TRI-AD

Data Scientist Challenge

Ankush Khullar
May 7, 2020

Part 1: Analysis

The dataset contains 11 features and 5,000 observations. The 11 features include 3 categorical features (sex, language, country), 2 continuous features (age, hours_studied) and 3 binary features (dojo_class, test_prep, pass). Three additional features include first name, last name and notes. The name of a student should not affect whether he or she passes the test and the notes column is mostly blank with random information for some entries.

1. Demographic Relationships

The demographic details that could be used to determine which students are most likely to pass the test include sex, language, country, and age. After transforming binary features into integers and the categorical variables into dummy variables, I examined the Pearson correlation between the features and whether a student passed the test. The feature with the highest correlation was sex (0.1873), followed by age (0.0594) and whether the student is from Mexico (0.0224). However, all of the correlations were very low, indicating that there is a weak relationship between the demographic features and whether a student passed the test. Furthermore, based on the p-values for the correlation coefficients, only the correlations for sex and age were deemed statistically significant at a 95% confidence level.

1.1. Sex

Males make up approximately 75% of the dataset. When calculating the correlation between `sex` and `pass`, males were coded as 1 and females were coded as 0. Therefore, the positive correlation indicates that males are more likely to pass the test than females. Examining the pass rate by sex shows that a higher percentage of males (82.5%) pass the test than females (64.7%).

In order to ensure that the difference in pass rates between males and females is statistically significant, a chi-square test of independence was performed. The chi-square statistic was 174.42 with a p-value of 7.99e-40. Based on the p-value, the difference in pass rates is statistically significant at a 95% confidence level.

1.2. Country

Japan is the most prevalent country in the dataset, followed by Italy and France. Pass rates by country range from 72.1% for Italy to 86.4% for Finland. Countries with the highest pass rates were Finland, Mexico and Australia and countries with the lowest pass rates were Italy, Spain, and France.

A chi-square test of independence was performed to determine if the difference in pass rates across countries is statistically significant. The chi-square statistic was 11.48 with a p-value of 0.2445. The high p-value indicates that the difference in pass rates across countries is not statistically significant at a 95% confidence level.

It may be surprising that a 14% difference in pass rates between students from Italy and students from Finland is not significant. The reason this is the case is because of the small sample size for all countries other than Japan. Small sample sizes translate to high standard errors. Therefore, although the difference between pass rates may seem large, the high standard errors mean that they cannot be deemed significant.

1.3. Language

Pass rates by language range from 72.1% for Italian to 86.5% for Finnish. Languages with the highest pass rates were Finnish, English and Spanish and languages with the lowest pass rates were Italian, French, and Japanese. This is similar to the results of ranking countries by pass rate, which showed that students from Finland had the highest pass rate, followed by students from Mexico and Australia.

A chi-square test of independence was performed to determine if the difference in pass rates across languages is statistically significant. The chi-square statistic was 6.78 with a p-value of 0.2379. The high p-value indicates that the difference in pass rates across languages is not statistically significant at a 95% confidence level. Again, this is due to the low representation of languages other than Japanese in the dataset.

1.4. Age

The final piece of demographic information that can be used to predict whether a student passes the test is age. The distribution of age has positive skew with a median value of 24. The average age of students that fail the test is 25.1, while the average age of students that pass test is 26.2. A two-sample t-test was performed to determine if the difference in the average age is statistically significant. The t-statistic was 4.12 with a p-value of 2.64e-5. The low p-value indicates that the difference is significant at a 95% confidence level, providing evidence that students that pass the class are, on average, older than students that fail the test.

Binning age by quartiles shows that 76% of students between 18 and 30, 79% of students between 24 and 30 and 81% of students between 30 and 50 pass the test. A chi-square test was performed to determine if the difference in pass rates across age bands is statistically significant. The chi-square statistic was 14.00 with a p-value of 0.0029, indicating the difference in pass rates across age bands is statistically significant at a 95% confidence level. However, the chi-square test only provides evidence that at least one of the age band pairs has a significant difference in pass rates.

In order to determine which pairs have a significant difference, a post-hoc test was performed by running a chi-square test for each pair of age bands. Since tests were run repeatedly, the Bonferroni adjustment was used to correct for issues associated with multiple testing. Based on the results of the post-hoc analysis, there is a significant difference in pass rates between students under the age

of 30 and students above the age of 30, with approximately 76.9% of those under 30 passing the test and 81.2% of those above 30 passing.

1.5. Summary

Based on the analysis of the relationship between demographic factors and pass rate, the following conclusions can be made:

1. Sex - males are more likely to pass the test than females;
2. Country - no significant relationship between country and pass rate;
3. Language - no significant relationship between language and pass rate; and,
4. Age - positive relationship between age and pass rate.

2. Efficacy of Interventions

The dataset contains binary features, `dojo_class` and `test_prep`, representing interventions intended to increase the chance of a student passing the test. Approximately 30% of the students participated in the Dojo class and 21% participated in the test prep course.

The correlation between students who took the Dojo class and students that passed the test is 0.1637, while the correlation between students who took the test prep course and students that passed the test is only 0.0163. Based on this information, it appears that students who take the Dojo class are slightly more likely to pass the test. On the other hand, there does not seem to be a relationship between students who took the test prep course and whether they passed the test.

2.1. Dojo Class

Approximately 88% of students who took the Dojo class passed the test, while only 74% of students who did not take the class passed. A chi-square test between `dojo_class` and `pass` resulted in a chi-square statistic of 131.52 with a p-value of 1.90e-30. The low p-value indicates that the difference in pass rates between students who took the Dojo class and students who did not take the Dojo class is statistically significant at a 95% confidence level.

As it was determined in the previous section that sex and age are related to pass rate, we must ensure that the effect from `dojo_class` are independent of the effects from the demographic features. That is, we must ensure that the higher pass rate for students who took the class is not because more males take the class than females or because students who take the class are, on average, older than students who do not take the class.

The distribution of males and females who took the Dojo class is representative of the sample as a whole. That is, approximately 75% of the students who took the class are male and approximately 75% of students who did not take the class are also male. Furthermore, a two sample t-test examining the relationship between the `age` and `dojo_class` resulted in a t-statistic of 1.00 with p-value of 0.3158, indicating that there is not a statistically significant difference in the average age of students who took the class and students who did not take the class. Therefore, we

can conclude the positive effect of `dojo_class` on pass rates is independent of the demographic factors.

2.2. Test Prep

Approximately 79% of students who took the test prep class passed the test and 78% of students who did not take the test prep class passed. A chi-square test examining the relationship between `test_prep` and `pass` resulted in a chi-square statistic of 1.23 with a p-value of 0.2672. The p-value indicates that the difference in pass rates between students who take the test prep class and students who do not take the test prep class is not statistically significant at a 95% confidence level. This is in line with the very low correlation `test_prep` and `pass`. Therefore, we can conclude that the test prep course does not affect the likelihood of a student passing the test.

3. Further Analysis

The previous sections provided evidence that `sex`, `age` and `dojo_class` are related to whether a student passes the test, but do not assess the magnitude or importance of these features. One way to do that is to perform a logistic regression and analyze the coefficients for each feature.

The results of the logistic regression reveal the following:

1. Sex - The odds that a male will pass the test are 2.53 times higher than the odds that a female will pass;
2. Age - A one-year increase in age increases the odds of passing the exam by 1.03; and,
3. Dojo class - The odds that a student who takes the Dojo class will pass the test are 3.04 times higher than a student who does not take the Dojo class.

Based on these results, Dojo class had the highest impact on whether a student passes the test, followed by sex and age. It is important to note that the relative importances are based on the logistic regression analysis, including any assumptions made by a logistic regression model (e.g., linear relationship between predictors and log odds of the response). Therefore, the relative importances could vary depending on the model used.