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

2020 年 3 月 31 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、令和元年 9 月 1 日から令和 2 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 2 年生対象の「時系列解析入門」の講義の演習の形に整備して学生の教育に活用することで、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、食品スーパーにおける 5 年分の酒類の月次販売データが提供された。このデータは滋賀大学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、今年度は時系列解析入門の講義でのレポート課題として用いることにより、学生達は講義で学習する時系列解析の手法がどのように実際のデータ分析に役立つかについて理解を深めることができた。これは滋賀大学データサイエンス学部が掲げる実践的なデータサイエンティスト育成の観点からも非常に有用であった。その点で本共同研究の意義は大きい。

２．滋賀大学データサイエンス学部の育成人材像

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。

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。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。小松秀樹氏には毎年ゲスト講師として講義をお願いしているが、今年度は 2019 年 11 月 29 日に時系列解析入門のゲスト講師として講義をしていただき、実務データの見方や扱い方に関するノウハウを学生に伝えていただいた。

滋賀大学データサイエンス学部は日本初のデータサイエンス学部として注目をあびているが、2017 年 4 月の開設以来この 4 月には 1 期生が 4 年生となり、あと 1 年でいよいよ初のデータサイエンス学部卒業生として社会に巣立っていく。滋賀大学データサイエンス学部の教育の成果が、社会でどのように評価されるかという時期を迎えており、今後も滋賀大学データサイエンス学部は実践的な教育を深化させていく必要がある。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滋賀大学が今後も展開していくべき活動である。

３．滋賀大学データサイエンス教育の展開

滋賀大学ではデータサイエンス教育を大学の戦略的な方針として位置付けており、データサイエンス教育研究拠点の整備を進めている。2019 年 4 月には、日本初のデータサイエンス研究科修士課程を開設した。修士課程の開設は、データサイエンス学部の開設から 2 年後のことであり、まだ学部からの卒業生が出ていない段階での「前倒し設置」であった。学部からの進学性がいないため、入学者は外部からの応募者である。定員 20 名のところを 23 名が修士課程 1 期生として入学したが、そのうち 19 名は企業派遣の社会人であった。このように多くの派遣社会人が入学したことの理由は、企業において社員のリカレント教育としてデータサイエンス教育の重要性が認識されるようになっており、そのような需要に応えた修士課程を設計したためと考えられる。

また 2020 年 4 月には博士後期課程も開設する。博士後期課程は当初の定員は 3 名と少数であり、入学予定者もちょうど 3 名となった。この 3 名はいずれも企業所属であり、博士課程での研究を通じて棟梁レベルのデータサイエンティストとなることを目指している。

このように 2020 年 4 月に、滋賀大学に学部から博士課程まで一貫したデータサインス教育体制が完成するが、これは学部から博士までずっと滋賀大学に在籍しまま博士号を取得する学生を想定しているわけではない。それよりは、学部卒で企業に入りしばらく実務経験を積んだ後に修士課程に戻り、さらにまた経験を積んだ後に棟梁レベルのデータサイエンティストを目指して博士課程に戻るというように、大学と企業を行き来しながらスパイラル型にデータサイエンティストとして成長するモデルを考えている。データサイエンス分野はこのように人的交流においても、大学と企業の連携が有効であると考えられる。

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

本年度にビュー・コミュニケーションズより提供されたデータは、関東圏食品スーパー100 店舗サンプルの販売金額合計値時系列データであり、2014 年 10 月から 2019 年 9 月までの 60 ヶ月分の月次データである。また販売品目は酒類 10 品目であり、具体的には、第３のビール、ビール 350ml、焼酎甲類、清酒パック、国産ワイン、輸入ワイン、国産ウィスキー、輸入ウィスキー、缶酎ハイ 350ml、発泡酒 350ml、であった。データはエクセルのファイルとして、すでに整形された形で提供されており、学生はデータを R や Python などの統計分析パッケージに読み込むことで、ARIMA モデルなどの時系列解析の手法を応用することができた。また、本年度提供されたデータは 5 年分のデータであり、それぞれの品目の売り上げの季節性なども見ることができるため、時系列解析の例題としては非常に実践的なものであった。次図は提供されたエクセルファイルの一部を示したものである。

	第3のビール	ﾋﾞｰﾙ 350ml	焼酎甲類	清酒パック	国産ワイン	輸入ワイン	国産ｳｲｽｷｰ
2014年10月	64,896,265	56,458,480	41,283,034	27,710,083	15,671,871	17,012,474	11,630,853
2014年11月	62,626,576	53,164,431	42,021,722	30,565,834	17,547,875	33,758,505	12,350,601
2014年12月	61,599,319	73,810,979	44,625,092	37,739,137	20,952,974	25,209,042	14,292,028
2015年1月	56,903,102	54,636,556	39,416,436	28,568,680	15,394,759	18,412,583	13,251,262
2015年2月	55,275,135	42,429,178	37,706,181	28,192,179	14,140,309	16,260,936	13,442,114
2015年3月	61,832,614	55,360,190	41,023,819	28,121,201	15,395,572	16,849,718	15,683,896
2015年4月	65,137,971	55,137,113	39,079,248	25,829,663	13,980,098	16,657,483	15,533,525
2015年5月	95,580,312	75,617,063	43,269,560	22,720,433	13,507,734	17,591,824	14,537,782
2015年6月	88,410,801	64,505,572	55,492,424	20,548,710	11,904,098	15,005,699	13,854,597
2015年7月	96,231,198	77,867,879	39,422,324	19,173,130	11,388,621	14,482,564	13,158,490
2015年8月	95,740,071	93,823,473	38,529,894	18,322,581	10,929,671	14,653,894	13,122,023
2015年9月	85,365,554	66,422,417	38,211,951	22,253,526	12,521,920	15,458,816	12,324,527
2015年10月	87,278,610	61,513,467	40,727,491	27,928,000	16,921,779	19,699,321	13,375,270
2015年11月	80,622,982	57,462,902	41,008,635	31,193,441	17,679,276	34,439,141	13,335,337
2015年12月	80,550,355	84,451,103	43,382,753	37,712,380	20,849,978	26,678,527	15,593,571
2016年1月	74,896,265	63,204,148	39,193,684	30,669,160	15,611,565	19,267,511	13,578,948
2016年2月	76,489,294	49,531,205	37,564,149	29,950,650	14,643,939	18,477,289	13,249,479
2016年3月	84,751,597	64,607,404	40,332,548	30,191,077	14,764,217	19,224,918	14,284,018
2016年4月	85,530,588	66,563,713	38,443,524	27,844,328	14,134,252	19,286,489	15,331,135
2016年5月	94,723,466	78,050,551	42,703,326	25,602,536	14,566,449	21,522,904	16,997,167
2016年6月	90,770,895	71,194,568	51,544,702	22,761,122	12,312,624	17,275,704	18,719,842
2016年7月	99,618,096	84,006,041	39,677,949	21,045,965	11,559,434	16,372,270	19,395,526
2016年8月	97,794,657	99,688,086	38,734,847	19,953,496	10,942,939	15,548,213	19,338,026
2016年9月	89,843,497	67,595,129	37,984,195	22,357,293	12,079,329	16,311,333	19,512,245

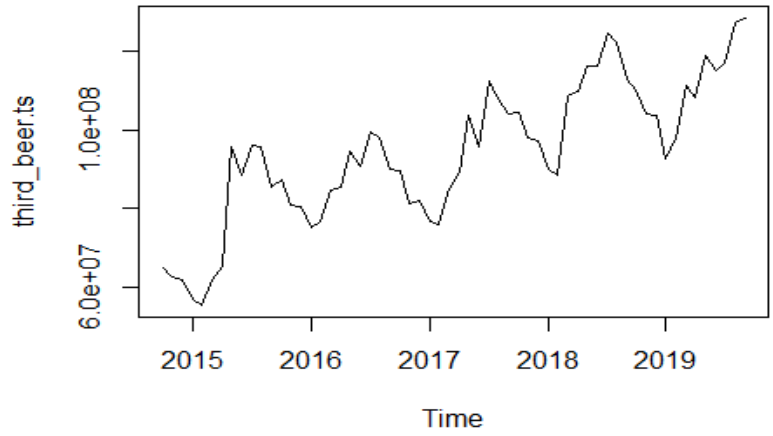
本年度提供データの一部

５．学生によるデータ分析の例

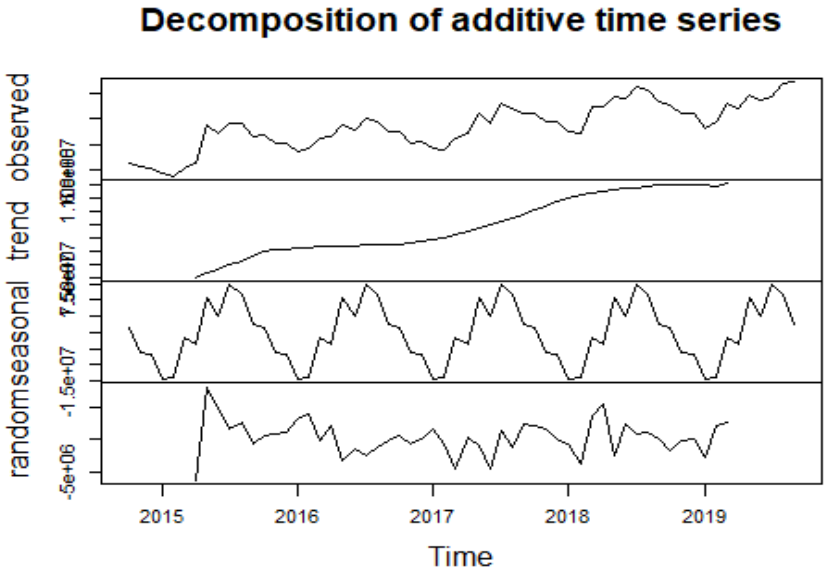
ここでは、レポート課題においてビュー・コミュニケーションズから提供されたデータを熱心に分析してくれた上野議博君のレポートの分析例を紹介する。各学生には 10 品目のうち 3 品目を選んで分析することを求めた。上野君のレポートでは第 3 のビール、ビール 350ml、国産ウィスキーの 3 品目を選んでいるが、以下では第 3 のビールについて紹介する。分析に使われたツールは R である。

```
#install.packages("openxlsx")
library(openxlsx)

## Warning: package 'openxlsx' was built under R version 3.6.2

# データの読み込み
data <- read.xlsx("演習用時系列データ_2019.xlsx")

third_beer <- data$"第３のビール"
beer <- data$"ﾋﾞｰﾙ.350ｍｌ"
whisky <- data$"国産ｳｲｽｷｰ"

third_beer.ts <- ts(third_beer, st=c(2014,10), end=c(2019,9), fr=12)
beer.ts <- ts(beer, st=c(2014,10), end=c(2019,9), fr=12)
whisky.ts <- ts(whisky, st=c(2014,10), end=c(2019,9), fr=12)

# とりあえずplot
plot(third_beer.ts, type="l")
```

第 3 のビールのプロットは以下である。

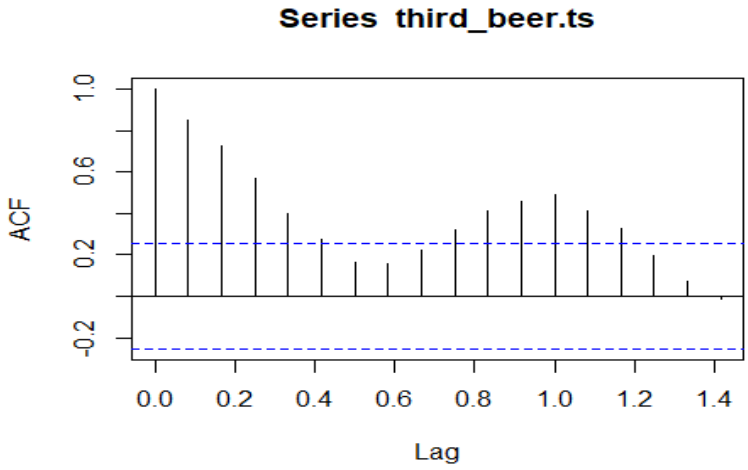


データのトレンド、季節成分、不規則成分への分解は以下のようになっており、第 3 のビールの販売量には、緩やかな上昇のトレンドがあることがわかる。また seasonal の部分より季節性があるように思われる。第 3 のビールは夏によく売れることが予想されるので、これは直感通りのことである。



データの自己相関関数は以下である。

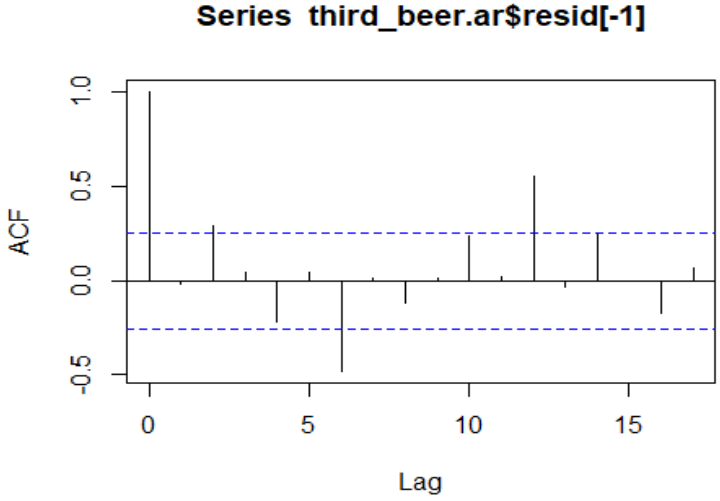
```
# 自己相関関数
acf(third_beer.ts)
```



まずは AR モデルをあてはめてみると、第 3 のビールの場合には AR(1)モデルが選ばれた。ラグ 1 の係数が 0.8515 と推定された。R の ar コマンドでは AR モデルの次数は AIC 基準などによって自動的に良いものが選ばれる。

```
# AR モデルをあてはめ
third_beer.ar <- ar(third_beer.ts)
## Coefficients:
##      1
## 0.8515
##
## Order selected 1  sigma^2 estimated as  9.237e+13
```

このモデルで季節性がまだ残っていることは、残差の自己相関(時図)を見ることで明らかとなる。



そこで、季節性を含む SARIMA モデルを当てはめることが考えられる。実際に SARIMA モデルを当てはめた結果は以下となり、SARIMA$(1,0,0)(1,1,0)_{12}$ モデルが選ばれ

た。式として展開すれば

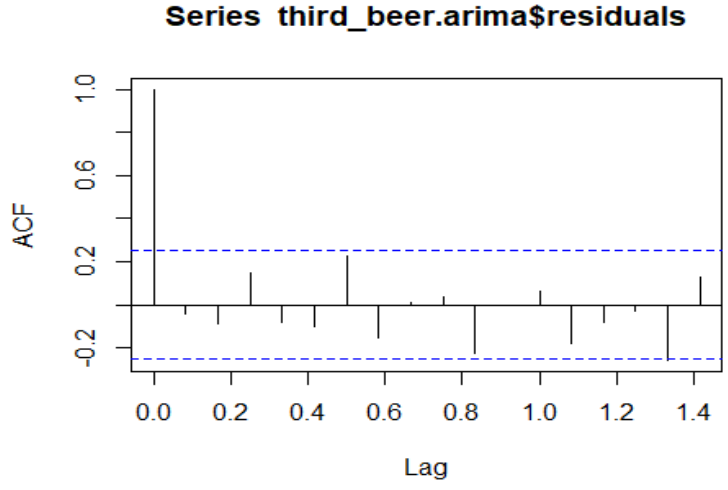
$$x_t = 820504 - 0.34x_{t-25} + 0.56x_{t-24} - 0.27x_{t-13} + 0.44x_{t-12} + 0.61x_{t-1} + w_t$$

である。ただしw_tはホワイトノイズ項である。

```
# SARIMA モデルの推定
third_beer.arima <- auto.arima(third_beer.ts, trace=F, stepwise=F, seasonal=T)
beer.arima <- auto.arima(beer.ts, trace=F, stepwise=F, seasonal=T)
whisky.arima <- auto.arima(whisky.ts, trace=F, stepwise=F, seasonal=T)
summary(third_beer.arima)
```

```
## Series: third_beer.ts
## ARIMA(1,0,0)(1,1,0)[12] with drift
##
## Coefficients:
##          ar1     sar1     drift
##       0.6119  -0.5620  820503.9
## s.e.  0.1351   0.1589   98594.5
##
## sigma^2 estimated as 2.276e+13:  log likelihood=-807.22
## AIC=1622.43   AICc=1623.36   BIC=1629.92
##
## Training set error measures:
##                    ME    RMSE     MAE        MPE     MAPE      MASE
## Training set 34957.13 4131772 2669697 0.04437455 2.676099 0.2731104
##                     ACF1
## Training set -0.04294248
```

残差の自己相関関数を見ると、以下のようになり、ホワイトノイズ性が見られ、分析としては適切なものと思われる。



このように、上野君の分析は、データのプロットから始め、単純なモデルの当てはめを経て適切なSARIMAモデルまで到達しており、模範的な分析例となっている。

６．まとめ

前節でみたように、本年度提供されたデータの分析をレポート課題としたところ、多くの学生が熱心に分析に取り組んだ。その理由の一つとしては、提供されたデータがわかりやすく学生が興味の持ちやすいデータであったことがあげられる。

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。修士課程も開設されており、より進んで手法の応用を含めて修士課程の演習としても活用していきたい。

滋賀大学は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文部科学省より選定されており、このような教育コンテンツを他大学に展開していくことも重要な課題である。