Springboard Data Science Career Track
Capstone Project 1 Final Report:
Predicting Student Test Outcome on an Online Tutorial

Cesur Dagli

July 22, 2019

Section 1: Introduction

Indiana University offers an online tutorial that aims at teaching how to recognize and avoid plagiarism. This online tutorial is free and has been used by hundreds of thousands of users worldwide. The majority of the users of this tutorial are undergraduate and graduate students, who are required by their instructors to complete the tutorial and take and also pass a test to earn a certification. In order to earn a certification, students take a 10-question test, which is randomly selected from a very large question pool. Only those who correctly answer 9 out of 10 questions pass the exam and earn a certification.

When students fail the certification test at their first attempt, they become frustrated. The university receives a significant number of emails with complaints about the tutorial and the school from students who fail the test at their first attempt. The designers of this tutorial would like to improve the pass rate, decrease the number of attempts, and help students pass a test and earn a certificate in their first attempts. Building models that predict whether or not a student will pass a certification test based on the data available, and identifying the variables that influence a passing versus a failing class could help the designers improve the pass rate and decrease the average attempt to earn a certificate.

From a technical point of view, machine learning models and various algorithms were used to address this supervised learning binary classification problem to predict the target (pass or fail). After implementing the phases of data wrangling, storytelling, inferential statistics, and machine learning, the best model found is able to predict 60% of the negative class (those who fail) and 63% of the positive class (those who pass) correctly. In other words, this model fails to label 40% of negative class and 37% of the positive class. Therefore, even though this model does a good job over the baseline model (logistic regression) and has a better performance than a random classifier (i.e., decision is made by “flipping a coin”) I would still consider this classifier as needing further improvement.

Section 2: Approach

2.1. Description of the Dataset

Anyone can access the tutorial. However, those who would like to earn a certification, have to complete the registration process, which collects demographic data. Also, students are asked to complete an optional 30-item survey, which seeks a student's perception of the quality and satisfaction of activities on the tutorial. And finally, students take a 10-item certification test, whose questions are randomly selected from a test item pool. Since students can take the certification test multiple times, the dataset may include multiple data points per person. Also, the tutorial provides two types of certification tests: one for advanced high school and undergraduate students, and the other one for graduate students, which is harder. Data were collected via two datasets.

2.1.1. Demographic Data: The first was a MySQL database used in the registration process to save user registration information, responses to the demographic survey, information concerning the last certification test passed such as test ID, certification date, and IP address used to take the passing test. This process stored all information of registered users regardless of participant or non-participant status. This dataset contains 571,963 users who registered for the tutorial.

- **Demographic data (required):** Test ID, age, education level, current level (of understanding of the concept of plagiarism), and reason for taking test.

2.1.2. Survey Data: To create the second dataset, extensive PHP code was developed, which stored data in sequential ASCII files created according to the users' actions. For instance, since the survey is optional, if an user takes the survey, the responses to the survey are appended to the appropriate ASCII files. The ASCII file for the survey results keeps records for each user's e-mail address, name, IP address, test type (UG or GR), test ID, date/time, minutes taken to pass, test attempts, number of right and wrong answers, and survey results. There are **three sub datasets** including more than 2 million data points with 74 columns.

Each unique user e-mail address was the primary key to cross-link other ASCII files as well as in the MySQL database. Both data storage mechanisms were backed up daily to other locations to help prevent accidental data loss. These data storage locations cannot be accessed through Web browsers. Test results and responses to the survey are stored in a secure location on the host site.

- **30-item survey (optional):** Satisfaction, quality, data related to activities, etc.
  - ***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.
  - ***Combined First Principles scale.*** This scale measures students' perceptions of overall presence of First Principles of Instruction in the tutorial.

- **Certification test (required for those who would like to earn a certification):** Test types (one for advanced high school and undergraduate students and one for graduate students), test outcomes (pass or fail)

2.2. Data Cleaning and Wrangling

Below are the main steps that were used to prepare the datasets to get them ready for further analysis.

1. Data collected from the MySQL database were exported as a CSV file, whereas MOO-TALQ data were stored by PHP scripts as three ASCII text files on the webserver. The CSV file provided rows for individual users and columns for the variables described in the data collection section. However, the text file contained one row for each attempt to take a certification test by each user. Thus, for each new attempt by a user a new row was added to the text file.
2. Both the CSV file and text files were then loaded into in Pandas Dataframe objects (using pd.read_csv and pd.read_txt).
3. All variables in both data frames were then properly defined and were assigned appropriate data types (Encoding objects to float and category using pd.astype()).
4. Next, three sub-datasets of Survey data were combined into one dataframe since all these three sub datasets have the same number of variables with a different number of data points (used pd.concat()).

5. Checked the null values were checked. There were four datapoints with no email addresses, which is my key value so that I dropped those data points.
6. Four negatively worded items in the survey were reverse-coded by computing new variables using pd.map to "recode," by which 5 became 1, 4 became 2, 3 stayed the same, 2 became 4 and 1 became 5.
7. And then unnecessary columns for the analysis were dropped, which are first name, last name, certification date, IP, the survey items needed to be reversed, etc.
8. Next, the demographic data and survey data were merged by using each user's unique email address as the case identifier (key) (using pd.merge). This merged file included multiple rows for those who made more than one attempt to take a certification test.
9. To handle missing data for any of the items of a scale for a given case, all data points with a missing value from the survey were dropped.
10. To interpret whether or not a test user agreed on the survey scales stated above, scale scores were formed based on participants' responses to the survey by calculating the mean scores for each scale for each case. Therefore, any mean scale score greater than 3.5 were categorized as "agreeing" on the presence of each first principle in the tutorial. Otherwise, it was categorized as disagreeing.
11. The final data frame was written to a comma-separated values (csv) file (using pd.to_csv()).

2.3 Exploratory Data Analysis

2.3.1. Descriptive Statistics. I used plots of type bar, box and scatter to display the frequencies of categorical variables and groups of numerical data (minimum, first quartile, median, third quartile, and maximum) through their quartiles (using seaborn boxplot). I checked the statistics for both subgroups of GR and UG. For instance, the plot in Figure 1 show that the average completion time for a test was shorter for the UG group than for the GR group. These box plots also show the outliers in the groups. I also observe that there is a greater variability for the mean of UG group as well as larger outliers.

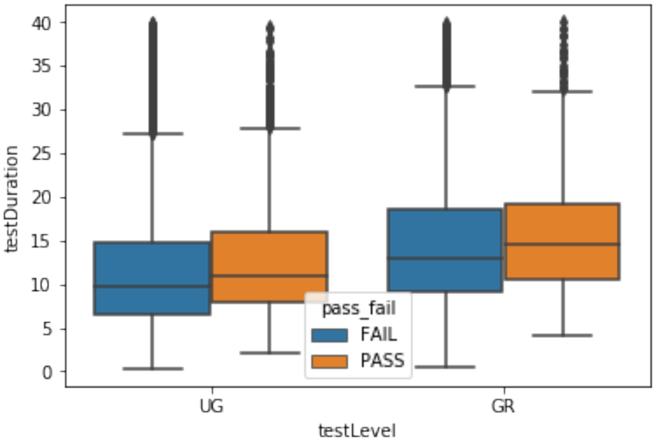

Figure 1.

I also used seaborn pair plots to show pairwise relationships between the features in a dataset. The plots revealed that there is a positive correlation between the scales of the survey. These plots below were used to visually summarize and review data for advanced analyses.

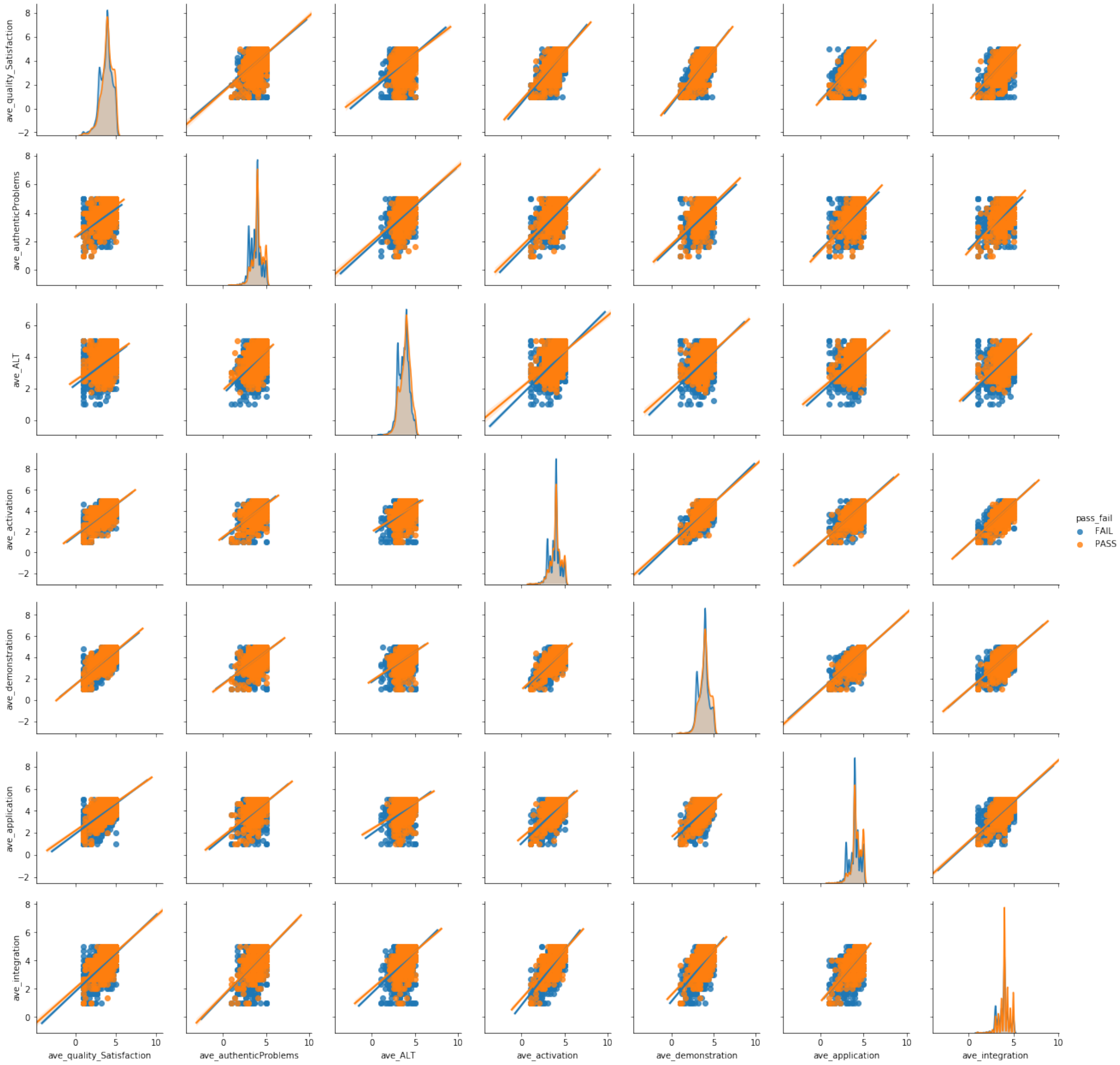

Figure 2.

2.3.2. Inferential Statistics

2.3.2.1. Anova. I run one-way ANOVA to compare the means of a continuous variable in two or more independent comparison groups. The hypothesis is as follows

- H0: There is no difference among students with confidence, some, and little level knowledge in the GR group based on their test duration, number of right questions, satisfaction with the tutorial respectively.

- H1: There is a significant difference between group means.

I run one-way ANOVA for both the UG and GR groups.

UG Group	**GR Group**

pvalue=0.1307604599473568
pvalue=7.590398905584018e-17
pvalue=5.895422160775933e-14
	pvalue=0.060285424876925564
	pvalue=4.002470700440738e-10
	pvalue=0.07289494126703458

I reject H0 for the second and third test. This means that I have statistically significant differences at α=0.05 in between the number of right questions and the satisfaction with the tutorial among students with the three levels of knowledge. I failed to reject H0 for the first one and therefore I cannot say whether or not the difference between test durations among students with three levels of knowledge is statistically significant with respect to α=0.05.	I reject H0 for the second test. I failed to reject H0 for first one and third one, in both cases using a threshold ofα=0.05.

However, where the difference between groups is unknown yet I calculated multiple t-tests.

UG Group	**GR Group**
The t-tests resulted in that there are statistically significant differences in mean of number of right questions and satisfaction between students with confident level and some level; between students with confident level and students with little level; between students with some level and student with little level.	The t-tests resulted in that there are statistically significant differences in mean of number of right questions between students with confident level and some level; between students with confident level and students with little level; between students with some level and student with little level.

2.3.2.2. Chi-Square Test. I also ran Chi-Square test to check the relationship between categorical variables. My hypothesis:

- H0: There is no statistically significant relationship between test result and experiencing first principles of instruction.
- H1: There is a statistically significant relationship between test results and experiencing first principles of instruction.

According to the theory of First Principles of Instruction (Merrill, 2002), those who experience the instructional principles are more likely to have better learning and performance. Therefore, I create a new column called **fpi** for those who stated they experience First Principles of Instruction by calculating each score of **['ave_authenticProblems', 'ave_activation', 'ave_demonstration', 'ave_application','ave_integration']** greater than 3.5 so that I can conclude that those with score greater than 3.5 experienced the First Principles of Instruction.

I first formatted the data into a frequency count table by using the **pd.crosstab()** function in pandas. And then I used **stats.chi2_contingency** to check the relationship between test results and First Principles of Instruction. I did these steps for both the UG and GR groups.

GR Group

Test Result	**First Principles of Instruction**			
	no	**partially**	**yes**	**All**
FAIL	143	701	1294	2138
PASS	12	210	509	731
All	155	911	1803	2869

I found p-value less than 0.05. Therefore, we can reject the null hypothesis. There is a statistically significant relationship between test results and First Principles of Instruction in the GR group.

As shown in the contingency table, there were 1803 out of 2869 GR test takers who agreed with having experienced all of the first principles. Given that the antecedent condition was true, the consequent (student mastery is ‘Pass’), was true in 509 of the 1803 cases, which yields a probability estimate of 509/1803 or 0.282. Thus, 28.2% of students who were rated as high masters

UG Group

Test Result	**First Principles of Instruction**			
	no	**partially**	**yes**	**All**
FAIL	906	2263	2627	5796
PASS	91	559	942	1592
All	997	2822	3569	7388

I found p-value less than 0.05. Therefore, we can reject the null hypothesis. There is a statistically significant relationship between test results and First Principles of Instruction in the GR group.

As shown in the contingency table1, there were 3569 out of 7388 UG test takers who agreed with having experienced all of the first principles. Given that the antecedent condition was true, the consequent (student mastery is ‘Pass’), was true in 942 of the 3569 cases, which yields a probability estimate of 942/3569 or 0.263. Thus, 26.3% of students who were rated as high masters by a certification test reported that they also agreed that first principles were experienced when using the tutorial.

Next, the pattern was investigated where the antecedent condition was no agreement with having experienced first principles and where the consequent was high mastery. For this pattern, the antecedent occurred 155 times, the consequent occurred in 12 of those 155 cases, for a probability estimate of 12/155 = 0.07. Thus, 7% of students who demonstrated high mastery of the objectives of the tutorial failed to agree that first principles occurred.

This finding can be further interpreted as follows: 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.**

by a certification test reported that they also agreed that first principles were experienced when using the tutorial.

Next, the pattern was investigated where the antecedent condition was no agreement with having experienced first principles and where the consequent was high mastery. For this pattern, the antecedent occurred 997 times, the consequent occurred in 91 of those 997 cases, for a probability estimate of 91/997 = 0.09. Thus, 9% of students who demonstrated high mastery of the objectives of the tutorial failed to agree that first principles occurred.

This finding can be further interpreted as follows: 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.**

2.4 Baseline modeling: Logistic Regression

As stated above, the goal is to predict whether not a student passes a certification test based on the data available. Therefore, Logistic Regression was conducted, which is a statistical method for predicting the probability of binary classes, for this classification problem to predict a student's test status belongs to what group: Pass or Fail.

Logistic Regression uses the Sigmoid function, which produces values between 0 and 1. In this study, the value 0.5 is used as the probability threshold to determine the classes. If the probability is greater than or equal to 0.5, we classify it as Class-1 (pass) (Y=1) or else as Class-0 (Fail) (Y=0).

2.4.1. Training and Test Datasets. First, I run a basic Logistic Regression by splitting the dataset into training and test sets. The training set is used to train the model. The test set is used to check how well the model created performs on this dataset in order to estimate how the model will perform on fresh unseen data. Splitting dataset into the training and test sets resulted in the training set with 7184 cases (75% of the dataset) and the test set with 2395, both of which includes 29 features and one target. After the model was fitted with the training set, and predictions were generated using the training set and the test set, I utilized several performance metrics to see how well the model performs. Below are the metrics I used:

2.4.2. Performance Metrics.

2.4.2.1. Accuracy Score. In classification problems, accuracy is a commonly used metric, which is the proportion of the total number of predictions that were correctly calculated. The main goal is to maximize overall accuracy by machine learning algorithms. In this case, the model's test accuracy is calculated as (0.78), which is pretty close to and slightly better than the training accuracy (0.77), then I can say that there is no "variance" between the training accuracy and the test accuracy. This is an indication that the model will "generalize well", which means that the model will be well-behaved when new data is presented to it. Nevertheless, accuracy alone is usually not a reliable metric that can be used to assess the performance of a classification model.

2.4.2.2. Confusion Matrix. In addition to the accuracy score, there are other metrics for the evaluation of the classification predictions: Precision, Recall and F-Measure, which can be calculated using the confusion matrix. The confusion matrix below shows the number of correct and incorrect predictions which are broken down by each class. Below is the interpretation of the confusion matrix: 1869 (1852+17) cases are correctly labeled either fail or pass. However, 526 (510+16) cases are incorrectly labeled. More specifically:

- Accuracy: Out of all 2395 cases, the classifier correctly predicted 1869 of them.
- 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.

	Predicted: Fail the test	Predicted: Pass the test
Actual: Fail the test	1852 (TN)	16 (FP)
Actual: Pass the test	510 (FN)	17 (TP)

‘True' or ‘False' indicate if the classifier predicted the class correctly or incorrectly, respectively, whereas ‘positive' or ‘negative' indicates if the classifier predicted the desired class (in this case, ‘positive' correspond to ‘Pass the test', and ‘Negative' is for ‘Failed the test'). Using this terminology, the scores of Precision, Recall, and F-Measure can be calculated as follows:

- Precision: The ratio of true positives to the sum of true and false positives (TP/ (TP + FP)).
- Recall: The ratio of true positives to the sum of true positives and false negatives (TP/ (TP + FN)).
- F1 Score: A weighted (harmonic) mean of precision and recall such that the best score is 1.0 and the worst is 0.0.

In this case, the goal is to have high precision (not many fail cases predicted as pass cases) and high recall (predicted most pass cases correctly).

2.4.2.3. Classification Report. The Precision-Recall and F1-Score can be obtained using the function classification_report instead of calculating each manually.

	precision	**recall**	**f1-score**	**support**
Fail	0.78	0.99	0.87	1864
Pass	0.41	0.02	0.05	531
avg / total	0.70	0.78	0.69	2395

Let’s interpret the classification report above:

Precision score shows the accuracy of positive predictions, which is defined as the ratio of the number of positive cases and predicted positive to all cases (either positive or negative in actual) that were predicted as positive. In my case, the accuracy of positive predictions for the FAIL class in the test set is found 0.78, which means that my classifier correctly marked 78 out of 100 fail cases as the fail cases who were actually the fail cases. However, the precision score for the pass cases for the test set is found low as 0.41.

Recall is the ability of a classifier to find all positive instances. For each class, it is defined as the ratio of true positives to the sum of true positives and false negatives. In my case, recall scores for the fail class in the test set is found 0.99, which can be interpreted as the classifier was able to get 99 percent of actual fail cases correctly. However, the recall score for the actual pass class is very low, 0.02, which means only two percent of the pass cases were correctly classified as the pass cases.

The **F1** score is a weighted harmonic mean of precision and recall. Here, F1 scores for the fail class is very high, 0.87. However, it is very low for the pass class, 0.05.

Overall, as indicated by the classification report, the model provides good scores for the fail class, which is the majority class. However, the low scores are found for the pass class, which is minority class. The model is biased towards the majority class. Also, the support in the classification report shows that classes are not represented equally (the majority (fail) class: 5536 and minority class (pass): 1648) see the classification report for the training set. the Ratio of Pass Class to Fail instances is 0.3. Thus, the classes in the dataset are disproportionately represented, which refers to an imbalanced dataset. In the next section, several techniques are used to handle this imbalanced dataset.

2.5 Extended modeling: Dealing with Imbalanced Dataset

There are various approaches to deal with an imbalanced dataset to improve the performance of the model. One of the main approaches is to balance classes by either increasing the number of instances of the minority class or decreasing the number of instances of the majority class so they both have nearly equal representation in the dataset. Below are the approaches used to handle the imbalanced data issue.

Package imbalanced-learn was installed and used, which is a Python package offering several re-sampling techniques commonly used to deal with imbalanced datasets. It is compatible with sci-kit-learn and is part of sci-kit-learn-contrib projects.

Note that all resampling approaches were completed on the train set after the main dataset had split into the train and test sets to avoid the model simply learn and memorize the specific observations and overfitting and poor generalization to the unseen test data.

After applying the resampling methods, I had an equal ratio of observation for each class. And then I run a logistic regression with the balanced training data.

2.5.1. Resampling Techniques

2.5.1.1 Under-sampling with RandomUnderSampler. Under-sampling is the approach of reducing the number of samples in the majority classes to balance the class distribution in the dataset. This technique is a good choice with lots of data. The disadvantage of this technique might be to cause to lose useful information. Also, the final sample chosen by reducing the number of samples in the majority class may be biased and cause the classifier to perform poorly on the new dataset.

2.5.1.2. Oversampling with RandomOverSampler. Oversampling is another way to deal with imbalanced data issue. This technique basically generates new samples in the minority class by randomly sampling with replacement of the currently available samples. Unlike under-sampling, this technique does not cause to lose useful information. However, it may increase the likelihood of overfitting due to a replication of the minority class observations.

2.5.1.3. Synthetic Minority Oversampling Technique (SMOTE). Instead of duplicating some of the original samples of the minority class, the Synthetic Minority Oversampling Technique (SMOTE) can be used as an alternative oversampling method. SMOTE generates synthetic samples by using n-nearest neighbors in the minority class for each of the samples in the class. Since SMOTE is also an oversampling technique there is no loss of useful information. Also, this technique decreases the likelihood of overfitting as the generated samples are not a duplication of the available observation in the minority class.

2.5.1.4. Adaptive synthetic sampling ADASYN. Adaptive synthetic sampling (ADASYN) is an improved algorithm of SMOTE to generate new samples. The main difference between SMOTE and ADASYN is that ADASYN focuses on the samples that are incorrectly classified using the nearest-neighbors rule, which is not the case for SMOTE.

2.5.2 Random Forest Classifiers. The other classification models I considered was tree-based algorithms. Random forest classifiers build multiple decision trees and ensembles them to get a more accurate and stable prediction. Random Forest prevents overfitting most of the time, by creating random subsets of the features and building smaller trees using these subsets. To increase predictive power the following hyperparameters are used to choose the best-performing ones.

- n_estimators: The number of trees in the forest. In general, a higher number of trees increases the performance and makes the predictions more stable, but it also slows down the computation.
- max_features: The maximum number of features to consider when looking for the best split:

Since the Random Forest classifier does not balance each class, I also used BalancedRandomForest to balance the class distribution.

Section 3: Findings

The results obtained with the baseline model suggest that an ideal goal for my business problem would be to build classification models with high precision and recall.

In my case, after running the methods mentioned above, the classification report for each classifier with the test set yielded the scores below, the recall score found by Balanced Random Forest and Easy Ensemble for the pass class are highest (0.67). In other words, these two classifiers much increased the recall score for the pass class to 0.67, which was found 0.03 with the baseline model, logistic regression. However, the recall score for the fail class dropped to 0.55 from 0.99. The classifier that returned the highest average recall score for both classes is OverSampling (0.6 for the fail class and 0.63 for the pass class). Also, this classifier returns the highest F1 scores for both classes over the rest of the classifiers.

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

3. 1. Performance of the Classifiers

An ideal classifier should produce high precision and recall, which predicts both classes correctly. In my case, my goal is to determine those who failed the pass so that they will be supported with feedback and guidance before taking a test. Thus, I aim to increase the pass rate by providing the feedback to help those who are predicted to fail the test so that they can pass the test. The recall scores found by the classifiers for both classes fall between 0.44 and 0.69. And after balancing the classes with oversampling I got the highest average recall scores for both classes, which are 0.6 for the fail class and 0.63 for the pass class. Based on these recall scores, the classifier does label 60% of the negative class and 63% of the positive class correctly. In other words, this classifier fails to label 40% of negative class and 37% of the positive class. Therefore, even though this classifier does a good job over the baseline model (logistic regression) and also should not be considered as a random classifier I would still consider this classifier as a candidate to be improved since it fails to label about 40 percent of both classes correctly.

3.2 Feature Importance Analysis

The implications of feature importance analysis: It is important to use the most useful features to create a better predictive model. Feature selection is a method that seeks to include the most relevant features and excludes unnecessary features in the data, which help to create more accurate predictive models. Also, fewer features reduce the complexity of the model. In this case, the dataset includes 29 features. I wanted to make sure which one is the most relevant to predict the target feature, which is the status of the test result: pass or fail. To find the most influential features, I examined the coefficients of the model and used p-values. As seen in the Table 3, the features reason for taking test, activation, demonstration, and integration task completion (in red) are found significantly related to the fail status. However, the authentic problem, applications, task completion, demonstration videos, practice tests (in green) are found significantly related to the pass status. In other words, students who watch the videos showing how to recognize plagiarism and take the practice tests are more likely to pass the certification test.

Table 3. Feature Importance Analysis

```
                         Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-------------------------------------------------------------------------
testLevel                 0.0133   0.0785   0.1693 0.8656 -0.1405  0.1671
highestEd                -0.0190   0.0354  -0.5373 0.5911 -0.0884  0.0504
reason                   -1.1745   0.0895 -13.1296 0.0000 -1.3498 -0.9992
level                     0.0594   0.0514   1.1566 0.2474 -0.0413  0.1601
ave_quality_Satisfaction -0.1158   0.0625  -1.8517 0.0641 -0.2383  0.0068
ave_authenticProblems     0.3739   0.0681   5.4869 0.0000  0.2404  0.5075
ave_ALT                   0.0519   0.0614   0.8456 0.3978 -0.0684  0.1723
ave_activation           -0.4345   0.0835  -5.2043 0.0000 -0.5982 -0.2709
ave_demonstration        -0.2423   0.0814  -2.9758 0.0029 -0.4020 -0.0827
ave_application           0.2638   0.0810   3.2587 0.0011  0.1051  0.4225
ave_integration           0.0769   0.0834   0.9221 0.3565 -0.0866  0.2404
taskCompleted             0.1402   0.0288   4.8720 0.0000  0.0838  0.1966
viewedVideoCases         -0.0453   0.0485  -0.9331 0.3507 -0.1403  0.0498
viewedDemVideos           0.1468   0.0464   3.1654 0.0015  0.0559  0.2377
integrationCompleted     -0.0934   0.0224  -4.1759 0.0000 -0.1372 -0.0496
practiceTestTaken         0.2127   0.0475   4.4816 0.0000  0.1197  0.3057
=========================================================================
```

Section 4: Conclusions and Future Work

The academic community highly values the acknowledgment of contributions to knowledge. It is crucial to know a proper citation in academic works. Indiana University has created an online instruction to help students to understand, recognize, and avoid plagiarism. This online tutorial has become an essential tool and used by worldwide students. The majority of students were required to take this tutorial, pass a test, and earn a certificate by their instructors. Therefore, to meet their class requirements, students seemed to pass a test and earn a certificate as fast as they could, who tended to ignore some parts of the tutorial (sometimes all of the tutorial) and quickly jump into taking the test. This minimal effort caused students to fail in the test so that students attempted several certification tests, which created frustration on students.

The goal of this project was to create a model that predicts whether or not a student passes a certification test. In order to be able to predict student's success on a certification test, demographic information of the student collected in the registration process and data gathered the survey measuring student's experience on the activities created to teach how to recognize and avoid plagiarism.

Various machine learning approaches were applied to create a model that does a best prediction of a test result. Since the problem required binary classification methods, Logistic Regression was used as a baseline model. This baseline model resulted in bias towards the fail class, which is the majority class. In other words, this trivial classifier that predicts every case as the majority class can still produce a high accuracy, which is a misleading score and not appropriate measure of performance. Having a bias towards one of class refers to the imbalanced dataset issue. In this case, the classes are not represented equally, where the dataset included 7400 students who failed and 2179 students who passed.

Several techniques were used to handle this imbalanced dataset. Sampling methods were utilized to balance both the classes either reducing the number of the failing cases or increasing the number of passing cases. Also, SMOTE and ADASYN were used to create synthetic samples to make a balance between the classes. Finally, the Random Forest method was applied as an alternative ensembling machine learning classifier to create a model that predicts student's test outcomes.

The best model found was found through the Logistic Regression with the oversampling method (RandomOverSampler). This model is able to predict 60% percent of those who will fail correctly and 63% of those who will pass a certification test. Compared to the baseline logistic regression model, this model does a better prediction for both classes. However, this model is considered as a poor model since it fails to predict 40% of students who will fail and 37% of students who will pass a test.

Further work is needed to create a better model that shows a balanced mix of precision and recall for both classes. Instead of relying on student self-reports, tracking the sequence of event occurrences about what each user actually does when interacting with the online tutorial may provide more reliable data on how exactly they complete the tutorial and activities.

To earn a certificate, users have to register first, and take a test and answered 9 out of 10 questions correctly. However, completing the survey is optional. Only 7% of those registered (571,963) completed the survey. And when I dropped those cases with missing values the final dataset included those who completed the survey and took a test, which resulted in 9,579 (2%) unique cases in the analysis. Also, 76% (434,691) registrants of the survey passed a test and earned a certificate. However, after applying the data cleaning and selection criteria, the final dataset used in this project included the size of the failing cases that is three times greater than the passing class. In my final dataset, 77% of the cases failed the test. This big passing rate difference happened because of several reasons. First, most of those who passed the test did not complete the survey. Therefore, I missed a significant amount of data about those who passed the test, therefore, the imbalance in the dataset is artificial. Second, users who failed in their first try but passed the test in their next attempts did not complete the survey since they already did. However, I considered them as fail cases since they completed the survey on their fail attempt, and I dropped their pass attempts because there was no survey data associated with their pass attempts. Thus, those who failed a test and completed the survey considered in the data analysis, while their pass attempts were dropped. In conclusion, more data is needed to have models to do better prediction. Therefore, action should be taken to get more cases with completing the survey.

Section 5: Recommendations for the Client

Make sure that everyone takes the survey: As discussed in the findings section, there is a lack of data in order to create a better model that classify both the pass and fail classes. One way to increase data points is to require everyone to take the survey, which may be held during the peak time of use until collecting sufficient data.

Reliability of Test Questions: The major concern with this online plagiarism tutorial is high test attempt rate (16%). Based on the data, a student needs to take about 7 attempts to take a test. Giving students multiple attempts to improve test scores provides a powerful learning opportunity. However, average 7 attempts to be successful is not an efficient way to help a student. There may be several reasons that caused higher average test attempts. First, students don't completely and carefully go through all parts of the tutorial. The survey results indicated that a relatively small proportion of users completed most parts of tutorials.
That is, 24% test takers completed most parts of the tutorial, which is consistent with data from Google Analytics. Both the survey results and Google Analytics show that the large majority of the users typically exhibited two patterns of initial use of the tutorial. First, after they logged in, they tried to pass a certification test, but most failed on their first attempt. And those students kept taking tests without any review of the tutorial. Also, it seems that some students did not complete any part of the tutorial and jumped in to take a test. Second, again students have to answer correctly 9 out of 10 questions to pass the test and earn a certificate. However, fatigue and lack of attention to detail can be obstacles to passing a test, even when students can recognize kinds of plagiarism and non-plagiarism. Students may be missing essential details in questions. Some questions may look similar to others but actually, differ in small but significant details from one test to the next. Note further, if the student version contains both word-for-word and paraphrasing plagiarism, the correct answer is word-for-word. Students need to read each question carefully, and not be hurried. Third, some questions might be very difficult or problematic so that the reliability of questions should be conducted and if exist those questions should be identified and updated or removed.