Q1. 社員属性推定

1. 概要

Q1. にて皆さんに取り組んでいただくのは「社員属性推定」です。

博報堂／博報堂ＤＹメディアパートナーズのオフィスが存在する赤坂 Biz タワーの 13 階には「100Tables」という名の社員食堂があり、多くの社員が同僚とのランチや息抜きの間食、簡単なパーティー、社外から訪問される方との打ち合わせなどを行っています。

今回、皆さんには 100Tables を博報堂／博報堂ＤＹメディアパートナーズ社員が利用した購買のデータをお渡しします。この「誰が」「いつ」「何を」「どれぐらい」食べたかの購買データをもとに、性別や年代、職種といった社員属性の予測に挑戦していただき、その予測精度で競い合っていただきます。

実はこのデータ、皆さんに出題するために初めて社内から集めました。そのため、我々社員ですらまだ分析したことがありません。ぜひ、我々も知らないような隠れたルールやパターンを見つけ出してください。

2. データ

`data/` ディレクトリには

- `train.xlsx`
- `test.xlsx`
- `sample_submission.xlsx`

という 3 つの xlsx ファイルが含まれています。それぞれ説明します。

2-1. train.xlsx / test.xlsx

`train.xlsx` と `test.xlsx` は社員食堂「100Tables」における 2018年11月1日から2019年1月17日までの各社員の購買履歴です。ふたつのファイルにはひとりの社員によってある利用時に購買された一種類の商品が一行ずつ記録されています。そのため、ある利用時に一種類の商品を複数個購買した場合には一行のデータが、複数種類の商品を購買した場合には複数行のデータが記述されています。

`train.xlsx` は 13 列 55945 行、 `test.xlsx` は 10 列 4095 行 (両ファイルともヘッダー込) で構成されています。

year	month	day	hour	minute	second	price	amount	menu_name	haku_id	gender	age	job
2018	11	01	10	00	10	360	1	フルーツヨーグルト	108e82	男性	30代	アカウントプロデュース職
2018	11	01	10	00	10	100	1	Ｓドリンク	108e82	男性	30代	アカウントプロデュース職
2018	11	01	10	00	56	160	1	Ｓドリンク	e3027d	男性	40代	マネジメントプロデュース職
2018	11	01	10	01	25	120	1	おにぎり	3d1303	女性	20代	ストラテジックプラニング職
2018	11	01	10	01	25	120	1	スープ(単品)	3d1303	女性	20代	ストラテジックプラニング職

この表は `train.xlsx` の一部を抽出・加工したものです。それぞれの列の説明は以下の通りです。

- `year` : 購買した年。
- `month` : 購買した月。
- `day` : 購買した日。
- `hour` : 購買した時。
- `minute` : 購買した分。
- `second` : 購買した秒。
- `price` : 購買した商品の価格。
- `amount` : 購買した商品の個数。
- `menu_name` : 購買した商品名。

- `haku_id` : 社員に対して一意に付与された ID。
- `gender` : 当該社員の性別であり、 `男性` と `女性` のいずれかで表記されています。
- `age` : 当該社員の年代であり、 `20代` 、 `30代` 、 `40代` 、 `50代` の 4 区分で表記されています。
- `job` : 当該社員の職種であり、 `PR職` 、 `ナレッジ開発職` 、 `クリエイティブ職` 、 `メディアプランニング職` 、 `メディアプロデュース職` 、 `アカウントプロデュース職` 、 `コンテンツプロデュース職` 、および `ビジネスディベロップメント職` の 10 種のいずれかです。

つまり、先程お見せした表は

- 30代・アカウントプロデュース職・男性の `108e82` さんが 2018/11/01 10:00:10 にフルーツヨーグルトとSドリンクをひとつずつ購買
- 40代・マネジメントプロデュース職・男性の `e3027d` さんが 2018/11/01 10:00:56 にSドリンクをひとつ購買
- 20代・ストラテジックプラニング職・女性の `3d1303` さんが 2018/11/01 10:01:25 におにぎりとスープ(単品)をひとつずつ購買

という 3 つの現象を意味しています。

`gender` 、 `age` および `job` (以降、これらの 3 列をまとめて **社員属性** と呼びます) は `train.xlsx` にのみ含まれる列であり、 `test.xlsx` には含まれていません (列数の差分はこれに由来しています)。今回予測していただきたいのは皆さんの手元には存在しない、 `test.xlsx` に含まれる `haku_id` 167 人分それぞれの社員属性です。

2-2. sample_submission.xlsx

`sample_submission.xlsx` は解答用のサンプルファイルです。このファイルは 4 列 168 行で構成されています。各列の説明は以下の通りです。

- `haku_id` : 予測対象である社員の ID。
- `gender` : 当該社員の予測した性別。初期値として全て `男性` としています。
- `age` : 当該社員の予測した年代。初期値として全て `20代` としています。
- `job` : 当該社員の予測した職種。初期値として全て `アカウントプロデュース職` としています。

2 列から 4 列目の社員属性をご自身の力で予測していただき、最もそれらしいと思われるものを記載し、そのファイルを投稿してください。投稿することで採点が行われます (採点方法については後述します)。

社員属性には前述した値のいずれかが入ります。 **つまり、実際の年代が `10代` や `90代` であったり、実際の職種が `社長` や `顧問` である、ということは発生しません。**

3. 投稿方法

皆さんには 2 種類のファイルを投稿していただきます。それぞれ説明します。

3-1. 属性の予測結果

予測した社員属性を記述したファイルは、投稿してはじめてどの程度予測が正しいのかの判定が行われます。予測した社員属性は `sample_submission.xlsx` と同じ形式で (つまり、 2 行目以降の 2 列目から 4 列目を書き換えて) `xlsx` ファイルに書き込んでください。 **列の順序や 1 行目の内容が変更されている場合。2 行目以降がどれだけ正しくてもスコアは 0 点です。**

予測結果を書き込んだファイルはマイページにおける **「Q1の課題を提出してください。 / 予測結果を提出してください。」フォーム** からアップロードしてください。ちなみに、投稿するファイル名を `sample_submission.xlxs` にする必要は無いので、ご自身で区別しやすいように命名してください。

3-1-1. 属性予測の採点方法

皆さんが予測した社員属性と実際の社員属性とにもとづいて **スコア** の採点が行われます。スコアは **全属性情報のうち、実際に正しく予測できたものの割合に100をかけたもの** で定義されます。

たとえば、 `test.xlsx` に予測対象の社員が 1 人しか存在していなかったとしましょう。もし、性別・年代・職種をそれぞれ `男性 / 30代 / アカウントプロデュース職` と予測し、実際の社員属性が `女性 / 30代 / クリエイティブ職` だった場合、正解しているのは年代のみなのでスコアは `1/3 * 100 = 33.4` です。

3-1-2. 採点結果の確認方法

投稿された予測は毎日午前 11 時にスコアの計算が行われ、正午に特設ページ (http://h-mp-recruit.jp/2020/special/) において

- スコア上位 5 名の 出場者 ID とスコアが記されたランキング
- 全参加者の 出場者 ID とスコアが記された xlsx ファイルが含まれた zip ファイル

のふたつが公開・更新されます。zip ファイルを展開するためのパスワードはマイページに記述されているものです (この問題文を読むために入力されたパスワードと同じものです)。

皆さんを識別するための出場者 ID の計算方法については特設ページ (http://h-mp-recruit.jp/2020/special/)を参照してください。また、ランキングに表示されている `sample_submission.xlsx` は `sample_submission.xlsx` をそのまま解答として投稿した場合に得られるスコアです。

3-2. どう取り組んだか

また

- どのように分析に取り組んだか
- データからどのような発見があったか
- スコアの改善に貢献した仮説、貢献しなかった仮説

などを説明するファイルを追加で投稿することも可能です。ファイルフォーマットは pdf で、5 ページ以内でお願いします。取り組みへの説明を記述したファイルはマイページにおける **「Q1の課題を提出してください。 / 予測した過程を記述したファイルを提出してください。」フォーム** からアップロードしてください。

こちらのファイルは毎日採点が行われるわけではなく、全ての解答を締め切ったのちに確認が行われます。

3-3. 投稿したファイルについて

ファイルは投稿するたびに上書きされ、また、最終的な皆さんの評価は最後に投稿されたファイルにもとづいて行われます。

そのため、たとえば予測を試行錯誤した場合、 **解答を締め切る 2019 年 3 月 31 日正午までにはもっとも良いスコアだったファイルを投稿することを忘れないでください。**

4. 注意

- **今回お渡しする購買データを Twitter や Facebook などの SNS 、掲示板、 GitHub などのコード共有サービスへアップロードすることを禁じます。**
- **解答の投稿は 2019 年 3 月 31 日正午まで可能です。**