17 Ultimate Data Science Projects To Boost Your Knowledge and Skills (& can be accessed freely)

MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/MACHINE-LEARNING/) PYTHON (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/PYTHON-2/) R (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/R/)

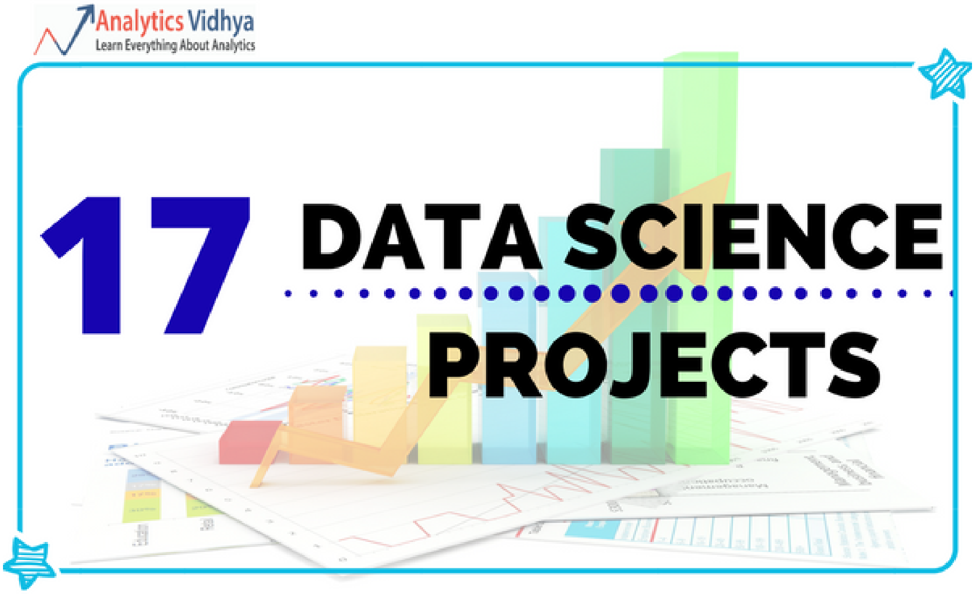
idhya.com/blog/2016/10/17-ultimate-data-science-projects-to-boost-your-knowledge-and-
t%20Your%20Knowledge%20and%20Skills%20(&%20can%20be%20accessed%20freely)) (https://twitter.com/home?
20Your%20Knowledge%20and%20Skills%20(&%20can%20be%20accessed%20freely)+https://www.analyticsvidhya.com/blog/2016/10/17-
s://plus.google.com/share?url=https://www.analyticsvidhya.com/blog/2016/10/17-ultimate-data-science-projects-to-boost-your-knowledge-
analyticsvidhya.com/blog/2016/10/17-ultimate-data-science-projects-to-boost-your-knowledge-and-
0/17-DATA-SCIENCE-
s%20To%20Boost%20Your%20Knowledge%20and%20Skills%20(&%20can%20be%20accessed%20freely))

Introduction

Data science projects offer you a promising way to kick-start your analytics career. Not only you get to learn data science by applying, you also get projects to showcase on your CV. Nowadays, recruiters evaluate a candidate's potential by his/her work, not as much by certificates and resumes. It wouldn't matter, if you just tell them how much you know, if you have nothing to show them! That's where most people struggle and miss out!

You might have worked on several problems, but if you can't make it presentable & explanatory, how on earth would someone know what you are capable of? That's where these projects would help you. Think of the time spend on these projects like your training sessions. I guarantee, the more time you spend, the better you'll become!

The data sets in the list below are handpicked. I've made sure to provide you a taste of variety of problems from different domains with different sizes. I believe, everyone must learn to smartly work on large data sets, hence large data sets are added. Also, I've made sure all the data sets are open and free to access.



Useful Information

To help you decide your start line, I've divided the data set into 3 levels namely:

1. **Beginner Level:** This level comprises of data sets which are fairly easy to work with, and doesn't require complex data science techniques. You can solve them using basic regression / classification algorithms. Also, these data sets have enough open tutorials to get you going. In this list, I've provided tutorials also to help you get started.

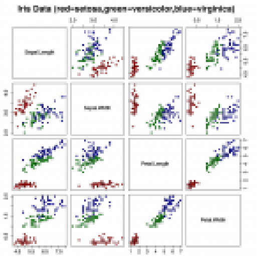
2. **Intermediate Level:** This level comprises of data sets which are challenging. It consists of mid & large data sets which require some serious pattern recognition skills. Also, feature engineering will make a difference here. There is no limit of use of ML techniques, everything under the sun can be put to use.
3. **Advanced Level:** This level is best suited for people who understand advanced topics like neural networks, deep learning, recommender systems etc. High dimensional data are also featured here. Also, this is the time to get creative – see the creativity best data scientists bring in their work and codes.

Table of Contents

Beginner Level

1. Iris Data Set

This is probably the most versatile, easy and resourceful data set in pattern recognition literature. Nothing could be simpler than iris data set to learn classification techniques. If you are totally new to data science, this is your start line. The data has only 150 rows & 4 columns.



Problem: Predict the flower class based on available attributes.

Start: Get Data (https://archive.ics.uci.edu/ml/datasets/Iris) | **Tutorial:** Get Here (http://www.slideshare.net/thoi_gian/iris-data-analysis-with-r)

2. Titanic Data Set

This is another most quoted data set in global data science community. With several tutorials and help guides, this project should give you enough kick to pursue data science deeper. With healthy mix of variables comprising categories, numbers, text, this data set has enough scope to support crazy ideas! This is a classification problem. The data has 891 rows & 12 columns.

Problem: Predict the survival of passengers in Titanic.

Start: Get Data (https://www.kaggle.com/c/titanic) | **Tutorial:** Get Here (http://trevorstephens.com/kaggle-titanic-tutorial/getting-started-with-r/)

3. Loan Prediction Data Set

Among all industries, insurance domain has the largest use of analytics & data science methods. This data set would provide you enough taste of working on data sets from insurance companies, what challenges are faced, what strategies are used, which variables influence the outcome etc. This is a classification problem. The data has 615 rows and 13 columns.

Problem: Predict if a loan will get approved or not.

Start: Get Data (https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/) | **Tutorial:** Get Here (https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/)

4. Bigmart Sales Data Set

Retail is another industry which extensively uses analytics to optimize business processes. Tasks like product placement, inventory management, customized offers, product bundling etc are being smartly handled using data science techniques. As the name suggests, this data comprises of transaction record of a sales store. This is a regression problem. The data has 8523 rows of 12 variables.

Problem: Predict the sales.

Start: Get Data (https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/) | **Tutorial:** Get Here (https://www.analyticsvidhya.com/blog/2016/02/bigmart-sales-solution-top-20/)

5. Boston Housing Data Set

This is another popular data set used in pattern recognition literature. The data set comes from real estate industry in Boston (US). This is a regression problem. The data has 506 rows and 14 columns. Thus, it's a fairly small data set where you can attempt any technique without worrying about your laptop's memory issue.

Problem: Predict the median value of owner occupied homes

Start: Get Data (http://archive.ics.uci.edu/ml/datasets/Housing) | **Tutorial:** Get Here (https://www.analyticsvidhya.com/blog/2015/11/started-machine-learning-ms-excel-xl-miner/)

Intermediate Level

1. Human Activity Recognition

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 students practice. It's your turn now. This is a multi-classification problem. The data set has 10299 rows and 561 columns.

Problem: Predict the activity category of a human

Start: Get Data (http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones)

2. Black Friday Data Set

This data set comprises of sales transactions captured at a retail store. It's a classic data set to explore your feature engineering skills and day to day understanding from your shopping experience. It's a regression problem. The data set has 550069 rows and 12 columns.

Problem: Predict purchase amount.

Start: Get Data (https://datahack.analyticsvidhya.com/contest/black-friday/)

3. Text Mining Data Set

été impossible en fait ou en droit. S'il s'agissait d'un enfant au berceau, d'un absent ou d'une femme, il était nécessaire que quelqu'un prît en main la cause de ce plaideur, qui ne pouvait ou ne devait pas présenter lui-même sa défense. Si l'État était directement intéressé à une affaire, si par exemple il s'agissait d'un procès de haute trahison, il ne fallait pas qu'en un cas aussi grave la défense de l'intérêt public fût abandonnée à l'initiative privée. La république désignait donc des orateurs, chargés de la représenter, et de soutenir l'accusation. Dans tous ces cas, il avait bien fallu déroger à la règle, et admettre que le plaideur se pourrait faire représenter par autrui.

De même, s'il s'agissait d'un laboureur, d'un vigneron, d'un matelot ou d'un soldat, dont la langue indocile se montrait rebelle à la parole, il devenait bien difficile d'appliquer la loi. C'eût été vraiment perdre le temps des juges, et aussi se moquer d'eux que de leur produire un plaideur absolument incapable d'exposer le premier mot de son affaire. N'était-ce pas d'ailleurs une injustice criante que de pauvres vieux soldats blanchis sous le harnais, qui dans maints combats s'étaient couverts d'une glorieuse sueur, fussent exposés sans défense aux attaques d'un vil sycophante, et se vissent ravir par une condamnation l'argent qui devait servir à leur assurer des funérailles décentes? Il avait paru juste que ce plaideur inexpérimenté pût appeler à son aide quelque

This data set is originally from siam competition 2007. The data set comprises of aviation safety reports describing problem(s) which occurred in certain flights. It is a multi-classification, high dimensional problem. It has 21519 rows and 30438 columns.

Problem: Classify the documents according to their labels

Start: Get Data (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html#siam-competition2007) | Get Information (https://catalog.data.gov/dataset/siam-2007-text-mining-competition-dataset/resource/794f14ae-8135-41d2-88c8-86bf8fad9cf6/proxy)

4. Trip History Data Set

This data set comes from a bike sharing service in US. This data set requires you to exercise your pro data munging skills. The data set is provided quarter wise from 2010 (Q4) onwards. Each file has 7 columns. It is a classification problem.

Problem: Predict the class of user

Start: Get Data (https://www.capitalbikeshare.com/trip-history-data)

5. Million Song Data Set

Didn't you know analytics can be used in entertainment industry also? Do it yourself now. This data set puts forward a regression task. It consists of 515345 observations and 90 variables. However, this is just a tiny subset of original database (http://labrosa.ee.columbia.edu/millionsong/pages/getting-dataset) of million song data. You should use data linked below.

Problem: Predict release year of the song

Start: Get Data (http://archive.ics.uci.edu/ml/datasets/YearPredictionMSD)

6. Census Income Data Set

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 hand dirty. The data set has 48842 rows and 14 columns. For guidance, you can check my imbalanced data project (https://www.analyticsvidhya.com/blog/2016/09/this-machine-learning-project-on-imbalanced-data-can-add-value-to-your-resume/).

Problem: Predict the income class of US population

Start: Get Data (http://archive.ics.uci.edu/ml/machine-learning-databases/census-income-mld/)

7. Movie Lens Data Set

This data set allows you to build a recommendation engine. Have you created one before? It's one of the most popular & quoted data set in data science industry. It is available in various dimensions (http://grouplens.org/datasets/movielens/). Here I've used a fairly small size. It has 1 million ratings from 6000 users on 4000 movies.

Problem: Recommend new movies to users

Start: Get Data (http://grouplens.org/datasets/movielens/1m/)

Advanced Level

1. Identify your Digits Data Set

This data set 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 an digit recognition problem. This data set has 7000 images of 28 X 28 size, sizing 31MB.

Problem: Identify digits from an image

Start: Get Data (https://datahack.analyticsvidhya.com/contest/practice-problem-identify-the-digits/)

2. Yelp Data Set

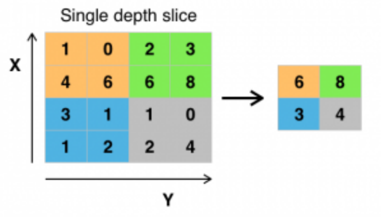
This data set is a part of round 8 of The Yelp Dataset Challenge. It comprises of nearly 200,000 images, provided in 3 json files of ~2GB. These images provide information about local businesses in 10 cities across 4 countries. You are required to find insights from data using cultural trends, seasonal trends, infer categories, text mining, social graph mining etc.

Problem: Find insights from images

Start: Get Data (https://www.yelp.com/dataset_challenge)

3. Image Net Data Set

ImageNet offers variety of problems which encompasses object detection, localization, classification and screen parsing. All the images are freely available. You can search for any type of image and build your project around it. As of now, this image engine has 14,197,122 images of multiple shapes sizing up to 140GB.



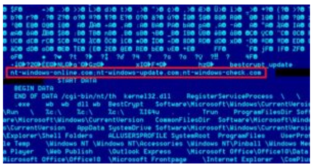
1	0	2	3
4	6	6	8
3	1	1	0
1	2	2	4

6	8
3	4

Problem: Problem to solve is subjected to the image type you download

Start: Get Data (http://image-net.org/download-imageurls)

4. KDD 1999 Data Set



How could I miss KDD Cup? Originally, KDD brought the taste of data mining competition to the world. Don't you want to see what data set they used to offer? I assure you, it'll be an enriching experience. This data poses a classification problem. It has 4M rows and 48 columns in a ~1.2GB file.

Problem: Classify a network intrusion detector as good or bad.

Start: Get Data (https://archive.ics.uci.edu/ml/datasets/KDD+Cup+1999+Data)

5. Chicago Crime Data Set

The ability of handle large data sets is expected of every data scientist these days. Companies no longer prefer to work on samples, they use full data. This data set would provide you 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 data set has 6M observations. It's a multi-classification problem.

Problem: Predict the type of crime.

Start: Get Data (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2) | To download data, click on Export -> CSV

End Notes