Personal Development Report

ADS-A

Brent Schoenmakers | 2018-2019

Summary

This will be filled in at the end of the semester.

Inhoud

Introduction

This personal development report is for me to document my experiences and growth as a data scientist. But first, I'm going to introduce myself.

My name is Brent Schoenmakers, and I live in a village called Oisterwijk. Oisterwijk is located near Tilburg, and it takes me about 40 minutes to travel to school. I chose to ICT as my study of choice, because I was always interested in how the computer really works. I was very interested in how certain programs work, and how they might work together with different programs, and how there is rarely an error presented to the user.

I'm currently in the 4th semester of Technology, meaning that I've already done a specialization route previous semester. This specialization was Game Design. But for this semester, I wanted to try something totally different. I specifically wanted to delve deeper into the world of machine learning, because I've always found it very interesting in how a machine makes predictions with the usage of algorithms. This is the reason I chose Applied Data Science as my second specialization route.

When starting this course, I had zero knowledge of Data Science. Basically, everything that I've learned this course is completely new to me.

Learning objectives in Applied Data Science

In ADS-A there are 8 different objectives which I need to convince the teachers that I've grown in over the course of this semester. These are:

- **Reporting**
  - You are able to report in a methodologically sound way about a data analysis (plan, process documentation, report of final results, etc.).
- **Machine Learning**
  - You are able to apply machine learning algorithms for classification and regression (supervised learning) to a given data set.
- **Data Driven Organization**
  - You are able to explain what a 'data driven organisation' is, are able to argue on the maturity level of such organisation and are able to translate this into a business case for the application of data science.
- **Business Requirements**
  - You are able to translate business requirements into a structured data analysis plan.
- **Cross Validation**
  - You are able to improve the quality of machine learning models using cross validation techniques and systematic searches of the model's hyper parameters.
- **Data Quality**
  - You are able to clean data sets according to theories of data quality, in such a way that the process of cleaning is repeatable and the final result is data set suitable for data analysis.
- **Data Ethics**
  - You are aware of, and are able to reflect on your own choices in terms of the fact that laws exist regarding digital data and can explain the term "data ethics".
- **Work Ethos**
  - You are an effective co-worker in project groups, and are able to guide your own study progression by asking for, interpreting and applying feedback by teachers, tutors, coaches and fellow students.

Below I will elaborate on every single topic of the ones named above. I will describe how I have grown over the course of this semesters, and how I achieved that growth.

Learning Objective	What did I learn	How did I learn
Reporting	• I learned how to create different types of graphs. • I became better at presenting • I've grown in documenting my findings.	.
Machine Learning	• I learned how the Knn algorithm works • I learned what different machine-learning types there are • I've learned how to use the decision tree algorithm • I've learned about the usage of the SVM algorithm	
Data Driven Organization	• I learned what it means to be data driven • I learned how to be data driven • I learned what it means to have clean and accessible data. • I learned about the hallmarks of a data driven organization. • Together with my group I came up with my own data drive organization with its own data driven infrastructure	
Business Requirements	• I've learned how to create a business case • I've learned how to analyze existing business case.	

Cross Validation	• I’ve learned about the importance of cross validating and what It actually means. • I’ve learned about Kfold cross validation	
Data Quality		
Data Ethics		
Work Ethos	• I learned the value of preparation work • I learned how to make a mindmap • I learned how to differentiate different types of analytics through groupwork. • I wrote a report using a questionnaire filled in by a company together with my subgroup • Every Thursday morning I have to work together with my group on a presentation. • Every Friday I work together with my group on the machine learning exercises.	

How did I learn?

REPORTING

To me, reporting is showcasing data to the unknowing. This can be achieved by visualizing the data, documenting or presenting. In the weeks 1-7, I've learned how to visualize data in a lot of different ways. Every week there were 1-2 new visualization methods presented to us. In week 7, I've learned to use the following visualization methods:

- Box plots
- Scatter plots
- Parallel coordinates
- Bar charts
- Line charts
- Scatter matrix
- Decision trees & random forest
- Support vector machines

Aside from visualizations, I've also improved my documenting work. In week 3-7, I was tasked, together with my subgroup, to interview a company about Data Science. After the interview, I had to make a report based on the answers of the contact person. I wrote about the strength and weaknesses of the company, and gave them my thought on where they could improve on. The company was very happy with my findings and said that the report that I and my group wrote for them, was of high value to the company.

Last but not least, I've also grown became more skilled in reporting via presenting. During the Thursday classes, the groups are tasked to make a little presentation about a certain topic that is being the center Point that week. Each week, our group appoints two people that present that week. I've personally had to present twice so far. To me, those times that I've presented had been proven to be really valuable. I always felt that presenting was a skill that I wasn't very good in. But, to become better at presenting, you simply have to present more. And that's exactly what I aim to do in the future.

MACHINE LEARNING

Machine learned to me is all about applying the correct algorithm. To know what algorithm to use in what situation, I first had to understand and learn about the different types of algorithms there are. I researched about the pro's and cons for each type of machine learning algorithm, and the general usage for each type.

I've also learned how to work with a few different algorithms myself. In week 1 we used the Knn-algorithm to classify different types of flowers, wines and computer parts.

In week 5 we used the decision tree algorithm to figure out characteristics about the passengers of the Titanic when it sank

In week 6 we used the SVM algorithm on the Iris dataset.

I've also found it important to notice that machine learning is divided in a number of steps:

1. Preparing the data
2. Analyzing and visualizing the data
3. Cleaning the data
4. Feature selection
5. Dividing your data into test and training sets
6. Training the algorithm
7. Applying machine learning
8. Evaluation

During the three cases where I worked with machine learning, some of the steps had already been done. For example, the preparation and cleaning of the data had already been done, because it was a dataset from internet. Nevertheless, the remaining steps all needed to be applied in every case. I think the three cases provided were great learning steps towards me being more skilled with machine learning.

DATA DRIVEN ORGANISATION

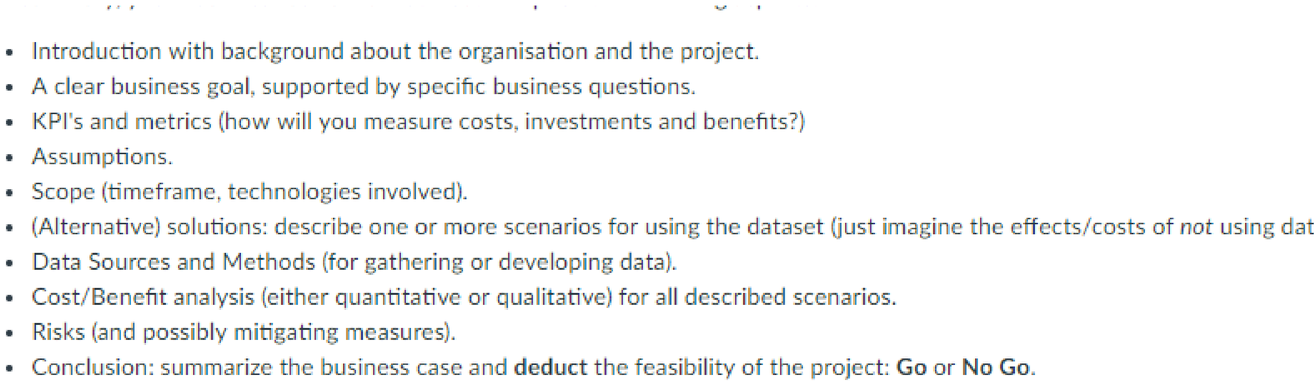
During these seven weeks I've learned a lot of how a data driven organization operates. In the early weeks I've read about what it takes to create a data driven organization and what the challenges are. What tools and people you need to be data driven, and how to handle problems.

Most of the stuff I did for this topic was reading up on documents, but I did learn some things in class as well.

During class me and my group were tasked to create our own Data driven organization. At this point we had all read up on the huge document by Carl Anderson; How to create a Data Driven Organization. Using the knowledge we gained from the document, we were able to present our idea of having a data driven organization. Generally speaking, we had a good idea, but missed some points.

BUSINESS REQUIREMENTS

This topic is all about writing business cases in my opinion. But, the business case has a lot of concepts mixed into it. With a business case, you also write about:



- Introduction with background about the organisation and the project.
- A clear business goal, supported by specific business questions.
- KPI's and metrics (how will you measure costs, investments and benefits?)
- Assumptions.
- Scope (timeframe, technologies involved).
- (Alternative) solutions: describe one or more scenarios for using the dataset (just imagine the effects/costs of *not* using dat
- Data Sources and Methods (for gathering or developing data).
- Cost/Benefit analysis (either quantitative or qualitative) for all described scenarios.
- Risks (and possibly mitigating measures).
- Conclusion: summarize the business case and **deduct** the feasibility of the project: **Go** or **No Go**.

During class, me and my group also had to present our own business case that we created from scratch. The business case that I presented was said to be headed the right direction, but it needed some thinking, because we weren't being very ethical with our business case.

For our project, we also have to write a business case. One that is not finished yet as of the data of me writing this version of the PDR. The struggle of writing a good business case become very clear in the project. We have a dataset that does not obviously point out failures within the company, so it is needed to explore the dataset exceptionally well in order to translate it to a business case.

To me, the difficulties of writing a business case are:

- Thinking about the KPI's and metrics of the company.
- Predicting what types of methods and data sources you are going to need.

CROSS VALIDATION

We only touched this topic as of last week, but I've learned a few things about cross validating in 1 week.

I've learned that cross validation is a method that is used to counter potentially overfitting your test set. Normally you create your test and training set, and train your algorithm. With cross validation, you create an additional set, called the validation set. First your train your training set, than you evaluate is being done on the validation set, and finally you evaluate your test set.

A risk of cross validation is that you now need to split up your data in three partitions, rather than 2. This means that your test and training data will be significantly smaller in size when using cross validation. Also the results can depend on a particular random choice for the pair of training and validation sets.

I've also studied on the usage of the Kfold function. This function splits your data in groups of equal size. This basically means that it validates your data using different groups of data each time.

As this is still a relatively new topic, I intend to learn more about this topic in the next coming weeks.

DATA QUALITY

WORK ETHOS

During the classes on Thursday and Friday, there is a lot of groupwork involved. Every Thursday morning, me and my group are tasked with making a presentation about the topic of that week. E.g: We had to create our own data driven company and presenting our company. On Friday mornings, I make the machine learning exercises together with my group.

For the first challenge, I was also tasked to work together. This time I was tasked to work together with my subgroup. We had to write a report about Mise en Place, a company that one of my subgroup members works for. During the challenge, I was keen on dividing the work evenly. This made the challenge a really relaxing and fun assignment, as it was genuinely fun talking to a company about Data Science, and not having to do too much, as the workload was divided really well.

There is also the project, where I work together with five other students. At the moment, we are still busy with exploring the data and creating the business case. I'm the project leader, and so it is my task to make sure everyone knows what to work on and what the goal is. We work on the project either on Thursday afternoon or Friday afternoon, depending on the week. When we begin working on the project, we do a little standup on the current situation and what everyone is going to do that day, and after that we all get to work. If someone is having a difficult time, we offer help to that person. In general, I think our group is working well together.