Predicting Performance on MOOC Assessments using Multi-Regression Models

Zhiyun Ren, Huzefa Rangwala, Aditya Johri

George Mason University
4400 University Drive, Fairfax, Virginia 22030
Outline

- Background
- Personal Linear Multi-Regression Models
- Feature selection
- Experiments and discussion
- Conclusion and future work
Background
### Video Lectures

Having trouble viewing lectures? Try changing your player in course preferences.

<table>
<thead>
<tr>
<th>Week 1 -- Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture 1.1: Human Computer Interaction (4:18)</td>
</tr>
<tr>
<td>Lecture 1.2: The Power of Prototyping (13:48)</td>
</tr>
<tr>
<td>Lecture 1.3: Evaluating Designs (12:15)</td>
</tr>
<tr>
<td>Lecture 1.4: The Birth of HCI (8:48)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Week 2 -- Needfinding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture 2.1: Participant Observation (12:55)</td>
</tr>
<tr>
<td>Lecture 2.2: Interviewing (11:37)</td>
</tr>
<tr>
<td>Lecture 2.3: Additional Needfinding Strategies (11:54)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Week 3 -- Rapid Prototyping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture 3.1: Paper Prototypes and Mockups (12:47)</td>
</tr>
</tbody>
</table>
Overview

- **Information we have:** MOOC server log

- **Things we want to do:** Predict student’s performance
Challenge

- Various kinds of participants
- High attrition rate
- Flexible timetable

Baselines we have tried: Linear regression model, meanscore
Personal Linear Multi-Regression Models

\[ \hat{g}_{s,a} = b_s + b_c + p_s^t W f_{sa} = b_s + b_c + \sum_{d=1}^{l} (p_{s,d} \sum_{k=1}^{n_F} f_{sa,k} \omega_{d,k}) \]

- \( l \) -- Number of regression models
- \( n_F \) -- Number of features

\[
\text{minimize} \quad \frac{1}{2N} \sum_{i=1}^{N} (g_{s,a} - \hat{g}_{s,a})^2 + \lambda (\|P\|_F^2 + \|W\|_F^2) + \gamma (\|P\| + \|W\|)
\]
Data structure

(a) Homework and quiz
(b) Video
(c) Study session
Feature selection

- quiz related features
- time related features
- interval-based features
- homework related features
Feature selection

- Video related features

- Session features
Experimental setup

- Different motivations part the data into two groups.
- Different models are applied for different data types.
Experimental protocol

- PreviousHW-based prediction

  HW1  HW2  HW3  HW4  ......

- PreviousOneHW-based prediction

  HW1  HW2  HW3  HW4  ......
Experimental baseline: KT-IDEM

Model parameters

\[ P(L_0) = \text{Initial Knowledge} \]
\[ P(T) = \text{Probability of learning} \]
\[ P(G_1 \ldots n) = \text{Probability of guess per question} \]
\[ P(S_1 \ldots n) = \text{Probability of slip per question} \]

\( n \) denotes the number of all questions.
Comparative Performance

- Prediction results with varying number of regression models for student group with continuous grade value
Comparative Performance

- Prediction results with varying number of regression models for student group with binary grade value
Comparative Performance

<table>
<thead>
<tr>
<th>HW#</th>
<th>Accuracy (↑)</th>
<th>F1 (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLMR</td>
<td>Baseline</td>
</tr>
<tr>
<td></td>
<td>Meanscore</td>
<td>KT-IDEM</td>
</tr>
<tr>
<td>2</td>
<td>0.641</td>
<td><strong>0.646</strong></td>
</tr>
<tr>
<td>3</td>
<td><strong>0.760</strong></td>
<td>0.580</td>
</tr>
<tr>
<td>4</td>
<td><strong>0.754</strong></td>
<td>0.710</td>
</tr>
<tr>
<td>5</td>
<td>0.867</td>
<td>0.809</td>
</tr>
<tr>
<td>6</td>
<td>0.730</td>
<td>0.678</td>
</tr>
<tr>
<td>7</td>
<td>0.716</td>
<td>0.675</td>
</tr>
<tr>
<td>8</td>
<td><strong>0.817</strong></td>
<td>0.762</td>
</tr>
<tr>
<td>9</td>
<td><strong>0.823</strong></td>
<td>0.794</td>
</tr>
<tr>
<td>Avg</td>
<td><strong>0.764</strong></td>
<td>0.707</td>
</tr>
</tbody>
</table>

- The comparison of the accuracy and F1 scores with baseline approaches.
Feature Importance

(a) AllStMed

(b) PartialStMed

(c) AllStLearn

(d) PartialStLearn

RMSE vs. Number of Multi-regression Models

Accuracy vs. Number of Multi-regression Models

Legend:
- Original results
- Without Homework Related features
- Without Session features
- Without Time Related features
- Without Quiz Related features
- Without Meanscore feature

George Mason University
Feature Importance

- IntervalLogin
- IntervalDailySession
- IntervalVideo
- IntervalQuizAttempt
- IntervalNumQuiz
- HwSessions
- TimePlayVideo
- TimeHwVideo
- TimeHwQuiz
- HWProblemSave
- VideoPctWatch
- VideoViewTime
- VideoNumPause
- VideoNum
- AvgQuiz
- NumQuiz
- AvgNumLogin
- AvgSessionLen
- NumSession
Conclusion and future work

- **Predict algorithm:** personalized multiple linear regression model.
- **Experimental results:** improved performance compared to baseline methods.
- **Other contribution:** analysis of feature importance.
- **Future work:** to set up an early warning system to help improve student’s performance.
Thank you!