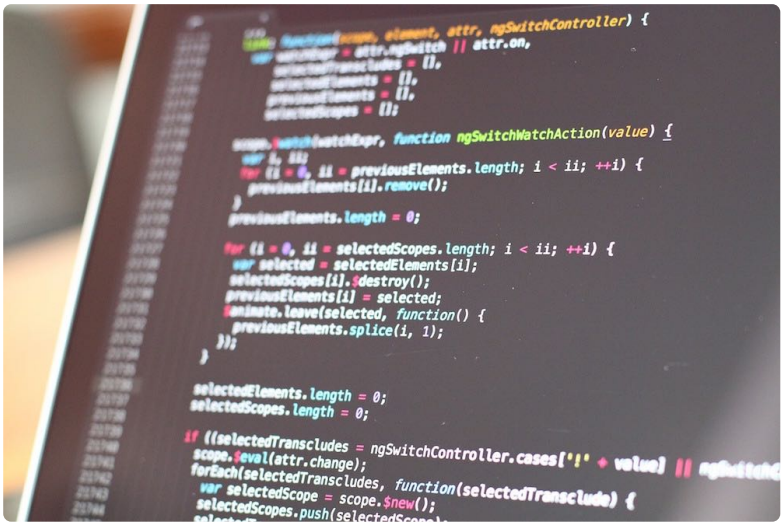


Data Science: A Kaggle Walkthrough – Introduction

FEBRUARY 22, 2016 / BRETT ROMERO / 0 COMMENTS

Search SEARCH

POPULAR POSTS

The argument for taxing capital gains at the full rate

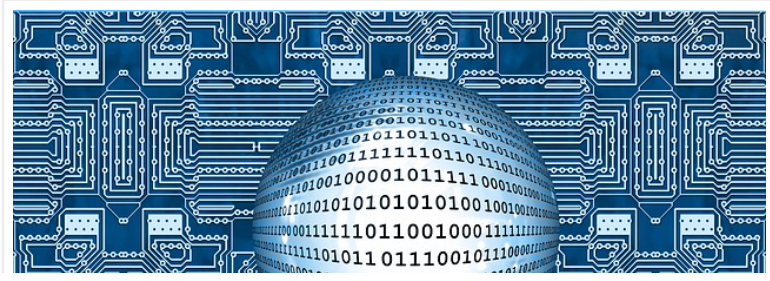
4 Reasons Working Long Hours is Crazy

Data Science: A Kaggle Walkthrough – Introduction

Why You Probably Don't Need a Financial Advisor

I have spent a lot of time working with spreadsheets, databases, and data more generally. This work has led to me having a very particular set of skills, skills I have acquired over a very long career. Skills that make me a nightmare for people like you. If you let my daughter go now, that'll be the end of it. I will not look for you, I will not pursue you. But if you don't, I will look for you, I will find you, and I will kill you.

The badassery of Liam Neeson aside, although I have spent years working with data in a range of capacities, the skills and techniques required for 'data science' are a very specific subset that do not tend to come up in too many jobs. What is more, data science tends to involve a lot more programming than most other data related work and this can be intimidating for people who are not coming from a computer science background. The problem is, people who work with data in other contexts (e.g. economics and statistics), as well as those with industry specific experience and knowledge, can often bring different and important perspectives to data science problems. Yet, these people often feel unable to contribute because they do not understand programming or the black box models being used.



Why Australians Love Foster's and Other Beer Related Stories

ARCHIVES

August 2017

May 2017

April 2017

January 2017

December 2016

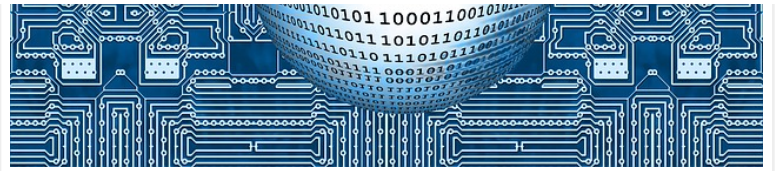
November 2016

May 2016

April 2016

March 2016

February 2016



Something that has nothing to do with data science

Therefore, in a probably futile attempt to shed some light on this field, this will be the first part in a multi-part series looking at what data science involves and some of the techniques most commonly used. This series is not intended to make everyone experts on data science, rather it is intended to simply try and remove some of the fear and mystery surrounding the field. In order to be as practical as possible, this series will be structured as a walk through of the process of entering a Kaggle competition and the steps taken to arrive at the final submission.

What is Kaggle?

For those that do not know, Kaggle is a website that hosts data science problems for an online community of data science enthusiasts to solve. These problems can be anything from predicting cancer based on patient data, to sentiment analysis of movie reviews and handwriting recognition – the only thing they all have in common is that they are problems requiring the application of data science to be solved.

The problems on Kaggle come from a range of

January 2016
December 2015
November 2015
October 2015
September 2015
August 2015
July 2015
June 2015
May 2015
April 2015
March 2015

CATEGORIES

Applications
Australia
Data Science
Development
Economics
Economy
Europe
Excel
Finance

sources. Some are provided just for fun and/or educational purposes, but many are provided by companies that have genuine problems they are trying to solve. As an incentive for Kaggle users to compete, prizes are often awarded for winning these competitions, or finishing in the top x positions. Sometimes the prize is a job or products from the company, but there can also be substantial monetary prizes. Home Depot for example is currently offering $40,000 for the algorithm that returns the most relevant search results on homedepot.com.

Despite the large prizes on offer though, many people on Kaggle compete simply for practice and the experience. The competitions involve interesting problems and there are plenty of users who submit their scripts publically, providing an excellent opportunity for learning for those just trying to break into the field. There are also active discussion forums full of people willing to provide advice and assistance to other users.

What is not spelled out on the website, but is assumed knowledge, is that to make accurate predictions, you will have to use machine learning.

Machine Learning

When it comes to machine learning, there is a lot of general misunderstanding about what this actually

Other

Technology

Transparency

United States

Visualization

involves. While there are different forms of machine learning, the one that I will focus on here is known as classification, which is a form of 'supervised learning'. Classification is the process of assigning records or instances (think rows in a dataset) to a specific category in a pre-determined set of categories. Think about a problem like predicting which passengers on the Titanic survived (i.e. there are two categories – 'survived' and 'did not survive') based on their age, class and gender[1].

Titanic Classification Problem

Show 10 entries Search:

Passenger	Age	Class	Gender	Survived?
0001	32	First	Female	?
0002	12	Second	Male	?
0003	64	Steerage	Male	?
0004	23	Steerage	Male	?
0005	11	Steerage	Male	?
0006	42	Steerage	Male	?
0007	9	Second	Female	?
0008	8	Steerage	Female	?
0009	19	Steerage	Male	?
0010	55	First	Male	?

Showing 1 to 10 of 12 entries Previous Next

Referring specifically to 'supervised learning' algorithms, the way these predictions are made is

by providing the algorithm with a dataset (typically the larger the better) of 'training data'. This training data contains all the information available to make the prediction *as well as* the categories each record corresponds to. This data is then used to 'train' the algorithm to find the most accurate way to classify those records for which we do not know the category.

Training Data

Show 10 entries Search:

Passenger	Age	Class	Gender	Survived?
0013	23	Second	Female	1
0014	21	Steerage	Female	0
0015	46	Steerage	Male	0
0016	32	First	Male	0
0017	13	First	Female	1
0018	24	Second	Male	0
0019	29	First	Male	1
0020	80	Second	Male	1
0021	9	Steerage	Female	0
0022	44	Steerage	Male	0

Showing 1 to 10 of 12 entries Previous Next

Although that seems relatively straightforward, part of what makes data science such a complex field is the limitless number of ways that a predictive model can be built. There are a huge

number of different algorithms that can be trained, mostly with weird sounding names like Neural Network, Random Forest and Support Vector Machine (we will look at some of these in more detail in future installments). These algorithms can also be combined to create a single model. In fact, the people/teams that end up winning Kaggle competitions often combine the predictions of a number of different algorithms.

To make things more complicated, within each algorithm, there is a range of parameters that can be adjusted to significantly alter the prediction accuracy, and these parameters will vary for each classification problem. Finding the optimal set of parameters to maximize accuracy is often an art in itself.

Finally, just feeding the training data into an algorithm and hoping for the best is typically a fast track to poor performance (if it works at all). Significant time is needed to clean the data, correct formats and add additional 'features' to maximize the predictive capability of the algorithm. We will go into more detail on both of these requirements in future installments.

OK, so now let's put all this into context by looking at the competition I entered, provided by Airbnb. The aim of the competition was to predict the country that users will make their first booking in, based on some basic user profile data[2]. In this case, the categories were the different country

options and an additional category for users that had not made a previous booking through Airbnb. The training data was a set of users for whom we were provided with the correct category (i.e. what country they made their first booking in). Using the training data, I was required to train the model to accurately predict the country of first booking, and then submit my predictions for a set of users for whom we did not know the outcome.

How?

The aim of this series is to walk through the process of assessing and analyzing data, cleaning, transforming and adding new features, constructing and testing a model, and finally creating final predictions. The primary technology I will be using as I walk through this is Python, in combination with Excel/Google Sheets to analyze some of the outputs. Why Python? There are several reasons:

1. It is free and open source.
2. It has a great range of libraries (also free) that provide access to a large number of machine learning algorithms and other useful tools. The libraries I will primarily use are numpy, pandas and sklearn.
3. It is very popular, meaning when I get stuck on a problem, there is usually plenty of material and documentation to be found online for help.

4. It is very fast (primarily the reason I have chosen Python over R).

For those that are interested in following this series but do not have a programming background, do not panic – although I will show code snippets as we go – being able to read the code is not vital to understanding what is happening.

Next Time

In the next piece, we will start looking at the data in more detail and discuss how we can clean and transform it, to help optimize the model performance.

[1] This is an actual competition on Kaggle at the moment (no prizes are awarded, it is for experience only).

[2] The data has been anonymized so that users cannot be identified

Data Science, Technology

ANALYSIS DATA DATA SCIENCE KAGGLE MACHINE LEARNING PYTHON

PREVIOUS POST
The argument for taxing capital gains at the full rate

NEXT POST
Data Science: A Kaggle Walkthrough – Understanding the Data

Leave a Reply

Your email address will not be published.

Name

Email

Website

Post Comment