共同研究「経営実務データを用いたデータサイエンティスト育成方法の研究」報告書

2018 年 3 月 26 日

国立大学法人滋賀大学データサイエンス教育研究センター長

竹村　彰通

１．共同研究の趣旨

本共同研究の研究題目は「経営実務データを用いたデータサイエンティスト育成方法の研究」であり、昨年度に引き続き、本共同研究の目的は企業の経営実務データを用いて販売予測や適性な在庫管理など、これからのデータサイエンティストに求められる能力を育成するための方法を研究することである。本共同研究の期間は、平成 29 年 9 月 6 日から平成 30 年 3 月 31 日までである。

特定非営利活動法人ビュー・コミュニケーションズは実際の販売データを教育目的のために適切に処理をおこなった上で国立大学法人滋賀大学に提供し、滋賀大学データサイエンス教育研究センターはデータをプロジェクト型の演習の形に整備する作業を担うことで、データサイエンティスト育成方法に関する共同研究をおこなった。

本年度はビュー・コミュニケーションズから、大型小売店舗の仕入れと販売実績の大規模なデータが提供された。このデータは滋賀大学に平成 29 年 4 月 1 日に開設された日本初のデータサイエンス学部の学生にとって、ビッグデータの一端を実感できる実務データであり、特に２年次以降の演習のために非常に有用である。その点で本共同研究の意味は大きい。

２．滋賀大学データサイエンス学部の育成人材像と演習の重視

滋賀大学データサイエンス学部では、データサイエンスの基礎要素技術としてのデータエンジニアリング(情報工学)及びデータアナリシス(統計学)に加えて、データからの価値創造の能力の育成を重視している。

この価値創造の能力を育成するためには、実際のデータを用いた演習の中で、試行錯誤や成功体験が必要でありさらに教科書的な理論と実際のデータとの乖離についても経験を積む必要がある。ビュー・コミュニケーションズから提供されたデータは、データ分析において試行錯誤を伴う作業を経験するためにも、また理論と実際との乖離を経験するためにも有用である。理論と実際の乖離については、ビュー・コミュニケーションズ副理事長の小松秀樹氏の著書「なぜあなたの予測は外れるのか－AI が起こすデータサイエンス革命」においても多くの例とともに示されており、データサイエンス教育において実際のデータを扱うことが非常に重要であることがわかる。小松秀樹氏には 2017 年 12 月 4 日に滋賀大学データサイエンス学部 1 年生向けの講義を担当いただき、実務データの見方や扱い方に関するノウハウを伝えていただいた

滋賀大学は、2016 年 12 月に「数理及びデータサイエンスに係る教育強化」の６拠点校の一校として、北海道大学、東京大学、京都大学、大阪大学、九州大学とともに文部科学省より選定を受けた。これは滋賀大学のデータサイエンス教育の先進性が高く評価されたためであると考えられる。特に、ビュー・コミュニケーションズから提供されたデータを用いた演習教材のような、価値創造につながる教育コンテンツの開発は拠点校として滋賀大学が求められている活動である。

３．滋賀大学データサイエンス学部における演習の設計

滋賀大学データサイエンス学部における演習の設計は「PPDAC サイクルを回す」ことにという考え方に基づいている。PPCAC サイクルとは、問題解決における各段階を Problem（問題）、Plan（調査の計画）、Data（データ）、Analysis（分析）、Conclusion（結論）に分けるという考え方である。ビュー・コミュニケーションズ提供のデータに基づく演習においても、データをそのまま学生に示すのではなく、例えば販売と仕入れの関係をどのように考えるか(Problem)、その問題を考えるときにどのようなデータが必要か(Plan)なども考えさせることとしている。これについては以下の 5 節に例示する。

データの分析手法や結論の導き方については、2 年次後半で履修する時系列解析入門や 3 年次前半で履修する時系列解析など手法を学んだ後のほうが、深い分析が可能となる。しかしながら、滋賀大学データサイエンス学部では、まず手法を学ぶのではなく、初年次から演習を重視しデータから出発してその分析の必要のために手法を学ぶこととしている。初年次は比較的小規模なデータを分析し、学年が上がるごとにだんだんと本格的な実務データを分析する経験を積むことが望ましい。本年度ビュー・コミュニケーションズから提供されたデータは 2 年次以降の演習に適切と考えられるため、本共同研究では、主に 2018 年度の 2 年生向けの演習の作成をおこなった。

４．本年度ビュー・コミュニケーションズより提供されたデータ

本年度ビュー・コミュニケーションズから提供されたデータは、とある大型小売店舗の販売実績と仕入れの時系列データである。データの時期は 2014 年 12 月 22 日から 2017 年 12 月 17 日の 3 年分であり、週次データである。本年度のデータは品目数が非常に多いものであり、ペット用品が 6,551 品目、日用品が 7,776 品目となっている。これらのデータを可視化することによって、実際の販売実績の様子を理解することができ、また仕入れがどのようにおこなわれているかについての分析を行うことができる。

５．提供データを用いた演習課題の例

以下の分析例は、滋賀大学データサイエンス教育研究センターの齋藤邦彦教授及び姫野哲人准教授による。

まずペット用品データの可視化をおこなう課題を考える。販売と仕入れの相関、販売額のトレンドなどを理解することが演習の目的となる。基本データとしては、以下の数値を抑えておく。

最大仕入れ；357 個／7 日　最大売上；200 個／7 日
最小仕入れ：-117 個／7 日　最小売上：-6 個／7 日
平均仕入れ：1.12 個／7 日　平均売上：1.21 個／7 日
全仕入れ：1,223,201 個　全売り上げ：1,240,645 個

ペット用品として消臭剤やケージ水槽･飼育セットなどがあるが、主な売り上げは飼料だと思われる。一例として、商品 90004 の仕入れ時系列データを折れ線グラフで描いてみる。ほとんどトレンドは見られない。ペット用品の売り上げが 20 単位で集計されていることが読み取れる程度である。

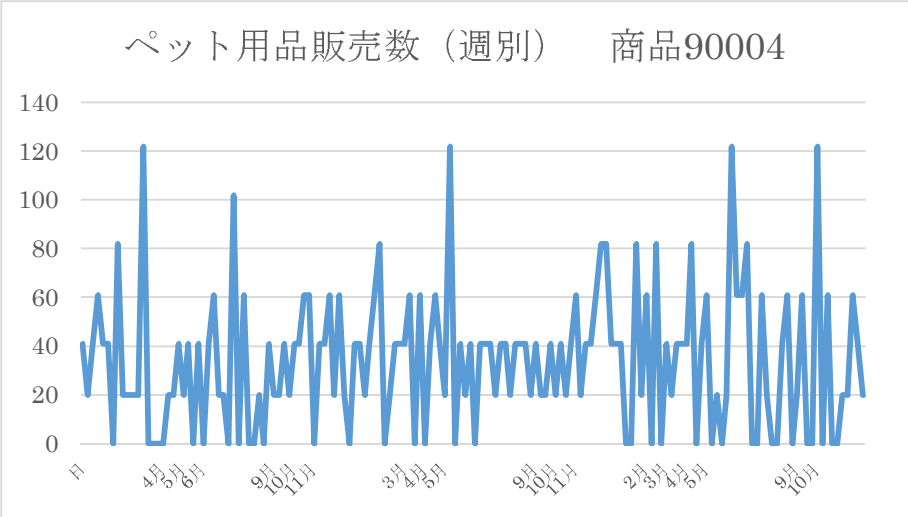

図 1

商品 90004 次に仕入れ時系列データと販売時系列データを組み合わせる。仕入れデータと数期先の販売データに相関がみられる。これは仕入れと販売の関係から自然である。

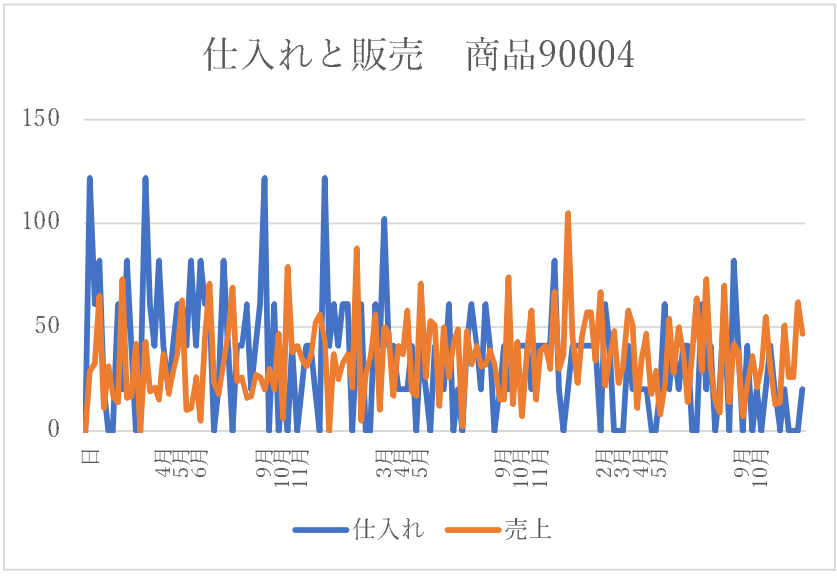

図 2

月別に全商品を集計すると、わかりやすい上昇トレンドが見ることができる。これは月ごとにまとまった販売が行われることが一つの要因であると思われる。ペット用品の性質上、消費者がある程度のサイクルで購入するものであり、消費者は安い時にまとめて買うものと思われる。例えば、月末の決算セールやポイント付与日などが考えられる。日ごと

のデータや、ポイント付与日の情報があれば、この点がよりはっきりと見ることができるだろう。

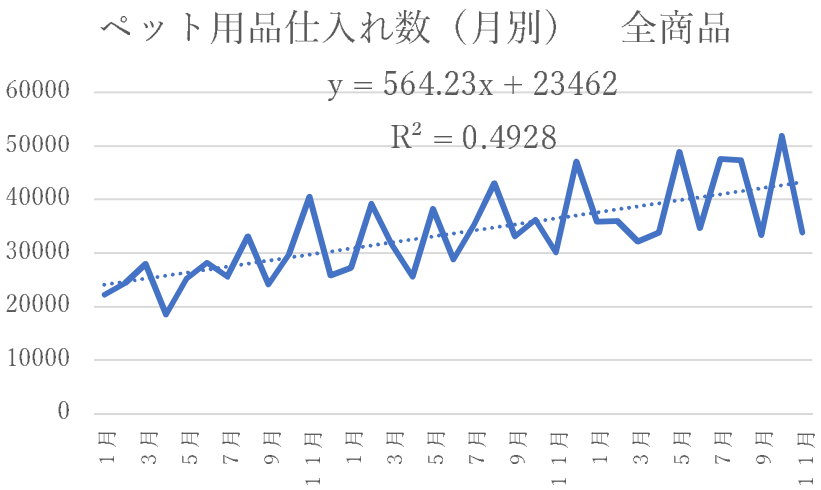

図 3

次に日用品の販売実績を分析する。以下の分析では、外れ値の処理、相関の可視化などを課題とする。

商品数が 7776 品あり、全商品の分析には手間がかかるため、3 年間の販売総数が 1000 以上の 433 品に絞って分析を行った。いくつかの商品について、販売数の間に強い相関がみられた。特に、商品番号が (130134, 131510), (130134, 131511), (131510, 131511), (131796, 131797), (132511, 132513), (132511, 136382), (133473, 133474), (133947, 133948), (134526, 134532), (134528, 134530), (134528, 134532), (134530, 134532), (135478, 145500), (135498, 135499), (135880, 136199), (136340, 136382), (136933, 136945) の 17 ペアについては、相関が 0.9 を超えた。しかし商品 (130134, 131510) のペアのように外れ値の影響によって相関が強くなっているペアもある。

そこで、外れ値の影響が少ないペアを探したところ、商品 (130134, 131510), (133473, 133474)のペアが見つかった。

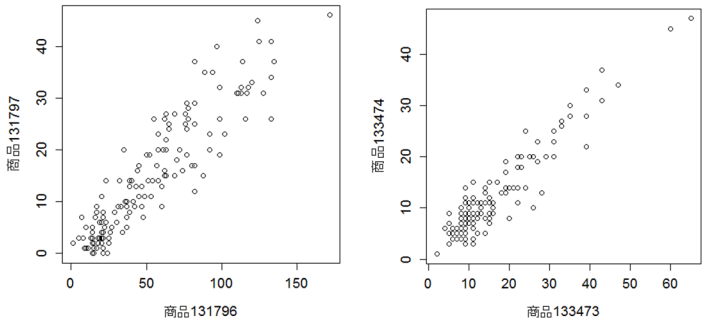

図 4

また、各週の 433 品全品の販売総数の時系列を調べたところ、以下のようになった。

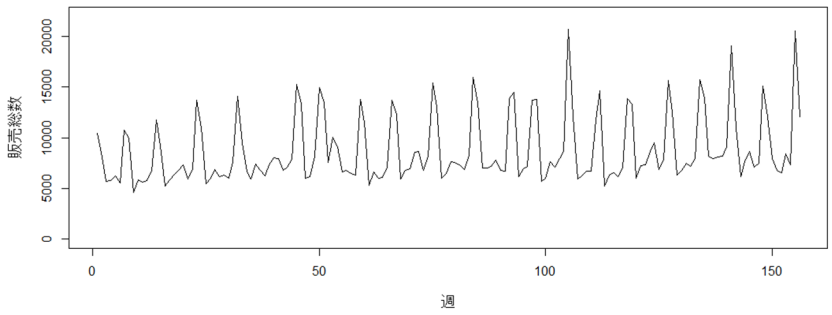

図 5

ところどころにピークが存在するが、このピークは不規則であり、祝日が多い週というわけでもない。これは、通常の週と比較し倍以上の差があり、この店舗がセールを行った週ではないかと考えられる。(しかし、ピークの週だけが売れているのではなく、ピークの週を含む 2～3 週間で販売数が増加している。毎回長期的なセールを行っていることが予想される。)

次に、在庫管理について調べてみる。(毎週の仕入れ数) − (販売数) の累積和を計算し、これを初週の在庫を 0 に基準化したときの在庫とみなす。この在庫について、標準偏差が 200 を超える 9 商品 (130075, 130423, 131592, 131597, 131655, 131927, 132158, 133694, 134164) について調べてみると、いくつかのパターンが見られた。商品 **131592, 131655, 133694, 134164** のように発売開始後に大量に仕入れ、その後在庫が 0 付近に近づいているものは在庫管理がうまくいっている例と考えられる。商品 **132158** については、在庫が急に増加しており、売れ行きが落ちているとみられる。商品 130075 と商品 131597 の在庫は異常な動きをしており、後者は在庫が増え続けており、これは賞味期限の短い食品等で廃棄が行われているものと予想される。一方、前者 (商品 130075) は在庫が減り続けており、この期間の直前に大量の在庫を抱えたことが予想される。

６．まとめ

以上に示したように、本年度ビュー・コミュニケーションズから提供された大規模なデータは、さまざまな観点から分析が可能であり、初年次生のうちから、さまざまにデータを試行錯誤的に分析することができ、教育的価値が非常に高いものである。

滋賀大学データサイエンス学部としては、今後もさらに興味深い演習課題をビュー・コミュニケーションズとの共同研究の形で開発して行きたいと考えている。また、数理及びデータサイエンスに係る教育強化の拠点校としては、このような教育コンテンツを全国展開していくことも重要な課題である。