Jackson King and Boone Tison

12/6/2021

CSC-372

Dr. Treu

Final Project: Movie Predictions

**Training Data:**

We started with the box office numbers to get our attributes, a class attribute of Gross and input attributes of Distributor and Number of Theatres. We then used IMDb to add more attributes, Budget, IMDb rating, Genre, MPAA Rating, Country, and Run Time. For our data, we added all the movies that were released in the same week from 2010-2019. We decided to ignore 2020, as it was an outlier in terms of movies released and their gross due to COVID. We also decided to initially ignore movies that were only released in 1-2 theatres. In the second version of our training dataset, we decided to add the movies that were released in 1-2 theatres. We also added a nominal version of Gross. This new Gross has the values of Worst, Poor, Good, and Best. Each of these, representing 25% of the numeric gross. In the final version of our dataset, we saw that a lot of our movies had missing values for the budget, especially documentaries and dramas that were in a couple of theatres. We decided to fill in their budgets with the average budget of other movies with similar genres, number of theatres, and gross. Other genres had 1 or 2 missing budget values, but we decided not to fill those in as none of the movies had other similar movies that we could average.

**Linear Regression:**

|  | Licorice Pizza | Encanto | House of Gucci | Resident Evil | Writing with Fire | For the Love of Money |
| --- | --- | --- | --- | --- | --- | --- |
| All attributes | -86,322,663 | 5,227,595 | 126,295,871 | 56,394,398 | 7,974,850 | -40,490,700 |
| Only Numeric | 2,622,516 | 26,842,635 | 52,332,432 | 25,667,659 | -15,781,599 | -127,839 |
| Only Nominal | -3,913,953 | 33,914,369 | 7,808,453 | 13,239,800 | 2,617,761 | -12,112,753 |
| Dist + Budget + Theatres | -49,113,534 | 14,128,621 | 27,513,573 | 12,804,663 | -1,466,650 | -51,127,699 |
| Actual | 537,798 | 45,220,641 | 26,687,819 | 10,518,747 | 5,551 | 354,086 |

The all attributes regression model used every attribute for the independent variables.

R-Squared: 0.741 MAE: 53,495,216.21

The numeric regression model only used the attributes that were originally numeric, Theatres, Budget, IMDB, and Run Time

R-Squared: 0.459 MAE: 12,890,887.66

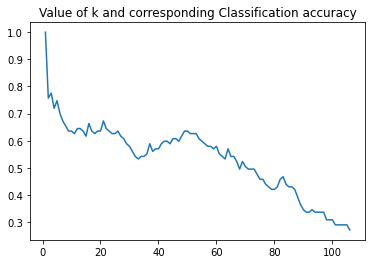
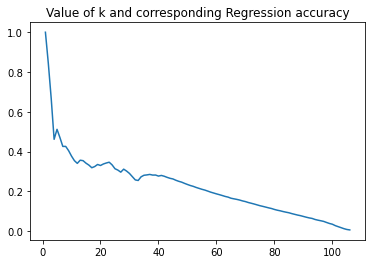
The nominal regression model only used the attributes that were originally nominal and then turned into numeric using one hot encoding, Distributor, Genre, Rating, and Country

R-Squared: 0.539 MAE: 8,769,582.03

The last model only used Distributor, Budget, and Theatres. Budget and Theatres were selected because they were the only numeric attributes with a correlation coefficient over 0.5 with Gross. Distributor was selected because it was the only nominal attribute that seriously affected the R-Squared when removed from the nominal only model.

R-Squared: 0.666 MAE: 22,771,501.76

**KNN Regression and Classification:**



|  | Licorice Pizza | Encanto | House of Gucci | Resident Evil | Writing with Fire | For the Love of Money |
| --- | --- | --- | --- | --- | --- | --- |
| Reg n=2 | 994,234 | 109,583,914 | 8,818,922 | 16,766,663 | 22,755 | 20,928 |
| Reg n=3 | 2,100,037 | 82,560,199 | 7,266,738 | 11,197,917 | 17,169 | 16,736 |
| Class n=2 | Good | Best | Best | Best | Poor | Poor |
| Class n=3 | Good | Best | Best | Best | Worst | Worst |
| Actual Reg | 537,798 | 45,220,641 | 26,687,819 | 10,518,747 | 5,551 | 354,086 |
| Actual Class | Good | Best | Best | Best | Worst | Poor |

KNN Regression model with neighborhood size of 2

MAE: 15,244,480.75

KNN Regression model with neighborhood size of 3

MAE: 9,921,836.05

KNN Classification model with neighborhood size of 2

Best [[3 0 0 0]

Good [0 1 0 0]

Poor [0 0 1 0]

Worst [0 0 1 0]]

KNN Classification model with neighborhood size of 3

Best [[3 0 0 0]

Good [0 1 0 0]

Poor [0 0 0 1]

Worst [0 0 0 1]]

**Decision Tree:**

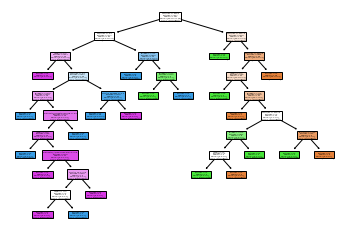
|  | Licorice Pizza | Encanto | House of Gucci | Resident Evil | Writing with Fire | For the Love of Money |
| --- | --- | --- | --- | --- | --- | --- |
|  | Poor | Best | Best | Good | Poor | Poor |
| Actual | Good | Best | Best | Best | Worst | Poor |

Best [[2 1 0 0]

Good [0 0 1 0]

Poor [0 0 1 0]

Worst [0 0 1 0]]



**Neural Network Regression:**

| Layers | MSE | RMSE | MAE |
| --- | --- | --- | --- |
| 3 layers: 10, 5, 1 | 479,356,770,058,240 | 21,894,217.731 | 13,878,006.22 |
| 2 layers: 10, 1 | 479,284,661,583,872 | 21,892,570.922 | 13,913,535.51 |
| 2 layers: 5, 1 | 479,427,637,018,624 | 21,895,836.066 | 13,915,609.18 |
| 5 layers: 10, 8, 6, 5, 1 | 472,387,447,422,976 | 21,734,476.010 | 13,371,190.71 |
| 5 layers: 30, 8, 6, 5, 1 | 364,313,051,136,000 | 19,086,986.434 | 12,176,200.4 |
| 7 layers: 30, 20, 10, 8, 6, 5, 1 | 199,650,799,779,840 | 14,129,784.138 | 12,110,085.63 |
| 9 layers: 50, 40, 30, 20, 10, 8, 6, 5, 1 | 239,750,157,959,168 | 15,483,867.668 | 11,399,032.45 |
| 7 Layers: 50, 40, 30, 20, 10, 5, 1 | 192,556,520,439,808 | 13,876,473.631 | 13,129,890.03 |

**9 layers: (Best Neural Network)**

| Movie | Predicted | Actual |
| --- | --- | --- |
| Licorice Pizza | $36,546 | $537,798 |
| Encanto | $56,159,600 | $45,220,641 |
| House of Gucci | $49,123,930 | $26,867,819 |
| Resident Evil | $39,097,450 | $10,518,747 |
| Writing With Fire | $25,548 | $5,551 |
| For the Love of Money | $6,453,258 | $354,086 |

**Discussion of Results:**

**Linear Regression:**

The first linear regression model used all of the input attributes. Then, feature selection was done to determine if any input attributes were negatively affecting performance. A model was created using only the numeric input attributes, then one by one, each of those was removed. In the end, the numeric model with all of the numeric attributes was the best, but worse than the all attributes model. Next, a model using only the nominal values was created. The nominal values also had each attribute removed one by one, but again the all nominal was best, but also worse than the all attributes model. For the final feature selected model, the numeric attributes with a correlation coefficient over 0.5 with Gross were used, Budget and Number of Theatres. In addition, Distributors was used as it was the only attribute that significantly affected the nominal only model when removed. This final feature selected model had a worse r-squared than all attributes model, but was better than nominal or numeric only. In terms of R-Squared, the best model was the all attributes model. Any of the attributes being removed caused the R-Squared to decrease. However, in terms of predicting our test values, the nominal only model had the lowest mean absolute error. So while it has a lower r-squared, we think the nominal only regression model is the best.

**KNN Regression:**

To determine the optimal neighborhood size, a KNN model was run from size 1 to size 107. The optimal neighborhood size was determined to be 2 or 3. Size 3 has the lower mean absolute error for the testing data, while size 2 has the better accuracy on the training data. We think neighborhood size of 3 is the best KNN regression model.

**KNN Classification:**

To determine the optimal neighborhood size, a KNN model was run from size 1 to size 107. The optimal neighborhood size was determined to be 2 or 3. These two sizes had the best accuracy on the training data. Each size got 5 out of the 6 test movies right. Size two predicted a “worst” as “poor” and size three predicted a “poor” as “worst”. We believe that both KNN classifications models are equally as good.

**Decision Tree:**

The decision tree was constructed using entropy and all of the input attributes. The tree has a 50% accuracy on the test movies. The tree predicted a “good” as “poor”, a “best” as “good”, and a “worst” as “poor”.

**Neural Network Regression:**

The neural network regression model was created with the class attribute being “Gross” to predict the gross profit of the test data movies. The data used was our training data set, where the data is cleaned by removing any NA’s and the columns “Movies” and “Gross Nominal”. We then raen one hot encoding on “Distributor”, “Genre”, “Rating”, and “Country” to turn the nominal attributes into numeric. Once the data was preprocessed correctly, the neural networks’ layers were chosen. In the models with less than 5 layers, the mean squared error was very high, averaging 479,356,350,000,000. The root mean squared error was also very high, with the lowest of the first 3 models being the 2-layer model at 21,892,570.922. Again, with these three models, the mean absolute error is the highest, averaging 13,902,383.64. As the layers increase, the MSE, RSME, and MAE decrease. The lowest MSE and RMSE were from the 7-layer model, with MSE = 192,556,520,439,808 and RMSE = 13,876,473.631. The best-fitting neural network based on the predicted gross profit and MAE was the 9-layer model with an input layer of 50, hidden layers of 40, 30, 20, 10, 8, 6, and 5, and an output layer of 1. The MSE = 239,750,157,959,168, RMSE = 15,483,867.668, and MAE = 11,399,032.45. Although this model has a higher MSE and RMSE than the lowest model, the MAE is the lowest. There are a high number of layers in this model making it complex and at risk of overfitting; however, because our data set is created to be similar to the testing data set, overfitting is not a concern. This model predicted the large-profit movies well and was on scale for, but over-predicted the lower-grossing movies.

**Analysis by movie:**

* Licorice Pizza is an MGM-distributed movie in 4 theaters, a budget of 40 million dollars, and an IMDB score of 8.7. The genre is comedy, rated R, made in the USA, and has a runtime of 133 minutes. The gross was $537,798 and “good”. The best models for predicting Licorice Pizza were KNN regression and classification. The regression model for neighbors = 2 predicted a gross profit of $994,234 with an actual gross profit of $537,798. The classification model with neighbors = 2 and neighbors = 3 predicted the nominal attribute “gross nominal” to be good, representing the second-highest gross profit, which is correct.
* Encanto is a Walt Disney-released movie in 3,980 theaters, a budget of 50 million dollars, and an IMDB score of 7.7. The genre is Animation, rated PG, made in the USA, and has a runtime of 102 minutes. The gross was $45,220,641 and “best”. The best models for predicting Encanto were nominal linear regression, KNN classification of N = 2 and N = 3, both predicting “best”, and neural network regression. The best model overall was the 9-layer neural network ,predicting a gross profit of $56,159,600 compared to an actual gross of $45,220,641. The second best model was the nominal linear regression model, predicting a gross profit of $33,914,369, with the actual gross of $45,220,641.
* House of Gucci is an United Artists-distributed movie in 3,471 theaters, a budget of 75 million dollars, and an IMDB score of 7.0. The genre is Crime, rated R, made in Canada, and with a run time of 158 minutes. The gross was $26,867,819 and “best”. The best regression model was the feature selected linear regression model. The decision tree and KNN classification model of both sizes correctly predicted the classification gross.
* Resident Evil is a Sony Pictures-distributed movie in 2,803 theaters, a budget of 25 million dollars, and an IMDB score of 5.6. The genre is Action, rated R, made in the USA, and with a run time of 107 minutes. The gross was $10,518,747 and “best”. The best regression model was KNN regression with a neighborhood size of 3, which predicted $11,197,917. The KNN classification models of both sizes correctly predicted the classification gross.
* Writing With Fire is a Music Box Films-distributed movie in 4 theaters, a budget of 5 million dollars, and an IMDB score of 7.9. The genre is Documentary, not rated, made in India, and with a run time of 92 minutes. The gross was $5,551 and “worst”. The best regression model was KNN regression with a neighborhood size of 3, which predicted $17,169. The KNN classification model with a neighborhood size of 3 correctly predicted the classification gross.
* For the Love of Money is a Freestyle Releasing-distributed movie in 519 theaters, a budget of 4.5 million dollars, and an IMDB score of 6.1. The genre is Drama, rated R, made in the USA, and with a run time of 99 minutes. The gross was $354,086 and “poor”. The best regression model for this movie was KNN regression with a neighborhood size of 2, predicted at $20,928. The decision tree and KNN classification with a neighborhood size of 2 both accurately predicted the classification gross.

**Final Model Selection:**

The overall best regression model was KNN Regression with Neighbors = 3. This repression model has the lowest mean absolute error, meaning that it fits the test data the best.

The best overall classification model was KNN Classification with N = 2 and N = 3, with an overall accuracy on training data equal between N = 2 and N = 3. The model performed 5 out of 6 predictions correctly on test data for both N = 2 and N = 3. For neighborhood size 2, the incorrect prediction was on a movie (Writing WIth FIre) with a gross profit near the next quantile, and the model predicted one discretized bucket away. For neighborhood size 3, the model also incorrectly predicted a movie (For the Love of Money) being lower, in the “worst” bucket.