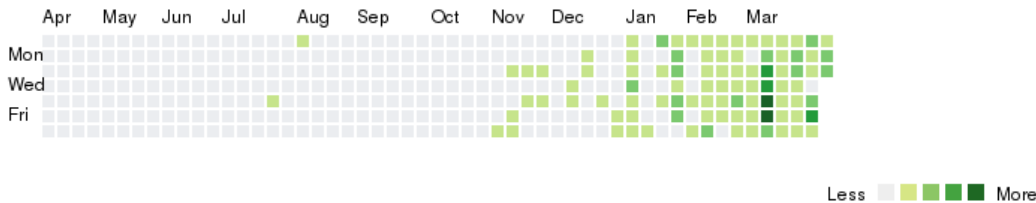




Drumato



2018-12-05

Kaggleに入門したい人におすすめするアプローチ法

Kaggle

Python

機械学習

30

1

ツイート

シェア

目次

- 目次
- 概要
  - 簡単な経歴
- Kaggleに対するアプローチ
  - 結果よりもプロセスを意識する
  - Python、R関係無く読む
  - コミュニティに入る
- オススメ環境
  - 高スペックマシンを扱えるカーネル
  - NoteBookの管理にGit
- 実際の流れ
  - 参加するコンペを選ぶ
  - Kernelを見る
  - 高スパンにカーネルを訪れて、最新の手法を真似し続ける
- Tips
  - よく使われるアルゴリズム
  - EDAの重要性
  - pd.get\_dummiesの実用性
    - その他機械学習についての記事
- コンペに使えるおすすめの本
  - 前処理大全
    - 追記
  - 完全独習 統計学入門
  - Scikit-learnとTensorFlowによる実践機械学習
  - 人工知能プログラミングのための数学がわかる本
- 総評

概要

著者の情報



Drumato (id:orangebladdy)

+ 読者になる 79

@drumatoさんをフォロー

Searching Articles

記事を検索

Q

Newest

2019-04-08  
1文字変数の確保・演算をgdbで

月別アーカイブ

▶ 2019 (48)

▼ 2018 (33)  
2018 / 12 (14)  
2018 / 11 (12)  
2018 / 10 (6)  
2018 / 9 (1)

カテゴリー

機械学習 (26)

Python (26)

ミニ記事 (23)

Go (21)

言語処理系 (11)

こちらは[Kaggle Advent Calendar 2018](#)の記事です。

英語・猛者揃い・応用力の塊と、高ハードル三拍子が揃ったKaggleで、  
私のような若輩者でも参加出来、  
また参加しづらい事を差し引いても余りあるメリットを紹介する記事です。

主に、

- **Kaggleおすすめのアプローチ**
- **オススメ環境**
- その他実際にどうするか(&Tips)

について話していきたいと思います。

## 簡単な経歴

記事を見るにあたって、**著者**の経歴は重要だと思うので書いておきます。  
私はこの記事を見ている**多くのKaggleに不安を持つ人達**と同じく、  
機械学習に対する**深い知見**や**ベテラン**ではありません。

今年の八月中旬に**0から作るディープラーニング**を読み、機械学習って面白いなと思いました。  
実際に本気で勉強し始めたのは10月の頭です。なので実歴は二ヶ月ぐらいですね。

SIGNATEの**マイナビコンペ**が初めてのコンペ参加でした。  
Kaggleの入門は只今開催中の**Eloコンペ**になります。

69	new	Drumato		3.715	17	3h
----	-----	---------	--	-------	----	----

今の順位はこんな感じ。  
後述しますが、**私自身の実力だけでここまで上がったわけではありません。**

## Kaggleに対するアプローチ

この記事を読んでいて**Kaggleを知らない人**はいないと思います。  
でも**参加することに抵抗がある**という人も多いのでは？

「勉強になるのはわかるけど何すればいいのかわからないよ...」  
「カーネル読めとか言うけど**全部英語**じゃん...」

という方の為に、いくつかオススメの利用方法を紹介しようと思います。

## 結果よりもプロセスを意識する

Kaggleといえばコンペ(=コンテスト)なので「結果が全て」というイメージがありますが、  
実際には**データサイエンスの総合プラットフォーム**という位置づけです。

コンペティションというのはKaggleの一項目に過ぎず、  
実際には**データサイエンティストたちのコミュニティ**としての意味合いもあります。

なので、日夜**活発な議論・意見**が飛び交っていますし、  
毎日**凄い人が沢山の知見**を発表しています。

これだけでもKaggleの利点がわかりますが、  
コンペティションそれぞれにも**Kernel**と**Discussion**があり、

- **凄い人が作ったチュートリアル**
- **コンペ参加者で盛り上げる議論**

等々、初心者が見るべきコンテンツがずらずらと並んでいます。

Ruby (9)
Linux (8)
Kaggle (7)
C言語 (6)
Gitlab (5)
Git (5)
書籍紹介 (5)
競プロ (3)
最低限 (3)
SIGNATE (3)
データ分析 (3)
PyTorch (3)
Vim・Neovim (2)
統計学 (2)
Xonsh (1)
V (1)
NW (1)
R (1)

### Hot

- Kaggleに入門したい人におすすめするアプローチ法
- Vimの補完プラグインであるcoc.nvimが強すぎる。
- SVMについて簡単に理解する。
- 2Dデータの可視化に最適なSeabornのチュートリアルを作成した
- 巷で噂のV言語のサンプルを読み解く。

特に**カーネル読め**と言われる理由はそれら知見を得られるためです。

**完全無料**で**実際のデータ分析の様子**が**いくらでも読める**というのは魅力ですね。

自分もそのコンペに参加することで、

**カーネルの意図**が深く読み取れる上にその知見を**コンペに流用**出来ます。

まずは臆せずチャレンジしてみるというかなあと思います。

なんてたって**参加も無料**ですから!(ルールはよく読みましょう)

---

また、コンペに取り組んで行くと**わからないことだらけ**になります。

- **この前処理**どうするん...
- **訓練アルゴリズム**どうしよう

等々数え上げたらキリがありません。

そこで上級者のコードを読むことで、

次回以降のコンペで**自力でそのコードが実装**出来るかもしれません。

特にデータの**前処理**は**沢山経験したもん勝ち**みたいなのがあるので、  
中途半端でもいろいろやってみることが大事だと思いますね。

オススメとしては、

- **カーネルをFork**してSubmit(やり方は後述)
- 実行した**カーネルをよく読んでよく考える**
- 自分で考えて**ちょっと変えてみる**

というプロセスを繰り返すと良いと思います。

私が本コンペ**いいスコアを出している**のも上記の方法を実践しているに過ぎません。

上級者のように**全部自分で考える**ほどの知見はなくとも、

**カーネルを読む**→**意味を考える**→**自分で試行錯誤**

することで、いいスコアを出すための方法を模索することが出来ますよね。

重要なことは、

「**凄いカーネル**を持ってきて**Submitのいい結果を見て満足**」して終わるのはXということです。

何のためにKagglerが**カーネルを"公開"**して、

多くの人が**Kaggleのカーネル読もう**と言っているのかを考えましょう。

## Python、R関係無く読む

カーネルに上げられるコードは、ほぼ全てが**Python**か**R**で書かれています。

大多数が**Python**のみ書ける状況だと思いますが、是非**R**も読みましょう。

Rの実行環境は**Kernel**で利用出来る(後述)ので、

一見よくわからなくても**区切って実行してみる**のが良いと思います。

セルの実行結果をいちいち確認すれば、わからなくてもやってることを理解出来ます。

## コミュニティに入る

Kaggleは一人でも出来ますが、

出来ればリアルタイムで同じコンペに取り組んでいる人と繋がれるとなお良いです。

コンペにおける情報共有は**チームメンバーのみ**承諾されている点には注意ですが、

実践している人の様子は何にも代えがたい参考資料になります。

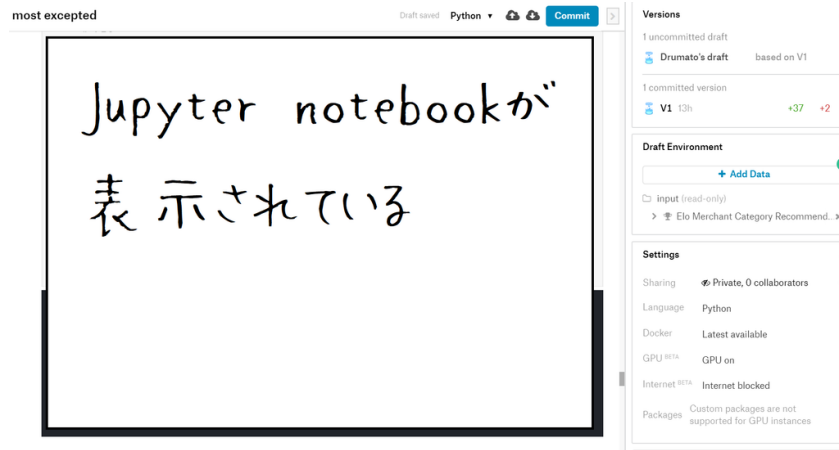
## オススメ環境

Kaggleに必要な**マシンリソース**(マシンパワー)ですが、  
公式の**Kernel**を用いる事で高パフォーマンスによるモデル構築が可能です。

以下使い方を示します。

## 高スペックマシンを扱えるカーネル

Web上で使えるVMのようなものだと思っていただければいいです。



こちらが実際の**Kernel**の画面です。

必ず使う機能を示します。

- **commit** ...現在のコードで実行してみます。実行に成功した場合はバージョンとしてどんどん追加されていきます。
  - つまりGitのような感じです。
  - 因みにコード内で**ファイルに出力**されれば、**Output**として保存されます。(これをそのままSubmitも出来ます。)
- **Download** ...**カーネル**のコードをダウンロード出来ます。(雲のマークに下矢印ついているやつ)
  - **最強カーネル**をローカルに落としてきて、**手元でお勉強**する時にも使えます。
- **Restart**...基本的に実行結果が保持されるノートブックで、**実行状態を初期化**します。
  - Jupyterで言えば**Restart Kernel**とかですね。

機械学習始めたての人は、**高スペックマシン**を持ってないと思います。

かくいう私も**現在はノートPC**しか持ってないので、どちらかというと劣悪環境寄りではあります。


**GCE**を使うという手段は**とても有効**ですが(特にスペック面)、  
お金がかかるという点で気後れしている人も多いはず。


**Kernelは無料で使える**ので、初心者の方の取っ掛かりにはうってつけですし、  
普段から使い慣れている**Jupyter Notebook**と使い方も酷似しています。

Kaggleのコンペに参加する為に**最適化**されているので、  
データの読み出しやSubmitも**とても簡単**なのでおすすめです。

## NoteBookの管理にGit

ご存知Gitですが、データ分析にももちろん効果を発揮します。  
使い方がわからない人はこちらをどうぞ。

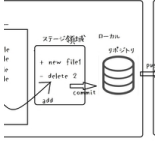



**Explore "Full-Stack" in depth!**  
id:orangebladdy

## 非ソフトウェアエンジニアも使えるべき最低限Git

概要 当記事の目的 対象読者 First Impact gitlab設定 Git初期設定 何故GitHubじゃないの? 更にセットアップを続ける いざ実践 リモートリポジトリの取得(git clone) ステージ領域への登録(git add) 口...

2018-11-17 11:28 **2 users**



[drumato.hatenablog.com](https://drumato.hatenablog.com)

先程の**ダウンロード機能**を使って、いいスコアを出したコードをリモートに保存してもいいですし、

**偉い人のカーネル**を保存してもいいですね。

しかしちゃんと**プライベートリポジトリ**にする様に!

データ分析コンペは基本的に**情報は非公開**です。

無料でプライベートを作れる**GitLab**がおすすめです。

## 実際の流れ

最後に、実際のコンペに参加した時の流れを書いてみようかなと思います。

本当に形式上の話にはなりますが、一応雰囲気はつかめると思います。

## 参加するコンペを選ぶ

ここが正直**一番大事**かもしれません。

画像認識系であれば**マシンリソース**はかなりのものを要求されますし、

**自然言語処理**や**時系列分析**等、コンペによってその特徴は様々です。

おすすめは**正統派なテーブルコンペ**です。

参加するコンペを決めたら、

**Data** の欄から、**データセット**のサイズを確認します。

小さいようなら**ローカルに落としても**良いと思います。

ローカルなら**ネットに繋がっていない時**でも作業できたりして効率がいいです。

Kaggleのコンペは比較的**データが大きい**ので、

ここではローカルに落とさない事を前提に話します。

## Kernel を見る

次に、**Most Votes** 等でソートして、高評価の**カーネル**を読みます。

Forkしてきて一度Submitしてみてもいいと思います。**現在のボーダーを把握出来ます。**

基本的には、以下のような手順で書かれていると思います。

- ライブラリのインポート(これの前に**カーネルの説明**が有る場合も)
- データの読み出し、Shape確認( `DataFrame.shape` 等)
  - `DataFrame.head()` とかが使われてることもあります。
- 前処理**
  - 欠損値**の確認、置換
  - データ型**を扱いやすい形に変更
  - カテゴリ変数**のダミー変数化等々
  - 取りうる値が少ない数値データを**離散データ**として扱ったり
- 特徴量エンジニアリング**
  - 文字列の日付**を日付型にして `day` , `month` 等取り出してみたり
  - データが**そのままでは扱いづらいもの**に対して**代表値**を取ってみたり

- その他**沢山の手法**が考えられる。
- データ準備
  - 複数のデータを**主キーで突き合わせたり**
  - 学習に用いないカラムを**取っ払ったり**
- **交差検証・KFold分割法**
  - 使われているモデルから**本コンペに適したモデル**を探るのも重要
  - 一般的に時間がかかるので**注意**
- **Submit**
  - `feature_importance_` を見たりすることもまあある
  - 提出ファイルの**フォーマット**は要確認。よくエラーが起きる部分

めちゃくちゃ簡単に書くとこんな感じですかね?

NoteBookには**Markdown記法**が埋め込めるので、  
Google翻訳を駆使して**英語を読む**と理解度が深まると思います。

出来れば**翻訳機を使わず**に読めるとなおいと思います。

このそれぞれのフェーズに対して、

- **何故この処理をしたのか**
- **この処理のありなしで結果はどう変わるのか**
- **学習の様子**

等、ありとあらゆる部分の知見を**盗もう**としましょう。

また**新規のKernel**を1つ起動して、  
手で実際に一つ一つ動かしてみるのも**凄いいい**と思います。

## 高スパンにカーネルを訪れて、最新の手法を真似し続ける

コンペの開催期間中には参加者達が常にデータを分析し続けるので、  
後々になって**ゴールデンフィーチャー**が見つかるなんてこともあります。

おすすめはしょっちゅう訪れて動向を確認しておくことですね。  
これは**好成績を残す**という意味でも、**勉強**という意味でもとても効果があります。

## Tips

ここではKaggle参加における豆知識集を紹介します。

## よく使われるアルゴリズム

- **XGBoost**...多分まだ一番人気。`feature_importance` が使えるのも強い。
  - **アンサンブルメソッド**を使っているのでスコアが良くなりやすい。
  - また**データの前処理**も厳密でなくていい。比較的受け入れてくれる。
  - **過学習**を起こしやすい!
- **LightGBM**...もう一つの勾配ブースティング木。
  - **高い汎化性能を持っている**...印象がある。
- **MLP (Multi Layer Perceptron)**

## EDAの重要性

カーネルで沢山のVoteをもらう**EDA系**ですね。

EDA(*Exploratory data analysis*)とは、データ分析において重要な**可視化**を含むフェーズの事です。

データの様子を**グラフにプロット**したり、  
**集計して関係**を見たり。

ここでの知見が**特徴量作成・選択**に大きく影響するので、  
めちゃくちゃ大事な部分と言えます。

**対象データセットの理解**にも一役買うので、  
コンペ参加した後最初に**EDAのカーネル**を読むのもおすすめです。

## pd.get\_dummies の実用性

必ずと言っていいほど行われる、  
**カテゴリ変数のダミー変数化**ですが、

**氏名**や**ユーザID**等、沢山の種類を取る変数がある場合には  
ダミー変数化によって大量のカラムが生まれ、  
**特徴量の爆発**によってモデル性能が下がる危険があります。

また**単純に計算時間が膨れ上がり**ます。

そのため、**特定のカラム**に対して処理を行うのがいいです。

```
train = pd.get_dummies(train, columns=['foo_cat', 'bar_cat'])
```

また二値のみで分かれる場合には、


```
DataFrame['column'].map({'YES':1, 'NO':0})
```

もいいと思います。


## その他機械学習についての記事

ネットにある**凄い人たちの記事**は既に読んだと思うので、  
私の記事を少しだけ...

あくまで参考程度にどうぞ。**マサカリ待ってます**。

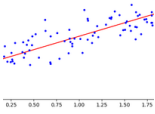


**Explore "Full-Stack" in depth!**  
id:orangebladdy

 Hatena Blog


**線形回帰モデルの訓練手法について。**

目次 目次 概要 対象読者 線形回帰モデルの定義とコスト関数MSE 使用するデータ 正規方程式 正規方程式の使い所と長短 正規方程式メソッド 正規方程式デメリット 勾配降下法(Gradient Descent...GD) ...




2018-11-02 13:30

[drumato.hatenablog.com](https://drumato.hatenablog.com)



**Explore "Full-Stack" in depth!**  
id:orangebladdy

 Hatena Blog

**正則化された線形回帰三種をまとめる。**

概要 使用するデータ 似た言葉(正則化、標準化) 正則化 標準化 本題: 正則化 Ridge回帰 Ridge回帰の閉形式解 Lasso回帰 Elastic Net 終わりに 余談 概要 サラマンダー本の4章は数式もガンガン出てきて、...

$$MSE(\theta) +$$

2018-11-12 23:21

[drumato.hatenablog.com](https://drumato.hatenablog.com)



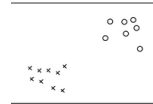
Explore "Full-Stack" in depth!  
id:orangebladdy



Hatena Blog

## SVMについて簡単に理解する。

概要 SVMの基本概念 線形識別モデルとSVMの比較 SVMの特徴 スケールの重要性 実装、SVM 非線形SVM分類器 多項式カーネル カーネル関数とは ガウスRBFカーネル SVM回帰 終わりに 概要 サラマン...



2018-11-17 23:46

[drumato.hatenablog.com](https://drumato.hatenablog.com)



Explore "Full-Stack" in depth!  
id:orangebladdy



Hatena Blog

## データ分析コンペのガイダンスを作成した。

概要 概要 記事のタイトルの割にはあまり内容の無い記事になります。機械学習に興味を持っている人が結構多くなってきた印象があります。かく言う私も機械学習をやろうとしてから一ヶ月程度が経ち、いろんな事に挑戦し始めました。そのうちの1つにデータ...

2018-11-18 21:58

[drumato.hatenablog.com](https://drumato.hatenablog.com)



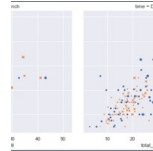
Explore "Full-Stack" in depth!  
id:orangebladdy



Hatena Blog

## 2Dデータの可視化に最適なSeabornのチュートリアルを作成した

目次 目次 概要 環境設定 Jupyter Notebookの場合 使用するデータセット 基本 統計的推定 カテゴリ変数の可視化 応用編 relplot()による統計データの関係 箱ひげ図 分布(ヒストグラム) 終わりに 余談 概...



2018-11-22 15:27 ★

[drumato.hatenablog.com](https://drumato.hatenablog.com)



Explore "Full-Stack" in depth!  
id:orangebladdy



Hatena Blog

## 記述統計の基礎を復習した(for統計学未学習者を含む)

目次 目次 概要 一章まとめ 記述統計と統計的推定 量的変数と質的変数 度数分布表とヒストグラム 2章まとめ 4つの尺度 代表値 平均値 中央値 最頻値 3章まとめ 散らばりの尺度 四分位偏差 分散 回帰...



2018-11-25 01:20 ★14

[drumato.hatenablog.com](https://drumato.hatenablog.com)

## コンペに使えるおすすめの本

### 前処理大全



#### 前処理大全[データ分析のためのSQL/R/Python実践テクニック]

作者: 本橋智光

出版社/メーカー: 技術評論社

発売日: 2018/04/13

メディア: 大型本

[この商品を含むブログ \(1件\) を見る](#)

正直一番有用なのは?と思う本です。

カーネルを読んで何だこのコードと思っても、

大抵この本に書いてあります。



既に色々調べている人なら分かると思いますが、  
コンペで特に重要なのは**データ加工・前処理**です。

この本はただコードを紹介するだけではなく  
**その処理を行う意味づけや、実行速度等、**  
とにかく**実践**に重きが置かれていて読んだことがそのまま利用できます。


私が現在参加しているコンペでも間違いなく**一番お世話になっています**。

一度書店に行ってみてペラペラめくってみるといいと思います。  
**名前のニッチさ**とは裏腹に、凄く読みやすいことがわかります。

## 追記

**前処理大全**の著者である@tomomoto\_LV3さんからオススをいただきました...!  
大変恐縮です、ありがとうございます!

**前処理大全**が特徴量エンジニアリング本ではないという旨のツイートをされています。


**tomomoto**  
@tomomoto\_LV3

記事で著書をご紹介いただき、大変ありがとうございます。記事は、Kaggle入門者全員にオススメです！


ちなみに本は特徴量エンジニアリング本ではなく前処理方法テクニック本なので、全員にオススメではないかも。店頭でご確認を！思った本と違うと割とネットで叩かれてます（笑）  
@siero5335さん情報thx!

**Drumato** 🍑 @drumato  
 Kaggle入門したい人に向けた勉強目的でのアプローチ法とその詳細の記事を書きました!#Kaggle#KaggleAdVentCalendar2018#はてなブログ  
 Kaggleに入門したい人におすすめするアプローチ法...drumato.hatenablog.com/entry/2018/12/...

♡ 22 10:09 - 2018年12月6日 ⓘ

👤 tomomotoさんの他のツイートを見る ➤

## 完全独習 統計学入門



**完全独習 統計学入門**  
 作者: 小島寛之  
 出版社/メーカー: ダイヤモンド社  
 発売日: 2006/09/28  
 メディア: 単行本 (ソフトカバー)  
 購入: 215人 クリック: 3,105回  
[この商品を含むブログ \(115件\) を見る](#)

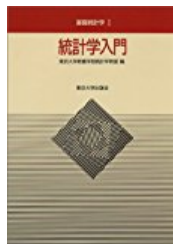
私自身は読んでいませんが、  
友人に見せてもらいましたし書店でもチラチラめくってみました。

**これ以上簡単な本は無い**というぐらい優しい言葉と図で説明されていて、  
絶対に理解させるという強い意思を感じます。

私は

### 統計学入門 (基礎統計学?)

作者: 東京大学教養学部統計学教室  
出版社/メーカー: 東京大学出版会



発売日: 1991/07/09  
メディア: 単行本  
購入: 158人 クリック: 3,604回  
[この商品を含むブログ \(79件\) を見る](#)

で統計学の入門をしています、かなり難しく手こずっています。

**段階を踏んで勉強していきたいなら、**  
この本は必ず助けとなってくれることでしょう。

統計の知識と機械学習の関連がわからない人は、  
騙されたと思って色々勉強してみたら**偉い人達の記事**を読みましょう。

きっと多くの記事が**統計の知識**を応用した手法を取り入れているはずです。  
例えば**データの代表値を特徴量として加えたり**。

またEDAにも統計は**大きく関連します**。

データの

## Scikit-learnとTensorFlowによる実践機械学習



**scikit-learnとTensorFlowによる実践機械学習**  
作者: Aurélien Géron, 下田倫大, 長尾高弘  
出版社/メーカー: [オライリージャパン](#)  
発売日: 2018/04/26  
メディア: 単行本 (ソフトカバー)  
[この商品を含むブログ \(1件\) を見る](#)

値段が高い分**網羅性も非常に高い**です。

どちらかというと**時間がたっぷり取れる時**に取り組むべきかな?

難しい部分も沢山ありますが、  
**機械学習の全体感**を持ちながら勉強出来、  
また身につく技術にも**確かなもの**があります。

先程の記事たちは**この本で得た知見**を主として書かれています。

## 人工知能プログラミングのための数学がわかる本

主に**数学サボってきちゃった人**の為の本です。

機械学習と数学が切っても切れない事はわかっている、  
**苦手意識**から中々踏み出せない方には最適だと思います。

義務教育範囲の数学を**簡単**に復習しながら、  
**それら数理が機械学習にどう関わるか**を知ることが出来ます。

数学やっていると、**これが機械学習の何のためになるんだ...?**となってしまうことがあります。

この本は常に**機械学習の為の数学**を忘れさせないようなアプローチを貫いていて、  
読んでいて迷子になることがありません。

## 総評

いかがでしたでしょうか。

Kaggleについてまだ知識の浅い私の記事でしたが、

一人でも多くの人に役立てば幸いです。

一緒にKaggleがんばりましょう！

Drumato (id:orangebladdy) 126日前

★+

30

1

ツイート

シェア

■週4可！■運用に特化したソーシャルゲーム会社で女性向けゲームのシナリオライター...  
大人気女性向けゲームのシナリオライターとして、...  
月収 22万円～...  
検索する

＜在宅＞オリジナル小説（ノベル）、漫画原案をご執筆いただけるクリエイターさん...  
自社開発のストーリーアプリ『Portie（ポルティ）...  
検索する

【広島県在住の方限定！】グルメ写真の撮影案件\カメラマン大募集！／  
【概要】大手グルメレストラン情報サイトに掲載す...  
検索する

【社員案件】スカパーの新規事業に携わる！Web開発ディレクターの募集！！  
日本で唯一、アジア最大の有料多チャンネル放送／...  
検索する

関連記事

- 2019-02-07

「Python機械学習プログラミング」はまさに"理論と実践"を学ぶのに最適な本

Python機械学習プログラミングの書評を熱く語りました。
- 2019-01-28

入門者に捧ぐ機械学習の勉強法&決定木の内部解説

目次 目次 概要 機械学習における理論と実践 本題:決定木 参考:...
- 2018-12-31

2018年を振り返る+2019年の目標を立てる

目次 目次 概要 今年のプログラミング総まとめ Ruby BookManage...
- 2018-12-18

1日1カーネルのススめ(サンプルあり)

目次 目次 概要 カーネルを読む意義 事例:Porto Seguro 概要 ラ...
- 2018-11-18

データ分析コンペのガイダンスを作成した。

概要 概要 記事のタイトルの割にはあまり内容の無い記事になり...

« (第一回)PyTorchの公式ドキュメント翻訳ハ...

ABC 114に参加しました。»

はてなブログをはじめよう！

orangebladdyさんは、はてなブログを使っています。あなたもはてなブログをはじめませんか？

はてなブログをはじめ（無料）

はてなブログとは

Explore "Full-Stack" in depth!

Powered by Hatena Blog | ブログを報告する