Second Nature

Data Challenge

WIRED NHS Forbes 30 UNDER 30 EUROPE Roche TechCrunch

SECOND NATURE

Brief

Investigation into churn of consumer users

Background

One challenge we face at Second Nature is people cancelling (**churning**) on the programme. We want to find out what factors influence people cancelling. Every user starts the Second Nature programme on a Monday, and they go through a 12 week programme (called Core) and then progress onto a lower-cost programme (called Sustain).

Brief

Investigation into churn of consumer users

Data

You have 3 datasets (in JSON format), which are taken from 4 different collections* of data. An explanation for each of the collections and the variables contained in them can be found later in the document.

The data is from x users who started on 16th, 23rd, and 30th May 2019. Users' messages and events are given for the first 6 weeks of the programme, with the response variable being whether they churned after the 6 week mark.

We have excluded any users who churned before the week 6 mark.

There are 10 users who have the value of NA for the variable "churnedAfterSix", these are the users we want you to make predictions for.

*collections is just the term for tables in our database type, MongoDB, if you want to find out more you can research it, but that shouldn't be necessary for this challenge.

Variable Descriptions

Investigation into churn of consumer users

MESSAGES	Each object is **1 message** sent by a user
user	Each user has a unique ID that is consistent across each table
messageType	Whether the message was sent in **group** or **private** chat (private chat is 1:1 with their mentor, group chat is in a group of other users and their mentor)
weekNumber	This variable says which week the user was in when they sent the message
sentiment	The calculated sentiment of the text sent by the message. We currently use the sentimentR package.
questionsAsked	The number of questions the user asks in the text of the message
emojisUsed	The number of emojis they used in the message
mentionedScales	The number of times they mentioned the word "scales" in the message
mentionedTracker	The number of times they mentioned the word tracker in the message

EVENTS	Each object is **1 event** performed by a user
user	Distinct User ID
weekNumber	This variable says which week the user was in when they did the event
title	The name of the event that they have performed

DEMOGRAPHICS	Information the user gives us when they sign up to the programme. Quiz flow can be found here Each object is **1 user**
_id	Distinct User ID
motivation	From first question in quiz flow
challenge	From second question in quiz flow
trigger	From third question in quiz flow
goalsMotivation	Fourth question
gender, age	Self explanatory
height, weight	Height (cm) & weight (kg)
churnedAfterSix	Whether or not the user churned after week six

Task

Investigation into churn of consumer users

Task

1. **Identify** variables which exhibit significant correlation with whether or not the user churns in core
2. **Prepare** a short report / presentation to communicate these differences to a non data savvy colleague. Must be a standalone document (i.e. no verbal presentation required)
3. **Predict** whether the 10 users who do not have “NA” for the variable “churnedAfterSix” will churn or not. (do not worry about the accuracy of your predictions, just looking for the process you follow).

Task

Investigation into churn of consumer users

What to submit

1. Please submit any and all code you used to your GitHub account and share with us - contact if unsure how to do this. ***(See next page for more detail. on code guidelines)***
2. Short report – Can be a markdown, pdf, word, powerpoint presentation. **(*Please include your name in the name of the file*)**
3. Predictions – Can be included in your report or submitted separately in any format you see fit.

For github, share with @boz-sn. If all other submissions could be attached in a reply to this email.

Task guidelines

1. Remember to answer ***all*** 3 questions asked in the task
2. Please use only **python** as your coding language and present code in a **jupyter notebook -** preferably this would be one notebook but, if multiple notebooks are needed, please make clear how they fit together in your README
3. For ***part 3***:
   - If you would like to use a Machine Learning algorithm please use either the ***LogisticRegression*** or ***RandomForestClassifiers*** from the scikit-learn package and any other elements you need from this package to make these work
   - Do not spend a large amount of time optimising (for example with Grid Search). You can simply explain what improvements you could make in your notebook
4. Treat the code as if it were going to be read by someone unfamiliar with exactly what you're doing or with the modelling techniques you are applying - i.e. explain what you're doing

Second Nature

Good luck!

WIRED NHS Forbes 30 UNDER 30 EUROPE Roche TechCrunch