



(/KIKUYA-Takumi)

@KIKUYA-Takumi (/KIKUYA-Takumi) 2018年04月05日に投稿



データサイエンティストを目指して勉強した1年間まとめ

Python(/tags/python) R(/tags/r)

機械学習(/tags/%E6%A9%9F%E6%A2%B0%E5%AD%A6%E7%BF%92)

ポエム(/tags/%E3%83%9D%E3%82%A8%E3%83%A0)

データ分析(/tags/%E3%83%87%E3%83%BC%E3%82%BF%E5%88%86%E6%9E%90)

はじめに

本記事では、データサイエンスについて学んだこと、データ分析業務に携わって、経験したこと、気付いたことをまとめています。特に、後半を中心にまとめています。前半についてはこちらの「データサイエンティストを目指して半年で学んだことまとめ (<https://qiita.com/KIKUYA-Takumi/items/531edd4e62c3e14a6d08>)」に書いています。ご興味があれば、読んでいただければと思います。

全てはビジョン（あるべき/ありたい姿）を明確にしてから始まる



(<https://camo.qiitusercontent.com/c935a27f3e391111e279349da019965cb633a89d/68747470733a2f2f71696974612d696d6167652d73746f72652e73332e616d617a6f6e6177732e636f6d2f302f31>)

34383133362f35613366376662652d386664312d373239622d643062642d3863376532373364643434362e706e67)

データ分析で最も重要になるのが、**ビジョン（あるべき/ありたい姿）の明確度**にあると感じています。ビジョンが明確であるほど、課題・目的も明確に設定でき、課題解決のための仮説検証、必要なデータの準備と、ビジョンの実現に向けたデータ分析ができるようになります。勿論、ビジョンが明確であれば良いというものではないかもしれませんが（必要なデータが集められない等）が、少なくとも、意味のない作業を減らすことは可能だと考えられます。

逆にビジョンが明確でない場合、まず、**何をやれば良いかわからない状態**になる可能性が非常に高いです。とりあえず、手元にあるデータから何か言えないかデータをいじくった結果、「**このデータから〇〇という傾向がわかる！**」と思っても、「**で？、だから何？**」となります。結局、**ビジョン（ゴールとも言える）**が定まっていないと何も解決できず、「とりあえず、やってみた！」に終わってしまうのです。

また、注意したいのが**良い感じの出力結果から目的設定**するスタイル。特に初心者（自分含め）に多いのではないのでしょうか？予想していた結果ではない（そもそも何も考えていない場合も...）が、特徴的な結果が出たときに、それを元に分析してしまうパターンです。このパターンに陥ると、（個人的経験によると）「やりたいことは、それじゃない」となります。目先の結果に囚われず、**常に、ビジョン・目的を達成できるのか否か**を基準にしなければなりません。最近では、**その結果は、何を説明できるのか？**を自分に問いかけることで、ビジョン・目的を見失わないようにしています。

ビジョンが大事と書いたものの、必ずしも、自らビジョンを設定できるとは限らないのが現状です。分析依頼者から「良い感じにして欲しい」と言われる場合は、「良い感じ」を聞き出すことで、ビジョンを明確にしていくことはできるかもしれませんが、**とりあえずAIで何かやってほしい！**と言われる場合は言っている人・タイミングの見極めが肝心、というのが個人的な見解です。

攻略！「とりあえずAI」

自分の知っている範囲ではありますが、「とりあえずAIで何かやりたい！」にはパターンがあると考えています。

とりあえずAI～AIは手段の一つ～

AIは手段の一つと考えているパターン。「AIでやると何が違うのか」を説明し、課題解決ができるのか、試す価値があるのかを理解してもらえると良い関係を築けるという印象。「AI（機械学習？）とは何か？どんなものを想像しているのか？」等、お互いのギャップを埋めつつ、ビジョンを設定していけるか、**対話次第**であると考えています。

機械学習等、いわゆる高度な分析でなくても、課題が解決できれば良いという印象が強い。

とりあえずAI～「AIでやってます」と言いたい～

次は、「AIでやってます」と言いたいパターン。特に、課題があるわけでもなく、何かに困っているわけでもないが「AIやってます」と言える何かをやりたい人達。「無理なら仕方がない」と思っている人もいる模様。ビジョンや課題は**こちらから提案**していく印象が強い。協力が必要な時に対応してもらえるかがポイントかなと。

データ収集が予想以上に大変と気づき、やめる印象があります。

とりあえずAI～AIは俺の意図通りの結果を出すんだろ？～

最後は、いわゆる**AI万能説**。特に、強い印象があるのが、「**Deep Learning**でやれば解決できるんでしょ？」という**Deep Learning特化型**と、「俺の求める分析結果を出して！AIのできるんでしょ？」という**シンプルなAI万能説信者**の2パターン。

Deep Learning特化型の特徴は昨今のAIブームの影響が非常に大きいというのは明らかです。確かに、Deep Learningが大きな成果を上げているのは事実ですが、どんな課題に対しても有効とはいえないこと、成果を出すだけのデータの量と質を用意できるのかという大きな課題があります。が、これらの課題を無視してでも、「とにかくDeep Learningでやりたい、ほかの手法ではダメだ、Deep Learningだ！」という人たちがこちらになります。もっと他にやるべきことがあるだろうに...

シンプルなAI万能説信者は、最近（自分の周りでは）減ってきている気がしますが、印象は一番強いです。特に、印象的だったのが、「データはあるんで、要求通りの予測結果出して！」の一言。思い通りに結果出るなら分析いらない説。

これらの**AIは自分の思い通りの結果を出す都合の良いもの**と考えている人はAIが全てやってくれろと思いついていて、**何を達成/解決したいか**という話にならないので、可能な限り全力逃走か、うまく回避したいところです。ただただ、消耗するだけです。

絶対回避！「とりあえず、やる」

自分で手を動かす時に避けたいのが**とりあえず、やる**。何をやれば良いかわからないときや、勉強したアプローチを使ってみたいときにやってしまうのですが、**結果の説明ができない**。「とりあえず、やってみた」とか「こんな結果出んですけど、どうですか？」くらいしか言えません。そして、注意したいのが**とりあえずやってみた結果が、稀に刺さる**ということです。

何か面白い結果が出た！と盛り上がってしまい、どんどん進めていった結果、**当初の目的を完全に見失う**なんてことも...結果が面白いかどうかには囚われてはいけなく、目的を達成できたかどうかを基準にしなければいけない。

「とりあえず、やる」をやめて、解決すべき課題は何か、何を主張できれば解決できるか、主張するためには根拠として何が必要か、その根拠はデータか導出できるか、**やってみると**、仮説検証をしっかりとやることで、「何か面白い結果」に左右されなくなってきた今日この頃であります。

前半と後半の違い

前半の半年は上記記事にも書いたように、機械学習を、アルゴリズムをどうやって活用するかを中心に学びました。技術検証を中心にやっていたこともあって、いわゆる**手段**として正しい使い方を勉強しました。一方、後半の半年は実案件にも携わるようになり、それまでの、データ、プログラム、数学・数式だけでなく、**顧客**とも向き合うようになり、**ビジョン**、**目的**の重要性を感じました。

上にも書いてますが、ビジョン・目的があって、それを達成するためにどんな手段が有効か？を考えるようになり、どのアプローチ・アルゴリズムを使うかではなく、何を主張できれば良いか（説得できるか）を先に決めるようになってきた気がします。

また、**コミュニケーション能力**は重要で、特に、最初の目的・課題設定を曖昧にしまうとできるものもできなくなったり、無駄なタスクが増えるので、目的、課題を具体的に設定する、分析結果から次のアクションへとつなげるための対話力が必要になります。

衝撃の一言 3選

- **お客さんがAIで何かやりたいと言っている、自分たちもAIで何かやりたい！意向がマッチしたから、よろしく！**

シンプルに**何か is 何**案件。結局、お客さんの業界調査して、課題を見つけ企画、提案しましたが...
一体、打ち合わせで何を議論したのか気になるところ...

- **データはどうでも良いから、Deep Learningで分析して！予測結果に対する原因とかわかれば良い！**

完全にDeep Learning万能説。そもそもデータを適当に用意（自作）して、どんな結果（原因含め）ができるか知りたいと言われましても、結果に何も意味ありませんよ？という話、しかもDeep Learning。やることに意義がある的なことを言われた記憶が...

- **お客さんに従来の分析とは違う分析を提案したい！今？Excelで何かやってるんじゃない？**

これは完全にコミュニケーションエラーです。**どういう目的の分析(what)**を聞いているのに、返事が**分析手段(how)**になっています。Excelでやった結果とRやPythonでやった結果が異なる(確率的要素は除く)ってそもそも何か問題があるのでは...という話。

これらから言えることは、分析依頼者、分析者の間に誰か（営業やマネージャー）が入る場合、その人も分析、データサイエンスの知識がないと余計な手間が増えるということです。最も避けたいのは、「この前〇〇ツールでこんなことができるって話したからこのツール使って！」という、無謀なるツール縛り。

データ分析では、データクレンジング・前処理という泥臭いタスクが大半とされていますが、データ分析をするために関係者を説得するというさらに泥臭いタスクがある場合もあるという現実と向き合っている今日この頃であります。

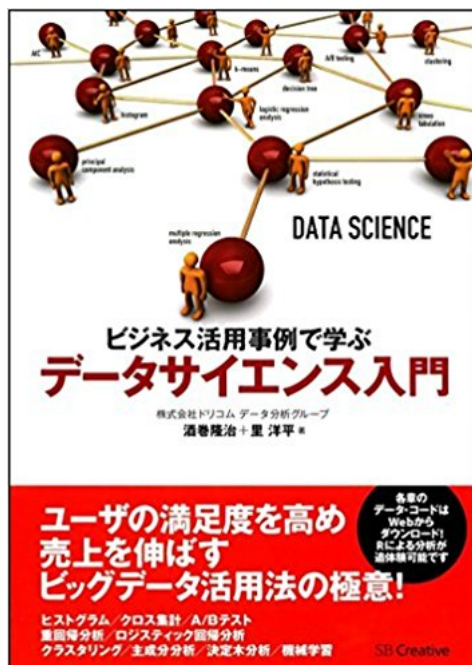
使用した参考書



(<https://camo.qiitusercontent.com/96123b964>)

6a46f515df72924e70512260796e783/68747470733a2f2f71696974612d696d6167652d73746f726
52e73332e616d617a6f6e6177732e636f6d2f302f3134383133362f38343862323435362d66386538
2d633765342d386638322d6533643266616437333266662e706e67)

文字通り、データ解析初心者向けの本。データ解析のプロセスを理解するのにおすすめ。個人的には分析者だけでなく、分析に携わる人（営業、マネージャー等）にも読む価値があると思っています。



(<https://camo.qiitusercontent.com/f57f9aa91>)

52038dc2267c8d4c9146dea142ecbef/68747470733a2f2f71696974612d696d6167652d73746f726
52e73332e616d617a6f6e6177732e636f6d2f302f3134383133362f31356134376364612d38386238
2d663636612d303362662d3431386466336331616633632e706e67)

こちらもプロセスを理解するために役立つ1冊。Rのサンプルコードもあるので、手を動かしながらできます。

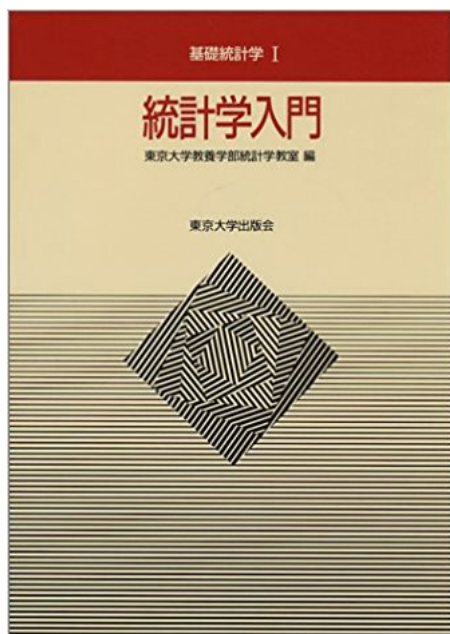


(<https://camo.qiitausercontent.com/d9df247f3>)

eea8dc86063b4c70f67972e128d4006/68747470733a2f2f71696974612d696d6167652d73746f726
52e73332e616d617a6f6e6177732e636f6d2f302f3134383133362f32383736656563662d32343762
2d633737612d383830342d3863366532316265353265362e706e67)

機械学習プロジェクトの進め方が書かれた1冊。機械学習のアプローチを理解してから読む本と
いう位置付け。

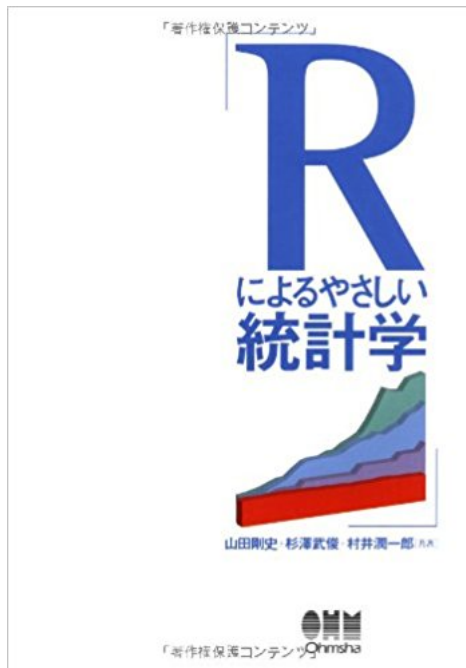
上記3冊は手法は理解したが、実践はどうすれば良いかわからないって人におすすめです。



(<https://camo.qiitausercontent.com/52567ac7ae8>)

f14ed77f17a700bb3a308388a6a25/68747470733a2f2f71696974612d696d6167652d73746f72652
e73332e616d617a6f6e6177732e636f6d2f302f3134383133362f64656436356635662d643537352d
623062392d333465652d6232653661333331376663312e706e67)

統計学でおすすめの本といえば、この1冊ではないでしょうか？



(<https://camo.qiitausercontent.com/6d8733ca8>

5d80d2680fb07e0f4cca74a66889ee9/68747470733a2f2f71696974612d696d6167652d73746f7265
2e73332e616d617a6f6e6177732e636f6d2f302f3134383133362f62616438313136342d316265322
d616266322d343035312d3433346234383832653738352e706e67)

統計学勉強ついでにRの勉強もできる1冊。



(<https://camo.qiitausercontent.com/cfbd18692>

826866e321969795588ba012acf6d4b/68747470733a2f2f71696974612d696d6167652d73746f726
52e73332e616d617a6f6e6177732e636f6d2f302f3134383133362f64306561663064352d66623436
2d633836342d353036332d3231303036653332636166612e706e67)

データサイエンスのための統計学ということで、従来の統計学では行うが、データサイエンスの分野ではやらない等、現代の統計学？とも言える内容。関連語なども書かれており、キーワードの整理にも役立つ1冊。



(<https://camo.qiitusercontent.com/e53f6519>)

b338d40a1c88ac9a8a3c8810c559574c/68747470733a2f2f71696974612d696d6167652d73746f72
652e73332e616d617a6f6e6177732e636f6d2f302f3134383133362f31313436383265362d6464393
32d383638622d643131332d3539633038393664396461662e706e67)

モデルの使い分けを学べる1冊。Rのサンプルコードあり。



(<https://camo.qiitusercontent.com/4f28>)

b5535385434be28f7c5dd5d9826e44b924fc/68747470733a2f2f71696974612d696d6167652d7374
6f72652e73332e616d617a6f6e6177732e636f6d2f302f3134383133362f39383139323161392d6539
30312d636134332d313661632d3966306531346335363233372e706e67)

Rは書くよりも読む方が圧倒的に多いので、書く練習のためにやってみました。



(<https://camo.qiitusercontent.com/205a1730>

b9092021d5e36750cad92ebeab29da10/68747470733a2f2f71696974612d696d6167652d73746f72652e73332e616d617a6f6e6177732e636f6d2f302f3134383133362f35656266646264622d613639382d326639662d373931332d3931316331313832663035632e706e67)

時系列モデルを扱う機会があったので、読んでみました。何気に実践→理論は初めてでしたが、「あのコードはこういう意味だったのか！」と気があったので、実践から入るのも悪くないと感じた1冊でした。



(<https://camo.qiitusercontent.com/bc5fb6e7f>

ef6e9308200b754759fd6f59a8790ea/68747470733a2f2f71696974612d696d6167652d73746f72652e73332e616d617a6f6e6177732e636f6d2f302f3134383133362f66323261336436652d336636642d633031392d623663362d6437666639386231343139312e706e67)

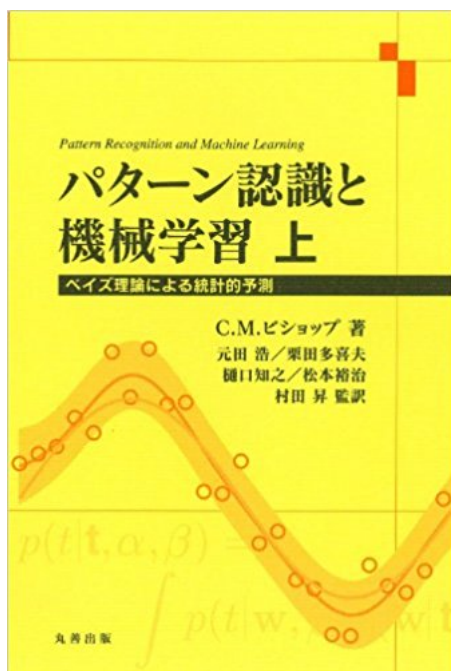
ログ解析で異常検知をやることになったので取り急ぎ読んだ1冊。モデルはまだ作っていない。



(<https://camo.qiitusercontent.com/ac0bec133>

98affad1da65b49fc31216cb862f99e/68747470733a2f2f71696974612d696d6167652d73746f7265
2e73332e616d617a6f6e6177732e636f6d2f302f3134383133362f35383264363265352d616539322
d663239612d386131372d3539656130356534383631302e706e67)

ベイズを勉強しようと思って、読みました。尚、確率論が弱い模様。



(<https://camo.qiitusercontent.com/dec4f6bfc13a>

9d850467878ca26480dc0dce7d05/68747470733a2f2f71696974612d696d6167652d73746f72652e
73332e616d617a6f6e6177732e636f6d2f302f3134383133362f32363133383861312d656661622d6
53137302d666631302d353637616636366333538612e706e67)

最後は、みんな大好きPRML。やはり確率論が弱いことに気付く。行列あたりも怪しいことに気
付く。一方で、何度も数式にだいが慣れていることにも気付く。少しずつだが、数式から処理を
イメージできるようになった気がする。

何冊も読むと似た内容があったり、本を読む→実装→再度、本を読むと最初わからなかったことが、2回目は理解できたり、気付きもなかったことに気付くようになったりと数をこなしたことで学んだことが多くあるはずなのに、やればやるほど、次やるべきことが見つかっていく...
改めて、多くのスキルが求められる分野だと実感しました。数学がプログラミングが苦手とか言っている場合ではない。

さいごに

1年間勉強してきてわかったことは、**ビジョン・目的を具体的に設定する力**も分析スキルと同様に必要であるということです。また、**次のアクションにつなげる主張**をデータ分析から導き出すことができるかが今後一番の（永遠の？）課題であると考えています。カギとなるのは**Feature Engineering**だと思っています。

では、これにて1年終了。

続


✎ 編集リクエスト (<https://qiita.com/KIKUYA-Takumi/items/162612ca42a9318cb1d8/edit>)

📁 ストック

✓ いいね済み

351

(<https://qiita.com/KIKUYA-Takumi/items/162612ca42a9318cb1d8/likers>)




(/KIKUYA-Takumi)

@KIKUYA-Takumi (/KIKUYA-Takumi)

データ分析・データサイエンスをやっています。自然言語処理もやることに... 発言は全て個人の見解であり、所属している企業、組織を代表するものではありません。

+ フォロー



(/organizations/tis)


TIS株式会社 (/organizations/tis)

創業40年超のSIerです。

<https://www.tis.co.jp/> (<https://www.tis.co.jp/>)

📌 コミュニティスポンサー広告 (<http://blog.qiita.com/post/176483510744/community-sponsor>)

関連記事 Recommended by (<https://www.logly.co.jp/privacy.html>)



by KIKUYA-Takumi

データサイエンティストを目指して半年で学んだことまとめ

(<https://qiita.com/KIKUYA-Takumi/items/531edd4e62c3e14a6d08>)

**AI,機械学習,Deep Learning わかりにくい流行り言葉をまとめる**

(<https://qiita.com/tomohiku/items/364a1b06b35ea4d514d6>)
by tomohiku

**ビジネス→エンジニア世界に入りたての私が実践したデータサイエンス系の学習したこと一覧**

(<https://qiita.com/lshio/items/a9fc377fef3f39bd110b>)
by lshio

**データサイエンスプロジェクトの成功に今一番必要な人**

(<https://qiita.com/KanNishida/items/1b3f25ff11e1d5a74353>)
by KanNishida

**"副業案件" を多数紹介！ Rails、Reactの即戦力エンジニア求む！** (https://dsp.logly.co.jp/click?ad=G8GmQLsVMLaQmqAPR_ilNSoWJ3oUjY2-VMXzmQL7tWqiEilL8q7lJxlx3T_uiR3Kgz-XXWieQm09ojkfsXrG3z27sGffbaNFyimOIUHEu1TvwqCex6f2hddfbco9pCXi5VO8GtKXas7S5MFwDO0GpddSfqRgZWx02kdBeotbDNU1TmCjOzADPVNWUEtxXch0hmFqJikAi5wmELi5Apc0bj9V1cVyGu8J8PylU2PObBzpWAHKXBLrlIqdZRHq0kdya5Z1bzipyaVDIQTBot7vjksDaMcalPoBUI4uPYUm2q_Uh6W9Esl2Uk)

PR シューマツワーカー

🔗 この記事は以下の記事からリンクされています



ブックマークしてあった、データサイエンスなどの記事約1年分のリンク集（2018年5月ごろまで）
([/simonritchie/items/e182b0eb54604518023f#_reference-6c02a4927842bb14fa5b](https://simonritchie/items/e182b0eb54604518023f#_reference-6c02a4927842bb14fa5b)) からリンク 10ヶ月前