Personal Development Report

ADS-A

Brent Schoenmakers | 2018-2019

Summary

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

Table of content

Summary 1
Introduction 3
Learning objectives in Applied Data Science 4
How did I learn? 8
- Reporting 8
- Machine learning 9
- Data driven organisation 10
- Business requirements 11
- Cross validation 12
- Data quality 13
- Data ethics 14
- Work ethos 15

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
Reporting	• I learned how to create different types of graphs. • I became better at presenting • I've grown in documenting my findings. • I completed my challenge where I reported my findings in a report
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	• I've learned what it means to have a clean or dirty dataset • I've learned how to deal with several data quality problems • I've cleaned a titanic dataset and came up with a function that cleans it
Data Ethics	• I've gained insight on what it means to handle data in an ethical way. • I've thought about data ethics in our project. • I've been given insight by someone that works for a data science company how much GDPR changed his life within the company.

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
- Regression algorithms

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.

In week 10 I was also tasked to make a report about thought up project that I had undertaken. I wrote a report about a project that predicted the usage of renewable energy in the future. This Data analysis report is supposed to be the final part of any data science project, where you report your findings to the client.

In the personal challenge I've also honed my reporting skills. There I was also tasked to make a data analysis report, where I reported my findings about video game sales in the future. There, I concluded that the North-american sales are prone to do the best in the near future and gave the client the advice to invest in north-american games in the Action Genre.

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.

In the last couple of weeks I've transferred my focus of machine learning from classification to regression. It came under my attention that I was doing okay in classification while I knew next to nothing about regression. In week 9 I was introduced to regression and since then I've been working with it constantly. First of all I was introduced to regression by making the week 9 machine learning algorithm exercise. Here I was introduced to the topic and gave me a simple basic understanding.

During my personal challenge it came under my attention that I had to predict sales, which were continuous values. Regression shines with using continuous values in predictions, so the personal challenge was a perfect way to improve my regression skills. In my challenge I've used Linear Regression, Support vector Regression and decision tree regression. After implementing those, I compared the three algorithms and came to the conclusion that SVR was not the way to go, and that Linear regression and decision tree algorithms were more suited for that particular task.

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.

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.

I've also applied a data maturity model to three different companies, assessing their maturity in different data driven aspects. This really set me thinking what it really means to be data driven, and that even the big names like Google and Amazon can be more mature in being data-driven.

In the first challenge, I was also tasked to interview a company and question them about their data driven ness. Then I wrote a report, giving them their current situation in their maturity in data driven ness, then giving them tips on how they could improve in the near future. The company in question was Mise en Place, which is an employment agency that just started to use data to improve internal processes. Things that stuck out was that there weren't many people that could actually read and handle data. Along with this, we gave Mise en Place some more feedback and they've replied that it was definitely something that they could use in the future, and thanking us for our time and our report.

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. The business case consists of the following parts:

- 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. 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. While this business case was generally accepted by the client, it came with a business solution that was very hard to achieve. Because of this, we came up with two new business solutions and had to rewrite the 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.

In the personal challenge I've also written a business case, talking about how I could tackle the problem of predicting the game sales.

CROSS VALIDATION

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.

I've also been trained on using cross validation by some Udacity courses. These went over the basics of cross validation, why it is used and how to implement it. These proved to be very useful. While we are nearing the end of the project, cross validation becomes a very good tool to check if we are not overfitting our machine learning algorithms. By detecting an overfitting, we can adjust our models to come up with better predictions with better scores.

DATA QUALITY

When obtaining datasets there is a big chance that some of the rows or columns contain data that you can't use in your visualizations and predictions. Think of data that has hidden whitespaces, empty fields or data that is in the wrong form. In order to use this data, it'll have to be cleaned. Data cleaning is a key part of every data science project. My experience with data cleaning started in week 8, where I received the task to clean a GRAIN dataset in one single function. This dataset came with some money values that were listed in different formats, it came with values that had hidden whitespaces and lines. The function that I created was able to fix all of those whitespaces and values.

The second time data cleaning was necessary was in the project. The project came with a dataset that had different methods for writing dates. Some came in the YYYY-MM-DD format while some came in the DD-MM-YYYY format. This was of course not practical in order for them to be used in predictions, so they needed to be converted into one single format. The format was chose to be YYYY-MM-DD because our database, an MYSQL database, was able to handle this format better than other formats.

In my personal challenge it was not necessary to implement data cleaning because I received my dataset from Kaggle and it was already clean.

Data quality for me also means that you should look into the source of the dataset. If the dataset is from Wikipedia for example, you can get a great accuracy. The thing with Wikipedia is that everyone can mess up the data, or put in data that is completely false. This might result in a good accuracy, but the data is from a source that can't be trusted, so in the end, can you good accurate result be trusted? I think not. A part of why I got my dataset from Kaggle is because I knew I could trust the source. Kaggle has been praised by many data scientists for providing very accurate and clean datasets. Kaggle is also a company that got bought by Google themselves, proving more that they can be trusted in providing good quality datasets.

DATA ETHICS

Being ethical is very important we working with data. Often you are working with very personal data, like names of people. Almost every time people do not prefer that this data is shared with other people, and they prefer to keep that information private. Data ethics is all about using your data in an ethical way. The experiences I've had with data ethics come from two situations; namely the workshop in week 7 and the project.

In the workshop in week 7 we were given information about what data ethics was all about, the recent changes with GDPR, and what to look out for when working with private data and how to avoid leaking it. GDPR has had a very large impact on the data science, with a lot of things that were allowed previously that are not allowed anymore. The consultant that came in our classroom in week 12 is a graduated student that recently started working in a data science company. He's told us how much GDPR changed his company when it was introduced. Beforehand, they were able to share data with other companies on a regular basis, but after GDPR, loads of contracts had to be signed in order to share data with other companies.

In the project we also had to deal with data ethics, as we were dealing with some personal data from the customers of the client. Luckily for us, the client has already thought about data ethics and has shielded us from various ethical problems. For example, they purposely left out the names of the clients, corrupted the days and months from the date of birth, and mixed up some of the locations of the customers. While these prove not to be very useful for us anyways, it was good from the client to shield us from using these columns in a unethical way.

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.