SEFA KAMBUR / IE 360 /

Statistical Forecasting and Time Series Homework 1

Content

1. Introduction

2. Data Preparation

3. Exploratory Data Analysis

4. Modeling

5. Model Evaluation

6. Model Improvement

7. Conclusion

8. References

1. Introduction

Decision makers use data to respond to changes faster or to make more accurate plans for the future. As part of the IE360 course, we learned how to analyze data and make a model. In this study, some data were analyzed and interpreted as far as I learned from the IE360 course. Code writing was done via R studio.

In this study, the data is as follows: The percentage of searches for the word "promotion" in Google trends, credit card expense index, card payment index, general clothing & food expenses.

The purposes of doing this analysis are to:

- To understand the underlying patterns and trends in the dataset.
- To identify key factors influencing the "promotion trend"
- To develop a predictive model for promotion trend percentage using regression analysis.
- To generate actionable insights and recommendations based on the analysis findings.

2. Data Preparation

Data preparation, which includes preparing the raw data for analysis by cleaning, converting, and organizing it, is an essential stage in any analysis process. In this study, first of all, "promotion" data was uploaded to Rstudio and this data was made ready for analysis and modeling.

```
file_path <- file.choose()
promotion_data  <- read_excel(file_path)
```

When the promotion data is examined, it is seen that it consists of 3 or 4 records per week. Since Google Trends web service does not provide this data in monthly format, promotion data has been converted to monthly format.

Raw Promotion Data :

07.01.2018	17
14.01.2018	19
21.01.2018	18
28.01.2018	19
04.02.2018	22
11.02.2018	21
18.02.2018	18
25.02.2018	20

Monthly Promotion Data:

2018-01-01	56
2018-02-01	81
2018-03-01	74
2018-04-01	85

Monthly promotion data was obtained with the following code:

```
monthly_promotion <- promotion_data %>%
group_by(Month) %>%
summarize(Promotion_Trend = sum(Promotion_Trend))
```

Another process was to give column names:

```
colnames(promotion_data) <- c("Date", "Promotion_Trend")
```

Finally, the NA value was checked. In the data we have, the NA number was obtained as zero.

```
missing_values <- colSums(is.na(monthly_promotion))
print(missing_values)
```

Three additional data were obtained from CBRT for this study. This data was mentioned in the introduction. The following steps were followed to make this data ready for analysis and modeling:

- Data was first uploaded to Rstudio with the upload function.
- It was observed that the date column was of char type and this column was converted to date format.
- It has been checked that the number of NA values in the CBRT data is zero.

```
file_path <- file.choose()
data <- read_excel(file_path)
data$Tarih <- as.Date(paste0(data$Tarih, "-01"), format =
"%Y-%m-%d")
missing_values <- sum(is.na(data))
```

The last step in this section is to combine the CBRT and monthly promotion data in two different tables into a single table.

```
merged_data <- merge(monthly_promotion, data, by.x = "Month",
by.y = "Tarih", all = TRUE)
```

3. Exploratory Data Analysis

An essential step in the data analysis process is exploratory data analysis (EDA), which aims to reveal patterns and relationships in the dataset and provide insights into it. Analysts use a range of methods to condense, illustrate, and analyze the data during EDA. In order to pinpoint important trends and patterns, this phase usually include looking at summary statistics, distributions, correlations, and linkages between variables. Visualizations like heatmaps, box plots, scatter plots, and histograms are frequently used to visually study data and identify any anomalies or outliers.

Essential libraries are loaded such as readxl, lubridate, zoo, dplyr, ggplot2, and forecast. These libraries provide functionalities for data manipulation, visualization, and time series analysis. The data is read from an Excel file.

```
library(readxl)
library(lubridate)
library(zoo)
library(dplyr)
library(ggplot2)
library(forecast)
```

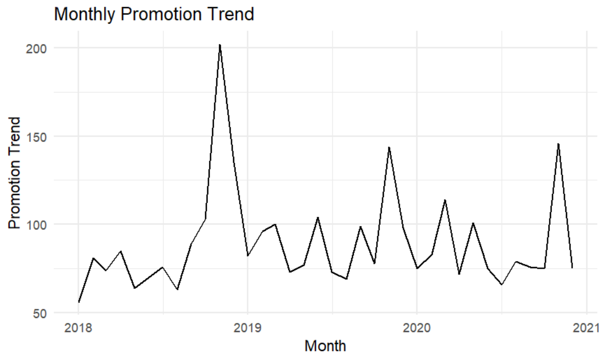
The structure of the dataset is inspected using the str() function to understand its variables and types.

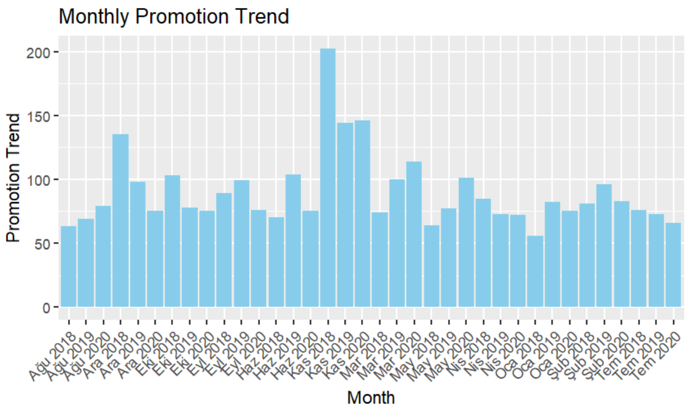
```
str(promotion_data)
```

Visualizations is created to explore the data further, including time series plots and bar plots of monthly promotion trends.

```
ggplot(monthly_promotion, aes(x = Month, y = Promotion_Trend))
+
geom_line() +
labs(x = "Month", y = "Promotion Trend", title = "Monthly
Promotion Trend") +
theme_minimal()
```

```
ggplot(monthly_promotion, aes(x = format(Month, "%b %Y"), y =
Promotion_Trend)) +
geom_bar(stat = "identity", fill = "skyblue") +
labs(x = "Month", y = "Promotion Trend", title = "Monthly
Promotion Trend") +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) #
Rotate x-axis labels for better readability
```





When you look at the plot images, you can see seasonality. We know that in the last months of each year, "legendary Friday" and December promotions increase people's online shopping tendencies. We see this situation clearly in the images.

We conduct time series analysis by converting the merged data into a time series object and exploring its characteristics.

```
ts_merged_data <- ts(merged_data[, -1], start = c(2018, 1),
                        frequency = 12)
```

4. Modeling

In this section, we perform linear regression analysis to model the relationship between the promotion trend and several predictor variables including credit card expense index, card payment index, general clothing & food expenses.

In the linear regression modeling process, we aim to establish a relationship between the promotion trend and several predictor variables. The model formulation involves expressing the promotion trend (X1, X2....Xn):

$$Y = B0 + B1X + B2X2.... BnXn + e$$

where

B0 is the intercept term,

B1X + B2X2.... BnXn are the coefficients of the predictive variables,

e is the error term representing unexplained variability.

The linear regression is fitted to our data using the lm() function in R. This function estimates the coefficients (B) that best fit the data to minimize the sum of squared residuals.

```
lm_model <- lm(Promotion_Trend ~ `Credit Card Expenses` +
`Kartlı Ödeme Endeksi` + `Consumer PI` + `TP KARTMETRE D1`,
data = ts_merged_data)
summary(lm_model)
```

```
Call:
lm(formula = Promotion_Trend ~ `Credit Card Expenses` +
`Kartlı Ödeme Endeksi` +
    `TP FG J03` + `TP KARTMETRE D1`, data = ts_merged_data)

Residuals:
   Min      1Q Median      3Q    Max
-35.10 -15.57  -2.16  10.27  65.23

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             -5.548e+01  1.180e+02  -0.470  0.64143
`Credit Card Expenses`   7.051e-06  4.215e-06   1.673  0.10446
`Kartlı Ödeme Endeksi` -4.464e-01  1.003e+00  -0.445  0.65946
`Consumer PI`          1.016e+00  3.609e-01   2.816  0.00837 **
`TP KARTMETRE D1`       -4.838e-01  2.823e-01  -1.714  0.09657
.
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.85 on 31 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.393,   Adjusted R-squared:  0.3147
F-statistic: 5.019 on 4 and 31 DF,  p-value: 0.003089
```

5. Model Evaluation

The model evaluation involves evaluating the performance and reliability of the linear regression model fitted to the data.

Residual Standard Error: The residual standard error measures the average deviation of the observed values from the fitted values by the model. In this case, the residual standard error is approximately 23.85, indicating the average discrepancy between the observed promotion trend and the values predicted by the model.

Multiple R-squared: The multiple R-squared value, also known as the coefficient of determination, represents the proportion of variance in the response variable (promotion trend) explained by the predictor variables . A value of 0.393 suggests that approximately 39.3% of the variance in the promotion trend is accounted for by the predictor variables included in the model.

Adjusted R-squared: The adjusted R-squared value adjusts the multiple R-squared value for the number of predictors in the model and provides a more accurate measure of the model's goodness of fit. With an adjusted R-squared value of 0.3147, we can infer that around 31.47% of the variability in the promotion trend is explained by the independent variables.

F-statistic: The F-statistic tests the overall significance of the linear regression model by comparing the fit of the model with a constant-only model (intercept-only model). In this case, the F-statistic is 5.019 with a corresponding p-value of 0.003089, indicating that the model as a whole is statistically significant at the 0.05 significance level.

Residuals: The residuals represent the differences between the observed promotion trend values and the values predicted by the model. The summary statistics of the residuals indicate that they range from -35.10 to 65.23, with quartiles at -15.57, -2.16, and 10.27. These values provide information about the spread and distribution of the residuals around the fitted line.

6. Model Improvement

The linear regression model was fitted to predict the log-transformed Promotion_Trend based on the predictor variables.

```
lm(formula = log_Promotion_Trend ~ Credit.Card.Expenses +
Kartlı.Ödeme.Endeksi +
    TP.FG.J03 + TP.KARTMETRE.D1, data = ts_merged_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32141 -0.19355 -0.00317  0.12922  0.43773

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.122e+00  1.112e+00   2.808  0.00855 **
Credit.Card.Expenses  5.444e-08  3.974e-08   1.370  0.18056
Kartlı.Ödeme.Endeksi -4.020e-03  9.457e-03  -0.425  0.67374
TP.FG.J03            9.263e-03  3.402e-03   2.723  0.01053 *
TP.KARTMETRE.D1     -4.008e-03  2.661e-03  -1.506  0.14216
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2248 on 31 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3615,  Adjusted R-squared:  0.2791
F-statistic: 4.388 on 4 and 31 DF,  p-value: 0.006305
```

Residuals: The residuals exhibit a mean close to zero and are approximately symmetrically distributed around zero, indicating that the model's assumptions regarding residuals are reasonable.

Coefficients: The coefficients indicate the direction and magnitude of the effect of each predictor variable on the log-transformed Promotion_Trend. However, only the coefficient for TP.FG.J03 (general clothing & food expenses.) is statistically significant at the 5% level.

R-squared: The multiple R-squared value of 0.3615 suggests that approximately 36.15% of the variance in the log-transformed Promotion_Trend can be explained by the predictor variables in the model.

F-statistic: The F-statistic tests the overall significance of the model. With a p-value of 0.006305, the model is statistically significant at the 5% level.

7. Conclusion

In conclusion, our discussion revolved around using linear regression to understand and predict promotional trends based on various factors. I started by preparing the data and then delved into exploring its patterns and trends. Building on this, I constructed a regression model to predict promotional trends, considering factors like credit card expenses and payment indices.

After fitting the initial model, we evaluated its performance and identified areas for improvement. One such improvement involved applying a log transformation to the response variable to address certain statistical issues. However, the model did not improve as a result of the log transform.