ClariTrics Technologies

(A BUDDIHealth Company)

Data Hackathon 2017, IIITDM Chennai

BUDDIHealth (formerly ClariTrics) is a disruptive, military grade secure, HIPAA compliant 'Deep Learning' based cloud platform focused in automating revenue cycle functions such as medical coding & billing for the Healthcare Providers, medical billing and coding vendors of all sizes. We are a health-tech focused artificial intelligence based algorithmic company based in New York City and have a cloud partnership with Microsoft for HIPAA compliance, PHI security and for scaling technology & business on Azure. We have developed a next generation revenue cycle platform to enable our provider clients to increase productivity, improve reimbursements, and improve efficiency and help providers focus more on quality care.

www.buddihealth.com

Rules of the Hackathon

- You are expected to solve any one of the above problems.
- The competition starts at the moment when you receive this document. The hackathon ends on 30th evening 10 PM IST. You are expected to send your reports before 10PM on 30th.
- You are free to use any tools and machines you have rightful access to. You can use any programming language or statistical software or libraries.
- At the end of the competition, the participants are requested to send a detailed report of their work. Report submission is mandatory.
- Code and supporting files are mandatory along with the report, else your participation will not be considered for this hackathon. For GUI based tools, submit zip file of snapshots of steps taken by you, else submit code file. Submit the source code and other appropriate files in .zip or .tar compressed archive. If using jupyter notebooks, you can submit them along with your report. To be more specific, the report must adhere to the format below.
- Report Format
    - Abstract - Problem Statement
    - Brief Literature survey
    - Data Characterization, Feature engineering
    - Model Selection
    - Model Parameter Tuning
    - Model Evaluation using F1 score with training and testing performance tabulations.
    - Conclusion(Along with any inferences from the data, model would be appreciated)
- Code of conduct
    - The students should not discuss the problem/solution with others.
    - The students should refrain from asking questions in online forums.
    - The students should refrain from copy-pasting code from the Internet. Using open-source libraries is entertained. Getting inspired by existing code is also ok, but a verbatim copy is forbidden.
    - Honesty is the best policy.
    - The students are expected to learn models from training set only. The testing set is provided with the labels, but the test set should not be used for training the model either in the inductive or transductive setting.
- Please submit your queries and final reports to hackathon@buddihealth.com

1. What's the Cuisine?

Some of our strongest geographic and cultural associations are tied to a region's local foods. If you're in Northern California, you'll be walking past the inevitable bushels of leafy greens, spiked with dark purple kale and the bright pinks and yellows of chard. Across the world in South Korea, mounds of bright red kimchi greet you, while the smell of the sea draws your attention to squids squirming nearby. India's market is perhaps the most colorful, awash in the rich hues and aromas of dozens of spices: turmeric, star anise, poppy seeds, and garam masala as far as the eye can see.

This challenge asks you to predict the category of a dish's cuisine given a list of its ingredients.In the dataset, we include the recipe id, the type of cuisine, and the list of ingredients of each recipe (of variable length). The data is stored in JSON format.

In the test data, the format of a recipe is the same as training data, only the cuisine type is removed, as it is the target variable you are going to predict.

Dataset Link

- Training Data
- Test Data

About the dataset

- Training Data - The training set containing recipes id, type of cuisine, and list of ingredients
- Test Data - The test set containing recipes id, and list of ingredients

2. Bike Sharing Demand

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city.

In this problem, you are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C. Moreover, this is a Regression with temporal component

Dataset Link

- Training Data
- Test Data

About the dataset

You are provided hourly rental data spanning two years. For this problem, the training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month. You must predict the total count of bikes rented during each hour covered by the test set, using only information available prior to the rental period.

3. Sentiment Analysis on Movie Reviews

Sentiment analysis is a challenging subject in machine learning. People express their emotions in language that is often obscured by sarcasm, ambiguity, and plays on words, all of which could be very misleading for both humans and computers.

The labeled data set consists of 50,000 IMDB movie reviews, specially selected for sentiment analysis. The sentiment of reviews is binary, meaning the IMDB rating < 5 results in a sentiment score of 0, and rating >=7 have a sentiment score of 1. No individual movie has more than 30 reviews. The 25,000 review labeled training set does not include any of the same movies as the 25,000 review test set. In addition, there are another 50,000 IMDB reviews provided without any rating labels.

Dataset Link

- Traning Data
- Test Data

About the dataset

- Training Data - The file is tab-delimited and has a header row followed by 25,000 rows containing an id, sentiment, and text for each review.
- Test Data - The tab-delimited file has a header row followed by 25,000 rows containing an id and text for each review. Your task is to predict the sentiment for each one.

4. Can you Handle the Imbalance?

You are given a dataset in which the instances of one class(majority) outnumber the instances of the other class(minority). The goal of this challenge is to learn a classifier

from the given dataset, and use it for prediction on the test dataset attaining maximum possible accuracy for both classes.

Seems simple, right? Anyways, we have some links for you to refer before you start playing around with the dataset.

Dataset Link

- Training Data
- Test Data

About the Dataset

Each row represents an instance and the final column in each row represents the class of the instance.

Reference for Class Imbalance Problem

- https://elitedatascience.com/imbalanced-classes
- https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/
- https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/