Applying Machine Learning Algorithms to Predict Ticket Prices using Concert and Artist Information

Ryan Rawitscher, Matt Engelken
Ryanrawitscher2019@u.northwestern.edu
EECS 349 Machine Learning Northwestern University
links: Dataset Analysis in R Data Collection

Value Proposition

A learner that can predict ticket prices for a given concert can be used in several ways. First and foremost, it can help give a better information of the pricing algorithms that vendors like Ticketmaster use and strategically buy tickets that would be seen as a "good deal" when compared to our model. The second application would be in budgeting for future concerts, such as when an there are no shows for a particular artist but our algorithm could help give a ballpark estimate for the price that attendees would be expected to pay. Third and perhaps most interestingly, our model is the first step at diagnosing which ticket vendor boasts the most customer friendly pricing algorithm. This information is incredible useful in helping consumers choose which vendor to buy event tickets from. Lastly, our model can be used as a music faced cost measure on city and genre of music for these events.

The Dataset

The dataset was collected from the Ticketmaster and Spotify APIs. Our set contains a list of the top 100 most popular artists (using billboard charts) and searched Ticketmaster for 20 of their upcoming shows. These key bits of information were collected for every event:

- Performance city and city population (population thanks to opendatasoft.com API)
- Artist Name and artist popularity score (Via Spotify API algorithm)
- Whether she show was performed on a weekend or not (binary)
- Month of performance (1-12)
- Genre of music being played
- Minimum ticket price offered for that event

After the data was pulled it needed to be cleaned. The first issue the dataset encountered was that some of the events had egregiously high prices since the artists were part of a multi-day music festival. These events were removed from the data. The second problem was that many of the cities actual populations were small but the surrounding area contained the majority of the audience on a regular basis (like Gillette Stadium in Foxborough, Massachusetts). To address this, events hosted only in America's top 1000 cities were kept. Lastly, the artist name was dropped from the dataset so that we could broaden the scope of our results.

Methodology

To get an idea for which models may be appropriate for the dataset, Weka trained the dataset on many different models and used 10-fold cross validation to test the accuracy. The models that performed the best were lbk, random tree and random forest. Next, the dataset was imported into R studio, a statistical machine learning program. After splitting the data into

90% training and 10% test sets, the data was ready to be fit. In order to decide which model was most appropriate for this predictive learner, the mean squared error on the test set should be compared. A Liner Regression, forward and backward stepwise regressions, ridge and lasso regressions, General Additive Model, decision tree, overgrown decision tree, random forest, random forest with bagging, and a boosted tree were all fit to the dataset for testing. The resulting Mean Squared Error measurements can be seen in Figure 1. The model that performed the best was the random forest with bagging (See Figure 3), achieving a mean squared error $33 less than the original random forest, the next best model. This indicates that the random forest, and likely many other of the models, suffer bad prediction accuracy due to high variance in the data and overfitting to the training set. The best predictors in the random forest with bagging were the population of the city that the concert was being performed in and the artist popularity score, as evidenced in Figure 2. This indicates that the popularity of the artist and the city of the concert were better at predicting the price of an event than the genre of music, the month of the performance, or the day of the week that the performance is being held.

Future Work

The next step in this process would have to be to use the Eventbrite, Eventful, Songkick, and SeatGeek Api to make similar analysis on the concert ticket pricing strategies. A more robust pricing model could help consumers find better concert ticket deals in the future. One important statistic that Ticketmaster lacks is the capacity of the concert venue, which might be an important indicator of how expensive the tickets are. Lastly, a more descriptive measure of concert tickets would likely be the median ticket price, which was not available via the Ticketmaster API.

Who Did What?

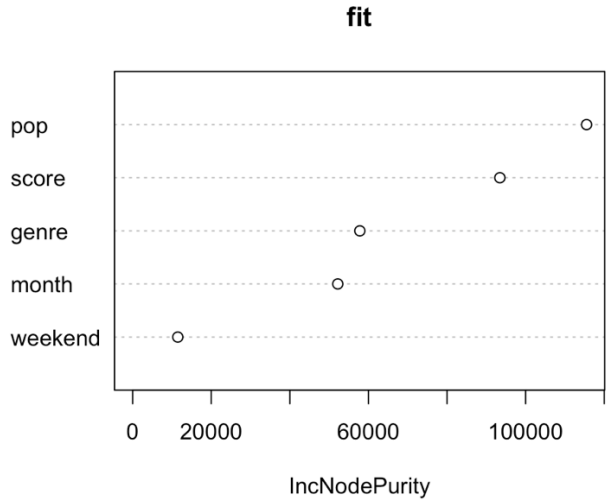
Data Collection—Ryan Rawitscher
Data Analysis –Ryan Rawitscher
API Research—Matthew Engelken
Website Development—Matthew Engelken
Website Design—Matthew Engelken & Ryan Rawitscher

Figure 1: Model type and resulting Mean Squared Error

Model	Mean Squared Error
Linear Regression	296.3087
Stepwise Model	296.5095
Ridge	302.556
Lasso	300.4573
General Additive Model	283.8466
Decision Tree	348.4048
Overgrown Tree	384.4649
Random Forest	248.0282
Random Forest with Bagging	215.9549
Boosted Tree	289.4104

Here the models applied to the dataset are compared using a test statistic, Mean Squared Error. The lower the MSE, the better the model.

Figure 2: varImpPlot of the Best Model, Random Forest with Bagging.



This model shows the difference between Residual Sum of Squres before and after the split on each variable.

Figure 3: Best Model, Random Forest with Bagging

```
Call:
 randomForest(x = yz[train, -1], y = shows$minprice[train], xtest = yz[!train,      -1], ntree = 100, mtry = 2)
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 2
```

This model was the most accurate in trying to predict the minimum ticket price of concerts from our dataset.