Predicting Student Performance from Admissions and Historical Performance Data

Alexander Panfilov
Matrikelnummer 5990087
alexander.panfilov@student.uni-tuebingen.de

Evgenii Kortukov
Matrikelnummer 5994382
evgenii.kortukov@student.uni-tuebingen.de

Abstract

We use a private dataset of anonymized data from students of one Russian University to assess how well student performance can be predicted from this data. The dataset contains admission data (state exam results, subject competition prizes, and sociodemographic information) and data of students' university-level performance. We conduct statistical testing, apply linear regression and dimensionality reduction techniques to provide a proof of concept of a student embedding.

1 Introduction

With the advent of MOOCs and personalized learning solutions universities have to adapt to maintain a competitive advantage and attract potential students. A promising innovation would be to implement data-driven personalized study plans tailored to the needs of every student. Such an approach has to rely on a sufficiently informative and predictive model of a student. Our main contribution is to show that available data is sufficient to construct embeddings with predictive power and provide examples of possible predictions. Code is available at the https://github.com/kortukov/learning_analytics.

2 Data

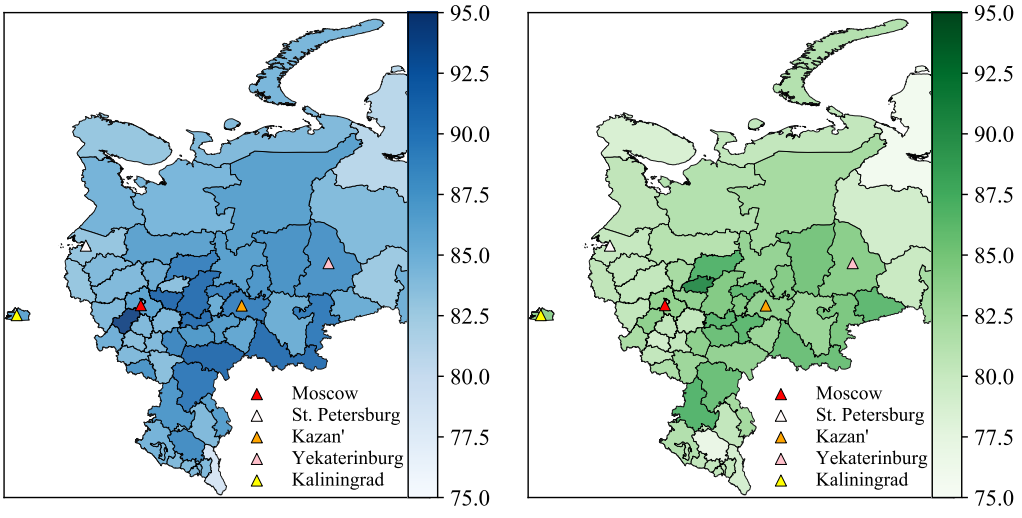
We perform our analysis on a dataset that contains information on approximately 20000 bachelor students of one Russian University. It contains a set of anonymized features which can describe a particular student. Students contributed to the dataset voluntarily and gave written consent for using their data for research purposes. Data consists of two main parts.

Admission data is the information that University knows about a student at the time of application process. In order to apply to University candidates must possess a school-leaving certificate and provide at least one of the following things:

State exam results: Each program offers a limited number of study places without a tuition fee on a competitive basis. Students with scores below the passing threshold can still be admitted, but on a tuition fee basis; School olympiads results: Prize holders of subject olympiads can be admitted to Universities without state exam results; In-University exam results: In special cases University can establish its own inner exams. These exams are commonly taken by foreign applicants who cannot provide state exam results.

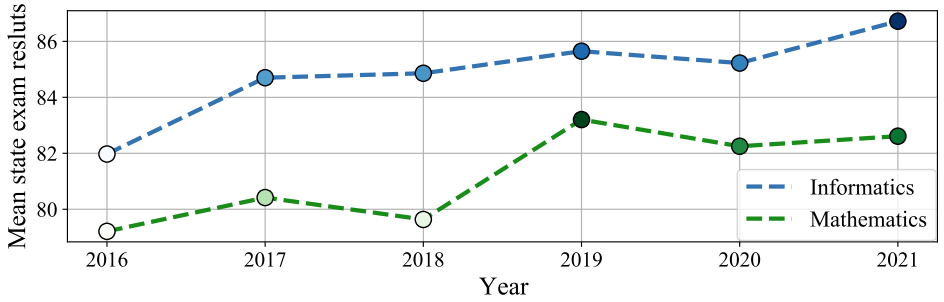
Thus, admission data contains exam results or olympiads, chosen faculty and program, fact of paying the tuition fee, country of origin and home region.

Project Report for *Data Literacy* 2021/22



(a) Mean informatics state exam results

(b) Mean mathematics state exam results



(c) Mean admitted state exam results over years

Figure 1: Geographical distribution of the state exam results in European part of Russia and trend for the last 5 years.

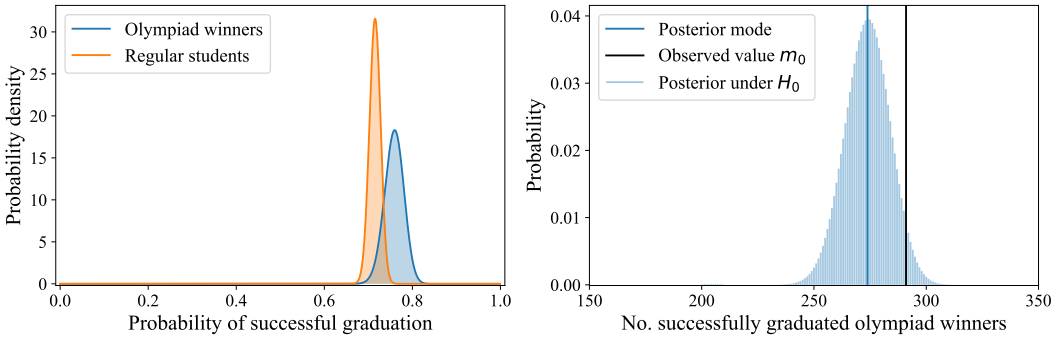
Universities employ this admission data to obtain the most talented students. Before the application period professors visit schools which usually have high exam results or prominent amount of olympiad winners and agitate schoolchildren to apply to their University. This recruitment process can be optimized with the help of data by recruiting from regions with high state exam result: in Figure 1a and Figure 1b we can see a heatmap of mean state exams results of the European part of Russia. Surprisingly the highest scores are not observed in the St. Peterburg's or Moscow's metropolitan areas, as one might expect.

Per-semester student performance data contains following features: GPA, number of failed exams, student group and faculty (which student may change during studies). Also, this data provides some possible target-features: fact of the successful graduation or fact of obtaining *summa cum laude*.

The dataset posed two main difficulties: large number of missing values and categorical features, which are problematic for linear models.

3 Hypothesis testing

A commonly held belief is that one specific feature separates highly performing students - olympiad participation. University admission offices provide benefits to olympiad winners in the admission process. We seek to justify this approach with data and also demonstrate that our student embeddings contain highly informative features. The metric of university-level performance we use is successful graduation from a study program. We test the null-hypothesis that olympiad winners have the same probability of successfully graduating from university as regularly admitted students.



(a) Posterior distributions of successful graduation probability π for each group of students.

(b) Posterior distribution of successfully graduated student count under the null hypothesis H_0 and the observed count in the olympiad winners group

Figure 2: Hypothesis testing results.

Our dataset includes 13109 bachelor students who either graduated or dropped out. Unfortunately, the exact admission type is known only for a subset of students who are admitted after 2016. There are **383** olympiad winners in this subset and **1275** regularly admitted students. We model each group of students as a series of independent identical Bernoulli experiments with unknown success probability π. To estimate this quantity we use Bayesian inference.

Let X_o be the set of $n_o = 383$ datapoints of olympiad winners with $m_o = 291$ successful graduates. Then X_r - set of $n_r = 1275$ regularly admitted students with $m_r = 912$ successful graduates. We want to estimate the probabilities of graduation in both groups π_o and π_r. Using binomial likelihood we get that the posterior distribution over π is then Beta distribution $\mathcal{B}(\pi; m+1, n-m+1)$. This leads us to a *maximum a posteriori* estimate of $\hat{\pi_o} = 0.760$ and $\hat{\pi_r} = 0.715$. Figure 2a shows the posterior distributions. Under our model we can say that with 95% probability π_o lies in the interval $[0.714; 0.800]$ and π_r lies in the interval $[0.690; 0.739]$.

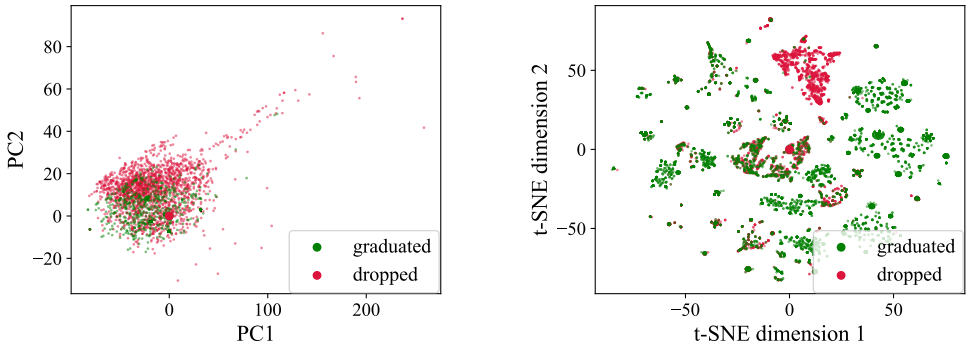
Now we test the hypothesis $\mathcal{H}_0$: *"Olympiad winners have the same probability of successful graduation as students admitted under the regular procedure."*. Note, that: 1. Under the null hypothesis $\mathcal{H}_0$ we have $\pi_r = \pi_o = \pi$. The likelihood of observing m_o successfully graduated olympiad winners given a particular probability $p(m_o|\pi)$ follows a binomial distribution. 2. The previously achieved Beta-posterior over π now acts as a prior, so $p(\pi|m_r, n_r) = \mathcal{B}(\pi; m_r+1, (n_r-m_r)+1)$. 3. To find the probability of observing m_o successfully graduated olympiad winners under the null hypothesis we marginalise over π: $p(m_o|\mathcal{H}_0) = \int p(m_o|\pi)p(\pi|m_r, n_r)\, d\pi$

This results in Beta-binomial distribution[1]. It is depicted in Figure 2b. We compute the probability of seeing m_o or more successfully graduated olympiad winners under the null hypothesis and achieve the p-value of $0.037 < 0.05$ which leads us to reject the null hypothesis.

We conclude that olympiad winners and regularly admitted students have statistically significant differences in graduation probability. This warrants further study of differences in performance between students groups.

4 Dimensionality reduction

In the previous Section 3 it was shown that, there are significant differences between some groups of students. We assume that if embedding is representative enough, we may visually distinguish graduated and dropped out groups of students in the 2D projection of the embedding. For this purpose we used PCA and t-SNE [3] algorithms. Although the groups are indistinguishable in the subspace of the first two principal components (Figure 3a), some structure can clearly be observed when using t-SNE algorithm (Figure 3b). There is a top red cluster with most of the dropped out students, whereas the right green clusters contain mainly graduated students. Other clusters in picture are more mixed with the prevalence of graduated students.



(a) PCA (b) t-SNE

Figure 3: Dimensionality reduction visualisation.

5 Regression

In this section we show how student performance can be predicted from admission data. Features described in 2 are independent variables, the response variable is average grade in the first semester. We standardize the data and apply linear regression with L1 regularization [2]. The training set is all data before before predicted year. The obtained results can be seen in Table 1. Grades range from 2.0 to 5.0. Prediction errors might be due to instability of exam results between years 1c and large number of missing values in the data 2.

Table 1: RMSE for predicting grades from admission data. Grades range from 2.0 to 5.0

Predicted year	2015	2016	2017
RMSE	0.520	0.581	0.589

6 Ethical considerations

Conducting an analysis of this kind raises a number of ethical questions. First, it poses privacy concerns. Participation must be fully voluntary and students should be provided same opportunities regardless of whether they gave out their data. Second, human supervision is needed to make sure that the use of resulting predictions is beneficial to students. Lastly, it is vital to make sure that the achieved model does not put groups of people at a disadvantage. We believe that given a careful data usage strategy, using such a model can be rewarding to both universities and students.

7 Conclusion

We have shown that it is possible to create a predictive student model from admission and semester performance data. Created model captures differences between student groups, contain discriminative features and can be used to predict future performance of a student.

References

[1] Philipp Hennig and Lukas Tatzel. Exercise sheet 5. *Data Literacy*, WS, 2021.

[2] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[3] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.