

Project description

The gym chain Model Fitness is developing a customer interaction strategy based on analytical data.

One of the most common problems gyms and other services face is customer churn. How do you know if a customer is no longer with you? You can calculate churn based on people who get rid of their accounts or don't renew their contracts. However, sometimes it's not obvious that a client has left: they may walk out on tiptoes.

Churn indicators vary from field to field. If a user buys from an online store rarely but regularly, you can't say they're a runaway. But if for two weeks they haven't opened a channel that's updated daily, that's a reason to worry: your follower might have gotten bored and left you.

For a gym, it makes sense to say a customer has left if they don't come for a month. Of course, it's possible they're in Cancun and will resume their visits when they return, but that's not a typical case. Usually, if a customer joins, comes a few times, then disappears, they're unlikely to come back.

In order to fight churn, Model Fitness has digitized a number of its customer profiles. Your task is to analyze them and come up with a customer retention strategy.

You should:

- Learn to predict the probability of churn (for the upcoming month) for each customer
- Draw up typical user portraits: select the most outstanding groups and describe their main features
- Analyze the factors that impact churn most
- Draw basic conclusions and develop recommendations on how to improve customer service:
 - Identify target groups
 - Suggest measures to cut churn
 - Describe any other patterns you see with respect to interaction with customers

Instructions for completing the project

Step 1. Download the data

Model Fitness provided you with CSV files containing data on churn for a given month and information on the month preceding it. The dataset includes the following fields:

- 'Churn' — the fact of churn for the month in question
- Current dataset fields:
 - User data for the preceding month
 - 'gender'
 - 'Near_Location' — whether the user lives or works in the neighborhood where the gym is located
 - 'Partner' — whether the user is an employee of a partner company (the gym has partner companies whose employees get discounts; in those cases the gym stores information on customers' employers)
 - 'Promo_friends' — whether the user originally signed up through a "bring a friend" offer (they used a friend's promo code when paying for their first membership)
 - 'Phone' — whether the user provided their phone number
 - 'Age'
 - 'Lifetime' — the time (in months) since the customer first came to the gym
- Data from the log of visits and purchases and data on current membership status
 - 'Contract_period' — 1 month, 3 months, 6 months, or 1 year
 - 'Month_to_end_contract' — the months remaining until the contract expires
 - 'Group_visits' — whether the user takes part in group sessions
 - 'Avg_class_frequency_total' — average frequency of visits per week over the customer's lifetime
 - 'Avg_class_frequency_current_month' — average frequency of visits per week over the preceding month
 - 'Avg_additional_charges_total' — the total amount of money spent on other gym services: cafe, athletic goods, cosmetics, massages, etc.

File path: `/datasets/gym_churn_us.csv`. [Download the dataset](#).

Step 2. Carry out exploratory data analysis (EDA)

- Look at the dataset: does it contain any missing features? Study the mean values and standard deviation (use the `describe()` method).
- Look at the mean feature values in two groups: for those who left (churn) and for those who stayed (use the `groupby()` method).
- Plot bar histograms and feature distributions for those who left (churn) and those who stayed.
- Build a correlation matrix and display it.

Step 3. Build a model to predict user churn

Build a binary classification model for customers where the target feature is the user's leaving next month.

- Divide the data into train and validation sets using the `train_test_split()` function.
- Train the model on the train set with two methods:
 - logistic regression
 - random forest
- Evaluate accuracy, precision, and recall for both models using the validation data. Use them to compare the models. Which model gave better results?

Remember to indicate the `random_state` parameter when dividing data and defining the algorithm.

Step 4. Create user clusters

Set aside the column with data on churn and identify object (user) clusters:

- Standardize the data.
- Use the `linkage()` function to build a matrix of distances based on the standardized feature matrix and plot a dendrogram. Note: rendering the dendrogram may take time! Use the resulting graph to estimate the number of clusters you can single out.

- Train the clustering model with the K-means algorithm and predict customer clusters. (Let the number of clusters be `n=5`, so that it'll be easier to compare your results with those of other students. However, in real life, no one will give you such hints, so you'll have to decide based on the graph from the previous step.)
- Look at the mean feature values for clusters. Does anything catch your eye?
- Plot distributions of features for the clusters. Do you notice anything?
- Calculate the churn rate for each cluster (use the `groupby()` method). Do they differ in terms of churn rate? Which clusters are prone to leaving, and which are loyal?

Step 5. Come up with conclusions and basic recommendations on working with customers

Draw conclusions and formulate recommendations regarding the strategy for customer interaction and retention.

You don't need to go into detail. Three or four essential principles and examples of their implementation in the form of specific marketing steps will do.