Capstone Project – Diagnostic Questions

Data Science Nanodegree

Audris Ločmelis

July 26th, 2020

Project Definition

Project Overview

This capstone project is a first attempt at predicting answers to diagnostic questions in a recently released machine learning competition Diagnostic Questions: The NeurIPS 2020 Education Challenge[1].

> *Digital technologies are becoming increasingly prevalent in education, enabling personalized, high quality education resources to be accessible by students across the world. Importantly, among these resources are diagnostic questions: the answers that the students give to these questions reveal key information about the specific nature of misconceptions that the students may hold. Analyzing the massive quantities of data stemming from students' interactions with these diagnostic questions can help us more accurately understand the students' learning status and thus allow us to automate learning curriculum recommendations.*

https://1drv.ms/b/s!AhaMcLyAxjaYiSYKKo43BDfRIzHr / https://arxiv.org/abs/2007.12061

Student provides a high-level overview of the project. Background information such as the problem domain, the project origin, and related data sets or input data is provided.

Problem Statement

The problem which needs to be solved is clearly defined. A strategy for solving the problem, including discussion of the expected solution, has been made.

Metrics

Metrics used to measure performance of a model or result are clearly defined. Metrics are justified based on the characteristics of the problem.

Analysis

Data Exploration

Features and calculated statistics relevant to the problem have been reported and discussed related to the dataset, and a thorough description of the input space or input data has been made.

[1] https://neurips.cc/Conferences/2020/CompetitionTrack

Abnormalities or characteristics about the data or input that need to be addressed have been identified.

Data Visualization

Build data visualizations to further convey the information associated with your data exploration journey. Ensure that visualizations are appropriate for the data values you are plotting.

Methodology

Data Preprocessing

All preprocessing steps have been clearly documented. Abnormalities or characteristics about the data or input that needed to be addressed have been corrected. If no data preprocessing is necessary, it has been clearly justified.

Implementation

The process for which metrics, algorithms, and techniques were implemented with the given datasets or input data has been thoroughly documented. Complications that occurred during the coding process are discussed.

Refinement

The process of improving upon the algorithms and techniques used is clearly documented. Both the initial and final solutions are reported, along with intermediate solutions, if necessary.

Results

Model Evaluation and Validation

If a model is used, the following should hold: The final model’s qualities — such as parameters — are evaluated in detail. Some type of analysis is used to validate the robustness of the model’s solution.

Alternatively a student may choose to answer questions with data visualizations or other means that don't involve machine learning if a different approach best helps them address their question(s) of interest.

Justification

The final results are discussed in detail.

Exploration as to why some techniques worked better than others, or how improvements were made are documented.

Conclusion

Reflection

Student adequately summarizes the end-to-end problem solution and discusses one or two particular aspects of the project they found interesting or difficult.

Improvement

Discussion is made as to how at least one aspect of the implementation could be improved. Potential solutions resulting from these improvements are considered and compared/contrasted to the current solution.

Deliverables

####

Write-up or Application

If the student chooses to provide a blog post the following must hold: Project report follows a well-organized structure and would be readily understood by a technical audience. Each section is written in a clear, concise and specific manner. Few grammatical and spelling mistakes are present. All resources used to complete the project are cited and referenced.

GitHub Repository

Student must have a Github repository of their project. The repository must have a README.md file that communicates the libraries used, the motivation for the project, the files in the repository with a small description of each, a summary of the results of the analysis, and necessary acknowledgements. If the student submits a web app rather than a blog post, then the Project Definition, Analysis, and Conclusion should be included in the README file, or in their Jupyter Notebook. Students should not use another student's code to complete the project, but they may use other references on the web including StackOverflow and Kaggle to complete the project.

Best Practices

Code is formatted neatly with comments and uses DRY principles. A README file is provided that provides. PEP8 is used as a guideline for best coding practices. Best practices from software engineering and communication lessons are used to create a phenomenal end product that students can be proud to showcase!