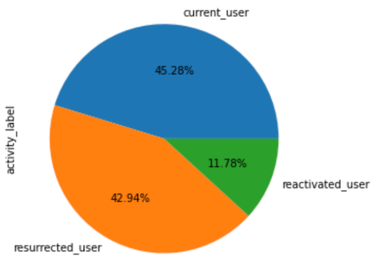
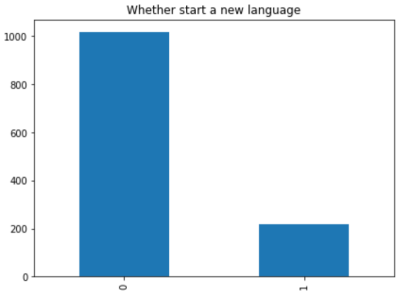
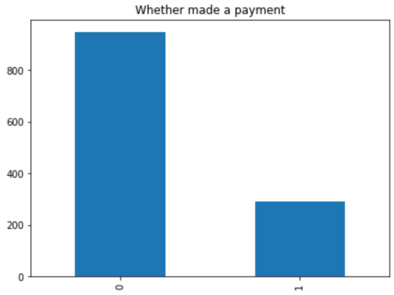
Yunbo Zhu – Duolingo Take Home Test Report

The issue of keeping users happy and satisfied is a perennial and age-old challenge. In this challenge, I used predictive modeling to predict resubscription of eligible users of Duolingo and my workflow consists of three major part:

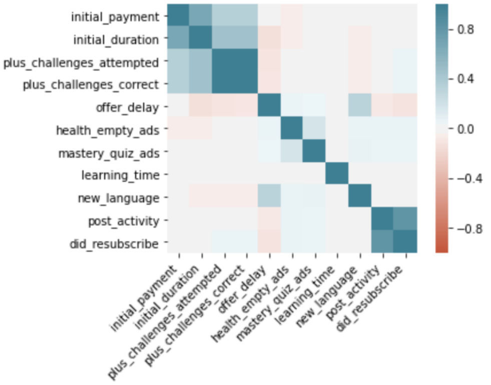
1. **Exploratory Data Analysis**: In this section, I explored the dataset by taking a look at the feature distributions, how correlated one feature is to the other and create some Seaborn and pairwise visualizations.
2. **Data preprocessing**: Handled missing values as well as encoded all categorical features into dummy variables.
3. **Implementing Machine Learning models**: I implemented and fine-tuned a Random Forest Classification Model and select features by importance, after which I examine the relationship between important features and our target resubscription.





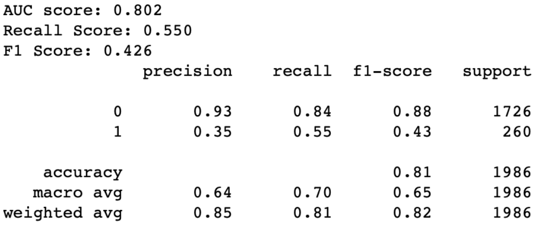


From EDA, among all the users in the dataset, over 80% did not resubscribe the second free trial and for those who did sign up for a second free trial, most of them are current users or resurrected users, 1/6 of them learn a new language and ¼ of them made a payment in initial tier. More information on initial tier tells that most of resubscribed users signed up for a 1-month or 12-month during their first free trial.



By plotting a correlation matrix, I had a very nice overview of how the features are related to one another. From the correlation heatmap, we can see that quite a lot of our columns seem to be poorly correlated with one another which is preferable when I train the predictive model. But what worth noticing is that number of challenges attempted is highly correlated with number of challenges correct. The more you attempted, the more likely you can get it correctly. Also, post activity is highly (positively) correlated with resubscription and we won't have any information of post activity if someone doesn't start a second free trial, so I exclude this feature in the prediction step.

Before implementing machine learning model, I use SMOT to oversample data due to skewness in target (about 87% not resubscribe while 13% resubscribe) and split data into training and test sets. To predict resubscription, I used Random Forest Classifier, an ensembled



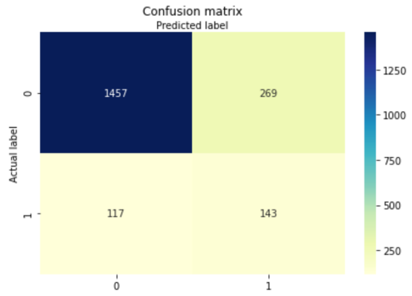
```
AUC score: 0.802
Recall Score: 0.550
F1 Score: 0.426
              precision    recall  f1-score   support

           0       0.93      0.84      0.88      1726
           1       0.35      0.55      0.43       260

    accuracy                           0.81      1986
   macro avg       0.64      0.70      0.65      1986
weighted avg       0.85      0.81      0.82      1986
```

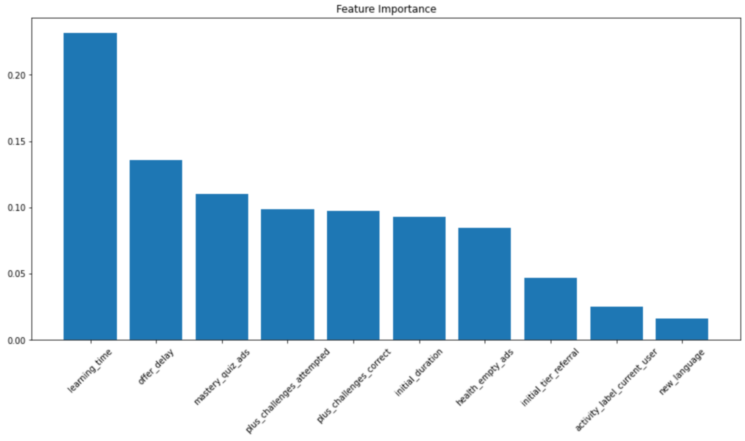
decision trees model. As our target is unbalanced, I use Recall Score as the performance measurement to tune my model by cross validation in order to find optimal model which maximizes recall (probability of correctly identifying truly resubscribed users) during training.

```
Accuracy of Random Forest Classifier on test set: 80.56%
```



As observed, our Random Forest returns an accuracy of approx. 80% for its predictions and on first glance this might seem to be a pretty decent performing model. However, when we think about how skewed our target variable, we see there's still room for improvement since test Recall is just slightly more than 55%. Comparing performances of different binary classification models like Logistic regression and XGBoost might be my suggestion. Also, it would be more informative to balance out the precision and recall scores as show in the classification report outputs. Where it falls down to the business considerations over whether one should priorities for a metric over the other - i.e. Precision vs Recall.

To tells which features within our dataset has been given most importance through the Random Forest algorithm, I selected top 10 of them by feature importance (gini importance in RF) and ranked them in order.



By observing the importance plot, we see learning time, offer delay and the mastery quiz ads are the top 3 important features that affect my model. Furthermore, by investigating relationship between features and target, I obtained that learning time and mastery quiz ads are positively correlated with resubscription while offer delay is negatively related. In other words, the more time you learned and more mastery quiz ads you see, the more likely you will resubscribe for a second Plus free trail and the longer the users waited for a second offer, the less likely they are going to resubscribe.

Recommendations:
Although correlation cannot tell causation, we at least have some clue of where to focus for driving more eligible users to resubscribe. For example, encourage longer learning time by providing reward such as a second offer when users reach certain level of learning.With potential inference on features, we need to conduct A/B Test to further prove the causation in order to make product feature improvements. Also, user activities could change over time. With potential improvements in product features, we may check on a regular basis to adjust the model and therefore to complete a full development circle.