Duolingo Take Home Task

## Ujjawal Madan

Business Objective:

Maximize the resubscription rate by understanding whether and why a resubscription-user will sign up for a second subscription.

## Dataset and Methodology:

Data Exploration:

* **Process:** Using Plotly, I was able to visualize the different variables in the given dataset, understand how the distributions/visualizations compare to one another when grouped by those who resubscribed and those who did not, and also create new variables using existing features to better understand the data.
* **Observations and results:**
  + The *post\_activity* variable is not information we would have during the implementation of our model and as such should not be utilized in our predictive model *– See figure 1.1*
  + New variables revealed inaccuracies within the data which should be investigated. *See Figure 1.2*

Model Prediction and Evaluation

* **Process:** The first step was to preprocess and prepare the data. After which, I then created 4 unique models which I ran on 3 different modified training datasets (totalling 12 distinct models). I then evaluated such models by plotting the TPR-FPR Curve for each.
* **Observations and results:**
  + The choice to train on imbalanced data, undersampled data or oversampled data does not significantly alter model performance. – *See figure 2.1*
  + The best performing model is XGBoost with an AUC Score of 0.80 (TPR-FPR Curve). The model threshold can be configured to meet business requirements. For example, if the minimum TPR needs to be 0.80, then the threshold will be adjusted to meet that requirement and, in this instance, FPR will be 0.3 – *See figure 2.1*
  + Further hyperparameter tuning will help optimize XGBoost Model.

Variable Importance and Insights:

* **Process:** The Random Forest and Gradient Boosting algorithms allow me to sort features by variable importance *– See figure 3.1*

**Product Insights:**

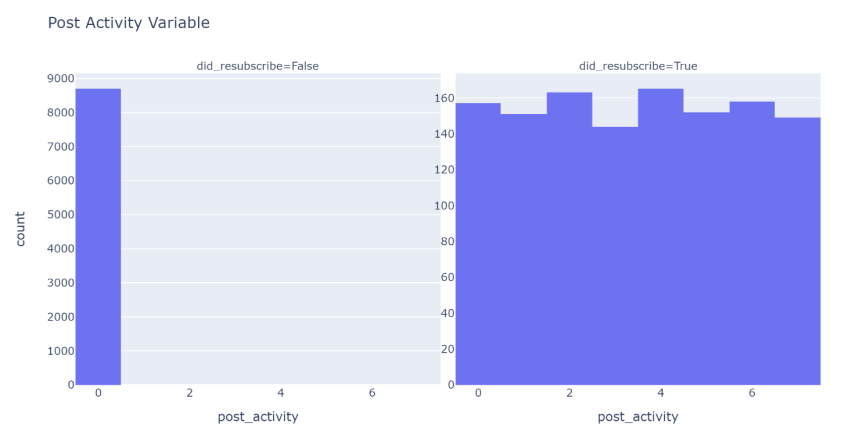
* + *Learning\_time* is the most influential feature based on our analysis accounting for 30% feature importance. The more time customers spend on average per day, the more likely they are to resubscribe. - *See figure 3.1*
  + The longer the initial subscription, the higher the probability the customer will resubscribe. Given by one estimate to have 20% feature importance.
  + The greater the number of questions answered or answered correctly, the higher the probability the customer will resubscribe. Given by one estimate to have 10% feature importance.
  + Customers who are introduced to the app by a referral have the lowest resubscription rate.
* **Key Insights:**
  + Three features of great importance which Duolingo has control over include *offer\_delay, mastery\_quiz\_ads and health\_empty\_ads* (they can choose exactly the number of ads to show, when to offer a second trial etc.)
  + We can use our existing machine learning model to find the optimal value for *offer\_delay, mastery\_quiz\_ads,* and *health\_empty\_ads* given a data point. Following which, we can employ a form of A/B Testing or a Multi-Armed Bandit Algorithm to test out our predictions and then modify our model to be more accurate. – *See figure 3.2*

Recommendations:

Given our objective to maximize the resubscription rate:

* Design, research and implement a form of A/B Testing which tests our predicted optimal *offer\_delay* value
* Define costs of running *mastery\_quiz* ads - If the cost of running the ads is less than the value of gaining a resubscription; design, research and implement an A/B test which experiments with our predicted optimal mastery\_quiz\_ads. Below is a simplified example:
  + The cost of running a *mastery\_quiz* ad is $1.
  + We have identified a segment of customers of whom 35% will resubscribe if shown the ad.
  + The acquisition cost of a resubscribe-eligible customer at this time is $5.
  + In this instance, it makes sense to run the increased number of ads given our costs.
* Increasing average time spent on the app should ultimately be given the highest priority as it is the most influential feature.

**Appendix:**

Figure 1.1 Post Activity

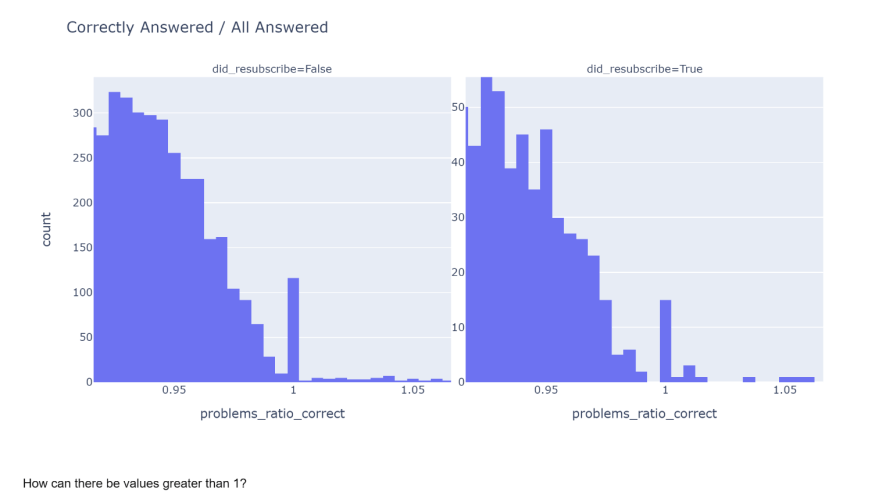
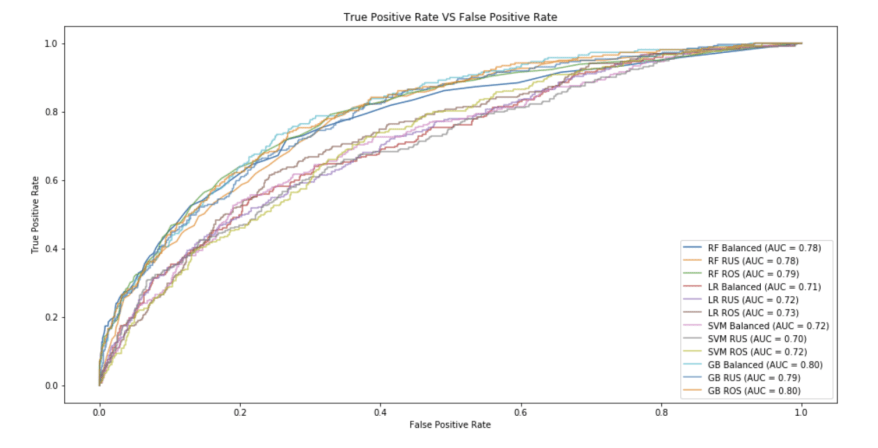
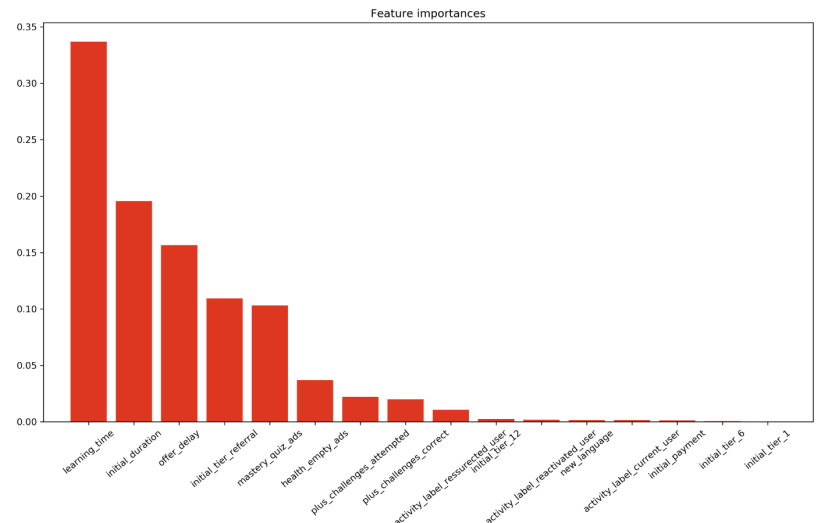
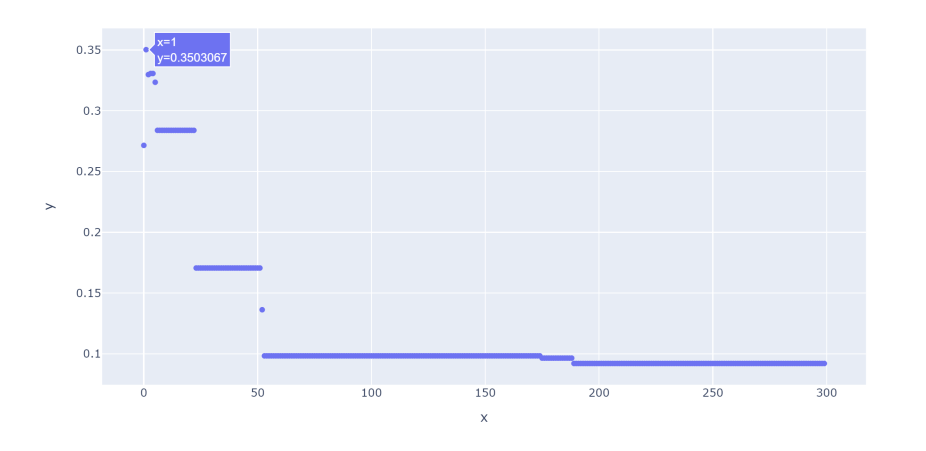
****Figure 1.2 Correctly Answered / All Answered

Figure 2.1 ****TPR VS FPR Curve

*Figure 3.1 Feature Importance (Random Forest)*

*Figure 3.2 Optimizing for Offer Delay*