The 2nd Big Data Analysis Contest
売上予測部門 レポート

一番シンプルなモデル

moemoe図鑑142種コンプ

背景

- 今回の目的:
  - CVSお菓子の販売数を予測する

- なぜそれが必要:
  - 仕入れ ＜ 需要 → 欠品による損失
  - 仕入れ ＞ 需要 → 廃棄による損失
  - 購入点数を予測することが営業利益に繋がる

最初に試した方法

- 線形回帰（機械学習）
  - 各属性のセグメントごとに、一つのモデルを構築
  - 説明変数：SNSデータ（お菓子種別の書き込み数など）
  - 被説明変数：翌月の販売数
  - 精度：0.21台
- 多項式回帰
  - 各セグメントの12データに対して、一つの多項式をフィッティング
  - 精度：0.23台
- ARMAモデル
  - 上記と同じく、一つのセグメントに対して、一つのモデルを構築
  - 精度：0.20台

考察

- どっちもベンチマークを下回っている
  - ベンチマーク：単純にセグメントの平均値で予測、精度：0.19台

- 今回はなぜ、上記の手法がうまく行かなかったのか
  - 訓練データが極めて少ない（一つのモデルに対して12個しかいない）
  - 今回の評価関数はRMSLEなので、大きな誤差に対する感度が高い

計算してみた					
実際値	5.0	6.0	7.0	6.0	RMSLE
予測1	6.0	5.0	6.0	7.0	0.1442
予測2	7.0	6.0	5.0	6.0	0.2034

  - 个平均絶対誤差は同じだが、RMSLEが違う（この例では50%の差がある）
  - 仮説：今回の場合、**精確に予測することより、大きな外れを出ないことを目指した方が正しい**かもしれない
  - 実際のビジネス場面でも、こうした方が得た利益が高い？（不明）

解決策

- ベンチマークようなシンプル&robustnessが高いモデルを構築した

- 各セグメントに対して、この式で予測値を算出

$$f = w_1 * x_1 + w_2 * x_2 + w_3 * x_{12} + w_4 * 1/9(x_3 + \ldots + x_{11})$$

f : 201606の与え

x_1 : 201605の与え

x_2 : 201604の与え

…

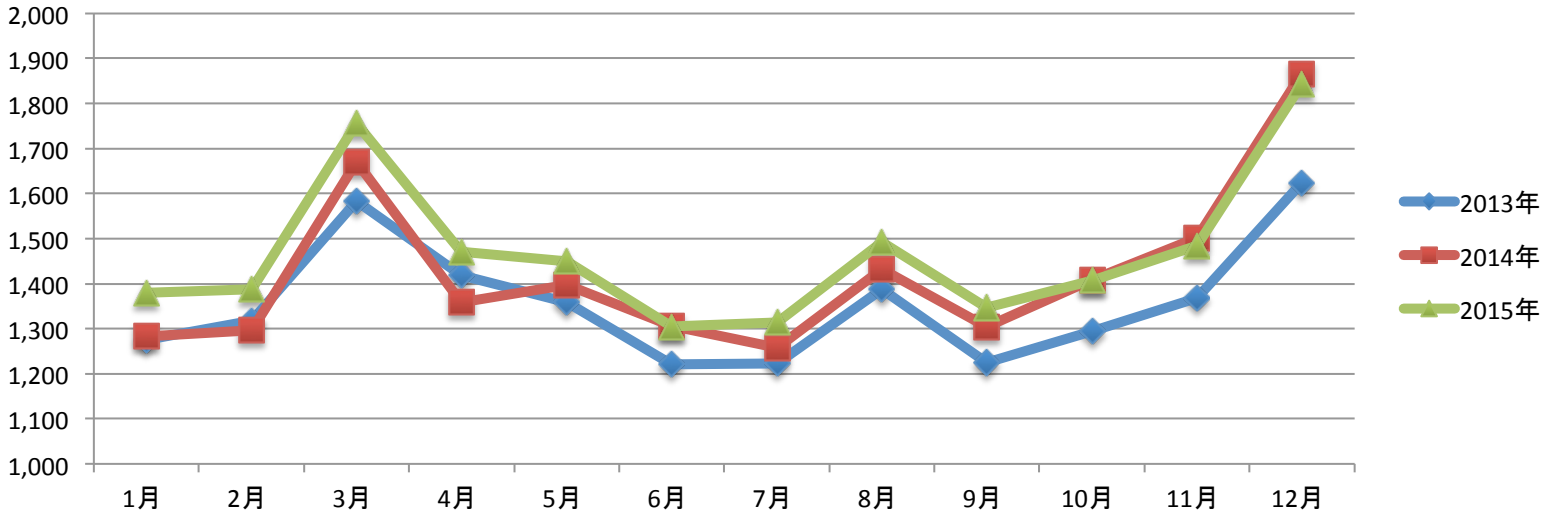
x_{12} : 201506の与え

$w_1 + w_2 + w_3 + w_4 = 1$

なぜこういうモデル

- リピート客が多く存在ので最近の販売数に強く関係している　→　2016年4月&5月のデータを利用

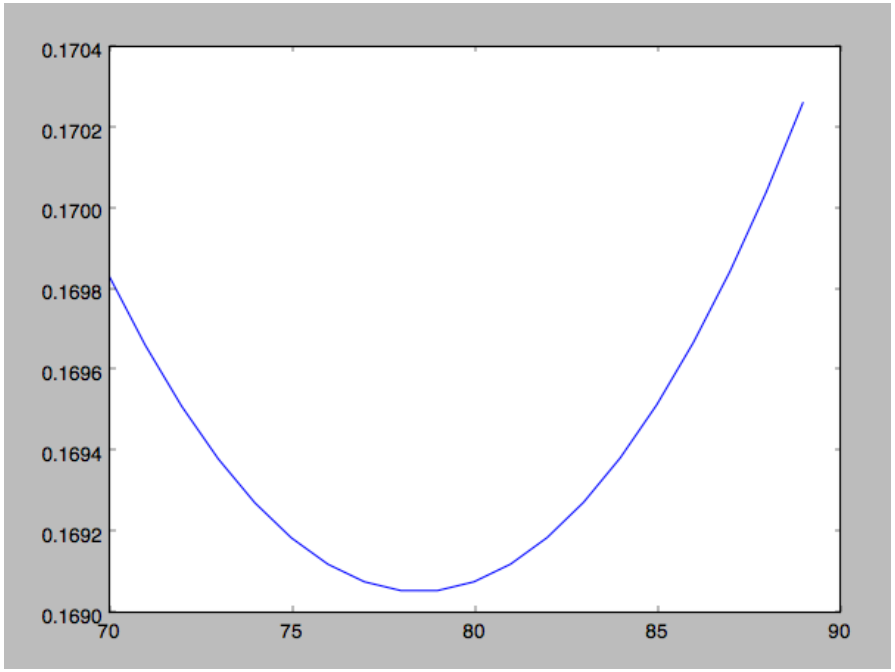
- お菓子の季節性がある → 2015年6月のデータも利用
  - コンビニが売っている五つカテゴリの総和は、毎年のトレンドが似ている
  - 3月と12月がピーク、6月、7月と9月が谷、コンビニのデータとほぼ一致



「1世帯当たり月別のお菓子支出金額」（総務省家計調査報告）より作成

重みをどうやって決めるか

- 過去のデータからw_1、w_2を学習
  - ある月を被説明変数にして、それ以前のデータを説明変数に
  - Brute-force searchingで一番いい精度になるw_1、w_2を探す
  - 計算量を減らしたい場合なら最急降下法でも行ける気がする



  - w_1対w_2がおよそ3対1の時、一番いい精度が得られる

細かい調整

- モデル安定性のため小さい割合(w_4)で平均値を入れた
- 2016年6月7日~2016年6月20日の間に、ナチュラルローソンでお菓子キャンペーンが行われたので、調整を行った
  - ナチュラルローソン部分の予測値を15%（直感）上げた
  - Reference: https://myunti.com/sale-campaign/lawson-natural1606.html
- 統計データによると、6月のお菓子の消費量が低いので、最後に全体的な調整を行った
  - できた予測値に0.95（家計調査データから計算した）をかけた

2015年7月以降発売された場合

- 発売月が2015年7月以降のセグメントに対して、欠損値があるため、下記の調整を行った

- 発売月が2016年6月の場合：
  - 極端に少ないので、適当に2.0にした
- 発売月が2016年5月の場合：
  - 201605の与えを1.6倍（データから計算）にした（月末発売の要素を考慮）
- 発売月が2016年4月の場合：
  - 201605の与えそのままで
- 発売月が2015年7月～2016年3月の場合：

$$f = w_1 * x_1 + w_2 * x_2 + (w_3 + w_4) * 1/(12-n)(x_n + \ldots + x_{11})$$

まとめ・課題

- 訓練データが少ないことに苦労した
  - 前回のモデルをそのまま適用することができない
- 販売数を予測するには、いろいろな要素に左右される
  - 競合他社の動き
  - 店内の環境・サービスなど
  - 人々食生活の変化
  - 外国人観光客の人数
  - 気候
- なので、今回はそこそこいい精度（0.15台）ができたが、いいモデルとはまだまだ言えない

疑問

- 月末発売の場合、その月の数字が極めて低いので、日数で割ったほうが扱いやすいかもしれない
  - お菓子の販売日を調べることができるが、その店での発売日は分からない

- 販売数が0の場合、そのコンビニが分母に入るかどうかは分からない、もし入っていない場合なら:

例えば、販売数が以下の場合					
	店A	店B	店C	店D	y
201603	3	3	3	1	2.5
201604	3	2	3	0	2.67

  - 201603のほうが売れているのに、数字的には201604のほうが高い…