Data Cleaning

All data cleaning and visualization was done in **R**, as that is the language I'm most comfortable exploring data in.

survey_data.csv

The survey data itself required relatively light cleaning, mainly just creating a category for missing data. Every question had some missing answers, so the way that this was dealt with was adding a new category called "No Response". This helped in visualizing the data and will also help when modeling it. The field `future_contact` had some missing values that were defaulted to "No" as to avoid harassing customers. The `other_resources` category was turned from a list into a number representing how many other resources they used, as this was more interpretable than the various categories. The only numeric column in the Survey data was `time_spent_seconds`, which had a number of outliers. These outliers (negatives times and completion times of up to a week) were identified and replaced by the median value (due to a non-Normal distribution).

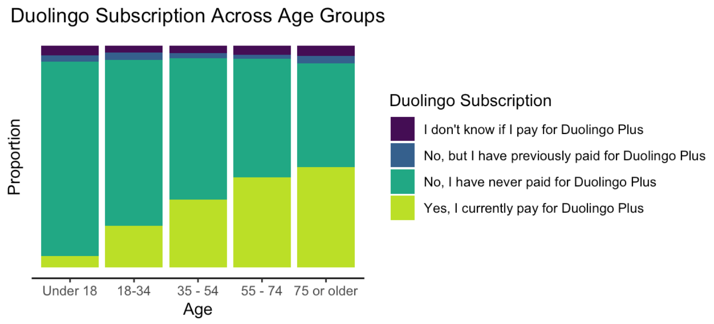
There were a number of duplicate surveys (around 74, or 0.1% of the data), and deduplication was not trivial—some surveys were drastically different from others, *despite* coming from the same user. I imagine this issue may arise from users sharing an account (only likely if it's a subscription account) or, more likely, a single user accidentally starting the survey and exiting, then later deciding to take it for real. For this latter "reason", I resolved the duplicate `user_id`'s by choosing the entries that had the longer completion time and removing the other.

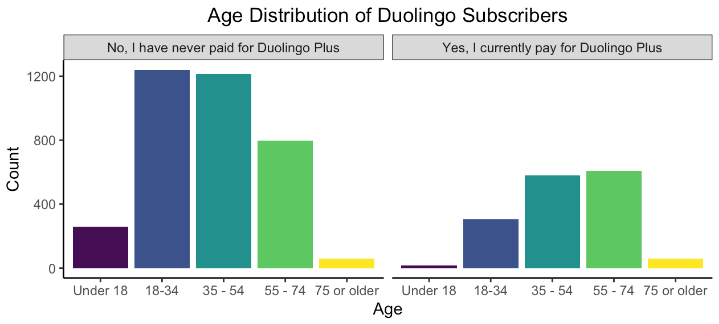
survey_users_app_usage.csv

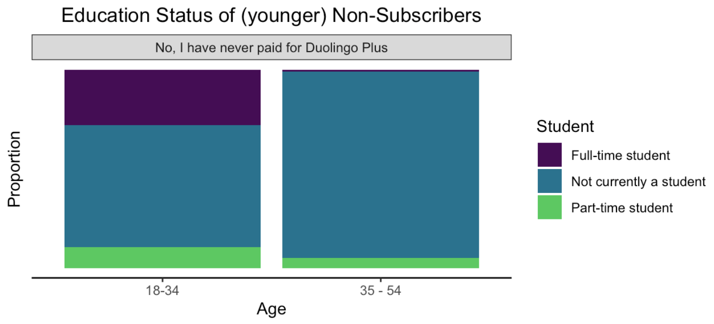
The usage statistics had many more inconsistencies. There were more erroneous and missing values, such as users having "streaks" that lasted 16 years (despite Duolingo being founded in 2011) and users making negative lesson progress. Erroneous and missing values were treated by replacing them with a sensible value, whether it was the median or some lower bound (i.e., `highest_crown_count` being set to 0). I decided to actually **drop** the `daily_goal` variable, as more than half of all users were missing it, and it would have likely confounded the analysis more than helped.

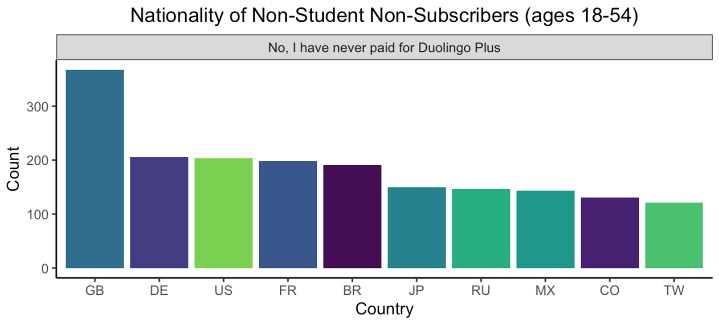
There were around 75 duplicate usage entries, which was quite odd, as one would have thought a single user could not have multiple `duolingo_start_date`'s or `lessons_completed`. These duplicates were resolved by preferring the usage entry with the most lessons completed. Both de-duplication decisions were heuristically made after manually viewing the duplications (lines 117 and 124 in `duolingo_analysis.r`).

Visual Exploration


Duolingo Subscription Across Age Groups


Age Distribution of Duolingo Subscribers


Education Status of (younger) Non-Subscribers


Nationality of Non-Student Non-Subscribers (ages 18-54)

Through visualizations I honed in on how well Duolingo fares with different audiences. From the upper left plot, it is clear that a subscription to the language learning application is increasingly popular as user age increases; however, the user base is primarily millenials and younger Gen X'ers. Users least likely to subscribe to Duolingo were young, low-income students from countries where English-speaking is widespread, which presents us with a target audience. Duolingo should try and incentivize 18-54 year olds in English-speaking countries to try their application, and then keep them active with the "crown" microcurrency in addition to highly incentivizing the maintenance of a streak, as those statistics are correlated with subscriptions. Of course, these users need a real incentive to actually *start* using the app.

Users who were learning from multiple outside resources were also more likely to purchase a subscription, so a good start is providing a good reason to begin learning in the first place. From looking over the `other_resources` column, it's evident that travel is a large motivator. Duolingo could perhaps organize collaborations with travel companies such that low-income users could get discounts if they happened to achieve a moderate amount of activity over a long period of time. This would have a ripple effect, as friends would wish to travel together and thus learn together (another motivator in itself!).

Statistical Exploration

I performed statistical analysis in both **R** and **Python**, as I'm more familiar with statistical analysis in the former, and machine learning and modeling in the latter. I generated correlation scatter matrices and plots and found that there was a subset of variables that correlated relatively strongly with one another, e.g., `highest_course_progress, highest_crown_count, n_active_days`. They were primarily activity indicators, and not overall novel or surprising in any way.

I used *mutual information* and *sequential feature selection* in order to rank individual variables as well as subsets of variables. After determining a subset of meaningful variables, I built two boosted models (*RandomForeset & AdaBoost*) that had an F1-score of around 62%, although the precision and accuracy were 70%+ (not exactly impressive, given class imbalance within the data). Despite this, the trained model presented two variables as having high feature importance: `n_lessons_started, longest_streak` (table to lower right). This echoes the importance of incentivizing daily activity from users (and given the Duolingo Owl's threatening emails, he knows this).

Cluster Analysis

I used hierarchical clustering on the full training data in order get a sense for user personas. I selected my cut-off to create 3 clusters, as that would provide more insight than the presumable "active" and "inactive" split that 2 clusters would show.

CLUSTER ONE:

- Long time users — this is by far the biggest distinction of this cluster
- Most likely to have a subscription! (40%)
- Very dedicated users - much longer streaks, utilizing more outside resources
- Middle-aged with disposable income (most likely to be employed full-time)
- Disproportionately from countries where English is relatively uncommon:
  - Russia, Japan, Taiwan

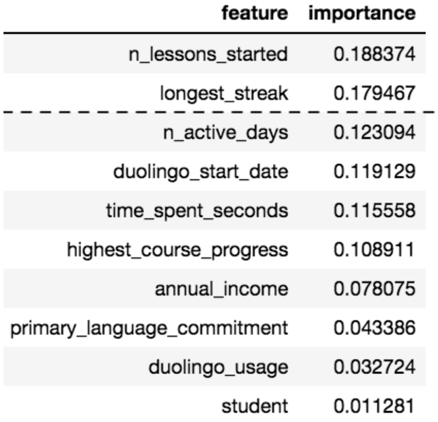
CLUSTER TWO:

- Largest Cluster (majority of users)
- Disproportionately low-income students (least likely to purchase subscription)
- Do not frequently use Duolingo (low overall activity)
- Disproportionately from countries where English is common
  - Germany, US, Great Britain
- Generally younger, not working and little disposable income

CLUSTER THREE:

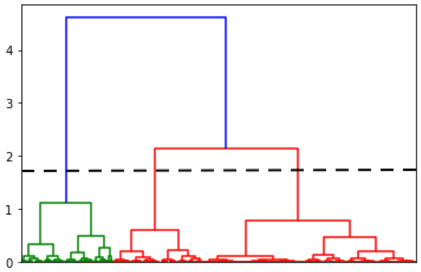
- Casual users (more active than student, less active than the dedicated users)
- From an even mix of English and non-English speaking countries
- Least likely to have taken the placement test
- Wide distribution of ages
- Seems to be everyone who doesn't fit well into cluster's one or two

RandomForest Feature Importances



feature	importance
n_lessons_started	0.188374
longest_streak	0.179467
n_active_days	0.123094
duolingo_start_date	0.119129
time_spent_seconds	0.115558
highest_course_progress	0.108911
annual_income	0.078075
primary_language_commitment	0.043386
duolingo_usage	0.032724
student	0.011281

Hierarchical Clustering Cut



These clusters echo the visual analysis. Cluster 2 is by far the largest (at around 3000, relative to the other clusters being 1300), and is composed of low-income millenials from English-speaking countries. If Duolingo is able to organize a collaboration with travel agencies that would allow these users to access the same opportunities as cluster 1 (with disposable income), that could create a larger volume of daily users and increase subscription rates. Thanks for reading!