Prediction Features to Increase Cinema Revenue

Trish Girmus

6/5/2021

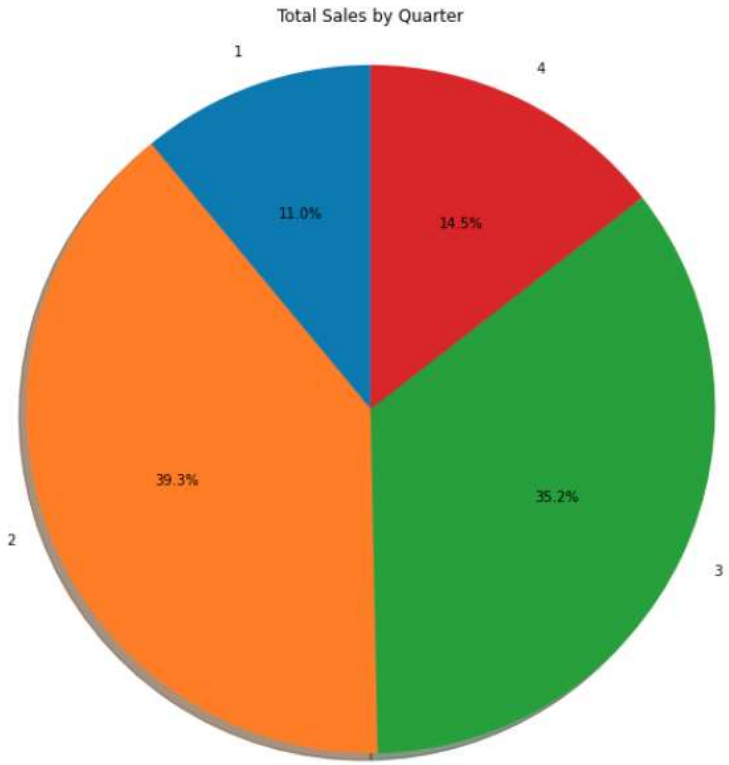
For this project, a dataset was selected that would help align with my current work as an Inventory Specialist/Data Analyst and give me the insight on how to apply with future projects (in my career). The topic of the dataset is cinema tickets. My decision to choose the cinema also stemmed from my love of movies as well as I wanted to work on a dataset that was a lighter topic in nature.

As I also love money, and work in a sales department, I am passionate about generating and growing revenue. My business problem for this dataset was, how can we predict/generate more revenue, or total sales based on the data at hand? For this business case, I am presuming life is “normal” again, and that we are not living through a pandemic.

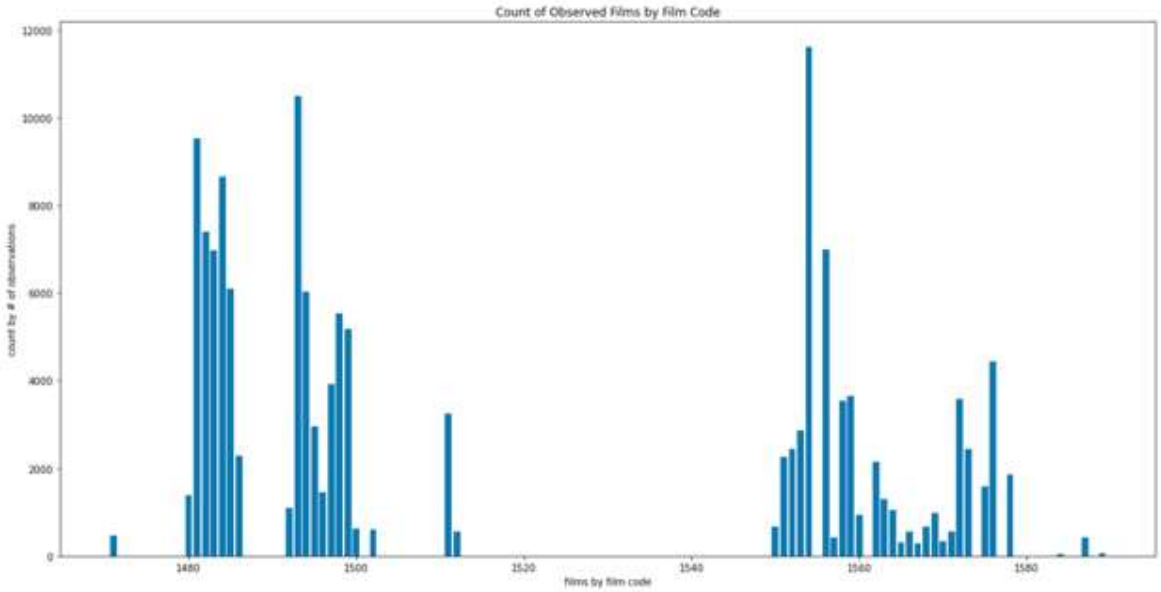
The dataset I selected is located on Kaggle and is titled cinemaTicket_Ref. The data covers roughly eight months of data in the year 2018, from March-November. It contains 142,524 rows and 14 columns of data. The 14 variables in this dataset include the following:

- film_code
- cinema_code
- total_sales
- tickets_sold
- tickets_out
- show_time
- occu_perc
- ticket_price
- ticket_use
- capacity
- date
- month
- quarter
- day

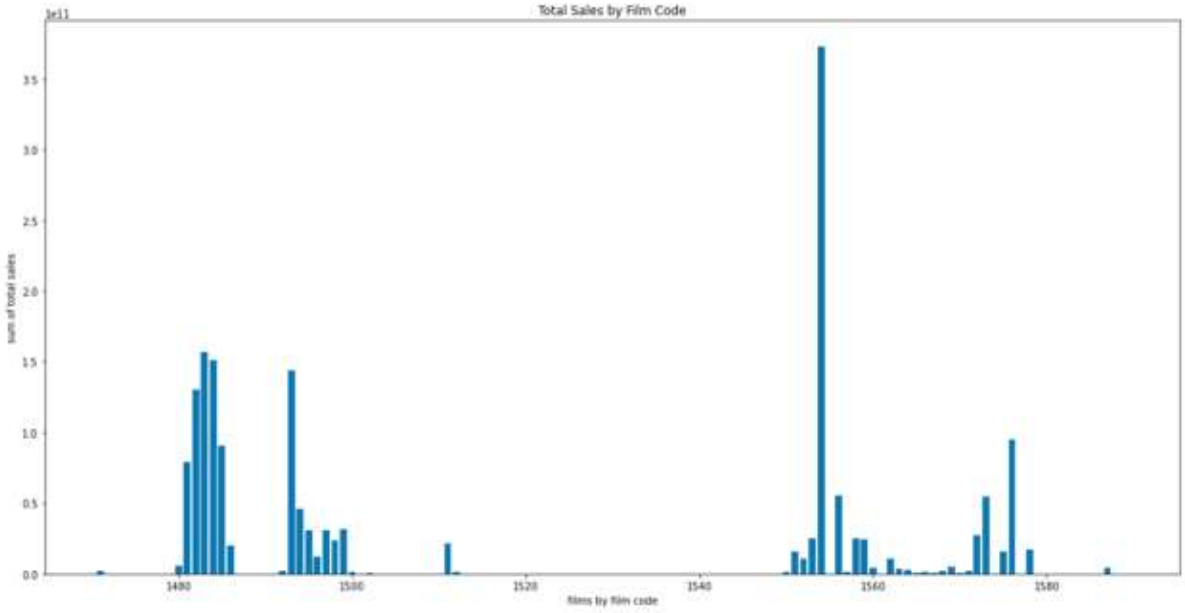
I performed some visualizations to begin to get a better idea of what the data told. In this graph, I did an analysis using a pie chart to determine total sales by quarter. As you will see below, almost 75% of total sales occur in the 2nd (39.3%) and 3rd (35.2%) quarters of 2018. Traditionally, the blockbuster movie premieres occur in the summer, so this was an interesting feature to review.



A second graph below shows a count of observed films according to film code. Initially, I wanted to review the film codes to determine which ones had the highest counts. As you can see, film codes 1493 and 1544 had the highest counts over other film codes, which indicates they were popular movies. Unfortunately, the author of the dataset chose to make the cinema titles anonymous and encoded the titles as these codes instead, which ultimately led me to drop this feature.



Another graph which I felt initially that was pertinent with this business case was total sales by film code. I believed this would show important information as to what films generated the highest total sales. As you can see, film code 1544 had the highest total sales. But as I mentioned in the previous paragraph, since the film codes were encoded, we do not know what the movie is and is not helpful with solving this business case as a result. I am also including this graph as I will elaborate on this in a future section of this paper, but the total sales variable is unknown, as far as unit of measurement. Could this be American dollars, or a foreign currency? According to imdb.com, *Avengers: Infinity War* had 2.05B worldwide total of sales, which is the highest of all total sales for 2018 that I have seen. The graph below contradicts this information with film code 1544 plotting a higher total.



The next phase of this project focused on preprocessing the data, feature reduction and feature engineering for modeling. I identified during preprocessing the data that there were 125 rows of missing data which I removed. I did this versus altering the data as it would not have been accurate with the rest of the analysis. I now have 142,399 rows of data.

For feature reduction and engineering, I identified the following features were unnecessary or irrelevant for later model building and therefore were dropped:

- show_time → this was later added back for modeling
- tickets_out
- tickets_sold
- film_code
- cinema_code
- occu_perc

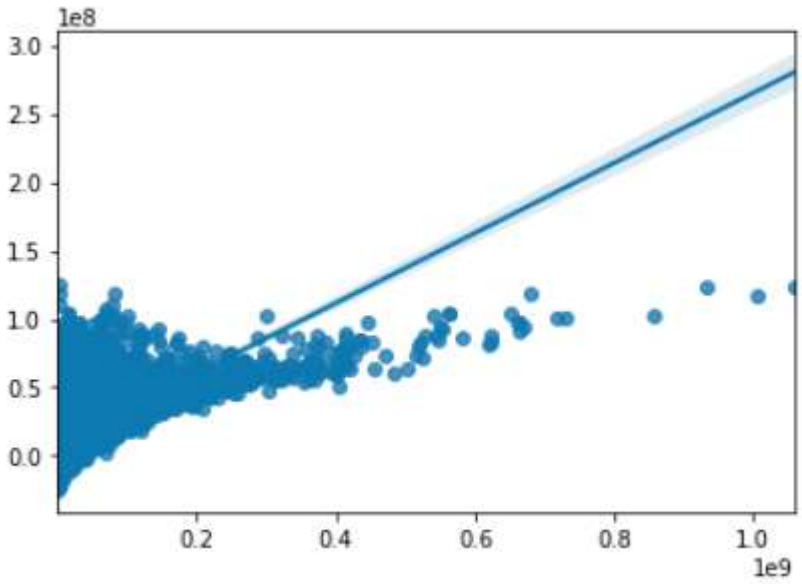
I added the features:

- majority_filled – to give a majority for occupancy over 50% with yes, under 50%, then no. (this feature was not included for my model)
- day_of_week – to show what day of the week had a higher or lower balance. This will help determine what day had higher ticket sales and tickets sold. I used dummy encoding to convert this categorical data into numbers for modeling.
- season – to determine when total sales and ticket sold to help with prediction of generating more revenue.

For my business problem, I identified that using a multiple linear regression model would best answer my question of how to generate more revenue. My dependent variable for this model was total_sales. I identified that the following variables were initially the independent variables for my model:

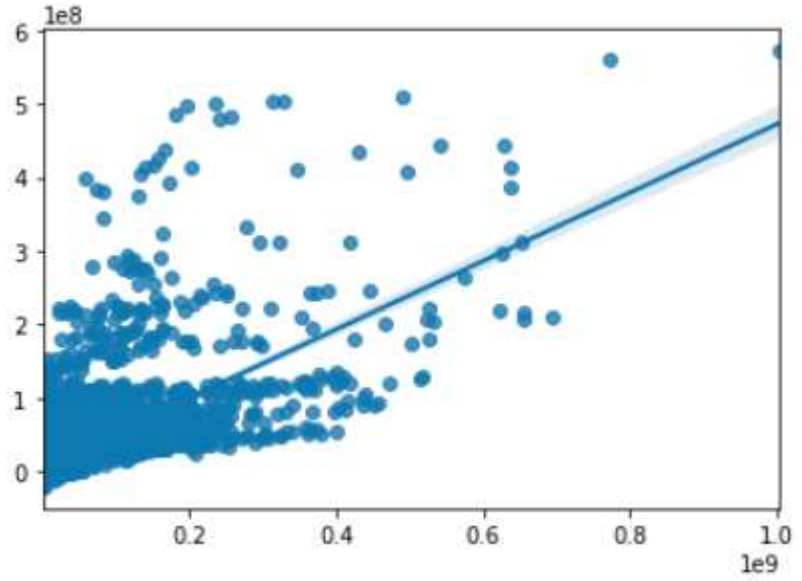
- ticket_price
- capacity
- quarter (this had to be converted into dummy variables for modeling)
- day_of_week
- season

I began the modeling process by creating a train and test dataset. I used a test size of 40% for the test dataset. When trying to predict the model, this is a visualization of what the data shows:



As you will see, this model is not good. The R-squared value was .258, which does not give a good explanation of the independent variables used for this prediction model. Again, this was using five independent variables (ticket_price, capacity, quarter, day_of_week, and season). I decided to add back the show_time feature to see if this would help improve my model.

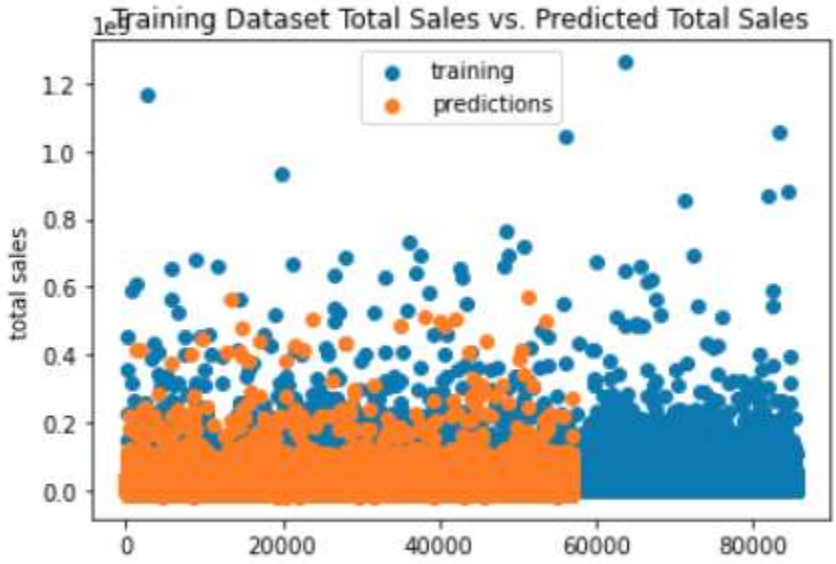
When adding back the show_time feature, I also had to dummy encode the variable and as a result, it created 51 new columns.



As you now see, the model looks somewhat improved, but not by much. The new R-squared value is .501. By adding in this additional feature, however, it does indicate that adding another feature

does help improve the model overall. The model is still not considered to be a good model, though, based on the R-squared score. A good model would indicate a score of at least 70% or higher.

I also plotted another graph to identify if outliers were present in the total sales versus the predicted sales. It appeared when plotting the test data that outliers were present. However, as you will see below, that is not the case with the training dataset. Neither the test data nor training data indicate that outliers are present.



This brings me to the conclusion of my findings with this project. Sadly, I have identified that I do not have a good model to present my business case to a group of stakeholders. In order to present my business case, I need to collect additional features to identify how to predict/generate revenue for total sales. I have several ideas for this business case. For example, how do concession sales impact total sales? What about marketing/advertising movies (film codes)? How are the theaters maintained which could determine sales over other locations (i.e., do they have air-conditioning? What are the employees' salaries? – this could impact the overall customer service experience). I have already shown that by adding back the feature show_time did improve my R-squared score, so I know that adding more features would improve my model.

Overall, I learned quite a lot while working on this project. When I began this process with even selecting a dataset, I learned that the data itself can have an impact on the analysis and answering the business case/question. I felt at the beginning my data would help solve my question, but I learned quickly through the process of the project that it hurt me more than anything. The author of the data encoded many of the features, and did not identify the values of the data, so it made it very difficult to interpret or use the data when building the model. I contacted the author several weeks ago as well as two other individuals, as we all had questions on interpretation of the units with the features. He did not reply to any of us. This was a big takeaway for me with finding a quality dataset. If the author had included the features of actual film_codes, cinema_code, show_time, and what the unit of currency was for total_sales, I feel I could have concluded with a good model and strong analysis for my business case. I know that this is all part of the learning process, but it was disappointing to have built a model did not provide a good accuracy score. I did enjoy the project and did learn a lot from it.

References

Mobius. (2020, October 29). Cinema Tickets. Kaggle. https://www.kaggle.com/arashnic/cinema-ticket.

Thedonovankidd. (2018, March 25). Highest grossing movies of 2018. IMDb.

https://www.imdb.com/list/ls064494766/.