	Training	1.10	1.21	6.32	3.18
	Hours	54.21	5.44	32.73	16.37
	Volumes	10.14	2.78	7.11	3.57
	Harbour	0.94	-	-	-
	HarbourNextDay	1.17	-	-	-
VARIANCE	**Payroll**		23.73	854.34	213.59
	Travel	355.45	3.71	126.94	32.11
	Training	110.77	1.21	39.89	10.10
	Hours	2938.47	29.31	1058.32	265.02
	Volumes	102.77	7.41	36.65	9.39
	Harbour	0.88	-	-	-
	HarbourNextDay	1.38	-	-	-
RANGE	**Payroll**	336.61	33.66	201.97	100.98
	Travel	109	11	65	33
	Training	69	7	42	21
	Hours	326	32	196	98
	Volumes	51	41	30	15
	Harbour	4.67	-	-	-
	HarbourNextDay	13.57	-	-	-

Table 3.1: Mean and standard deviation value for each of the variables

In table 3.1 it may be observed that both the standard deviation and the mean of the target variable present low value. About the same variable, the range is not extremely low but the one for the *Harbour* variable is, which demonstrates that the only observation that highers the range of

HarbourNextDay is the last day of production (The only different value between the two variables). That means that the value of the target variable is around the value of the mean almost always.

3.5 Data Preparation

Before the predictions are made the data must be pre-processed in order to avoid misleading results. Action need to be taken according to the format of data. Starting with the original data set, several changes are applied to the format of data. Feature engineering techniques were applied with the goal of creating a data set that is better suited to the algorithms.

Data set 1 is described in section 3.4.2. It is the original data set provided by *Volkswagen Autoeuropa* (plant data set). It serves as the main basis to build other data sets with different features. **Data set 2** was provided by *Volkswagen Autoeuropa* that described the domain separation by car model produced. The data set building process is described in section 3.4.3. Adding information to the data set could improve the results. Due to the lack of improvement in the results, the domain separation was not taken into account in the following experiments. Another data set was built trying to expose more information that could positively influence the results. It characterized the day of production, reporting the day of the week, the week and the month of that day (**Data Set 3**). To put more emphasis in the past information, a data set was created joining Data Set 3 and the information about the day before to each observation (**Data Set 4**). **Data sets 5, 6, 7 and 8** are equal to Data Set 4 except that each observation also includes the information about the past two (**Data Set 5**), three (**Data Set 6**), four (**Data Set 7**) and five previous days (**Data Set 8**).

Given that all these data sets resulted in experiments that have a very similar predicted value for the target variable, data sets were created that put emphasis on the differences between consecutive days. The target variable is the difference between the value of an observation in the *Harbour* column and the day ahead observation in the same column. **Data Set 9** is equal to data set 3 plus the variables that represent the difference of values between the previous observations and the values on the day before. **Data Sets 10, 11, 12 and 13** are equal to data set 9 except that each observation also includes the variable differences between the current day and two, three, four and five days respectively. **Data set 14** is based on the data set 3 except that it adds variables representing the differences between each the values of each observation to a given variable and the mean value of the same variable.

Data Set	Information
Data Set 1	Plant Data Set
Data Set 2	Data Set 1 and domain separation by car model produced
Data Set 3	Data set 1 plus days description (month, week, day of the week)
Data Set 4	Data set 3 plus information about 1 day before each observation
Data Set 5	Data set 3 plus information about 2 days before each observation
Data Set 6	Data set 3 plus information about 3 days before each observation
Data Set 7	Data set 3 plus information about 4 days before each observation
Data Set 8	Data set 3 plus information about 5 days before each observation
Data Set 9	Data set 3 plus differences between the value of each observation and the previous one
Data Set 10	Data set 9 plus differences between the value of each observation and 2 days before
Data Set 11	Data set 10 plus differences between the value of each observation and 3 days before
Data Set 12	Data set 11 plus differences between the value of each observation and 4 days before
Data Set 13	Data set 12 plus differences between the value of each observation and 5 days before
Data Set 14	Data set 3 plus differences between the value of each observation variables' mean value

Table 3.2: Compilation of the information per observation in the built data sets

Chapter 4

Results

In this chapter, the experimental setup is clarified in section 4.1, exposing the way data is split (4.1.1) and the way parameters are tuned (4.1.2). The obtained results for each of the tested data sets and algorithms is presented in section 4.2.

4.1 Experimental Setup

4.1.1 Data Splitting

The data set is a time series and that time dependence must be taken into account when the splitting is made. The technique that was used is *sliding window*. The process consists on orderly slide the train set and the test set at each prediction made, has represented on fig. 4.1. The initial window has fifty observations and the horizon has one observation.

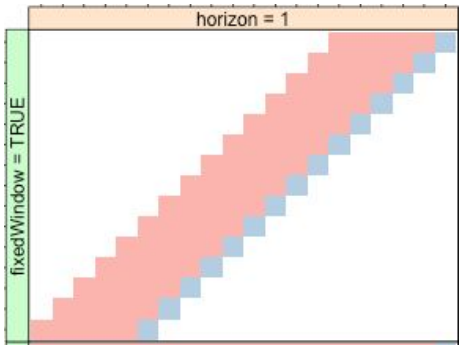


Figure 4.1: Sliding window process [10]

The first trained set has fifty observations and another one is predicted at each iteration

4.1.2 Parameters

The parameters are tuned using the R language. The main package used to train the algorithms is the *Caret* package. [10]

Random Forest needs to receive as input the number of variables sampled as candidates at each split (section 2.8.2).

The PLS algorithm receives as an input the number of components to be used in the model (section 2.8.4).

To build the model of M5, three steps may be implemented. However, they are not necessary for the algorithm to work. For that reason, the algorithm is tuned choosing between implementing or not each one of those steps (section 2.8.3).

The kNN algorithm receives the number of neighbors to be accountable for the experiment (section 2.8.5).

4.2 Results

A baseline had to be created to validate or not the obtained results. To do that, it was assumed that the efficiency tomorrow would be the same as it is today. It was called *Naive Prediction*.

Data set 1

Parameters: default

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors	**Naive Prediction**
RMSE	1.28	1.24	1.23	1.27	1.52
MAE	0.85	0.80	0.79	0.85	1.10

Table 4.1: Error results from **data set 1** with the default tuned parameters

Parameters:

- Random Forest (RF) [39]

  mtry: varies between 1 and the number of predictors of the data set, 1 by 1

- Partial Least Squares (PLS) [40]

  ncomp: varies between 1 and 6, 1 by 1.

- M5 [32]

  pruned: varies between "Yes" and "No"
  smoothed: varies between "Yes" and "No"
  rules: varies between "Yes" and "No"

- k-Nearest Neighbors (k-NN) [10]

  k: varies between 1 and 15, 1 by 1.

The parameters are always tuned this way along the project.

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.27	1.24	1.23	1.25
MAE	0.84	0.801	0.80	0.84

Table 4.2: Error results from **data set 1** with tuned parameters

It may be observed that the tuning of parameters does not reflect a significant improvement on any of the algorithms. However, the parameters kept tuned the same way on the following experiments.

Data set 2

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.27	1.25	1.24	1.25
MAE	0.85	0.81	0.81	0.85

Table 4.3: Error results from **data set 2** with tuned parameters

The addition of the domain separation by car models, as observed on table 4.3, does not improve the results. For that reason, the additional variables of this data set relative to **data set 1** are not kept into account on the following experiments.

Data set 3

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.24	1.25	1.26	1.25
MAE	0.80	0.82	0.83	0.84

Table 4.4: Error results from **data set 3** with tuned parameters

The characterization of production days (data set 3) does not increase or decrease the quality of results. In spite of it, the number of variables in this case is not as large as it is in data set 2 and therefore the risk of curse of dimensionality is smaller. For that reason the variables in data set 3 are kept along the next experiments.

Data set 4

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.24	1.24	1.25	1.26
MAE	0.80	0.81	0.82	0.84

Table 4.5: Error results from **data set 4** with tuned parameters

The information about the past on each observation represents a small decrease of the error using the *Random Forest* algorithm. The other algorithms kept the same results. More experiments were practised, increasing the information about the past by one day at each data set in **data sets 5, 6, 7** and **8**.

Data set 5

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.24	1.24	1.26	1.26
MAE	0.81	0.82	0.83	0.84

Table 4.6: Error results from **data set 5** with tuned parameters

Data set 6

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.24	1.25	1.26	1.27
MAE	0.81	0.82	0.86	0.86

Table 4.7: Error results from **data set 6** with tuned parameters

Data set 7

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.24	1.26	1.29	1.27
MAE	0.81	0.83	0.88	0.85

Table 4.8: Error results from **data set 7** with tuned parameters

Data set 8

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.24	1.26	1.33	1.26
MAE	0.81	0.83	0.90	0.84

Table 4.9: Error results from **data set 8** with tuned parameters

As observed on **data sets 4, 5, 6, 7 and 8**, the data set with information at each observation of the past does not reflect a significant improvement, oscillating between increases and decreases of the error results.

Data set 9

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.30	1.27	1.28	1.53
MAE	0.88	0.84	0.83	1.11

Table 4.10: Error results from **data set 9** with tuned parameters

Data set 10

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.31	1.34	1.31	1.55
MAE	0.88	0.93	0.89	1.11

Table 4.11: Error results from **data set 10** with tuned parameters

Data set 11

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.31	1.36	1.32	1.55
MAE	0.89	0.95	0.90	1.12

Table 4.12: Error results from **data set 11** with tuned parameters

Data set 12

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.29	1.37	1.33	1.56
MAE	0.87	0.97	0.89	1.12

Table 4.13: Error results from **data set 12** with tuned parameters

Data set 13

	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.28	1.31	1.31	1.55
MAE	0.86	0.91	0.87	1.13

Table 4.14: Error results from **data set 13** with tuned parameters

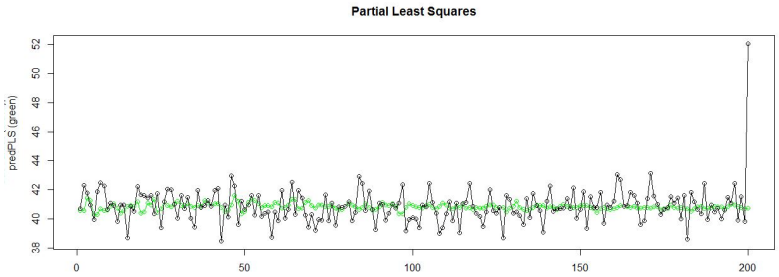
From the experiments with data sets 9 to 13, it may be concluded that the focus on the differences between observations (instead of the values of the observations itself) increases the value of the error. However, that happens due to the fact that the algorithms tested with other data sets are spotting a very similar result on the objective variable at every observation, making the predicted values constant. Due to that, and observing fig. 4.2 it becomes clear that data sets 9 to 13 are more suited than the other ones, following the variation that the true values suffer. It may also be observed that the *kNN* algorithm loses its validity when the data set is built this way (all the error values are above the baseline ones for this algorithm). Besides that, adding information considering the differences between more than one day, does not add value to the results (data sets 9,10, 11, 12 and 13 present similar results).

Data set 14

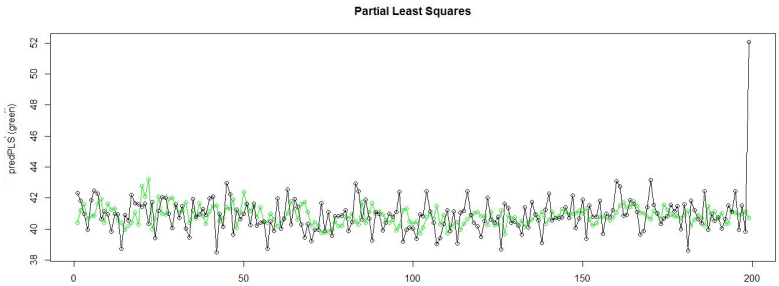
	Random Forest	Partial Least Squares	M5	K Nearest Neighbors
RMSE	1.23	1.25	1.26	1.25
MAE	0.80	0.82	0.83	0.84

Table 4.15: Error results from **data set 14** with tuned parameters

Experiments with **data set 14** led to the same conclusions taken from the experiments carried out with data sets 1 to 8. The fact that the difference between the value of each observation and the value of the mean for the target variable present a low value, the predicted values for the target variable do not follow the variations of the true values. The plots resulting from the experiments carried out with data set 14 are similar to the one presented on fig. 4.2a.



(a) Example of resulting plot from a partial least squares experiment from one of the first eight experiments



(b) Example of resulting plot from a partial least squares experiment from one of the last five experiments

Figure 4.2: Examples stating the differences between plots from one of the first eight experiments (top) and one of the last five experiments (bottom). The black line represents the real values and the green line the predicted ones

Support Vector Machines (SVM) and Artificial Neural Networks (ANN) were also tested with **data set 1**. SVM took a long time processing and that, conjugated with the fact that its results were similar to the ones obtained by the other algorithms(RMSE=1.25, MAE=0.83), led to the abandonment of the experiments with this algorithm. Artificial Neural Networks tested with data set 1, besides the long processing time, had results much worse than the expected (RMSE=40.17, MAE=40.16), reasons enough to quit the experiments with this algorithm.

Chapter 5

Conclusion

The permanent change that occurs on complex businesses makes the reactive thinking insufficient to be sustainable. A proactive thinking is required and to achieve that, the future needs to be predicted as accurately as possible.

Considering this, an empirical study of six different algorithms to predict an efficiency indicator value in the automotive industry one day ahead was developed. The original data set was changed in order to obtain one that led to better results. This fact conducted to the creation of fourteen different sets of variables were tested to improve the obtained results.

The fact that the error values are almost always lower than the ones from the baseline, seems to give full validity to the results. However, the results were contradictory: a low error was achieved because the prediction kept its values almost constant. On the other hand, when the variations of the true values were followed by the prediction, the error increased.

There are some ways of improving this project. More algorithms may be tested and more feature engineering techniques may be applied in order to lead to more accurate results. The analysis of results may be extended. Larger data sets with different values should be tested. Besides that, the predictions that are accurate enough may be implemented in a performance management software to facilitate its access.

References

[1] Kent Bauer. Predictive analytics: The next wave in kpis. *Information Management*, 15(11):68, 2005.

[2] Soares Carlos de Oliveira, Sérgio and Filipe Ferreira. Performance management analytics for the automotive industry. 2015.

[3] Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39, 2000.

[4] Regression. *Oracle Database Online Documentation*, 11g Release(4), 2016. URL: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm.

[5] Cuong Nguyen, Yong Wang, and Ha Nam Nguyen. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*, 6(5):551, 2013.

[6] Jacob T Bjerrum, Ole H Nielsen, Yulan L Wang, and Jørgen Olsen. Technology insight: metabonomics in gastroenterology—basic principles and potential clinical applications. *Nature Clinical Practice Gastroenterology & Hepatology*, 5(6):332–343, 2008.

[7] Christos Stergiou and Dimitrios Siganos. Neural networks.

[8] Daniel Shiffman, Shannon Fry, and Zannah Marsh. *The nature of code*. D. Shiffman, 2012.

[9] K Padmavathi and K Sri Ramakrishna. Detection of atrial fibrillation using autoregressive modeling. *International Journal of Electrical and Computer Engineering (IJECE)*, 5(1):64–70, 2015.

[10] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, and Can Candan. *caret: Classification and Regression Training*, 2016. R package version 6.0-64. URL: https://CRAN.R-project.org/package=caret.

[11] Performance management – definition. *University of Berkeley, California*. URL: http://hrweb.berkeley.edu/guides/managing-hr/managing-successfully/performance-management/concepts.

[12] *Yale University Performance Management Guide*. Yale University, 2010.

[13] Nenad Stefanovic. Proactive supply chain performance management with predictive analytics. *The Scientific World Journal*, 2014.

[14] Sean McPheat. Performance management. *MTD Training & Ventus Publishing UK, pp8*, 2010.

[15] Florian Melchert, Robert Winter, and Mario Klesse. Aligning process automation and business intelligence to support corporate performance management. *AMCIS 2004 Proceedings*, page 507, 2004.

[16] US Department of Health, Human Services, et al. Health resources and services administration. *Critical Care Workforce Report. Requested by Senate Report*, 2008.

[17] Andy Neely, Mike Gregory, and Ken Platts. Performance measurement system design: a literature review and research agenda. *International journal of operations & production management*, 15(4):80–116, 1995.

[18] Mark Wilcox and Mike Bourne. Predicting performance. *Management Decision*, 41(8):806–816, 2003.

[19] Michael Notté. Defining actionable business-driven kpi's – a practical methodology. 2008. URL: :http://www.kaizen-analytics.com/2008/11/defining-actionable-business-driven.html.

[20] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[21] Thomas M Mitchell. Machine learning. *Boston et al*, 1997.

[22] Bill Palace. Data mining: What is data mining. *Retrieved fro m http://www. anderson. ucla. edu/faculty/jason. frand/teacher/technologies/palace/datamining. htm*, 1996.

[23] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Elsevier, 2011.

[24] Anna Palczewska. Association rule learning. 2012.

[25] What is the crisp-dm methodology? 2016. URL: http://www.sv-europe.com/crisp-dm-methodology/.

[26] Training and testing data sets. URL: https://msdn.microsoft.com/en-us/library/bb895173.aspx.

[27] Subject: What are the population, sample, training set, design set, validation set, and test set? URL: ftp://ftp.sas.com/pub/neural/FAQ.html#A_data.

[28] J Gama, A Carvalho, M Oliveira, K Faceli, and A Lorena. Extração de conhecimento de dados, data mining. *JC Gama, Extração de Conhecimento de Dados, Data Mining*, page 101, 2012.

[29] *e-Handbook of Statistical Methods*. NIST/SEMATECH, 2009. URL: http://www.itl.nist.gov/div898/handbook/.

[30] Susan Holmes. Rms error. *Retrieved June*, 21:2012, 2000.

[31] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79, 2005.

[32] Kurt Hornik, Christian Buchta, and Achim Zeileis. Open-source machine learning: R meets weka. *Computational Statistics*, 24(2):225–232, 2009.

[33] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[34] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[35] John R Quinlan et al. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348. Singapore, 1992.

[36] Sebastian Racshka. Principal component analysis in 3 simple steps. 2015.

[37] Kee Siong Ng. A simple explanation of partial least squares, 2013.

[38] Volkswagen. volkwsagen autoeuropa - 'quem somos'. 2013.

[39] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL: http://CRAN.R-project.org/doc/Rnews/.

[40] Bjørn-Helge Mevik, Ron Wehrens, and Kristian Hovde Liland. *pls: Partial Least Squares and Principal Component Regression*, 2015. R package version 2.5-0. URL: https://CRAN.R-project.org/package=pls.