報　告

Practical Data Science - Final Report

DSS-2022A Rakuten Card Group

Wang Boyu　Graduate School of Frontier Sciences
6747418421@edu.k.u-tokyo.ac.jp

Xu Wangjie　Graduate School of Frontier Sciences
wjiexu@g.ecc.u-tokyo.ac.jp

Valentin Fontanger　Information Science and Technology
valentin.fontanger@gmail.com

Shengjie Fang　Graduate School of Frontier Sciences
jiesheng811@g.ecc.u-tokyo.ac.jp

Boyu Wang　Graduate School of Engineering
wang-boyu128@g.ecc.u-tokyo.ac.jp

1. Introduction

The rise of cashless payment methods, such as QR payments, e-money payment and transportation IC payment, has the potential to strongly impact the traditional credit card payment business. As consumers increasingly adopt cashless payment options, the demand for credit cards may decline, leading to lower revenue for credit card companies. On the other hand, credit card companies can adapt to the shift towards cashless payments by offering their own cashless payment solutions, integrating with existing cashless payment systems, or partnering with fintech companies to offer new services. The impact of cashless payments on the credit card industry will depend on how well credit card companies are able to respond to the changing market.

In this project, we aimed to show how cashless payment can affect Rakuten credit card business and give advice on how to modify credit card business. We firstly described the dependence between credit card payment and cashless payment and then analyzed regional specific effects. We also used ARIMA, LSTM and Prophet to predict the future trend of credit card business.

2. About the data

The data was provided by Rakuten Card company and consisted of the sum spent amount and unique user count of different payment methods, including credit card payment and cashless payment. Cashless payment includes three different types: QR code payment, e-money payment, and transportation card payment. The time span covered is from 2019-09-01 to 2022-09-01.

3. Exploratory Data Analysis (EDA)

3·1 Distribution Visualization

Before performing in-depth analysis on the data, it is important to understand the basic characteristics and statistical features of the data. To show the distribution of sum spent amount and unique user count of four different payment methods, a hist plot was used to visualize it intuitively. The result can be seen in **Figure 1** below. In this plot, we can see firstly the sum spent amount and unique user count of credit card is much more than those of cashless payments. Secondly, for cashless payments, e-money payment sum amount is the largest. The distributions of QR payment and transportation IC payment unique user counts are wider than e-money payment.

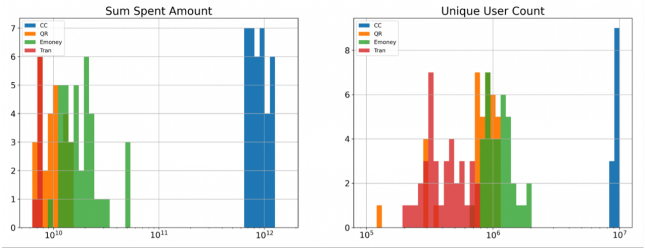

図 1 Distribution of sum spent amount and unique user count of four payment methods

	Credit Card	QR code	E-money	Transportation IC
Totoal Spent	東京都, 愛知県, 大阪府, 埼玉県, 千葉県, 兵庫県, 神奈川県, 北海道, 茨城県, 福岡県	東京都, 愛知県, 大阪府, 埼玉県, 千葉県, 神奈川県, 兵庫県, 福岡県, 岐阜県, 静岡県	東京都, 愛知県, 大阪府, 千葉県, 埼玉県, 兵庫県, 神奈川県, 北海道, 沖縄県, 茨城県	東京都, 千葉県, 埼玉県, 神奈川県, 大阪府, 愛知県, 茨城県, 兵庫県, 栃木県, 群馬県
Average Spent	東京都, 愛知県, 千葉県, 神奈川県, 埼玉県, 大阪府, 茨城県, 沖縄県, 群馬県, 石川県	福井県, 和歌山県, 神奈川県, 香川県, 石川県, 岐阜県, 広島県, 岡山県, 京都府, 富山県	沖縄県, 高知県, 富山県, 秋田県, 青森県, 東京都, 北海道, 神奈川県, 和歌山県, 千葉県	東京都, 千葉県, 神奈川県, 埼玉県, 茨城県, 栃木県, 群馬県, 青森県, 山梨県, 奈良県

表 1 Top 10 prefectures of payment sum spent and average spent

3·2 Regional Distribution Visualization

In order to explore the different payment habits of customers from different regions of Japan, we visualized customer payment data of 47 prefectures including the sum spent amount and average spent amount of four different payment methods. The average spent amount was defined as $\frac{sum\ spent\ amount}{unique\ user\ count}$. Results can be checked in **Figure 2** and **Figure 3**. From Figure 2 and Figure 3, we can see that the distribution of sum spent amount of four different payments among 47 prefectures are similar, however the average spent amounts show some difference. To clarify the difference more clearly, we summarized top 10 sum spent amount prefectures and top 10 average spent amount prefectures in **Table 1**. From Table 1, we can clearly see the difference of payment habits of customers from different regions which suggests the need for a regional business plan to promote credit card business.

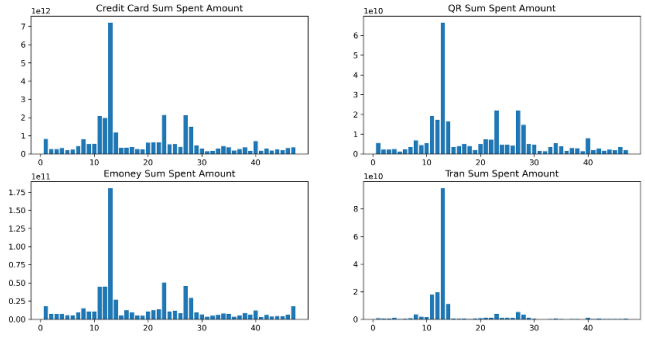

図 2 Distribution of payment sum spent amount among prefectures

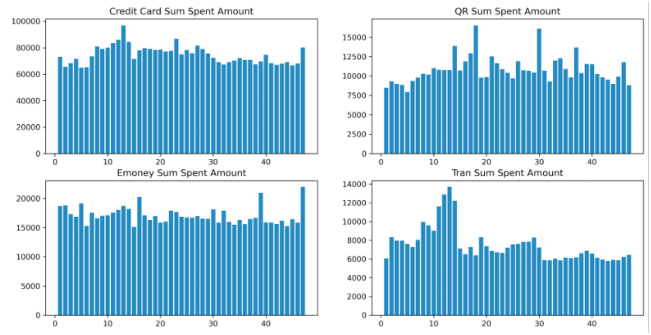

図 3 Distribution of payment average spent amount among prefectures

3·3 Time Series Analysis

In order to build credit card business growth forecast models, we first analyze the whole country scale time series data of four payment methods. Visualization results can be seen in **Figure 4** and **Figure 5**. From these two plots, we can see that overall, the sum spent amount and unique user count of four payment methods continued growing from September 2019 to September 2022. We then

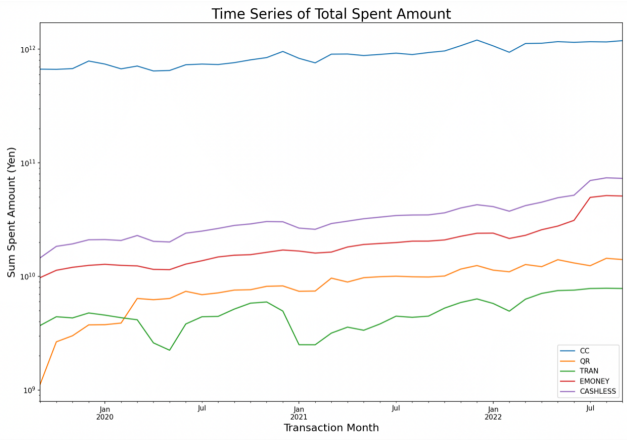

図 4 Time series of payment method sum spent amount

concentrated on credit card sum spent amount. We found there were strong seasonal features of the growth from **Figure 6**, so we than did the seasonal decomposition. The multiplicative model represents the time series as:

$$Y_t = Trend_t \times Seasonal_t \times Irregular_t \qquad (1)$$

We used LOESS smoothing to decompose time series into trend items, seasonal items and random items. R package was utilized to do this task. Results will be showed in later part.

Also we plot the sum spent amount time series of all prefectures for all payment methods including credit card shopping, credit card cashing, QR code, e-money and transportation card payment (**Figure 6**). We can see the trend

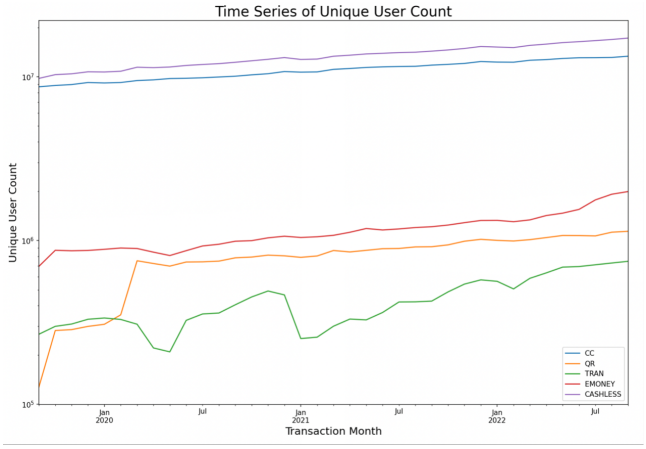

図 5 Time series of payment method unique user count

of each sum spent amount are similar for different prefectures besides QR code payment. Especially the red line in **Figure 6 (c)** (Kanagawa prefecture) shows a peculiar behavior which means some unique events happened there. We would like to discuss this anomalous move later and to find out the relationship of credit card shopping amount and each payment methods.

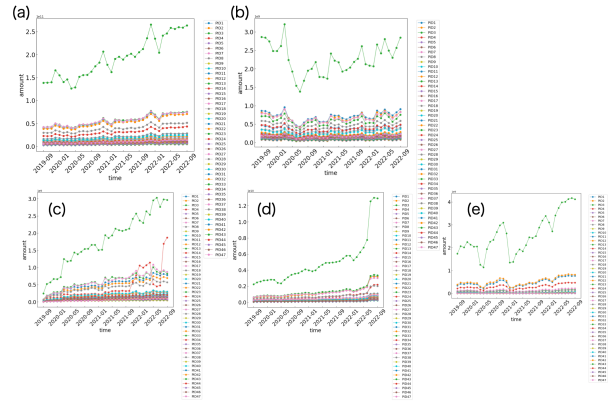

図 6 Sum spent amount time series of all prefectures for (a)credit card shopping, (b)credit card cashing, (c)QR code, (d)e-money, (e)transportation card payment.

4. Correlation Analysis

One of the most important tasks of this project is to describe the dependence of cashless payment and credit card payment. For better interpretability and the complexity of the model, we chose **linear regression** to show the linear correlation. Linear regression is a statistical method used to model the relationship between two continuous variables, the dependent variable (y) and the independent variable (x). It aims to find the linear equation that best fits the observed data points and can be used to make predictions about the dependent variable based on the independent variable. The equation takes the form of $y = mx + b$, where m is the slope of the line and b is the y-intercept. The method assumes a linear relationship between the variables and uses least-squares optimization to find the best line of fit. T-test was used to check the significance of the m and b parameters.

4·1 Single Predictor Regression

We set single predictor regressions. First, we used the credit card sum spent amount as a response and cashless payments sum spent amount as predictors respectively. The results can be seen in **Figure 7**, **Figure 8** and **Figure 9**.

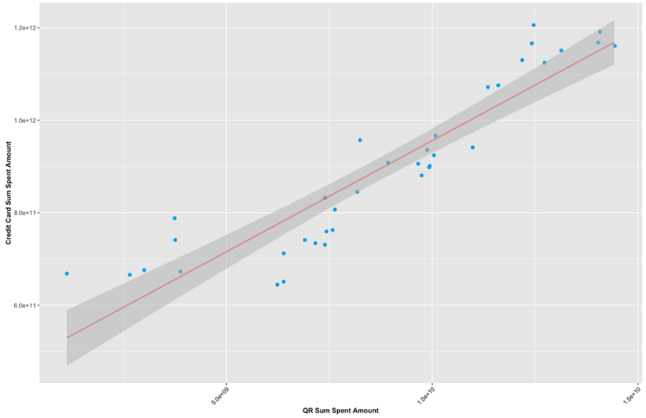

図 7 Regression of sum spent amount of credit card with the predictor of QR code

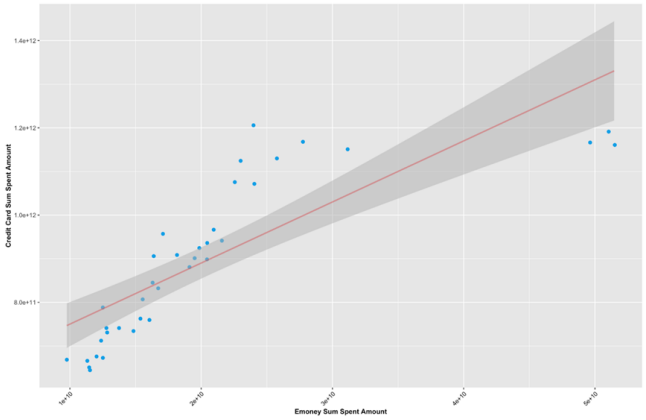

図 8 Regression of sum spent amount of credit card with the predictor of e-money

We also calculated correlation factors which can be checked in **Table 2** to verify our regression results. Among three cashless payment methods, QR payment sum spent amount has the strongest correlation with credit card payment sum spent amount. The correlation between e-money and transportation IC sum spent amount and credit card sum spent

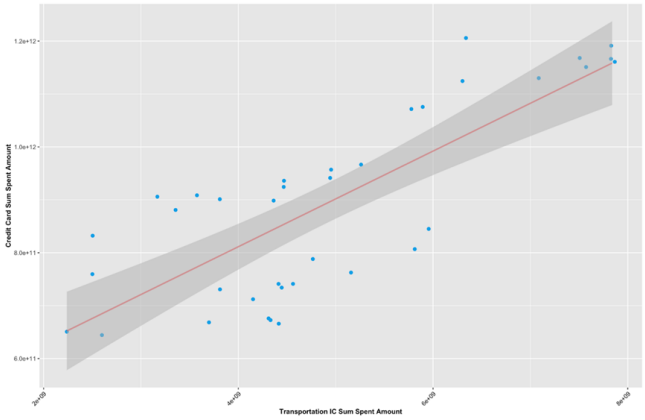

図 9 Regression of sum spent amount of credit card with the predictor of IC card

amount are also significant. The result of single predictor linear regression of sum spent amount suggests that the increase of cashless payment is positively correlated with the increase of credit card payment amount. Then, we did similar analysis based on unique user count. The results can be seen in **Figure 10**, **Figure 11** and **Figure 12**.

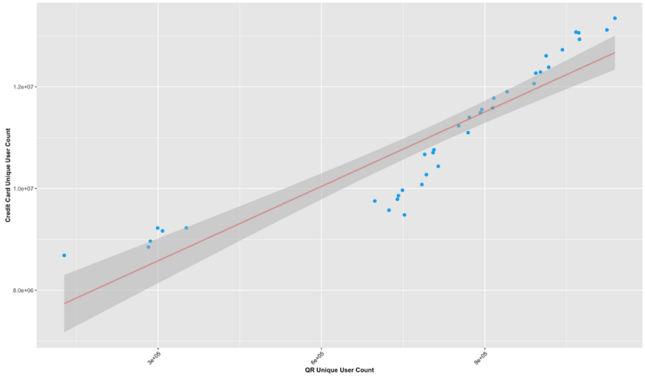

図 10 Regression of unique user count of credit card with the predictor of QR code

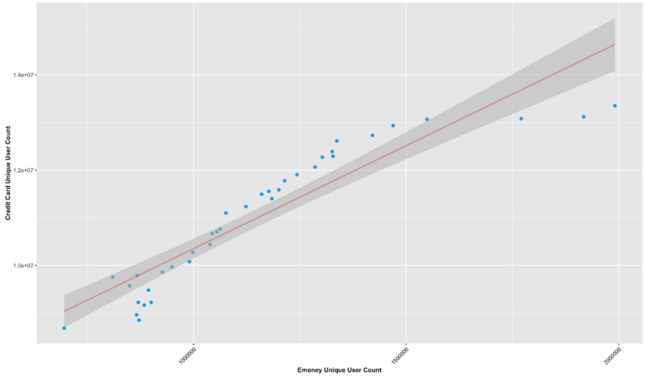

図 11 Regression of unique user count of credit card with the predictor of e-money

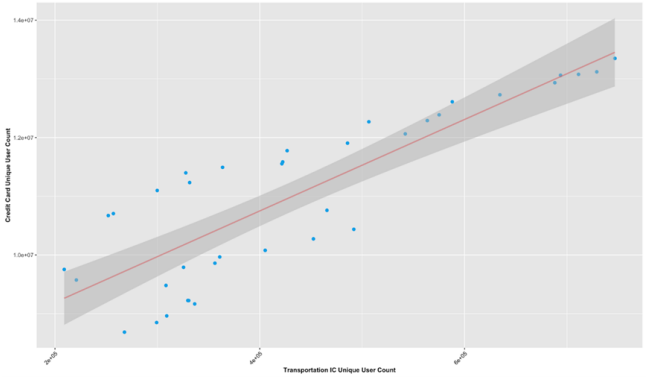

図 12 Regression of unique user count of credit card with the predictor of IC card

	QR Code	E-money	Transportation IC
Total Amount	0.916	0.811	0.792
Unique User Count	0.910	0.922	0.848

表 2 Correlation factors between credit card and cashless payments

Correlation factors in **Table 2** can help to verify our regression results. Among three cashless payment methods, E-money payment unique user count has the strongest correlation with credit card payment unique user count. The correlation between QR and transportation IC unique user count and credit card unique user count are also significant. The result of single predictor linear regression of unique user count suggests that the increase of cashless payment users is positively correlated with the increase of credit card payment users.

4·2 Multiple Predictors Regression

Based on the results of single predictor linear regression, we would like to fit multiple predictor regression models to show how multiple cashless payment methods are correlated with credit card payment. We used subset selection method to find the best model. For sum spent amount, the best linear model has adjusted R2 = 0.8971 with predictors QR sum spent amount and transportation IC sum spent amount (**Table 3**). For unique user count, the best linear model has adjusted R2 = 0.9323 with predictors QR unique user count and e-money unique user count (**Table 4**).

	Estimate	Std. Error	t value	Pr(|t|)
(Intercept)	3.876e+11	3.213e+10	12.063	7.78e-14
qr_sum_spent_amount	3.652e+01	3.719e+00	9.818	1.86e-11
tran_sum_spent_amount	3.829e+01	8.086e+00	4.736	3.77e-05

表 3 Regression results between QR code and transportation IC

	Estimate	Std. Error	t value	Pr(|t|)
(Intercept)	6.004e+06	2.421e+05	24.794	2e-16
qr_unique_user_count	2.556e+00	3.778e-01	6.764	8.92e-08
emoney_unique_user_count	2.574e+00	3.289e-01	7.826	4.15e-09

表 4 Regression results between QR code and emoney

5. Regional Specific Analysis

We have found some special features of credit card sum spent amount and average spent amount in previous analysis. Developing a regional specific business strategy is good for credit card business growth because it can better meet regional needs and the needs of specific customer segments. It can offer more attractive and valuable credit card products and build partnerships with regional merchants to offer customized offers and rewards. At the same time, through the understanding and analysis of the local market, formulating regional strategies can also promote and market credit card products more effectively, thereby attracting more customers to join. We would like to figure out how cashless payment can affect credit card business in different regions and give some business advice based on our findings. In this part, we pay attention to shopping data.

5·1 Regional Specific Correlation Analysis

Firstly, we visualized QR, e-money and credit card shopping payment amount of 47 prefectures with a 3D plot **Figure 13**. In this plot, we can see those prefectures with high credit card shopping amount always have high QR and e-money shopping amount. To show the correlation between credit card shopping payment and cashless shopping payment of all prefectures, we calculated the correlation factors respectively and visualize them in **Figure 14**. In this figure, we can see that the correlation between credit card shopping payment and QR shopping payment and correlation between credit card shopping payment and e-money shopping payment are stable among different prefectures. However, the correlation between credit card shopping payment and transportation IC shopping payment shows strong regional specificity. For some of the prefectures, the correlation is over 0.8 which suggests a strong positive correlation. For some others, the correlation is close to or even less than 0 which suggests there is no correlation.

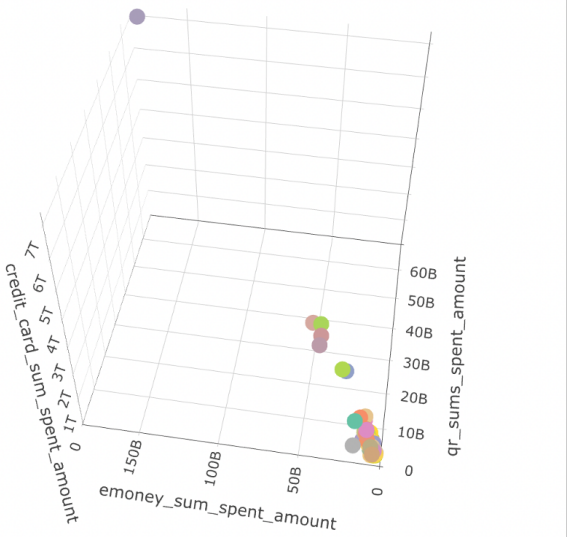

図 13 Distribution of shopping payment spent amount of credit card, QR and e-money sum spent amount of 47 prefectures

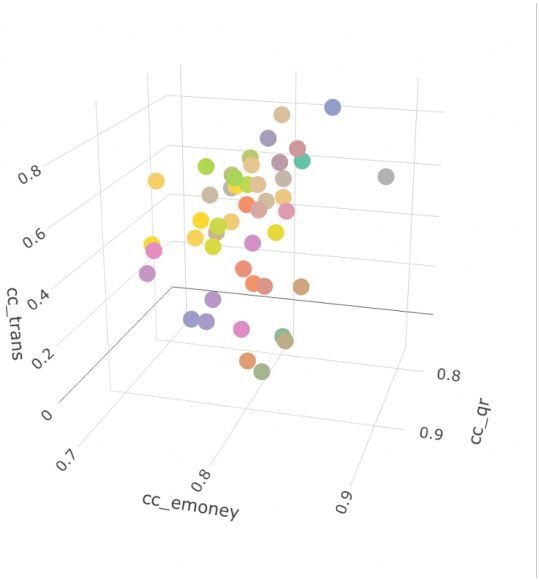

図 14 Distribution of correlation factors between of credit card shopping payment and cashless payment amount of 47 prefectures

5·2 Regional Association Study

To clarify how much the cashless payment can affect credit card payment for all prefectures, we performed regional association study. In this study, we set linear regression with cashless payment shopping amount as predictors and credit card shopping amount as response. We used the effect size of cashless payment to measure the effect. The results are shown in **Figure 15**. From these plots, we can find out that credit card shopping spent amount of those prefectures with lower credit card payment were affected by e-money and transportation IC shopping spent amount more than those prefectures with higher credit card payment. In addition, credit card shopping spent amount of Kyushu region were affected by e-money shopping spent amount more than other areas. Credit card shopping spent

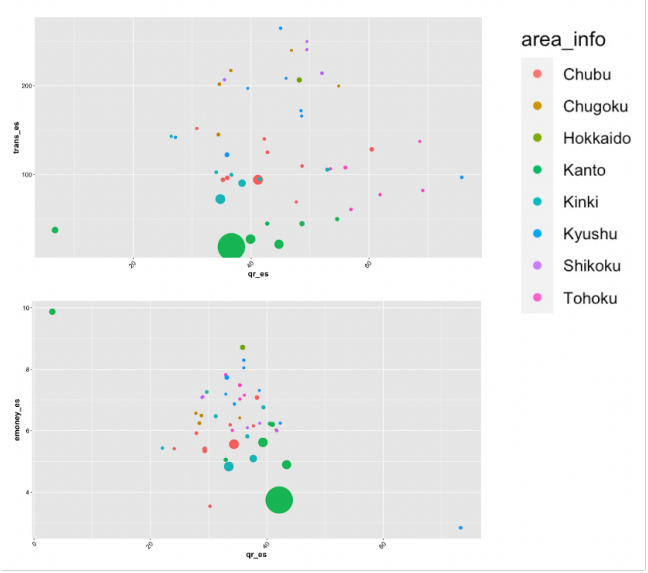

図 15 Bubble plot of regional specific regression result. Size of point represents shopping payment of credit card of the prefecture

amount of Shikoku and Hokkaido were affected by transportation IC shopping spent amount more than other areas. Credit card shopping spent amount of Kanto area was affected by transportation IC shopping spent amount less than other areas. We did similar work based on unique user count of all payment methods. The results are shown in **Figure 16**. Payment unique user count also showed some regional specific features. For example, for Kanto

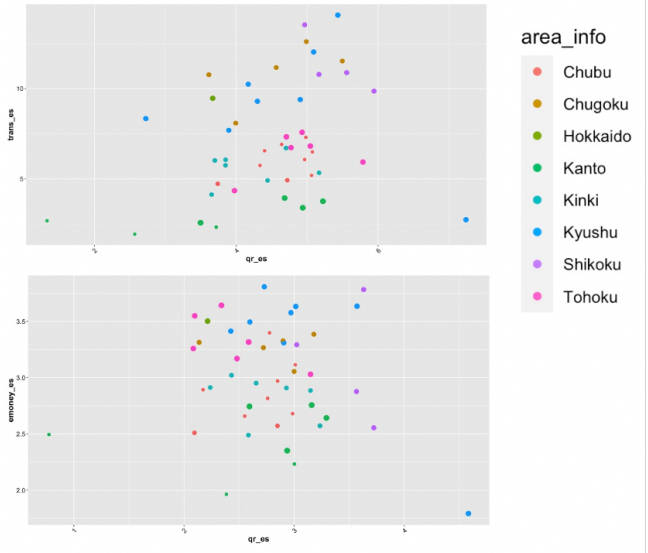

図 16 Bubble plot of regional specific regression of unique user count results. Size of point represents unique user count of credit card of the prefecture

area, credit card unique user count was less affected by transportation IC user count than others. For Kyushu area, credit card unique user count was more affected by e-money user count than others.

6. Time Series Analysis & Prediction

6·1 Time Series Analysis

The time series analysis depicts the trend of the sum spent amount of credit card over time.

The overall trend of the spent amount appears to be an upward trajectory(**Figure 17**), suggesting that it is increasing with time, and the slope, which reflects an increasing rate, is increasing, indicating that the development of the Rakuten credit card business is growing steadily. In

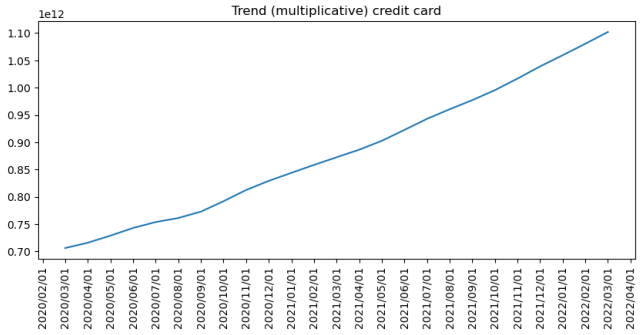

図 17 Trend of Sum Spent Amount of Credit Card

addition to the overall trend, the seasonality figure also reveals seasonal characteristics of the data. The data appears to have cyclical fluctuations, where changes in each year have a similar pattern. These fluctuations could be related to seasonal factors such as changes in demand, weather patterns, consuming enthusiasm, and other potential factors(**Figure 18**, **Figure 19**).

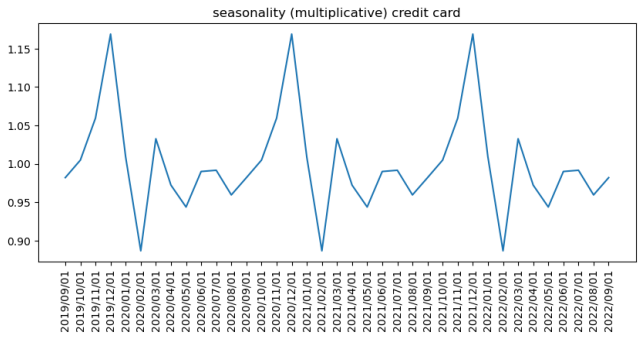

図 18 Seasonality of Sum Spent Amount of Credit Card

It appears that in December, the spent amount on the credit card will come to a peak, and correspondingly, the spent amount will suffer a rapid decline in January and February and reach the lowest point of the spent amount in February. We can observe that the consuming enthusiasm starts raising significantly around September, which

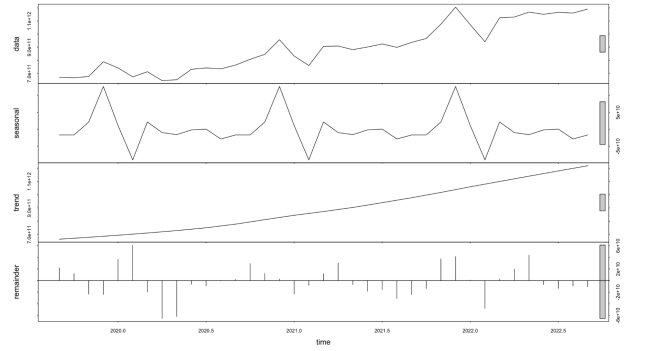

図 19 Time Series Analysis of Sum Spent Amount of Credit Card

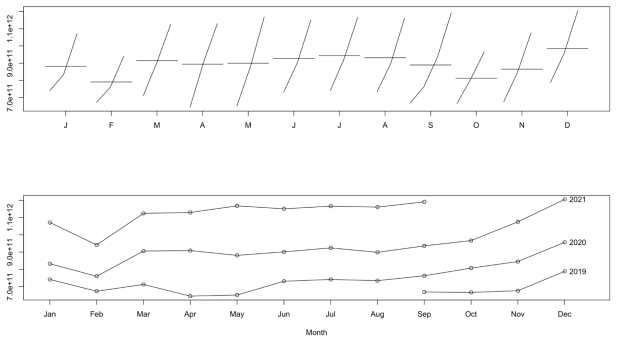

図 20 Detailed Seasonality of Sum Spent Amount of Credit Card

is also the start of autumn. What's more, in the spring and summer, the volatility of spent amount decreases and enters the stable periods until the coming of autumn.

The remainder part also highlights certain times, which appear to be outliers compared to the rest of the data. These times could be associated with holidays, special events, or other unique circumstances that deviate from the typical seasonal trends.

Based on the above time series analysis conclusions, several personal opinions about making full use of time tendency are listed as follows:

(1) Anticipate increased demand: Preparing for a significant increase in demand for credit card services in September and being ready to handle the increased volume of transactions would be helpful to the preparation for peak season. This could include increasing the capacity of their systems and having contingency plans in place to handle any unexpected issues.

(2) Offer seasonal promotions: The credit card company could offer special campaigns and incentives for cardholders during the peak season to encourage customers to use their credit cards more frequently.

(3) Increase marketing efforts: To reach potential customers and encourage them to use their credit cards, the company could increase its marketing efforts during this time. This could include targeted advertisements, email campaigns, or social media promotions to drive awareness and usage.

(4) Prepare for the post-peak decline: After the peak in December, the credit card company should anticipate a significant decline in demand and prepare accordingly. This could include cutting back on marketing efforts and adjusting their systems to better handle the reduced volume of transactions.

More detailed seasonality of the sum spent amount of credit cards states the information that:

(1) The spent amount on the credit card in one specific month of each year increases with the years going by.

(2) The moving patterns in autumn and winter are clear, while the tend from Mar to Jun differs every year, which also makes the result of predicting in these months less accurate. The range in these months also indicates that the spent amount in the spring and summer is more dependent on the situation of that specific year.

6·2 ARIMA Model

AutoRegressive Integrated Moving Average (ARIMA) is a popular model for time series prediction because it can effectively capture the underlying structure of the data. The "auto-regressive" aspect of the model captures the dependency of the current value on past values, allowing it to account for trends and patterns in the time series. The "integrated" aspect of the model accounts for any non-stationarity in the data, such as a trend or seasonality, by transforming the data into stationary form. The "moving average" aspect of the model captures the residual component of the series and allows it to model any short-term fluctuations in the data. These three components work together to provide a comprehensive and flexible model that can be adapted to a wide range of time series data. We used time series data from September 2019 to June 2022 as training data, July 2022 to September 2022 as test data. The prediction result was shown in **Figure 21**. Regularized MSE for this prediction is about 6.04×10^{-5}. We also did the similar analysis based on credit card average spent amount. The prediction result was shown in **Figure 22**. Regularized MSE for this prediction is about 7.39×10^{-5}.

6·3 Attention + LSTM

§ 1 Introduction

LSTM(Long Short-Term Memory network) [1]) is a type of recurrent neural network (RNN) that is designed to han-

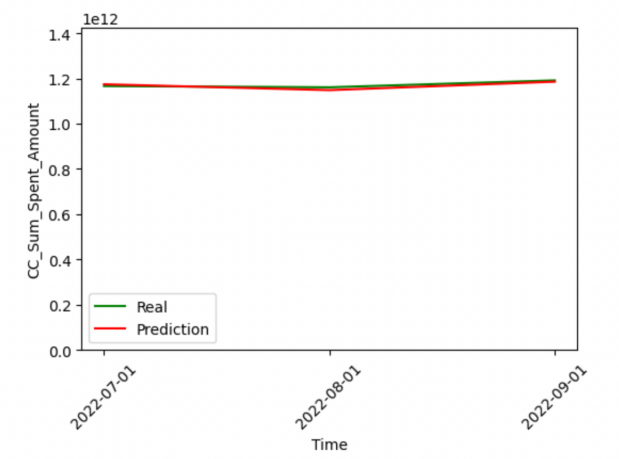

図 21 Prediction of credit card sum spent amount from 2022/7 to 2022/9

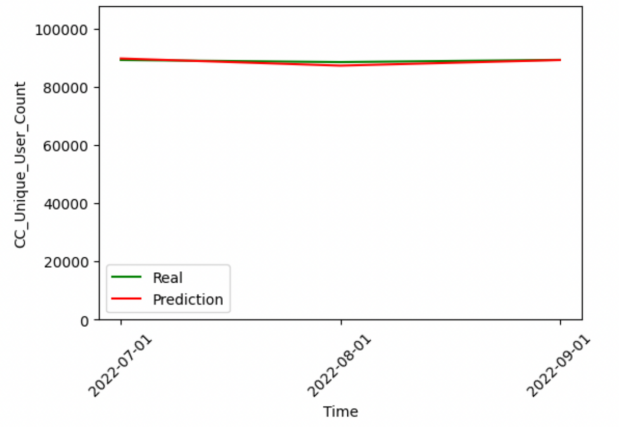

図 22 Prediction of credit card average spent amount from 2022/7 to 2022/9

dle the problem of vanishing gradients in traditional RNNs.

The basic structure of an LSTM network is composed of memory cells, input gates, forget gates, and output gates. The memory cells are responsible for maintaining information over a long period of time, while the input, forget, and output gates control the flow of information into and out of the memory cells.

The basic equations that define the LSTM network are as 2:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\ c'_t &= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c \\ c_t &= f_t \odot c_{t-1} + i_t \odot c'_t \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{2}$$

where x_t is the input vector at time step t, h_t is the hidden state, i_t is the input gate, f_t is the forget gate, c_t is the memory cell state, c'_t is the candidate memory cell state, o_t is the output gate, and W and b are the weight matrices and bias vectors, respectively.

Attention-LSTM Network is an extension of the LSTM network that incorporates an attention mechanism [2]. The attention mechanism allows the network to selectively focus on certain parts of the input sequence, which can improve the performance of the network. Moreover, Attention-LSTM can handle sequential dependencies in a more sophisticated manner than LSTM, as it can selectively focus on different parts of the input sequence based on their relevance.

The basic structure of the Attention-LSTM network is similar to that of the LSTM network, but with an additional attention layer that computes the attention weights for each time step in the input sequence.

The basic equations that define the Attention-LSTM network are as follows:

$$\begin{aligned} h_t &= \text{LSTM}(x_t, h_{t-1}) \\ a_t &= \text{softmax}(W_a h_t + b_a) \\ z &= \textstyle\sum_{t=1}^{T} a_t h_t \end{aligned} \tag{3}$$

where h_t is the hidden state of the LSTM network at the time step t, a_t is the attention weight for time step t, W_a and b_a are the weight matrix and bias vector of the attention layer, and z is the weighted sum.

§ 2 Analytic Results

In our Rakuten data, we have a small set of data but abundant variables. In this case, it is necessary to execute appropriate modifications on the model.

When running deep learning models in a relatively small set of data, it may occur that a model learns the training data too well and starts to memorize it, leading to poor performance on unseen data.

Therefore, we add the dropout function to randomly drop out neurons during each iteration of the training process, thus avoiding the overfit of model.

$$y = x * \text{Bernoulli}(p) \tag{4}$$

where x is the input to the neuron, p is the dropout probability, and Bernoulli(p) is a Bernoulli distribution with probability p.

In our case, we set $p = 0.1$.

Moreover, regularization on the loss function is applied with L2 penalty.

$$\mathcal{L} = \mathcal{L}_{\text{data}} + \lambda \sum_{i=1}^{n} w_i^2 \tag{5}$$

where $\mathcal{L}$ is the total loss, $\mathcal{L}_{\text{data}}$ is the loss from the data, n is the number of parameters, w_i is the i-th parameter,

and λ is the regularization parameter. In this case, we set $\lambda = 0.001$.

In this prediction task, the input parameters are the past 12 months of data in both sum_spent_amount and unique_user_co of credit card, emoney, transport card and qr code. The outputs of the prediction model are sum_spent_amount and unique_user_counts of credit card.

The whole length after this processing model would be 25, we set training size as 22, to predict the sum spent amount and unique users count of credit card in final 3 months.

In this prediction task, the mechanism to predict the future data is as follows:

(1) Using the current data to predict the data of next month.

(2) Shifting data for one step forward, and add the predicted data to the input parameters of next month's prediction.

(3) Predict the data of next month.

This method helps us make full use of existing data, to get the best performance of prediction.

The results of the prediction are as follows.

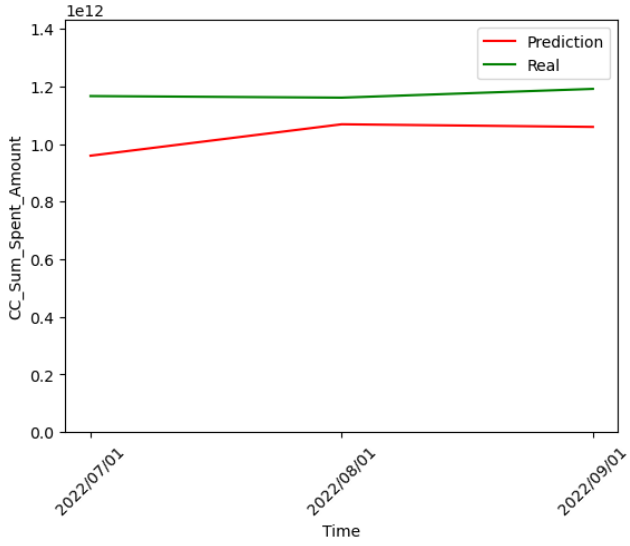

図 23 Prediction of Sum Spent Amount of Credit Card

The results of prediction are measured by MSE(Max Squared Error) after normalization. And the MSE score is 0.073 for the sum spent amount of credit card.

Based on the above results, we also calculate the average spent amount of credit card, whose MSE is 0.003, which is much smaller than the MSE of both sum spent amount and unique user counts. This indicates that the results from Attention-LSTM model have a better prediction of the tendency than the actual value.

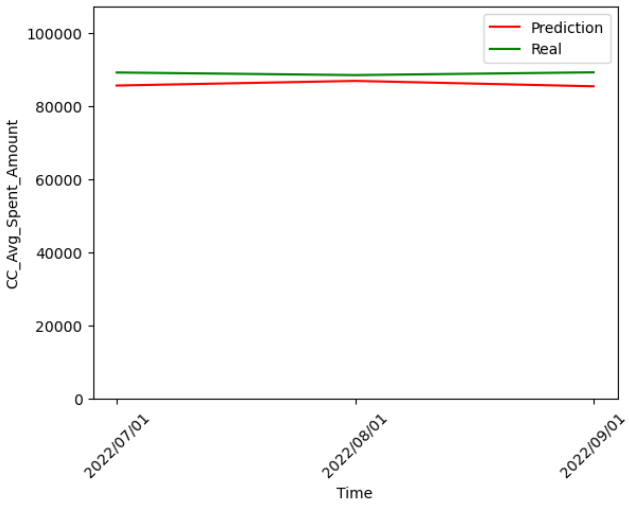

図 24 Prediction of Average Spent Amount of Credit Card

6·4 Prophet

In the task of predicting the total sum spent amount of credit cards for each prefecture in Japan, we chose to use Facebook Prophet due to its ability to handle seasonal trends in the data. Despite the challenge of having a small amount of data, the high seasonality in the data allowed us to observe meaningful results. Prophet's ability to model non-linear trends and seasonality made it a suitable choice for this task. Overall, despite the challenge of having limited data, the use of Prophet provided valuable insights into the spending patterns of credit card users in Japan. The strength of Prophet is the results it can achieve on messy data. It is robust to outliers, missing data, and dramatic changes in your time series, which is likely to happen on short time series.

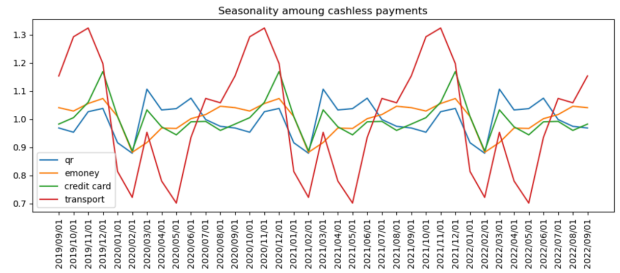

図 25 Seasonality plot for each payment type

For each prefecture, we fitted a Prophet model using the 25% last data point of the time serie as test data, and the first 75% first data point of the time serie as Training data. We repeat this process 3 times, with 3 different set of input features. Due to the automatic nature of Facebook Prophet, we found out that, depending on the given features, the quality of the prediction could be dras-

tically different for a specific prefecture. Indeed, we believe that further hyper parameters tuning would have lead to better results when adding more features to the model. Prophet finds automatically the 'best' hyper parameters, which might not be optimal in certain cases.

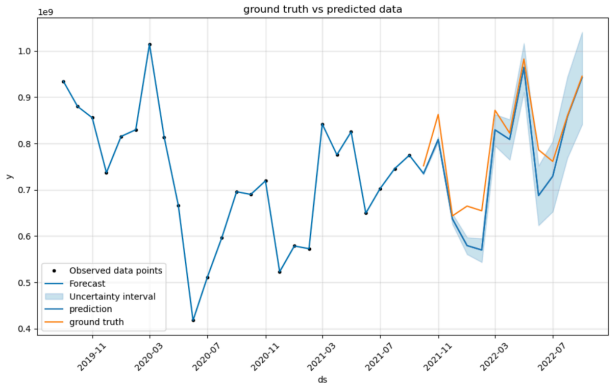

図 26 Prediction on prefecture 1 with sum_unique_user as input

We can see that our prediction generally follows the trend of the ground truth data. The uncertainty interval generally contains the ground truth data. We achieve an RMSE of 50700987. When adding the unique_user_count_qr feature, our RMSE decreases to 46769018.

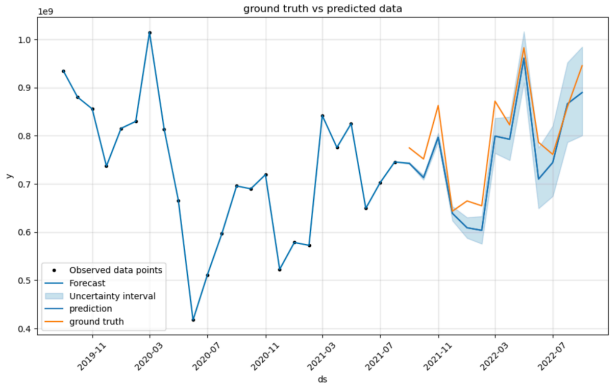

図 27 Prediction on prefecture 1 with sum_unique_user and unique_user_count_qr as input

As we can see, adding unique_user_count_emoney prevented the model to fit correctly. We believe that hyper parameter tuning would allow an increase in performances.

6·5 Comparison

When the dataset is small, ARIMA performs better than Attention-LSTM for several reasons:

(1) ARIMA assumes that the time series data is stationary, meaning that its statistical properties, such as mean and variance, do not change over time. This makes it easier to model the data and obtain accurate predictions. In contrast, Attention-LSTM is more suitable for non-stationary time series data, as it can capture complex relationships between the inputs and outputs.

(2) ARIMA models the relationships between the data points directly, making it easier to understand the relationships between the inputs and outputs. Attention-LSTM, on the other hand, models the relationships between the inputs and outputs through a complex network of neurons, making it more difficult to interpret the relationships between the inputs and outputs.

(3) In a small dataset, there is a high risk of overfitting in Attention-LSTM, as the model has many parameters to be trained on limited data. ARIMA, on the other hand, is less prone to overfitting, as it has fewer parameters to be trained.

In conclusion, when the dataset is small, ARIMA may perform better than Attention-LSTM due to its simplicity, efficiency, and ability to handle stationary data.

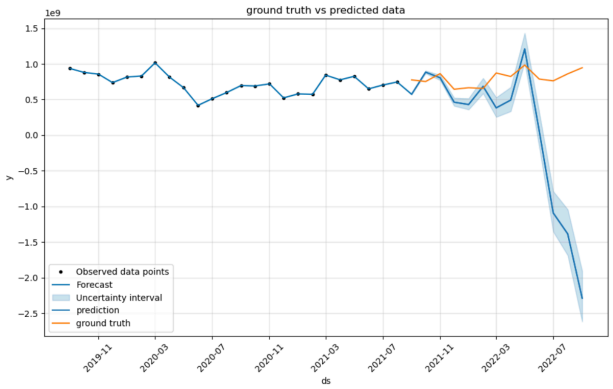

図 28 Prediction on prefecture 1 with sum_unique_user, unique_user_count_qr and unique_user_count_emoney as input

7. Causal Analysis

As mentioned in the Exploratory Data Analysis section, the sum spent amount time series of Kanagawa prefecture for QR code payment showed anomalous increase in the period from 2021-09 (**Figure 29**). To understand this behaviour, we searched the campaign which have held in Kanagawa prefecture and found the event called "Kanagawa pay". This event was a cash-back program for QR code payment users in limited area and the period was 2021-10 to 2022-04 and the second campaign started from 2022-07. Since these periods have matched the increase of the red line in **Figure 29**, we concluded the anomalous increase attributed to the kanagawa pay campaign. We are

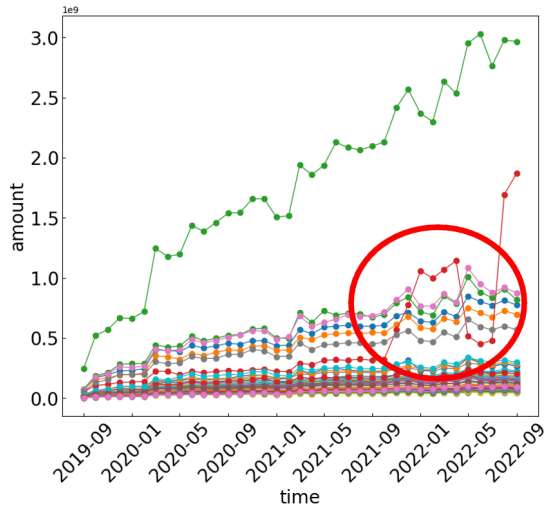

図 29 The time series of QR sum spend amount for different prefectures.

now interested in whether the increase of QR code sum spent amount affect the credit card sum spent amount. In this section, we would like to conduct a causal analysis toward credit card sum spent amount and various payment methods. Also, additional campaign data sets were collected.

Beforehand, the correlation of the credit card sum spent amount and payment methods are shown in **Figure 30**. We can see the correlation of the cashing credit card sum

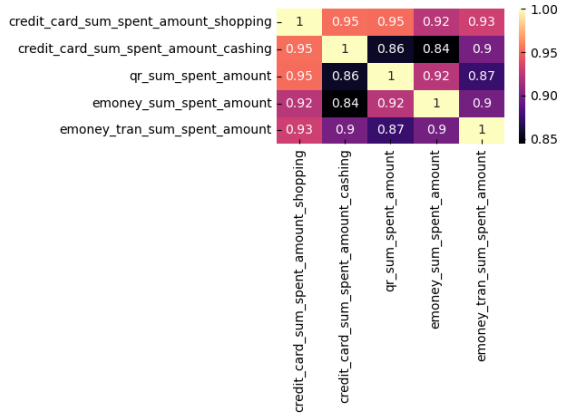

図 30 The correlation of credit card sum spent amount and payment various methods.

spent amount and QR coed sum spent amount exhibited the highest number. At this point, this implies the campaign for the QR code payment has a positive impact for the credit card sum spent amount.

Also, for the data set shown in **Figure30**, we conducted the LightGBM model to find out feature importance. As a result, e-money sum spent amount have shown the highest feature importance which was different from the correlation analysis.

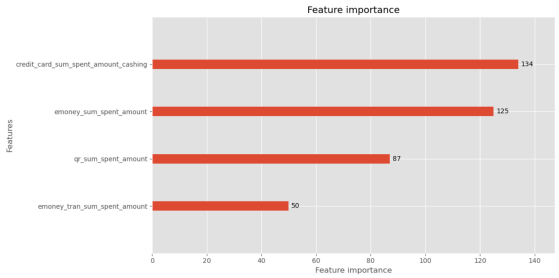

図 31 Feature importance by LightGBM model.

7·1 Data used in Causal Analysis

As there is no readily available dataset to use for analysis, a simple dataset has been constructed based on information gathered from the Internet. This dataset includes one Rakuten card campaign for new users and six-point campaigns of the Rakuten market which are considered to affect the user amount of Rakuten cards and the consumer behaviours of Rakuten card users respectively. The data is sourced from the following two websites [3] [4]. This dataset consists of the number of days when there is a campaign in each month in the duration of the Rakuten dataset from 01-09-2019 to 01-09-2022. The **Figure 32** is the plot of the dataset whose x-axis is the time sequence and the y-axis is the number of days.

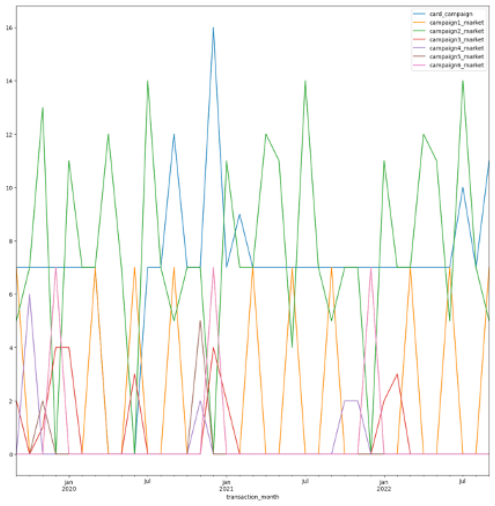

図 32 The Rakuten campaigns data

7·2 LINGAM

This analysis focuses on the causal relationship between the campaign and card data by the Linear non-Gaussian

Acyclic Model (LINGAM). So far, most statistics and machine learning have been based on correlations to understand relationships. However, what we want to know: what causes this variation, is uncertain. Causal search refers to a method of inferring causal structure from data. It involves the direction and magnitude of causal relationships and is a way of finding causes, which is different from traditional statistics and machine learning. The LINGAM is a typical approach to statistical causal search and the model of LINGAM is generally described as the following equation [5]:

$$\boldsymbol{x} = \boldsymbol{C}\boldsymbol{x} + \boldsymbol{e} \tag{6}$$

In this equation, $\boldsymbol{x}$ is a vector containing n events, $\boldsymbol{C}$ is the adjacency matrix and the element $\boldsymbol{c_{ij}}$ of $\boldsymbol{C}$ is the causal factor from the event j to the event i. The matrix $\boldsymbol{C}$ allows for a causal diagram to be drawn. The causal diagram can show the causal relationship with a direct view.

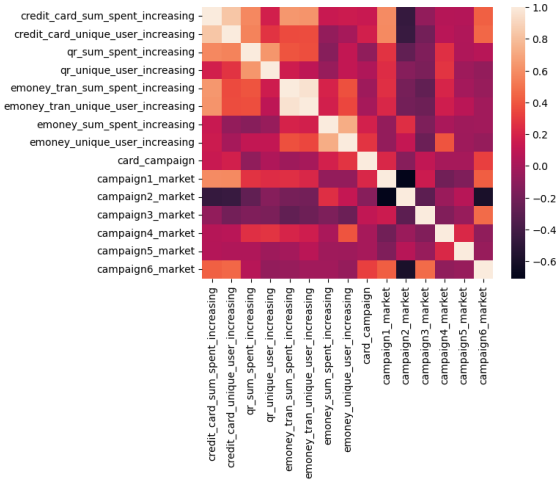

図 33 The correlation colour map of campaign data and Rakuten data

7·3 Analysis Result

Before the causal analysis, we did not know the causal relationship between the original dataset and the campaign data. To analyze the casual relationship, consider the increase in the spent amount and the increase in user count of the four payment types including credit card, QR, transportation IC and e-money respectively and the campaign data. The **Figure 33** is the colour map of the correlation coefficient among features. This figure shows that there is a correlation between the data relating to promotions and the growth in spending and user count, although the correlation coefficient does not indicate a strong correlation.

Due to the LINGAM model, the adjacency matrix of features is calculated, and **Figure 34** is the colour map of

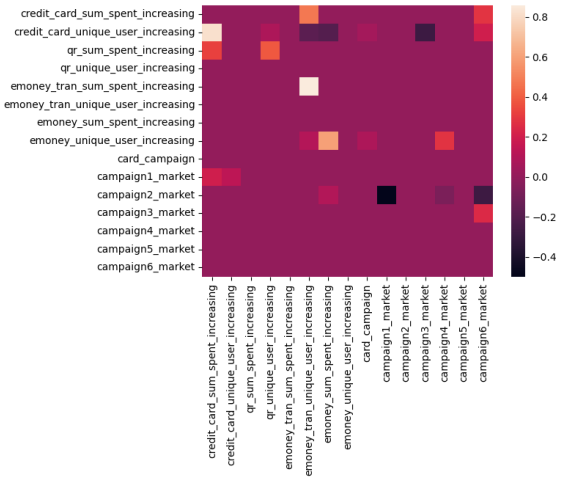

図 34 The colour map of adjacency matrix

this adjacency matrix. From this figure, we can get a rough view of the causal relationship between the features.

図 35 The colour map of adjacency matrix

However, in **Figure 35** the causal diagram shows how the campaigns influence the increase of payment data. Obviously, we know that the campaigns of Rakuten cards do positively influence the user count of Rakuten cards and e-money payment way. And campaign 3, campaign 4 and campaign 6 have made the users count of credit cards, users count of e-money payment and spent amount of credit cards increase respectively.

Due to the dataset of campaigns including some general campaigns every month and every year, this dataset cannot earn a good enough result or explain the increase of data in the dataset provided by Rakuten. Although, we' found some causal relationship between these campaigns and the changing trends of the original dataset. To find a better explanation of unnormal fluctuations of data, building a complete campaign dataset is necessary for future work and may improve the causal analysis result.

8. Conclusion

In this paper, we conduct an overall analysis of the payment data provided by Rakuten, including exploratory data analysis, correlation analysis, regional analysis, time series analysis, and prediction.

We analyzed the relationships between credit card pay-

ments and different cashless payments and found that the spent amount of QR code payment has the strongest relationship with the spent amount of credit card payment. At the same time, differences in regions could also bring differences in the relationship between different payment methods.

In the time series analysis, we explore the seasonality and trend of the total spent amount of credit card payment, and further give suggestions regarding the seasonality patterns. Then we used three different models to conduct the prediction of sum spent amount and the average spent amount of credit card payment. As a result, ARIMA shows the best performance of prediction and reaches a result with normalized MSE score of 7.39×10^{-5} because of its great ability to fit in small datasets.

Moreover, the causal analysis is conducted for the abnormal data growth, we found the clear relationship between campaigns and spent amount changes.

References

[1] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[2] A. Vaswani et al., “Attention is all you need,” in Advances in neural information processing systems, 2017, vol. 30.
[3] https://matsunosuke.jp/rakutencard-campaign-past/
[4] https://matome-search.com/useful/rakuten-sale/#regular-sale
[5] Shimizu, Shohei, et al. ”A linear non-Gaussian acyclic model for causal discovery.” Journal of Machine Learning Research 7.10 (2006).

About the person in charge of work

Responsibilities in each section of this report are as follows:

(1) **Wang Boyu**: Exploratory Data Analysis, Correlation Analysis, Regional Specific Analysis, Time Series Analysis, ARIMA Model.
(2) **Xu Wangjie**: Time Series Analysis, Attention-LSTM Model, Correlation Analysis, Comparisons
(3) **Valentin Fontanger**: Prophet Model
(4) **Shengjie Fang**: Exploratory Data Analysis, Causal Analysis
(5) **Boyu Wang**: Causal Analysis