Explicitness on Spotify

Johannes M. Halkenhaeusser

Minerva Schools at KGI

IL181.006 Prof. Scheffler

Fall 2020

Introduction

Spotify has revolutionized the music industry by offering a subscription service that allows users to choose from millions of songs and international artists. Analyzing a sample of these songs will help us understand what can make a hit song on Spotify. Over time there has also been a dramatic increase in the amount of explicit language used in songs. While songs in the 1950's used barely any explicit language and had artists like Elvis Presley became famous for a suggestive dance, today's songs are characterized by explicit language.

Need for CATE

The "treatment" of explicit language may be highly dependent on various other characteristics such as its age. For example, a 2020 song's popularity is likely to gain or be unaffected by explicit language. On the other hand, a 1940s piece may not be popular if it includes explicit language. This dependence of other variables on the effect of explicitness on the popularity of a song is means that we are dealing with a conditional average treatment effect (CATE).

Data

The data contains 169,909 individual songs from Spotify, including audio track characteristics such as the length, year, valence, the artists. It also contains a popularity score, which is a function of the number of plays and the recentness of these plays. According to Spotify, plays that lie further in the past contribute less to popularity compared to more recent plays. A table of the feature definitions, as given by Spotify, is given in the Appendix. The dataset was obtained from User Yamaç Eren Ay, who created it in 2020: https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks.

There is a small subset of the data for which the genre or multiple genres are available. The genre definition is, upon inspection, not particularly telling because songs are associated with multiple genres without an indication which genre fits the most or about the overlap of the genres. Hence, while it would be a valid predictor, this analysis will assume that what makes genres unique is also captured by the other predictor features.

There may be significant sample selection depending on the age and popularity of a song. For example, a song that was produced in 1948 may not be added to Spotify at all because its features did not warrant popularity in 1948, even though it could be popular today with the exact same features meaning our sample is biased through which songs were selected into it. Further, at most 2000 songs fro each year have been added to the dataset. We will have to assume that they are chosen at random.

Estimating the effect of explicitness

Looking at the ratio of explicit songs per year in Fig. 1, we can see that the ratio of explicit songs has steadily increased over time. This points to artists realizing that more explicit songs are more popular with their audience, meaning more explicit songs are produced. We are making the assumption that explicitness is not a function of whether a song has gone under review for explicitness but that they have been correctly classified. In the years prior to 1950, there is only one song from 1933 that has a non-zero popularity score[1], which further shows why the year of the song may be a highly relevant predictor of the treatment effect of explicitness. Note also how the dataset is imbalanced, with there being about ten times more non-explicit than explicit songs in the dataset.

However, could popularity just be explained by it being a function of the age of the plays which is naturally lower for older songs. As shown in Fig. 2, popularity in 2020 increases over time, which may lead to the correlation between explicitness and popularity we see in Fig. 3 being just that: a correlation.

It is crucial to understand popularity as representing popularity at the time of data retrieval in 2020. However, this is not terrible since we care about what makes a song popular in today's terms, so being able to disentangle out what we cannot change anyways (age) will be helpful.

[1] A song called *Shave 'Em Dry II* by Lucille Bogan.

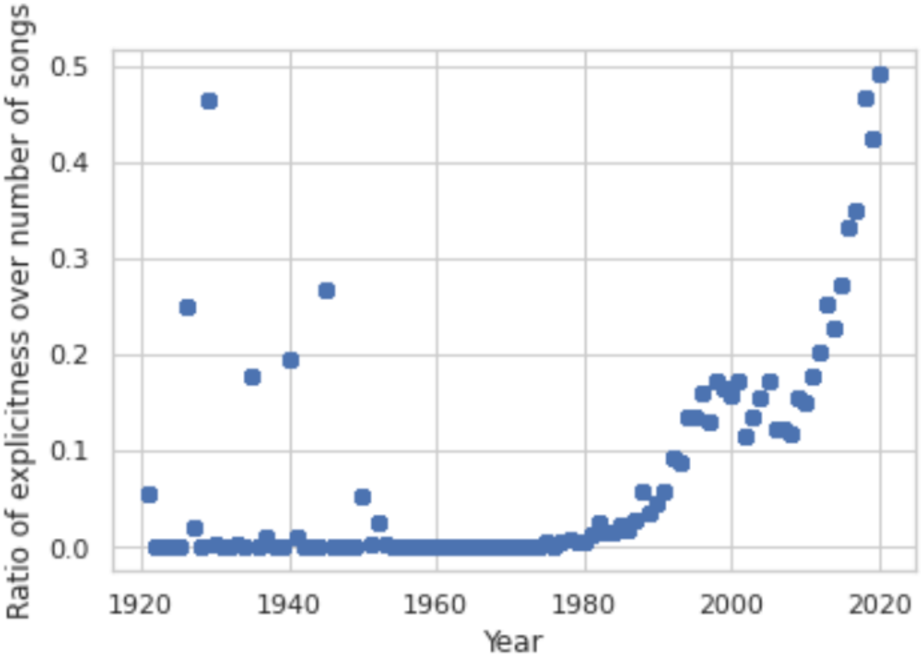

Figure 1. Ratios of songs per year rated "Explicit." Over time explicitness has markedly increased exponentially.

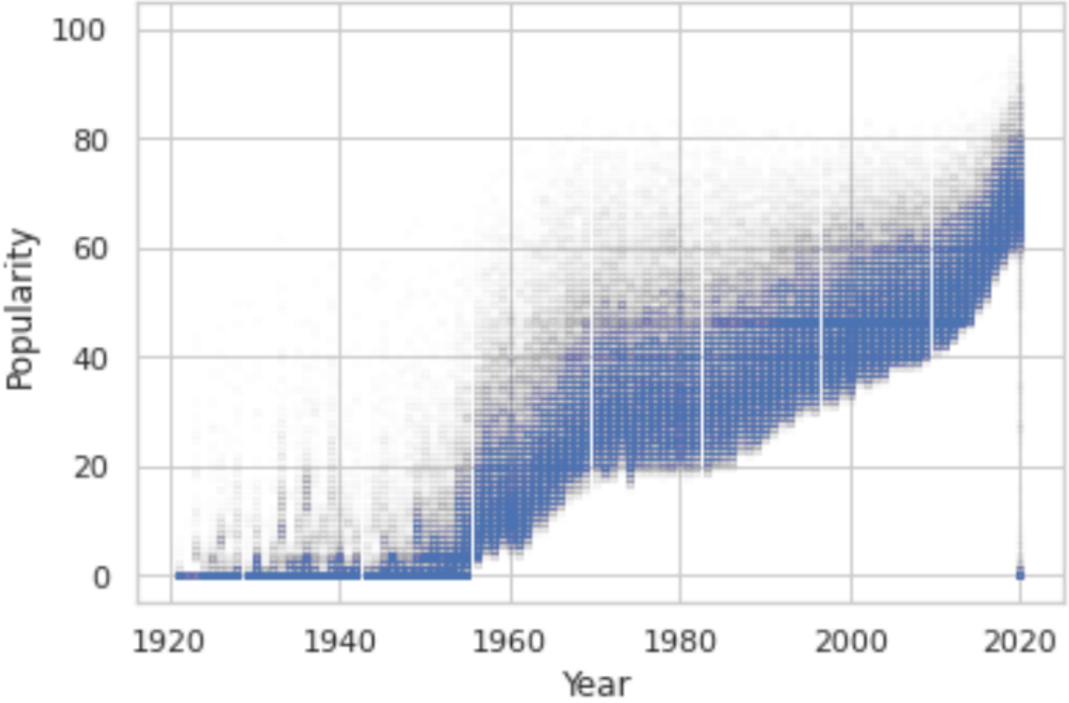

Figure 2. Popularity of songs in 2020 plotted against their year. Generally, the older a song is the less popular it is in 2020. Note, that the opacity of markers is very low in this plot to show the general trend of the data.

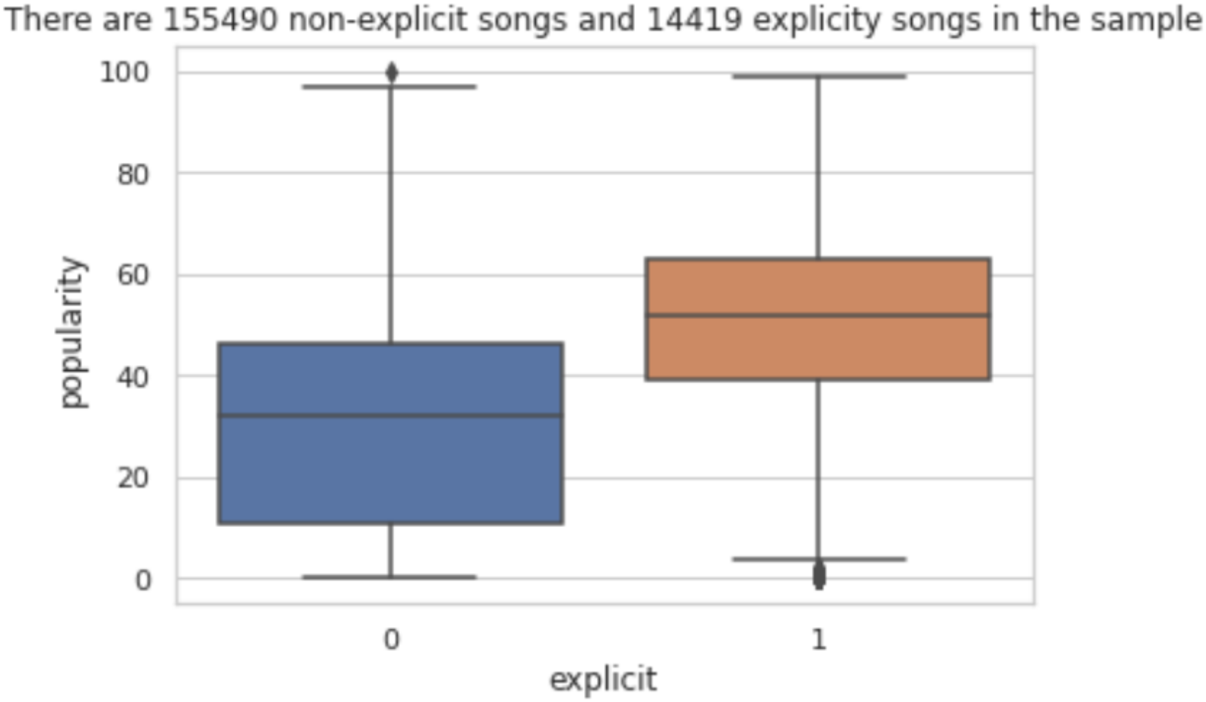

Figure 3. Boxplot of popularity in the non-explicit (left) vs. explicit (right) songs.

Causal Modeling of the effect of Explicitness on Popularity

We suspect that the effect of explicitness on popularity is affected by a variety of factors such as speechiness, tempo, energy, and the year it was released.

Therefore, we have to go through the necessary steps to perform a causal inference estimation:

- *model*
- *identify*
- *estimate*
- *refute*

Model

Using DoWhy we can estimate the heterogeneous treatment effect of explicit lyrics. We will assume that we have measured all the relevant confounders given that these are the different audio track features measured by Spotify, who should hopefully know what they are doing. To make a valid estimation of the CATE this is the strongest assumption because we cannot control for unobserved confounders and it fundamentally affects our model. We cannot magically observe another confounder, even if it is there, but we can test the strength of our estimate if there was another confounder in our *refute* section. The causal model that is constructed is based on our understanding of the causal mechanisms and no statistical software is capable of making causal connections. The causal model we construct here is shown in Fig. 4 below.

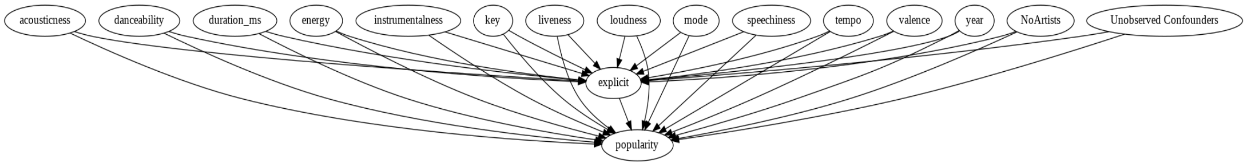

Figure 4. Causal Model. Explicit is predicted by the other observable covariates and together with these covariates predicts the popularity of a song.

Identify

We can have DoWhy give us the identification strategy based on our DAG. It tried this backdoor way which tries to control for the different covariates. It also tried finding an instrument but did not find one.

```
Estimand name: backdoor

Estimand expression:
```

$$\frac{d}{d[explicit]}(\mathrm{Expectation}(\mathrm{popularity}|X,W))$$

, where X and W are effect modifiers and common causes.

Estimate

When we examine the flowchart for which causal inference method we should use, we can follow along the non-experiment path. We assume that all confounding drivers of treatment assignment have been measured, and that there are only a few (4) features that affect treatment responsiveness. The term "few" is taken as relative to the number of samples. Given that we have 160+ thousand samples and only 4 features, we can with certainty answer that there are only few features that affect the treatment effect.

We do not know the nature of the heterogeneity (i.e., its linearity). With the uncertainty, EconML advises using, among others, forest estimation models. Within the forest methods is a Doubly Robust Machine Learning method which is usually preferable over Double Machine Learning methods, unless there are regions in the control space that have low probability of treatment assignment.

The ForestDRLearner estimates the conditional average treatment effect by using two stages. In the first stage, nuisance functions are estimated in a cross-fitting manner. In the second stage, the CATE is estimated. There are two first stage estimation functions, one estimating treatment the other the outcome.

$$E[T \mid X, W] \tag{1a}$$

$$E[Y \mid X, W, T] \tag{1b}$$

(1 a) estimates the treatment given the observed confounders and effect modifiers. Because this estimates a binary treatment outcome, we can use any classification model that we would regularly use to solve this propensity score estimation task. Here, we use a GradientBoostedClassifier. This method is suitable because it allows for modeling the non-linearity in the data.

The second first-stage estimation uses the features, confounders, and treatment to estimate the outcome variable (1 b). However, because popularity is continuous, we use a gradient boosted regressor and not a classifier. The final stage is left at the default and allows for the estimation of the multi-task regression.

Double Machine Learning works almost the same but predicts the outcome just from the controls but performs better when the controls do not overlap in the treatment assignment. Double Machine Learning only requires overlap on average. This overlap problem could be the case with how the explicit ratio increases over time. Hence, we will compare both methods as a robustness check using forest estimators. We would expect both methods to perform relatively similar.

The forest estimators are chosen because it allows for computing valid confidence intervals and is still suitable when there is uncertainty about the model (Microsoft Research, 2019). DML and DRL are estimated with trees to allow model comparability.

Results

The CATE of explicit on popularity seems to be negligible. The conditional average treatment effect of explicit given age, speechiness, tempo, and energy estimated by the DML is around 0.0586. The DRL estimates it at -0.04415, pointing to a non-robust treatment effect of "explicit".[2] When we plot the DML tree into an interpretable tree format (Fig. 6), we find that at any node, the CATE standard deviation renders all of the CATE estimates insignificant. Similarly, when we plot the effect on the observations against the different features, we see that none of them affect the effect of explicitness on popularity (Fig. 7).

The Doubly Robust Learner performed surprisingly bad in comparison to the Double Machine Learner and its estimate is likely invalid. The models can be compared using the mean squared error score that is returned by EconML. The lower the score/error, the better the model. The DRL has an MSE of 6113.159 while the DML's MSE is only 81.105. Given that the range of popularity is zero to one hundred, these MSEs are unsatisfactory and we will hence only try to refute the better model (DML). A possible explanation is the lack of overlap in treatment probability in the covariate space which leads to the DML method being able to better extrapolate and achieve lower MSE (Microsoft Research, 2019). We can check if this is the case by looking at the first stage propensity scores and plotting it against our suspect for missing covariate overlap (Fig. 8). There is a portion of the data between 1950 and 1980 (around a third of that variable range) that has no overlap which is a suspected source of the flaw in this method.

[2] Estimating the confidence interval was unfeasible with the amount of data and model complexity. The SDs in the tree graphic are calculated using only few trees.

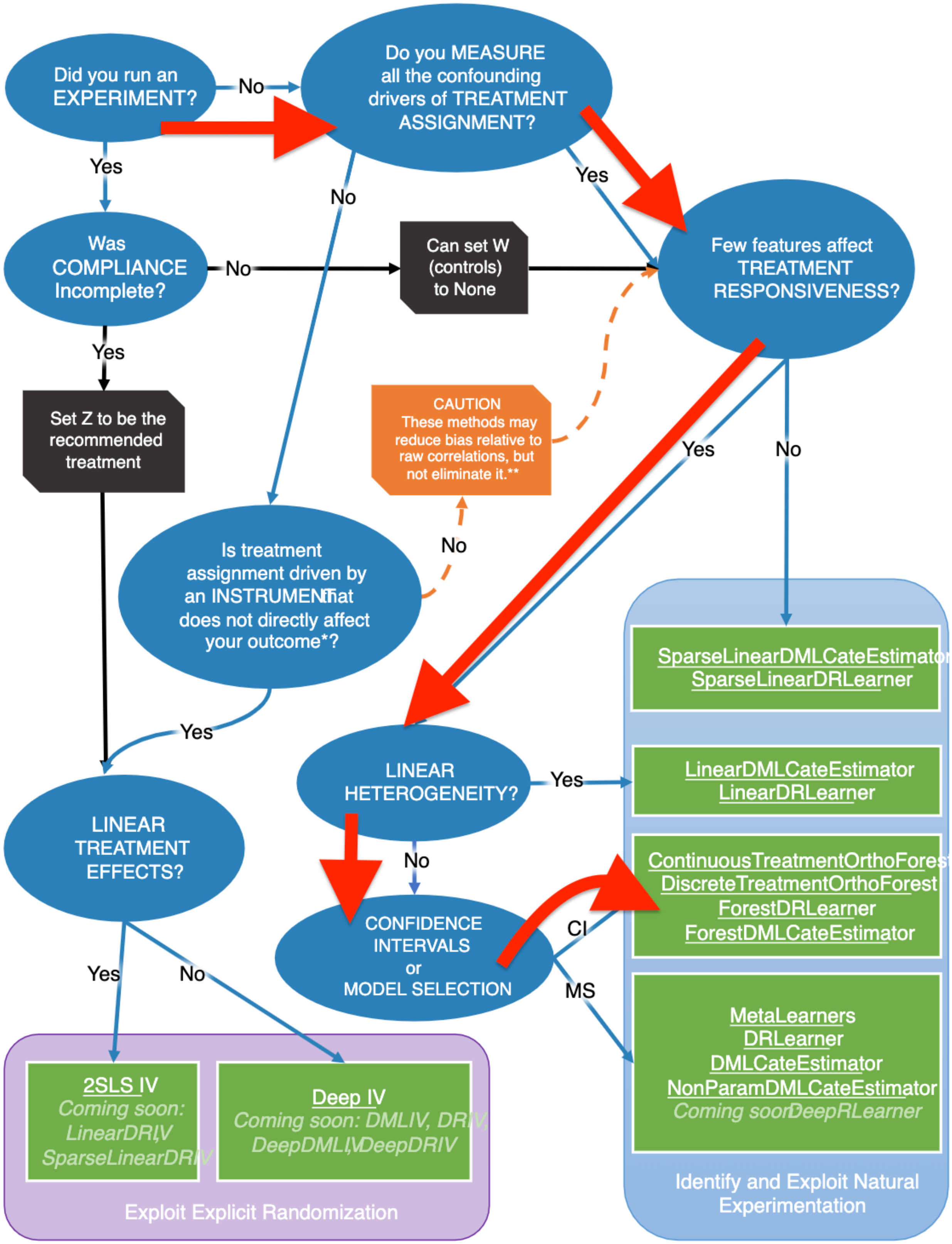
Did you run an EXPERIMENT?
No
Do you MEASURE all the confounding drivers of TREATMENT ASSIGNMENT?
Yes
No
Yes
Was COMPLIANCE Incomplete?
No
Can set W (controls) to None
Few features affect TREATMENT RESPONSIVENESS?
Yes
Set Z to be the recommended treatment
CAUTION These methods may reduce bias relative to raw correlations, but not eliminate it.**
Yes
No
No
Is treatment assignment driven by an INSTRUMENT that does not directly affect your outcome*?
SparseLinearDMLCateEstimator
SparseLinearDRLearner
Yes
LINEAR HETEROGENEITY?
Yes
LinearDMLCateEstimator
LinearDRLearner
LINEAR TREATMENT EFFECTS?
No
ContinuousTreatmentOrthoForest
DiscreteTreatmentOrthoForest
ForestDRLearner
ForestDMLCateEstimator
CONFIDENCE INTERVALS or MODEL SELECTION
CI
MS
Yes
No
MetaLearners
DRLearner
DMLCateEstimator
NonParamDMLCateEstimator
Coming soon: DeepRLearner
2SLS IV
Coming soon: LinearDRIV SparseLinearDRIV
Deep IV
Coming soon: DMLIV, DRIV, DeepDMLIV, DeepDRIV
Exploit Explicit Randomization
Identify and Exploit Natural Experimentation

Figure 5. *EconML flowchart. The red arrows indicate the path we are taking at each decision node (Microsoft Research , 2019).*

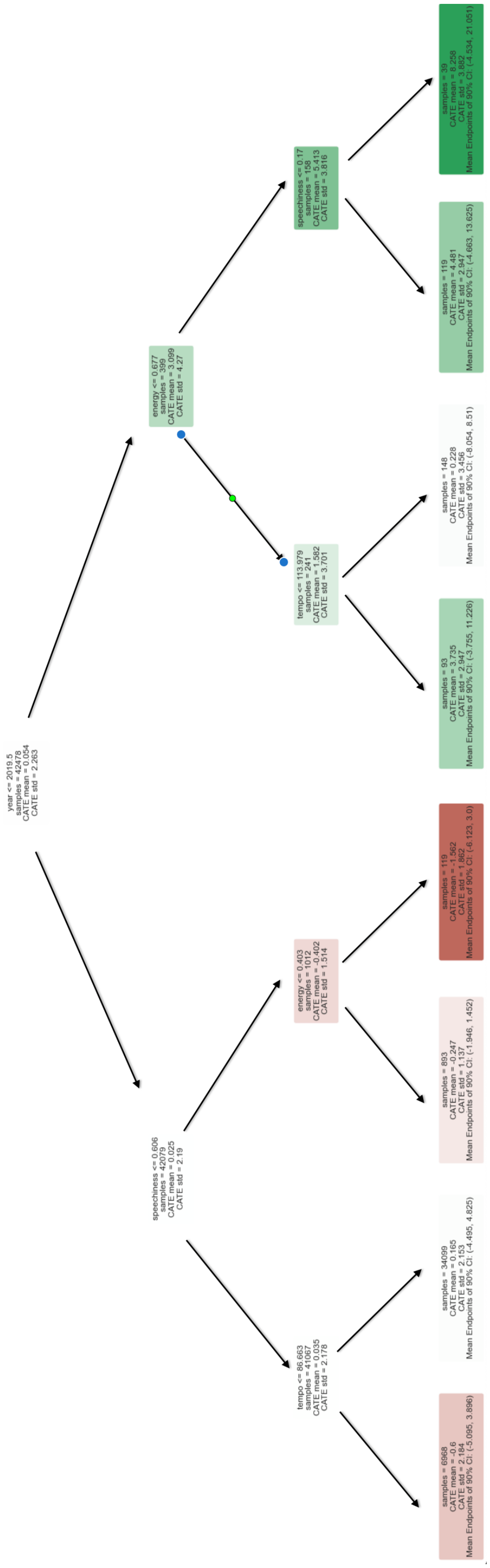

Figure 6. Tree interpretation of the DML model. We only visualize the DML because the DRL's MSE is too high. See the code notebook for a clearer visualization.

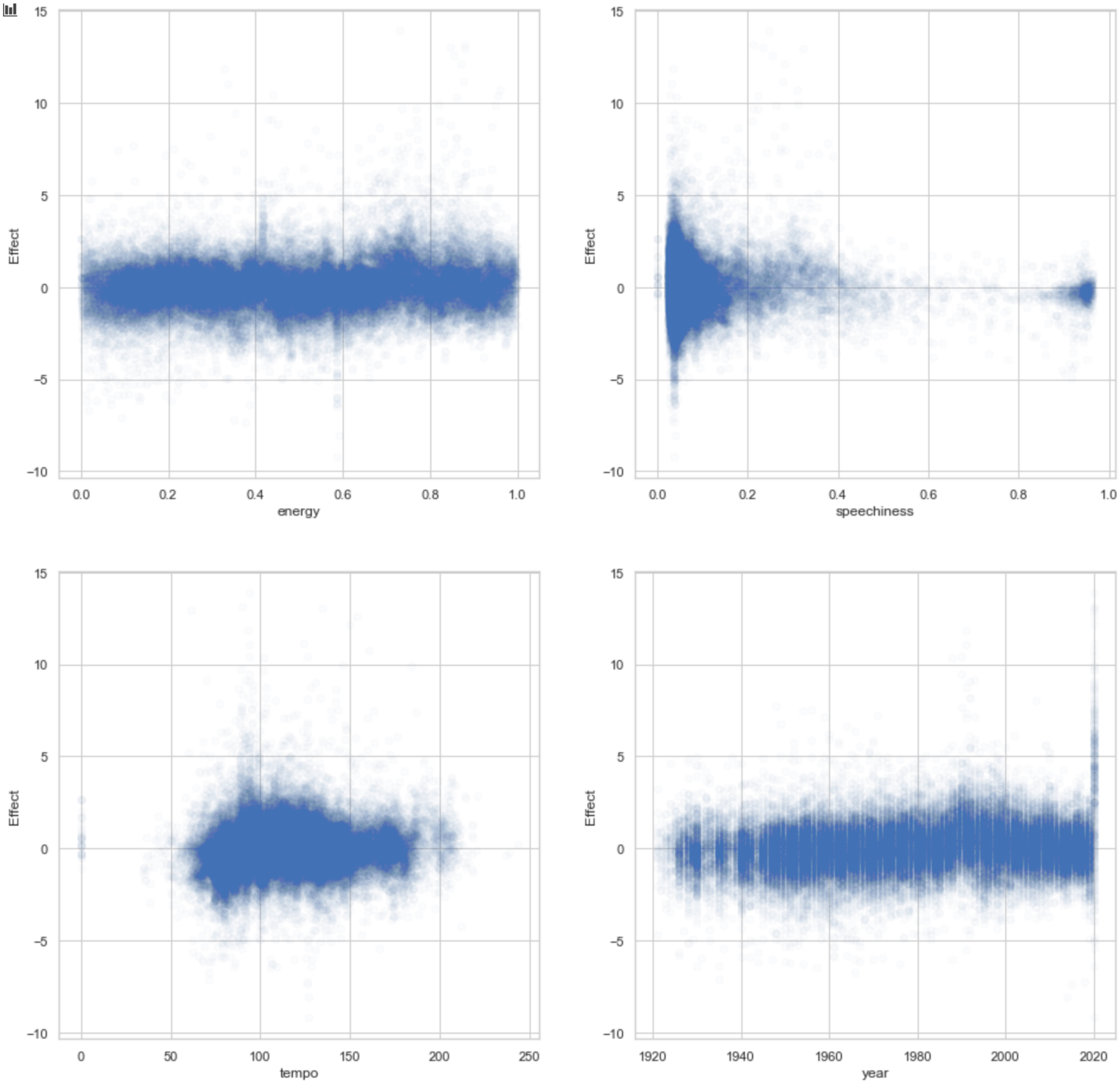

Figure 7. CATE as a function of the different features. None of them is a relevant predictor of how explicit affects popularity. [3]

[3] #dataviz: With a large number of observations it is difficult to make meaningful graphs that tell a story. However, the data visualizations all serve to make a point (e.g. insignificance in effect size) or like Fig. 8, a visual showcasing of the lacking propensity score overlap.

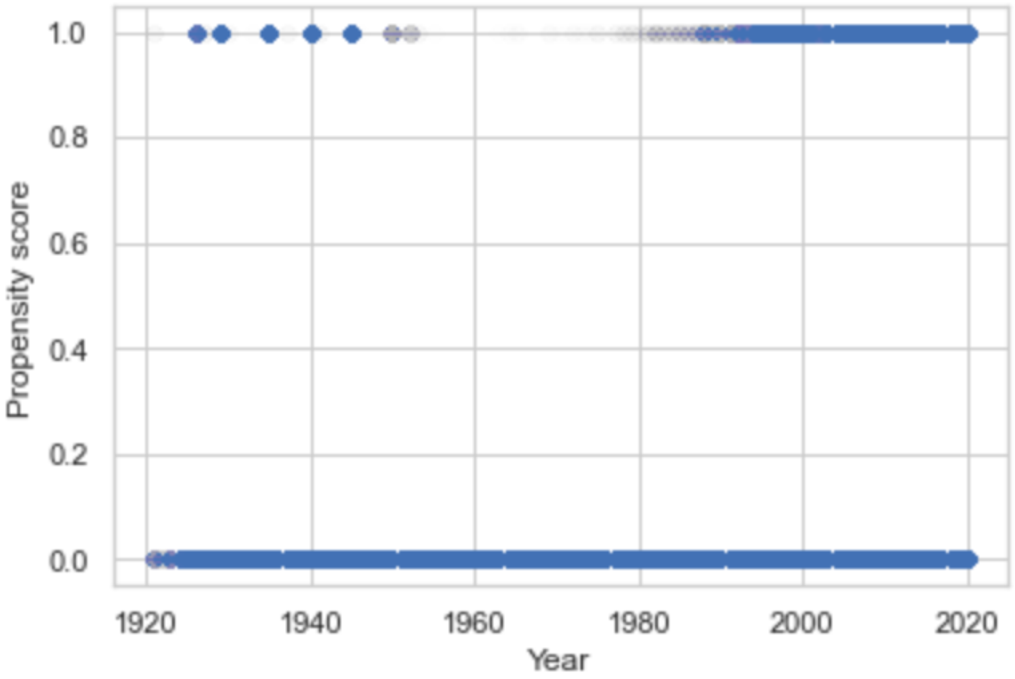

Figure 8. Not sufficient overlap in treatment probability for the full covariate space of the feature year. *The opacity of datapoints is set at 0.02 to not let individual datapoints skew the interpretation of the figure that contains roughly 120,000 individual propensity score predictions.*

Refute

The DRL model has effectively been refuted by comparing it to the DML model that has outperformed it on the MSE. Both models are trained using a large dataset (~120,000 samples), so let us check how the model behaves when randomly delete 20% of the data. Given that there is imbalance in the dataset, deleting data will let the model hinge on fewer treated observations, which in a non-robust model could lead to an effect of explicit being even less easily identifiable. When we run this robustness check, the effect is almost unchanged and the p-value is almost at 0.5, confirming the highly insignificant CATE of explicit.

Another robustness check is adding a random common cause. If the model is not robust, we will see that a CATE should be estimated as it will pick up the random common cause, as the treatment effect. This flips the treatment effect's sign, showing again that explicit is not a relevant predictor.

Conclusion

As a relieve to educators, explicit language does not seem to influence the success of music when using our models. To estimate this insignificance, we have built a causal model predicting the popularity of songs based on the treatment of having explicit language when conditioned on energy, tempo, speechiness, and release year. By comparing a doubly robust and double machine learning model we have found that both models show no significant effect of explicit language. Yet neither model is particularly great, having a large MSE which may be due to underlying other confounders or sample selection that was not measured. SUTVA may be violated because the popularity of a song by one artist may affect the popularity of that artists next song. In future models, we should account for both the sample selection through, e.g., a heckit model (Heckman, 1976) and the SUTVA violation by changing the unit of observation to an artist or album level.

Works Cited

Microsoft Research . (2019). *Library Flow Chart*. Retrieved from EconML: https://econml.azurewebsites.net/spec/flowchart.html

Microsoft Research. (2019). *Forest Based Estimators*. Retrieved from EconML: https://econml.azurewebsites.net/spec/estimation/forest.html

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 475–492.

Ay, Y. E. (2020, 10). *Spotify Dataset 1921-2020, 160k+ Tracks*. Retrieved from Kaggle: https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

Spotify AB. (2020). *Get A Track*. Retrieved from Spotify for Developers: https://developer.spotify.com/documentation/web-api/reference/tracks/get-track/

Microsoft Research. (2020). *Customer Segmentation: Estimate Individualized Responses to Incentives*. Retrieved from GitHub: https://github.com/microsoft/EconML/blob/master/notebooks/CustomerScenarios/Case%20Study%20-%20Customer%20Segmentation%20at%20An%20Online%20Media%20Company%20-%20EconML%20%2B%20DoWhy.ipynb

(Microsoft Research, 2020)

Appendix A – Feature Definition

These definitions are taken straight from the Spotify API documentation (Spotify AB, 2020)

Feature	Detail
Acousticness	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
Danceability	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
Energy	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
Duration_ms	The track length in milliseconds.
Explicit	Whether or not the track has explicit lyrics (true = yes it does; false = no it does not OR unknown).
Instrumentalness	Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
Key	The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on.

Liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
Loudness	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
Mode	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
Popularity	The popularity of the track. The value will be between 0 and 100, with 100 being the most popular. The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Duplicate tracks (e.g. the same track from a single and an album) are rated independently. Artist and album popularity is derived mathematically from track popularity. Note that the popularity value may lag actual popularity by a few days: the value is not updated in real time.
Speechiness	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.

Tempo	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
Valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).