

Individual Project : Popularity Score

For the individual project assigned in Data Mining 614 I was tasked to develop a predictive model that could help a music streaming company more advantageously determine whether certain songs are more likely to be popular upon future release. If this can be achieved, the company can make better use of its recommender systems and simultaneously increase revenue and lower costs. I took a strategic approach to incorporate three different types of classification in my analysis that included decision trees, logistic regressions, and k-nearest neighbor.

When implementing my decision tree analysis there were certain steps I took beforehand. After uploading the appropriate software packages, I then made sure to clean the data we were using by removing any missing values in the dataset. I then created a subset of all the available data in the dataset so that I could narrow my analysis using better explanatory variables and edited it accordingly to fit into our potential models (true, false, = 1, 0). A training set of 60% of the data and testing set of 40% of the data was then created. Decision tree modeling is a top-down method that uses statistical measures gained to select which test attributes should be used. I was genuinely shocked as to how poorly many of my models performed. None of the models I tried to fit for the data ever received an F1 score above 0.2. I did notice that lower values for sample splitting resulted in marginally better results for a fit, but the model outcomes were still too low for serious consideration.

There also seemed to be a soft cap for tree depth at around 15. Running cross-validation confirmed this. And when visualizing our trees there did not seem to be “too” many nodes that created high Gini levels, and yet, our results were still disappointing. The best fit I could achieve using decision tree modeling was with tree depth equal to 15 and sample splitting equal to 2. It yielded these results, recall: 0.17, precision: 0.17, F1_score: 0.17. I did get slightly higher results when I set the depth equal to 500, but this would have not been useful due to overfitting.

With my logistic regression models, the experience only worsened. I was expecting to have the most success using this type of classification. Logistic regression models are extremely useful when you are dealing with categorical data, especially if your dependent variable is categorical. There were 20 independent variables with which I was working. They ranged from the duration of the song, danceability, loudness, and genre, etc. Before fitting my initial model, I had to be sure to normalize the data since there were different measurements for the quantitative data. I then ran my first logistic regression model using all 20 independent variables and the variable “hot” (popularity above 75) as our dependent variable.

I was dissatisfied with those initial results, but after running a validation I finally began to see a glimmer of hope! The validation results came back with a precision score of 1.00! Now, granted the recall was only 0.01, at least it felt like progress. But after running a cross-validation all those numbers dropped to zero. The recall, precision, and F1 score for my logistic regression model were 0.00. This didn’t make much sense. There actually is a defining reason for these outcomes, more on that later. So, I decided to try building a more accurate model by removing all the variables that had no relevance in terms of correlation.

While two of the variables that I decided to keep in my second model did technically have p-values above 0.05, almost all of the 20 independent variables in the initial logistic regression model had p-values so high that I wanted to see if removing the really high ones would affect or rather lower the p-values of the independent variables that I did decide to keep in my second model. From the initial model I decided to keep the independent variables - explicit, with a p-value of 0.084, loudness, with a p-value of 0.068, and R&B which had a p-value of 0.007. After running the second model, the p-values for explicit and loudness both increased. After sequentially removing them and using the only variable left with any correlation to our dependent variable of popularity over 75, the final results produced another model with recall, precision, and F1 scores of 0.00.

The third type of classification I performed during my analysis was k-nearest neighbor. This is a classification rule assigned to a test sample to identify the category label of its nearest training samples, with k defining the number of neighbors, using some form of distance measurement. The results I got from these models were bad, just “less bad” than the other classification methods I used for my analysis. Since my k-nearest neighbor models seemed to be the lesser of three evils, the final model I chose for this project

would be a k-nearest neighbor model for prediction with $k = 1$. Yes, I understand that this could lead to misleading information because of overfitting. But we were tasked with choosing a final model, and from a financial perspective for the company, this model would produce the least amount of costs. It should be noted that in a real-world scenario, none of these models should be applied in a professional setting, and this is a very poor dataset to use for the purposes outlined by our client.

When initially running my KNN models, I first reloaded the original data since I had amended the dataset while running logistic regression, and then normalized the data. I began with a standard $k = 3$ model. This model produced a recall of 0.13, a precision of 0.25, and an F1 score of 0.17 after cross validation. I then decided to plot and try to find the best k that would fit the model. This wasn't immensely helpful; the plot indicated a high level of accuracy across a wide range of k values. After running a number of different k models, I finally decided that $k = 1$ produced the best results. My final model produced a recall of 0.28, a precision of 0.25, and an F1 score of 0.26. After prolonged and dubious considerations regarding how much this dataset failed to help me produce any meaningful information for my client, I decided that the problem might lie with how the data is constructed. I decided I would take one last approach and try something different.

For my final "final" model I decided to reengineer the data (I did not make any changes to any values in cells, all the reengineered data was taken from the original dataset). The idea to approach the project this way came to me when I noticed there were some songs in the data under the genre columns (pop, rock, hip hop, etc.) that did not have any 1s at all. I noticed this was because there was no column for metal music. So, I made a new column labeled "metal" and then for each row that had a metal tag in the genre column I placed a 1 in the metal column, otherwise I placed a 0. I did this so I could apply conditional formatting to the column to identify metal songs more easily. This could also have been done to add an additional explanatory variable to the dataset.

I then partitioned the data by genre and made a new dataset that only included rows for songs that had a "metal" tag in their genre column. The results produced a dataset with 65 rows that could be used to predict popularity for only the genre of metal music. There were 13 popular songs amongst the 65 records. I decided to change the parameters for popularity to a score of 70 to help prevent any overbalancing. Considering 70 is still a good score, and for the purpose of helping our client achieve a net profit instead of a net loss, this decision was warranted because it produced a result of 29 popular songs among the 65 records.

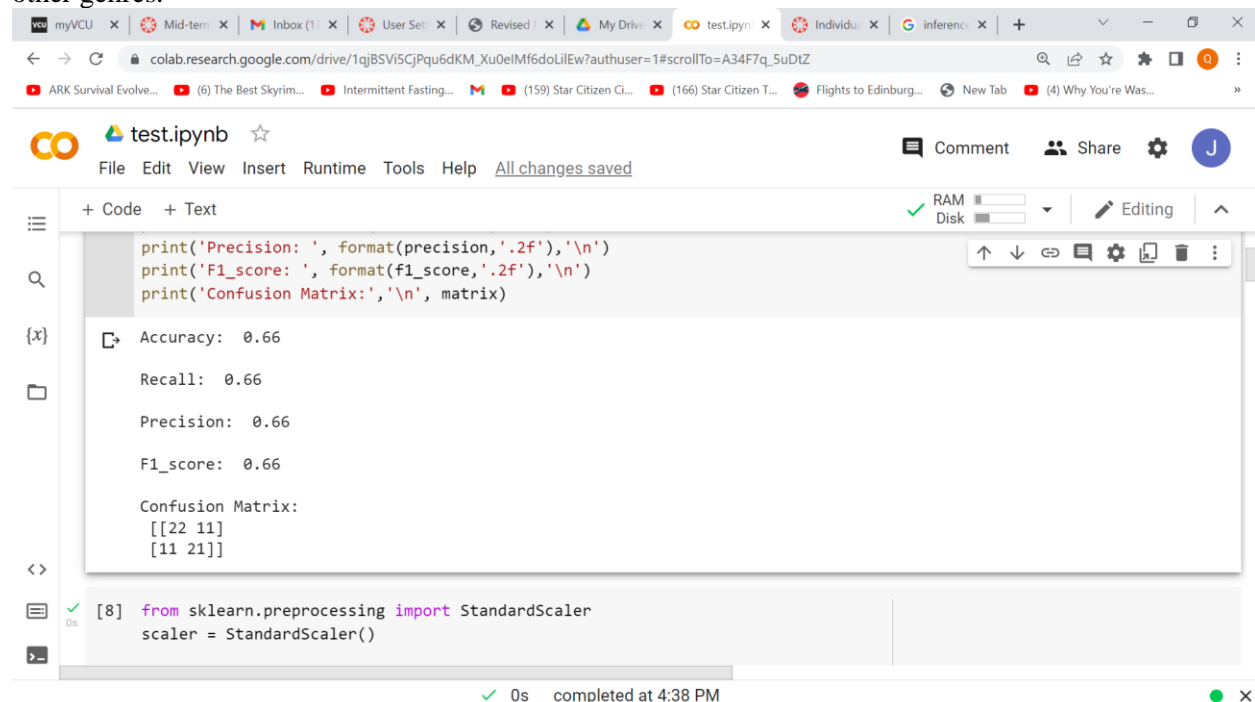
Amongst the explanatory variables indicated that we use, for the purposes of this dataset, I decided not to use hip hop, dance, folk, R&B, and Latin as explanatory variables. All the rows produced zeros for these columns so these variables would not give us very much additional information. I also decided to edit this dataset in excel beforehand, instead of editing it in python. I did this for two reasons. First, it is a smaller dataset, and I can keep track of my progress. Second, I did not want there to be any confusion regarding the first dataset I used which produced poor results. The code for the editing of explicit and removal of missing values only applies to the songs_utf.csv dataset. I changed true/false to 1/0 using the IF function and then copied those values to a new column labeled explicit and manually removed the variables stated above (it would have been easier to do it this way, but I actually decided to leave them in, just in case anyone wanted to view this dataset in its entirety; I excluded them from the subset I produced in python). I saved this dataset and named it "metal."

Considering how poor the results were with my previous data, I was optimistic for finding some meaningful insight using this new approach. First, I ran a decision tree with maximum depth set to 15 and minimum split set to 2. After cross-validation this model produced the following results: a recall of 0.66, a precision of 0.62, and an F1 score of 0.64. This is a huge improvement from our previous models. And while the accuracy for this model did drop to 0.63, which on the surface sounds bad, we must remember that there is no benefit associated with correctly predicting non-popular songs. We are only concerning ourselves with how many popular songs we can predict. So yes, I had found my best model so far. Increasing the minimum number of splits to 3 improved all our results across the board to 0.66.

After running a logistic regression model with my "metal" dataset, validation produced less favorable results. But our cross-validation produced remarkably similar results to our decision tree model

with an accuracy of 0.66, a recall of 0.59, a precision of 0.68, and an F1 score of 0.63. Now it was time for me to try KNN. First, I plotted and wanted to try and find the best k we could use for the model and used a range from 3 to 39. The plot showed that our highest level of accuracy was around k = 21. But after running models with different k's, I found that k = 5 was the best performing KNN model with an accuracy of 0.62, a recall 0.59, a precision of 0.61, and an F1 score of 0.60.

The final model I chose for my “metal” dataset is a decision tree model with max depth 8 and minimum split 3. Cross-validation confirmed the model produced a recall of 0.66, a precision of 0.66, and an F1 score of 0.66. Evaluation of the confusion matrix illustrated that the model produced 11 false negatives (\$1,200 cost each), 11 false positives (\$500 cost each), and 21 true positives (\$1,000 revenue each). $\$21,000 - \$13,200 - \$5,500 = \$2,300$, so this model does, indeed, produce a profit for this dataset (the inference of the values in the confusion matrix produced in the python code were confirmed by Professor Kim, python lists 0 before 1 which is the opposite of what was on the slides during our class sessions). Scaling this model appropriately with more metal records could produce even better results. This type of reengineering for the metal genre demonstrates that it is feasible to apply these same practices to other genres as well to achieve similar results. **I can not** speculate that the results for the metal genre will **certainly** produce the same favorable results for other genres, only that the success of this approach after dealing with so many failed models should warrant further investigation for the use of this approach for other genres.



The screenshot shows a Google Colab notebook interface. The top bar includes the Colab logo, the notebook name 'test.ipynb', and various icons for file management, comments, and sharing. Below the top bar is a menu bar with options: File, Edit, View, Insert, Runtime, Tools, Help, and 'All changes saved'. The main workspace is divided into two sections: 'Code' and 'Text'. The 'Code' section contains a Python code cell with the following output:

```
print('Precision: ', format(precision, '.2f'), '\n')
print('F1_score: ', format(f1_score, '.2f'), '\n')
print('Confusion Matrix:', '\n', matrix)
```

The output of the code cell is displayed in a separate box:

```
Accuracy: 0.66
Recall: 0.66
Precision: 0.66
F1_score: 0.66
Confusion Matrix:
[[22 11]
 [11 21]]
```

Below the code cell is a console output showing the execution of a command:

```
[8] from sklearn.preprocessing import StandardScaler
     scaler = StandardScaler()
```

The bottom status bar indicates that the execution was successful, with a green checkmark, '0s' execution time, and 'completed at 4:38 PM'.

There were many things I observed about this dataset during my analysis. The first and most obvious thing that popped out at me, and in large part – it might have been the only strong indicator, was that there were many popular hip hop and R&B songs. Pop as a genre gets thrown around a lot and can vary in taste and sound. This did not come as much of a surprise. Most R&B songs are about love and relationships, and anyone in marketing can tell you, sex always sells. And, furthermore, most R&B singers happen to be quite attractive and this only compounds towards their popularity and the potential of the popularity of one of their songs. It would be worth testing to see if R&B should be removed from the dataset because it might be an obvious correlation.

I also noticed that we were instructed to remove “Artist” as one of our explanatory variables. On the surface, this sounds like we might be missing some useful information. There are some bands that develop cult followings, and those die-hard fans might support a particular song just because it was released

by their favorite band. Bands such as Creed and Nickelback had this effect on their fans. You also see this effect take hold in the film industry. That is why Dwayne “The Rock” Johnson gets paid so much for his movie roles. Kids will show up to see the movie if he is in it, regardless of whether it got good reviews or not.

Meanwhile, the most crippling aspect of the dataset was the imbalance of nonpopular songs to popular songs. This is why some of our initial models had very low recall scores but high accuracy scores (our initial logistic regression model exhibited this behavior). There were only 192 popular songs out of the 1500 entries. This left us with a popular to nonpopular ratio of 1 to 7.8 (roughly). The only reason those models had higher accuracy is because they were correctly predicting nonpopular songs, which is not helpful for the purposes of this assignment. There is no associated revenue for predicting nonpopular songs.

Something I could not wrap my head around conceptually is how this dataset might fail to account for culture shifts over time. Some of the things that influence pop culture can change drastically and quickly. Something that made a song popular in 2000 might not necessarily have the same effects on a song in 2022. And these types of influences would affect some of the more “raw” data such as duration or loudness. My point being is that I am not entirely certain that randomly sampling from this data would produce correlations that have fruitful meanings (since the records came out during different years). Just because one correlation is produced from the overall dataset does not necessarily mean that correlation would remain true for any specific given year, again, because of the rapid shifts associated with pop culture.

I do understand that is why the variable “year” was instructed to be excluded from the list of our explanatory variables. But it still does not change the fact that those particular songs still were released in different years under very different cultural circumstances. I also didn’t think that the popularity variable should be included in the dataset (you would still keep “hot” included, but how to measure it needs to be challenged). Yes, I understand we excluded it from our models, but the score itself is misleading in its nature. Some genres of music are naturally more popular than others, and vice versa, some genres are naturally less popular (folk music for example). If the idea for our client is to target certain customers and determine if they will buy a record in the future, then these popularity scores need to reflect how popular a song is amongst the people who actually listen to that genre of music and not how popular a song is amongst the entire population.

For example, a folk song using the dataset’s metric might receive a popularity score of 28. But if most people don’t like folk music to begin with, then that number is misleading. It just might be that of the people who do listen to folk music, 98% of those individuals do like the song. So, it seems most appropriate to me that building models to predict whether these songs will be popular or not should be partitioned by genre. And last, but not least, how would you really use the variable “energy” as a predictive variable? Those scores are subjective. They won’t reveal anything concrete when being included with quantitative variables that use more accurate scaling systems. Duration is a count, loudness is a count, tempo is a count, but the “energy” variable is just something that a person is perceiving. There needs to be more clarity on how that variable was recorded.

	Accuracy	Recall	Precision	F1 Score
--	----------	--------	-----------	----------

DT depth=4, spilt=5	0.86	0.02	0.13	0.04
DT depth=5, spilt=2	0.85	0.03	0.14	0.04
DT depth=15, spilt=2	0.79	0.17	0.17	0.17
DT depth=500, spilt=2	0.78	0.21	0.19	0.20
Log Regression	0.87	0.00	0.00	0.00
Log Reg w/var removed	0.87	0.00	0.00	0.00
KNN k=3	0.84	0.13	0.25	0.17
KNN k=1	0.80	0.28	0.25	0.26
DT(metal) depth=8, spilt=3	0.66	0.66	0.66	0.66
Log Reg(metal)	0.66	0.59	0.68	0.63
KNN(metal) k=5	0.62	0.59	0.61	0.60

https://colab.research.google.com/drive/1qjBSVi5CjPqu6dKM_Xu0eIMf6doLilEw?authuser=1#scrollTo=b02oXg1gDNfs&line=1&uniqifier=1