Hackathon

Problem Statements

By

GREYATOM

Table of Content

How will this Hackathon work?

Structure of the Hackathon:-
The Hackathon is scheduled for two days. Each team can spend 30 mins with the mentor on both days to sound off ideas/approaches. Mentor will NOT give away solutions. Students will be provided with these problem statements on Friday. **Each team need to select one problem statement**

Day 1: Saturday, Nov 03, 2018

- Hackathon kick-off by the mentor
- Data Exploration Graphs, Correlation, Feature Engineering should be completed by 4.30PM

Day 2: Sunday, Nov 04, 2018

- Feature Selection, ML Models must be ready by 11 AM
- Understanding and fine-tuning ML models by 3 PM
- Rest of the time on proper code documentation, packaging repository for anybody to use and presentation
- Final Presentations - 3:00 PM - 6.00 PM

Note: You need to hand over your GitHub repo details and presentation files to the mentor before the presentation.

1. H1B Disclosure Dataset: Predicting the case Status

Problem Statement:

The H-1B Dataset selected for this project contains data from employer’s Labor Condition Application and the case certification determinations processed by the Office of Foreign Labor Certification (OFLC) where the date of the determination was issued on or after October 1, 2016, and on or before June 30, 2017.

The Labor Condition Application (LCA) is a document that a prospective H-1B employer files with U.S. Department of Labor Employment and Training Administration (DOLETA) when it seeks to employ nonimmigrant workers at a specific job occupation in an area of intended employment for not more than three years.

The goal for this project is to predict the case status of an application submitted by the employer to hire non-immigrant workers under the H-1B visa program. The employer can hire non-immigrant workers only after their LCA petition is approved. The approved LCA petition is then submitted as part of the Petition for a Non-immigrant Worker application for work authorizations for H-1B visa status.

Download the dataset:
https://drive.google.com/drive/folders/1sIjRnbrIvrDaSkj8TIi-myErAx258iGf?usp=sharing

Data Set Information:
The H-1B dataset from OFLC contained 40 attributes and 528,147 instances.

2. Predict if the client will subscribe to direct marketing campaign for a banking institution

Problem Statement:
The data is related to direct marketing campaigns of a Portuguese banking institution. Predict if the client will subscribe to a term deposit based on a marketing campaign

Data Set Download:
https://drive.google.com/drive/folders/1urwTQPkUypJ6dGDJgS9Gszb83bfXEG6z?usp=sharing

Data Set Information:
The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.
There are four datasets:

1. bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
2. bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
3. bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with fewer inputs).
4. bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with fewer inputs). The smallest datasets are provided to test more computationally demanding machine learning algorithms

Goal:- The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

3. Indian Startup Funding: Predict How much funds does startup generally get in India?

Possible questions which could be answered are:

- How does the funding ecosystem change with time?
- Do cities play a major role in funding?
- Which industries are favored by investors for funding?
- Who are the important investors in the Indian Ecosystem?
- How much funds do startups generally get in India?

Data Set Download:
https://drive.google.com/drive/folders/1PuD_5PUfk9x9noE5im0ZsZvMegh1gkKi?usp=sharing

Data Set Information:
This dataset is a chance to explore the Indian startup scene. Deep dive into funding data and derive insights into the future!
SNo: Serial number

Date: Date of funding in format DD/MM/YYYY.

StartupName: Name of the startup which got funded.

industry vertical: Industry to which the startup belongs.

SubVertical: Sub-category of the industry type.

CityLocation: City which the startup is based out of.

InvestorsName: Name of the investors involved in the funding round.

investment type: Either Private Equity or Seed Funding.

AmountInUSD: Funding Amount in USD.

Remarks: Other information, if any.

4. Predict the Segment - Hotstar

- 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 behavior. Participants are free to use any open source external data.

Data Set 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 an aggregation of all records corresponding to the user as described below.

Description ID: a unique identifier
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

Download dataset from here

- https://drive.google.com/file/d/1YPVPife9QCCQPBcT5yOYi-gCpV6NABMR/view?usp=sharing

Helpful techniques

- Try different clustering techniques to segment users and use those outputs as features to your machine learning models.
- Data cleaning and feature creation should be novel. Go with your intuition.

5. Predict the probability of whether an ad will get clicked or not.

A leading affiliate network company from Europe wants to leverage machine learning to improve (optimise) their conversion rates and eventually their topline. Their network is spread across multiple countries in europe such as Portugal, Germany, France, Austria, Switzerland etc.
Affiliate network is a form of online marketing channel where an intermediary promotes products / services and earns commission based on conversions (click or sign up). The benefit companies sees in using such affiliate channels is that, they are able to reach to audience which doesn't exist in their marketing reach.
The company wants to improve their CPC (cost per click) performance. A future insight about an ad performance will give them enough headstart to make changes (if necessary) in their upcoming CPC campaigns.
In this challenge, you have to predict the probability of whether an ad will get clicked or not.

Data Set

- https://drive.google.com/file/d/1oJCM7LJ5oCcI4PuQH6vTFypAzwsMA0Jz/view?usp=sharing

Data Description
You are given three files to download: train.csv, test.csv and sample_submission.csv Variables in this data set are anonymized due to privacy.
The training data is given for 10 days (10 Jan 2017 to 20 Jan 2017). The test data is given for next 3 days.

Features Given

ID	Unique ID
datetime	timestamp
siteid	website id
offerid	offer id (commission based offers)
category	offer category
merchant	seller ID
countrycode	country where affiliates reach is present

browserid	browser used
devid	device used
click	target variable

Evaluation Metric

Submission will be evaluated based on AUC-ROC score. Higher the better.

Helpful techniques:

- Mean encoding of variables will help a lot
- Combine two, three, four variables and then generate mean encoding with output variable. This will help in understanding user behavior and make it easier for ML models to learn a variety of features.
- Use all standard machine learning models. GBMs and its variants should work well.
- Try different ensembling and stacking techniques to improve the score.