We apply a method of applying semi-supervised learning to data from an online test preparation site in order to predict students' GMAT scores. We leverage a large amount of unlabeled data by using a variational autoencoder to extract feature embeddings for each student. We compare these feature embeddings with hand-engineered features on a linear regression model and show that the feature embeddings perform comparably. We also show that combining the feature embeddings with the hand engineered features increases the performance of the model with hand engineered features alone.

Dataset
We use a dataset from TAL Education Group, an online education company that provides online test prep for various exams including the GMAT. The original dataset consists of 6,681 questions with category labels (verbal, quantitative, integrated reasoning); 90,831 students; 1,600,324 student-question interactions consisting of a student, a question, the time the question was attempted and whether or not the student got the question correct; and 4,581 students labeled with self-reported GMAT scores. Out of the 458 students with self-reported GMAT scores, 372 used the test prep system before the date of their reported exam. Out of these 372 students, we consider the 354 students who attempted at least 50 questions. Similar to out of the 90,831 overall students, we only consider the 19,941 students that have completed at least 50 questions. The overall number of questions that these students attempted is 7,808.

References