26.05.2021

PROJECT II: FORECASTING CINEMA SALES

Salah ABUKAR (N°: 19439561) – Ilana COHEN (N°:17433293) – Edmundo PAEZ RAMIREZ (N°: 17324823) –

Julie TIMMERMANS (N°: 17203993) – Siwei YU (N°: 17587213)

INTRODUCTION

In the last two years and especially since Covid-19, a clear shift in consumer's behaviour has been observed. In 2020, 36% of the individuals declared they would prefer to stream a film at home rather than visit a cinema. With the arising success of streaming platforms such as Netflix, one might doubt in the future of movie theatres. Therefore, forecasting the sales level can be valuable for cinemas. It gives them the ability to make informed business decisions, develop data-driven strategies, estimates the costs, and undertake actions to meet their target profit level.

The **goal of our project** is to forecast the number of tickets sold in November 2018 for the best seller film for the cinema with the highest sales. To answer this question, thanks to the data provided by the cinema, two models will be used: the ARIMA and the regression model.

Description of the data

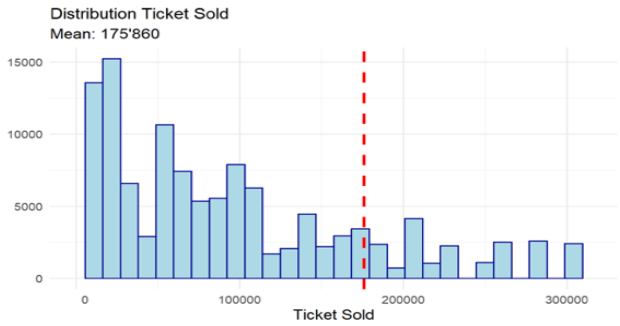
The data contains daily sales information about 246 cinemas and 48 different movies in 2018. In this project the following 5 variables will be used:

- Date: date the sales take place
- Tickets sold: number of tickets sold in a day for a certain movie in a certain cinema
- Ticket price: price of one movie ticket for a certain movie in a certain cinema
- Show time: the number of screening hours for a certain movie in a certain cinema
- Capacity: number of seats available in one room of a certain cinema.

EXPLORATORY DATA ANALYSIS

The exploratory data analysis has been performed using the data provided by the 246 cinemas for the 48 different movies.

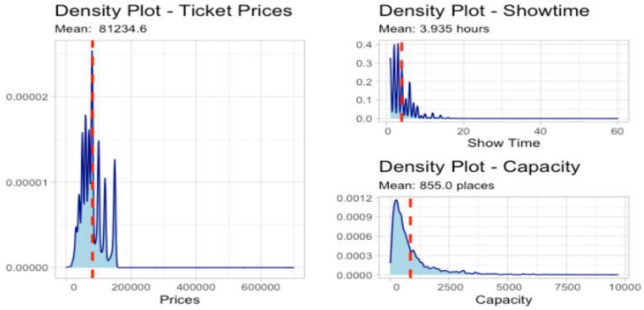
Distribution Tickets Sold



This first graph represents the distribution of the number of tickets sold. We can see that the maximum number of tickets sold in one day is 311’469.

The minimum number of tickets sold in one day is 0. The average number of tickets sold in one day is 175’860 tickets (red dashed line). There is a lot of variance in the number of tickets sold.

Distribution Ticket Price – Capacity – Show Time



The first graph represents the variable ticket price, which is ranged between 483.9 and 700'000. The second graph represents the variable show time and is ranged between 1 and 60 hours. The third graph represents the capacity and is ranged between 10 and 9'692 seats. From these graphs, several assumptions are done.

First, due to the high range of prices, we assume it is a foreign currency. Secondly, as the maximum number of screening hours exceeds 24 hours, we will assume that one movie can be shown in several rooms of one cinema simultaneously. The last assumption is that the capacity represents the number of seats available in one cinema room.

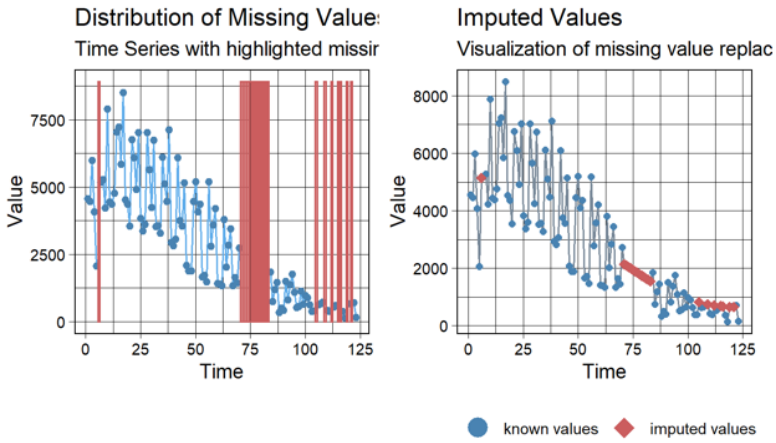
Correlation

The variable ticket sold is positively correlated to all the other variables. Ticket sold and ticket price are positively correlated (0.104), which can seem counterintuitive. The correlation between ticket sold and the screening hours is equal to 0.522 and the one between ticket sold and capacity is equal to 0.425. In conclusion, an increase in the capacity, ticket price and show time is associated to an increase in the number of tickets sold by the cinema.

ANALYSIS

The analysis and prediction part will be dedicated to the bestselling cinema for one film (best seller film).

Missing data

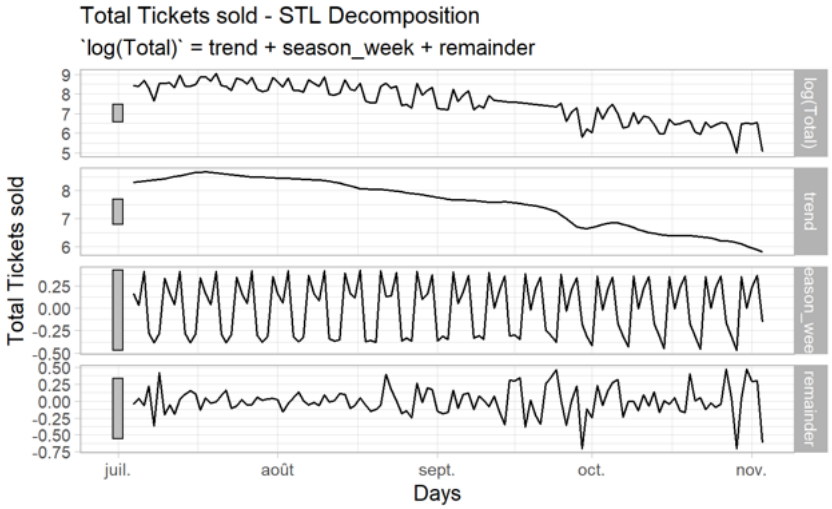


Initially, the time series contains 102 observations from the 4th of July 2018 until the 3rd of November of the same year. However, as seen on the first left graph, the time series contains some missing data.

For some days, the number of tickets sold was not provided by the cinema. Therefore the gaps were filled with a model called auto.arima (see second graph). However, missing data influence negatively the performance of the forecasting models.

Seasonal and Trend Decomposition

To identify the seasonal and trend pattern in the time series, the following time series decomposition is represented.



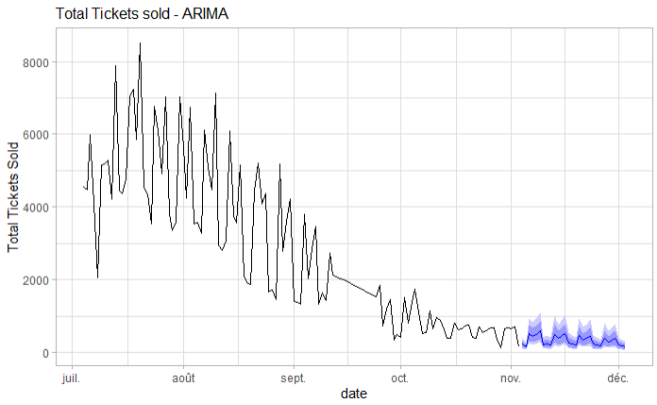
The whole time series is represented on the first graph, the lack of information in September can easily be identified. The second one represents the trend component, which decreases on the whole period. It shows that the sales continuously decrease over time. On the third graph we can observe the strong additive seasonality. The regular spikes represent the two highest sales day of the week, namely Tuesday and Friday.

There is no observable pattern in the last graph (remainders), meaning that most of the information about the sales is captured by the seasonal and trend component. This is necessary for the implementation of forecasting models.

FORECASTING METHODS

To predict the number of tickets sold in November 2018, two models will be implemented and compared based on their results.

I. ARIMA Model

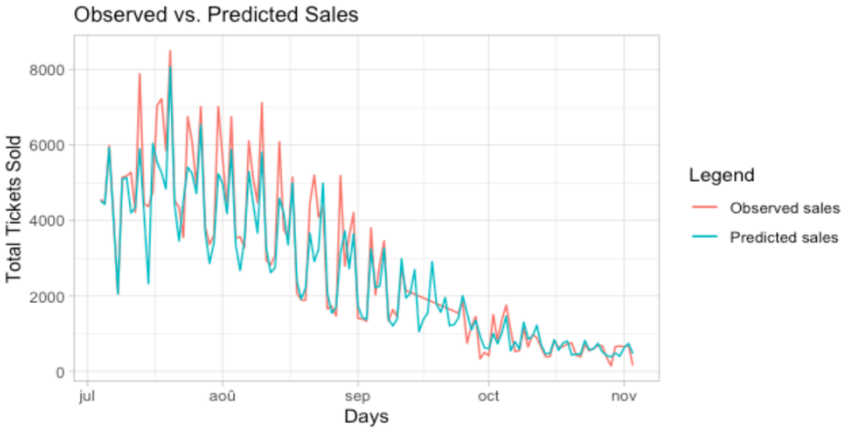


The results of the first model (ARIMA model) can be visualized on this graph.

The blue line represents the forecasts of the tickets sold for November 2018. The seasonal pattern of the previous month is considered meaning that the predicted sales will be higher on Tuesdays and Fridays. In addition, the trend of the forecasted sales is stagnating which is due to the slowly decreasing trend in the past.

The confidence interval (represented by the blue range on the graph) is relatively narrow, indicating that there is 95% probability that confidence interval contains the real future sales.

Performance of the model



By analysing the predicted vs the observed sales one can assess the goodness of fit of the model. The closer they are, the more the model fits the data. The forecasted sales follow well the real sales fluctuations and the decreasing trend of the sales. Nevertheless, there are some differences between the predicted and observed sales (for example, between September and October). This can be explained by the lack of data provided by the cinema.

Furthermore, a statistical test (Ljung Box test) was computed to confirm the validity of the model to ensure that most of the relevant information about the cinema sales were captured by the implemented model.

ARIMA model conclusion

The ARIMA model is valid and predicts the tickets sold for November 2018 well. However, the model uses only past values of tickets sold to make predictions. Indeed, the number of tickets sold by the cinema could be related to other variables as mentioned in the EDA part. In the next section, a regression model is implemented to forecast the daily number of tickets sold using the capacity, ticket price and show time as predictors.

II. Regression model

The aim of this model is to forecast the sales using all the information provided by the cinema and to compare them with the previous model that used only the sales data. By doing so, the cinema will be able to assess whether these different variables influence its sales level.

Coefficient's interpretation

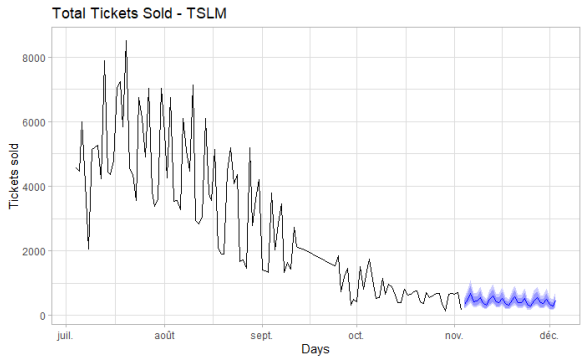
	term	pceffect
1	(Intercept)	132955.0173
2	trend()	-0.9636
3	Capacity	0.0499
4	Price	-0.0002
5	Show_time	-5.5716
6	weekdaydimanche	12.6387
7	weekdayjeudi	58.4011
8	weekdaymardi	56.2992
9	weekdaymercredi	24.5649
10	weekdaysamedi	7.637
11	weekdayvendredi	64.587

The table shows the coefficients (in %) of the regression model. The trend coefficient shows that every day, the number of tickets sold decrease by 1%. The first predictor capacity is positive, an increase in cinema capacity is associated to higher tickets sales. The price coefficient is negative and very small meaning that if the ticket price increases by 1000 monetary units, the number of tickets sold decrease by 0.2%. The negative show time coefficient shows that an increase in the movie show time is associated to lower tickets sales.

The weekdays variable has 7 levels, one for each day of the week. The reference level is Monday. As we can see the coefficients are positive for all days of the week, meaning that switching from Monday to any other day of the week is associated with higher tickets sales. More precisely, when switching from Monday to Friday, the number of tickets sold increases the most (on average by 65%). To sum up, high sales are associated to high-capacity rooms, low ticket price and low screening time.

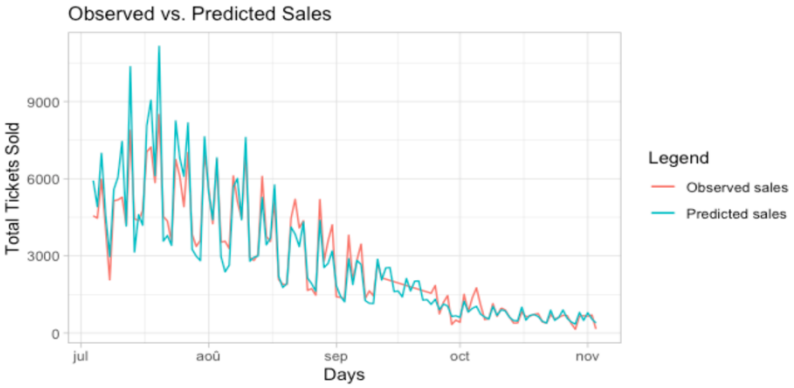
To build a regression model, the ticket prices, show times and room capacities of the cinema need to be predicted for November 2018 as they were not provided by the cinema. The two following methods will be used: ex-ante forecast using ARIMA and scenario-based forecast.

Ex-ante forecast with ARIMA



Here, once each predictor is forecasted with an ARIMA model, the regression model is applied to predict the number of tickets sold. From the graph, we can see that the seasonal pattern is captured by the model. Moreover, the predicted sales seem to be slightly higher compared to the previous month and the sales predicted with ARIMA. Indeed, this model is more optimistic about the future sales of the cinema. As the confidence interval is even lower than with the ARIMA model, it means that this model is more precise.

Performance the model

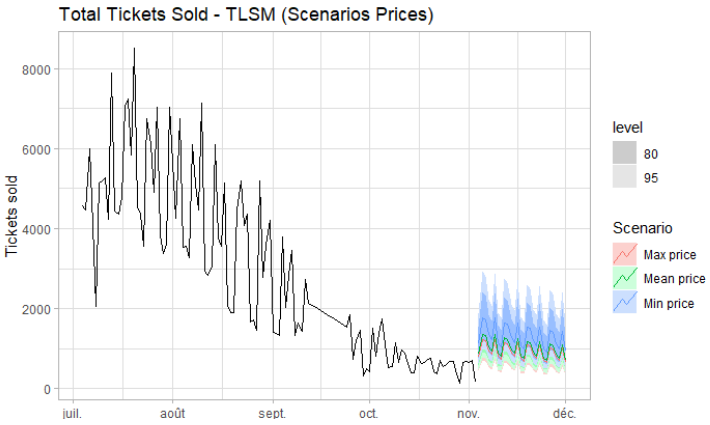


As previously the observed and predicted sales are compared. From the graph, we can see that between mid-July and mid-August the fluctuations in the predicted sales are higher than in the real sales. Overall, the regression model tends to over-estimate the number of tickets sold.

As for the previous model, a statistical test was conducted to ensure the validity of the model.

Scenario based forecasting

As mentioned above, higher price is associated to lower sales. Therefore, three different price scenarios will be analysed.



Each scenario represents a different ticket price (minimum, mean and maximum price) while the other predictors (show time and capacity) are constant for the predicted period. The graph shows that the sales are the highest for the minimum ticket price. This assumption makes sense and illustrate the negative coefficient related to the price we saw above for the regression model. However, the confidence intervals are large. The predicted sales of each scenario are much higher than for the previous month and do not follow the decreasing trend. The previous regression model (Ex-ante with ARIMA) is more realistic and preferred to predict the cinema's future sales.

RESULTS: ARIMA vs Regression model

Forecast's comparison

Forecasts

Date	Tickets sold TSLM	Tickets sold ARIMA
2018-11-04	349	230
2018-11-05	515	156
2018-11-06	690	522
2018-11-07	429	460
2018-11-08	442	503

The following table presents the forecasts for beginning November 2018 with both models. As observed during the analysis, the forecasts of the regression model are slightly higher compared to the ARIMA model (for example from 04/11/18 to 06/11/18).

The predicted number of tickets sold is approximatively constant for the whole month of November (while maintaining the high sales days on Tuesday and Friday). According to our analysis, the cinema should not worry about the future sales, they will not continue to decrease in November 2018.

Accuracy comparison

To evaluate the accuracy of the models, forecasts errors are computed. By comparing the two models, the regression model performs better in terms of accuracy. We can conclude that the regression model is more accurate and suitable to predict the sales of the cinema. In fact, the capacity, screening time and ticket price are valuable information to consider when predicting the cinema sales.

CONCLUSION

As useful and necessary forecasting sales might be, it remains a tedious task. In our analysis, two different models were used and compared to predict the future sales of the cinema. The ARIMA model based on past values and the regression model enabled us to add all the information provided about the cinema. Even though the regression model is more accurate and precise, it requires more information. Furthermore, it might be a more uncertain and complex model.

This analysis could be continued in multiple ways, one possibility could be to make forecasting combination: average the forecasts of the combined models or add other predictors (type of movie, weather, number of cinemas in the area...). Finally, it is important to keep in mind that the forecasts made for the cinema are an honest projection of what will happen given the information available. The aim of forecasting is to know the future to be able to do something about it. In the case of this cinema, the sales predicted for the next 30 days for the best seller film are approximatively constant, meaning that some marketing actions for example could be undertaken so that the real future sales are higher than predicted.

SOURCE

The Dataset was source from Kaggle Cinema Dataset