CSC8631 Report

Muzaffer Senkal - 210351491

3/12/2021

Introduction

Learning analytics refers to acquisition all data about learners and analysis improving learners outcomes. Most universities and online course platforms are starting their own learning analytics programs. It is expected that many differences will be created in education by using big data. FutureLearn is one of these platforms. They keep data such as demographic data of their users and activities within the courses in their own data warehouses. They want to support student development by using this data. For this, all the data of a cyber security course which is presented by researchers from Newcastle University's School of Computing Science were shared. There are various instructional sections such as videos, articles, quizzes in this course. Since the number of articles in the course is higher than the others, I thought it would be more useful to do a study on the article. Therefore, in this study, how the reading time of the articles changes according to learners will be analyzed and a model will be developed that can predict the average article reading time for each individual by using this data. The important thing for science is the **reproducibility** of the study. In order for the study to be a science, tools and methods were used to provide us with reproducibility.

Materials and Methods

It is known that data mining process includes various complex steps and requires many tools, skills. To overcome the challenge, it needs a structured methodology , an effective project management and collaborative working method.

CRISP-DM Methodology

CRISP-DM is a structured process model that provides practitioners with a complete blueprint for the project. The purpose of CRISP-DM is to make data mining projects low cost, reliable, repeatable, manageable [1]. It is broken down six phases;

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

CRISP-DM is not a linear process, it is a life cycle. It is possible to take a step back and apply multiple iterations to achieve business goals. I think it is the right approach for the methodology to focus on the business problem and needs rather than starting the analysis at the beginning. Understanding the customer's views and needs is crucial to the progress of a project.

On the other hand, I think there will be some coordination problems when applying CRISP-DM methodology within large teams. There may not be enough efficiency when assigning tasks to teams. Therefore, in this study , CRISP-DM methodology was combined with Kanban.

Project Template

Scientific results only become stronger when they can be reproduced by the researcher [2]. These studies should be in a certain order and well organized. Project template is a really powerful tool that help us for systematizing how we organize on our project. Installation and usage is really simple and not complicated.

By using this library in our project, it both accelerated us and enabled us to be organized. Even the auto-load data feature can speed up a data scientist's analysis. The project structure created was quite suitable for a data science project. But I think it needs to be developed a little more. For example, if I need to save the model outputs to a file, we need to put it in the data folder. Thus, input and output data can be mixed in a file. In general, I would like to say that it is really efficient when viewed as a design pattern.

Git

Managing our data ensures that you can always find our data and ensure the quality of scientific practice. One of the tools that allows us to do this is Git. It is an open source version control system that will speed us up as we develop our projects, large or small, and help us maximize efficiency. The versioning can improve the reproducibility of our scientific analyses. This study was carried out using Github which is a web-based storage service for projects that use Git as a version control system.

Feature branch workflow was chosen instead of classical branching models. Branches are created on the based on tasks and features. Instead of committing directly on their main or dev branch, a new feature needs to be created for a new feature or task. Feature branch name should be unique and clear like feature-1234. Then, a merge request is opened for the feature branch to be review and merge to the develop branch. After reviewing, the feature branch is merged into the dev branch not into master. When dev branch reaches a stable point and is ready to be release, the dev branch is merged into the master branch. Thus, master branch will never include broken code. The disadvantage of this is that it can cause confusion and conflicts when there are too many merge requests.

GitHub help us organize and prioritize our work. Github boards are used for planning and organization. Furthermore, GitHub provides makes it easy to automate all software workflows. It is called Github Action. 2 Github actions have been defined in this project. First, it runs tests within the project when pull requests are made to the dev branch. The second one, when merged to master branch, converts the report from Rmd file to html file and sends it to Firebase.

Conclusions

In conclusion, estimated reading time generally is calculated traditional ways which only use simple arithmetic and a single output is produced for everyone. However, the reading time of scientific articles varies according to each person. In this study, reading times were analyzed how the article reading time changed according to the individuals, and a model was produced that could predict for each person as a new approach. The CRISP-DM methodology which gives advantages to practitioners was implemented. It starts with business understanding and it is a life cycle. As in the study, some steps were taken back to the previous step. It definitely gives flexibility and reproducibility. On the side, using the Git tool has made a great contribution to making the project repeatable and reproducible.

References

1. Wirth, R., & Hipp, J. (n.d.). CRISP-DM: Towards a Standard Process Model for Data Mining. http://www.cs.unibo.it/~montesi/CBD/Beatriz/10.1.1.198.5133.pdf
2. Overview of Reproducible Research — The Turing Way. (2020). Netlify.app. https://the-turing-way.netlify.app/reproducible-research/overview.html