Duolingo User Persona Analysis and Subscription Prediction

Shiyu Liu
Data Science Initiative, Brown University
shiyuliu001@gmail.com

1 Introduction

Along the globalization of business and education as well as the development of technology, the demand of learning a new language online has surged during the past decade. In addition to extra time, more people are willing to invest money in learning in pursuit of a better experience and a more convenient learning environment. The aim of this study is to analyze the personas of the users who are more willing to subscribe pay-for-use plans on a language-learning platform, Duolingo. Taking usage data and user demographic as known information, we are able to investigate the users characteristics who are willing to subscribe for pay-for-use plans. Two data science pipelines are developed to predict the subscription intention of the users. One of them focused on predicting whether an existing user have the intention of subscribing by taking both demographics and platform usage data. The second pipeline focuses on unregistered individuals that are likely to subscribe pay-for-use plans for better service on the language-learning platform. Since it is impossible to have the usage data for potential customers, we only take the accessible demographic data only for the second data science.

2 Data and Preprocessing

The data was collected by Duolingo based on the activity and information of its users. There are two datasets used in this study. Each dataset was first implemented with data cleaning, missing value imputation and expletory data analysis.

2.1 Survey Data

The Survey Data was collected at the end of 2018 through a survey for 6187 Duolingo users. The survey asked users a series of questions about demographics (e.g., country, age, employment status), and motivation (e.g., primary reason for studying a language). The responses of the questions are on a voluntary basis. For the missing values caused by such unanswered sections, a new feature category named as "Not Specified" was filled.

2.2 Usage Data

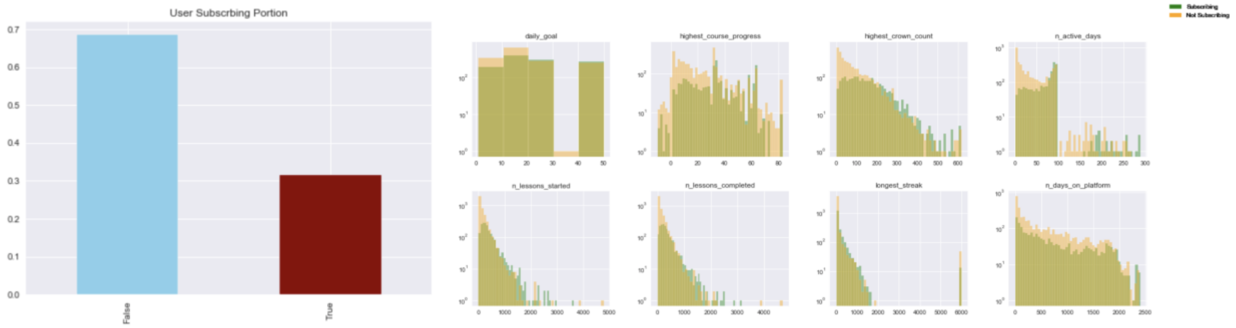


Figure 1. The left panels shows the proportion of each class in the target variable, indicating an imbalanced dataset; The right panel shows the two target distributions based on numerical features in Usage Data. It is hard to find big difference between the two distributions based on single value.

Collected from the users who were included in the survey, the Usage Data reflects the user activity and achievement from August 1, 2018 to November 5. Missing values were imputed using multivariate imputation with a method of random forest regressor. As shown in Figure 1, the subscription status, which will be used as target variable later, has two classes: while approximately 69% of the users were not subscribing, approximately 31% of them purchased Duolingo Plus. We can only qualitatively find slight differences of distributions of the subscription status by single numerical features.

3 Method

3.1 Target Variable

The two datasets were first merged on user ID In order to simultaneously consider their usage and demographical information. The target variable used in the study is the subscription status during the usage data collection period.

3.2 Features and Persona Selection

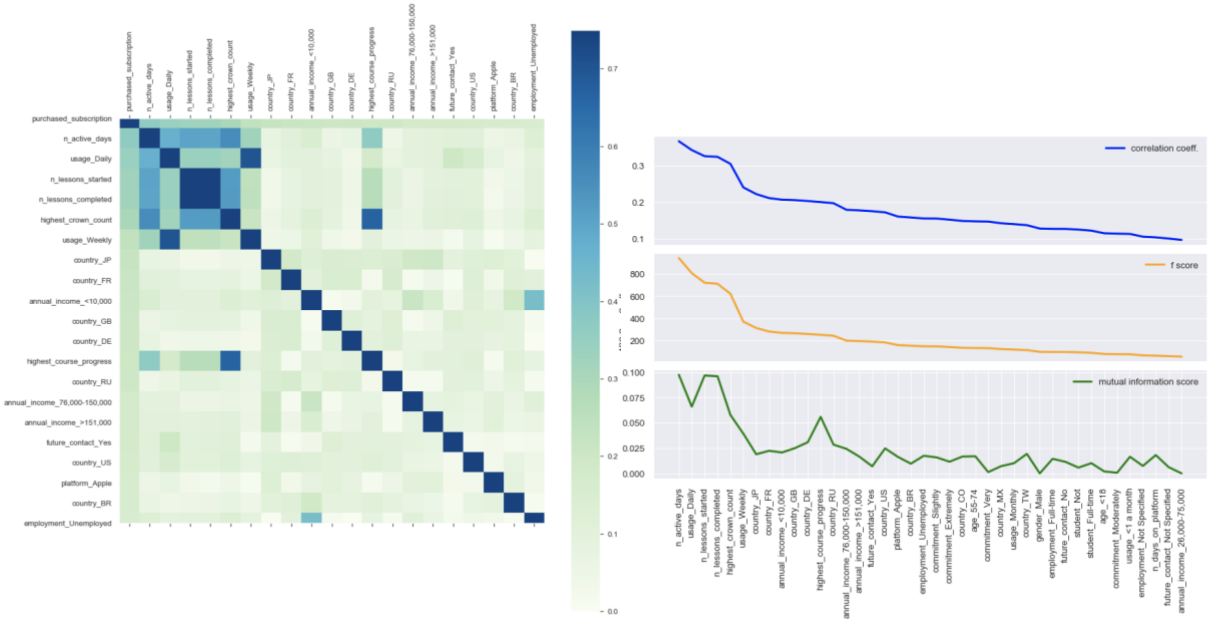


Figure 2. The trend of the three metrics measuring relationship strength between target variable(subscription) and the features. The trends are the same for f-value and mutual information score, based on the feature order according to the value of correlation coefficients. 40 personas were shown here, taking 0.05 of correlation coefficient as a threshold.

There are 85 features viable features after encoding categorical variables. To quantitatively measure the correlation between target variable and such features, the feature selection methods include correlation coefficients, ANOVA F-value and mutual information. As shown in figure 2, such three metrics show similar result for the features having top correlations with the target variables. Among the 85 features, the top 40 features were chosen and included in the following models for further investigation. Such decision is made at a correlation coefficient threshold absolute value of 0.05 between a particular feature and the target variable, subscription status. After such features are quantitatively selected as personas, a model takes them as independent variables. Numerical personas are then standardized.

3.3 Models

3.3.1 Two Models Analyzing Personas for Existing Users

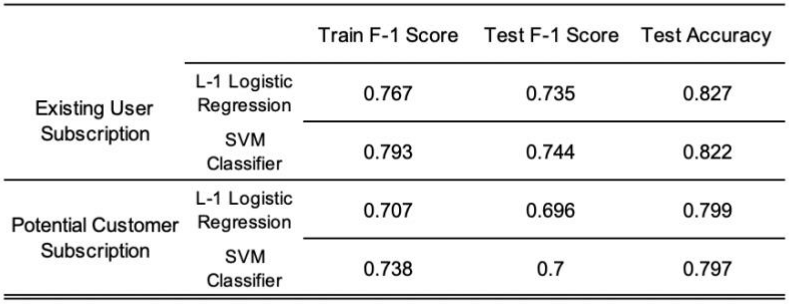
A l1-regularized logistic regression model and a support vector machine classifier(SVC) model are presented to predict the subscription. The independent variables are the top 40 features presented above, while the target is the subscription status of the corresponding users. Instead of accuracy, F1 Score is used as the metric in order to deal with this imbalanced class distribution(to avoid high false negative rate). After the regularized logistic model with best hyperparameters is selected, the estimated coefficients can show the feature importance.

3.3.2 Two Models Analyzing Personas for Potential

Since the marketing actions might be implemented on potential customers, this pipeline focused on predicting the user subscription intention based on demographic information.
Similar to the first pipeline, A l1-regularized logistic regression model and a support vector machine classifier model are presented to predict the subscription intention. The independent variables are 37 categorical features that can be acquired without Duolingo account register. F1 Score continues to be used as the metric. After the regularized logistic model with best hyperparameters is selected, the estimated coefficients can show the feature importance.

4 Result

Table 1. Model Performance Summary



		Train F-1 Score	Test F-1 Score	Test Accuracy
Existing User Subscription	L-1 Logistic Regression	0.767	0.735	0.827
	SVM Classifier	0.793	0.744	0.822
Potential Customer Subscription	L-1 Logistic Regression	0.707	0.696	0.799
	SVM Classifier	0.738	0.7	0.797

The results for the four models are shown in Table1. According to test f-1 score, it clear that there is a slightly better performance for support vector machine classifier in predicting subscription, for both existing user and potential customers.

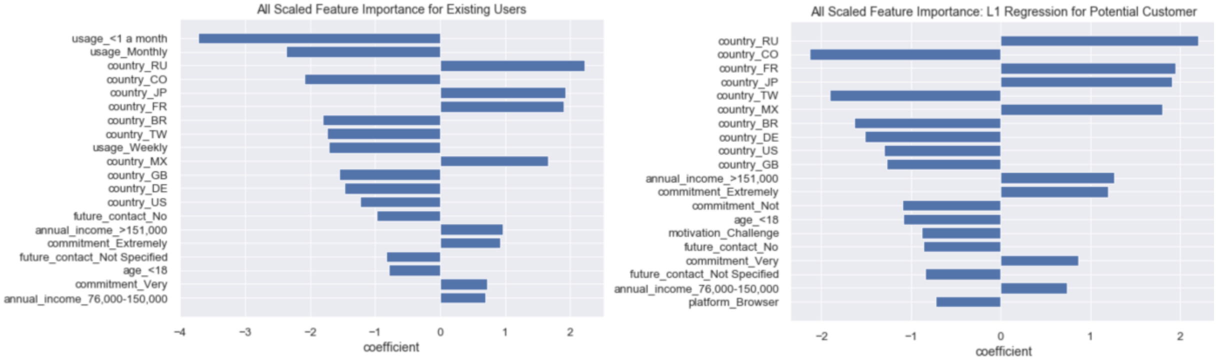


Figure 3. Feature importance for the two existing users(left panel) and potential customers(right panel). While bars to the right represent negative effects on the subscription status, bars to the right represent positive effects on the subscription status. A full version of this plot with 40 personas is attached in the appendix.

The feature importance based on model coefficients are shown in Figure 3. The features importance trend are almost the same for the two conditions by excluding the features not considered for potential customers. A large proportion of significant fs falls on country, indicating that country is a polarized factor when determining whether a user is willing to spend money for the platform. Other common personas that are significant include annual income, young-age user and commitment. Generally, an individual with higher income, higher age and strong commitment of learning primary language tends to purchase the subscription.

5 Discussion

5.1 Marketing Insight Discussion

The feature importance from our models can directly influence the marketing insight for Duolingo's product. However, the actionable marketing product can be different for existing users and potential customers. For existing users, it is convincing that there is a strong positive relationship between subscription and long-term/frequent learner. It is still unclear the causal inference between such behaviors. It is likely that users who subscribed Duolingo Plus has a higher motivation to learn more frequently and consecutively. Such customer stickiness, however, can be utilized to retain the subscribers. Potential marketing actions can be constantly launch and recommend new courses to those subscribers so that they are willing to spend more time and money on language-learning.

Country and annual income are both identified as top features/personas for the two group of people. Future marketing campaigns for existing users might benefit from conduct sales in low-subscribed countries including Taiwan, Brazil, Germany, US and UK in order to let the users experience such pay-for-use service. For potential customers, it is more reasonable to implement more Duolingo product promotion in Russia, France, Mexico and Japan. As a purpose, we expect more people will know Duolingo, start using it and have a high proportion subscribing. Compared to the sales in low-subscribed countries, this action in high-subscribed countries might surge the revenue in a short period.

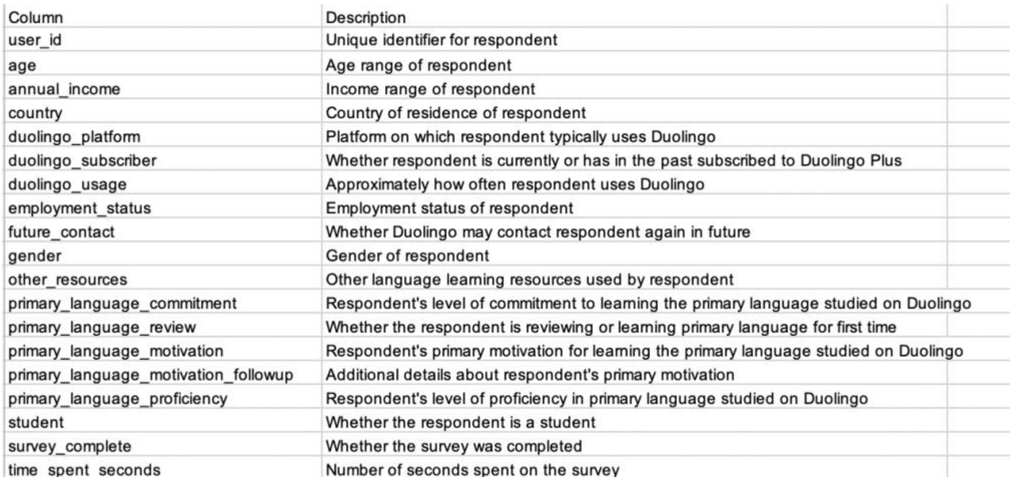
Generally, people with high annual income, older age and high commitment tend to subscribe. Such common finding among both models might further consolidate the marketing positioning of Duolingo at wealthy and senior classes.

5.2 Machine Learning Models Discussion

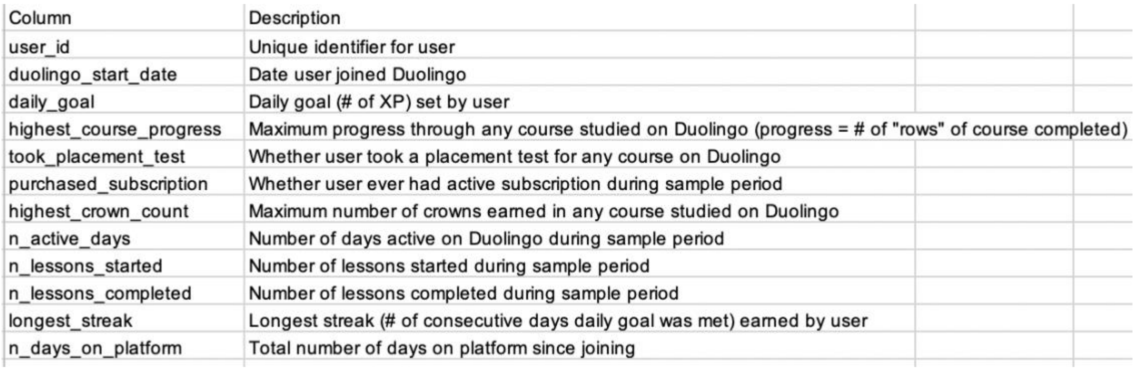
Although the machine learning models yield adequate performance in this study to predict the subscription status of users, there are huge potential improvements. First, future studies might benefit from further training the model with more intact dataset. The high proportion of missing value of both dataset can be a problem influencing accuracy. Second, a more precise of time series dataset for subscription status is necessary for product retention analysis, which is defined as the cases that continuing subscribing the service. Considering machine learning models are widely used in retention prediction, we can expect a high accuracy predicting the subscription intention by time.

Appendix

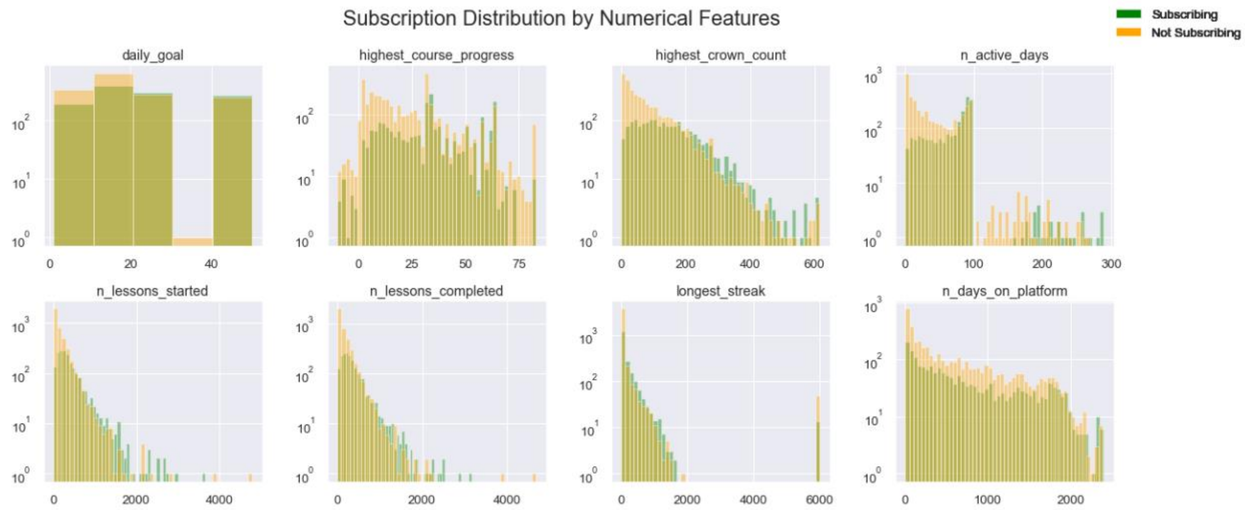
Features:



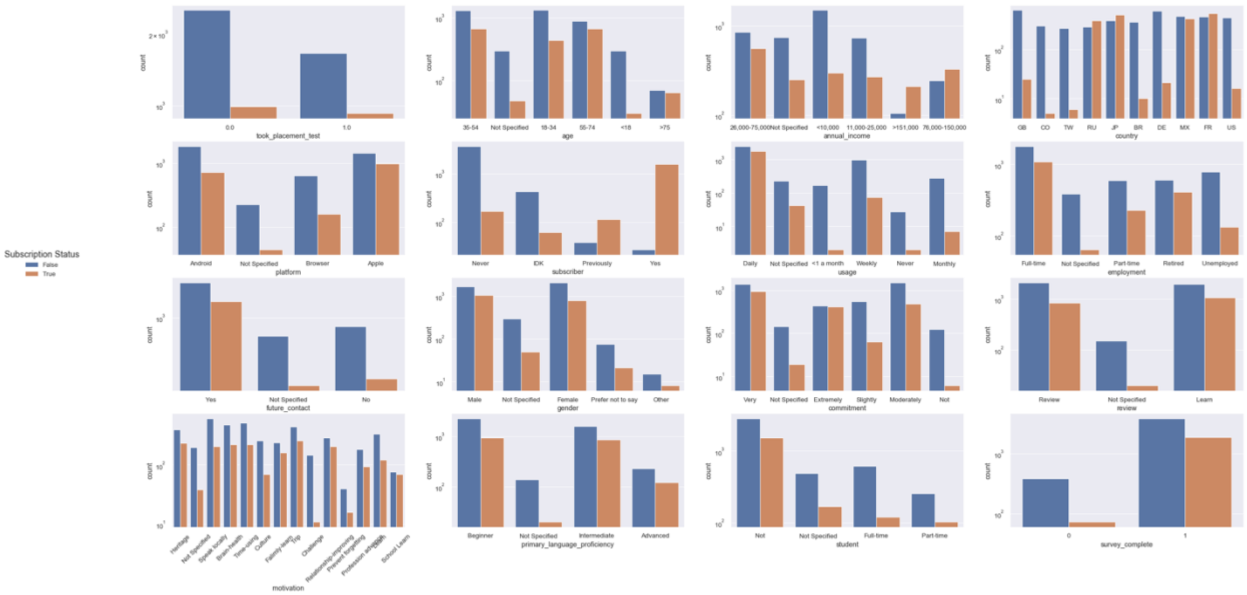
Column	Description
user_id	Unique identifier for respondent
age	Age range of respondent
annual_income	Income range of respondent
country	Country of residence of respondent
duolingo_platform	Platform on which respondent typically uses Duolingo
duolingo_subscriber	Whether respondent is currently or has in the past subscribed to Duolingo Plus
duolingo_usage	Approximately how often respondent uses Duolingo
employment_status	Employment status of respondent
future_contact	Whether Duolingo may contact respondent again in future
gender	Gender of respondent
other_resources	Other language learning resources used by respondent
primary_language_commitment	Respondent's level of commitment to learning the primary language studied on Duolingo
primary_language_review	Whether the respondent is reviewing or learning primary language for first time
primary_language_motivation	Respondent's primary motivation for learning the primary language studied on Duolingo
primary_language_motivation_followup	Additional details about respondent's primary motivation
primary_language_proficiency	Respondent's level of proficiency in primary language studied on Duolingo
student	Whether the respondent is a student
survey_complete	Whether the survey was completed
time_spent_seconds	Number of seconds spent on the survey



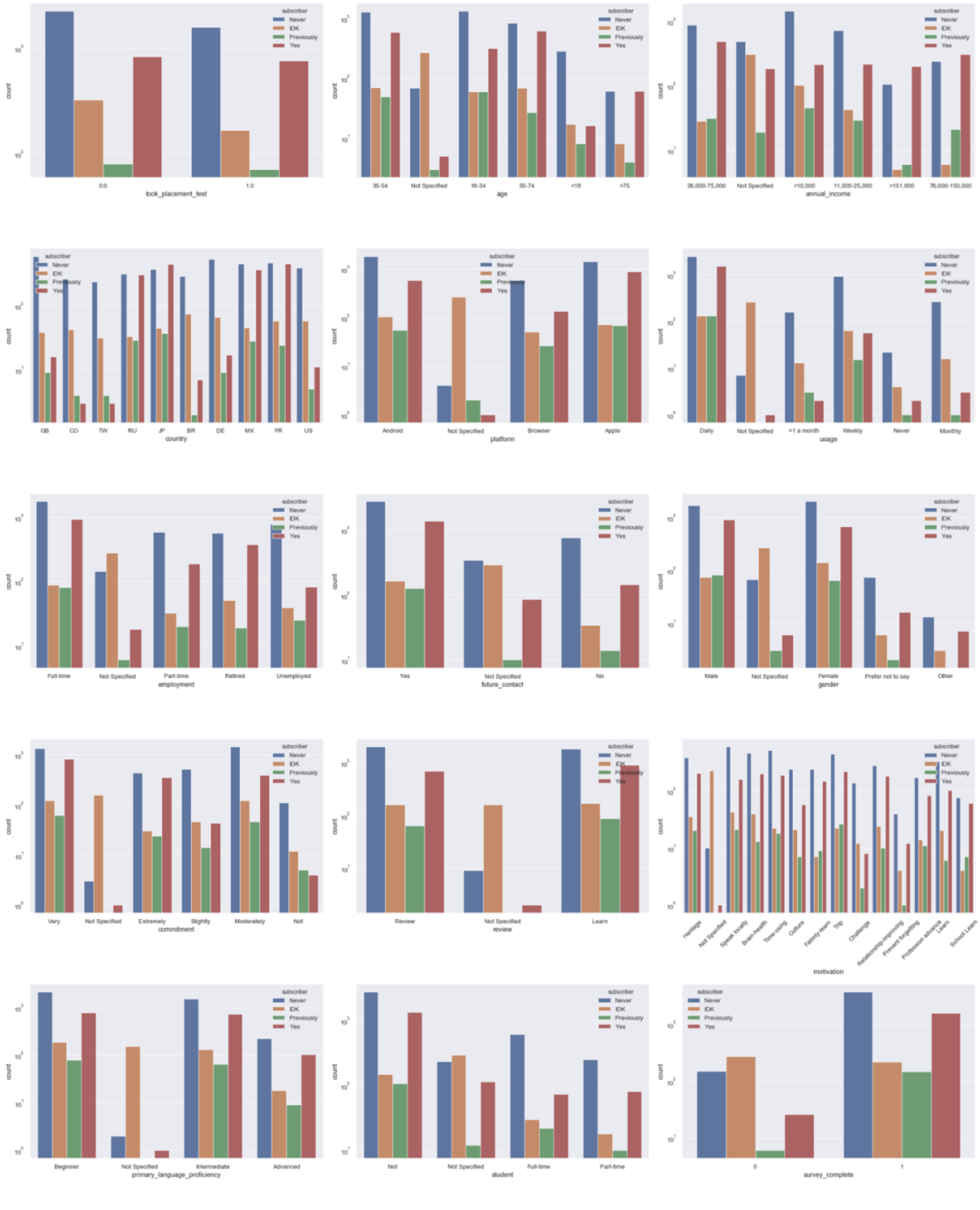
Column	Description
user_id	Unique identifier for user
duolingo_start_date	Date user joined Duolingo
daily_goal	Daily goal (# of XP) set by user
highest_course_progress	Maximum progress through any course studied on Duolingo (progress = # of "rows" of course completed)
took_placement_test	Whether user took a placement test for any course on Duolingo
purchased_subscription	Whether user ever had active subscription during sample period
highest_crown_count	Maximum number of crowns earned in any course studied on Duolingo
n_active_days	Number of days active on Duolingo during sample period
n_lessons_started	Number of lessons started during sample period
n_lessons_completed	Number of lessons completed during sample period
longest_streak	Longest streak (# of consecutive days daily goal was met) earned by user
n_days_on_platform	Total number of days on platform since joining

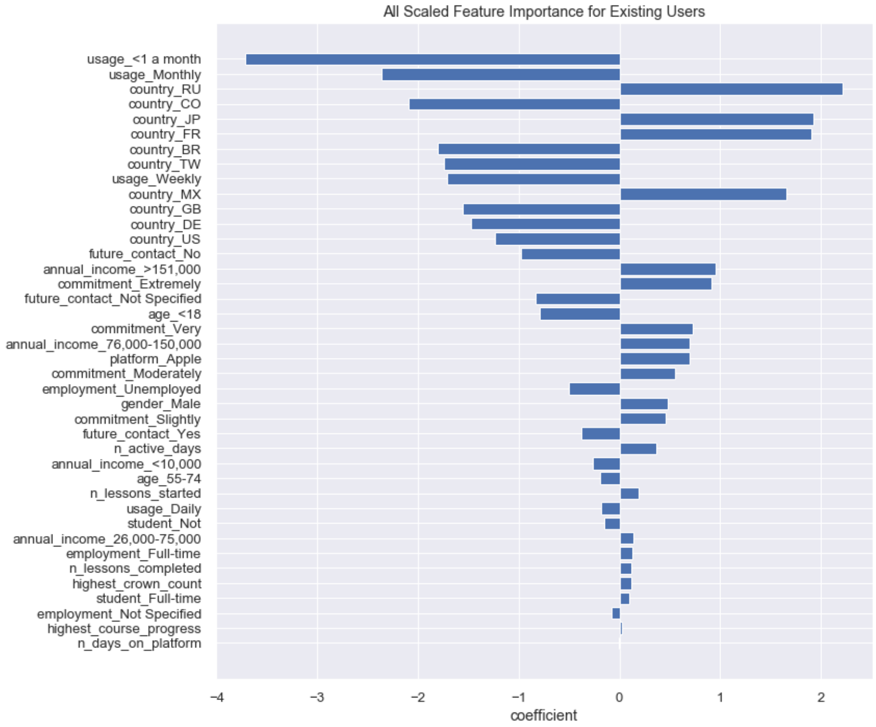

Subscription Distribution by Numerical Features

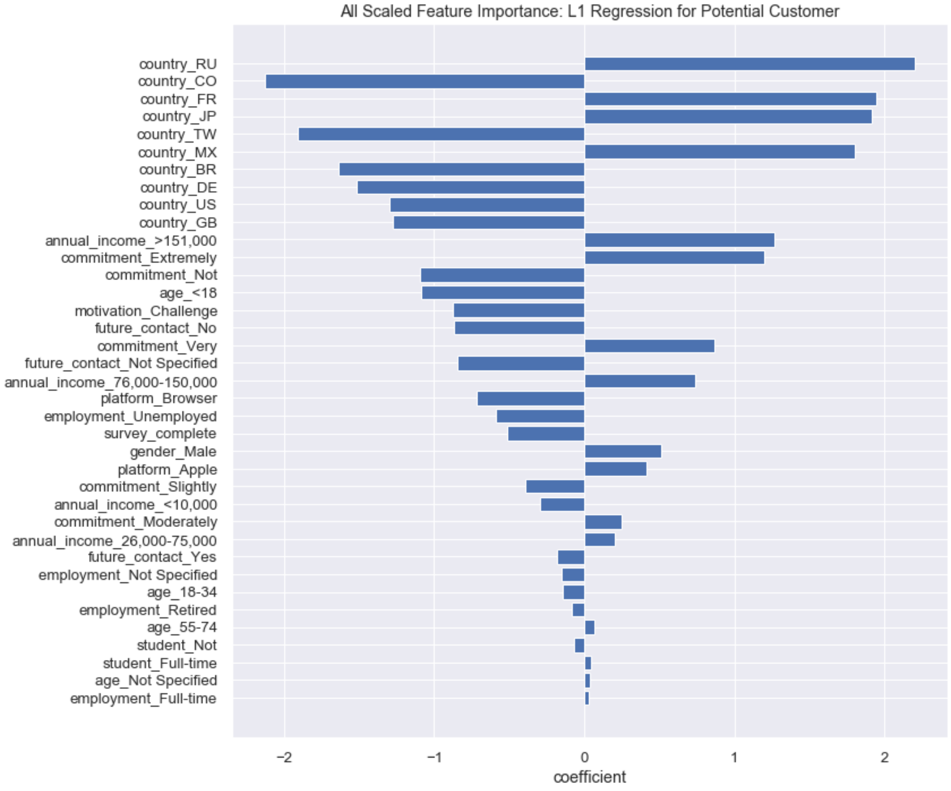
Subscription Distribution by Categotical Survey Features



Subscription History Distribution by Categotical Survey Features




All Scaled Feature Importance for Existing Users


All Scaled Feature Importance: L1 Regression for Potential Customer