CSE 158 Assignment 2 Report

1A. DATASET

For this assignment, we decided to use the following four datasets: *Hot 100 Audio, Hot Stuff.csv, Features.csv*, *UNRATE.csv*, and *InflationRate.csv*. The first two datasets are based on the Billboard Hot weekly charts while the latter two are based on unemployment rate and inflation rate, respectively. Our motivation is to find out if there is a correlation between the type of songs that are popular on the Billboard Hot 100 and unemployment and inflation rate. The "Billboard Hot 100 Audio Features" dataset provides a detailed view of the audio characteristics of chart-topping songs from the Billboard Hot 100, combined with Spotify audio analysis. Spanning multiple years, it includes attributes such as song title, artist, rank, and publication date, alongside musical features like danceability, energy, loudness, tempo, valence, and more. The "Billboard Hot Stuff" dataset provides information about the rank of a song on the Billboard Hot 100. It identifies songs based on the title of the song and who performs it. Regarding the rank of the song, the dataset records the position of the song on the Billboard Hot 100 for a specific week, the position for the previous week, its peak position, as well as how many weeks the song has been on the Billboard Hot 100. The unemployment rate dataset records the unemployment rate each month from 1948-2024, while the inflation rate dataset records the annual inflation rate from 1960-2023.

1B. EXPLORATORY ANALYSIS

Starting off with a basic analysis of our datasets, the two datasets related to the Billboard Hot 100 are created using data from 1963-2021. The Billboard Hot 100 Audio Features contains 29503 entries, while the Billboard Hot Stuff has 327894 entries. As stated earlier, the unemployment rate dataset covers the period from 1948-2024, whereas the inflation rate dataset encompasses the years 1960-2023. Since the unemployment rate dataset records monthly it consists of more entries than the inflation rate dataset which records annually. The unemployment rate dataset sits at 922 entries compared to the 64 entries in the inflation rate dataset.

A notable observation of the Billboard Hot 100 dataset is that highly ranked songs tend to exhibit greater danceability and energy, reflecting audience preferences for dynamic and engaging tracks. This dataset not only offers insights into the changing landscape of mainstream music but also highlights the elements that resonate most with listeners, making it an essential resource for understanding trends in musical composition and popularity. Comparing the unemployment rates to the inflation rates throughout the years, we notice that both tend to increase and decrease with each other. The most significant changes in the rates occurred all throughout the 1970s and 1980s as well as briefly in 2010 and 2020.

2A. PREDICTIVE TASK

Our predictive task was as follows: given the average characteristics of the most popular songs for a particular year, what was the inflation rate? The motivation for this idea stems from the thought that certain characteristics of a song could make it more popular during a certain period of time. For instance, what made disco popular in the 1970s, could it be because people experienced economic hardship during this time so they listened to upbeat music to lighten up?

We decided that an approach based on regression would make the most sense for this predictive task. As this was a numerical prediction, we chose to evaluate the performance of our model by its mean squared error. This is a standard method by which regression models are evaluated, and it has the benefit of penalizing large errors more than small ones. As such, a model with a low MSE may not be perfectly accurate, but it is close enough to be considered insightful.

To aid in our MSE-based evaluation, we created a simple model to use as a baseline. The way this baseline model operated is simple: it used the average of the inflation values from the training data as its prediction for every data point. We found that this baseline model had an MSE of about 3.0594. The validity of our model was determined by comparing it to this baseline: if the MSE of our model's prediction set was worse than this value, that would have indicated our model was a poor predictor. However, if our model produced a significantly lower MSE value, that would indicate the validity of our approach.

2B. FEATURES

To obtain the features for our model, we had to combine and process data from multiple datasets. The two datasets related to the Billboard Hot 100 were merged, and any rows that lacked sufficient data were dropped. This produced a dataset consisting of the songs that were at the top of the charts for each week, as well as their audio characteristics. However, the inflation dataset listed the inflation rate on a yearly basis, and so we had to adjust our features accordingly. Audio characteristics of the songs in the weekly top 100 were grouped by year and then averaged to create the features used by our model, and the inflation rate for each year was chosen as the objective. This dataset was split into training and testing data at a 4:1 ratio, where the former was used to train the model and the latter was used to produce the associated MSE value.

As a part of the exploratory analysis, we also performed regression with these same features to predict the monthly unemployment rate. For this analysis we averaged the audio characteristics by month instead, and used the unemployment rate for each month as the objective. The results of this exploratory analysis indicated that at least half of the audio features were statistically significant predictors for the unemployment rate, and so we felt it was worthwhile to proceed with our inflation rate analysis.

3. MODEL

Our predictive task consisted of using several numerical features to obtain a single number as output, and so we decided to experiment with a regression-based approach. We considered several models and compared their performance to arrive at our final result. The three models we chose to compare were as follows: a model that used AIC based stepwise selection to choose the best features for use with OLS regression, a model that performed p-value based feature selection to use with OLS regression, and a model that scaled features for use with ridge regression. Each of these models was compared to the baseline model to determine their validity. Feature selection was considered because, as per our exploratory analysis, not all features appeared to be statistically significant. AIC and p-value selection both seemed promising, as it was unclear which would necessarily provide the best possible selection of features. We also wanted to consider an approach that implemented feature scaling and weighting, and so we applied ridge regression to the predictive task.

4. RELATED LITERATURE

Our datasets are sourced from several places. The datasets related to music come from the Billboard Hot 100 weekly singles chart, with the song characteristics being sourced from Spotify. Typically, this data is studied to discover and predict trends in listening habits. The economic datasets are sourced from Federal Reserve Economic Data. This data is typically used to study and make predictions about the economy. There are very few examples of these datasets being used in combination, so finding literature related to this project was difficult.

There is literature claiming the existence of the correlation between music styles and the economy. For example, a recent article published by CNBC discusses the concept of ‘recession pop,’ a musical genre associated with the 2008 financial crisis that is apparently making a resurgence (Dickler and Solá). The most relevant literature to our project came from a project presented at the Discovery Summit Americas 2023, which used data sourced from Spotify and shared several of the same features as our own model. This project also implemented regression, but unlike our model, the objective in this project was the popularity of each song, and economic factors (including the unemployment rate and GDP) were implemented as additional features (Sun et al.). The results of this project agreed with our own findings.

5A. RESULTS

Each of the three models applied to our predictive task were evaluated on their MSE in comparison to that of the baseline model and to each other. Their ranked performances, along with that of the baseline, are as follows:

Rank	Model	MSE
1	Ridge Regression	2.1071
2	AIC based OLS	2.7797
3	Baseline Model	3.0594
4	P-value OLS	3.2001

It can be seen from these results that ridge regression provided the most accurate predictions, followed by the AIC based approach, and the p-value approach gave the least accurate predictions. The p-value approach resulted in a model that only incorporated three features: loudness, liveness, and valence. Ultimately, this model performed even worse than the baseline. The AIC based approach included the same features, but it also incorporated time signature, energy, mode, acousticness, and speechiness, for a total of eight features. This model improved on the baseline, but it was not the top performer.

Of each approach we tried, the model based on ridge regression produced the lowest MSE. It is likely that this outcome is related to multicollinearity in our dataset. For example, it would make sense if danceability and energy are highly correlated in music. Given this assumption, it appears that ridge regression was a good choice for our predictive task. That is because ridge regression is designed to penalize the weighting of correlated features, which means that it is well-suited for prediction tasks with datasets that exhibit multicollinearity.

5B. CONCLUSION

Our analysis indicated that there is indeed correlation between the characteristics of popular music and the state of the economy. From our exploratory analysis, we found that certain attributes of popular music were statistically significant indicators of unemployment rates. By focusing on these characteristics, we were able to build several predictors for the yearly inflation rate. In comparing these models, we found that ridge regression gave the most accurate predictions, likely because it is designed to account for multicollinearity in datasets. The MSE associated with this model when evaluated on the testing data was just 2.1071, about 0.95 less than that of the baseline. Considering that there is natural variability in popular music that can’t be accounted for, this is a strong outcome overall, and so we were successfully able to apply this model to our predictive task.

REFERENCES

Dickler, Jessica and Ana Teresa Solá. "'Recession pop' is in: Why so many listeners are returning to music from darker economic times." *CNBC*, https://www.cnbc.com/2024/07/21/recession-pop-explained-how-music-collides-with-economic-trends.html

"Inflation, consumer prices for the United States." *FRED*, https://fred.stlouisfed.org/series/FPCPITOTLZGUSA

Miller, Sean. "Billboard Hot 100 Weekly Charts with Spotify Audio Features." *Kaggle*, https://www.kaggle.com/datasets/thedevastator/billboard-hot-100-audio-features

Sun, Yoon Hye, et al. "Relationship Between the Type of Music and Economic Conditions." *Discovery Summit*, https://community.jmp.com/t5/Abstracts/Relationship-Between-the-Type-of-Music-and-Economic-Conditions/ev-p/738898

"Unemployment Rate." *FRED*, https://fred.stlouisfed.org/series/UNRATE