Woven Planet Data Scientist Challenge - Xan Varcoe

Part 1: Analysis

Part 1.1 Demographic Details

Summary
In terms of demographics we can see that as student age increased, the likelihood of passing also increased. The USA was the country with the highest pass rate and for language it was people studying French.

Fig 1 - Histogram showing frequency distribution of student ages for test passes (blue) and fails (orange)

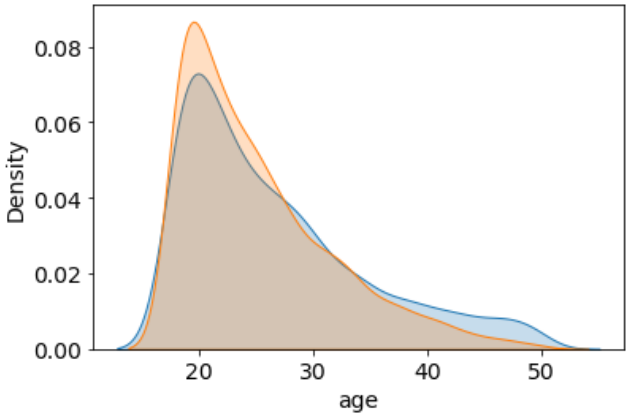


Fig 1 shows that older students were slightly more likely to pass than younger ones. The pass rate was lowest for students aged 20. The pass rate then gradually increases with age, peaking for students between 47 - 50. Students younger than 36 were slightly more likely to fail than pass whereas students older than 36 were slightly more likely to pass than fail.

Fig 2 - Pass rate by language (0-1)

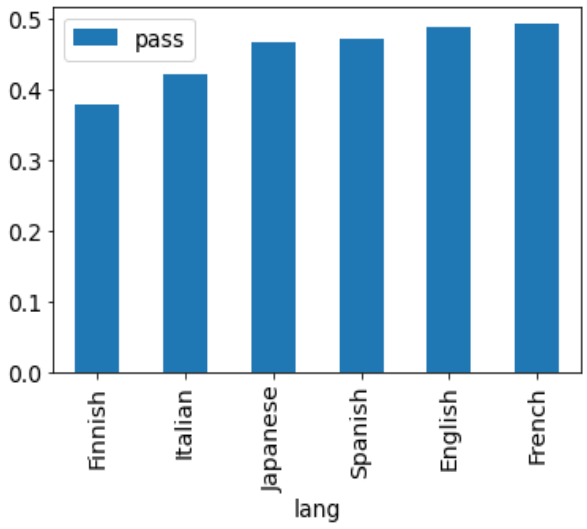


Fig 3 - Pass rate by country (0-1)

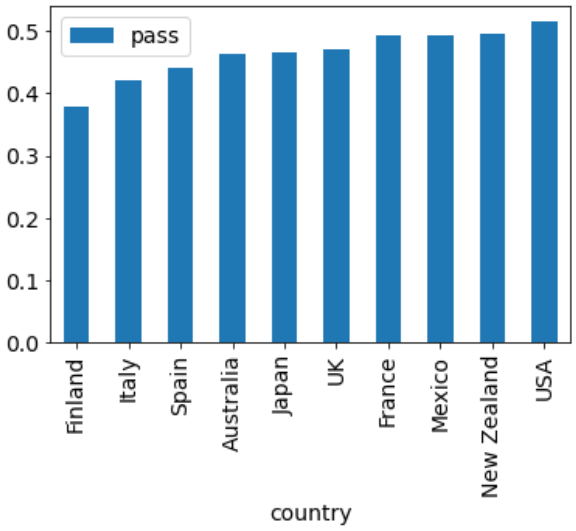


Fig 2 demonstrates the likelihood of passing depending on language, the range of pass rates were between 0.38 and 0.49. The lowest pass rate was Finnish and the highest was French.

Fig 3 demonstrates the likelihood of passing depending on country, the range of pass rates were between 0.38 and 0.51. The lowest pass rate was Finland and the highest was the USA.

Part 1.2 Intervention Efficacy

Summary

At face value the dojo class appears to be a significantly better intervention than the test preparation class at increasing the pass rate of students. Nevertheless, further data collection should be made to identify clear causality.

Fig 4 - Pass rate by dojo attendance Fig 5 - Pass rate by test_prep attendance

Blue = Pass, Orange = Fail

Attendance = 1.0, Absence = 0.00

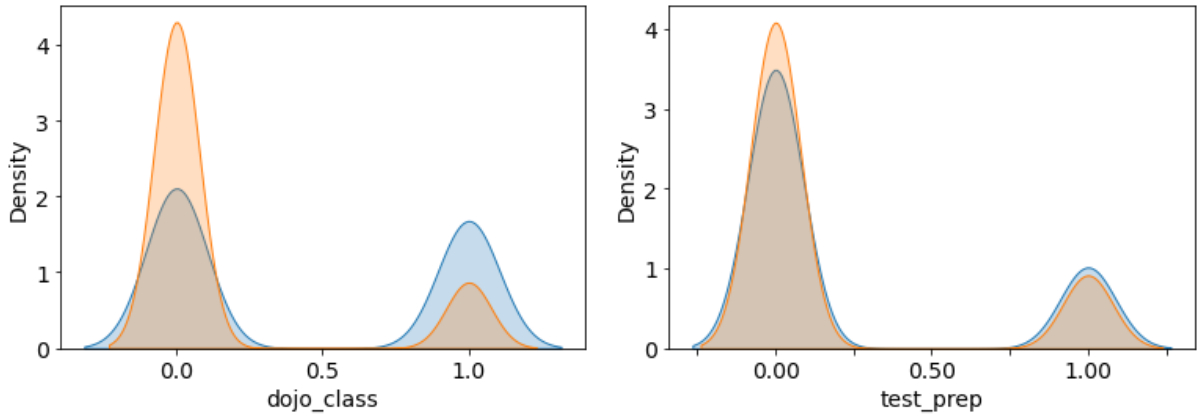


Fig 4 shows the attendance of the dojo class, for those that passed the test and those that failed. We can see that the majority of students did not attend the dojo class and that most of these absentees failed the test. Nevertheless, the majority of students that did attend the dojo class passed the test. This indicates that the dojo class was a highly effective intervention in increasing the pass rate of students. However, we need to consider the possible impact of extraneous variables. It is possible that only students who were already high performers attended the dojo class, this may explain why dojo attendees were more likely to pass. This could only be confirmed by gathering additional data, such as questionnaires to the students or a control group.

Fig 5 shows the attendance of the test prep class, for those that passed the test and those that failed. We can see that the majority of students did not attend the test prep class and that slightly more of these absentees failed than passed. For those that did attend the test prep class, there was an approximately equal amount of passes and fails. This is a slightly higher pass rate than for the absentees which suggests that the test prep class has a small positive impact on the pass rate.

Part 1.3 Interesting trends

Fig 6 Heatmap of correlations between variables

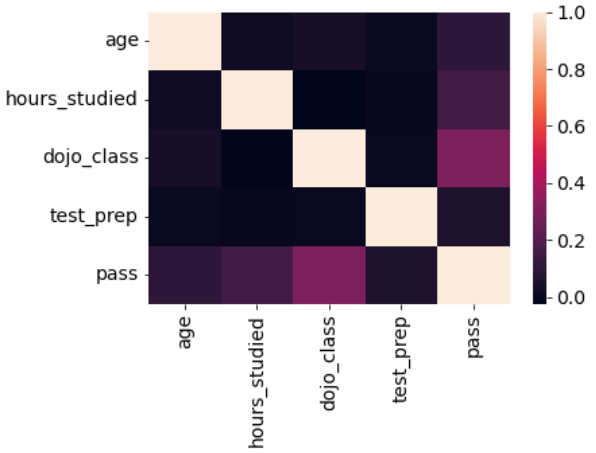


Fig 7 Student age vs hours studied

Fig 6 shows us that most variables are not correlated. The highest correlated variables are as follows (descending order):

1. Dojo Class attendance with Pass Rate
2. Hours Studied with Pass Rate
3. Age with Pass Rate

These would appear to be my primary features influencing Pass Rate. I will investigate this further with permutation importance and on the basis of this I may be able to reduce the dimensionality of the data that I use for my model. In addition to these I will also investigate the role of the test prep course in relation to pass rate, as the challenge instructs. I will pay particular attention to these features during data pre-processing and model tuning.

Fig 7 plots student age against hours studied. It demonstrates that older students were more likely to study for longer periods than younger students. This indicates why older people are more likely to pass than younger people. Nevertheless, the correlation between age and hours studied is not very strong so we cannot infer clear pass causality based on age or hours studied. Although correlated features can sometimes mask interactions between features when building machine learning models, I have decided not to remove either of these variables because their correlation is not strong enough to warrant concern.

Part 2: Technical Explanation

Feature Imputation

- Hours studied was the only relevant column with large amounts of data missing. I decided to impute this data because it was clearly highly correlated with the pass rate and as such was too important to drop. I initially used the median to impute values because the data has a large variance and some skew. Nevertheless, I eventually found that imputing the mean offered the greatest accuracy level.

Scaling

- I decided to use the Robust Scaler for my numerical values because there was a large amount of variance, particularly in the number of hours studied. The Robust Scaler stops the outliers influencing my model too heavily.

Encoding

- I found that by One Hot Encoding my country and language variables I was able to get the highest model accuracy values. These were higher than if I was to remove the values altogether.

Feature Selection

- I used permutation importance to understand which features had the greatest importance to the target. When tuning my models later on I tried multiple variations based on this. Nevertheless, I got my highest accuracy levels from using my full feature set.

Algorithm

- I tested a number of classification algorithms such as Logistic Regression, KNeighborsClassifier, RandomForestClassifier and XGBoostClassifier. Although baseline scores for XGBoost were higher than for RandomForest, I found that tuning RandomForest had the highest accuracy.

Model Testing and Tuning

- I initially tried adjusting the max_features and n_estimators manually to see if I could improve my accuracy score. This had very little effect and eventually decided to use a RandomSearchCV to check the accuracy of my model. I offered a wide range of hyper-parameter variation for the search and eventually found that the best estimator this found was higher than any other I had received.

Prediction

- I evaluated the model a second time by fitting the model with the training data and then predicting the X_test and then comparing the outputted predictions with Y_test using rmse.
- I also used the model for predictions by manually inputting values into the main function of the predict_model.py file.

Using the model

- If I was going to deploy this model I would enhance the input and output functionality so that data could be inputted through the command line, an API or web app so that we could use the model efficiently. Nevertheless, as this challenge only requested the model's evaluation, this is all that I print to the command line when the program is run.

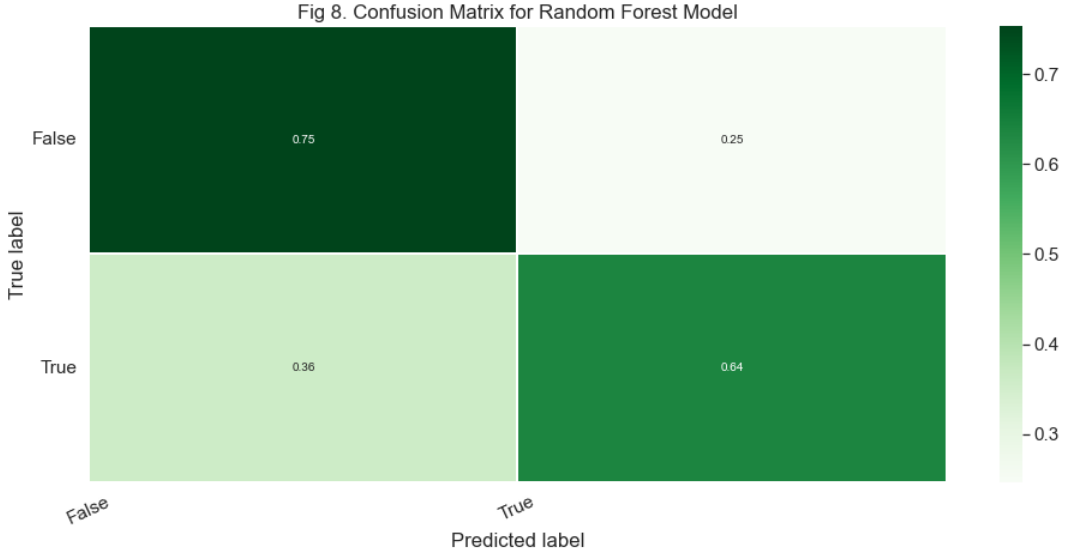
Refactoring

- I refactored my model from a notebook and into an object oriented folder structure so that the model is portable, cross compatible and can be installed as a library or run with a Makefile.

Part 3: Reporting

Part 3.1: Create visualizations to show the efficacy of your model.

Fig 8. Confusion Matrix for Random Forest Model



The above confusion matrix (Fig 8) expresses how many of my classifier's predictions were correct, and when incorrect, where these mistakes were made.

- More true negatives than false positives.
- 36% chance of predicting someone will fail when they will actually pass
- 25% chance of predicting someone will pass when actually they will fail
- We may not identify some students who need additional help.
- Low recall is a major weakness with my model

Fig 9. Classification report for test data and prediction

	precision	recall	f1-score	support
False	0.71	0.75	0.73	681
True	0.68	0.64	0.66	569
accuracy			0.70	1250
macro avg	0.70	0.70	0.70	1250
weighted avg	0.70	0.70	0.70	1250

- The model performs better at identifying students who will not pass (False) than students who will pass (True).
- Recall for the "False" class is particularly high, the model avoids classifying Fails as Passes.

- This is a risky mistake so it is good that of all of the metrics we have here this is the best performing.

Part 3.2.1: How we might help more people pass the test

- The lowest hanging fruit is to increase the rate of students doing the dojo
- If resources are a concern we can offer separate our students into groups based on necessity
    - The people who need help the most can be prioritised for the dojo:
        - Youngest students should be given priority because they have the lowest pass rates.
        - Students who have studied the least hours should also be given priority as these have lowest pass rates.
- Another intervention would be to encourage students to increase the number of hours they studied asd these has a clear correlation with increased pass rates.

Part 3.2.2: How we might create more accurate models based on your findings.

- Gathering more data could significantly improve the model. Our current model does not take into account student ability or past performance in whether they pass. If we had an idea of their current academic score or their history of passes and fails we would be able to isolate the impact of the dojo, the test prep and the hours studied more clearly.
- Existing data could also be improved. The language and country data is insufficient to make sense of certain edge cases as the vast majority of observations are Japanese tests from Japan. Other demographic data is insufficient. If we had more data we may be able to find interactions from the countries and languages which currently do not have much data. A lack of data is also problematic for the hours studied as this feature had a large amount of null values.
- There was a slight imbalance in classes (2665 vs 2334), I could get over this by resampling my data or by attempting to collect more.
- I may be able to model better by increasing the model complexity. I could do this by stacking ensemble methods to include logistic regression and or clustering in my stack.