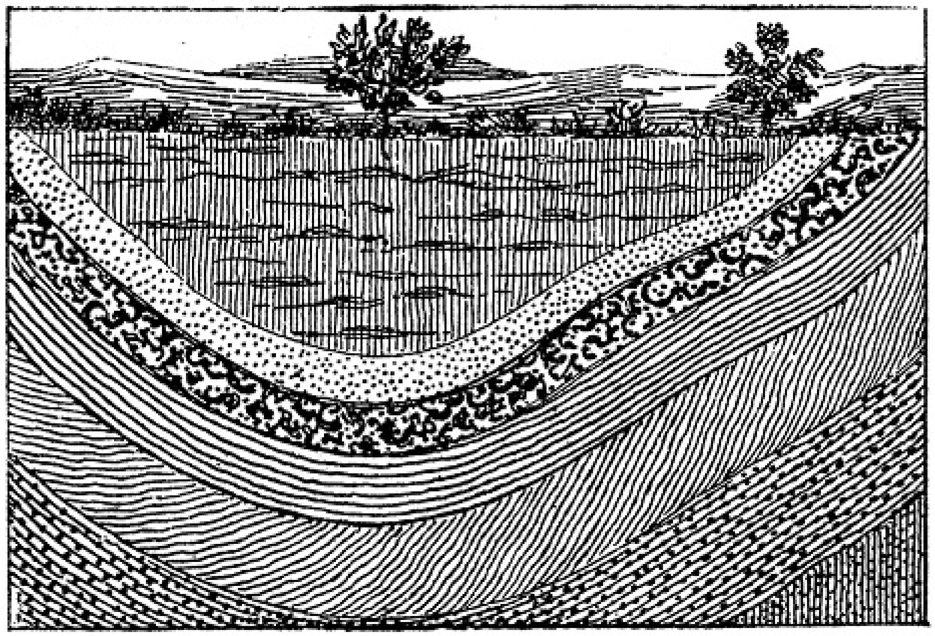
Recent Kaggle Winner Discusses Statistical Machine Learning Methods for his Winning Soil Property Predictions

Features

Author: Lillian Pierson, P.E.
Date: 05 Jan 2015

Recently, predictive modeling platform Kaggle hosted an Africa Soil Prediction Challenge. African Soil Information Service (AfSIS) sponsored this competition with the main goal that competitors would “predict physical and chemical properties of soil using spectral measurements”. With these predictions in hand, AfSIS would have access to the information it needs for improved decision-making for ecological planning. Data-informed decision-making with respect to ecological planning results in improved crop yields, more sustainable agriculture, cleaner water, cleaner air, and increased plant growth rates. In a place like Africa, with strained soil systems, poor agricultural productivity, poor sanitation, and the cycles of poverty that these factors generate, AfSIS is on a critical, life-saving mission here.

Interestingly, among the 1233 other competitors, it was a smart, young software engineer and data scientist that emerged as the competition winner. In an exclusive interview for Statistics Views website, winner Yasser Tabandeh describes what steps he took to make the winning prediction, and how it was possible for him to beat other data scientists who had a much stronger environmental background than that of his own.



You’re a software engineer and data scientist, yet you’re also the recent winner of Kaggle’s Africa Soil Property Prediction Challenge. From this win, it’s obvious that you’ve had tremendous success in applying your software and data expertise to solve an environmental problem, but how did you do this? Did you need to learn enough about soil properties to help you understand the data insights you generated, or did you work based on empirical data-based evidence alone? In other words, how could a software engineer solve an environmental problem better than competing environmental data scientists? It’s remarkable!

Successful software engineers have "algorithmic" minds and this often helps them make good decisions when solving almost any type of complicated business problem. Algorithmic thinking makes it easier for software engineers to extend their programming skills in order to find shorter and faster ways to solutions. On the other hand, data scientists can understand data, and the hidden facts among data, better and faster than others. Experience is a data scientist's most valuable asset. Without enough experience, a smart data scientist is still like a little child.

Experience is a data scientist's most valuable asset. Without enough experience, a smart data scientist is still like a little child

So data scientists with programming skills, and software engineers with enough background knowledge in statistics and machine learning, are not confined to using only pre-built tools and packages. Rather, they can write their own codes to bring the flexibility and power of programming into data science.

Understanding the problem is the first and most important step in solving machine learning problems. It's helpful if one really takes time to understand the data and also studies related works where people were already successful in solving similar types of problems.

Uncertainty is a big factor in data science. There is no warranty on which person can do a data science task better than another. For example, a physician might solve an environmental problem better than both a software engineer and an environmental data scientist! All things are possible in this game.

Uncertainty is a big factor in data science. There is no warranty on which person can do a data science task better than another. For example, a physician might solve an environmental problem better than both a software engineer and an environmental data scientist! All things are possible

Can you share a bit about the machine learning algorithms that you used to solve this problem?

For the Africa Soil Property Prediction Challenge, I used R for training and prediction. Machine learning algorithms entered in two phases, as such:

1. Pre-processing phase - Some pre-processing transformations and filters were applied to the data before the modeling phase.

a) *Savitzky-Golay filter*: This filter is used for smoothing the data.
b) *Continuum removal*: For normalization and handling outliers.
c) *Discrete wavelet transforms*: For discrete sampling and data reduction.
d) *First derivatives*: In some cases, this increases prediction quality.
e) *Unsupervised feature selection*: Standard deviation was used to select top features for some algorithms.

2. Modeling phase - Four different types of modeling algorithms were trained to predict soil property.

a) Neural Networks
b) Support Vector Machines (SVMs)
c) Multivariate Regression
d) Gaussian Process

AfSIS team sampled data from most accessible regions of Africa. There were about 1200 training cases, which was too low compared to number of predictors (3600 predictors). The data size was too small, so many competitors had problems with over-fitting. I think the sponsors could have gotten better results and more stable models if they had provided a larger number of training instances.

More information about my code and documentation can be found at the Kaggle site here.

With these excellent skills in data science, the sky is truly the limit… What's next for you? What are your dreams and aspirations for the future of your career?

Thanks, but in my opinion my skills are not even good enough for me to be considered a "real" data scientist. There are many other smart people, both within and outside of Kaggle competitions. These people are sharing their ideas and making improvements to state-of-the-art methods. I just apply and try their works. There are many areas within data science where I still need to improve my skills.

Unfortunately, due to political sanctions on my country, I couldn't be paid the prize money that was offered for the Africa Soil Property Challenge. I don't let this disappoint me though. I love machine learning and data analysis. My big dream is to have a small role in the growth of data science.

About Yasser Tabandeh: Mr Tabandeh has a Master's degree in software engineering from Shiraz University in Iran. His thesis was on weighting methods for feature selection. He has about 10 years of experience working in software development and about 5 years of experience working in data science. He is currently employed at Golgohar Mining and Industrial Company in Sirjan, Iran. He can be contacted through LinkedIn or email.

Related Topics

Biostatistics -
General

Engineering -
Machine Learning

Environmental -
General
Environmental Modeling

Methods -
Multivariate
Data analytics

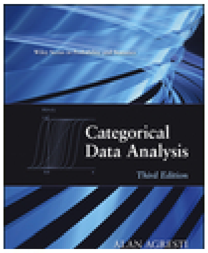
Social Sciences -
General

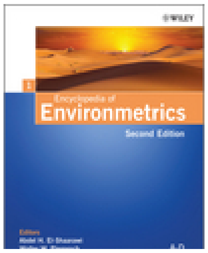
Related Publications

Books & Journals

Books
Journals

Books





ıvironmetrics, 2nd Edition, 6 Volume Set

ıalysis, 3rd E

tent

ote Big Data dev

: Climate Change

per 2014 issue ju

A point process model for tornado report climatologyjournal article

Stat

Significance

New Journal: Stat – The ISI's Journal for the Rapid Dissemination of Statistics Researchfeature
Presenting Data: Understanding the User by RSS Treasurer Ed Swires-Hennessyvideo
13th Islamic Countries Conference on Statistical Sciencesevent
The Future of Statistical Sciences Workshopwebinar

Statistics Views

Main Site Navigation