（Fanbox用）Pixivでウケる小説の条件

0. この記事の要約

- 疑問：どういう作品がウケるのか。また、読者はどういう作品を求めているのか。
- 手法：閲覧数と公開期間に基づく「作品の人気度」を定義した。また、Pixivの自作品を線形回帰モデルで分析し、「作品の人気度」に寄与する性質が何であるのかを調べた。
- 結果：文字数が多い作品・女性上位モノの作品のウケが良い。また、フォロワーの多いイラストレーターの方に挿絵を描いてもらうのが効果的である。

1. はじめに

- これまで6年近くPixiv上で細々と創作を続けてきて、幸いにも多くの方々の手に取ってもらえるような作品を書いてこれた。
- その時のテンションで筆の赴くままに作品を書いてきたが、個人的にうまく書けたと思ってもブックマーク数（ブクマ数）に伸び悩んだり、逆に手ぐせで書き殴ったような作品が思いのほかよく読まれるといった現象に遭遇してきた。

- そして同時に、たとえば「ドMホイホイ」のタグをつけると閲覧数（PV数）が跳ね上がるような現象を見るたび、ウケる作品に一定の傾向があるのではないかと思うようになった。
- 読者に求められている作品とは何であるのか・どういった要素がウケる作品に欠かせないのかを探るために、これまで書いてきた作品を分析することにした。

2. 人気な作品の要素とは何かを列挙してみる

- Pixivから作品に関するデータを集めてくる前に、何が人気度に貢献しているのかを予想し、とりあえず思いつく限り挙げてみる。
- 大きく分けて2種類あり、細かく分けると19項目ある。これらを説明変数の候補とし、その中から人気の度合いを説明するのに必要なものを後で選ぶ。
- 「数量」は適当な数字（正の実数値）を取るもので、「バイナリ」は0か1の値しか取らないものを指す。例えば、HEIZENはバイナリだが、もしこれが0だったら「その作品にはHEIZENの要素が含まれていない」、1だったら「その作品にはHEIZENの要素が含まれている」、ということを表現している。

作品の内容に関わらないデータ

- ln_char_num（数量）→作品の文字数について自然対数をとったもの。
- ln_illustrator_follow（数量）→有償リクエストで依頼し、ツイッター上で自分が画像ツイートで宣伝した、もしくはイラストレーター本人に宣伝orRTしてもらった場合、そのイラストレーターのフォロー数の自然対数をとったもの。
- ln_illustrator_follower（数量）→有償リクエストで依頼し、ツイッター上で自分が画像ツイートで宣伝した、もしくはイラストレーター本人に宣伝orRTしてもらった場合、そのイラストレーターのフォロワー数の自然対数をとったもの。
- ln_follower_to_follow（数量）→有償リクエストで依頼し、ツイッター上で自分が画像ツイートで宣伝した、もしくはイラストレーター本人に宣伝orRTしてもらった場合、そのイラストレーターのフォロワー数をフォロー数で割った値の自然対数をとったもの。

作品の内容に関わるデータ

- sexaroid（バイナリ）→アンドロイドに、性行為を目的とした機能が搭載されているか否か。
- HEIZEN（バイナリ）→一見知性があるように見えたり、外見が人間そっくりだったりするものの、実際は外部からの刺激に従って適切な行動をとるだけの機械にすぎないか否か。
    - 竿役から辱めにあったり、公衆の面前で異常な挙動や性的な行為をしても、全く動じず人間らしい対応を取ることができていない場合に限定する。つまり、機械的な挙動で搾精する、機械的な言動をとっていかにもロボットらしさを見せるといった作品はHEIZENとはみなさず、人間らしい反応が期待されるような外見・振る舞いをしながらも、全く人間らしくない不自然な反応を見せるような作品にする。
- femdom（バイナリ）→アンドロイドに竿役が犯されるような描写があるか否か。
    - 「逆レイプ」や「ドMホイホイ」といったタグに対応している。
- age（数量）→アンドロイドの設定年齢。
    - 設定のない二次創作や、オリジナル作品であっても細かく決めていない場合は想像で適当に補完した。
- lolita（バイナリ）→アンドロイドの設定年齢が15歳以下か否か。
- vaginal（バイナリ）→アンドロイドに女性器が搭載されていて、作中でその描写があるか否か。
- handjob（バイナリ）→作中に手コキの描写があるか否か。
- blowjob（バイナリ）→作中にフェラの描写があるか否か。
- titfuck（バイナリ）→作中にパイズリの描写があるか否か。
- intercrural（バイナリ）→作中に素股の描写があるか否か。

- sex（バイナリ）→作中に挿入描写があるか否か。
- template_mecha（バイナリ）→ロボットらしい体言止めを多用したセリフがあったり、アンドロイドが壊れている・メカバレしているような、ロボ娘作品にありがちな描写があるか否か。
- TPS（バイナリ）→作品が客観的な視点で書かれているか否か。
- original（バイナリ）→作品が二次創作ではなく、設定を自分で考えたか否か。
- descriptive_title（バイナリ）→タイトルを読めば、作中で何が書かれているか大体想像できるか否か。
    - 例：「人身御供」はこれに該当しない
    - 例：「女子陸上選手型アンドロイドを騙して素股させたりフェラしてもらう話」は該当する

3. データの収集

- Pixivから、アンドロイドに関連する作品のデータを手作業で集めた。宣伝やサンプル、台本、オムニバス作品、翻訳作品は除いた。合計47件になった。
- 今後モデルの学習を行うため、訓練用のデータ42件、テスト用のデータ5件に分けた。テスト用のデータはランダムに選んだ。モデルの訓練は訓練用のデータで行い、性能評価はテスト用のデータで行う。

▼ データの形式について

- データはCSV形式で格納した。以下のような感じで、値がカンマ区切りで区切られている形式なのでComma Separated Values（CSV）という名前が付いている。CSV形式だとpandasというライブラリで扱いやすい（データを読む時はpandas.read_csv、書き出すときはpandas.to_csvで簡単に書ける）。

```
title,bookmark,PV,showing_period,char_num,illustrator_follow,illustrator_follower,follo
愛玩人形リサ−1−,69,3112,2209,1684,0,0,0.0,1,0,0,12,1,1,0,1,0,0,0,0,0,1,0
愛玩人形リサ−2−,24,1558,2209,1880,0,0,0.0,1,0,0,12,1,1,0,0,0,0,0,0,0,1,0
愛玩人形リサ−3−,33,2047,2207,5540,0,0,0.0,1,0,0,12,1,1,0,0,0,1,1,0,0,1,0
あるバー(？)にて,94,3990,2178,5917,0,0,0.0,1,1,0,18,0,1,0,0,0,0,0,1,1,0,1,0
カジノにて,207,5578,1890,5396,654,863,1.3,1,0,0,14,0,1,0,0,0,0,1,0,0,1,0
```

- テストデータは以下の5つ

```
title,bookmark,PV,showing_period,ln_char_num,ln_illustrator_follow,ln_illustrator_follower,ln_foll
人身御供,383,7712,1084,9.242613932526478,1.3862943611198906,7.310550158534422,6.213207117506563,1,0
一緒に寝ていたオナニーサポート用アンドロイドのロリ美少女に、金玉カラカラになるまで絞られました。,421,8875,899,9.50
オナニーするだけのバイトと聞いていたはずなのに、気づいたら人型自慰行為補助機械のお姉さんの奴隷になっていました。,1248
戦場から遠く離れたところでは,111,2551,2137,9.487820581943367,0.0,0.0,0.0,1,0,1,20,0,1,1,1,0,1,1,1,0,1,0
家事手伝いアンドロイドで精通した11歳の男の子のオナニー風景,212,6482,1455,7.798112628829788,0.0,0.0,0.0,0,1,0
```

4. 人気作品の指標を計算する

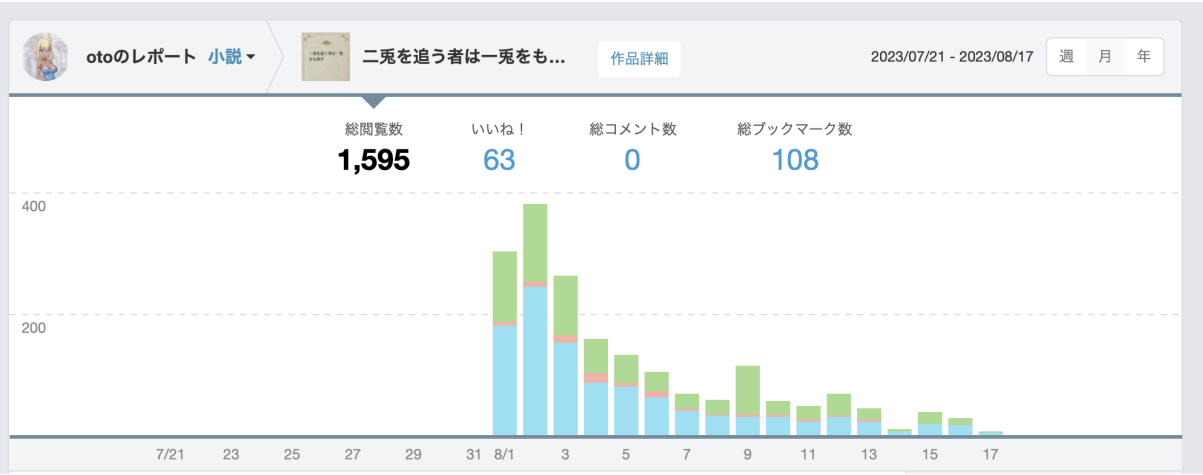
人気の指標としてブクマ数は使えるか？

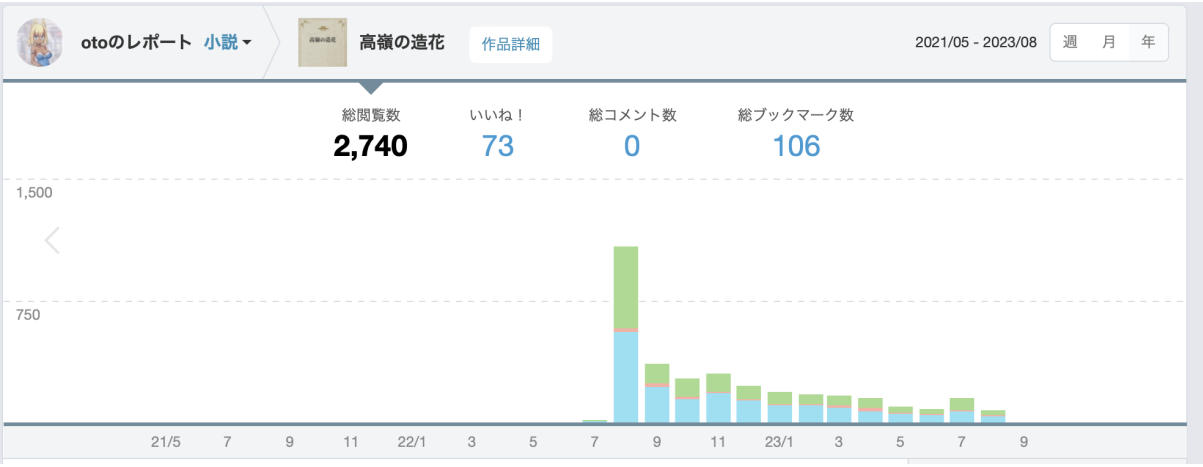
- さて、データも集めたし、ブクマ数を目的変数として回帰分析でもしようかとなるが、ちょっと立ち止まって考えてみたい。
- ブクマ数や閲覧数は確かに人気を表しているが、公開日数に依存する部分がある。このため、同じくらい人気がある作品でも、公開日数が短い方がブクマ数が少なくなり、ブクマ数で単純比較すると人気の有無が正確に見て取れないことになる。作品に潜在的に備わる人気度を測る上で、公開日数に依存しないような人気度の指標が欲しい。
- つまり、人気度を表す変数に「経過日数」が絡まないようにしたい。経過日数を固定した上で、同じ経過日数の作品ならどうなったらより人気があると言えるかが分かれば、経過日数を無視した指標ができる。

- 閲覧数が多いことが人気作品の条件とする。そして今、「公開期間がこれくらいだと、大体閲覧数がこれくらいになりそうだな」ということが予想できるとする。そうすれば、たとえば公開日数が10日で予想閲覧数が1000回だった時、実際の閲覧数が1500回だったら「お、この作品はウケてるな」と考えられたり、閲覧数が500回だったら「この作品はウケがよくなさそうだな」と考えることができるはず。
- つまり、公開期間（日数）から予測される、平均的な閲覧数がどうなるかが分かれば良い。この方針にシフトして話を進めていく。

人気を示す指標を、公開期間と閲覧数から計算する

- 平均的な作品の一日あたりの閲覧数yがべき乗則に従うと仮定する。つまり、ある定数kとaを使い、公開期間（日数）xを使ってy=kx^a と表せるとする。例えばk=1, a=-2だったとすると、リリース直後の1日目は360人に読まれた時に、リリースしてから2日目は1 * 360 * (1/4) = 90人、3日目は1 * 360 * (1/9) = 40人、といったように、一日あたりの閲覧数が減少していく。
    - ここで、xが増えるとyは減るので、aは0未満であるとする。またx=1の時y=kなので、kは「平均的な『1日目の』閲覧数」と捉えることができる。また、kとaは独立であると仮定する。つまり、kとaが同じファクターに依存していて、そのファクターに伴ってkとaが一緒に変化する、みたいな関係性はないとする。
    - 実際、閲覧数が急激に減少しているのが以下の図から見て取れる。このような性質をべき乗則（ロングテール性）というが、これは色々なところで見られる。





- 一日あたりの閲覧数のデータは得られないが、総閲覧数のデータは得られるので、公開日数と総閲覧数の関係性を捉えることを目標とする。一日あたりの閲覧数y=kx^aを積分した値が、総閲覧数になる。一日あたりの閲覧数をD日目まで積分すると、以下のようになる。

$$y = \int_1^D kx^a dx = \frac{k}{a+1}(D^{a+1} - 1)$$

- つまり、公開期間x日における総閲覧数yの予測値は、定数kとaを使って、以下のように表せる。

$$y = \frac{k}{a+1}(x^{a+1} - 1)$$

- さて、手元には作品ごとの公開期間と閲覧数の訓練データ{(x_1, y_1), …, (x_m, y_m)}がある（今回はm=42である）。これを使ってkとaを求めたい。最小化したい目的関数は平均二乗誤差を使う。つまり、以下のように、真の値から予測値を引いた値を誤差とし、誤差を2乗した値の平均をとったものである。つまり、kとaがいい感じなら、i番目のデータの閲覧数y_iと上の式で計算した閲覧数の予測値yが近くなり、目的関数が小さくなる、という感じ。

$$f(k,a) = \frac{1}{2m}\sum_{i=1}^{m}\left(y_i - \frac{k}{a+1}(x_i^{a+1} - 1)\right)^2$$

- パラメータkとaを更新するには、最急降下法を使う。つまり、上記の目的関数をkとaについて偏微分し、その結果に学習率をかけて、kとaのそれぞれから引いて値を更新する。

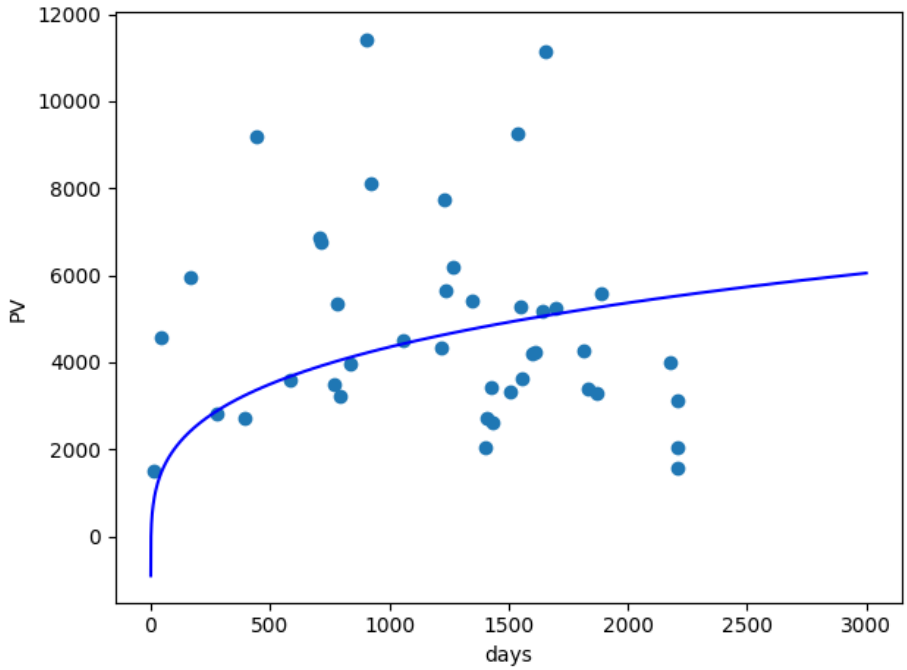
$$k \leftarrow k - \alpha\frac{\partial f(k,a)}{\partial k} = k - \frac{\alpha}{m}\sum_{i=1}^{m}\left(\frac{1 - x_i^{a+1}}{a+1}\right)$$

$$a \leftarrow a - \alpha\frac{\partial f(k,a)}{\partial k} = a - \frac{\alpha}{m}\sum_{i=1}^{m}\left(-kx^a - \frac{kx^{a+1} + k}{(a+1)^2}\right)$$

- 初期値の探索は、候補となるkとaの値を総当たりで組み合わせて探すグリッドサーチで行い、ベストな値を求める。kが「平均的な『1日目の』閲覧数」である以上、kはだいたい数百くらいだろうと想像される（これは人によって変わるが、自分の場合はだいたい200人くらいかな、と見積もった）。このようにしてアタリをつけた上で、kとaを探す。最終的に収束する値は、初期値と学習率でかなり変わるので、何度も試すのが重要。
  - 1回目のグリッドサーチ→kを100~300まで10刻み、aを-0.5~-1.5まで0.5刻みで動かす。合計20*3=60個の組み合わせについて、目的関数の値が最小となるkとaの組み合わせを求める。
    - 結果として、k=229.9999235816114, a=-0.7510436347605907と得られた。（初期値はk=230.0, a=-0.5。平均二乗誤差は、最小値がbest_mse=3366215.395081603）
  - 2回目のグリッドサーチ→kを220から240まで1刻み、aを-0.4から-0.6まで0.05刻みで動かす。合計20*5=100個の組み合わせについて、目的関数の値が最小となるkとaの組み合わせを求める。
    - 結果として、k=229.99996126196385, a=-0.7448186941765937と得られた。（初期値はk=230.0, a=-0.55。平均二乗誤差は、最小値がbest_mse=3349427.6655487986）
- 以上より、公開期間xまでの総閲覧数yの予測値はだいたい以下のようになった（k=230, a=-0.745としたものなので、求めた値とは厳密には異なるが、こう書くとわかりやすいので書いてみた）。

$$y = 901.96(x^{0.255} - 1)$$

- 以下の図は、横軸に公開日数、縦軸にPV数をとっている。青い点は、訓練データのデータを表している。図中の曲線は、求めたkとaをもとに閲覧数を予測したものであり、上記の公開期間xまでの総閲覧数yの予測値の式を表している。この曲線より上の作品が「思ったよりよく読まれている」、つまり「人気がある」作品であり、曲線より下の作品が「思ったより読まれていない」、つまり「人気のない」作品である。



- 実際の総閲覧数を、この回帰式で得られた予測値で割った値を「**人気度（popularity）**」と定義する。値自体は正の値をとるが、1未満であれば「人気がない」、1以上であれば「人気がある」とざっくり捉えられる。では、実際それぞれの作品がどれくらいになっているかを見てみよう。
  - 人気度0~1：まあそんなに人気ないので言及するほどでもない。
  - 人気度1~2：ちょっとくらいはウケてる。多くは人気度0.5~1.5に収まるので、1.5を超えたあたりからまあまあ良い作品みたいな感じが出てくる。

タイトル	人気度	公開日数	PV数	予測PV数
なんにも知らないデストロイヤー	1.7320267080936438	1539	9242	5335.945431333569
女子陸上選手型アンドロイドを騙して素股させたりフェラしてもらう話	1.6502101928480575	709	6868	4161.894060384324
クレーム対応代行アンドロイド	1.2433399506410854	779	5336	4291.66616680231

  - 人気度2~3：準エース級。自分でも何回も読み返したくなる。なかなか書けないが、これくらいが書けると結構ウケたな～と思えるくらいの目安。

タイトル	人気度	公開日数	PV数	予測PV数
治安維持機関所属機械人形による性欲処理業務の一風景	2.57887007114972	442	9177	3558.5352293102073
痴漢被害引き受けアンドロイド（金髪褐色ギャル型）	2.342042015100076	168	5951	2540.9450221778843

  - 人気度3以上：代表作級。メチャクチャよく書けていたり、イラストが1000RT以上されていたりするとこれになる。内容的に面白い上に、運がよくないと多分難しい。

タイトル	人気度	公開日数	PV数	予測PV数
堕ちた偶像	3.12992318047358	41	4582	1463.933692873162

- では、この求めた回帰式から、テストデータの5作品の人気度がどのような感じになっているか確認する。テストデータは回帰式の算出には使っていないので、未知の作品ということになる。自分の体感の人気度とマッチしているだろうか？　結果は以下の通りで、大体一致している感じがした。特に「オナニーするだけの～」

は人気度4.2という破格の数字が出ていたが、多分これまで書いた作品の中でも一番ウケているので、この数字が出るのも納得といったところか。

タイトル	人気度
人身御供	1.728718
一緒に寝ていたオナニーサポート用アンドロイドのロリ美少女に、金玉カラカラになるまで絞られました。	2.107545
オナニーするだけのバイトと聞いていたはずなのに、気づいたら人型自慰行為補助機械のお姉さんの奴隷になっていました。	4.226200
戦場から遠く離れたところでは	0.465920
家事手伝いアンドロイドで精通した11歳の男の子のオナニー風景	1.328440

具体的な作品を通じて人気度を確認する

人気度トップ5

タイトル	人気度（PV数/予測PV数）	公開日数	PV数	予測PV数	一日あたりのPV数（公開日数/PV数）
堕ちた偶像	3.12992318047358	41	4582	1463.933692873162	111.76
治安維持機関所属機械人形による性欲処理業務の一風景	2.57887007114972	442	9177	3558.5352293102073	20.76
高身長イケメン女子なオレっ娘性欲処理専用アンドロイドは、俺のことが大好きすぎる性欲モンスターでした。	2.5354639662996092	905	11423	4505.289821440978	12.62
痴漢被害引き受けアンドロイド（金髪褐色ギャル型）	2.342042015100076	168	5951	2540.9450221778843	35.42
性のお勉強	2.042099103135243	1652	11142	5456.150479128874	6.74

人気度ワースト5

タイトル	人気度	公開日数	PV数	予測PV数	一日あたりのPV数（公開日数/PV数）
愛玩人形リサ-1-	0.5209191661415659	2209	3112	5974.055481679624	1.41
家庭用アンドロイドN-9H8397D1EJ2MA（介護機能オプション付き）	0.5037492221105312	1431	2627	5214.896390298725	1.84
Innocent	0.3946763987980405	1400	2044	5178.926346305124	1.46
愛玩人形リサ-3-	0.3427446374794778	2207	2047	5972.376446363996	0.92
愛玩人形リサ-2-	0.26079436402588674	2209	1558	5974.055481679624	0.70

- 訓練データの中にある人気度トップ5とワースト5を並べてみたところ、かなり直感的な理解がしやすい結果が得られた。人気度トップ5に並ぶものはどれも自分の作品で人気があるものだし、逆にワースト5は初期の作品や、知名度が低い二次創作、文字数の少ない作品が並んだ。
- この指標なら、「閲覧数を経過日数で割った値」が持つ欠点を解消できる。つまり、リリースしたての作品が、その作品の良し悪しに関わらずよく読まれるので、本当に人気がある作品よりも一日あたりの閲覧数が大きくなってしまう問題を回避できる。実際、テストデータを含めた人気度1位は「オナニーするだけのバイト

と聞いていたはずなのに、気づいたら人型自慰行為補助機械のお姉さんの奴隷になっていました。」だが、「閲覧数を経過日数で割った値」は19.2である。一方で人気度2位の「堕ちた偶像」は、「閲覧数を経過日数で割った値」が111.7と圧倒的に高い。「堕ちた偶像」は直近に出た、かつ人気もある作品だが、それよりも「オナニーするだけの〜」が高い人気を保ち続けているため、人気度という数字で測ることが作品がどれだけウケているかを真に表しているといえそうである。

まとめ

- 作品の人気を示す指標として閲覧数と公開期間から計算できる「人気度（popularity）」を定義した。人気度の大小が実際の肌感と合っていることを確認したので、以降はこれを作品の人気度を測る物差しと考え、分析を進めていくことにする。

5. 「分析」とは具体的に何をするのか

線形回帰

- ここまで、作品がどれだけウケているかを測る「人気度」を計算してきた。いよいよ、2章で考えた作品の性質のうちどれが、人気度に寄与しているのかを探っていく。そこで使うのが線形回帰モデルである。
- 例えば、あるコンビニに一日あたりに入るお客さんの数を予測したいとする。客の数は「コンビニの周囲100m以内に住んでいる人間の数」と「コンビニの周囲100m以内にある別のコンビニの数」で決まるとする。
- この仮定の元では、ある決まった数aとbを使って、（客の数）= a * （人間の数）+ b * （コンビニの数）という式で表されると想像できる。例えば、a = 0.4、b = -10 とすると、住んでいる人間が200人、コンビニの数が2件だった場合、一日の客の数は 0.4 * 200 + (-10) * 2 = 80 - 20 = 60人、と計算できる。
- このように、「何らかの事象を説明する数量に適当な数（回帰係数）をかけていき、それらの総和を取って目的の変数を計算する」という考え方を「線形回帰」と呼ぶ。回帰係数はそれぞれの要素の重要度を表していて、「コンビニの数が一件増えると、一日あたりの客が10人減る」といった関係にあることを示している。
- aやbがわかれば、先ほどのように住んでいる人間の数やコンビニの数を代入するだけで、一日あたりの客の数が求められるということになる。このaやbなどの「回帰係数」は、手元にあるデータから求めることができる。回帰係数を求める作業が、「モデルを作る」ことに相当する。
- 計算される対象の「客の数」は「目的変数」と呼ぶ。目的変数を計算するために使われる、住んでいる人間の数や周囲のコンビニの数などは「説明変数」と呼ぶ。
- 今回は、2章で挙げたような変数を説明変数として、目的変数である人気度を予測する。

6. 人気度を線形回帰する

- 人気度の計算は4章で終わったので、以下の流れに従って線形回帰をする。
  1. 目的変数の予測に使う説明変数を決定する。
  2. モデルを作り、性能を評価する。
  3. 回帰係数を眺めて、どの説明変数が人気度にどれだけ影響しているかを調べる。

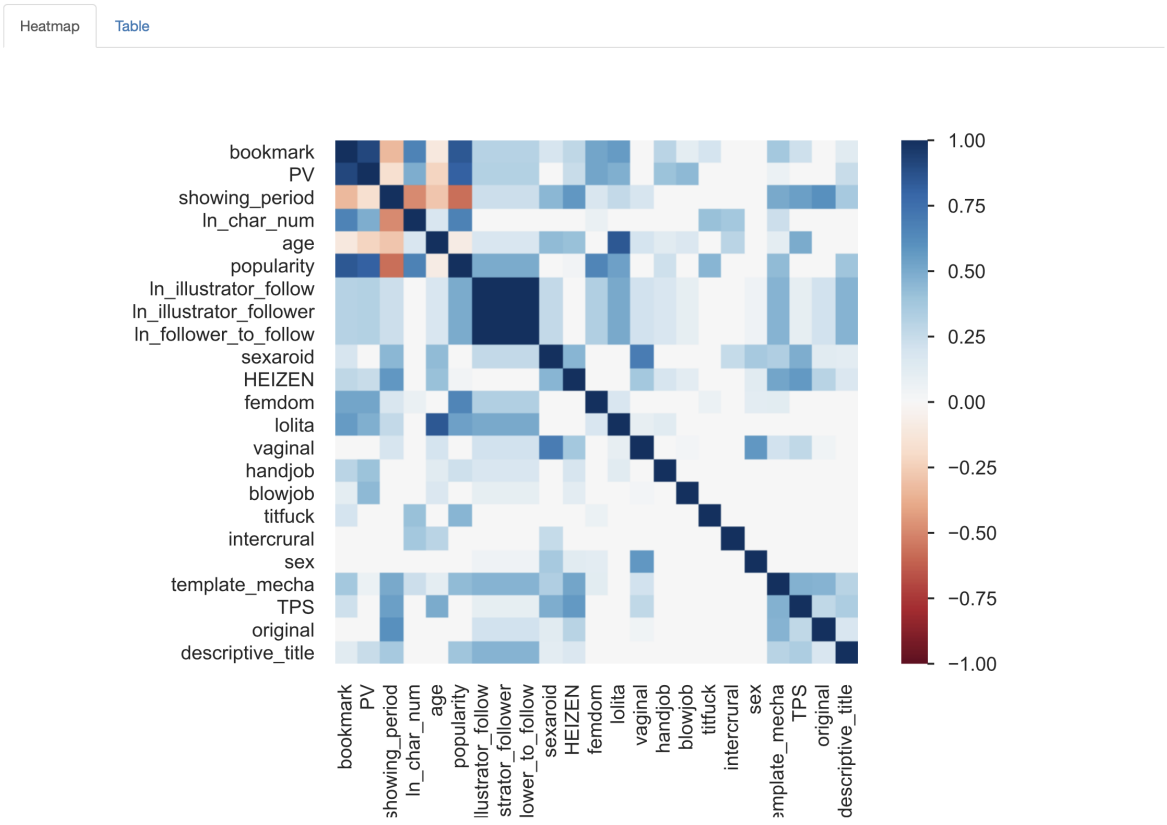
ステップ1：説明変数の選定

- 機械学習をするときに、訓練データの中身を見る工程がある。目的変数の予測に本当に役に立ちそうな変数を残し、不要なものを使わないようにしたり、逆に新たに役立ちそうな変数を作ったりするのである。この工程は「探索的データ分析（Exploratory Data Analysis: EDA）」と呼ばれる。
  - ある事象を説明するために置く仮定はシンプルな方がいいとか、モデルは単純な方が好ましい、みたいなことはこの分野でよく言われることである。（cf. オッカムの剃刀）
- 流れは以下の通りである

- ステップ3-1：ydata-profilingで、説明変数の分布や相関関係を眺める。やることは2つ。①：目的変数と相関がない（0の）説明変数は除外する。②強い相関がある説明変数は、そのうち一つだけを残す。
- ステップ3-2：分散拡大係数（VIF）を計算し、（ステップ3-1で取りこぼした）強い相関がある説明変数を整理する。

ステップ1-1：ydata-profilingでデータを観察し、人気度と相関がない説明変数を取り除く

- ydata-profiling というライブラリを使うことで、データがどう分布しているのかや、変数の間の相関関係を見たりできる。これを使って、強い相関がある変数、つまり同じような傾向を示す説明変数を見つけ、必要なもののみ残す。これはどういうことかというと、ある物事を説明する上で、同じようなデータがあったらどちらか一つで十分で、二つ用意する必要はないと理解すれば良い。
- 結果はこんな感じ。ヒートマップを載せる。（bookmark, PV, showing_periodは使わないことを念頭に置く）



- まず、popularityと相関がないもの（相関係数rが、-0.3 ≤ r ≤ 0.3となるようなもの）を列挙し、これらは使わないようにする。これらは、手元のデータでは相関関係が分からなかったもの、もしくは本当にpopularityと相関がないものなので、popularityの予測に貢献しないものだと考えられる。
  - sexaroid
  - HEIZEN
  - vaginal
  - handjob
  - blowjob
  - intercrural
  - sex
  - TPS

    - original
- 上記以外で高い相関（相関係数が0.7以上or-0.7以下）があるものを調べる。これらは多重共線性の原因になるので、いずれか一方を取り除きたい。次のステップでどれを残すか決める。また、他の説明変数についても、次のステップで取捨選択をする。
    - ageとlolita（0.848）
        - いずれも年齢に関係があるので相関する。次のステップで確認する。
    - ln_illustrator_followとln_illustrator_followerとln_follower_to_follow（1.0）
        - どれか一つのみ残さなければならない。次のステップで確認する。
- また、ここで見つけたわけではないが、パイズリ（titfuck）はそもそもデータ数が少なすぎて説明変数として使うのがかなり危ないと思ったので、除外することにした。

ステップ1-2：決定係数を確認し、他の説明変数で回帰できる説明変数を取り除く

- ある変数が、他の（複数の）変数と相関があるかどうかを確認するためには、分散拡大係数（VIF）を確認すれば良い。分散拡大係数を求めることと、決定係数という指標を求めることは同じなので、ここれでは決定係数を求める。VIFが5や10より大きい（つまり決定係数が0.8や0.9より大きい）時、ある変数は他の変数で説明できるということを示している。

$$\text{決定係数を}R^2\text{として、}VIF := \frac{1}{1-R^2}$$

- 決定係数は、線形回帰で自分のモデルがどれだけ真の値に近い値を予測できているかを測るための指標である。
- ステップ1-1で消さなかった説明変数に対し、それぞれ他の説明変数で線形回帰し、決定係数が0.8を超えているものを削除する。このステップを、決定係数が最も大きいものについて繰り返し行う。
- 1回目。ln_illustrator_followerの決定係数が最も大きいので、これを除く。

```
R^2 of ln_char_num: 0.4009954848093723
R^2 of ln_illustrator_follow: 0.99873223939539 79
R^2 of ln_illustrator_follower: 0.9995559253309438
R^2 of ln_follower_to_follow: 0.9984612608759696
R^2 of femdom: 0.3649195478711862
R^2 of age: 0.6531828680257779
R^2 of lolita: 0.6815651779575986
R^2 of template_mecha: 0.5445006011623095
R^2 of descriptive_title: 0.34117464048209445
```

- 2回目。決定係数が0.8を超えているものがないので終了。ageとlolitaについては、lolitaの方がR^2が大きかったので、より粒度の高いageを残し、lolitaを捨てることにした。

```
R^2 of ln_char_num: 0.4009946366389256 5
R^2 of ln_illustrator_follow: 0.4790183204890843
R^2 of ln_follower_to_follow: 0.654578414532959
R^2 of femdom: 0.36488246640463384
R^2 of age: 0.622762832050372
R^2 of lolita: 0.6296035523004566
R^2 of template_mecha: 0.5444669718426327
R^2 of descriptive_title: 0.3411277216172639
```

最終結果

- 以上より、人気度の予測に使う説明変数はこの7個に絞られた。
  - **文字数（ln_char_num）**
  - **絵師のフォロー数（ln_illustrator_follow）**
  - **絵師のフォロワー：フォロー比（ln_follower_to_follow）**
  - **女性上位（femdom）**
  - **設定年齢（age）**
  - **機械的な言動やメカバレ（template_mecha）**
  - **説明的なタイトル（descriptive_title）**

ステップ2：モデルの構築・評価

線形回帰モデルの種類

- 人気度を線形回帰するための準備が整った。いよいよモデルを作って、どのような結果になるのかを観察する。
- 先ほども少し述べたが、線形回帰は「何らかの事象を説明する数量に適当な数（回帰係数）をかけていき、それらの総和を取って目的の変数を計算する」という考え方である。コンビニの客の数を予測する式が以下のように書けるという仮定を置いて、aとb（回帰係数）を求めるのが線形回帰でやるべきことである。

$$\text{客の数} = (a \times \text{周りに住む人間の数}) + (b \times \text{周囲のコンビニの数})$$

- 線形回帰のモデルでよく使われるものは、リッジ回帰・ラッソ回帰・ElasticNetの3種類がある。名前が違うだけで、上の式自体は同じものを求める。ただ、何を目指すか（何を目的関数とするか）の違いで、呼び方が違う。
- 基本的に、線形回帰における目的関数は二乗誤差（実際の値と予測値の値）の総和であることが多い。この誤差が小さくなるように、最急降下法などの方法を使ってパラメータを変更していくのが、モデルを作る時に使われる手段である。

$$f(x_i)\text{を予測値、}y_i\text{を真の値として、}L = \sum_{i=1}^{n}(f(x_i) - y_i)^2$$

- しかし、リッジ回帰などは、単に二乗誤差を目的関数とするのではなく、この誤差に「正則化項」をつける。パラメータの絶対値が大きくなるのを防ぎ、ある説明変数が他の説明変数を圧倒してしまうような影響力を持つのを防ぐ役割がある（ある現象が一個だけの変数から説明できてしまう、なんて面白くない結果になるのを防いでくれる）。また、多重共線性で解がもとまらない問題を回避することもできる。以下はそれぞれ上から順に、リッジ回帰・ラッソ回帰・ElasticNetの目的関数である。（二乗の項はL2正則化項、絶対値の項はL1正

則化項と呼ぶ。λは正則化項が全体に及ぼす影響を制御するハイパーパラメータであり、適度な大きさにすることが求められる）

w_jをj番目の説明変数の重みとして、$L = \sum_{i=1}^{n}(f(x_i) - y_i)^2 + \lambda \sum_{j=1}^{m} w_j^2$

$$L = \sum_{i=1}^{n}(f(x_i) - y_i)^2 + \lambda \sum_{j=1}^{m} |w_j|$$

$$L = \sum_{i=1}^{n}(f(x_i) - y_i)^2 + \lambda_1 \sum_{j=1}^{m} w_j^2 - \lambda_2 \sum_{j=1}^{m} |w_j|$$

線形回帰モデルの評価

- 決定係数（R^2）と平均二乗誤差（MSE）を使う。訓練データとテストデータで分けたので、訓練データで作ったモデルをテストデータで評価する（ホールドアウト検証をする）。ラッソ回帰・リッジ回帰・ElasticNetの3種類のうち、決定係数が最も良いものを取り上げて分析する。
- テストデータに対する結果：（MSEがベストなものでも同じ結果になった）

	R^2	MSE	λ（正規化項の係数）
リッジ回帰	0.2196629862853423	1.2237899270887673	29
ラッソ回帰	0.2753903729959457	**1.1363935671560643**	0.06
ElasticNet	**0.27643140627096774**	1.1554019728873566	0.12

- ElasticNetの決定係数が最もよく、R^2 = 0.276であった。これは相関係数がだいたい0.525ということなので、まあまあ相関がある（モデルがテストデータに対する予測ができている）と言える。

ステップ3：回帰係数の解釈

- ElasticNetについて見ると、以下のようになった。

変数名	変数の種類	回帰係数
ln_char_num	数量	**0.24015759493652022**
ln_illustrator_follow	数量	**0.08853729530468853**
ln_follower_to_follow	数量	**0.10815679247466005**
femdom	バイナリ	**0.17172651522465532**
age	数量	**-0.003879890320590044**
template_mecha	バイナリ	**0.0**
descriptive_title	バイナリ	**0.0**

- template_mechaとdescriptive_titleの係数は0になった。これは、メカ表現や説明的なタイトルにしてもpopularityに影響が出ないことを示している。他の5つについて詳細に見ていく。

数量

- 文字数（char_num）→**0.24015759493652022**
  - 人気を決定づけるのに一番重要、土台となる指標。多い方が良い。2000文字、10000文字、20000文字の作品を比較する。（lnは自然対数をとっていることを表している）
    - 0.24*ln(2000)=1.824
    - 0.24*ln(10000)=2.210
    - 0.24*ln(20000)=2.377

    - つまり、2000文字の作品と10000文字の作品は、人気度が0.4違う。10000文字と20000文字は0.16違う。→文字数を2倍にすれば0.16、5倍にすれば0.4増えると言える。（文字数は対数をとっているので、人気度への寄与はn倍（n%）の変化で捉えるべき）
- 絵師のフォロー数（ln_illustrator_follow）→**0.08853729530468853**
- 絵師のフォロワー：フォロー比（ln_follower_to_follow）→**0.10815679247466005**
    - 以下の表は、上記の結果から導出したイラストレータのフォロー・フォロワー数に応じた人気度の寄与をまとめたものである。
    - フォロワー数が10倍になると人気度が0.25増える。また、たとえばフォロー数100人、フォロワー数10000人の人に依頼をすると、人気度が0.90上乗せされる→文字数5~6倍上乗せしたのと同じくらいの効果があるので、かなり効果が大きいと言える。というかフォロワー10000人以上の人に依頼するだけで自分の普通の作品と比べてよく読まれることが確約されるという感じ。
    - 正直フォロワー・フォロー比はあんまり気にしなくていい気がする。大事なのはフォロワーの数。

行：フォロワー・列：フォロー	10	100	1000
100	0.4529048744007904	-	-
1000	0.7019450924589932	0.656769530743378	-
10000	0.9509853105171959	0.9058097488015808	0.8606341870859654
100000	1.200025528575399	1.1548499668597834	1.1096744051441683

- 設定年齢（age）→**-0.003879890320590044**
    - 若い方が良い。12歳、24歳、40歳で比較する。
        - 12歳→-0.0468
        - 24歳→-0.0936
        - 40歳→-0.156
    - そんなに大きな影響を与えるわけではないが、とはいえ特にこだわりがないならロリ～10代後半くらいに絞って書いた方が良さそう。

バイナリ

- 女性上位（femdom）→**0.17172651522465532**
    - あったほうが良い。あるだけで0.17の上乗せになる。
    - これは文字数を倍にして得られる人気度と同じくらい大きな幅。

シミュレーションしてみる

- 文字数5000文字のネタがあるとする。まだイラスト頼んでないしマゾ向けでもない。年齢も適当に24歳くらいにしようかな……→この時の人気度は、popularity = 0.24 * ln(5000) + 0.089 * 0 + 0.108 * 0 + 0.172 * 0 + 24 * (-0.0038) = 1.733
    - 女の子の設定年齢を16歳くらいにして、逆レイプされる話にすると、popularity = 0.24 * ln(5000) + 0.089 * 0 + 0.108 * 0 + 0.172 * 1 + 16 * (-0.0038) = 2.155
    - 文字数を10000字まで増やすと、popularity = 0.24 * ln(10000) + 0.089 * 0 + 0.108 * 0 + 0.172 * 1 + 16 * (-0.0038) = 2.321
    - イラストを頼む。フォロー数500人、フォロワー数10000人の絵師さんに頼むと、0.24 * ln(10000) + 0.089 * ln(500) + 0.108 * ln(10000/500) + 0.172 * 1 + 16 * (-0.0038) = 3.447
- ちょっと多く見積もられている感じはあるが、逆レにする・文字数を増やす・イラストを頼むことがそれぞれ人気度に寄与していることが見て取れると思う。

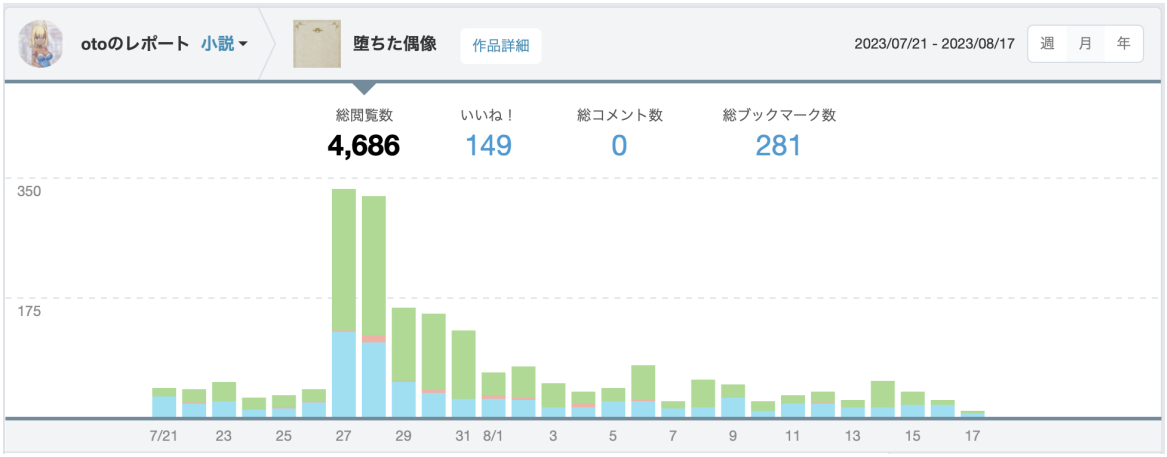
7. まとめ（作品を読んでもらうために大切な4箇条）

とにかく文字数は多くしよう！

- 内容を濃くする・導入からしっかり書くことを心がけよう！　10000文字を目安に書いていくといい。

イラストを描いてもらおう！

- やはりイラストレーターさんに挿絵を描いてもらうのは超効果的。Skebなりでフォロワーの多い絵師にイラストを描いてもらおう！
- でもこれは絵師さんがツイッターで宣伝してくれる場合に限った話をしているので、ツイッター上でちゃんとRTなりしてくれる人じゃないと当てはまらないことに注意しよう。（Skebで依頼する前に、その人がちゃんと描いたイラストを一般公開しているかどうかをちゃ～んと確認すること。FANBOX限定公開とか、そもそも公開すらしない人とかもいるのでそこは自己責任で確認しよう）
    - イラストレーターの方がイラストをRTするとPV数が跳ね上がる例。イラストを依頼するときは、自分のイラストをたくさんセルフRTする人にお願いすると良い。



マゾ向けの話を書こう！

- 女性上位やドMホイホイのタグをつけよう！　思っている以上に読まれます。

アンドロイドの設定年齢は未成年から20代に絞ろう！

- ロリは最強。

参考

- EDAから学習までの流れ
    - https://qiita.com/tk-tatsuro/items/793bc3ec181e30803157
- 少量データに対する機械学習のイロハ
    - https://www.kaggle.com/code/rafjaa/dealing-with-very-small-datasets
- 人気度の回帰について、最急降下法のお手本
    - https://zero2one.jp/learningblog/machine-learning-polynomial-regression/
- 目的関数と相関があるもののみを説明変数にするという考え方のリファレンス
    - https://istat.co.jp/ta_commentary/multiple_03
- 決定係数の概要

    - https://bellcurve.jp/statistics/course/9706.html
    - https://manabitimes.jp/math/1016
    - https://qiita.com/shnchr/items/fc321bfe4fca2b8565d7
    - https://ja.wikipedia.org/wiki/決定係数
- 決定係数を非線形回帰で使ってはいけない理由
    - https://bmcpharma.biomedcentral.com/articles/10.1186/1471-2210-10-6
    - https://qiita.com/FukuharaYohei/items/b7a3a7562c3c892b6d8e
- 線形回帰の概要
    - https://aizine.ai/ridge-lasso-elasticnet/
    - https://qiita.com/g-k/items/d3124eb00cb166f5b575
    - https://qiita.com/Takayoshi_Makabe/items/8f6dcb25124b9dcb1ae8