Week 1

T.Nakao

1.
Competitionの内容を知る

Overview

Description

[原文]
This challenge serves as final project for the "How to win a data science competition" Coursera course. In this competition you will work with a challenging time-series dataset consisting of daily sales data, kindly provided by one of the largest Russian software firms - 1C Company.

We are asking you to predict total sales for every product and store in the next month. By solving this competition you will be able to apply and enhance your data science skills.

[Google翻訳]
この挑戦は、「データサイエンス競争に勝つ方法」コースラコースの最終プロジェクトとなります。
この大会では、ロシア最大のソフトウェア企業である1C Companyから親切に提供された日々の販売データからなる挑戦的な時系列データセットを使用して作業します。

次の月にすべての製品と店舗の合計売上を予測するように求めています。この競争を解決することにより、データサイエンススキルを適用し、強化することができます。

Overview

Evaluation

[原文]
Submissions are evaluated by root mean squared error (RMSE). True target values are clipped into [0,20] range.

Submission File

For each id in the test set, you must predict a total number of sales. The file should contain a header and have the following format:

```
ID,item_cnt_month
0,0.5
1,0.5
2,0.5
3,0.5
…
```

[Google翻訳]
提出物は、二乗平均平方根誤差（RMSE）によって評価される。 真の目標値は[0,20]の範囲にクリップされます。

提出ファイル

テストセットの各IDに対して、売上の合計数を予測する必要があります。ファイルにはヘッダーが含まれていて、次の形式を持つ必要があります。

Data

Competition Data

• item_categories.csv	3.49KB	85行(*ヘッダ[項目名]行含む)
• items.csv	1.5MB	22,171行(〃)
• shops.csv	2.91KB	61行(〃)
• sales_train.csv.gz	12.89MB	
(sales_train_v2.csv	90.2 MB)	2,935,850行(〃)
• sample_submission.csv.gz	468.75KB	
(sample_submission.csv	2.14 MB)	
• test.csv.gz	1.01MB	
(test.csv	3.03 MB)	214,201行(〃)

Data

Data Description

[原文]
You are provided with daily historical sales data. The task is to forecast the total amount of products sold in every shop for the test set. Note that the list of shops and products slightly changes every month. Creating a robust model that can handle such situations is part of the challenge.

[Google翻訳]
日々の過去の販売データが提供されます。 タスクは、テストセットのために各店舗で販売された製品の総量を予測することです。 店舗や製品のリストは毎月わずかに変更されることに注意してください。 そのような状況に対応できる堅牢なモデルを作成することが課題の一部です。

Data

File Description

File	Description
sales_train.csv	the training set. Daily historical data from January 2013 to October 2015.(トレーニングセット。 2013年1月から2015年10月までの日別履歴データ。)
test.csv	the test set. You need to forecast the sales for these shops and products for November 2015.(テストセット。 2015年11月には、これらの店舗や商品の売上を予測する必要があります。)
sample_submission.csv	a sample submission file in the correct format.(正しい書式のサンプル提出ファイル。)
items.csv	supplemental information about the items/products.(商品/商品に関する補足情報)
item_categories.csv	supplemental information about the items categories.(アイテムカテゴリに関する補足情報。)
shops.csv	supplemental information about the shops.（店舗の補足情報)

Data

Data fields

Data field	Description
ID	an Id that represents a (Shop, Item) tuple within the test set(テストセット内の（ショップ、アイテム）タプルを表すId)
shop_id	unique identifier of a shop(店の一意の識別子)
item_id	unique identifier of a product(製品の一意の識別子)
item_category_id	unique identifier of item category(明細カテゴリの一意の識別子)
item_cnt_day	number of products sold. You are predicting a monthly amount of this measure(販売された製品の数 この措置の月額を予測しています)
item_price	current price of an item(商品の現在の価格)
date	date in format dd/mm/yyyy(書式dd / mm / yyyyの日付)
date_block_num	a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33(便宜のために使用される連続した月番号。 2013年1月は0、2013年2月は1、...、2015年10月は33)
item_name	name of item(アイテムの名前)
shop_name	name of shop(店名)
item_category_name	name of item category(明細カテゴリの名称)

2.
とりあえずやってみる。

予測するもの：

・各店舗で販売された製品の総量

・test.csv … 2015年11月の店舗×製品毎の ”月間” 販売数

ID	shop_id	item_id	item_cnt_month
0	5	5037	？？？(予測する値)
1	5	5320	？？？(予測する値)
…	…	…	…
5100	4	5037	？？？(予測する値)
…	…	…	…

コレをsubmit(提出)する

学習/検証用データ：

・2013年1月～2015年10月までの “日別” 販売履歴

・sales_train.csv ＋ 各マップcsv(shops.csv, items.csv, item_categories.csv)

date	dt_blck_nm	shop_id	item_id	item_price	item_cnt_day	shop_name	item_name	itm_category_id	itm_category_name
02.01.2013	0	59	22154	999.0	1.0	Яр…	ЯВ…	37	Кино - Blu-Ray
03.01.2013	0	25	2552	899.0	1.0	Мо…	DEEP…	58	Музыка - Винил
05.01.2013	0	25	2552	899.0	-1.0	Мо…	DEEP…	58	Музыка - Винил
…	…	…	…	…	…		…	…	…

sales_train.csv

各マップcsv(shops, items, item_categories)

予測モデル：

・製品の総量の予測(数値の予測)なので、回帰モデル。

$$f$$

✓ 線形回帰系
- ☐ 線形回帰 ← 今回はコレ使う
- ☐ Ridge回帰
- ☐ Lasso回帰
- ☐ Elastic Net

✓ 決定木(回帰木)系
- ☐ 決定木, ランダムフォレスト, 勾配ブースティング回帰木(GBRT), …

✓ 時系列系
- ☐ AR … 自己回帰(Auto Regressive)
- ☐ MA … 移動平均(Moving Average)
- ☐ ARMA … AR + MA
- ☐ ARIMA … ARMA + Integrated(和分[積分])
- ☐ SARIMA … ARIMA + Seasonal(季節変動)

✓ ニューラルネット系
- ☐ NN, RNN(LSTM)

あたりがありそう。。。

その他備忘：

- Note that the list of shops and products slightly changes every month. Creating a robust model that can handle such situations is part of the challenge.

[Google翻訳]
店舗や製品のリストは毎月わずかに変更されることに注意してください。 そのような状況に対応できる堅牢なモデルを作成することが課題の一部です。

＞＞ 店舗や製品を特定する(強く依存する)特徴量は使用を控える、ってことかな(？)

- 学習/検証データ　　→ “日別”値
  予測　　→ “月間”値

＞＞ 学習/検証データを”月間”値に加工するのが良さげか(？)

- True target values are clipped into [0,20] range.

- あとは1Weekの「Feature Preprocess…」の内容でデータ加工すれば何となくいける？

最終的にやったこと：

- データは以下を用意しました。

Data field	Description
ID	Testデータのみの項目。モデルには使用せず、使用は提出時のみ。
date_block_num	連続した月番号。 時系列の数値変動を予測するためには時間項目が何らか必要と考えて使用。 Testデータには値34(予測月の2015年11月の月番号)を付与。
i_ctgry_flg_00 ～ 83	アイテムカテゴリをダミー変数化(0,1)したもの。 アイテムの定性項目は予測に寄与すると考えて使用。
item_price	アイテムの価格。 アイテムの価格は予測に寄与すると考えて使用。Trainデータはそのまま使えばよいが、Testデータには項目がないため作成する必要あり。(今回はTrainデータでitem_id毎にitem_priceの平均値を取得して作成した。欠損は0で補完。)
item_cnt_month	今回の予測対象。Trainデータのみの項目。Testデータに対してはモデルの出力値。 元のデータはitem_cnt_day(日単位)だったので、課題で要求されている月単位の数値にするために、date_block_num(月番号)×shop_id×item_id毎にitem_cnt_dayをsumして作成した。

特徴量として使用

← 予測対象

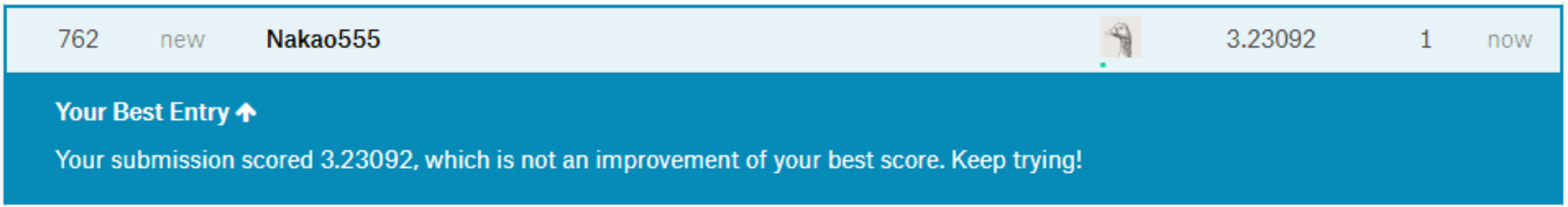
＊ その他、ショップの定性項目も予測に使えそうと考えて、ショップ名(ロシア語)をGoogle翻訳して所在の都市名などを抽出して項目化しようとしたが、数行やったら大変だったので断念しました。。。(でも何らか出来そうでした)

最終的にやったこと：

- モデルは線形回帰を使いました。LinearRegression()

- 交差検証(モデルの汎化性能を検証)するため、Trainデータを7:3に分割して、7をモデル作成に使用しました。
  7側の精度　→ RMSE＝6.4439…
  3側の精度　→ RMSE＝6.8019 …　　※大差ないと判断し検証OKとしました。

- 精度は置いといて。。。 (ランキング：ワースト20位圏内、wow!)



  ✓ とりあえずKaggleの雰囲気が体験できました。
  ✓ 個人的にはpanda, numpyの知識が全然足りないのでそこが課題です。
  データ加工がほとんどできないです。。。(今回は仕方ないのでMSAccessを使ってデータ加工しました)

以　上