Ticketmaster Price Predictions

Fall 2022

1 Introduction

1.1 Overview

Pricing predictions can be used to understand how to accurately and fairly price products based on previous data and feature similarities. When selling tickets, being able to predict a price is important in being able to correctly gauge market value for events. It's also important for resellers who must know the value of what they are purchasing and consumers who must be aware of a reasonable price to pay. In this assignment, I have generated a new dataset with data pulled directly from ticketmaster and employed a number of predictive models to predict ticket prices using features such as seating info, artist popularity, genre, offer types, and city of event.

1.2 Project Background

The inspiration for this project stemmed from this month's highly publicized Ticketmaster Taylor Swift Era's tour debacle. For those unfamiliar, the situation can be explained here:
https://business.ticketmaster.com/business-solutions/taylor-swift-the-eras-tour-onsale-explained/

I had hoped to pull pricing data for the Taylor Swift Era's Tour with the intent to predict pricing based on factors such as seat selection, city, offer type, and if possible the dynamic pricing feature; dynamic pricing raises prices as inventory decreases. Unfortunately, Ticketmaster (smartly) shut off API developer access to the events. Instead, I turned to other popular artists who were touring and started building my dataset.

2 Dataset

2.1 Overview

The dataset used in this analysis was obtained primarily from Ticketmaster using the Ticketmaster developer API with additional features sourced from Spotify. Ticketmaster is the world's largest ticket distribution company, essentially holding a monopoly on ticket distribution in the United States. The dataset contains ticket pricing for concerts with information such as artist, artist genre, artist popularity, city concert is being held in, concert venue, pricing zones, offer types, and seating areas. Artists were selected objectively based on currently touring artists promoted by Ticketmaster.

Artist	Monthly Listeners	Popularity Bin	Genre
The Weeknd	79,051,730	8	Hip-Hop/Rap
Ed Sheeran	75,868,690	8	Pop
Joji	33,434,420	4	R&B
Jack Harlow	31,900,694	4	Hip-Hop/Rap

P!nk	27,305,507	3	Rock
Lizzo	25,520,081	3	R&B
The Killers	18,454,651	3	Rock
Eagles	17,991,705	2	Rock
Paramore	16,305,005	2	Rock
Blink 182	15,632,660	2	Rock
Nelly	15,460,608	2	Hip-Hop/Rap
Bruce Springsteen	15,457,577	2	Rock
Chris Stapleton	12,443,131	2	Country
Shania Twain	11,091,990	2	Country
Carrie Underwood	8,400,353	1	Country
Phoebe Bridgers	8,383,731	1	Rock

Table 1: Artist Categorization

2.2 Data Cleansing

Data with a currency value other than USD were removed to avoid having to factor in an exchange rate. Tickets with prices of $0.00 were removed, as well as 114 observations with ticket prices higher than $950 which were deemed as outliers. The resulting dataset was N = 26,303.

Statistic	Ticketmaster Dataset
# of Artists	12
# of Concerts	190
# of Price Zones	32
# of Seating Areas	28
# of Offer Types	104
# of Locations	89
# of Venues	114
Minimum Ticket Price	$7.00
Maximum Ticket Price	$950.00
Average Ticket Price	$233.18

Table 2: General cleaned dataset statistics

2.3 Data Exploration

Data exploration was done in order to determine potential features. Features such as area and pricing zones showed a potential linear fit. An interesting property of the data is that Shania Twain tickets account for 28.9% (N = 7652) of the data points. This can be attributed to the number of shows Shania Twain is playing in addition to the number of ticket variations offered for her shows. Popularity bin 2 and the rock genre showed the highest priced tickets, however this could be skewed as the dataset is primarily made up of these two groups. This was not intentional, but due to current artists touring within the states. Weekend tickets - Friday, Saturday, and Sunday - were the highest day of week tickets. Graphs are included in the appendix.

3 Predictive Task

In this assignment, I predicted ticket prices for concerts. While I explain more in depth in Section 4 Model Selection, my initial instinct was to use regression when modeling this problem as I was predicting continuous values.

3.1 Evaluation Metric

To evaluate my predictions, I calculated the Mean Absolute Error (MAE) between my predictions and test data. I came to the decision of using MAE over Mean Squared Error (MSE) as my data is not uniformly distributed and has a sparse amount of higher priced tickets which will be difficult to predict. MAE allows for a longer tail of errors that MSE would severely punish my model for. As a baseline, I'll look at the MAE of predicting the mean ticket price. My train, validation, and test split was 80:20:20. The validation set allowed me to tune my model before final predictions and avoid potential overfitting.

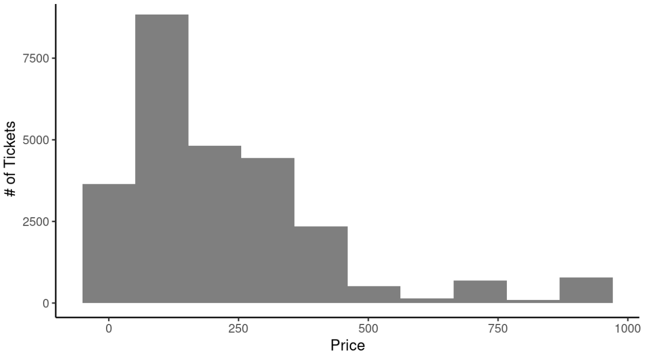


Figure 1: Ticket Price Distribution

3.2 Feature Selection

When creating my dataset, I selected a number of features that I believed to be most influential in determining ticket price. For artists included in this dataset, I pulled all upcoming concerts and recorded the concert city, venue, and date. I created an additional day of week feature from the date, with the belief that events held on the weekend might be more expensive. Using the concert id, I was able to pull pricing for all variations of seating (price zone and area) and offer codes at each concert. Artists were bucketed based on monthly listeners with bin sizes of 10 million to create the popularity feature; there are 9 possible bins, as the artist with the highest monthly listeners on Spotify is Taylor Swift with 84.1 million monthly listeners. The final features used in the model include: *price_zone, area, offer_type, city, venue, DOW (day of week), artist popularity, artist genre.* Price zone and area refer to the seating area within the venue. Offer types were codes that indicated seating characteristics such as ADULT, AISLE, VIP, VIP1,VIP2, etc. Once I had my final dataset, I created dummy variables for all of my features. My resulting feature vector had 318 features.

4 Model Selection

4.1 LightGBM Regressor

LightGBM is a gradient boosting framework that uses tree based learning algorithms to solve a number of machine learning tasks, including regression. As discussed in the *Personalized Machine Learning* textbook:

> *Decision trees straightforwardly facilitate learning non-linear classifiers that capture complex interactions among features, e.g. we can straightforwardly learn that a low price is associated with a positive review for young people, while a high price is associated with a positive review for older people: such an association is difficult for a linear classifier to learn if neither the ‘age’ nor ‘price’ feature is individually correlated with the outcome.* (McAuley 53)

I tried the LightGBM model hoping to pick up some complex interactions between my features; for instance, perhaps we find ticket price for a certain genre fluctuates depending on which city the concert is in, something which would be difficult for a standard linear model to pick up on. I ultimately selected this model as I had the ability to tune hyperparameters, which resulted in a significantly decreased MAE.

I optimized my model by performing hyperparameter tuning using 3-fold cross validation using RandomizedSearchCV. Cross validation is used to reduce bias and avoid overfitting. Parameters in the final model were *num_leaves = 20, learning_rate = 0.05, max_depth = 10, min_data = 3, n_estimators = 500, bagging_fraction = 0.9*.

Parameters	Interpretation
num_leaves	This is the maximum tree leaves for base learners.
learning_rate	This controls the speed of iteration.
max_depth	This describes the maximum depth of the tree. It is capable of handling model overfitting.
min_data	This is the minimum number of the records a leaf may have. It is also used to deal with overfitting.
bagging_fraction	This specifies the fraction of data to be used for each iteration and is generally used to speed up the training and avoid overfitting.

Table 3: LightGBM Parameters (Sun et al.)

Advantages: LightGBM allows for faster training and high efficiency while maintaining low memory usage. Tree based models can understand complex interactions that are overlooked by standard linear regressions. LightGBM is also compatible with large datasets.
Disadvantages: Because LightGBM splits the tree leaf-wise as opposed to depth-wise it can be highly sensitive to overfitting and is best not used with small datasets.

4.2 Other Model Attempts - Linear / Ridge / Gamma Regression

The first model I used was a simple linear regression. Linear regressions are often the simplest model when predicting a continuous variable and assume a relationship between features X and labels y as $y = X\theta$.

I also tested a ridge regression which is a regularized linear regression used when data might suffer from multicollinearity using an alpha = 1. I believe my ridge regression performed similarly to my initial linear regression model as there wasn't high correlation between the features I was using. Both the standard linear regression and ridge regression gave MAEs significantly better than the baseline but not comparable to LightGBM.

Finally I tried a generalized linear model (GLM) for a gamma distribution. This model type and distribution specification is typically used in modeling continuous, non-negative and positive-skewed data. As ticket prices cannot be negative and data was skewed towards the lower priced tickets, I thought this would be better suited than a typical linear regression. This model ended up performing the worst.

Advantage: Linear regressions are simple to implement and easy to understand when predicting continuous variables. Linear regressions can be modified to handle non-uniform data, e.g. gamma distribution takes into account the 0 bound which was present in my dataset.
Disadvantages: Linear regressions can only represent linear relationships so they are not successful at picking up complex interactions between features if they do not directly correlate to the outcome.

5 Relevant Literature

While I didn't find much literature relating specifically to ticket pricing prediction, there are a number of kaggle datasets and papers on using machine learning methods to predict housing prices, stock prices, cryptocurrency price trends, and more.

5.1 Price Prediction Using Machine Learning Regression

This *Towards Data Science* article (Kumar) focuses on a dataset from Mercari, one of the biggest community-powered shopping websites in Japan. The goal of the project was to use machine learning regression models to allow Mercari to automatically suggest the right product price to sellers on the app. The post used root mean squared logarithmic error (RMSLE) as its performance metric and discussed data cleansing and feature engineering before evaluating four different models: a ridge regressor, SVM regressor, random forest regressor, and lightGBM regressor. In this article, the lightGBM model performed the best and is what drove me to investigate the lightGBM regression model.

5.2 A novel cryptocurrency price trend forecasting model based on LightGBM

This (Sun et al.) is one of the first articles I found when investigating lightGBM regression models. They were using lightGBM to forecast cryptocurrency market trends. The paper primarily uses lightGBM as a way to classify if the cryptocurrency market is falling or rising based on prices, but also discusses LightGBM as a tool for different types of machine learning tasks, such as classification, regression, and ordering. The paper shows lightGBM models to have higher accuracy and better robustness in forecasting, which was also the case with my dataset.

5.3 Personalized Machine Learning Textbook

My starting point stemmed from textbook material; it also gave me the idea to try out the gamma distribution as my data was skewed. If I had had a different dataset, such as ticket purchases by individuals with purchaser characteristics, I would have used models such as BPR or a latent factor model mentioned in this book to recommend the next ticket purchased or tickets to recommend.

6 Results & Conclusions

Ticketmaster Dataset	Baseline (Average)	Linear Regression	Gamma Regression	Ridge Regression	LightGBM Regression
Validation	145.59	64.48	124.25	64.489	28.24
Test	143.39	65.73	123.89	65.722	28.74

Table 4: Performance of models in terms of MAE

Out of the 4 models attempted, the LightGBM Regression, in terms of MAE, performed significantly better with a test set MAE of 28.74, decreasing the MAE 56.2% compared to the next best model. An MAE of 28.74 means that on average our pricing prediction errors were around $28.24. The dataset ranged in price from $7 to $950 and was not evenly distributed making predictions more complex, so I was happy with this outcome. While the linear and ridge regression performed much better than the baseline, I believe what set the LightGBM regression model apart was its ability to detect complex relationships within the data.

Features such as price zone (Appendix Figure 2) and area (Appendix Figure 3) did show trends that were useful in a linear model and the Day of Week feature (DOW) (Appendix Figure 4) showed concert tickets to be significantly higher on weekends, which made sense. I originally believed we would see a trend between popularity score and ticket pricing, however I saw the largest range of ticket prices in the 2nd popularity bin. I think this is due to dataset construction and the overwhelming amount of data that fell into the 2nd popularity bin.

If I was to redo this analysis, I would start by creating a more balanced dataset. I limited my dataset collection to well known artists who were currently touring, but it ended up providing a disproportionate number of artists in the same genre and popularity bin. In reality, all of these artists are at a level in which they can command high ticket prices and the pricing bins I assigned might have little effect on pricing at this level. Another feature that could have been improved was offer types. There were 104 unique offer types, this feature might have been a better predictor if grouped into more generalized offers.

Citations

Kumar, Arun. “Price Prediction using Machine Learning Regression — a case study.” *Towards Data Science*, https://towardsdatascience.com/mercari-price-suggestion-97ff15840dbd. Accessed 26 November 2022.

McAuley, Julian. *Personalized Machine Learning*. Cambridge University Press, 2022.

Sun, Xiaolei, et al. “A novel cryptocurrency price trend forecasting model based on LightGBM.” *Finance Research Letters*, vol. 32, no. 101084, 2020, p. NA. *Science Direct*, https://www.sciencedirect.com/science/article/pii/S1544612318307918#tbl0001.

Appendix

Figure 2: Scatter Plot - # of Observations at Ticket Price vs Price Zone

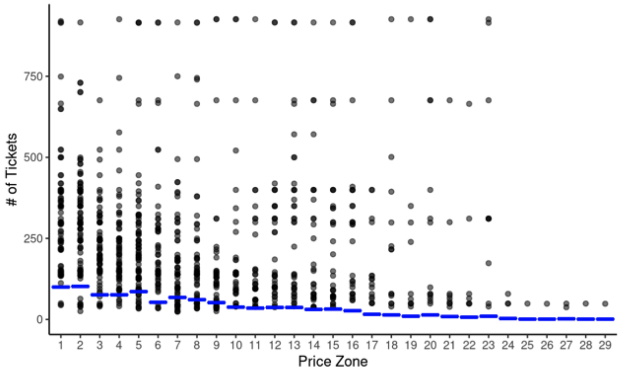


Figure 3: Scatter Plot - # of Observations at Ticket Price vs Area

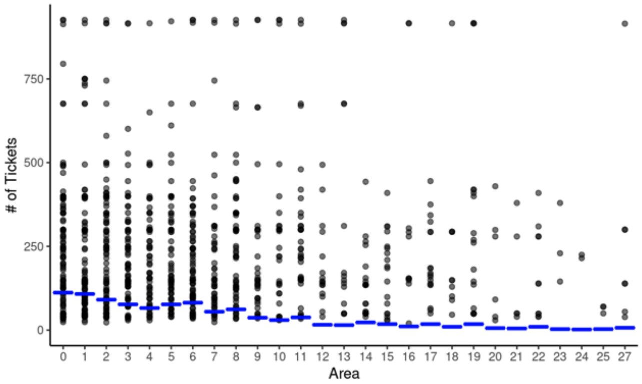


Figure 4: Scatter Plot - # of Observations at Ticket Price vs DOW

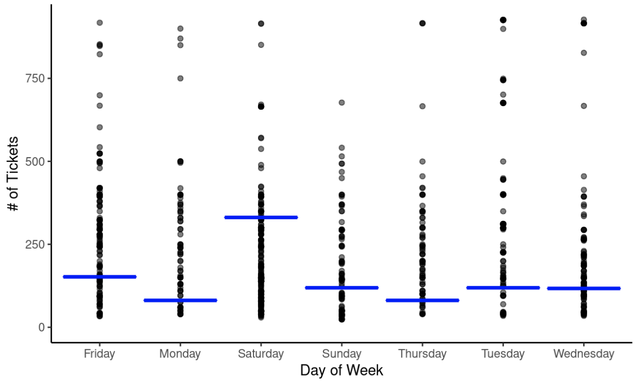


Figure 5: Scatter Plot - # of Observations at Ticket Price vs Popularity Bin

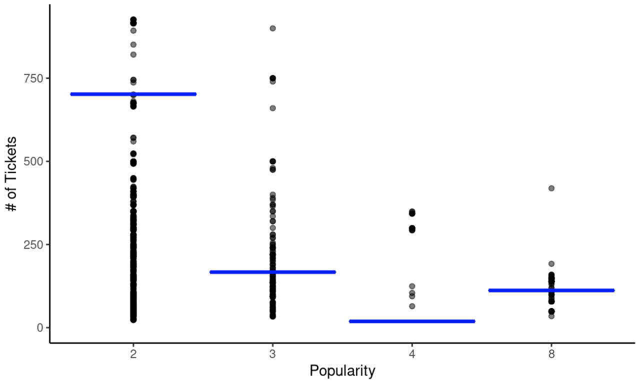
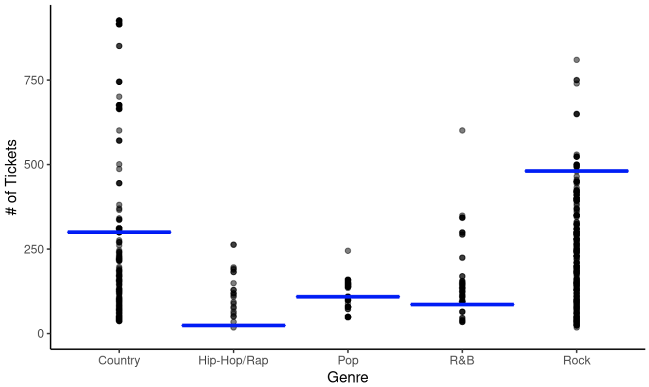


Figure 6: Scatter Plot - # of Observations at Ticket Price vs Genre



750
500
250
0
# of Tickets
Country
Hip-Hop/Rap
Pop
R&B
Rock
Genre