# Predicting Song Popularity

Edric Kyauk, Edwin Park, James Pham

## Introduction

Music has been an integral part of our culture all throughout human history. In 2012 alone, the U.S. music industry generated $15 billion. Having a fundamental understanding of what makes a song popular has major implications to businesses that thrive on popular music, namely radio stations, record labels, and digital and physical music market places. Many private companies in these industries have solutions for this problem, but details have been kept private for competitive reasons. For our project, we will predict song popularity based on an song's audio features and metadata.

## Dataset and Features

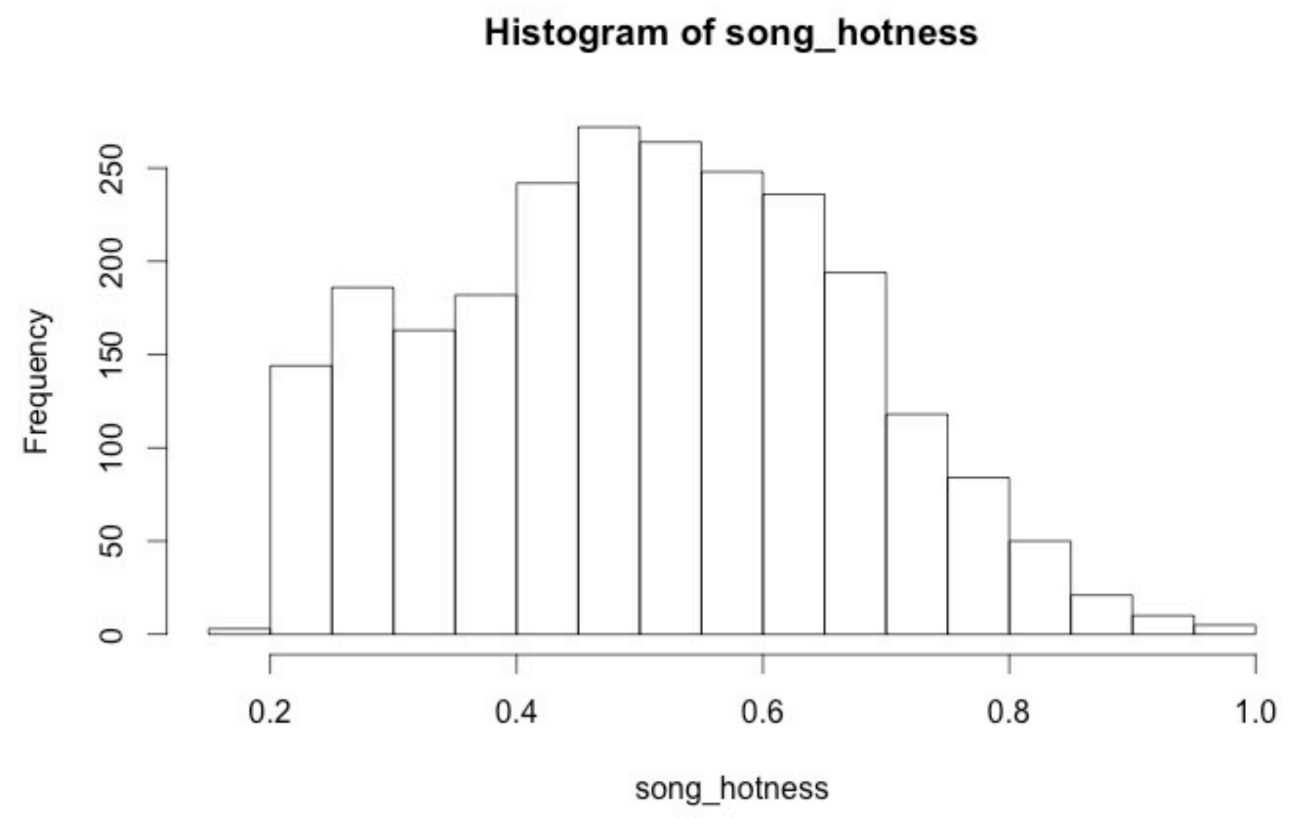
We used music data from The Million Song Dataset. The Million Song Dataset is an exhaustive collection audio features and metadata for a million songs up to 2011. The audio features include attributes about the music track itself, such as duration, key, year. The metadata use more abstract features, such as danceability, energy, or song hotttnesss, generated from The Echo Nest, a music intelligence platform. Our project uses a subset of this data.

### Feature Removal

We also removed features that we knew should have no prediction influences, such as song IDs and artist IDs.

### Metrics

The Echo Nest provides social field for a song called "song hotttnesss" which we will use as our metric of popularity. While the exact calculation of this field is not released, the metric is generally based upon total activity they see for the songs on the thousands of websites that Echo Nest uses.



## Methods & Results

### Baseline Model

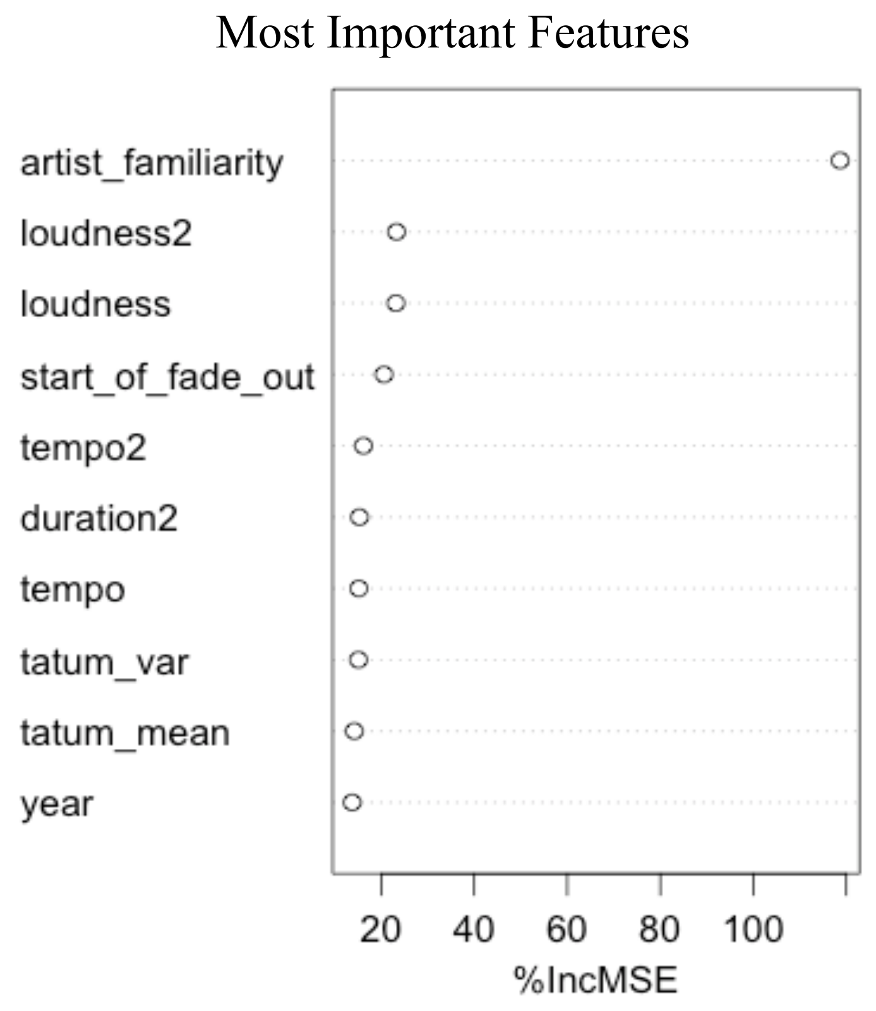
Features included in the baseline model were simply the original features of the million song dataset.

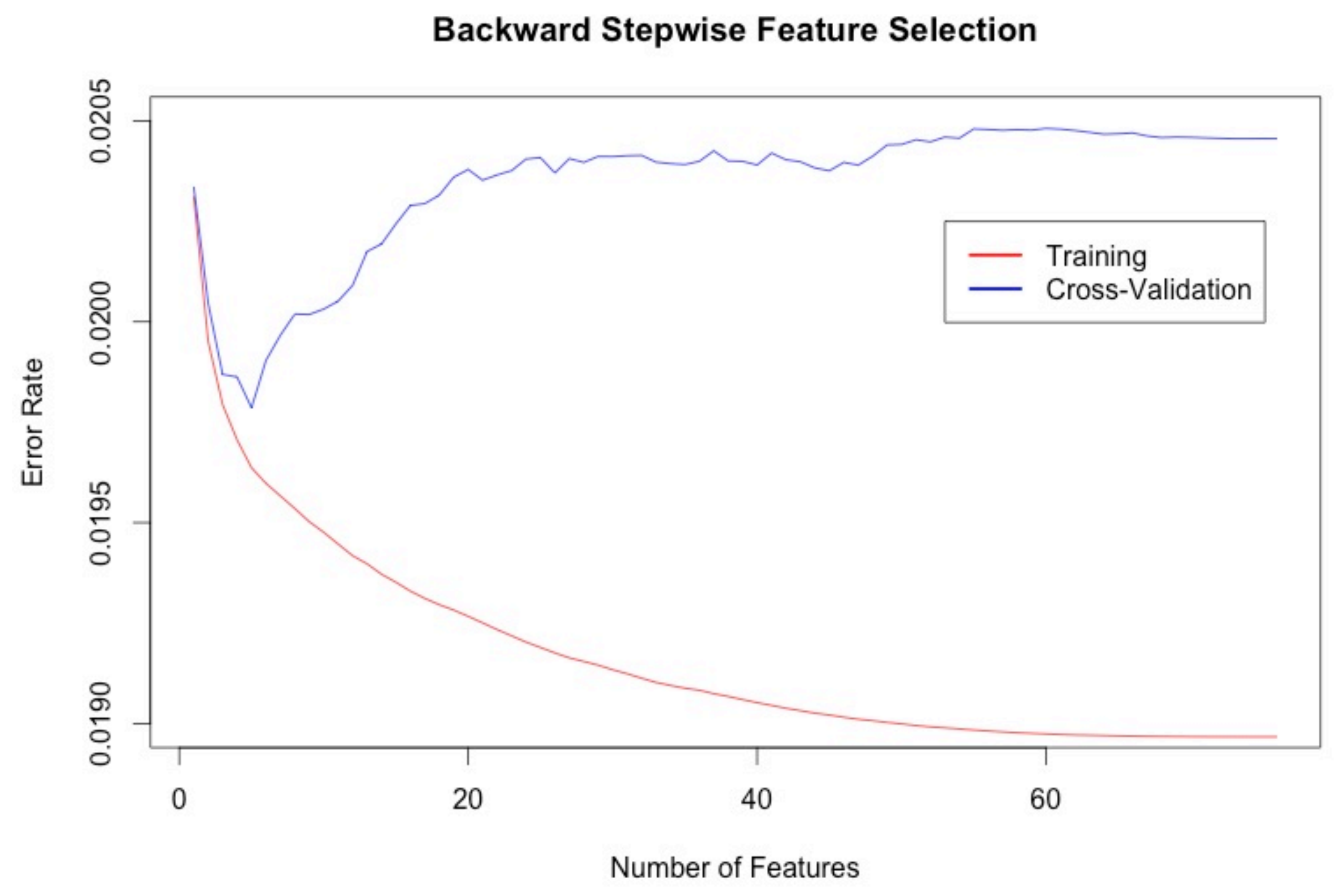
### Full Model: Baseline + Non-linear and Interaction Terms

Recognizing that our baseline model would be unable to capture non-linear relationships, we added the square of certain features, such as loudness and popularity. We also took the product of features such as key and tempo to capture their relationship.

### Feature Selection

Through different feature selection methods, we were able to identify important features in predicting the popularity of a song. Some features were consistently included in the feature set, namely artist familiarity, timbre_mean 5, loudness$^2$, and loudness.





### Classification vs. Regression

We first approached this as a classification problem; training examples were binarized using output values based on a threshold for popularity; however, this approach loses valuable information about the song popularity. As a result, we also used regression to predict the values of the popularity.

**Classification Results**

| Model | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|
| Baseline (SVM) | 0.6268 | 0.5000 | 0.2022 | 0.5 |
| Full Model (SVM) | 1.0000 | 0 | 0 | 0.7465 |
| Recursive Feature Selection (SVM) | 0.6277 | 0.1401 | 0.2282 | 0.7597 |
| Logistic Regression | 0.6765 | 0.3382 | 0.4510 | 0.8042 |
| Gaussian Discriminant Analysis | 0.6250 | 0.2941 | 0.4000 | 0.7902 |

**Regression Results**

| Model | Mean Squared Error | RMSE |
|---|---|---|
| Baseline | 0.0254 | 0.1594 |
| Full Model | 0.0187 | 0.1369 |
| Forward Stepwise Selection | 0.0181 | 0.1346 |
| Backward Stepwise Selection | 0.0176 | 0.1327 |
| Lasso Regression | 0.0185 | 0.1361 |

## Future Work

In addition to using the "song hotttnesss" metric, we can also create our own metric of popularity, which we can define as the number of downloads on iTunes or the number of plays on Spotify. We could also use an n-gram type model and use sequences of pitches/loudness as features.