**Predict the Segment – Hotstar**

**Problem Statement**

Determining the demographics of customers is one of the most key tasks in the advertising domain. Advertisers usually want to target customers based on demographic attributes. However, it is difficult to get demographic data from all the customers since that can add friction to the user experience.

At Hotstar, we have detailed information on all the content that customers watch, let’s call it “watch patterns” and we’d like to use this signal to fine tune demo-targeting.

We are seeking a machine learning based solution using which we can learn patterns from customers whose watch patterns are already known. In this competition, the task is to generate predictive models that can best capture the behaviour. Participants are free to use any open source external data.

**Data Information**

A zipped file containing train, test and sample submission files are given. The training dataset consists of data corresponding to 200,000 customers and the test dataset consists of 100,000 customers. Both training and test data is in the form of json dict, where key is masked user ID and value is aggregation of all records corresponding to the user as described below.

| **Variable** | **Description** |
| --- | --- |
| ID | unique identifier variable |
| titles | titles of the shows watched by the user and watch\_time on different titles in the format “title:watch\_time” separated by comma, e.g. “JOLLY LLB:23, Ishqbaaz:40”. watch\_time is in seconds |
| genres | same format as titles |
| cities | same format as titles |
| tod | total watch time of the user spreaded across different time of days (24 hours format) in the format “time\_of\_day:watch\_time” separated by comma, e.g. “1:454, “17”:5444” |
| dow | total watch time of the user spreaded across different days of week (7 days format) in the format “day\_of\_week:watch\_time” separated by comma, e.g. “1:454, “6”:5444” |
| segment | target variable. consider them as interest segments. For modeling, encode pos = 1, neg = 0 |

**Submission**

The user has to submit a csv file with the ID and predicted probabilities. Check sample submission file.

ID,segment

test-23855,0.2

test-23854,0.1

test-23857,0.6

test-23856,0.3

test-23851,0.7

**Evaluation Metric**

Submissions will be evaluated based on AUC ROC score. For more information about this metric, [read\_here](https://en.wikipedia.org/wiki/Receiver_operating_characteristic).

**Note** : This is Round 1 of India Hacks Machine Learning 2017. In this round, two problems are given. Make sure you solve both of them. Go to Problems (See above).