Industry Insider

Justin Session, Troy White, Savada Wilson

*Computer Science Department,Georgia State University*

1. Introduction

From the opulent display of jewelry, to the time invested into negotiating fees and clearing records, a lot of money goes behind musicians that is sourced from the big-name record labels that back them. Too often the most successful projects are the ones people least expect to do well, while others with the most accelaides and talent don’t sell well. In such a case, Data Mining can provide novel insights into the music industry via using its predictive power to foretell the changing tides and trends in genres of music, while also recognizing the musicians who’d produce the best ROI short-term. Successfully implementing Data Mining methods and algorithms to accomplish both goals are our group’s project. Therefore our project has two distinct modules: for determining ROI Random Forest Regression was used, Logistic Regression was used for determining the trends in music genres through examining the relation between number of reviews vs actual rating.

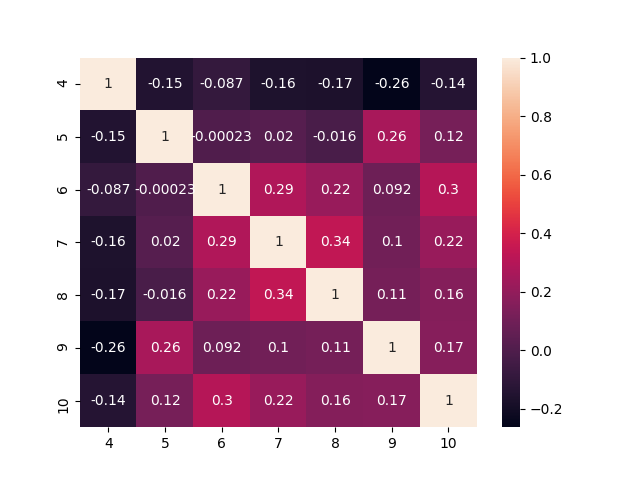
1. Random Forest Regression & Calculating Revenue

* *Getting Data*

For data to be valid in the use of calculating the target “revenue”, the target must already be present in the data along with the respective artist and body of work the revenue is a result of. This is the minimum requirement to work from. Next, one should ensure the time length the revenue is generated in is standardized for all observations. If not then one observation point might have revenue produced from 2 months vs another observation that only measured 1 month of revenue. Other features can be searched for separately but it is better to have at least these prerequisites met in any dataset under consideration. Thankfully, instead of scouring the web and piecing together single columns of data, a valid dataset was obtained from [1]. This dataset contains 11 features which are : month, position, artist, album, indicative revenue, US peak ranking, UK peak ranking, Dutch peak ranking, France , Canada , and Australia. Let it be known that all features are standardized in 1-month periods, the month after the album’s initial release. While this dataset did not make things simple, as there are many missing values, and incompatible features, it serves as a good input for our model, therefore it’s aptly called “maininput.csv”.

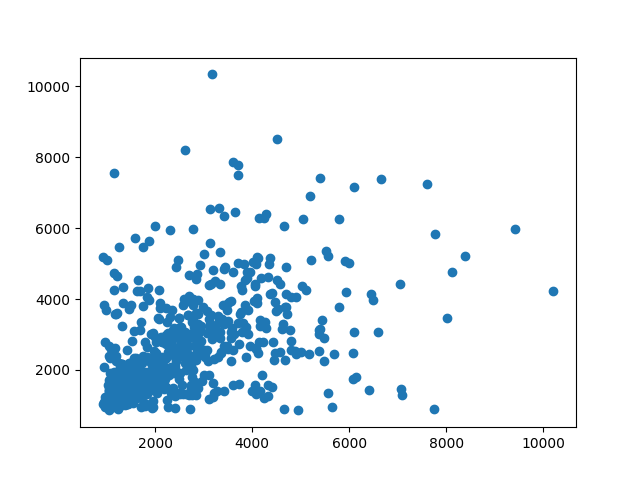
* *Data Analysis*

Before we moved on to preprocessing the data, two things were tested. Foremost we wished to see the Heat Map of the features in relation to the target variable as shown below.



The data used for this heatmap had the names of the features replaced with their row numbers where: 4 is the revenue, 5 is US, 6 is UK, 7 is Dutch, 8 is France , 9 is Canada , and 10 is Australia. On the heatmap, a negative correlation is observed between the target ,”4” and the rankings of the many countries. This is expected as a high ranking is one that approaches zero, therefore the lower the ranking, the higher the revenue. This relation is shown to be strongest with the US rankings.

The second objective that was tested is the significance of the data alone. In a Tree the Raw data was numerically encoded without processing and passed to a Decision Tree Regressor, the Actual vs Predicted Plot is shown below.

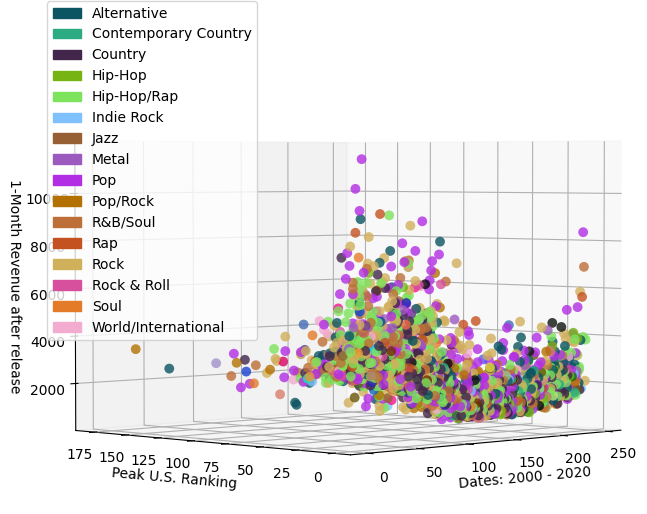


The Actual is on the X-axis and the Predicted is on the Y-axis. From this type of plot one can infer prediction accuracy of a continuous value. An actual vs predicted plot is supposed to take the shape of a diagonal line running from the bottom left to the top right, where each data point has a coordinate of “(Actual value, Predicted Value) , the closer the coordinates are to each other's value, the more accurate. Moving on, from the raw data input there was a slight formation of a line although not significant enough to warrant anything as even with an error margin of 15% the actual value, the accuracy still teteetered around 19% when testing with a Decision Tree Regressor on unseen data. Additionally, upon doing multiple iterations, a high amount of variance was observed in the location of points on the plot. As a result, the notion of simply using a Decision Tree was determined to not be sufficient and instead a Random Forest Regressor was implemented in all tests and models beyond this point. But first the data had to be processed.

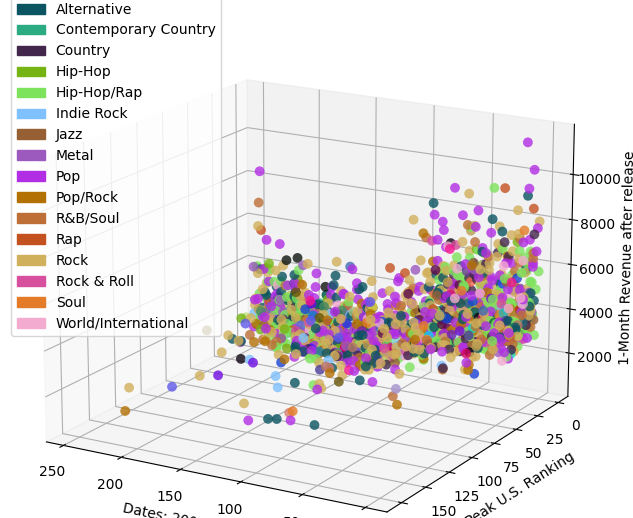
* *PreProcessing*

To start with, one of the most important aspects of music wasn’t present in the “maininput” data, i.e. the genres were missing. Therefore a large 10,000 list of artists with their respective genres was procured [2], it included superfluous data that was not selected to be a feature and was run through a filter that matched the index of the artist name in “maininput” with their respective sub genre of music. Many artists' names were not matched to their genres despite the size of the list being enormous, due to grammatical differences in the names or them simply not being present. As genres are a necessary feature, rows of artists without genres were dropped. Moving on, the dataset had many missing values in the rankings of albums. Decision Trees are known to be tolerant to missing values and make accurate predictions despite not having them, however for the best possible results one is to replace them with a close approximation. Therefore, each ranking feature was pushed through a “filter” which took the mean of the ranking which were present and replaced the missing positions with the averages. Subsequently, it was determined that the names of the musicians couldn’t be left in a single column as there’d be no way for the model to use them unless they were transformed into numbers, however even if that were done one would quickly realize that the model will ascribe weights to those arbitrary numbers to make inferences; as if artists’ names were continuous variables rather than categorical. To circumnavigate this issue, “One-hot-encoding” was implemented via the “get\_dummies” method and rectified the issue for the artist and genre features alike. Next up, the dates needed to be converted to numbers. They were converted into an incremental system where albums released the same month and year had the same value and increased by 1 in correlation with the passing months. Finally, the column “Album” which kept the names of the albums was dropped, because if one wanted to predict the potential revenue of a new album, its name wouldn’t wouldn’t be a pre-existing category as opposed to the name of the artist themself. The end result of preprocessing created a csv file “ProcessedData” with 2164 rows and 720 columns.

* *Data Analysis II*

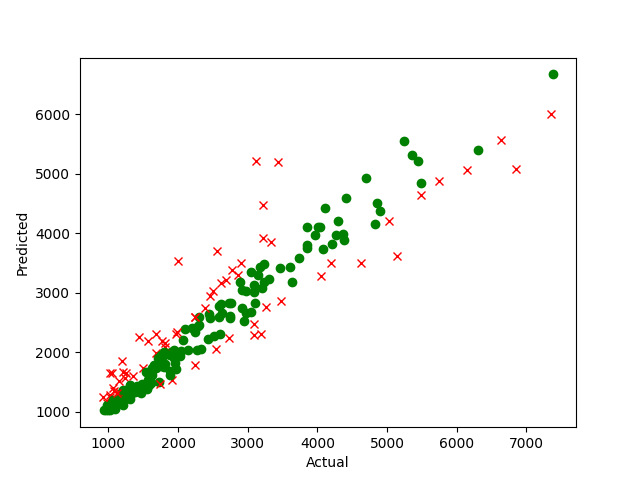
At This Stage the model in use for calculating short term ROI of an artist was a RandomForest Regressor. As exclaimed prior, RandomForest Regressors significantly cut the inherent amount of variance present between iterations of Decision Trees. The improvement was substantial but to better analyze how to increase accuracy even more, after all features were processed they were plotted once again so one could get an idea of what was being “seen” by the model and if it leaned toward a cluster that favored one feature vs another. The image below depicts the: genre, revenue, Us ranking, and Time of release for all data points in the “ProcessedData.csv” set.

In this image there are over 2000 data objects. Each one represents a single album. As one can see, there is a Downtrend in revenue from the year 2000 to the present. In regards to the music genre of the artist/album there is not a section of the graph filled with a predominant type. Albeit, it must be said that there is an observable pattern of pop music being an outlier in terms of revenue. If one were to observe the above figure and the one below while focused on the US rankings, they would see an amalgamation of different genres with a US ranking of “1”. The problem is that there is clearly variance in the “indicative revenue” despite the month or year. Therefore revenue and US rankings correlate strongly as the Heatmap indicated but, the variance is still quite high. This suggests that rankings are insufficient to calculate revenue with, and that a new feature would add a new depth or “dimension” in the analysis of revenue.

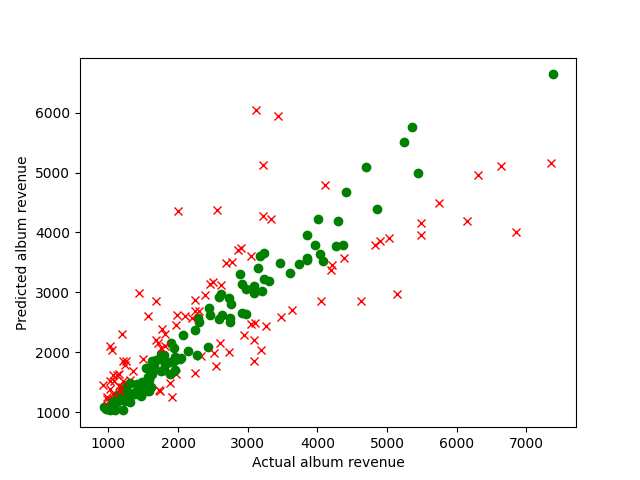


* *Results & Improvement*

The RandomForest Regressor was tested to make novel observations about how best to proceed with this model in the future. An error margin of 15% was the maximum allotted margin to still be deemed “accurate”. This had little effect on projected revenue for lower amounts, but gave a lot of slack for high revenue holding data points. The max accuracy without creating a split for testing and training teetered around 72% and produced the scatter plot below, where accurate predictions are green and inaccurate are red X’s. The plot is shown below for seen data. Like the solid color scatter plot in the first data analysis section, the desirable outcome is, X = Y for the data points plotted with a domain of X’s and range of Y’s i.e. (X,Y).



In contrast, the model performed significantly poorly on unseen data, with an accuracy hovering around 55%. The plot for unseen data is below.



To conclude, the “Calculate-Short Term Module could use many tweaks to get accuracy up for unseen data. In particular it would be beneficial to procure more relevant features to merge with the existing data and perform in depth feature selection. Great results and insights were gleaned thus far therefore this module was considered a success.

III. Regression and Ratings

* Overview and Goals

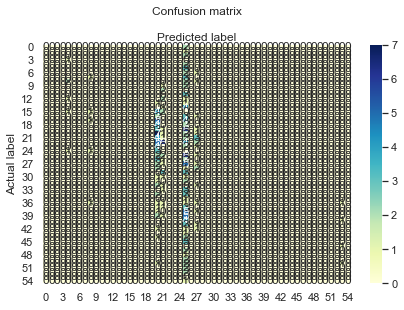
The original goal of using certain regression algorithms such as Logistic, Linear and Decision Tree to define the trends of certain genres popularity, while the revenue was determined in the previous section this section had to vary so instead of popularity we wanted to find the correlation between the number of ratings and reviews to the actual rating.

* *Data and Preprocessing*

The data for this portion of the project did not need too extensive of a preprocessing. Some general exploration was done but the dataset we used for this was pretty complete. There was still a need to check for missing values and data types of the required columns. This was required for each of the regression models used as certain ones do not accept float values.

* *Logistic Regression and Results*

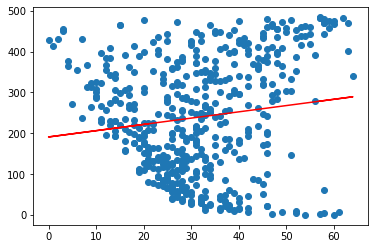
The first attempt on finding the correlation was Logistic regression, this algorithm was mostly chosen to set a baseline for the other few because as we know logistic regression is better for categorical features. Since this portion will mostly work with numbers using Logistic Regression will not yield very high results.



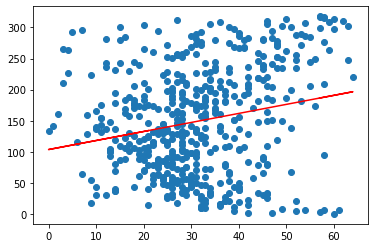
The above heatmap is the result of running Logistic regression, this visualization as you will notice is extremely clustered and doesn’t show results very well. The accuracy was extremely low, around .06, which was to be expected.

* *Linear Regression and Results*

Linear Regression was the next attempt. This was expected to yield much higher results. Linear Regression is geared towards the numerical and binary values. The data preprocessing for this one is different as well as many columns that were kept and separated by the features column in Logistic regression are dropped all together. This is due to the fact that the second parameter of Linear regression requires a reshaped array of one column. The Y values are reshaped and Linear regression is performed.



**Actual v.s Number of Ratings (Above)**



**Actual v.s Number of Reviews (Above)**

The above graphs are the results of Linear Regression ran on the actual rating v.s first the number of ratings and secondly the number of reviews. By Looking at the graph above we can see that the steeper slope belongs to the actual rating v.s number of reviews. This shows a higher correlation between these two columns than the other. While these results are okay, they could definitely be improved upon by better grouping of the dataset.

* *Factor Analysis and Results*

While this was not a part of the classes coursework, factor analysis was another method used to determine the impact of factors or columns on a target value. I will not spend too much time on it as coursework functions take priority but this did however yield the best results from the three.

(array([2.17882081, 1.22394878, 0.06717052]),

array([0.5447052 , 0.3059872 , 0.01679263]),

array([0.5447052 , 0.8506924 , 0.86748503]))

The above matrix on the last row shows variance between the factors for the actual rating. The last value of the array is the overall, while the other two are for the two factors used.

* *Overall Results and Improvement*

Here we take a glance over each method used, the results and how to improve. Firstly, Logistic Regression yielded by far the worst results as trying to use a function meant for categorical features will always perform poorly. Next, Linear Regression performed the best out of the methods used in class. It showed that the number of reviews mattered more than the number of ratings even though both were sloped positively. Lastly, Factor Analysis yielded the best results, but since this was not a part of our core classwork no details will be specified. To improve upon this, generalizing the genres to allow for more expansive function use would help, and also visualization changes would also help to understand results better.

IV. Conclusion

In conclusion the overall objective of our project, which was to analyze the macro trends in music using data mining was satisfied. Both sections: Random Forest Regression, and the inspection of the number of reviews/ratings on the actual rating, compliment each other. In Random Forest Regression, the model performed well albeit it was concluded that perhaps finding new features to be added to the data would in fact increase the predictive power and give more insight into the data graphically. From the Regression and Ratings section, a strong correlation was found between the number of reviews and the actual rating, This suggests that the number of reviews may be a good candidate to add to the RandomForest model’s data. If the project were to be continued, procuring review data specific to each album would be the next step. Subsequently, while training and testing the Random Forest model with the additional feature, a different, new feature would be tested with regression and also SVM to see if it is a good candidate for addition as well. To summarize, our initial goal was met however there are various things we would do to excel past our first goal and have a more complete predictive power for all aspects of music.

Citations

[1}https://chart2000.com/data/chart2000-albummonth-0-3-0060.csv

[2}https://gist.github.com/mbejda/9912f7a366c62c1f296c