

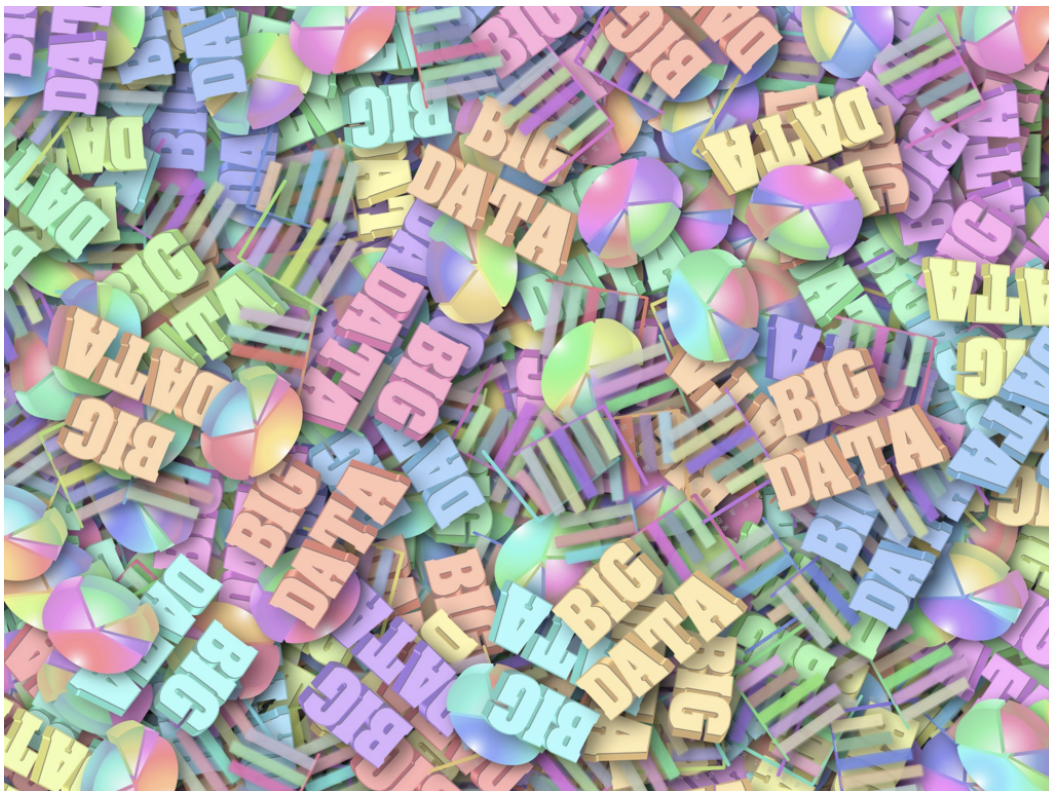
六本木で働くデータサイエンティストのブログ

元祖「銀座で働くデータサイエンティスト」です / 道玄坂→銀座→東京→六本木

2019年版：データサイエンティスト・機械学習エンジニアのスキル要件、そして期待されるバックグラウンドについて

データサイエンティスト 機械学習エンジニア 統計学 機械学習 データ分析 人材 ビジネス キャリア

2019-02-19



(Image by Pixabay)

この記事は、以前の同様のスキル要件記事のアップデートです。



六本木で働くデータサイエンティストのブログ

id:TJO

データサイエンティスト&機械学習（人工知能）エンジニアのスキル要件と、過熱する人...

(Image by Pixabay)この記事は去年はてブ1100以上ついてしまった与太記事の続編です。その時はタイトルを読んで字の如く「データサイエンティスト」と「機械学習エンジニア」の満たすべきスキル...

2018-02-07 19:00 ★7 550 users

Hatena Blog



正直言って内容的には大差ないと思いますが、今回は2つ新たな軸を加えることにしました。一つは「ジュニアレベル（駆け出し）」と「シニアレベル（熟練職人）」とで分けるということ、もう一つは「データ分析以外の業界知識（ドメイン知識）」にも重きを置く、ということです。

プロフィール



id:TJO

Takashi J. OZAKI, Ph.D.
Data Scientist

English: <https://tjo-en.hatenablog.com/>

ブログの内容は個人の意見・見解の表明であり、所属組織の意見・見解を代表しません。またブログ内容の正確性については一切保証いたしません（誤りを見つけた場合はコメント欄などでお知らせいただけると有難いです）。

また、ブログの中で取り上げられているデータ分析事例・データセット・分析上の知見など全ての記述はいずれもいかなる実在する企業・組織・機関の、いかなる個別の事例とも無関係です。ブログ記事内容は予告なく公開後に改変されることがあります。改変した事実は明示されることもあれば明示されないこともあります。

現在、講演依頼・書籍執筆依頼・メディア出演依頼等は全てお断りしております。悪しからずご了承ください。



ビジネスに活かすデータマイニング

作者: 尾崎隆

出版社/メーカー: 技術評論社

発売日: 2014/08/22

メディア: 単行本 (ソフトカバー)

[この商品を含むブログ \(6件\) を見る](#)

というのも、空前の人工知能ブームが予想よりも長く続いていることで、人材マーケットを観察する限りではデータサイエンティスト・機械学習エンジニアとも求人数が高止まりしているように見えるのですが、その結果としてこのブログの過去のスキル要件記事で挙げたような「完成されたデータ分析人材（熟練職人）」に限らず「駆け出し」でも良いからデータ分析人材が欲しいという企業が増えているように感じられるからです。

その一方で、かつては主にwebマーケティング業界に集中していたデータ分析専門職が、今や非常に幅広い相異なる業界に多岐に渡って分布しており、むしろバックグラウンドとなる業界ドメイン知識についても細分化していくこと、そしてそれによるアドバンテージも考慮する必要が出てきているようにも思います。今回の記事ではその点についても少し触れています。

いつもながらの断り書きですが、言うまでもなくここに並べた内容はあくまでも僕の個人的な意見にして、なおかつ僕自身がこれまでの経験と見聞に基づいて「これまで自分が属してきた組織やチームにおけるデータサイエンティストや機械学習エンジニアがこうであったら良かったかも」という最大公約数的な願望を書き並べたものに過ぎません。よって何かの組織や団体の意見を代表するものではありませんし、況してやauthorizeされた意見として見られるべきものでもないという点、予めご了承くださいm(_ _)m
願わくば、[前回記事でリンクした榊さんの記事](#)のように「うちの組織ではこういう考え方をしている」というようなご意見（ご異見）をお寄せいただけると有難いです。

また、機械学習エンジニアのスキル要件については僕自身が現在エンジニア部門に所属していないこともあり、チェックも兼ねて友人知人で実際に機械学習エンジニアとして働いている人たちにレビューしてもらっており、実際スキル要件の内容にもレビューコメントを反映させています。ただし、それでもその人たちの経験と見聞の範囲に留まるという点にご留意ください。

- [基本的な考え方](#)
- [ジュニアレベル（駆け出し）のスキル要件](#)
- [シニアレベル（熟練職人）のスキル要件](#)
- [細かい説明など](#)
 - [ジュニアレベル](#)
 - [シニアレベル](#)
 - [プログラミングスキルに関する要件は？](#)
 - [「多々ますます分ず」では現実離れしやすい](#)
 - [「技術的スキルなんて適当で良い」ではおしゃべり課題解決コンサルおじさんに墮する](#)
- [最後に](#)

基本的な考え方

前回のスキル要件記事ではまとめて論じたポイントですが、今回はスキル要件とは別に以下のように明示しておきます。

- [データサイエンティスト：「アナリスト」の発展版](#)
- [機械学習（人工知能）エンジニア：「エンジニア」の発展版](#)

異論があることは承知していますが、僕個人の考えでは「データサイエンティストはアナリスト」「機械学習エンジニアはエンジニア」だと見ています。言い換えると、前者は「アナリストの仕事に統計学や機械学習を持ち込んだもの」、後者は「エンジニアの仕事に機械学習を持ち込んだもの」という理解です。

プロフィール：[LinkedIn](#)

ご質問など：[Quora](#)

業績一覧：[Google Scholar Citations](#)

科研費情報：[KAKEN](#)

ご連絡は出来るだけLinkedInメッセージでお願いいたします。

Copyright © Takashi J. OZAKI
2013 All rights reserved.

読者になる 2488

検索

記事を検索

カテゴリー

[データサイエンティスト \(73\)](#)

[人材 \(27\)](#)

[機械学習 \(130\)](#)

[機械学習エンジニア \(14\)](#)

[キャリア \(9\)](#)

[統計学 \(103\)](#)

[ビジネス \(55\)](#)

[データ分析 \(69\)](#)

[生TensorFlow七転八倒記 \(11\)](#)

[TensorFlow \(11\)](#)

[Python \(29\)](#)

[私事 \(7\)](#)

[昔話 \(6\)](#)

[雑感 \(13\)](#)

[Deep Learning \(19\)](#)

[R \(126\)](#)

[時系列分析 \(35\)](#)

[書評 \(23\)](#)

[書籍 \(29\)](#)

[旅行記 \(12\)](#)

[論文 \(10\)](#)

[マーケティング \(12\)](#)

[BUGS/Stan \(18\)](#)

[エイプリル fools \(5\)](#)

[統計的因果推論 \(10\)](#)

[研究 \(2\)](#)

[graph/network \(8\)](#)

アナリストは「オンデマンドで情報を分析し結果をレポートして意思決定に貢献する」のが仕事で、エンジニアは「システムを開発して何かしらの自動化されたアウトプットを出して事業に貢献する」のが仕事だとすれば、それぞれを統計分析や機械学習によってさらにブーストさせる。それが、データサイエンティストまたは機械学習エンジニアという仕事なのかなと考えています。

ただし、先に「データサイエンティスト」という職種がブームになり、その後から「機械学習（人工知能）エンジニア」という職種が空前の人工知能ブームに伴って人気を呼ぶようになったという経緯があるため、前者はある程度ざっくりとした「データ分析」全体を担うことが多い一方で後者は割とエンジニア領域に特化していることが多いというのが僕の認識です。この点を踏まえた上で、スキル要件について触れてみたいと思います。

ジュニアレベル（駆け出し）のスキル要件

まず「駆け出し」とも言えるジュニアレベルのデータサイエンティスト・機械学習エンジニアのスキル要件について書いてみます。ここで想定しているのは、「新卒でデータ分析職に就こうというキャリア初期の若手」であったり「他のキャリアからデータ分析に関連する勉強をしてきてデータ分析職としてのキャリアに新たにジョブチェンジしようとしている人たち」と言った層です。

データサイエンティスト

- 1. 一般的なアナリストとしてのスキル
 - BIツールなどを用いたインサイトレポートが出来る
 - A/Bテストなど効果検証とそのデザインが出来る...etc.
- 2. 東京大学出版会の統計学シリーズ3巻分に該当する統計学の知識
- 3. はじパタに該当する一般的な機械学習の知識

機械学習（人工知能）エンジニア

- 1. 一般的なエンジニアとしてのスキル
 - システム設計が出来る
 - テストや運用が出来る
 - システム開発手法に秀でている...etc.
- 2. 代表的なフレームワーク（scikit-learnやTensorFlowなど）を使って機械学習を扱える程度のコーディング力と知識*1
- 3. Goodfellow本レベルのDeep Learningの知識

二者共通の要件

- 1. SQL文法を含むデータベース操作の技術
- 2. クラウドの知識
- 3. データ前処理・特徴量エンジニアリングの技術
- 4. 何かしらのビジネス領域における若干年数の実務経験

一応、「これだけの知識があればある程度現場で手を動かしながら一般的な統計分析and/or機械学習の業務ができる」レベルのスキル要件を挙げてみました。総論としては、エントリーレベルのデータサイエンティストや機械学習エンジニアが担う業務というと「アナリストやエンジニアとしてのタスクにデータ活用が加わった」ぐらいのレベル感かなと思われるので、「既存業務+アドオンとしてのデータ分析」という枠組みの範囲内でデータ分析業務をこなすに足るだけのスキルセットを並べてみた次第です。

DeepLearning実践シリーズ (5)

MCMC (10)

異常検知 (4)

SQL (1)

DLM (7)

お知らせ (8)

サンプルデータで試す機械学習シリーズ (16)

データマイニング (31)

データ分析実践編 (4)

PR (4)

最適化計画 (2)

Matlab (2)

カンファレンス (8)

状態空間 (1)

テキストマイニング (1)

日常 (3)

アナリティクス (10)

Hadoop (2)

Hive (2)

Excel (1)

自己紹介 (2)

Facebook (2)

IT (1)

最新記事

2019年版：データサイエンティスト・機械学習エンジニアのスキル要件、そして期待されるバックグラウンドについて

生TensorFlow七転八倒記 (10)：テキストデータをTF-Hubでfeature vectorに直してからt-SNEにかけてみる

研究者を辞めた時のこと、そしてその後のこと

生TensorFlow七転八倒記 (9)：TF-Hub embeddingを利用して感情分析してみる

単純なK-meansと{TSclust}のDTWによる時系列クラスタリングとではどう違うのか実験してみた

機械学習システム開発や統計分析を仕事にしたい人にオススメの書籍初級5冊&

シニアレベル（熟練職人）のスキル要件

そして、「熟練職人」としてある程度指導的立場に立つこともあり得るシニアレベルのスキル要件です。こちらはジュニアレベルのスキルがあることを大前提とした上で、そこに専門家及びシニア人材としてのバリューをどれだけ上乗せしていけるかという点を重視しています。

データサイエンティスト

- 1. ジュニアレベルのスキル全て
- 2. [ベイジアン統計モデリング](#)の知識と確率的プログラミングのスキル*2
- 3. ジュニアレベルの機械学習エンジニアの機械学習に関するスキル
 - 代表的なフレームワーク（scikit-learnやTensorFlowなど）を使って機械学習を扱える程度のコーディング力と知識
 - Goodfellow本レベルのDeep Learningの知識
- 4. 統計的因果推論の知識と技術*3
- 5. 統計分析をアナリスト業務に用いる上で生じる解釈の問題や意思決定プロセスへの関与の仕方について詳しいこと

機械学習（人工知能）エンジニア

- 1. ジュニアレベルのスキル全て
- 2. [講談社MLPシリーズ](#)・黄色い本(PRML)及びカステラ本(ESL)に相当する汎用的な機械学習の知識
- 3. 各種トップカンファレンスや[arXiv](#)の論文含めてDeep Learning諸系統の最先端の研究開発動向に詳しいこと
- 4. [機械学習](#)を実システムで運用する上で生じるシステム設計の問題や技術的負債及びバイアスについて詳しいこと

二者共通の要件

- 1. 一般的なシニアアナリストorシニアエンジニアとしてのスキル
- 2. 専門とするビジネス領域において各種データ分析手法を適用してきた & 適用せずにあえて見送った実務経験

「これだけのスキルがあれば、個々のデータ分析の現場で『一人前』として戦うことができ、場合によっては指導的立場に立ってチームを引っ張ることが出来る」レベルのスキルをリストアップしてみたつもりです。このレベルになるとズバリ「[データ分析の特質・特性を活かした新規プロジェクトをゼロから立ち上げる](#)」ことが求められると思うので、それを可能にするスキルセットとなると大体上記のような感じになるはずです。

細かい説明など

TL;DR 以下、個々のスキル要件についての細かい説明を書いていきます。気が付いたら物凄いボリュームになってしまったので、超長文ご容赦くださいm(_ _)m

ジュニアレベル

イメージとしては、データサイエンティストであれば一般的なA/Bテスト・効果測定分析・回帰モデルなどの統計分析や、ちょっとした機械学習を用いた要因分析を行う感じでしょうか。ロジスティック回帰でコンバージョンへの寄与要因を探し出すとか、ランダムフォレストなどの[非線形分類](#)手法で変数重要度を

中級10冊 + テーマ別9冊
(2019年1月版)

終わりのなき学びと、社会実装と

『新版 統計学のセンス』は統計学を「使う」人なら必携の書

データサイエンティストや機械学習エンジニアが、可能な限り統計学や機械学習やプログラミングを使って課題を解決するべき3つの理由

シンガポール旅行まとめ (2018年秋版)

リンク

Kaggle

人工知能に関する断創録

AnalyticBridge

月別アーカイブ

- ▼ 2019 (6)
 - 2019 / 2 (2)
 - 2019 / 1 (4)
- ▶ 2018 (32)
- ▶ 2017 (33)
- ▶ 2016 (33)
- ▶ 2015 (44)
- ▶ 2014 (61)
- ▶ 2013 (100)

忍者アナライズ

見るなんてこともあるかと思われます。機械学習エンジニアであれば、一般的な機械学習全体の知識があった上でさらにある程度Deep Learningの実装をやれた方が良いのかなと。というのも、日進月歩を通り越して秒進分歩の勢いでDeep Learningは研究としても技術基盤としても進歩しており、特に画像認識や自然言語処理などを扱う場合にはいきなりDeep Learningを投入してしまう現場も増えているように見聞するためです。

加えて、二者の共通要件として以前から挙げている「DB基盤技術（SQL含む）」「クラウド」「前処理&特徴量エンジニアリング」を入れています。これらに関してはもはや論を俟たないでしょう。データを武器として戦う専門職なので、DBを使いこなせれば困りますし、データが膨大で扱いづらい場合にクラウドでスケールさせるということが出来なければ仕事にならないことも多いはずです。前処理と特徴量エンジニアリングは「これがなければモデルが回らない」代物なので、ジュニアレベルの時点で身につけて欲しい大切なスキルの一つです。

そして、データサイエンティストを志すならば「アナリスト」としての経験が、機械学習エンジニアを志すならば「エンジニア」としての経験が、それぞれあった方が良いでしょう。これは先に述べた通りです。

まずデータサイエンティストについては、アナリストとしての一般的な分析業務が出来ることを前提としています。BIツールなどダッシュボードに基づくインサイトレポートや、A/Bテストなどの効果検証の運用とそのデザインが出来て欲しいところです。機械学習エンジニアについてはレビューコメントで「そもそもエンジニアとしてシステム開発の仕事が出来なければ機械学習に関わる仕事もこなせない」という指摘があったため、先にエンジニアとしての一般的なシステム開発スキルを持ててきてあります。これは僕自身がプロトタイプを書いたシステムがプロダクトに移行していく過程を見ても納得がいく話で、POCやプロトタイプとして「一応動くもの」とプロダクトとして「本番で動くもの」とは勝手が全く違います。

ちなみに機械学習エンジニアの方には「Goodfellow本レベルのDeep Learningの知識」を挙げてありますが、要はこの本を一通り読んでいるということですね。



深層学習

作者: Ian Goodfellow, Yoshua Bengio, Aaron Courville, 岩澤有祐, 鈴木雅大, 中山浩太郎, 松尾豊, 味曾野雅史, 黒滝紘生, 保住純, 野中尚輝, 河野慎, 富山翔司, 角田貴大

出版社/メーカー: KADOKAWA

発売日: 2018/03/07

メディア: 単行本

[この商品を含むブログ \(1件\) を見る](#)

これはレビューコメントで提案されたものですが、個人的にも納得がいきます。というのも、Deep Learning周りの「基礎知識」を「全て」網羅的に概観しようとする、これ一冊で済むという本は正直言ってなかなかありません。その点Goodfellow本なら大体のところを概観できるので、レベル感としてはちょうど良いと思います。

で、ここからが過去のスキル要件記事とは異なる点で、もう少し具体的に「〇〇のアナリスト」「〇〇のエンジニア」という「出自」をもっと重視しても良いのかなと考えています。例えば、僕の現在の専門分野はデジタル広告・デジタルマーケティング・広告マーケティング戦略なので、僕がデータサイエンティストを探すとすればやはり「広告業界のアナリスト」経験者を、機械学習エンジニアを探すなら「アドテク分野のエンジニア」経験者を、それぞれ選ぶことになると思います。

こう言った「データ分析専門職としてどういうバックグラウンドを専門に持っているかを重視する」という流れは恐らくどんな業界にも多かれ少なかれあるはずで、個人的に見聞する範囲でも

- ・ ゲームアプリ
- ・ SNS
- ・ Eコマース
- ・ ロボティクス
- ・ 戦略・ITコンサルティング
- ・ 金融（保険・ヘッジファンドなど）
- ・ 会計・監査
- ・ 医療・創薬・ヘルスケア
- ・ リスク管理
- ・ 製造業（自動車・航空機エンジンなど）

などの分野で、個別にそれぞれの業界に詳しいデータサイエンティストや機械学習エンジニアを探し求めている印象があります。この中で、例えばヘルスケアのスタートアップでデータサイエンティストを探すとなれば、ある程度ヘルスケア分野に詳しく何かしらの業界経験（製薬企業や医療機器メーカーでの勤務歴など）を持っている候補者を探す、という流れになるのはごく自然なことだと思われます。

元々の専門領域に近い分野であれば、より正確で緻密で有益なデータ分析が出来るというのはどう見ても当然のことです。会計・監査分野の出身者がいきなり広告マーケティング戦略分野のデータ分析を任せられても一体どこを分析すれば良いか見当もつかないことの方が多いと思いますが、広告マーケティング分野の出身者であればアトリビューションとかブランド効果といったmetricsについて検証&分析を行えば〇〇が分かるという「勘所」が物を言うことでしょう。

ということで、データサイエンティストにせよ機械学習エンジニアにせよ、元々の専門領域に近いデータ分析を手掛けてこそそのバリューという側面があり、バリューを出すためにこそ元々の専門領域を生かしたデータ分析業務を手掛ける、という方向性を重視するべきだと考えています。


シニアレベル

まずデータサイエンティストですが、強調したいのは3点です。「ある程度機械学習も扱える」ことと「統計的因果推論に長じている」こと、そして「統計分析の結果をビジネスに反映させる術を知っている」ことです。一点目の理由は簡単で、データサイエンティストとして統計分析メインの仕事をしていても中にはある程度「これなら自動化すればいいじゃないか」というものが出てきます。そういう時に「細かく統計分析するよりも機械学習でざっくり自動化した方が効率的」というケースがままあり、そこである程度機械学習システムに移行させるための若干の開発作業が出来る*4ことが重要だと考えています。

二点目の統計的因果推論ですが、これについては過去にシリーズ記事を組んだのでそちらもお読みいただければ良いかなと。

統計的因果推論 カテゴリーの記事一覧 - 六本木で働くデータサイエンティストのブログ

元祖「銀座で働くデータサイエンティスト」です / 道玄坂→銀座→東京→六本木

 tjo.hatenablog.com



これは機械学習でも当てはまることですが、特にマーケティング分野の統計分析で扱うようなデータは

往々にして何かしらの交絡要因に冒されていることが多く、適切に対処しないと間違った結論に達することがあります*5。しかしながら、因果推論のアプローチはまだまだオールインワンのパッケージにしてクリック一発で全て解決というわけにはいかない要素が多いため、面倒でもある程度体系的に因果推論について知っておく必要があると考える次第です。

三点目は「シニア」の立場であれば絶対に出来なければいけないこととして挙げています。往々にして生じるのが「統計分析結果の解釈の問題」。何であれ、統計分析した結果をビジネス上の意思決定層に正しく伝えるのは、経験者なら分かると思いますが至難の業です。加えて「統計分析で意思決定プロセスに貢献する」となるとデータサイエンティスト個々人ではどうにもならないことがあり、多くの場合で意思決定プロセスへの関与の仕方自体を考えなければいけないものです。これは必ずしも技術的なものだけでなく場合によっては「政治的」な工夫も必要になったりしますが*6、そこまで考慮した上で最適なエビデンスが得られる統計分析アプローチを選んだり、可視化を工夫したり、ストーリーを練る、と言ったデータサイエンティスト側で出来る工夫を行うこともまた重要です。


なお優先度という意味ではそこまで高くしませんでした。 「ベイズモデリングの素養」は現代における複雑な統計モデリングの多くがMCMCや粒子フィルタなどを利用したベイズ統計の枠組みのもとで実践されることを考えれば、シニアレベルのデータサイエンティストならば身につけておいて損はないと思っています。

次に機械学習エンジニアですが、強調したいのは2点。「最先端の研究開発動向にも通じている」ことと「機械学習の実システム開発&運用のバッドプラクティスを熟知している」ことです。一点目は既に多くのメディア記事でも喧伝されていますが、今や多くのテクノロジー企業がDeep Learning含む機械学習技術の研究開発にしのぎを削っており、気付いたらライバル企業が最先端の機械学習アルゴリズムを実装してこれまでとは一味違ったプロダクトの使い心地を実現していた、なんてことも珍しくありません。また新たな機械学習技術の登場で、これまで実現が難しかった〇〇の機能が作りやすくなった、なんてこともあったりします。そういう「最先端」をキャッチアップし続けるためにも、各種トップカンファレンスやarXiv論文などを常時サーベイするというのは重要なことでしょう。勿論自ら研究を行って論文を出すというのもアリかと。

二点目ですが、これはレビューコメントで指摘されたポイントであると同時に、有名な「機械学習の技術的負債」論文に触発されたものです。ちなみに論文については有志の方が日本語でまとめた要約記事もあるようです。


Hidden Technical Debt in Machine Learning Systems

Electronic Proceedings of Neural Information Processing Systems

 papers.nips.cc **8 users**

機械学習システムにおける「技術的負債」とその回避策 - Qiita

#はじめに 空前のAIブームだった2017年、Yahooニュースでは毎日のように『〇〇が△△の出しが目立ちました。2018年は『AIの**運用**』の時代になるとも言われています。 **しがを...

 qiita.com **53 users**

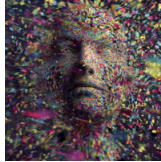
これを見れば分かりますが、通常のエンジニアリングとしてのシステム開発と同じような感じで、尚且つ原因がコード以外にも依っているような複雑な技術的負債が機械学習システムでは容易に発生し得るということです。「隠れたフィードバックループ」「データ依存性コスト」「グルーコード」「隠れた実験コードパス」辺りは僕でも思わず苦笑いしてしまうようなあるあるネタですが、そうであるならば実際の機械学習システム開発の現場ではもっとキツイ現実が待っているということでもあります。

また、B2Cのシステムに機械学習システムを適用する場合はバイアスというか「公平性」の問題が生じることもあります。これについては以下の記事がよくまとまっていると思います。

機械学習で危険なバイアスを避ける3つの方法 | TechCrunch Japan

歴史における現在の時点で人間のバイアスがもたらす危険に目を無視することは不可能だ。コンピューターがこの危険を増幅している。われわれは機械学習を通じて忍び込む人間のバイアスの危険...

tc.jp.techcrunch.com 17 users



個人的な意見ですが、「現実の社会が何かしらの差別を（遺憾ながら）行っているのであればそこから得られたデータに基づいて作られた機械学習システムもまた同じ差別を働く」のだと思っています。なればこそ、シニア機械学習エンジニアともなればバイアスや公平性の問題についても熟知しているべきだと考える次第です。

最後に二者の共通要件として、純粋に「シニア」としてのスキルがあることと「社会実装」の実務経験があることを求めています。前者はある程度スキル要件としても書いておきましたが、やはり組織の中でデータ分析の責任者として振る舞うに当たって「鼎の軽重」を問われる場面はどうしてもあります。そういう時に純粋にアナリストとしてもエンジニアとしても「シニア」として個々のシチュエーションに対応できることは極めて重要です。そして何度かこのブログでもコメントしていますが、やはり「社会実装」の実務経験があることが重要なと。



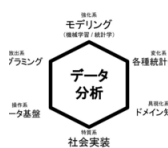
六本木で働くデータサイエンティストのブログ
id:TJO

Hatena Blog

HUNTER×HUNTERの念能力6系統で喩えるデータ分析スキル

HUNTER×HUNTER モノクロ版 36 (ジャンプコミックスDIGITAL)作者: 富樫義博出版社/メーカー: 集英社発売日: 2018/10/04メディア: Kindle版この商品を含むブログを見るみんな大好きHUNTER×HUNTE...

2018-10-09 19:00 ★56 232 users



ネタ記事で恐縮ですが、それでも「社会実装」の重みを実体験していることは大きいです。スキル要件の中にも書きましたが、例えばデータ分析手法を「適用した」経験と同じくらい「あえて適用せずに見送った」経験があれば、それにも大きな価値があります。

これだけ「ビッグデータ」「人工知能」「データサイエンティスト」がバズワード化している昨今では、必要性も分からないのに「人工知能ありき」というプロジェクトが立ち上がることも巷では珍しくありません。データ基盤も整備されていなければ、うっかりするとデータ自体もめっちゃくちゃ。それこそ「共有サーバーに大量のExcelファイルが放置されていてそれをビッグデータと言い張っている」なんてとんでもないケースもあったりします。そういう時に例えば「人工知能プロジェクトなんてやっている場合ではない、データ基盤の整備が先だ」と毅然としてかつ説得力を持って*7言い切れる専門家がいることが大切です。

目の前の課題を解決するために、要りもしない統計分析や機械学習システム実装を強行するのではなく、必要がなければ見送る。逆に当初計画ではそれらが要らないことになっていたとしても、必要が生じたら直ちに投入する算段をつけて実行に移す。そういった「やるorやらない」「やるならどうやるか」まで采配を振るえてこそ、シニアレベルのデータサイエンティストなり機械学習エンジニアなりなのだと思います。

プログラミングスキルに関する要件は？

上記のスキル要件を満たすための「インタフェース」は今後多様化していくことが予想されるため、レビューコメントも踏まえてRが良いとかPythonを使うべきとかC++は書けた方が良いとかプログラミングがこの程度出来るべきと言った「プログラミング言語要件」は全面的に外しました。その代わりアウトプットとして何が出来るかという点を重視しています。言い換えると「これくらいのアウトプットが出せるなら必然的に相応にプログラミングは出来るはず」*8というのを根底に置いています。

とは言え、このご時世にFORTRAN一本で頑張るとか、特にデータ分析関連のパッケージが強いわけでもないRubyやPerl一本で頑張るとか、そういうのは流石に辛いかなと。。

「多々ますます弁ず」では現実離れしやすい

この手のスキル要件談義になると、現在でも例えば「データサイエンティストや機械学習エンジニアを名乗るならKDDやらNeurIPSやらトップカンファレンスに毎年論文を載せられるくらいの人材でなければダメだ」「一流のデータサイエンティストなら研究論文も書いてビジネスも回せるべきだ」などという声が度々挙がるのを見聞します。上記のシニアレベル機械学習エンジニアのスキル要件で挙げたように、確かにそういう実力の持ち主であれば尚良いというのは事実だと思います。

しかしながら、そういう議論で忘れられがちなのが「どれくらいのスキルがあればどれくらいのレベルの仕事が務まるのか」という「最大公約数」の発想です。あれもこれもと要件を付け足していけば勿論理想的で優秀な人材像に近付いていきますが、やり過ぎればオーバースペックになってしまいます。例えば、ユーザー数が数百万人オーダーに上る人気スマホアプリに機械学習に基づく新機能を複数実装するための数十人規模の新規開発チームを作るとした時に、そのチームのリーダーに求められるのはどういう要件でしょうか？ 上記のシニアレベル機械学習エンジニアのスキル要件のうち1・4番目は重要でしょうが、残りはケースバイケースになるはずです。けれども人材採用に当たって3番目（研究動向の強力なキャッチアップ）を満たすことにこだわってしまい、それ以外はぴったりの人材を逃し続けてしまったら、いつまで経っても良いリーダーを迎えられず新チームは船出する前に沈没してしまうかもしれません。

また、これはデータ分析チームの組織作りの方法論・組織論にも関わる話なので深入りは避けませんが、1人のシニアレベル人材が「全部」のスキル要件を満たしていなくても良いはずですよ。何なら、2〜3人で互いにスキルをオーバーラップさせながら「合わせて全部」の要件を満たすようにチームを作っても機能すると期待されるからです。そういう柔軟で合理的な発想をせずにひたすら「多々ますます弁ず」という態度で人材を探し続けても、そんな人材はよっぽどの幸運にでも恵まれない限りは見つからないでしょう*9。あまり極端にスペックにばかり拘らず、その時点でのデータ分析チームに必要なバランスの良い人材を探すという考え方も必要と考えられます。

また、データ分析チームには高スペックな専門家が必要だというので採用を頑張ったら研究者もしくは研究マインドの強い尖った人材ばかりになり、みな研究には熱心だが実務のエンジニア仕事は誰もやらないのでビジネス上の成果がほとんど出ず、結局チームは解散させられてしまったという悲しい話を過去に何度か聞いたことがあります。研究部門ならばそういう採用の仕方でも良いと思いますが、開発部門でそれでは必ずしもうまくいかないかもしれません。チームの目的に合わせて、スキル要件の捉え方には柔軟になるべきかなと思っています。

スキル要件の捉え方に柔軟であるべきというのは、何もシニアレベルに限った話ではありません。ジュニアレベルでも、例えば「単位根過程も分からないような奴にデータサイエンティストの仕事は任せられない」「特異値分解の仕組みも分からない奴に機械学習エンジニアの仕事は任せられない」というように杓子定規に切り捨てても仕方がないので*10、まずは担当してもらう仕事のレベルに応じてhiring barを上げ下げすれば良いのではと思う次第です。

「技術的スキルなんて適当で良い」ではおしゃべり課題解決コンサルおじさんに堕する

上記の指摘とは真逆の話ですが、これは以前の記事でも書いています。



六本木で働くデータサイエンティストのブログ

id:TJO

データサイエンティストや機械学習エンジニアが、可能な限り紮やプログラミングを使って課題を解決すべき3つの理由

(Image by Pixabay)しばらく前のことですが、旧知のTakayanagi-sanがこんなブログを書いて
ネス上の課題を解決していくことは当然必須であるが、データ分析者としてのキャリアを
ータ分析に関係のない仕事はできるだけ避けたほうが良い。このような環境で職務経験を

2018-12-04 19:00 ★35 94 users

一般に、**何事であれヒトは易きに流れるもの**。大抵のビジネス実務の現場では、放っておくと「小難しい統計学や機械学習のような代物を使ってデータをこねくり回すよりも神Excelのようなまだ素人が見ても分かりやすいような方法で分かったつもりにさせてくれるようなデータ分析をしてくれた方がマシ」みたいな声が挙がりがちだと、個人的な経験や見聞からは感じています。

そうするとデータサイエンティストや機械学習エンジニアにも「技術的スキルよりもむしろビジネススキル」を求めるべきという話になりがちで、実際にランダムフォレストもまともに使いこなせないけどクロス集計と相関分析ぐらいは出来るというくらいのスキルセットで、あとはとにかく営業トークがうまくてそれっぽいビジネスプランの提案をするのは上手みみたいな自称データサイエンティストばかりズラリと並べているチームがある、という話も過去に何度か耳にしたことがあります。他にも、初歩的な汎化性能の概念も分からず単にネット上に転がっているscikit-learnのサンプルコードをコピペしてくる以外には何も出来ない、自称人工知能エンジニアが沢山いて「人工知能システム」の開発をバンバン請け負うチームとかいうホラーみたいな話も最近では出てきているようです。

勿論ビジネスとしての成功を考えるならばそれもアリだとは思いますが。ただ、それならデータサイエンティストとか機械学習エンジニアとか名乗る（名乗らせる）のはやめた方が良いのではないのでしょうか。上記の以前の記事にも書いたように、技術的スキルを重視しないのであればそれはただの「おしゃべり課題解決コンサルおじさん」でしかないのです。

こう書くと「別におしゃべり課題解決コンサルおじさんでも課題を解決してくれるならいいじゃないか」という声が飛んでくるかもしれません。ところが、例えばおしゃべり課題解決コンサルおじさんが、付け焼き刃の重回帰分析で後から後からどんどん新しい説明変数を付け足して更新していったモデルで「年々

精度の高いモデルになって役員会でのビジネスプランに役立ちます！」と会社の役員会の場で喧伝した挙句、実は学習データへの当てはまりだけが良くなっていて汎化性能の低い過学習したモデルになっていると気づかず、ある日突然全く予測が当たらなくなって困った、なんてことになったらどうしますか？ また、交差検証の概念を知らないコピペエンジニアが付け焼き刃のSVMで「（学習データを）予測させたら99%のaccuracyが出たのでこのモデルをレコメンデーションに使ってください」とか言い出して、企画部門のお偉いさんが鵜呑みにしてしまったらどうしますか？*11 統計学や機械学習を誤用したことによる副作用の多くは、いざビジネスの場に使われてしまった後では笑い話では済まないのです。

前のサブセクションで「多々ますます弁ず」は良くないと書きましたが、同じように「技術スキルなんて適当で良い」もやはりダメなのです。バランスが重要だということを最後に強調させてください。

最後に

レビューコメント踏まえながら色々細大漏らさないように書いてみたら、嘘みたいに長大なスキル要件と太記事になってしまい、これこそ「多々ますます弁ず」なのではないかと反省しておりますorz 最低でも次回からはデータサイエンティストと機械学習エンジニアとで記事を分けることにします。。。

*1: はじパタぐらいのレベルはあって欲しいところ

*2: アヒル本及び岩波DS当該巻に相当するレベル、時系列モデリングも含む

*3: 岩波DS3及び関連書籍の内容に該当するレベル

*4: 本番環境まで書くのかプロトタイプ程度が書ければ良いかは現場の開発体制次第ということで

*5: 岩波DS3のTVCMのアプリマーケティング効果測定事例では「TVCMを見れば見るほどアプリを使う時間が短くなる」という結論になる例が紹介されている

*6: 先に役員会レベルを説得すべきか、それとも現場レベルに納得してもらう方が優先か、などなど

*7: つまり言動に説得力を与えられるだけの経験と、ある種の「権威」が必要だということ

*8: ただし競技プログラミングでトップランカーになったりする必要までは流石にないかも

*9: 実際どこかのナントカ協会がかつて提唱したシニアレベル以上のスキル要件は「こんなスーパーマンなんて日本はおろかUSどころか世界全体を見渡しても見つからないのでは」みたいな代物だったのを思い出します

*10: 時系列分析をやらないのであれば前者はどうでも良いし、行列分解でレコメンデーションとかゴリゴリやるのもなければ後者は後から勉強してもらっても良い話

*11: どちらも完全なフィクションだったら良かったんですけどね

TJO 3日前



345

80

ツイート

G+



送る



関連記事



2018-12-04
データサイエンティストや機械学習エンジニアが、可能な限り統計学や機械学習やプログラミングを使って課題...
(Image by Pixabay)しばらく前のことですが、旧知のTakayanagi-



2018-02-07
データサイエンティスト&機械学習（人工知能）エンジニアのスキル要件と、過熱する人工知能ブームが生み出...
(Image by Pixaby)この記事は去年はてブ1100以上ついてしまった...



2017-06-25
データサイエンティストもしくは機械学習エンジニアになるためのスキル要件とは（2017年夏版）
この記事は2年前の以下の記事のアップデートです。前回はとりあ...



2016-09-21
データサイエンティスト（本物）は決して幻の職業などではない
かつて拙著出版の際に大変お世話になった技術評論社（技評）さ...



2015-05-29
海の向こうでも日本でも「データサイエンティスト」は雌伏の時
「データサイエンティストはつらいよ」、注目職種も求人が多く...

コメントを書く

生TensorFlow七転八倒記(10)：デ
キストデ... 》

 六本木で働くデータサイエンティストのブログ

Powered by Hatena Blog | ブログを報告する