Predicting and Understanding what leads to High Knowledge Levels

Adam Walmsley, Morgan Dean, Tracy Wang, Yuexiang Ni

Declaration

Our research project utilized the DSCI100 project (Anthony (2022)).

It was originally completed by one of our members, with consent from all his teammates.

Summary

In this study we look at how exam performance and time spent studying affect user knowledge on a specific application area. We see if we can build a regression model to accurately predict this user knowledge, and which if any of the two features are more important in building an accurate model.

Introduction

Understanding how study time and exam performance contribute to a student's knowledge level is crucial in educational research (Timbers (2022)). In this study, we analyze data from the User Knowledge Modeling Dataset retrieved from the UCI machine learning repository database with the CC BY 4.0 license (Kahraman (2013)). It was collected from undergraduate students in the Faculty of Technology at Gazi University. The dataset, available from the UCI Machine Learning Repository, was originally developed to assess intuitive knowledge classifiers using study behaviors and performance metrics (Kahraman (2012)).

The dataset consists of six key variables:
- STG: Degree of study time for goal object materials
- SCG: Degree of repetition for goal object materials
- STR: Degree of study time for related objects with goal object
- LPR: Exam performance of the user for related objects with goal object
- PEG: Exam performance of the user for goal objects
- UNS: Knowledge level of the user (categorical target variable)

For this study, we focus on STG (study time for goal object materials)** and PEG (exam performance for goal objects)** as primary predictors of student knowledge, represented by the UNS variable, which is classified into four categories: Very Low, Low, Middle, and High. Our objective is to determine whether time spent studying or actual exam performance is a stronger indicator of student knowledge. In other words, from this data we can use study time, and exam performance to predict our own user's knowledge level and test which is a better determining factor to knowledge level, study time or exam performance. We decided to use these two variables to determine user knowledge because we were curious on what mattered most for students' knowledge level, study time or how they performed on the exam. This will also in turn let us understand if exams taken place actually evaluate what they are supposed to. If proven we will be able to implement our model onto similar evaluations at any university to determine if exams are in fact relevant to real-world learnings.

Question:

Which habit/result is more indicative of a student's knowledge level: the time they spent studying or their actual exam results?

Methods and Results

Preliminary exploratory data analysis:

The dataset being used is User Knowledge Modeling Data Set and includes factors that measured the study time of more and less specific information for an exam as well as study repetition of this material. Exam performance on specific and less specific information was also measured. These factors were recorded on a scale from 0.00 to 1.00. These observations were then used to determine the target value of the knowledge level which was categorized with very low, low, middle and high.

To begin answering our question, we began by loading the required packages.

Since the data set was an excel file and came from the web, we used the `download.file` function to download and convert the data from its original format to something we could work with and analyze. The excel also contained several sheets and was already split up into training and test data, which we used to create our training and test variables. The word format was inconsistent between sheets; the training data's user knowledge level was typed in lowercase and with underscores, and the test data's user knowledge level was typed in regular case and with spaces. So to ensure consistency between data, we converted the "Very Low" value into "very_low." The data was already standardized and so this step was not required here.

With our data now loaded in, we select the specific columns (which include our predictors and classifier variables) from the data set. These columns include the User Knowledge Level

(UNS, the classification variable), Study Time for Goal Object Materials (STG, a predictor), and Exam Performance for Goal Object (PEG, another predictor). Because the UNS column contains the classification groups, we mutated this from a string to an ordinal factor data type. We loaded the first couple of rows from each table to observe these changes. (See Table 1 and Table 2 below)

Table 1: First six rows of the Training Data

STG	PEG	UNS
0.00	0.00	very_low
0.08	0.90	High
0.06	0.33	Low
0.10	0.30	Middle
0.08	0.24	Low
0.09	0.66	Middle

Table 2: First six rows of the Testing Data

STG	PEG	UNS
0.00	0.05	very_low
0.05	0.14	Low
0.08	0.85	High
0.20	0.85	High
0.22	0.90	High
0.14	0.30	Low

Next, we wanted to see the distribution of knowledge levels within the training and testing data, so we grouped the data by the knowledge level, found their counts, and then calculated the percent of each class. From Table 3, we can see that 9% of users have a very low knowledge level, 24% have a high knowledge level, 32% have a low knowledge level, and 34% have a middle knowledge level. Table 4 (testing data) on the other hand, shows that 18% of users have a very low knowledge level, 27% have a high knowledge level, 32% have a low knowledge level, and 23% have a middle knowledge level. We will have to keep these differences in mind during the randomization process and as we conduct further analysis.

Table 3: Count and Percent of User Knowledge Levels in the training data.

UNS	count	percentage
very_low	24	9.302326
Low	83	32.170543

Table 3: Count and Percent of User Knowledge Levels in the training data.

UNS	count	percentage
Middle	88	34.108527
High	63	24.418605

Table 4: Count and Percent of User Knowledge Levels in the testing data.

UNS	count	percentage
very_low	26	17.93103
Low	46	31.72414
Middle	34	23.44828
High	39	26.89655

We still wanted to understand our data a bit more. To do so, we wrangled the training data a bit more to generate a table of the mean STG and PEG and the minimum and maximum values of the predictors for each knowledge level. This helped us get a sense of the boundaries for each class and the variance within each predictor (See Table 5 below).

Table 5: Means, Minimums, and Maximums of Selected Variables in Training Data

UNS	count	mean_STG	mean_PEG	max_STG	max_PEG	min_STG	min_PEG
very_low	24	0.3057917	0.0908333	0.68	0.24	0.00	0.00
Low	83	0.3211446	0.2376265	0.73	0.35	0.02	0.01
Middle	88	0.3999773	0.5423864	0.80	0.83	0.06	0.25
High	63	0.4216508	0.7725397	0.99	0.93	0.00	0.47

All these tables helped us understand the data but still required attentive interpretation. So we now created a visualization of the User Knowledge Level distribution based on the Exam Performance and Study Time variables. The plot below shows how the knowledge levels are stacked vertically like layers, where study time can vary from 0 to 1 for each class, but the exam performance imposes somewhat of a boundary on each category (See Figure 1 below).

Main Analysis and Results

In this analysis we chose to use the machine learning method of ordinal logistic regression. Since this is a multi classification problem we could not use simple binary logistic regression. We begin by fitting our model and looking at the coefficients on each feature.

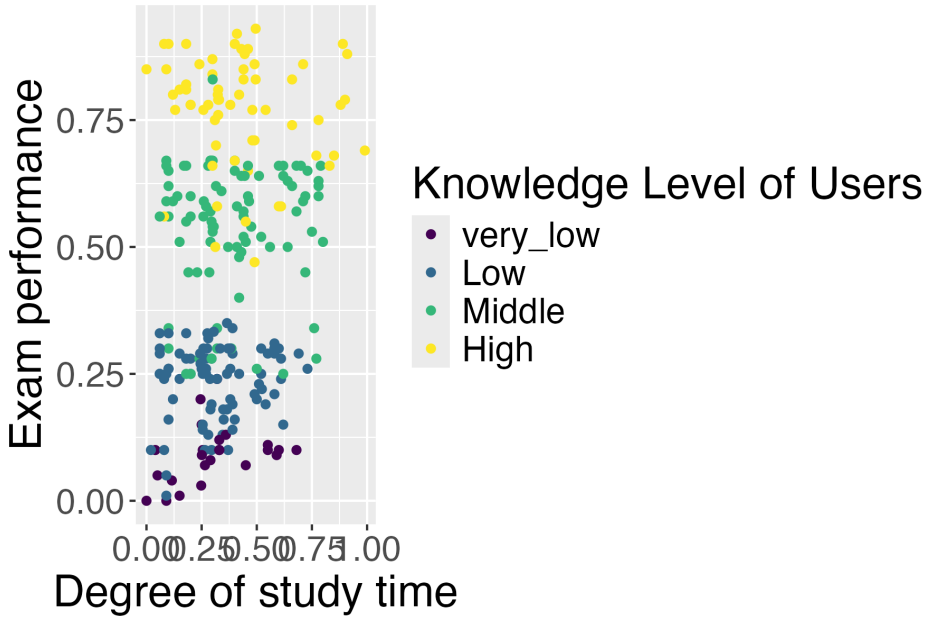

Figure 1: Visualization of the distribution of Knowledge Levels based on Exam Performance and Study Time.

Table 6: Summary across included features in our fitted model

Call:
polr(formula = UNS ~ STG + PEG, data = knowledge_train_data, Hess = TRUE)
Coefficients:
Value Std. Error t value
STG 0.4481 0.8315 0.5389
PEG 23.3257 2.4173 9.6495
Intercepts:
Value Std. Error t value
very_low|Low 2.6636 0.5232 5.0912
Low|Middle 8.3128 0.8372 9.9291
Middle|High 16.0228 1.6559 9.6762
Residual Deviance: 232.5871
AIC: 242.5871

From Table 6 we see that exam performance has a coefficient nearly fifty times the degree of study indicating the feature is much more indicative of user knowledge. Now we have to see if this model performs well on the test data

Table 7: Model Accuracy

[1] 0.8206897

Our model produces a mean accuracy of 82% which is promising (see Table 7). This means that really only using exam grade as a feature allows ordinal regression to achieve high accuracy across this complex problem.

Next we produce some visualizations to get a better sense of the data.

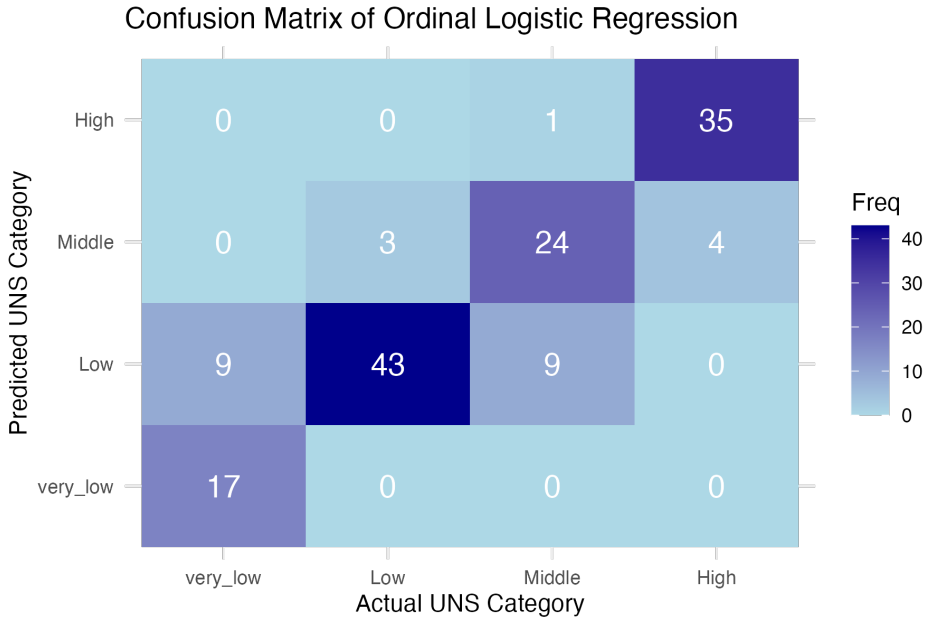

Figure 2: Confusion Matrix

This confusion matrix (Figure 2) tells us where our model is predicting correctly and incorrectly, as well as the frequency to which it does that. We can see that across all possible classes it is fairly successful with some slight confusion being introduced when comparing the 'low' and 'very low' classes.

In Figure 3 we again look at how much importance our model gave to each of the features we included. As seen during fitting, exam grades is given the vast majority of the weighting and is practically the sole predictor for our model.

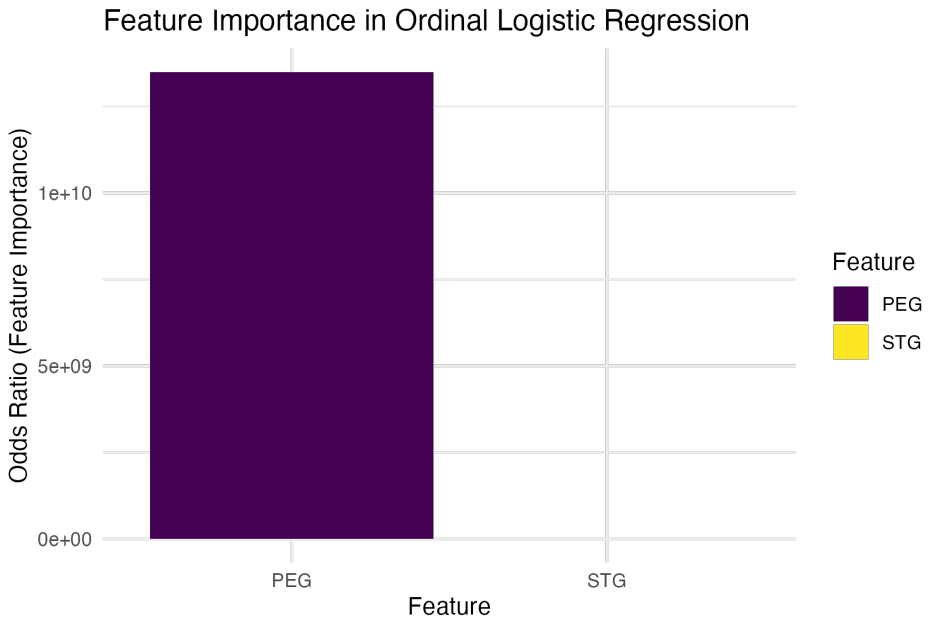

Figure 3: Feature Importance

Discussion

During this analysis we found a model that successfully predicts user knowledge separated by four distinct classes. We saw that when using degree of study time and exam performance as features, exam performance was by far the most useful predictor and dominated the model. This was an unexpected result in terms of the magnitude of importance. We thought that exam performance would probably be the most useful feature out of the two, but not by so much. This shows that the exams taken where this study was conducted is indeed a tell tale sign of user knowledge and that time spent studying, which may increase exam performance, does not tell us much about the users knowledge.

The impacts this study could bring about are vast. By performing a similar result at other institutions it could be a way of confirming if their testing procedures are representative of real life skills. This is something every test maker hopes to achieve and could be substantiated by a similar result as in this study.

In the future this could lead to questions regarding what else is indicative of user knowledge other than exam performance. Are there other indicators more important than exam performance?

We would also like to state that to improve this studies reliability we could make some or all of the following improvements. We could use cross validation to increase the confidence of our

results, include more features to build a more robust and accurate model, and test out other regression techniques.

Assumptions and Limitations

In this study we determined a model that can classify students into one of four categories for student performance based on a few metrics. We acknowledge that our model was only applied to a single dataset from one university and extracting these results and methods to new datasets will yield different results that are out of our scope. We also assumed that all data was taken in a random, unbiased manor from Gazi University.

Reference

Kahraman (2012) Kahraman (2013) Timbers (2022) Anthony (2022)

Anthony, Prohaska, E. 2022. "DSCI100 Classification Project." Vancouver, Canada: The University of British Columbia.

Kahraman, Sagiroglu, H. T. 2012. "The Development of Intuitive Knowledge Classifier and the Modeling of Domain Dependent Data." https://www.sciencedirect.com/science/article/abs/pii/S0950705112002225?via%3Dihub.

———. 2013. "User Knowledge Modeling Data Set." Ankara, Turkey: UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling#.

Timbers, Campbell, T. 2022. "DSCI100 Textbook." Vancouver, Canada: The University of British Columbia. https://datasciencebook.ca/classification2.html.