PREDICTING TEST OUTCOMES ON AN ONLINE TUTORIAL

Cesur Dagli

July 22, 2018

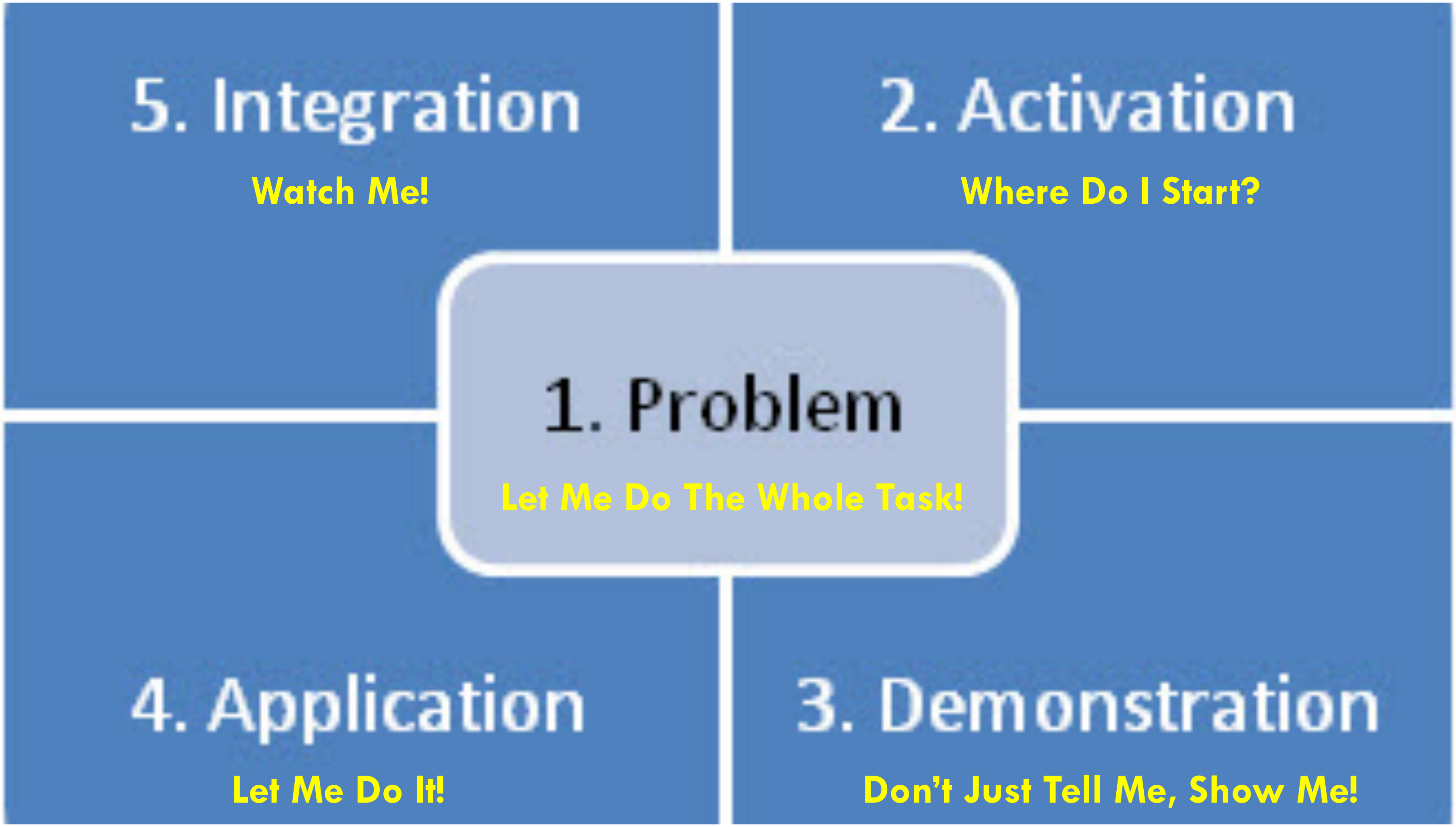
Springboard

Capstone Project Final Project

Background

- **Plagiarism** is crucial in academic setting.
- **Online tutorial** aiming at teaching how to recognize and avoid plagiarism.
  - Free
  - Used by hundreds of thousands of users worldwide
  - Designed based on a specific instructional theory (First Principles of Instruction by Merrill, 2002)
- Audience
  - **Undergraduate** and **graduate** students
  - Majority of students **required by their instructors** to earn a certification

First Principles of Instruction (Merrill, 2002, 2013)



5. Integration

Watch Me!

2. Activation

Where Do I Start?

1. Problem

Let Me Do The Whole Task!

4. Application

Let Me Do It!

3. Demonstration

Don’t Just Tell Me, Show Me!

Problem Statement

- Predict student test results (pass or fail) based on available data to provide support to those who are more likely to fail
- Determine key features related to target variables to close design deficiencies and improve tutorial

Data Collection

- Data were collected between January 2016 November 2018.
- 571,963 worldwide users registered for the tutorial
- 2,202,281 data points included in datasets

Description of Dataset

- Demographic Information
  - Required to take a certification test
  - Age, age, education level, current level (of understanding of the concept of plagiarism), and reason for taking test
- Survey
  - Optional (before taking a certification)
  - Designed to assess instructional quality of the tutorial
    - A 30-Item survey
- Certification Test
  - Two types of certification tests test outcomes
  - To earn a certificate, nine out of 10 question should be correctly answered.

Scales of Survey

- ***Overall quality & satisfaction scale:*** This scale measures student perceptions of overall quality of and student satisfaction with the tutorial.
- ***Task engagement and task success scales***. The engagement scale measures student perceptions of the time and effort they invested into learning to recognize plagiarism with the tutorial.
- ***Authentic problems scale.*** This scale measures student perceptions of overall presence of authentic plagiarism tasks arranged from simple to complex.
- ***Activation scale.*** This scale measures students' perceptions of overall presence of connecting past learning or experience with what is to be newly learned in the tutorial.
- ***Demonstration scale.*** This scale measures students' perceptions of overall presence of demonstration of what is to be learned in the tutorial.
- ***Application scale.*** This scale measures students' perceptions of their opportunity to use newly acquired knowledge or skills, to try themselves.
- ***Integration scale.*** This scale measures students' perceptions of overall presence of incorporation of what is learned in the survey into students' own lives.

Scale Construction

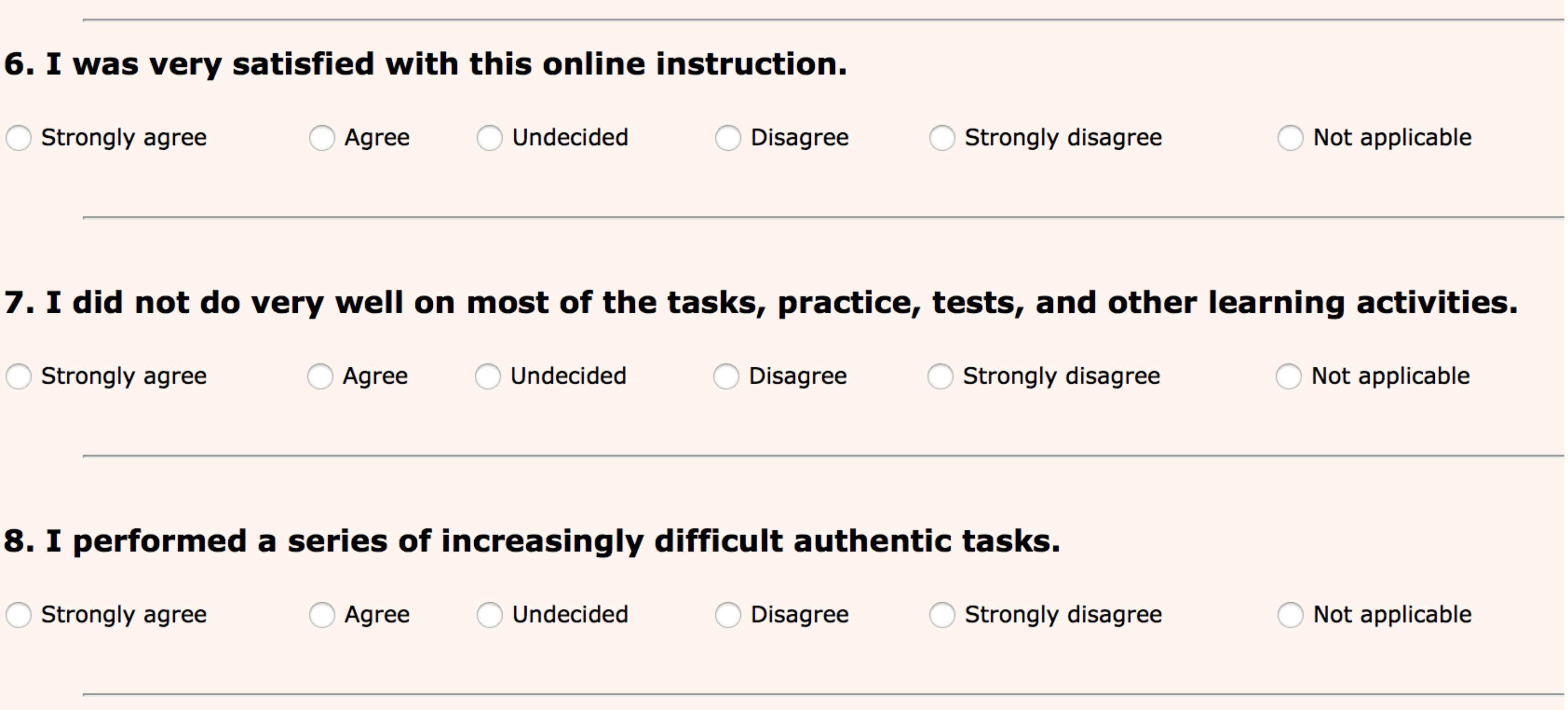
- Calculated by average across items for each scale:

• E.g., for a case for Activation:

(3 + 4 + 5) / 3 = 4.0 = scale score

- If a scale score is greater than 3.5, this was categorized as ‘**agreeing**’.

Examples of the Survey Items



6. I was very satisfied with this online instruction.

○ Strongly agree ○ Agree ○ Undecided ○ Disagree ○ Strongly disagree ○ Not applicable

7. I did not do very well on most of the tasks, practice, tests, and other learning activities.

○ Strongly agree ○ Agree ○ Undecided ○ Disagree ○ Strongly disagree ○ Not applicable

8. I performed a series of increasingly difficult authentic tasks.

○ Strongly agree ○ Agree ○ Undecided ○ Disagree ○ Strongly disagree ○ Not applicable

Data Selection Criteria

- Cases were excluded for those
  1. who did not take a test
  2. who did not complete the optional survey
  3. who did not rate all items on the survey

- Final dataset includes 9,579 data points.
  - 7400 fail and 2,179 pass cases

Descriptive Statistics

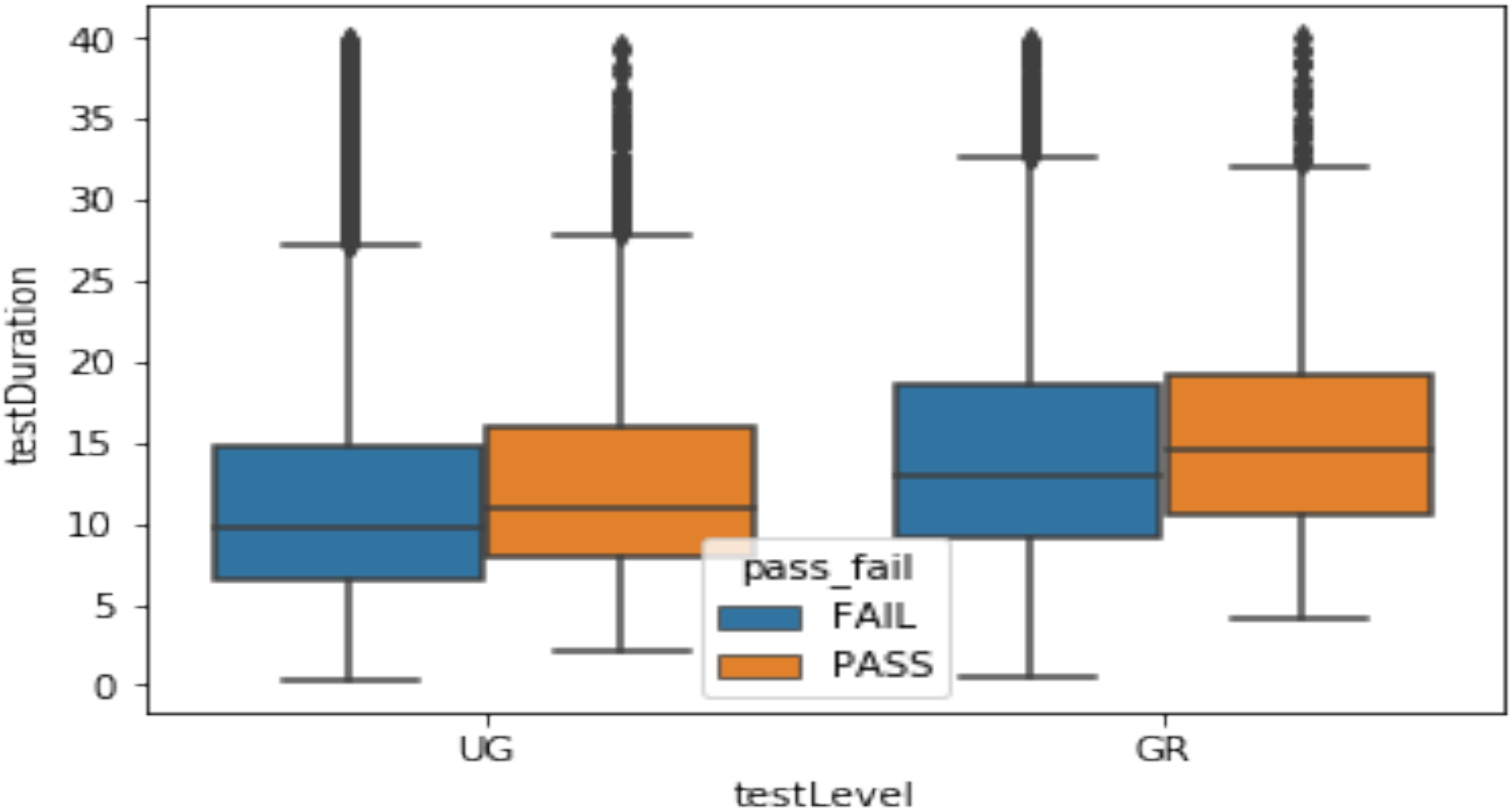
The High school or undergraduates (H&UG) Group (*N* = 6,827)

- □ 64% were **between 18 and 22 years** old.
- □ 64% holds a **high school diploma.**
- □ 96 % did the tutorial as **an assignment.**
- □ 49.8% rated their level of understanding of plagiarism as **confident.**

The Graduate (GR) Group (*N* = 2,752)

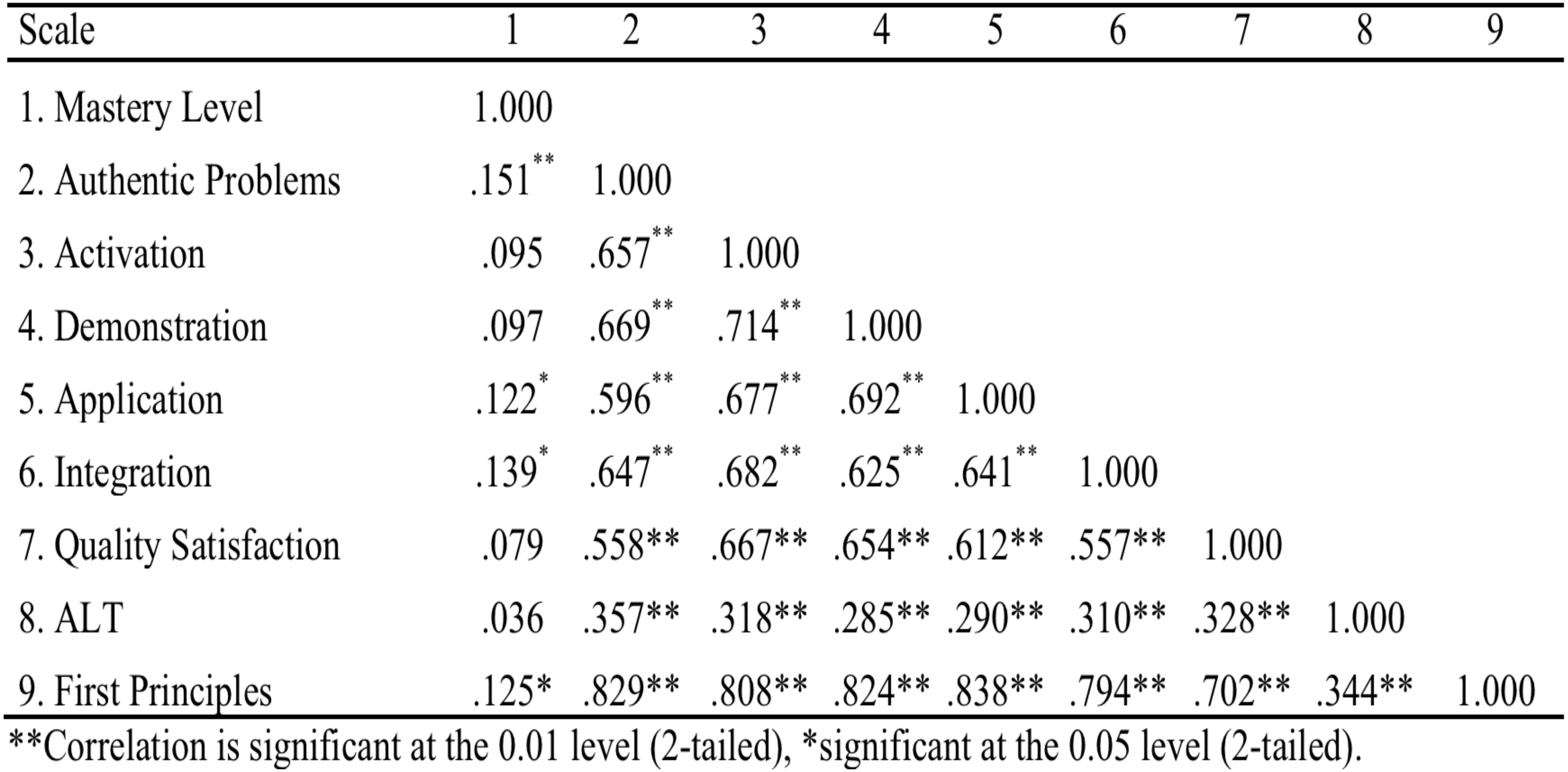
- □ 87% were **older than 22.**
- □ 68% holds a **bachelor's degree.**
- □ 93.3 % did the tutorial as **an assignment.**
- □ 51% rated their level of understanding of plagiarism as **confident.**

Average Completion Time by Groups



Inferential Statistics: Correlations

- *Correlations among survey scales*



Scale	1	2	3	4	5	6	7	8	9
1. Mastery Level	1.000								
2. Authentic Problems	.151**	1.000							
3. Activation	.095	.657**	1.000						
4. Demonstration	.097	.669**	.714**	1.000					
5. Application	.122*	.596**	.677**	.692**	1.000				
6. Integration	.139*	.647**	.682**	.625**	.641**	1.000			
7. Quality Satisfaction	.079	.558**	.667**	.654**	.612**	.557**	1.000		
8. ALT	.036	.357**	.318**	.285**	.290**	.310**	.328**	1.000	
9. First Principles	.125*	.829**	.808**	.824**	.838**	.794**	.702**	.344**	1.000

**Correlation is significant at the 0.01 level (2-tailed), *significant at the 0.05 level (2-tailed).

Inferential Statistics: ANOVA

- *Results for the Anova Query for the* ***GR Group***
  - When students perceived that first principles occurred; **they were more than four times as likely (0.282/0.07 = 4.02) to be high masters compared to those participants who did not agree that first principles were experienced in the tutorials.**

- *Results for the Anova Query for the* ***UG Group***
  - When students perceived that first principles occurred. **They were about three times as likely (0.263/0.09 = 2.92) to be high masters compared to those participants who did not agree that first principles were experienced in the tutorials.**

Baseline modeling: Logistic Regression

- A statistical method for **predicting the probability of binary classes**
- If the probability is greater than or equal to 0.5, a case is classified as Class-1 (pass) (Y=1) or else as Class-0 (Fail) (Y=0).
- Run a basic Logistic Regression, splitting the dataset into training and test sets.
  - Training Set (75% of the dataset): **train the model**
  - Test set 25% of the dataset: check how well the model will perform on fresh unseen data

Baseline modeling: Performance Metrics (1)

- Accuracy Score
  - Test accuracy: 0.78
  - Training accuracy: 0.77
- Confusion Matrix
  - True Negative (TN): 1852 cases failed the test and are predicted to fail the test.
  - False Positive (FP): 16 cases failed test but are predicted to pass the test.
  - False Negative (FN): 510 cases passed the test but are predicted to fail the test.
  - True Positive (TP): 17 cases passed the test and are also predicted to pass the test.

Baseline modeling: Performance Metrics (2)

- Classification Report
  - Precision: 0.78 for Fail Class and 0.41 for Pass Class
  - Recall: 0.99 for Fail Class and 0.02 for Pass Class
  - F1 Score: 0.87 for Fail Class and 0.05 for Pass Class

- Imbalanced Dataset
  - The majority (fail) class: 7400 and minority class (pass): 2179)
  - The model is biased towards the majority class.

Extending Modeling

- Resampling Techniques
  - Under-sampling with RandomUnderSampler
  - Oversampling with RandomOverSampler
  - Synthetic Minority Oversampling Technique (SMOTE).
  - Adaptive synthetic sampling (ADASYN)
- Ensembling Techniques
  - Random Forest Classifier
  - Balanced Random Forest Classifier (BalancedRandomForest)

Extending Modeling: Findings

Models	Classifiers	Accuracy	Class	Precision	Recall	F1 Score	Support
Baseline Model	Logistic Regression (LR)	0.78	Fail	0.78	0.99	0.88	1868
			Pass	0.52	0.03	0.06	527
Re-Sampling	LR with Under_Sampling	0.59	Fail	0.85	0.58	0.69	1868
			Pass	0.30	0.63	0.40	527
	LR with Over Sampling	0.60	Fail	0.85	0.60	0.70	1868
			Pass	0.31	0.63	0.41	527
	LR with SMOTE	0.60	Fail	0.85	0.59	0.69	1868
			Pass	0.30	0.63	0.41	527
	LR with ADASYN	0.58	Fail	0.85	0.56	0.68	1868
			Pass	0.30	0.66	0.41	527
Ensembling	Random Forest	0.75	Fail	0.79	0.94	0.86	1868
			Pass	0.27	0.08	0.12	527
	Balanced Random Forest	0.58	Fail	0.86	0.55	0.67	1868
			Pass	0.30	0.67	0.41	527

Feature Importance Analysis

```
 Coef.                    Std.Err.    z     P>|z|
----------------------------------------------------
testLevel                 0.0133   0.0785    0.1693
highestEd                -0.0190   0.0354   -0.5373
reason                   -1.1745   0.0895  -13.1296
level                     0.0594   0.0514    1.1566
ave_quality_Satisfaction -0.1158   0.0625   -1.8517
ave_authenticProblems     0.3739   0.0681    5.4869
ave_ALT                   0.0519   0.0614    0.8456
ave_activation           -0.4345   0.0835   -5.2043
ave_demonstration        -0.2423   0.0814   -2.9758
ave_application           0.2638   0.0810    3.2587
ave_integration           0.0769   0.0834    0.9221
taskCompleted             0.1402   0.0288    4.8720
viewedVideoCases         -0.0453   0.0485   -0.9331
viewedDemVideos           0.1468   0.0464    3.1654
integrationCompleted     -0.0934   0.0224   -4.1759
practiceTestTaken         0.2127   0.0475    4.4816
===================================================
```

Conclusion and Future Works

- Poor Classification Model
- Feature importance analysis should be taken into consideration
- The imbalance in the dataset is artificial
- More data is needed to have models to do better prediction
- Instead of relying on student self-reports, use tracking the sequence of event occurrences about what each user actually does as data

Recommendations

- Make sure that everyone takes the survey
- Reliability of Test Questions