Predicting Whether a Listener Will Love a New Song
Data-Driven Marketing in the Music Industry

STAT 4249 Applied Data Science

Team 26
Benjamin Seiyon Lee, Sou Min Sonia Lee,
Mengying Li, Wuwei Tang, Xiaolei Xu

Table of Contents

1. Executive Summary
2. Background
3. Research Question and Hypotheses
4. Data
5. Pre-processing
6. Our Approach
    1. Regularized (Elastic Net) Multinomial Logistic Regression
    2. SVM
    3. Random Forest
    4. Other methods (Naive Bayes)
7. Results
8. Conclusion
    1. Hypotheses
    2. Recommendations
    3. Further Considerations
9. Appendix

Executive Summary

Motivation

Can we use someone's personal information to predict whether or not they will like a song? Music titan EMI compiled a comprehensive dataset (EMI One Million Interview Dataset) containing the interests, attitudes, and behaviors of listeners around the world. For their 2012 Music Data Science Hackathon, EMI released a subset of the data set and tasked data scientists to "predict if a listener will love a new song." We decided to take on this challenge. This report presents the results from four different classification methods (Elastic Net Regression, Support Vector Machine, Random Forest and Naïve Bayes).

Results

Using the EMI sample, our team predicted users' ratings of songs into three preference categories (Like, Indifferent, and Dislike) using four classification methods.

- The **random forest** model "grows" numerous random classification trees to optimally classify our records. It was our best model with a misclassification rate of **31.4%**
- **Support Vector Machine (SVM)** classifies new observations depending on which side of a hyperplane it lies on. The SVM model had a misclassification rate of **33%.**
- The **Elastic Net Regression** is a more efficient version of a standard logistic regression model except that it selects the optimal predictors and shrinks their coefficients using an L1 and L2 penalty. This model returned an error rate of **46.1%.**
- Other models included the **Naive Bayes Classifier (52.8%).**

Model Selection and Assessment: Accuracy, Interpretability, and Complexity

While each model was able to beat a random guess (67% misclassification rate), there are other concerns when it comes to choosing the "right" model. The random forest model is the most accurate, but it does not provide the "direction" of a predictor's influence on the dependent variable. The SVM method is less prone to over-fitting, but it computationally intensive and it is a "black box" predictor with low interpretability. The multinomial logistic regression model with elastic net regularization allows us to "fine tune" a logistic regression model using a penalty parameter (λ), and it allows us to interpret the influence of the predictors. However, it is less accurate and less stable when the response variable has more than two classes.

Findings

- Familiarity leads to favorability. Our analysis suggests users who have heard songs by an artist are more inclined to like that artist.
- Certain demographic variables (e.g. gender, age) do not significantly improve a model's predictive power. It is more important to focus on users' music consumption behaviors (hours listened to music per day) than their demographic information.

Recommendations

- We recommend that EMI use the random forest model in conjunction with the multinomial logistic regression using elastic net regularization. The random forest model provides more accurate predictions, which is better for targeting likely fans. The interpretability of the logistic regression model with elastic net regularization allows us to identify influential predictors.
- We also recommend that EMI invest more resources into maximizing airplay because users are more likely to rate familiar songs positively than unfamiliar ones.
- For their marketing efforts, we suggest that EMI use variables that measure users' music consumption behaviors rather than generic demographic variables because the former are better predictors of music preference.

1. Background

How do you know if someone will love or hate a song? This is the billion dollar question for record labels that pour millions into marketing their next hit single or artist. In the past, managers with "a deep knowledge of the [music] industry but little data to draw on" (The Economist, 2012) made these crucial marketing decisions. Now, record companies have data on music listeners' habits, demographics, and preferences. This means that record labels can allocate their marketing budget efficiently by targeting specific audiences instead of blindly pestering unreceptive consumers.

EMI Music, the record label that launched artists such as the Beatles and the Rolling Stones, is no exception when it comes to the need for smart data-driven marketing. EMI compiled the world's largest dataset (EMI One Million Interview Dataset) on the interests, attitudes, and behaviors of listeners across the globe. In 2012, they released a subset of the data for their "Music Data Science Hackathon" asking participants "Can you predict if a listener will love a new song?"

We decided to take on this challenge. This report presents the results from four classification models (Elastic Net Regression, Support Vector Machine (SVM), Random Forest, and Naïve Bayes Classifier) generated by using the data provided by EMI.

2. Research Question and Hypotheses

Research Question

Using listeners' (a) demographics, (b) artist and track ratings, (c) responses to questions about music preferences, (d) and adjectives that listeners used to describe EMI artists, can you predict whether or not an individual likes a track they just heard?

Hypotheses

H1: Incorporating information on listeners' demographics, music habits, and how people describe artists into a classification model will yield better predictions than randomly guessing

- Can the data provided by EMI in fact be used to answer their question: "Can you predict if a listener will love a new song?" If so, then the models that we generate should perform better than randomly guessing. (Our team categorized the song ratings into three categories, "Like", "Indifferent", "Dislike", and randomly guessing should yield a correct prediction rate of 33%).

H2: Individuals will rate an artist/song more positively if they have heard of the artist before

- Does familiarity lead to fondness? If so, sheer repetition (excessive radio airplay), rather than song quality, can strongly influence listeners to like a song. Thus, record labels would be incentivized to invest most of their resources in earning more airtime on various media, rather than targeting specific demographic groups or regions.

H3: Models including user demographic information (e.g. gender, age etc.) will perform significantly better than models without this information

- Is traditional segmentation that uses demographic data still relevant for record labels or should companies be paying attention to attitudinal factors (e.g. 'I like to be at the cutting edge of new music)?

3. Data Files

The data files made available by EMI are a subset of the EMI Million Interviews Dataset. The sample in the subset is limited to individuals residing in the United Kingdom.

(a) train.csv: This file contains approximately 190,000 ratings from 50,000 individuals of 184 different songs from 50 artists. The song ratings range from 0 to 100 and are our dependent variable of interest (Please see next section on our treatment of the dependent variable).

(b) users.csv: This file contains demographic information about users (gender, age, working status, region of residence), the importance of music to users, estimates of how many hours listeners spending listening to music, and answers to 19 questions about music habits (See Exhibit 1 in Appendix; responses range from 0 - 100 indicating how strongly individuals agree).

(c) words.csv: This file contains answers from 50,928 listeners on questions about whether they have heard of an artist, how much they like an artist, and what words (e.g. “good lyrics”, “fun”, “mainstream”) they would use to describe the artist. There are a total of 82 words and listeners were presented with a different subset (44 - 54 words) during interviews.

4. Pre-processing

Merging data files: The training dataset came in the form of three relational tables with a unique identifier (userid) used to join the tables. Using the *sqldf* package in R*,* we selected the unique user records present in the *train*, *user* and *words* files (174,779 in total).

Missing data: Records with missing values in any of the predictors were removed. Due to time considerations, we did not impute the missing values using SVD++ or multiple imputations. Ultimately, we ended up with 103,236 total records with no missing values.

K-Fold Cross Validation: The remaining 103,236 observations were split into test and training sets. We used *K*-fold cross validation with a *k* of 5. This method was preferred over the “leave-one-out” cross validation because the latter is much more time-consuming.

Dependent variable: Originally, the dependent variable was continuous (distribution shown in Table 1). However, we believe EMI marketing executives would be more interested in whether a listener likes, dislikes, or is indifferent towards a song rather than a specific numeric score. Therefore, we thought it would be best to convert our dependent variable (ratings) into a categorical variable with 3 levels based on the ratings of the song: Dislike (0-33.9), Indifferent (34-66.9), Like (67-100).

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
0.00	14.00	32.00	36.34	50.00	100.00

Table 1. Summary of Ratings

Music habit questions: There were 19 continuous numeric variables in our dataset, which described a user’s music habits (e.g. “I am out of touch with new music”). Each question had very similar distributions with 4 modes and a thin right tail (see Figure 1). Since differences between two close scores are arbitrary (e.g. 22 vs. 26), we changed the numeric response of each question on music habits to a categorical variable with 10 categories using the cut-off points 10, 20, ..., 90, 100.

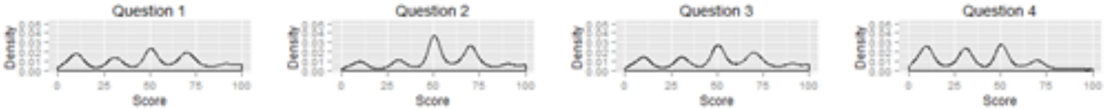


Figure 1: Distribution of Scores for Each Question
(See Appendix: Exhibit 2 for all Q1 – Q19 distributions)

Artist Cluster: After executing the previous pre-processing tasks, there were 33 unique artists remaining in the dataset. The remaining artists were grouped into 5 clusters based on the sentiment of the words that users chose to describe the artists. First, we categorized the 82 words into three groups - negative, neutral and positive (see Appendix: Exhibit 3). Next, we counted how many times a positive, negative, and neutral word was used to describe an artist (see Appendix). Then, we computed the distance between these artists using Kullback–Leibler measure based on the sentiments counts from the previous step. Finally, we used the "complete linkages" method to hierarchically cluster them into 5 groups (see Figure 2). Figure 3 shows the distribution of ratings (Like, Indifferent, Dislike) for each artist cluster (See Appendix: Exhibit 6). We can see that all the artist clusters have a relatively high spike of "Dislike" ratings and that Cluster 5 seems to be the least favorable one.

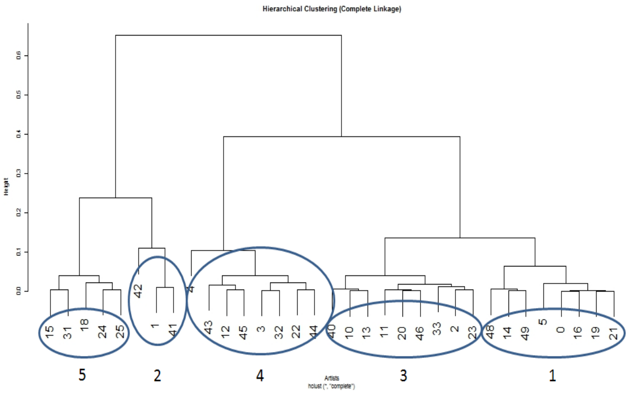


Figure 2: Hierarchical Cluster of Artists

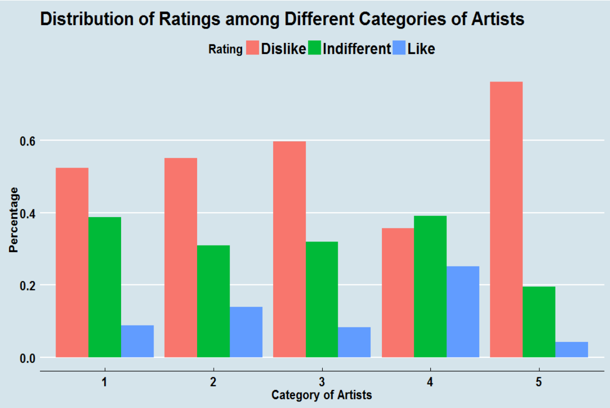


Figure 3: Distribution of Scores for Each Artist Category

Selection of predictors: In our merged data files, we have 30 independent variables in total. Since "User id" and "Track num" are used only for identification and not useful for rating prediction, we excluded them and kept the remaining 28 independent variables in our model. These variables consist of users' attitudes towards artists and a specific track, along with user's demographic information like gender, area, working status and their music habits.

5. Approaches

For this project, we tested four classification models, which varied in complexity, interpretability, and performance: Elastic Net Logistic Regression, SVM (Support Vector Machine), Random Forest, and Naive Bayes Classifiers.

To choose our preferred model, we carefully weighed the advantages and disadvantages for each model. For example, a random forest model is comprehensive and provides information about the importance of different variables but does not inform us about the direction of the relationship. Elastic net regression provides information about whether a variable's effect on ratings is positive or negative. However, it may select redundant (highly correlated) variables. Thus we considered these trade-offs when selecting our final models.

Our analysis was conducted using both R and Python. Our code for the analyses can be found at https://github.com/columbia-w4249-spr2014/team_26/tree/master/Final-Project

I. Regularized Multinomial Logistic Regression (Elastic Net)

Description

In statistics and, in particular, in the fitting of linear or logistic regression models, the elastic net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. The elastic penalty provides a compromise between ridge and lasso, which is in the form of:

$$\lambda \sum_{j=1}^{p}(\alpha\beta_j^2 + (1-\alpha)|\beta|).$$

The elastic-net selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge. It has computational advantages.

Results

After adding a penalty term for regression, the predication accuracy reaches 53.94% (misclassification rate of 46.06%). Specifically, we used the combination of L1 and L2 penalty, which is called elastic net, with tuning parameters $\alpha = 0$ and $\lambda = 1000$.

According to the coefficient plot, we see that (a) having listened to an artist's music *recently* and (b) having listened to an artist's music *ever* both have a significant and *positive* effect on song ratings. Likewise, never having heard of an artist's music before has a significant *negative* effect on ratings.

Actual Prediction	Dislike	Indifferent	Like
Dislike	6966	2943	509
Indifferent	3955	4096	2052
Like	22	40	64

Table 2: Classification Results for Elastic Net Logistic Regression

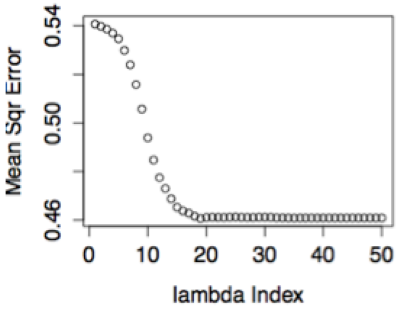


Figure 4: Tuning Parameter

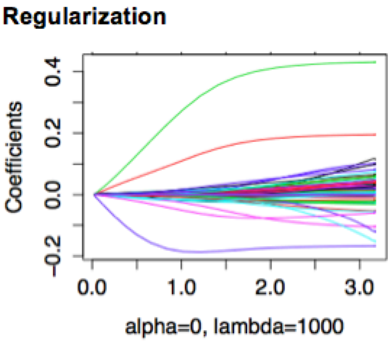


Figure 5: Coefficient Profiles for df(λ)

Model Summary

The elastic net is a novel shrinkage and selection method, which produces a sparse model with good prediction accuracy, and exhibits grouping effects. Using elastic net, our model predicts the correct rating category ("Like", "Indifferent", "Dislike") more than half of the time.

Advantages: Similar to the lasso method, elastic net simultaneously does automatic variable selection and continuous shrinkage. Unlike the lasso method, it is able to deal with situations where p, the number of features, is larger than, n, the number of observations. Also, with elastic net, highly correlated predictors will have similar regression coefficients (i.e. grouping effect). In contrast, the lasso process randomly selects one variable among the highly correlated ones.

Disadvantages: The elastic net estimator is a two-stage procedure: for each fixed α, we find the penalty parameter λ. One disadvantage is that it appears to incur a double amount of shrinkage, which does not help to reduce the variances much and introduces unnecessary extra bias, compared with pure lasso or ridge shrinkage.

II. SVM (Support Vector Machine)

Description

The support vector machine is an extension of the support vector classifier that results from enlarging the feature space in a specific way, using kernels. For a two-class response, the support vector classifier classifies a test observation depending on which side of a hyperplane it lies on. The hyperplane is chosen to correctly separate most of the training observations. This method provides the solution to the optimization problem listed in the Appendix (Exhibit 4).

For the EMI data here, we have three classes for the response variable Rating. The one-versus-all approach is used. That is, we fit three SVMs, each time comparing one of the three classes to the remaining two classes. We assign the observation to the class that has a high level of confidence that the test observation belongs to, rather than any of the other classes.

Because of the huge dataset we have, it is difficult to tune the parameters “Cost” and “gamma”[1], at the same time with the whole data. We choose a subset of the data (10%) to get the best parameters, and fit the model with the selected ones. The cross validation method is used when tuning the best parameters.

Results

The SVM method reaches an accuracy of 67.0% for prediction, which is twice the accuracy of a random guess (33%). The best parameters selected are cost = 1000, gamma = 0.33, with a radial kernel. However, this is achieved from a local range of gamma = (0.33,1,2,3,4)*cost = (0.1,1,10,100,1000), because of the large computational question, we cannot test a global range.

Actual Prediction	Dislike	Indifferent	Like
Dislike	8952	2782	737
Indifferent	1817	3806	812
Like	174	491	1076

Table 3: Classification Results for Support Vector Machine

Model Summary

The SVM model is more accurate than the previous model with a misclassification rate of only 33%. However, it is a “black box” prediction model, so it does not lend much to interpretability.

[1] Cost stands for how much violation is accepted by the model and gamma is the degree of the exponential expression when choosing a “radial” kernel

Advantages: First, it has a regularization parameter, which prevents over-fitting. Second, it uses the kernel trick, so one can build in expert knowledge about the problem via engineering the kernel. Third, an SVM is defined by a convex optimization problem for which there are efficient solutions.

Disadvantages: The SVM model has higher prediction ability but is less interpretable. The SVM function does not explicitly output the coefficients of the decision boundary obtained when the support vector classifier is fitted, nor does it output the width of the margin. In addition, SVM only covers the determination of the parameters for a given value of the regularization and kernel parameters. Also, it is computationally intensive.

III. Random Forest

Description

Random forest is a classification method wherein multiple classification trees are grown and each tree employs a random sample of predictors chosen from the entire set. Unlike a simple classification tree and bagging methods, a random sample of predictors (smaller than the total remaining predictors) is considered at each split. In other words, the random forests method allows us to "de-correlate" our trees and then average them (Complete Description in Appendix). This renders the average of the trees less variable and thereby, more reliable.

Results

We first used simple tree method as a "test run" before running the random forest. We found that the variable 'Heard_of' (i.e. "Have you heard of and/or heard music by this artist?") was the strongest predictor. In fact, the 'Heard_of' variable had such a strong influence that the other variables appeared less useful in our model (they were not in the branches of the tree). As we previously mentioned, simple trees use stronger predictors at higher splits, so we need to randomize the predictors at each split. Therefore, we chose to use random forests which ultimate "de-correlates" trees by choosing a random subset of predictors at each split.

Since there are 30 different variables in total, we chose a subset size of 5 and doing so de-correlated the trees significantly. After running the random forest model on our training set, we obtained these results:

Actual Prediction	Dislike	Indifferent	Like
Dislike	8996	2568	455
Indifferent	1752	3916	918
Like	195	595	1252

Table 4: Classification Results for Random Forest

The overall error rate was around 31.4%, which is the smallest error rate we've obtained.

In addition to accuracy, the random forest model gauges the importance of each predictor in the model. The importance is calculated based on 'mean decrease accuracy' or 'mean decrease Gini index', which is defined as:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

According to the Variance Importance Plot (see Figure 6), ‘Heard_of’ still outweighs the other variables significantly if we use accuracy as a measurement of purity. However, if we use the ‘Gini Index’, then ‘List_Back’ (estimate for the hours per day that the respondent spends listening to background music/music they have not chosen) is the most significant variable. The ‘Gini Index’ also tells us that the variables ‘Heard_of’, ‘List_Own’ and ‘Working’ are the most significant predictors.

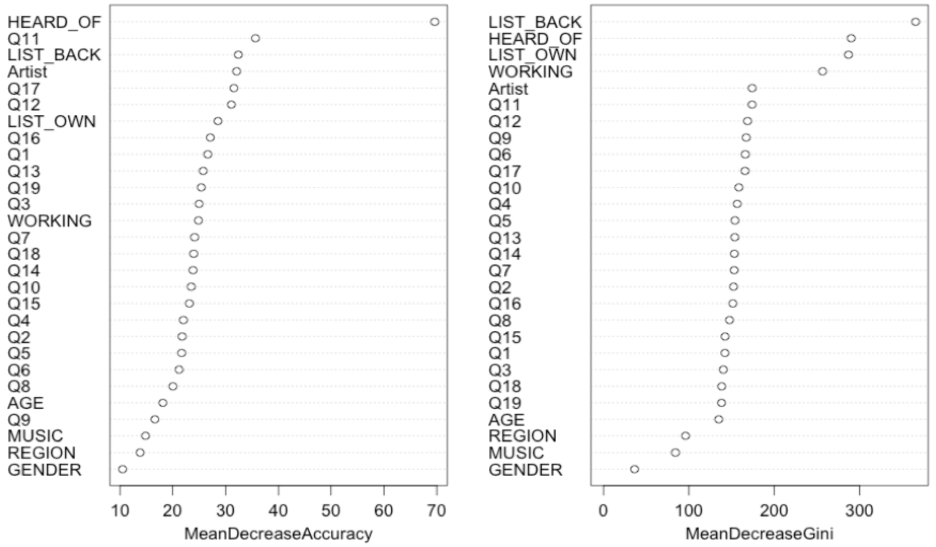


Figure 6: Variance Importance Plots (Accuracy and GINI)

Unfortunately, since the results don’t give us coefficients for each variable, we cannot determine if the effect is positive or negative; we just know that it is a significant predictor.

Model Summary

The random forest model had a relatively low misclassification rate (31.4%) and the Variance Importance Plot tells us that popularity (‘Heard_of’) of an artist may be the most significant predictor. The variables “List_back” and “List_own” are also significant.

Advantages: We decided to use the random forest model because it is known to be one of the more accurate learning algorithms for classification. Second, random forests can handle a large number of input variables without having to resort to variable deletion. And finally, this method gives us estimates as to which variables are important for classification.

Disadvantages: As we mentioned above, the model tells us which variables are important for classification. However, it does not tell us whether the variable will have a positive or negative influence on the dependent variable. For example, our model states that “Heard_of” is a significant predictor, but we do not know whether a user will like a particular song if they heard the songs by the particular artist before.

IV. Other Methods

Other traditional methods for classification include Discriminant Analysis (LDA and QDA) and Bayesian Classifier. Discriminant Analysis assumes a Gaussian distribution of underlying data, and is suitable for continuous exploratory variables. However, most of our independent variables are categorical, so it is not applicable in our case. So, only the Naive Bayesian Classifier approach will be discussed.

Description

Naïve Bayesian analysis is used in predicting categorical responses from mostly categorical predictor variables. The basic idea of naïve Bayesian method is to search over the training dataset to find cases that match exactly the values of predictor variables of input data, and then use the most frequent response of the matched cases for deciding which category the prediction should be in (in our case "Dislike", "Indifferent" and "Like"). A comprehensive description can be found in the Appendix (Exhibit 5)

Result:

The overall misclassification rate of naive Bayesian approach is **52.8%.**

Actual Predict	Dislike	Indifferent	Like
Dislike	6523	2788	630
Indifferent	1583	2356	1129
Like	2837	1935	866

Table 5: Classification Results for Naïve Bayesian Classifier

Model Summary

The Naive Bayesian classifier does not yield a satisfying result. However, it outperforms random guess (67% error rate for three categorical response variables).

Advantage: No stringent assumptions are needed for the distribution and structure of the data.

Disadvantage: Naive Bayes Classifier assumes features are conditionally independent. In our case, a lot of variables, such as Q11 "Pop music is fun" and Q12 "Pop music helps me to escape", might be highly positively correlated. This also explains why Bayesian model didn't perform well compared to our other models.

7. Results

Table 6 below summarizes results from the different models that we generated from our training set to predict whether listeners liked, disliked, or were indifferent to songs in the test set. All models performed better than random guessing, suggesting that information about listener demographics, music habits, and perception of different artists (i.e. words used to describe artists) provide information about listeners' preferences.

Furthermore, we tested naive models that performed even worse than random guessing. This highlights the importance of applying appropriate models to the data. For example, using multinomial logistic regression without regularization, we obtained an accuracy rate of only 14%.

Table 6: Summary of Misclassification Rates for All Models

Model	Error (Misclassification) Rate
Elastic Net Logistic Regression	46.1%
Random Forest	31.4%
SVM	33.0%
Naive Bayes	52.8%

8. Conclusion

Hypotheses

H1: Incorporating demographic and music preference variables into a classification model will yield lower misclassification rate than a random guess.

- All of our models performed better than a random guess for predicting whether a user will like a song. The two best models were the random forest model with an error rate of 31.4% and the SVM model with 33%. Even our worst model (Naive Bayes) beat a random guess with an error rate of 52.8%.

H2: Individuals will rate an artist/song more highly if they have heard the artist before

- The regularized multinomial logistic regression (elastic net) model generates coefficients with the direction of the relationship between variables and song ratings. The two variables with the largest positive coefficients are "Heard_OF_and_listened_to_recently" (0.4266) and "Heard_OF_ and_listened_to_ever" (0.1927).
  Thus, having heard of songs by the artist recently is positively associated with being in a higher category of liking an artist, although we cannot conclude causality.

H3: Models including user demographic information (e.g. gender, age etc.) do NOT perform significantly better than models without this information

- In our models we did not find evidence to indicate that certain user demographic information greatly improves the prediction accuracy of our models. The Variance Importance plot from our random forest model indicates that gender, age and region of residence have low importance for rating prediction.

Recommendations

Maximize airplay (for certain genres): Our analysis is consistent with previous literature that suggests that familiarity with songs increases fondness (Peretz et al., 1998). For artists that appeal to a broader audience (e.g. pop or easy listening), we suggest investing resources to maximize exposure.

Focus on attitudes and behaviors, rather than demographics: Our models suggest that age, gender, or work status are not strong predictors of whether an individual will like a certain song or not. Instead, record companies should invest in learning about individuals' attitudes and perceptions of artists, which is possible through methods such as conducting surveys and analyzing text from social media. However, it should be noted that possibly important variables, such as ethnicity, were not included in the dataset released by EMI.

Further Considerations

Missing Values: Due to time constraints, we decided to omit records with missing values. Future studies should try to "fill in" these missing values using imputation methods.

Downward Bias: The ratings distribution in Table 1 and Figure 3 show that users were more inclined to give artists lower ratings than higher ones. According to a study done by Koenigstein et al. (2011) there might exist downward bias when it comes to users rating songs. In future studies, we might try to use statistical methods to remove any potential downward biases.

Geography: The EMI dataset is restricted to music consumers from the UK. However, the UK is only the third largest music market when it comes to total retail value ($1.3 billion) according to the IFPI 2013 annual report. Future studies could focus on the two largest music markets - United States ($4.48 billion) and Japan ($4.42 billion).

References

Koenigstein, N., Dror, G., & Koren, Y. (2011, October). Yahoo! music recommendations: modeling music ratings with temporal dynamics and item taxonomy. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 165-172). ACM.

Peretz, I., Gaudreau, D., & Bonnel, A. M. (1998). Exposure effects on music preference and recognition. *Memory & Cognition*, *26*(5), 884-902.

The Economist. (2012). EMI Music: data-driven marketing. Retrieved from http://www.economistinsights.com/technology-innovation/analysis/big-data-0/casestudies.

The New York Times. (2013). <http://www.nytimes.com/2013/12/15/arts/music/lorde-rules-a-year-end-list.html?_r=0>

Zhang, Y., & Rajapakse, J. C. (2009). *Machine learning in bioinformatics* (Vol. 4). John Wiley & Sons.

Appendix

Exhibit 1. Music Habit Questions

Q1	I enjoy actively searching for and discovering music that I have never heard before
Q2	I find it easy to find new music
Q3	I am constantly interested in and looking for more music
Q4	I would like to buy new music but I don't know what to buy
Q5	I used to know where to find music
Q6	I am not willing to pay for music
Q7	I enjoy music primarily from going out to dance
Q8	Music for me is all about nightlife and going out
Q9	I am out of touch with new music
Q10	My music collection is a source of pride
Q11	Pop music is fun
Q12	Pop music helps me to escape
Q13	I want a multi media experience at my fingertips wherever I go
Q14	I love technology
Q15	People often ask my advice on music - what to listen to
Q16	I would be willing to pay for the opportunity to buy new music pre-release
Q17	I find seeing a new artist / band on TV a useful way of discovering new music
Q18	I like to be at the cutting edge of new music
Q19	I like to know about music before other people

Exhibit 2: Distribution of Scores for Each Question (All questions)

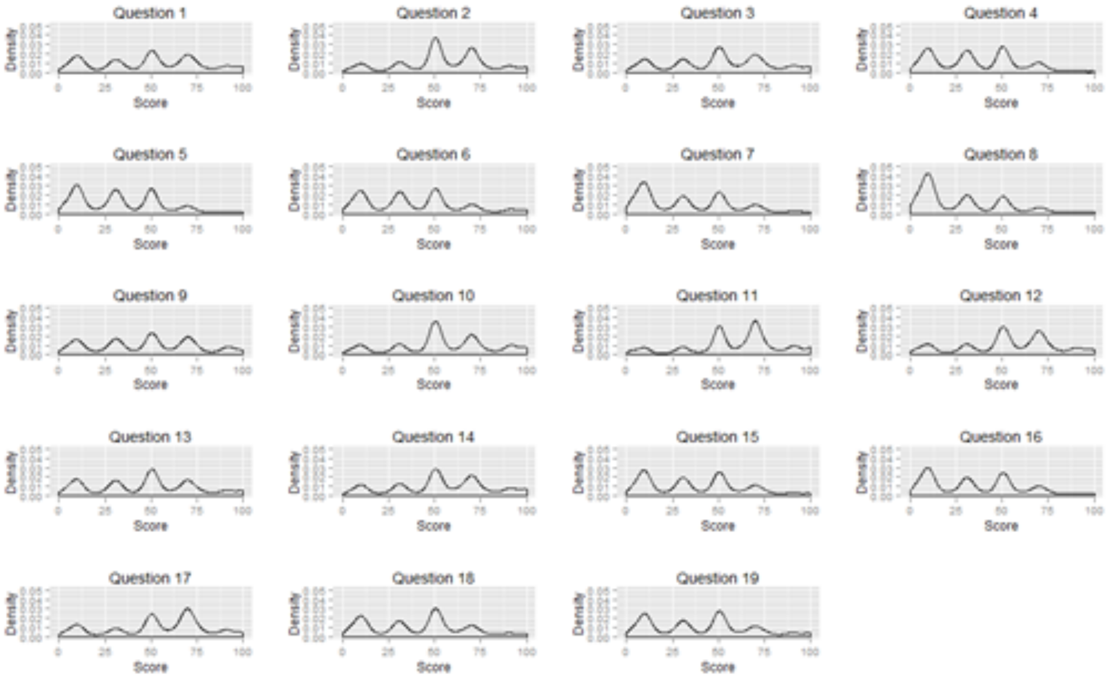


Exhibit 3. Categorization of Words by Sentiment (82 words in total)

- **Positive words**: Sophisticated, Edgy, Sociable, Laid.back, Wholesome, Uplifting, Intriguing, Legendary, Free, Thoughtful, Good.lyrics, Confident, Youthful, Colourful, Stylish, Heartfelt, Calm, Pioneer, Outgoing, Inspiring, Beautiful, Fun, Authentic, Credible, Cool, Catchy, Passionate, Good.Lyrics, Timeless, Original, Talented, Distinctive, Approachable, Genius, Trendsetter, Upbeat, Relatable, Energetic, Exciting, Nostalgic, Progressive, Sexy, Popular, Superstar, Relaxed, Iconic, Classic, Playful, Warm, Soulful
- **Negative words**: Uninspired, Aggressive, Unattractive, Old, Boring, Cheap, Irrelevant, Way.out, Superficial, Annoying, Dark, Not.authentic, Depressing, Noisy, Over, Fake, Cheesy, Intrusive, Unoriginal, Dated, Unapproachable, Arrogant
- **Neutral words**: Outspoken, Serious, Current, Sensitive, Mainstream, Background, Worldly, Emotional, None.of.these, Rebellious

Exhibit 4. SVM: Optimization Problem

SVM solves the optimization problem
"Where C is a nonnegative parameter, M is the width of the margin."

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M$$

$$\text{subject to} \ \sum_{j=1}^{p}\beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1-\epsilon_i),$$

$$\epsilon_i \geq 0, \ \sum_{i=1}^{n}\epsilon_i \leq C,$$

Exhibit 5. Addition Information about Naïve Bayesian Classifier

<u>Naive Bayesian Classifier</u>

Naïve Bayesian analysis is used in predicting categorical responses from mostly categorical predictor variables. The basic idea of naïve Bayesian method is to search over the training dataset to find cases that match exactly the values of predictor variables of input data, and then use the most frequent response of the matched cases for deciding which category the prediction should be in (in our case "Dislike", "Indifferent" and "Like").

The underlying theory of Naïve Bayes is based on the Bayes formula, which can be written as

$$P\,(y=1 \mid X_1=x_1, X_2=x_2, \cdots, X_k=x_k) = \frac{P\,(X_1=x_1, X_2=x_2, \cdots, X_k=x_k \mid y=1)\cdot P\,(y=1)}{\sum_i P\,(X_1=x_1, X_2=x_2, \cdots, X_k=x_k \mid y=i)\cdot P\,(y=i)}$$

This result is exact, and follows basic conditional probability rules. But also this solution is difficult to implement, because with a fine categorization of predictor variables it will be difficult to estimate the conditional joint probabilities $P\,(X_1=x_1, X_2=x_2, \cdots, X_k=x_k \mid y=1)$ and $P\,(X_1=x_1, X_2=x_2, \cdots, X_k=x_k \mid y=0)$. The prior probabilities, $P\,(y=1)$ and $P\,(y=0)$, on the other hand, are easy to estimate. We can use the frequencies from the training set.

For conditional joint probabilities, the Naive Bayesian approach assumes that, the predictors are independent if we condition them on the response. Under this assumption, we can write the Bayes formula as:

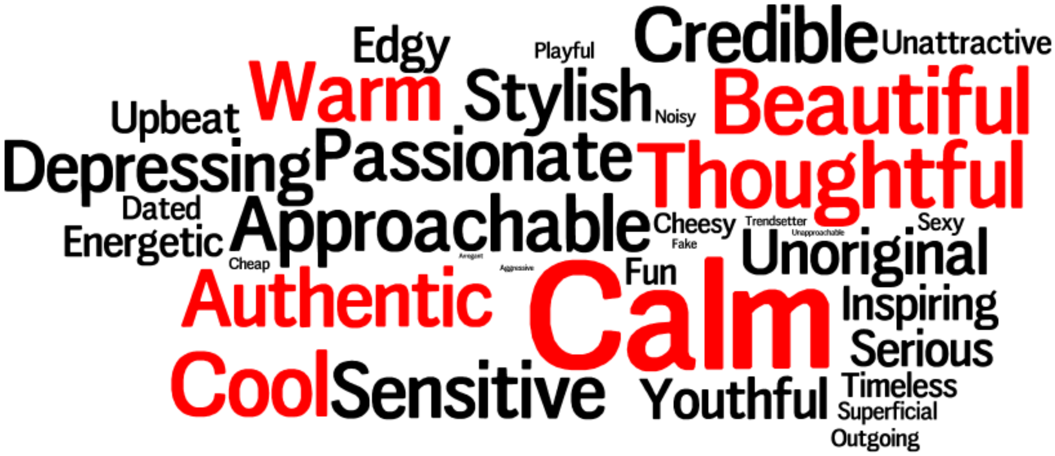
$$P(y=h \mid X_1=x_1, X_2=x_2, \cdots, X_k=x_k)=\frac{P\,(y=h)\prod_{i=1}^{k} P\,(x_i \mid y=h)}{\sum_{j} P\,(y=j)\prod_{i=1}^{k} P\,(x_i \mid y=j)}$$

Exhibit 6. Word Clouds

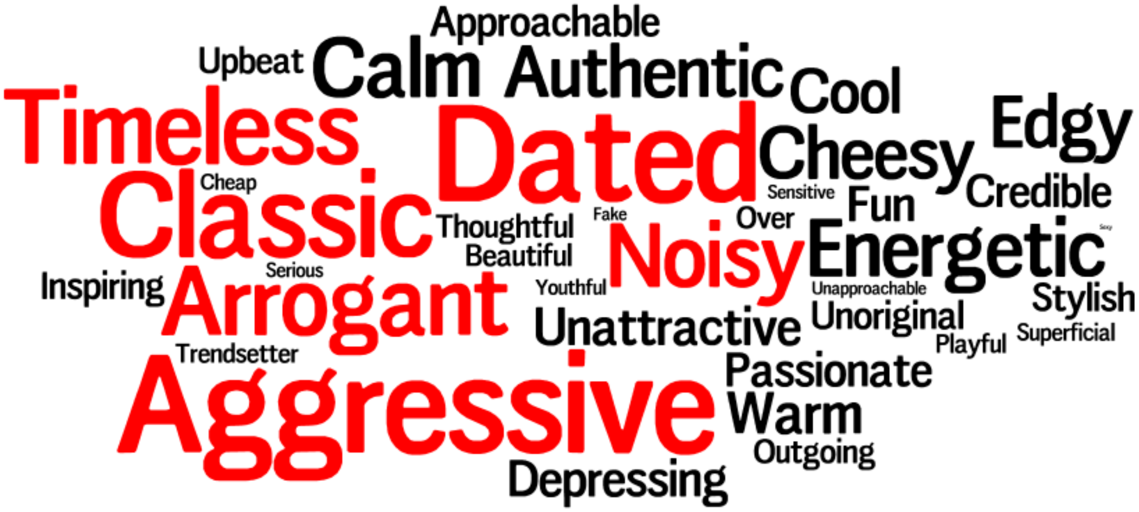
The following words were shared across clusters; therefore, they were excluded from the visualizations below:
Distinctive, Confident, Catchy, Current, Talented, Good.lyrics, Original, Good.Lyrics

The top 5 words used to describe the artists in each cluster are highlighted in red.

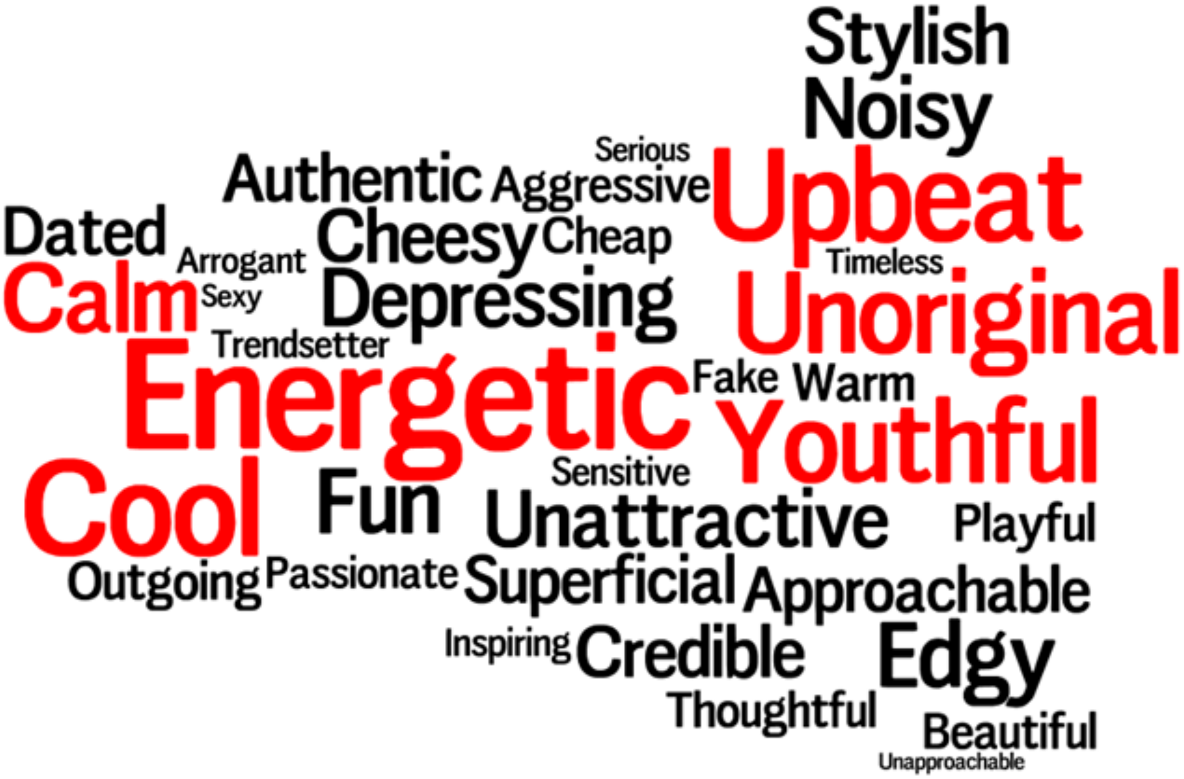
Cluster 1



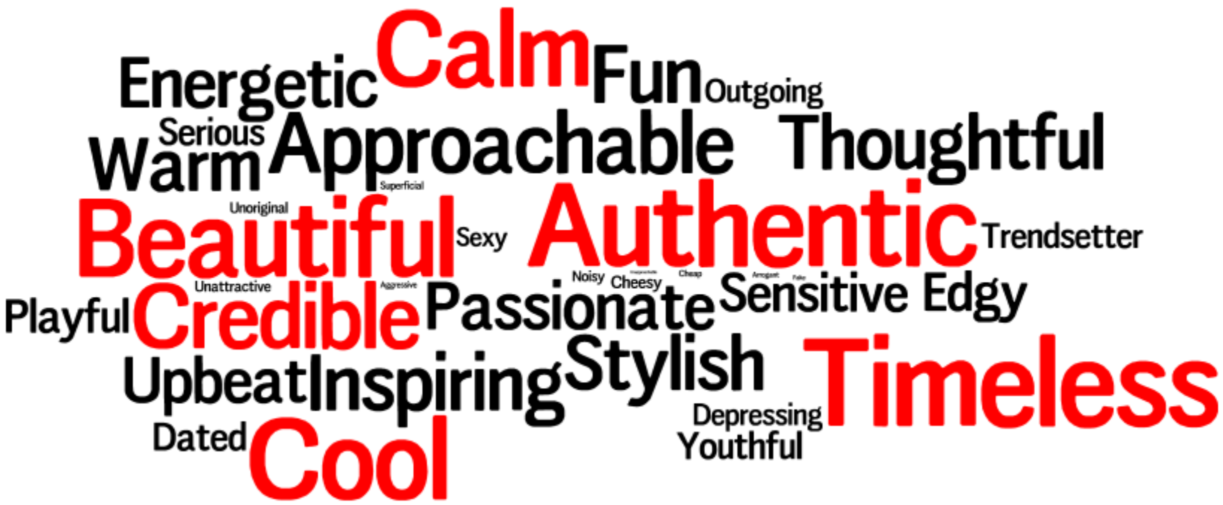
Cluster 2



Cluster 3



Cluster 4



Cluster 5

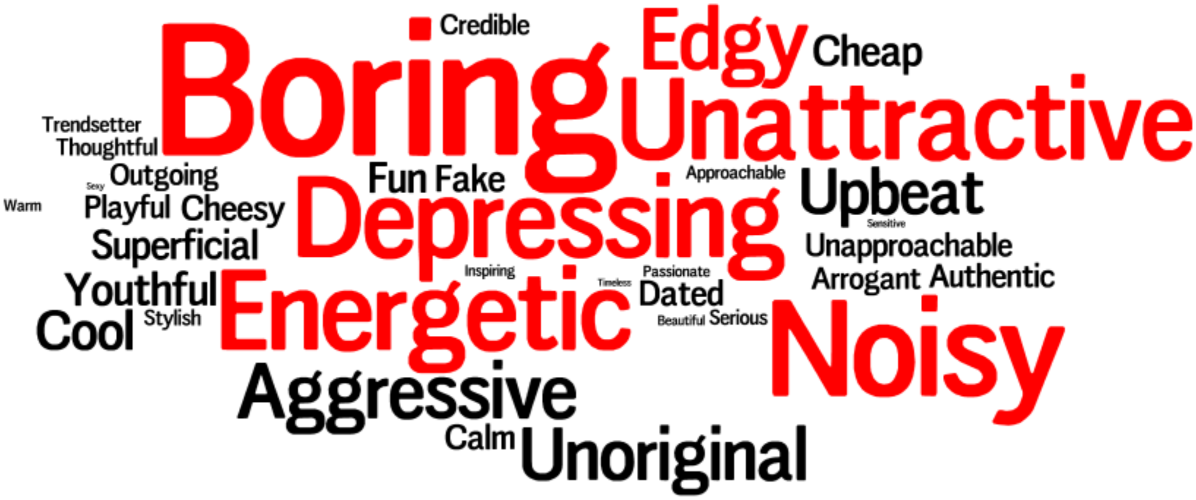


Exhibit 7. Mean GINI Index within the Random Forest Model

The other advantage of using random forest is that it provides us with the importance of each variable. The importance is calculated based on ‘mean decrease accuracy’ or ‘mean decrease Gini index’, which is defined as

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

Exhibit 8. Explanation on parameters selected for Regularized Multinomial Logistic Regression

In Regularized logistic regression, alpha is the parameter for the combination of using L1 norm and L2 norm. Here, the best alpha happens to be 0, so basically it's a lasso.
Lambda in it is the penalty put on adding more parameters. And it's 1000. So actually a lot of parameters were dropped when the model runs.

Exhibit 9. Explanation on parameters selected for SVM

In the SVM model, gamma is the degree we find when using the radial kernel (it's an exponential expression) and c stands for cost, which shows how much violation is accepted. Here cost is selected at 1000, means we do not accept a lot of violations when finding the hyperplane to separate different classes.