Machine Learning Engineer Nanodegree

Capstone Proposal

Arwa Alwani
Febrary 10, 2019

Proposal

Domain Background

Universities around the world provide different courses for their students to educate them in pecific fields after their generic studies in schools. Universities nowadays are better and available in all countries. They are making big efforts to boost their students' education and performances by evaluating them in mny metrics and try to provide best solutions for them. This project information is collected by Ernest Fokoue and Necla Gunduz from students from Gazi University in Ankara (Turkey). This dataset is releated to the students evaluation of the courses and instructors in many means. I found this dataset interesting because it has many valuable information data that is interesting to be explored. around the world tries to boost their students attendance to boost their performance. Many universities have applied some restrictions to lessen students' abscence and other methods. There might be many factors related to the course and it's instructor that cause One of uncontrolled problems is students' absence of classes for a reason or another. The instuctor attitude, the course difficulty, unsufficiency of resource, how enjoyable is the course..etc are all factors that can affect the student attendace to the course.

Problem Statement

In order for universities to boost their students' performance, many universities try to boost students attendance to courses. Many reasons can affect students' attendance. some of the reasons that might make a student a course less than another one ight be The instuctor attitude, the course difficulty, unsufficiency of resource, how enjoyable the course is..etc. Thus, the target variable of this dataset is the attendance. The project will find the most factors the student point that affect their attendance to a course for the sake of universities to focus on improving the related aspects to these factors and enhance attendance of stuents to courses.

Datasets and Inputs

The dataset contains of 5820 instances and is collected from students from Gazi University in Ankara (Turkey). The dataset is numerical. It contains the

following features as described by the providers:

instr: Instructor's identifier; values taken from {1,2,3} class: Course code (descriptor); values taken from {1-13} repeat: Number of times the student is taking this course; values taken from {0,1,2,3,...} attendance: Code of the level of attendance; values from {0, 1, 2, 3, 4} difficulty: Level of difficulty of the course as perceived by the student; values taken from {1,2,3,4,5} Q1: The semester course content, teaching method and evaluation system were provided at the start. Q2: The course aims and objectives were clearly stated at the beginning of the period. Q3: The course was worth the amount of credit assigned to it. Q4: The course was taught according to the syllabus announced on the first day of class. Q5: The class discussions, homework assignments, applications and studies were satisfactory. Q6: The textbook and other courses resources were sufficient and up to date.
Q7: The course allowed field work, applications, laboratory, discussion and other studies. Q8: The quizzes, assignments, projects and exams contributed to helping the learning.
Q9: I greatly enjoyed the class and was eager to actively participate during the lectures. Q10: My initial expectations about the course were met at the end of the period or year. Q11: The course was relevant and beneficial to my professional development. Q12: The course helped me look at life and the world with a new perspective. Q13: The Instructor's knowledge was relevant and up to date. Q14: The Instructor came prepared for classes. Q15: The Instructor taught in accordance with the announced lesson plan. Q16: The Instructor was committed to the course and was understandable. Q17: The Instructor arrived on time for classes. Q18: The Instructor has a smooth and easy to follow delivery/speech. Q19: The Instructor made effective use of class hours. Q20: The Instructor explained the course and was eager to be helpful to students. Q21: The Instructor demonstrated a positive approach to students. Q22: The Instructor was open and respectful of the views of students about the course. Q23: The Instructor encouraged participation in the course. Q24: The Instructor gave relevant homework assignments/projects, and helped/guided students. Q25: The Instructor responded to questions about the course inside and outside of the course. Q26: The Instructor's evaluation system (midterm and final questions, projects, assignments, etc.) effectively measured the course objectives. Q27: The Instructor provided solutions to exams and discussed them with students. Q28: The Instructor treated all students in a right and objective manner.

Q1-Q28 are all Likert-type, meaning that the values are taken from {1,2,3,4,5}

The dataset is covering many important aspects that can be directly relevant to students' attendance. All the features are going to be used. an F-score will be used in case the target variable is not normally distibuted.

Solution Statement

Finding the most relevant reasons for students abscence of a course more than another could help univrsities better understand the problem. After the problem is understood better, this can help in improving courses and help in decision making of courses' changes in order to not make them reflect negatively on students' attendance and performance. The Gradient Boosting classifier model will be used in order to classify the data.

In this section, clearly describe a solution to the problem. The solution should be applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, describe the solution thoroughly such that it is clear that the solution is quantifiable (the solution can be expressed in mathematical or logical terms) , measurable (the solution can be measured by some metric and clearly observed), and replicable (the solution can be reproduced and occurs more than once).

Benchmark Model

The benchmark model will be the Decision Tree modl which will be applied as well to the same data. The result of the Gradient Boosting model(the primary model) will be compared with the Decision Tree model in terms of time consumed by every model for training, accuracy score and F-score scored by both algorithms as well.

Evaluation Metrics

The evaluation metrics will be used to evaluate the Gradient Boosting model(the primary model) and Decesion Tree model(the benchmark model) are three metrics. The first metric is time consumption be ech model. The second metric is the accuracy score scored by each metric. The third metric will be the F-score scored by each metric. These metrics will be the tools to evaluate the both models.

Project Design

(approx. 1 page)

The project architecture will be as the following:

1- Data Preprocessing:

- Download the Turkiye Student Evaluation dataset
- Explore data

2- Develop Gradient Boosting Model:

- Create model
- Fit the model
- Prediction
- Calculate performance
- Plot graphs
- Find most affecting features

3- Develop Decision Tree Model:

- Create model
- Fit themodel
- Prediction
- Calculate performance
- Plot graphs
- Find most affecting features

4- Conclusion

- Comparison of both models

Resources:

https://archive.ics.uci.edu/ml/datasets/Turkiye+Student+Evaluation#