1. **Recommendation Engine**

This is an advanced recommendation system challenge. In this practice problem, you are given the data of programmers and questions that they have previously solved, along with the time that they took to solve that particular question. As a data scientist, the model you build will help online judges to decide the next level of questions to recommend to a user.

**Problem:** Predict the time taken to solve a problem given the current status of the user.

https://datahack.analyticsvidhya.com/contest/practice-problem-recommendation-engine/

1. **Age Detection of Indian Actors Dataset**

This is a fascinating challenge for any deep learning enthusiast. The dataset contains thousands of images of Indian actors and your task is to identify their age. All the images are manually selected and cropped from the video frames resulting in a high degree of variability interms of scale, pose, expression, illumination, age, resolution, occlusion, and makeup. There are 19,906 images in the training set and 6,636 in the test set.

**Problem:** Predict the age of the actors.

**Dataset:** <https://datahack.analyticsvidhya.com/contest/practice-problem-age-detection/>

1. **Chicago Crime Dataset**

The ability to handle large datasets is expected of every data scientist these days. Companies no longer prefer to work on samples when they the computational power to work on the full dataset. This dataset provides you a much needed hands-on experience of handling large data sets on your local machines. The problem is easy, but data management is the key! This dataset has 6M observations. It’s a multi-classification problem.

**Problem:** Predict the type of crime.

<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>

1. **Vox Celebrity Dataset**

Audio processing is rapidly becoming an important field in deep learning hence here’s another challenging problem. This dataset is for large-scale speaker identification and contains words spoken by celebrities, extracted from YouTube videos. It’s an intriguing use case for isolating and identifying speech recognition. The data contains 100,000 utterances spoken by 1,251 celebrities.

**Problem:** Figure out which celebrity the voice belongs to.

http://www.robots.ox.ac.uk/~vgg/data/voxceleb/

Ref: https://www.robots.ox.ac.uk/~vgg/publications/2017/Nagrani17/nagrani17.pdf

1. **Urban Sound Classification**

When you start your machine learning journey, you go with simple machine learning problems like titanic survival prediction. But you still don’t have enough practice when it comes to real life problems. Hence, this practice problem is meant to introduce you to audio processing in the usual classification scenario. This dataset consists of 8,732 sound excerpts of urban sounds from 10 classes.

**Problem:** Classify the type of sound from the audio.

<https://datahack.analyticsvidhya.com/contest/practice-problem-urban-sound-classification/>

1. **Identify your Digits Dataset**

This dataset allows you to study, analyze and recognize elements in the images. That’s exactly how your camera detects your face, using image recognition! It’s your turn to build and test that technique. It’s a digit recognition problem. This data set has 7,000 images of 28 X 28 size, totalling 31MB.

**Problem:** Identify digits from an image.

https://datahack.analyticsvidhya.com/contest/practice-problem-identify-the-digits/

1. **Twitter Classification Dataset**

Working with Twitter data has become an integral part of sentiment analysis problems. If you want to carve a niche for yourself in this area, you will have fun working on the challenge this dataset poses. The dataset is 3MB in size and has 31,962 tweets.

**Problem:**Identify the tweets which are hate tweets and which are not.

<https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/>

1. **Census Income Dataset**

It’s an imbalanced classification and a classic machine learning problem. You know, machine learning is being extensively used to solve imbalanced problems such as cancer detection, fraud detection etc. It’s time to get your hands dirty. The data set has 48,842 rows and 14 columns. For guidance, you can check this imbalanced data project.

**Problem:** Predict the income class of US population.

**Dataset: http://archive.ics.uci.edu/ml/machine-learning-databases/census-income-mld/**

**Ref 1:** <https://www.analyticsvidhya.com/blog/2016/09/this-machine-learning-project-on-imbalanced-data-can-add-value-to-your-resume/>

**Ref 2:** https://cseweb.ucsd.edu/~jmcauley/cse190/reports/sp15/048.pdf

1. **Human Activity Recognition Dataset**

This data set is collected from recordings of 30 human subjects captured via smartphones enabled with embedded inertial sensors. Many machine learning courses use this data for teaching purposes. It’s your turn now. This is a multi-classification problem. The data set has 10,299 rows and 561 columns.

**Problem:** Predict the activity category of a human.

Dataset: <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>

Ref: <https://rstudio-pubs-static.s3.amazonaws.com/291850_859937539fb14c37b0a311db344a6016.html>

1. **Black Friday Dataset**

This dataset comprises of sales transactions captured at a retail store. It’s a classic dataset to explore and expand your feature engineering skills and day to day understanding from multiple shopping experiences. This is a regression problem. The dataset has 550,069 rows and 12 columns.

**Problem:** Predict purchase amount.

Dataset: <https://datahack.analyticsvidhya.com/contest/black-friday/>

1. **[Text Mining Data Set](https://www.kaggle.com/c/whats-cooking" \t "_blank) (What’s Cooking? on Kaggle)**

In simple words, text mining means analysing data within text. Large amounts of unstructured data is found within natural language.  Mining this unstructured data from sources such as  e-mails, text messages and other platforms like Facebook and Twitter, can help companies gain business insights about customers, and their patterns and topics of interest.

Data sets from the famous competition, **What’s Cooking? on Kaggle,** can help you get started in the area of text mining. The goal is to use recipe ingredients to categorize cuisines. Text mining data sets test skills on classification and clustering. Occasionally, regression analysis may be required.

**Objective:** Classification and categorisation based on tags or labels.

**Dataset:** https://www.kaggle.com/c/whats-cooking

1. **Chatbot - 1**

This will be a **Retrieval-based chatbot**, made using the Google's DialogFlow framework. The bot would be deployed on a cloud platform (Firebase or Heroku)

1. **Chatbot - 2**

This will be a **Generative-model-based** chatbot, made using the RASA or similar framework. The bot would be finally deployed on a cloud platform (Firebase or Heroku) for serving the APIs over WhatsApp.

1. **Chatbot - 3**

This will be a **rule-based chatbot**, made using the DialogFlow/RASA or similar framework. The bot would be finally deployed on a cloud platform (Firebase or Heroku) for serving the APIs over WhatsApp.