Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2014:11**
http://www.math.su.se/matstat

Prediktion av studieframgång inom kursen Matematik I på Stockholms universitet

Jonathan Pollack*

September 2014

Sammanfattning

I denna uppsats använder vi oss av data insamlad under höstterminen 2013 på Stockholms universitet i syfte att prediktera studieframgång och genomströmning i kursen Matematik I. Vi utvecklar prediktiva modeller med hjälp av linjär och logistisk regression, naiv Bayesiansk klassifikation samt klassifikationsträd. Modellerna jämförs sedan och analyseras, särskilt utifrån deras AUC, Brier score och precision. Vi anpassar även modeller på äldre och yngre studenter separat. Av metoderna som används finner vi att logistisk regression och klassifikationsträd är särskilt bra lämpade, och att den prediktiva styrkan allmänt är högre för yngre studenter i jämförelse med äldre. Vi jämför även våra modeller med den allmänt tillämpade urvalsmodellen där man går enbart efter gymnasiebetyg och finner att även denna metod har en godtagbar prediktionsförmåga, särskilt för kursmomentet *analys*.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: jhpollack@yahoo.com. Handledare: Martin Sköld.

Abstract

In this paper we use data collected from students studying the first-year math course 'Matematik I' at Stockholm University in the fall of 2013, with the aim of predicting first-year retention rates. Predictive models are built and developed using linear and logistic regression, naive Bayes classification and decision trees. The models are compared and analyzed, especially in terms of their AUC, Brier score and precision. We also fit models on older and younger students separately. Logistic regression and classification trees are found to perform especially well on the dataset, and the models developed on younger students are found to have higher predictive strength compared to those on the older students. We also compare our models to the common selection procedure in higher education, in which only grades from secondary education are used to predict student success rate. We find that this model too makes predictions at an acceptable level, especially for the *calculus* component of the course.