Logo

Description automatically generated

**PROJECT**

**PROPOSAL**

Project: optimizing travel time

Class: AI44

Mentor: Martijn Lamers

Client: Ordina

Members:

Alex Svetoslavov

Antonia Dineva

Bas van Mil

David Lima Ruas van Rijthoven

Eva Bijker

Ryan van Wegen

[1 Introduction 3](#_Toc101442095)

[2 Project Statement 4](#_Toc101442096)

[2.1 The company 4](#_Toc101442097)

[2.2 Our group (Pythonatic) 4](#_Toc101442098)

[2.3 Stakeholders 4](#_Toc101442099)

[2.4 Problem Description 5](#_Toc101442100)

[2.5 Project goal 5](#_Toc101442101)

[2.6 Scope 6](#_Toc101442102)

[2.7 Project deliverables & Non deliverables 6](#_Toc101442103)

[2.8 Project Constraints 6](#_Toc101442104)

[2.9 Project Risks 8](#_Toc101442105)

[2.10 Proposal solution 10](#_Toc101442106)

[2.11 Data sourcing 10](#_Toc101442107)

[2.12 Target variable 10](#_Toc101442108)

[3 Methodology and phasing 12](#_Toc101442109)

[3.1 Methodology 12](#_Toc101442110)

[3.2 Approach 13](#_Toc101442111)

[4 Planning 14](#_Toc101442112)

[4.1 First 12 weeks 14](#_Toc101442113)

[4.2 Week 13-16 14](#_Toc101442114)

[5 Required skills 16](#_Toc101442115)

[6 MOSCOW for Jupyter notebook 17](#_Toc101442116)

[7 Conclusion 18](#_Toc101442117)

[8 Appendix 19](#_Toc101442118)

[8.1 Contacts 19](#_Toc101442119)

[8.2 Domain understanding 19](#_Toc101442120)

# Introduction

In the morning, a lot of people need to get to their work. They can use the car; they can go by train or use another type of transportation.

Using the train is a lot more sustainable than using a car using fossil fuel. It is also a struggle when there is traffic in the morning rush, which means they will be late for their work or must leave early.

Nowadays an electric car is a lot more popular and is more sustainable than using fossil fuel. But this is still less sustainable than the train.

It is not possible to use a train whenever you would like, because trains have a tight schedule they follow. This means travelers sometimes must wait for their next train. It can also happen that a train has a delay or even will not drive anymore.

This project is meant for Ordina to optimize the travel for their employees using train departure times.

# Project Statement

## The company

Ordina is a local, independent IT service company, in Benelux with approximately 2.650 employees. Before the Corona crisis started, most consultants worked on projects and assignments at the customers, spread out over the Netherlands.

The head office in the Netherlands of Ordina, is in Nieuwegein. It is located directly next to the highway, and it is easily accessible by car and public transport.

## Our group (Pythonatic)

Pythonatic is an IT consultant organization, established in Eindhoven. The group consists of 6 students of Fontys University of applied science - Eva Bijker, Antonia Dineva, David Lima Ruas van Rijthoven, Bas van Mil, Alex Svetoslavov and Ryan Wegen. Contact details can be found in the appendix. We also have a couple of roles in our group. Antonia is the Secretary; Eva is the Finisher and the contacts person, and Ryan is the Leader. Our secretary makes sure that the minutes of the meeting are made after every meeting. Those notes are made if we have a meeting with the client or teachers. The finisher makes sure that everyone does their tasks and checks up on group members. The leader makes sure that the group work is going smoothly, and the contacts person is the one who will be the contact person for the client.

The organization aims to supply efficient solutions and manage problems by offering an IT product. Establishing a good connection with the client is the main priority as well as delivering wonderful experience and solutions that best fit the clients' needs. We focus on cooperation with the client and through the constant process of collecting feedback to understand the client's needs and to improve our solution based on that.

## Stakeholders

The project will go through several phases. They will be marked by meeting with the client after which feedback will be collected and considered. The stakeholders are everyone that is involved in the project either as a developer, customer, or user. Those roles and their representatives are outlined below.

* Our client is Ordina company. Our project tutor is Mr. Lamers. They are the people who will be reported to and keep informed about the project development. They are also the people whose feedback will be collected before the final product is presented and ready to be used by the end user.
* The end users are employees of Ordina. They are the target audience and the one who will use the product.
* Pythonatic is the group working on this project to deliver the product to meet the stakeholders' needs in the end. Based on the planning outlined in this document they will work to deliver the agreed upon product.
* NS/the train companies are also an indirect stakeholder of our project. This is because if we can accurately predict which trains are going to have a delay then they would also want to investigate it so they can improve their services.

## Problem Description

Before the Corona crisis most consultants at Ordina worked on projects and assignments at the customers who are all spread out over the Netherlands. Ordina wants to stimulate their employees to travel sustainably such as by train. The goal is to optimize the travel of employees who are going by train and to figure out how the train departure data can help with that.

## Project goal

Our goal is to improve an aspect of Ordina employees’ travel. We have been given carte blanche to decide for ourselves what we want to improve in their employees’ commute.

We decided to improve Ordina employees’ travel times in the Netherlands by giving them proper advice for which trains they should avoid based on how often that route experiences a delay.

## Scope

The main goal that we were given by Ordina is to optimize travel for the employees. We have decided to focus on the train travel by the employees and how to optimize that. Travelling by electric car could be an option but this is out of our scope, and we will not be researching that option further.

## Project deliverables & Non deliverables

Phase 1 Deliverables:

* Project Proposal Document (with Exploratory Data Analysis Notebook attached)

Phase 2 Deliverables:

* Data Storage Solution
* Jupyter notebook (with the prepared data)

Phase 3 Deliverables:

* Jupyter notebook (with an effective model included)

Phase 4 Deliverables:

* Final Delivery (Delivery document with Demonstration Materials attached)

Non-Deliverables:

* Minutes of Meeting documentation
* Peer-to-peer review

## Project Constraints

Project constraint will be outlined for the general limitations to be considered during the life cycle of this project.

1. Time constraint. We have 16 weeks (about 3 and a half months) to present a solution that fits client needs.
2. Budget. We are doing this project for study purposes the project has no budget.
3. Scope. What the project includes and what it does not are in the project scope. The scope of this project refers to the specific deliverables that are agreed upon by the stakeholders. The team will deliver those deliverables and any other deliverables that are not included in the scope even though desired from client will be left out of scope.
4. Quality. The focus here is on the characteristics of the delivered product. It will be evaluated on how closely it matches expectations. As we are university students, we have our knowledge and skills to depend on when delivering the product. That is why its quality will match our knowledge and will be outlined when defining scope.
5. Resources. The available resources, both human and material, are considered. The human resources are the 6 members, and the materials are the hardware (laptops) that they will use for developing and delivering the product.

## Project Risks

In this table you can see an overview of some project risks.

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Probability** | **Impact** | **Overcome** |
| Someone gets ill (covid etc.) | Medium | Low | Do your tasks as soon as possible and if someone is ill and can’t work on it the teammates will take over |
| Miscommunication | Low | Big | We try to meet as often as possible at a location and always double check before sending the mail. We will also do a peer-to-peer review every now and then (see planning) |
| We focus on the wrong part of the problem | Low | Big | Ask for feedback from the teachers and clients as often as possible |
| New lockdown for covid | Low | Medium | If possible, meet at someone's place and otherwise often have Teams meetings with the camera on. Also try to do fun activities together. |
| Laptop breaks down/ laptop failure | Low | Big | We will save every doc also on teams so nothing will get lost. Teammates can take over work if needed, but that team member is responsible for buying a new laptop/getting it fixed. |
| We can’t find enough space for our dataset | Medium | Big | Make the data set as small as possible and probably find an online space to save it in if it's too big to save locally |
| Our notebook gets super slow because our dataset is too big | Medium | Big | Make multiple small notebooks and save the cleaned dataset so it can be loaded in other documents. |
| Poor time management/mistake planning | Low | Medium | If we find out that we are behind schedule we would have to make over hours at the weekend to keep up or/and scope down the scale of our project |

## Proposal solution

Based on the analysis that Pythonatic made, the solution the team offers is a Jupyter notebook. The notebook will be a balance between text descriptions and visualizations. The Jupyter notebook can be run on web browser like Firefox or Google Chrome, using the Jupyter notebook application.

The notebook could be used to make the current process of Ordina employing travel more efficient. In the notebook the client can find an analysis of the current travel time of trains and predictions which train rides cause the most delay and therefore may be avoided (based on the start location).

The employees of Ordina company will be able to travel to work by train faster if they choose to use the insights gained from the analysis in the notebook. The insight can be used to make a website/application in the future. Which could be accessed by the employee at any time and where they could impute their start destination and departing time and get estimate on the train delays.

## Data sourcing

We got our train data from the client, and it is from the following website: <https://trein.fwrite.org/idx/DVS.html> . This is an open-source data source filled with all the train data from every day from 2013 to 2016 in the Netherlands. From this link we are only going to investigate the data from 2016. We are also going to need a dataset with the vacation's dates in the Netherlands.

## Target variable

Using the train travel data, we will try to predict train delays. We might also detect patterns we can use to give advice to Ordina employees.

This means we will try to predict train delays based on the following features:

**Time of the day** will have a strong influence on travel times and delays, we can notice things like rush hour trends

**Day of the week** will help us determine days more likely to experience delays, and it can be paired with Time of the day for more in-depth conclusions.

**Whether the day is a state holiday** can have a significant impact on train travel and even what train routes are available, we should keep this in mind when looking at trends.

**Day of the month** will help us determine the seasonal trends

**Delay** will help us determine the average delay for a certain ride and finally predict the trains rides not to take

**Begin Station,** this will help us indicate the different train rides and help us filter the prediction for the user

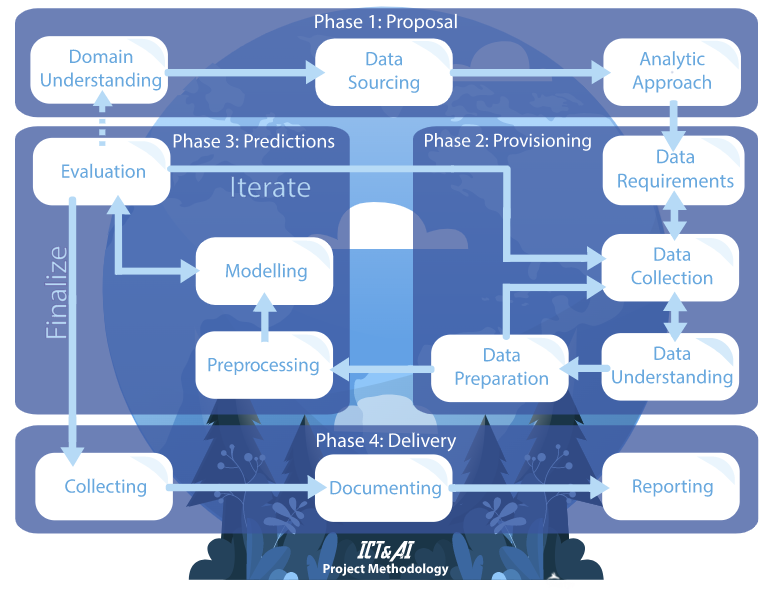
**End Station,** this will help us indicate the different train rides

# Methodology and phasing

## Methodology

The methodology we are going to use for the project is based on the IBM Foundational Methodology for Data Science.

The diagram below shows a general overview of the AI project methodology and its steps, divided into 4 phases.

In the first phase we will deliver the project proposal document, which has the domain understanding, data sourcing and the analytic approach. In the second phase, we will gather data requirements, explain the data collection, and understand and prepare the dataset. In the third phase, the machine learning model will be prepared with the dataset and the results will be evaluated. Lastly, in the fourth phase we will report and deliver our work.

During the development of the project, we will pass through all phases. However, in case the prediction results are evaluated negatively, and the accuracy score is not high enough there is possibility to work iteratively going back from phase 3 to phase 2.

## Approach

In the first 12 weeks of the project development, our team gathers at the TQ building for one full day per week. The next 4 weeks, we will spend five full working days per week and deliver the final product by the end of week 16.

In case a team member cannot attend the building, an online meeting in Microsoft Teams will be created, so he/she can still communicate with the rest of the group. If he/she gets sick, he/she must notify the other team members and his/her work will be done by the others if needed.

Communication with the client is done on a demand basis, such as when proposal is done, or considerable progress is made. In the meantime, if questions arise, the client person will contact the clients via email or will propose scheduling a short meeting with the group. Technical questions will be sent to the teachers.

The algorithms we are going to try will be SARIMAX, which stands for Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors model. And we will also try a classification and regression model. As mentioned in 2.10, we want to predict the train delays based on past data, including relevant indicators such as the time when the trains departure for each day and week, days where there were delays and whether there is a holiday.

# Planning

## First 12 weeks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Week** | **Phase** | **Main activity** | **Document** | **Side Activity** |
| 1 | 1 | Setting up the team |  |  |
| 2 | 1 | Brainstorming | Domain understanding  (Research doc) | Figuring out the data |
| 3 | 1 | Domain understanding | Domain understanding  (Research doc) | Figuring out the data |
| 4 | 1 | Domain understanding | Domain understanding  (Research doc) | Figuring out the data |
| 5 | 1 | Domain understanding & data sourcing | Domain understanding  (Research doc)/  Proposal | Figuring out the data |
| 6 | 1 | Data sourcing/Peer-to-peer review | Proposal | Figuring out the data |
| 7 | 1 | Data sourcing | Proposal | Figuring out the data |
| 8 | 1 | Data sourcing & analytic approach | Proposal | Figuring out the data |
| 9 | 1 | Analytic approach | Proposal | Figuring out the data |
| 10 | 1 | Analytic approach | Proposal | Figuring out the data |
| 11 | 2 | Data requirements | Notebook | Presentation client/teacher |
| 12 | 2 | Data requirements/Peer-to-peer review | Notebook | Presentation client/teacher |

## Week 13-16

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week** | **Day** | **Phase** | **Main activity** | **Document** | **Side Activity/** **back-up planning** |
| 13 | Mon | 2 | Data collection | Notebook |  |
|  | Tue | 2 | Data collection/ Data understanding | Notebook |  |
|  | Wed | 2 | Data understanding | Notebook |  |
|  | Thu | 2 | Data preparation | Notebook / Prepared dataset |  |
|  | Fri | 2 | Data preparation/Peer-to-peer review | Notebook/Prepared dataset |  |
| 14 | Mon | 3 | Preprocessing | Notebook |  |
|  | Tue | 3 | Preprocessing/ Modelling | Notebook |  |
|  | Wed | 3 | Modelling | Notebook |  |
|  | Thu | 3 | Evaluation | Notebook |  |
|  | Fri | 3 | Evaluation  /Peer-to-peer review | Notebook |  |
| 15 | Mon | 2 & 3 | Redo things if needed | Notebook | collecting |
|  | Tue | 2 & 3 | Redo things if needed | Notebook | collecting |
|  | Wed | 2 & 3 | Redo things if needed | Notebook | collecting |
|  | Thu | 2 & 3 | Redo things if needed | Notebook | Documenting |
|  | Fri | 2 & 3 | Redo things if needed/Peer-to-peer review | Notebook | documenting |
| 16 | Mon | 4 | Collecting | Delivery document | documenting/reporting |
|  | Tue | 4 | Collecting &documenting | Delivery document | reporting |
|  | Wed | 4 | Documenting | Delivery document | reporting |
|  | Thu | 4 | Reporting | Delivery document | Presentation client/teacher |
|  | Fri | 4 | Reporting/Peer-to-peer review | Delivery document | Presentation client/teacher |

# Required skills

Our team members will need:

* Critical analysis skills to understand our problem and its data
* Basic understanding of Python
* Basic understanding of Jupyter notebooks
* Knowledge of adequate data analysis and visualization methods
* Knowledge of existing algorithms we can apply to our problem
* Formal writing skills for professional documents
* Soft skills to communicate with our client

# MOSCOW for Jupyter notebook

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Project Features** | **Must** | **Should** | **Could** | **Won't** |
| Ability to predict which trains not to take | **X** |  |  |  |
| Ability to predict the average delay for a specific train ride | **X** |  |  |  |
| Visualization of probability of delay (per train ride) | **X** |  |  |  |
| Ability to display graphs after saved | **X** |  |  |  |
| Ability to display description to the plots |  | **X** |  |  |
| Processed and clean data used for the plots | **X** |  |  |  |
| Easy access (on mobile device) |  |  |  | **X** |
| The visualization depended on time and location |  |  | **X** |  |
| Ability to load within a few seconds |  |  | **X** |  |
| Mock-up application/website |  |  | **X** |  |

In the MOSCOW above you can see the features our end product must, should, could or won't have. It is useful to get an overview of what we are planning to do.

# Conclusion

Based on the information we got and the analysis we performed, this project and the proposed solution seems to be a realistic one for the time frame and other constraints that are presented in this document. Pythonatic believes that given the fact that Ordina agrees on cooperating we can achieve optimized travel time for the company’s employees with the means of using the Jupiter Notebook insights. An advice will be given to the employee: how big is the chance of a delay on the route and/or should they take another route because there can be a slighter chance of any delay. This mainly depends on what time the employee travels and the route they take. The project will create a clear view on ‘normal’ delays and extreme delays, which will be used to make the best advice as possible for the employee.

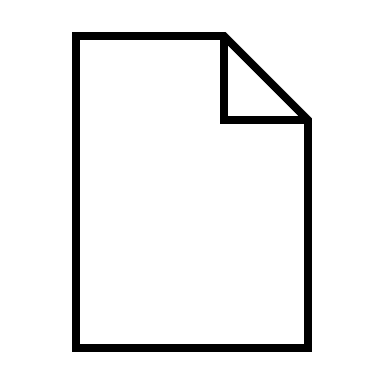
Cooperation between client and the team is crucial for the final delivery as quality and key stakeholders’ satisfaction is a main priority in this project. For further questions do not hesitate to contact any of the team members whose contacts are given in the appendix below.

# Appendix

## Contacts

* David van Rijthoven, [392425@student.fontys.nl](mailto:392425@student.fontys.nl)
* Alex Svetoslavov, [a.svetoslavov@student.fontys.nl](mailto:a.svetoslavov@student.fontys.nl)
* Antonia Dineva, [a.dineva@student.fontys.nl](mailto:a.dineva@student.fontys.nl)
* Bas van Mil, [bas.vanmil@student.fontys.nl](mailto:bas.vanmil@student.fontys.nl)
* Eva Bijker, [e.bijker@student.fontys.nl](mailto:e.bijker@student.fontys.nl)
* Ryan van Wegen, [r.vanwegen@student.fontys.nl](mailto:r.vanwegen@student.fontys.nl)

## Domain understanding



See Domain understanding document