Education Analytics with R and Cortana Intelligence Suite

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

Education is a relatively late adopter of predictive analytics and machine learning as a management tool. A keen desire for improving educational outcomes for society is now leading universities and governments to perform **student predictive analytics** to provide better-informed and timely decision making.

Student predictive analytics often aims to solve two key problems:

1. Predict student academic outcomes so as to better target support.

2. Predict students at risk of dropping out so as to prevent attrition.

Education systems face enormous diversity across regions and countries. Two case studies demonstrate the novel and unique landscape for machine learning in the education world.

• A mixed effects regression model has been developed in conjunction with an Australian education department to measure the influence of student characteristics and to predict student test scores in the presence of variation across students and schools. The model was implemented using R and then integrated with Azure Machine Learning for deployment to production through Power BI.

• A predictive model for student drop out has been developed in conjunction with an Indian state government using machine learning two-class boosted decision trees. For deployment an end-to-end pipeline was built using Azure services including Azure SQL Database, Azure ML and Azure Data Factory

Microsoft Data Scientists assisted with the analysis in both cases and we present details below with R code provided in a git repository to replicate the modelling on artificial data.

Student Performance

NAPLAN is a standardized testing system used by all schools in Australia to assess students' basic skills—reading, writing, grammar, spelling and numeracy. A majority of students take the five tests in years 3, 5, 7, and 9. A goal of this use case is to identify talent based on NAPLAN test scores and to set individual targets across school cohorts.

The data were collected from 83,000 students across almost 140 schools in a major city. The data included information about yearly NAPLAN testing, student demographics, school records and school attributes.

We addressed the task as a regression problem, taking random effects for student and school into account. The `lmer` function from the R package lme4 was used to fit a mixed effects regression model to the NAPLAN test score data.

With this 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. It is observed that students/schools with very similar characteristics can perform quite differently in NAPLAN tests. Also, a school/student with poor or with good NAPLAN scores can be characterized by combinations of variables exposed in the data.

The model was deployed into a cloud solution exposing this customized R model. Through the interface the education administrators can now easily gain insight into student performance. For example, trends can be detected for student scores over multiple years, key factors affecting academic achievement become apparent, comparative quality of education across schools can be explored, and talent can be identified and shared.

Student Drop-Out

Many governments have a focus on reducing the number of school dropouts and thereby increase the overall skill levels of the citizens so as to increase the human capital. This is certainly the case in Andhra Pradesh and other states in India.

To achieve this objective complex data covering student performance, socio-economic situation, school infrastructure, and teacher skills is combined with external sources from NGOs and government agencies working in education.

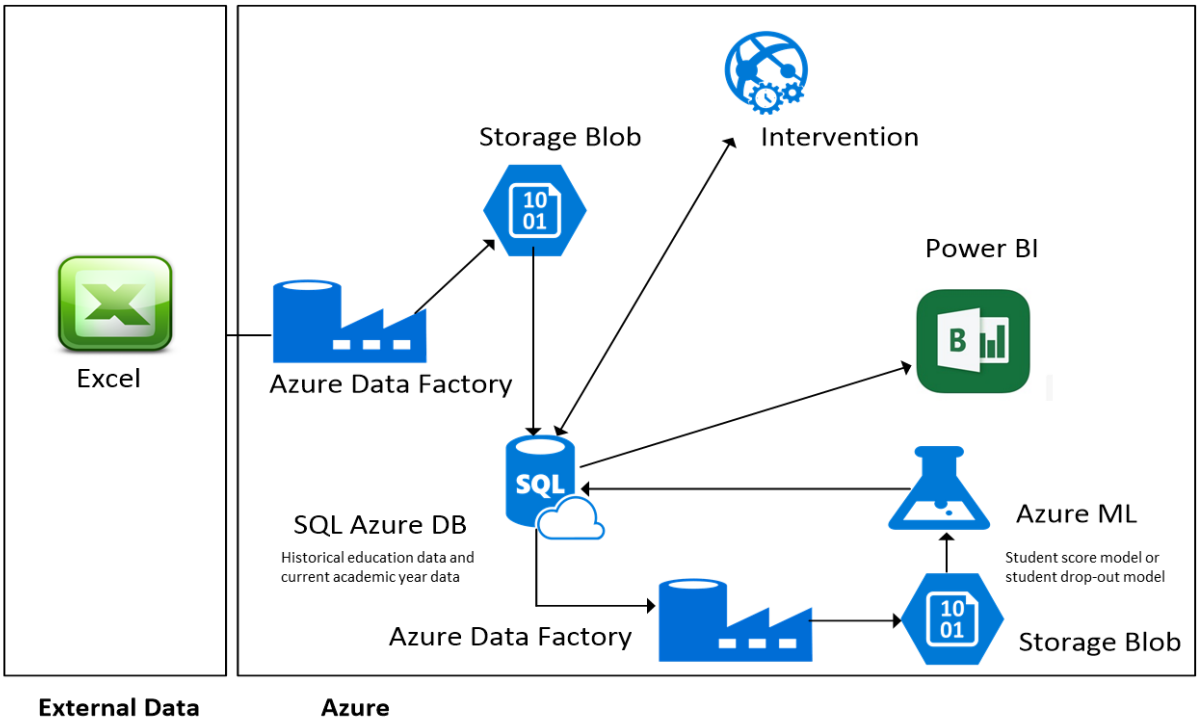
Microsoft's solution has involved building and deploying machine learning models for binary classification that can predict the likelihood of a student dropping out, in addition to other educational outcomes at the school, district and state-levels.

R-based models using the latest advances in boosted decision trees were implemented within Azure Machine Learning Studio to achieve model performances with accuracy of 89%, precision of 94%, recall of 62%, F1 score of 75% and an AUC of 89%. With such high accuracy for predicting student drop-out governments are taking proactive measures to generate effective and targeted strategies for reducing student attrition. Non-academic characteristics often external to the school are also identified as playing significant roles in drop-out rates and support a focus for strategies which ensure a solid educational base for the future prosperity of the community.

Sample Solution Architecture

The student predictive analytics solution runs in the Azure cloud and leverages Cortana Intelligence Suite components. Prior to using Azure for insights into student performance or drop-out, the education data has been acquired by data integration components in a prescribed format, transformed, merged and cleansed. Azure SQL Database is used to support storage of both academic year data as well as

historic data. We then leverage an Azure Machine Learning model (with R customization) trained on historic data to predict a student's NAPLAN test score or whether a student in current academic year is likely to drop out. An Azure Data Factory pipeline is developed and deployed to automatically drive the gathering of data, transforming data into a format suitable for Azure ML model, and loading the processed data (with prediction results) back to the target Azure SQL Database for reporting by Power BI dashboard.



The student predictive analytics solution we've shown here demonstrates how to extend the capabilities of R with the Cortana Intelligence Suite by integrating custom R code with Azure ML studio, to solve problems in the education world. For a quick guide on how to use R in Azure ML studio, see these document and online tutorial. For additional examples of Cortana Intelligence based solutions, see the Cortana Intelligence Gallery.

Data Science Design Pattern for Education Analytics

Based on our experience with these use cases and indeed from future cases we learn of or are involved in we have developed and maintain the Data Science Design Pattern for Education Analytics. This includes implementations of both Student Score Modeling and Student Drop-Out Prediction. This pattern provides a starting point for a data scientist exploring a new dataset in the education world using R. The GitHub repository includes a sample dataset and R scripts to build the models described above. Data Scientists working in the Education domain can replicate this modelling approach using their own internal datasets.

By no means is this 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 the provided R Markdown

and Jupyter Notebook files. We welcome feedback on this pattern, and you are welcome to comment and contribute at the GitHub repository linked below.

Fang Zhou (GitHub): Data Science Design Pattern for Educational Analytics