Education Analytics with R and Cortana Intelligent Suite

By Fang Zhou, Microsoft Data Scientist; Hong Ooi, Microsoft Senior Data Scientist; Graham Williams, Microsoft Director of Data Science

**Overview**

Education is a relatively late adopter of predictive analytics or machine learning as a management tool. Nowadays, a desire for stronger operations is leading universities or government to perform student predictive analysis, which helps in better-informed and faster decision making.

Student predictive analytics aims to solve two key problems:

1. Identify students more likely to perform better academically, which helps in optimizing support.

2. Foresee students at high risk of dropping out thus timely preventing attrition.

Education systems face an enormous diversity across regions and countries. The case studies in Asia show a novel and unique landscape for machine learning in the education world.

• An Australian Education Department

Built a mixed effects regression model to measure the influence of student characteristics and predict students’ NAPLAN test scores in the presence of variation in students and schools, using R with integration into Azure ML and Power BI for production.

• Indian State Government

Leveraged a ML model, two-class boosted decision tree, to predict the likelihood of a student dropping out, moving next to build an E2E pipeline using Azure services, including Azure SQL Database, Azure ML, and Azure Data Factory, etc.

**An Australian Education Department**

****

NAPLAN is a standardized testing system for all schools in Australia to assess students’ basic skills - reading, writing, grammar, spelling and numeracy. Majority of students take the 5 tests in years 3, 5, 7, and 9. An Australia Education Institute wanted to identify talent based on each NAPLAN test scores and set individual targets across school cohorts.

The data was collected from 83,000 students across almost 140 schools in one of the major Australian cities. The data included information about yearly NAPLAN test history, student demographics, school records and school attributes.

We addressed the problem as a regression problem, but taking random effects for student and school into account. We fit a mixed effects regression model for each NAPLAN test score.

Our model was implemented using the function ‘lmer’ from the R package ‘lme4’.

With such mixed-effects regression model, we can measure the influence of the fixed effects in the presence of variation in students and schools, as well as fairly assess the quality of a student or a school while taking other factors into account.

• Two students or two schools with the same characteristics might perform totally different in NAPLAN test

• A school or a student with poor/good NAPLAN scores starts looking better/worse when considering their conditions, like indigenous, parent occupation, disability etc.

Benefiting from our cloud solution with customized R model, this Australian Education Department could now easily gain insights on various aspects, whether it is to detect the trend of the student scores over year, to identify key factors affecting academic achievements (e.g., indigenous, parent occupation), to compare the quality of education across schools, or to discover talents.

**Indian State Government**



An Indian State Government is embarking on a mission to reduce the number of school dropouts and increase the overall skill levels of the citizens of the state to increase the human capital.

To achieve this objective, we proposed a cloud solution to make data-driven intelligent inferences by processing complex data on education. The data includes data on student performance, school infrastructure, teacher skills, and data from external sources like NGOs and government agencies working in education space.

This solution involves building and deploying machine learning models for binary classification that can predict the likelihood of a student dropping out and other education outcomes at school, district and state-levels.

Our models were implemented by leveraging Azure ML modules as well as R and achieved good model performance metrics, where Accuracy=0.888; Precision=0.941; Recall=0.618; F1 Score=0.746; and AUC=0.891. Such a high accuracy in predicting student drop-out helps the government take proactive actions to generate effective and targeted strategy in reducing student attrition.

Microsoft machine learning empowered cloud solution is helping make education equal to every child who has a dream and a future.

**Data Science Design Pattern for Education Analytics**

Based on our experiences with these use cases we developed an initial version of the Data Science Design Pattern for Education Analytics. This includes both Student Score Modeling and Student Drop-Out Prediction. This pattern provides a starting point for the data scientist exploring a new dataset in the education world by using R. By no means is it the endpoint of the data science journey. The pattern is under regular revision and improvement and is provided as is. To try out this pattern please download its R Markdown and Jupyter Notebook files from this github repository [EduDP](https://github.com/ZhouFang928/EduDP).