|  |
| --- |
|  |
| **DataTeam**  **Purple I**  Members: Nadya Cheperkova  Mihai Alexandru Tudor  Max Post  Niek Sleddens  Rob Hendrikx |

****



Data Quality &

Machine Learning

Date: 27 October 2019

# CONTENT

[CONTENT 2](#_Toc26108449)

[VERSION INDICATION 3](#_Toc26108450)

[INTRODUCTION 4](#_Toc26108451)

[THE PROJECT 4](#_Toc26108452)

[Current situation 4](#_Toc26108453)

[Problem description 4](#_Toc26108454)

[Project goal 4](#_Toc26108455)

[Team 4](#_Toc26108456)

[Mitigations 4](#_Toc26108457)

[Metrics 4](#_Toc26108458)

[DATA REQUIREMENTS 4](#_Toc26108459)

[Sources 4](#_Toc26108460)

[Format 4](#_Toc26108461)

[Quantity 4](#_Toc26108462)

[ANALYTIC PREDICTIONS 4](#_Toc26108463)

[Model choice 4](#_Toc26108464)

[Model tests 4](#_Toc26108465)

[Model 4](#_Toc26108466)

[EVALUATION 4](#_Toc26108467)

[Deployment 4](#_Toc26108468)

[Metrics 4](#_Toc26108469)

[Performance 4](#_Toc26108470)

[CONCLUSTION AND RECOMMENDATIONS 4](#_Toc26108471)

# VERSION INDICATION

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Editor | Change |
| 0.1 | 16/10/2019 | Niek | Initial layout  Added content for Project Statement, Client, Team, Current Situation, Problem Description, Project deliverables and non-deliverables, constraints, risk |
| 0.2 | 23/10/2019 | Max | Changed Version history to correct format  Edited  Changed layout |
| 0.3 | 23/10/2019 | Nadya | Change completely the whole layout  Change completely the content for Project Statement, Client, Team, Current Situation, Problem Description, Project deliverables and non-deliverables, constraints, risk |
| 0.4 | 24/10/2019 | Mihai & Niek | Review and make some changes on current situation and problem description |
| 0.5 | 24/10/2019 | Rob & Max | Analysis the data and looking for correlations, generates some graphs |
| 0.6 | 24/10/2019 | Max | Generates graphs |
| 0.7 | 25/10/2019 | Nadya &  Mihai | Write Project Goal and KPI’s and metrics |
| 0.8 | 25/10/2019 | Rob & Max | Generates more graphs focused on delivery time |
| 0.9 | 25/10/2019 | Niek | Generates graph for showing the reasons of returning the items |
| 0.10 | 25/10/2019 | Nadya &  Mihai | Create Phasing and Gantt Chart |
| 0.11 | 26/10/2019 | Nadya | Reporting EDA |
| 0.12 | 2/10/2019 | Mihai | Write more KPI’s and metrics |
| 0.13 | 27/10/2019 | Rob | Create a graph which shows which products are sold the most for the last year |
| 0.14 | 27/10/2019 | Everyone | Reviewed |

# INTRODUCTION

# THE PROJECT

## Current situation

One of Informa’s customers, a web shop focused on ‘Garden & Lifestyle’, has reached the point where the data volume is getting bigger, more complex and more versatile. They require a machine learning /deep learning solution which should be applicable to any similar data format. The web shop’s IT department has comprised a dataset containing all the information that the web shop collects.

Informa asked our team to look into the given dataset and return with a solution supported by correlations and give insights. Based on those findings they can decide how their organization can benefit from Data Science and how to expand their partnership.

## Problem description

The company has received a request from a customer with a running web shop, which focuses on products related to gardening and garden maintenance. The web shop owner requested help with improving the sales of the shop. By looking at the data it can be noticed that the number of orders is quite steady over the last years. Moreover, a significant number of orders are being returned due to delivery delays. This cost the company a lot of money and customers.

## Problem justification

## Project goal

The web shop’s Sales Department, Marketing Department and Logistics Department can benefit from our final conclusions for the sake of improving the delivery time of the orders. Therefore, the goal of this project is to find different correlations and insights for improving the customer satisfaction, the sales and logistics with the main focus on the expected delivery time. The result of that can help the decision-makers of Informa to take further steps in creating the final solution for the improvement of their customer’s web shop. We noticed that one of the reasons of complaints is the actual delivery time being later than the expected delivery time by a large margin. Therefore, it is one of the significant reasons of why people are returning their orders. The following areas are the ones that can benefit from our findings:

**• Customer Service**

The customer satisfaction can be improved by compensating for the orders that were not delivered on time. Machine Learning algorithms can be used to create a personal profile based on the customer’s previous orders. Once that is created, the information can be used to increase the sales of low selling items by giving a discount on these items (selected items) for their next purchase.

**• Sales**

With the help of Machine Learning algorithm, historical selling can be used for improving the accuracy and scale of sales forecasts. Being able to give a buying history can accurately predict the future buying levels. Forecasting demand for products can improve the availability in the stock and the delivery time can be reduced.

**• Logistic Department**

The Logistics Department can improve its Management System with the use of Machine Learning. Based on Machine Learning algorithms, the customer of the web shop will be able to see the real-time changes of the estimated delivery time as well as how many products are available in the stock. This can be beneficial for the organization because the customer will be able to monitor, in real-time, the time of their delivery and the availability in the stock.

# TEAM

The project team consists of five young and motivated specialists who are going to work together on this project in order to achieve its objectives. It consists of project leader, project members, and other members who are not directly involved with management but are responsible for monitoring the work. The team members are with different technical and business knowledge but with the required skill set to carry out the work for the project. The structure and characteristics of the project team usually vary, but the project leader’s role remains constant.

## Project leader

The team leader is Nadya Cheperkova who is a young female specialized in Data Science.

Contact Information

**Nadya Cheperkova**

Email: n.cheperkova@student.fontys.nl

## Team members

Contact Information

**Max Post**

Email: m.post@student.fontys.nl

Contact Information

**Niek Sleddens**

Email: n.sleddens@student.fontys.nl

Contact Information

**Rob Hendrikx**

Email: rob.hendrikx@student.fontys.nl

Contact Information

**Mihai Tudor**

Email: m.tudor@student.fontys.nl

## Mitigations

In the business proposal we discussed some of the risks that come with the project. Unfortunately, a few of those risks happened while working on the project.

**The resources were limited.**

While working on the project we found out that the dataset is entirely randomly generated which made finding correlations and making accurate predictions impossible. Our solution to this was to ignore this and started making the models anyway. These models predict the delivery time for orders. These results itself aren’t very useful as there is nothing to predict due to lack of correlations, but the models would be able to accurately predict delivery times if the data was real.

## Metrics

We all know that expectation management is really important, hence, we chose to focus on predicting the actual delivery time for the orders. Therefore, we can help close the gap between the predicted and the actual delivery time. Studies have shown that customers would rather wait longer for their order to arrive if they expect it to take that long, than receive their order sooner but later than expected.

In order to predict this, we used the latitude and longitude values of the place the customer lives in and the product number

**Completed tasks**

Firstly, we set up the data requirements. We defined which data is needed and what this data should look like formatting wise. Afterwards we cleaned the dataset based on these requirements so that we could start analysing it and making predictions based on the data.

When the data was ready to be used, we started thinking about which models should be used for the prediction. We picked linear regression, polynomial regression, decision tree regression, support vector regression and random forest regression. We chose them because they are the most suitable ones for the regression problem we are facing.

Based on the results from these models we were able to evaluate, draw conclusions and give recommendations.

# DATA REQUIREMENTS

## Sources

**Informa:** The dataset provided by Informa contains the orders made in the last years.

*The dataset contains the following features:* Customer Code, DoB (Date of Birth), Gender, City, Order-Nr, Order Date, Product-Nr, Sub Category, Category, Amount, Price, Expected Delivery Time, Actual Delivery Time, Reason of Return, Rating. This data will need to be seen as very trustworthy since the client provided this set.

**Longitude & latitude:** The [longitude latitude set](https://github.com/bobdenotter/4pp) consists of 2384 towns in the Netherlands.

*The dataset contains:* City, Longitude, Latitude. The dataset is missing some towns compared to Informa set. The set is trustworthy since all longitude and latitude are within those of The Netherlands and a couple were checked and found correct.

## Format

The existing columns that will be used are Customer number, Product number, Subcategory, Category, Actual Delivery Time, Latitude, Longitude.

Two extra columns have been added that determine the latitude and longitude of the towns and cities, the values are expressed in decimals. These columns are named: Latitude, Longitude.

*Order date – Date (dd-mm-yyyy) -> it2*

**Product number:** The column Product number contains the product codes. These codes are defined as strings, their format will be changed to numbers. We will need this feature in case we are going to predict delivery times per product.

**Subcategory:** The column Subcategory describes the different subcategories that are available in the shop. It is of type string and will remain so. We will need this feature in case we will predict delivery times per subcategory.

**Category:** The column Category describes the different categories available in the shop. It is of the string type and will remain so. We will need this feature if we want to predict delivery times based on category.

*Count – Integer -> it2*

**Actual delivery time:** The column Actual Delivery Time describes the number of days it took for the order to arrive at the customer’s address since ordering. It is of type integer and will be kept as such. This will be our target for the model.

**Latitude**: The column Latitude contains the latitude of the towns in which the customers reside. It was created based on the initial information regarding towns. Decimals will be used to describe these values.

**Longitude:** The column Longitude contains the longitude of the towns in which the customers reside. It was created based on the initial information regarding towns. Decimals will be used to describe these values.

## Quantity

The original dataset contains 2.604.981 rows of orders. Out of these we will need 20.000 of high-quality entries for training and adjusting the model. The reason for choosing the number 20.000 is because this amount is representative of the full set for training and eliminates variance while having a reduced processing time.

# ANALYTIC PREDICTIONS

## Model choice

## Model tests

## Model

\

\

# 

# EVALUATION

# “Mean absolute error measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.”

# “The mean squared error tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. The squaring is necessary to remove any negative signs. It also gives more weight to larger differences. It’s called the mean squared error as you’re finding the average of a set of errors.”

# “R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.”

## Linear regressor

## 

# mean\_absolute\_error: 2.4622382022314824

# mean\_squared\_error: 20.16856488839667

# r2\_score: -0.0007712555701668844

# Linear regression does not seem to be actively predicting values and always predicts delivery time to be between 3.6 and 3.8 days rounded as 4 which is not a real predictor.

## Polynomial Regression

## 

# mean\_absolute\_error: 2.4604015741518914

# mean\_squared\_error: 20.169757293887347

# r2\_score: -0.0008304231483589053

# Polynomial regression is not varying much for lower in the left of the graph (middle/left of The Netherlands) and varies more on the scores and starts to vary more if people live further from the middle/left.

## Decision Tree

## 

## Afbeelding met tekst, kaart Automatisch gegenereerde beschrijving

Representation of the model when limited to only 40 nodes.

## mean\_absolute\_error: 3.4930027988804477

## mean\_squared\_error: 43.04438224710116

## r2\_score: -1.1358773271695162

## The decision tree seems to be making some more accurate predictions over a more spread out number of days. The model seems to be stacking predictions on other values and therefore looks very promising.

## Support vector Regressor

## 

mean\_absolute\_error: 2.126736824689831

mean\_squared\_error: 21.06009768751201

r2\_score: -0.045009425399788805

The SVR model is spreading is only making predictions between 1 and 6 days and does not reach any of the days above 6 and rarely above 4.

## Random Forest Regressor

## Afbeelding met schermafbeelding Automatisch gegenereerde beschrijving

mean\_absolute\_error: 2.8882167609146823

mean\_squared\_error: 24.74920754572321

r2\_score: -0.22806434899836137

## The Random Forest classifier is following the values really well however does seem to have some trouble to see when a order will be delivered in one day.

# CONCLUSTION AND RECOMMENDATIONS