УДК 513.7

Т. Г. Ємель'яненко

Дніпровський національний університет імені Олеся Гончара

ПОБУДОВА ПРОГНОЗІВ З УРАХУВАННЯМ ДОДАТКОВИХ ДАНИХ, ЩО ВПЛИВАЮТЬ НА ПОВЕДІНКУ ЧАСОВОГО РЯДУ

Описано шляхи модифікації методів прогнозування з урахуванням додаткових факторів, таких як день тижня, прогноз погоди, свята. Розглянуто комбінований підхід, який враховує використання багатовимірної регресійної моделі та моделі ARIMA.

Ключові слова: *прогнозування, часовий ряд, прогноз, ARIMA, лінійна регресійна модель, фіктивні змінні.*

Рассмотрены пути модификации методов прогнозирования с учетом дополнительных факторов, таких как, день недели, прогноз погоды, праздники. Рассмотрен комбинированный подход, учитывающий использование многомерной регрессионной модели и модели ARIMA.

Ключевые слова: *прогнозирование, временной ряд, прогноз, ARIMA, линейная регрессионная модель, фиктивные переменные.*

When constructing forecasts, especially when it comes to time series of sales, there is the task of considering the day of the week, since orders on weekend, for example, may not be accepted or accepted, but not processed. If we have additional information, such as state holidays or school holiday we can use this information and include additional parameters to the model. For some time series we can include to the model such parameter as weather characteristics. For example, if we analyze time series such as the number of visitors to the restaurant, then it is likely that the number of visitors will be less than during the same period when the weather was more favorable. Linear regression allows considering both the information about the previous behavior of the time series and additional factors that may affect the time series change. If we want to include to the model such qualitative feature as, whether there is a holiday or not, it is necessary to include an additional dummy variable "Holiday", which will equals to 1 on the day of the holiday and 0 on all other days. If we want to consider the day of the week when constructing the forecast, then six dummy variables are taken into consideration. Variableswill equal to 1 on the corresponding day of the week and for the last day of week all the variables will be equal to 0.It is supposed that the errors of a linear regression model contain autocorrelation, so the forecast was constructed for

the combined regression model and ARIMA. For the forecasting a separate forecast for the regression model and the ARIMA model should applied, and then the results should combined. This approach was illustrated on the sales forecasting for one of the Rossman store. The time series was retrieved from the Rossman store sales data presented at the Kaggle competition.

Keywords: *forecasting, time series, forecast, ARIMA, linear regression model, dummy variables.*

Вступ. Під час побудови прогнозів, особливо якщо це стосується часових рядів продаж, виникає задача урахування додаткових факторів, таких як проведення промоакцій, день тижня, оскільки у вихідні, наприклад, можуть не прийматися замовлення, або прийматися, але не оброблятися. Можна враховувати наявність свят у період, що розглядається. На деякі процеси, що вивчаються, можуть впливати такі сторонні фактори, як погода, наприклад, якщо розглядається часовий ряд кількості відвідувачів ресторану, то імовірно, що кількість відвідувачів буде меншою, ніж у цей самий період, коли погода була сприятливішою.

Аналіз останніх досліджень і публікацій. Питання побудови прогнозу, який враховує, окрім часового ряду, додаткові фактори, розглянуто в роботі [1]. У роботі [2] наведено детальний розбір динамічних регресійних моделей, які представляють комбінацію ARIMA та регресійних моделей.

Постановка задачі. Запропонувати можливі варіанти урахування додаткових факторів під час побудови прогнозів.

Основний матеріал. Припустимо, що будується проста прогнозна модель на основі лінійної регресії

$$u_t = \beta_0 + \beta_1 t + \varepsilon_t.$$

У разі необхідності урахування такої якісної ознаки як наявність свят або їх відсутність, треба ввести додаткову фіктивну змінну «Свято», яка буде приймати значення 1 в день свята і 0 в усі інші дні. І в цьому випадку будується багатовимірна лінійна регресійна модель у такому вигляді:

$$u_t = \beta_0 + \beta_1 t + \beta_2 h_t + \varepsilon_t.$$

Якщо ми бажаємо врахувати день тижня під час побудови прогнозу, то у розгляд вводяться шість фіктивних змінних, які приймають значення 1 у відповідний день тижня і нульові значення в останній, сьомий, день.

Тоді модель набуває такого вигляду:

$$u_t = \beta_0 + \beta_1 t + \beta_2 w_{1t} + \beta_3 w_{2t} + \beta_4 w_{3t} + \beta_5 w_{4t} + \beta_6 w_{5t} + \beta_7 w_{6t} + \varepsilon_t.$$

Таблиця 1

Фіктивні змінні для урахування дня тижня

День тижня	w_{1t}	w_{2t}	w_{3t}	w_{4t}	w_{5t}	w_{6t}
Понеділок	1	0	0	0	0	0
Вівторок	0	1	0	0	0	0
Середа	0	0	1	0	0	0
Четвер	0	0	0	1	0	0
П'ятниця	0	0	0	0	1	0
Субота	0	0	0	0	0	1
Неділя	0	0	0	0	0	0

У разі, якщо необхідно врахувати інший кількісний показник, наприклад, прогноз погоди, то необхідна додаткова змінна включається в регресійну модель.

Використання багатовимірної регресії дозволяє врахувати як інформацію про попередню поведінку часового ряду, так і про додаткові фактори, що можуть впливати на динаміку зміни ряду. Розглянемо, як можна урахувати додаткові фактори в моделях, що дозволяють більш тонке налаштування, наприклад, у моделях авторегресії.

Вважатимемо, що похибки багатовимірної лінійної регресійної моделі містять автокореляцію

$$u_t = \beta_0 + \beta_1 t + \cdots + \beta_k t + \delta_t,$$

де δ_t відповідає ARIMAмоделі, наприклад, ARIMA(1,1,1).

$$(1 - \varphi_1)B(1 - B)\delta_t = (1 + \theta_1 B)\varepsilon_t,$$

де ε_t – білий шум.

Під час оцінки параметрів моделі нам необхідно мінімізувати суму квадратів ε_t, якщо ми замість цього мінімізуємо суму квадратів δ_t, то результатом буде те, що оцінки параметрів $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ не будуть ефективними; результати перевірки моделі за статистичними критеріями, наприклад t-тест, будуть невірними; у більшості випадків p-значення оцінки значущості коефіцієнтів моделі будуть заниженими і деякі пояснюючі змінні будуть здаватися важливішими, ніж вони є насправді.

У разі, якщо мінімізуються суми квадратів похибок ε_t, цих проблем вдається уникнути.

Для побудови прогнозу за комбінованою моделлю регресії та ARIMA слід виконати прогноз окремо для регресійної моделі та моделі ARIMA, а потім об'єднати результати.

Розв'яжемо задачу прогнозування для одного з магазинів Rossman з

урахуванням дня тижня і наявності шкільних канікул. Дані отримані з бази даних, яка була надана компанією Rossman для проведення Kaggle змагання по прогнозуванню обсягу продажів. На рис. 1 наведено часовий ряд обсягів продажів за період з січня 2013 по липень 2017 року. Як бачимо, часовий ряд набуває нульових значень у неділю, коли магазин був зачинений. Вигляд часового ряду свідчить про наявність сезонності.

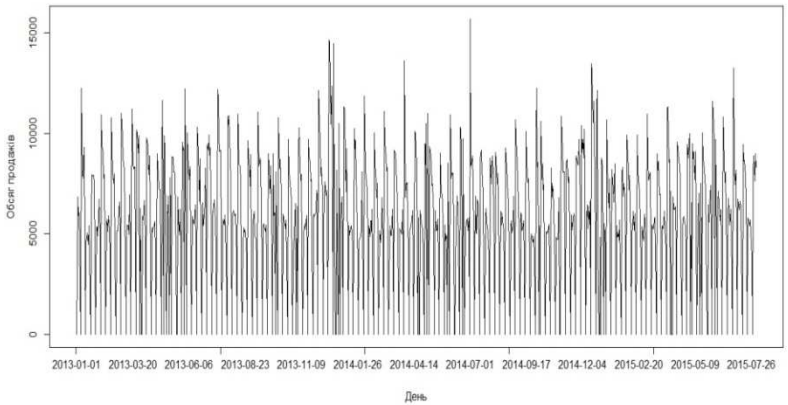


Рисунок 1 – Графік часового ряду обсягів продажу

За графіком автокореляційної функції часового ряду (рис. 2) можна зробити висновок про наявність сезонності з періодом 7. Тому будемо підбирати сезонну модель ARIMA.

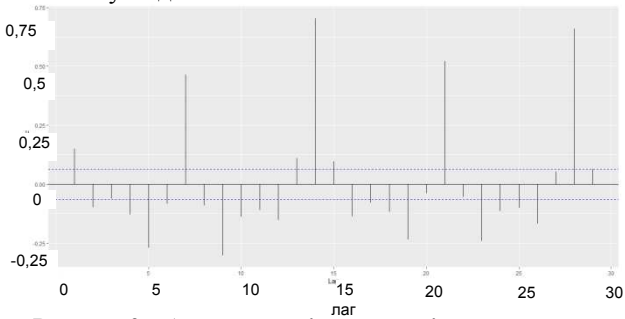


Рисунок 2 – Автокореляційна функція часового ряду

На рис. 3 наведено результати підбору моделі, таким чином, модель має вигляд:

$$u_t = 10651.6572 - 1191.8571w_t - 774.4742h_t + \delta_t,$$

де залишки δ_t представляють собою сезонну модель ARIMAз параметрами ARIMA(2,0,0)(2,0,0)7, w_t – день тижня, h_t – наявність шкільних канікул.

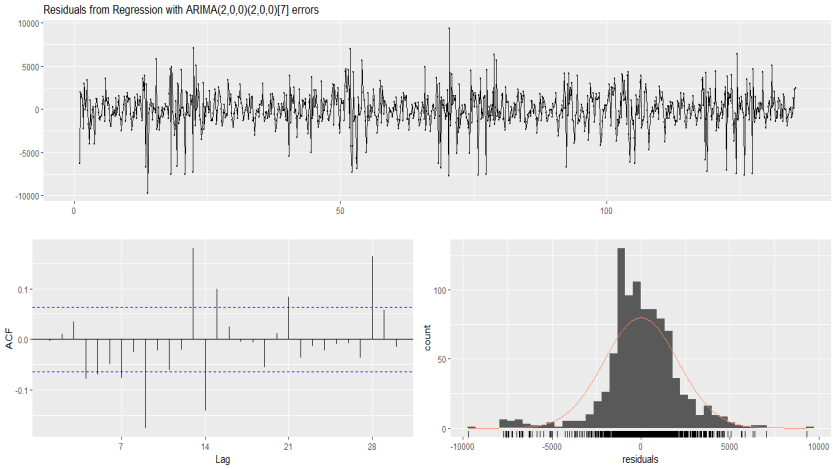
```
Regression with ARIMA(2,0,0)(2,0,0)[7] errors

Coefficients:
         ar1     ar2    sar1    sar2   intercept         Day  SchoolHoliday
      0.2082  0.1077  0.1010  0.4673  10651.6572  -1191.8571      -774.4742
s.e.  0.0344  0.0344  0.0305  0.0305    392.7868     78.1646       231.7195

sigma^2 estimated as 4347747:  log likelihood=-8534.33
AIC=17084.67   AICc=17084.82   BIC=17123.45
```

Рисунок 3 – Результати підбору сезонної моделі ARIMA

На рис. 4 представлено графік залишків та автокореляційної функції залишків, а також результати перевірки тесту Льюнга – Бокса, а на рис. 5 – прогноз на 10 днів, побудований за обраною моделлю.




```
        Ljung-Box test

data:  Residuals from Regression with ARIMA(2,0,0)(2,0,0)[7] errors
Q* = 104.08, df = 7, p-value < 2.2e-16

Model df: 7.   Total lags used: 14
```

Рисунок 4 – Результати перевірки моделі

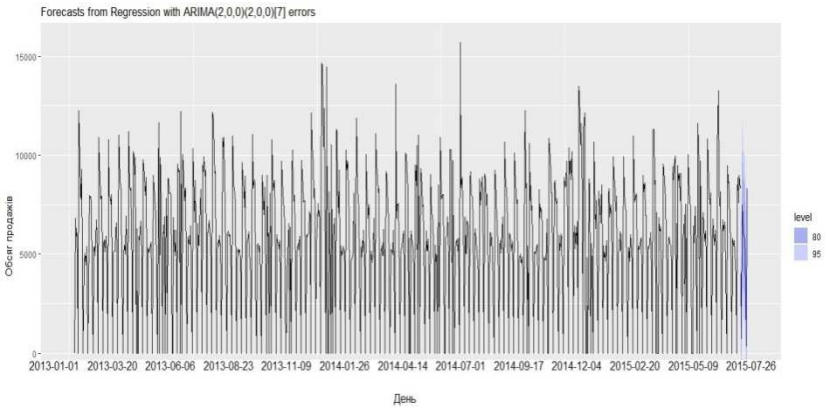


Рисунок 5 – Побудований прогноз

Висновки. Розглянуто підхід побудови прогнозів для часових рядів при наявності додаткових даних, які можуть впливати на поведінку часового ряду. Описаний підхід дозволяє будувати комбінований прогноз з використанням регресійних моделей та моделей ARIMA.

Бібліографічні посилання

1. Hyndman R. J., Athanasopoulos G. Forecasting principles and practice. OTexts, 2018. 382 p.
2. Pankratz A. E. Forecasting with dynamic regression models. N.Y.: John Wiley & Sons, 1991. 400 p.

Надійшла до редколегії 15.11.2019.