PROJECT TITLE: Regis University Practicum II – Capstone Project: A Prediction of Oscar winning Best Original Song

PROJECT SUMMARY

I enjoy movies and I enjoy music. It occurred to me this past Oscar season that every year we are given 4-5 songs that all share greatness and then 1 of those songs is ultimately chosen as the Best Original Song. I thought it would be great to predict the winner, but I could only do that if I was given some features of the song. Perhaps its length, or joyousness, or its key, or time signature, or danceability, or energy measures, if I could just get these things, then I could predict Oscar winning songs from any song really, but ultimately from the nominees. Then Spotify released an R package called spotifyr which allows you through an API to access the features of any song that has a unique Spotify number. So I will compile my own dataset of songs that have been nominated and that have won the Oscar for Best Original Song and collect features of these songs and lyrics to perform a prediction of which song will win. I will also be performing a lyric analysis (sentiment analysis, word counts, ngrams) to produce numeric values of text analytics to use as part of my predictor variables. Winning the Oscar will be my target variable.

MILESTONES (Bullet point STEPs in your project. You can tag “DONE” for things are done)

(Example: You can edit for yours. You can also make it more detailed if you want.)

* Data collection (done)
* Text analysis features collected (done)
* Exploratory data analysis (doing)
* Text Analysis features (doing)

This week’s progress:

I have collected most of the data that I will be using for the text and exploratory analytics. I have started to analyze the features and look for interesting patterns. I am also starting to on my organization and making my plots production ready. Last practicum I waited to do this until the end. Text and exploratory analysis will most likely be a two week venture at this point.

Issues and Discussions:

The only issue I really ran into was handling of missing values, some of the songs (4 to be exact) could not be found on Spotify. So because of this, I was missing around 17 columns worth of data for each, so I ended up excluding them from the analysis. I still have over 140 songs that I will analyze, so I do not think it is crucial to keep them at this point.

To Do

* Feature Selection
* SVM Model, Kmeans, and I want to try a Neural Net
* Co-linearity check
* Further exploratory data analysis
* Github repo creation and organization
* Data Dictionary

Resources

* <https://rpubs.com/BreeMcLennan/music_lyric_analysis>
* <https://www.tidytextmining.com/sentiment.html>
* <http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010> AFINN reference
* <http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm> nrc reference

Week 3 Update

PROJECT TITLE: Regis University Practicum II – Capstone Project: A Prediction of Oscar winning Best Original Song

PROJECT SUMMARY

I enjoy movies and I enjoy music. It occurred to me this past Oscar season that every year we are given 4-5 songs that all share greatness and then 1 of those songs is ultimately chosen as the Best Original Song. I thought it would be great to predict the winner, but I could only do that if I was given some features of the song. Perhaps its length, or joyousness, or its key, or time signature, or danceability, or energy measures, if I could just get these things, then I could predict Oscar winning songs from any song really, but ultimately from the nominees. Then Spotify released an R package called spotifyr which allows you through an API to access the features of any song that has a unique Spotify number. So I will compile my own dataset of songs that have been nominated and that have won the Oscar for Best Original Song and collect features of these songs and lyrics to perform a prediction of which song will win. I will also be performing a lyric analysis (sentiment analysis, word counts, ngrams) to produce numeric values of text analytics to use as part of my predictor variables. Winning the Oscar will be my target variable.

MILESTONES (Bullet point STEPs in your project. You can tag “DONE” for things are done)

(Example: You can edit for yours. You can also make it more detailed if you want.)

* Data collection (done)
* Text analysis features collected (done)
* Exploratory data analysis (doing)
* Text Analysis features (done for the moment)
* Organization and explanation of plots (ongoing)
* General organization and write up (ongoing)

This week’s progress:

I started this with week with my exploratory data analysis. With this, I have just been trying to find general trends in the data, I used the correlation plot, pairs plots, and histogram plots, just to get an idea of what my data is doing. Another area that I am trying to explore is the differences in my data in regards to my target variable (songs that won and songs that were just nominated). This hopefully will inform some of my analysis down the road as I try to discover what makes a winning song. I have visualized these differences using boxplots, but I will also analyze them using a t-test to confirm whether my visual hunches were correct. I have already identified visually a handful of features that seem to have significant different medians (means I am unsure of at the moment) but I will use the t-test on those to ensure that the differences I observe are in fact significant.

Issues and Discussions:

I am actually not running into too many issues. I have had to change the format of my data twice, which I have not had to do before. Twice I have had to pivot my data, once to make it wider with my text analysis. I found the sentiment aligned with each word (joy, anger, sadness, anticipation, etc..) and then I calculated what percent of the songs words fit that sentiment and then pivoted out each sentiment to get a column and a value for each song. But when I wanted to make a boxplot of each feature compared to the target variable, I again had to mutate the dataframe to make it a longer dataframe that had three variables; target, feature, value. The feature column had all my original features repeated throughout the column and then the value was the corresponding value.

For example:

Target Feature Value

1 Year 2019

0 Year 2018

1 Acousticness .2

0 Acousticness .1

1 Anger .05

1 Anger .15

To Do

* Feature Selection
* SVM Model, Kmeans, and I want to try a Neural Net
* Co-linearity check
* T-test for significant differences
* Github repo creation and organization
* Data Dictionary

Resources

* <https://rpubs.com/BreeMcLennan/music_lyric_analysis>
* <https://www.tidytextmining.com/sentiment.html>
* <http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010> AFINN reference
* <http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm> nrc reference

Week 3 Update

PROJECT TITLE: Regis University Practicum II – Capstone Project: A Prediction of Oscar winning Best Original Song

PROJECT SUMMARY

I enjoy movies and I enjoy music. It occurred to me this past Oscar season that every year we are given 4-5 songs that all share greatness and then 1 of those songs is ultimately chosen as the Best Original Song. I thought it would be great to predict the winner, but I could only do that if I was given some features of the song. Perhaps its length, or joyousness, or its key, or time signature, or danceability, or energy measures, if I could just get these things, then I could predict Oscar winning songs from any song really, but ultimately from the nominees. Then Spotify released an R package called spotifyr which allows you through an API to access the features of any song that has a unique Spotify number. So I will compile my own dataset of songs that have been nominated and that have won the Oscar for Best Original Song and collect features of these songs and lyrics to perform a prediction of which song will win. I will also be performing a lyric analysis (sentiment analysis, word counts, ngrams) to produce numeric values of text analytics to use as part of my predictor variables. Winning the Oscar will be my target variable.

MILESTONES (Bullet point STEPs in your project. You can tag “DONE” for things are done)

(Example: You can edit for yours. You can also make it more detailed if you want.)

* Learned about class or target imbalance and how that can affect your model
* Created and tested an SVM model
* Cleaned and organized all previous code chunks to make them production ready
* Tested different methods for class imbalance on the SVM model

This week’s progress:

I started with an SVM model on the data to predict winning songs. I realized that my model’s performance was right around 75%, which is right around what my data would be. So, if I predicted that every song lost, then I would be right about 75% of the time. I had high specificity numbers (true negatives) but a very low sensitivity (true positives). I will have to keep exploring with the imbalance techniques to see if they improve the SVM, but I will also try different models/algorithms to see if I can improve from random guessing.

Issues and Discussions:

Besides my model not being better than guessing the target or class imbalance I am not running into too many issues. I do think I will need to study up on how to improve the data imbalance.

To Do

* Kmeans, and I want to try a Neural Net and decision tree
* T-test for significant differences
* Github repo creation and organization
* Data Dictionary
* Comparison of models

Resources

* <https://rpubs.com/BreeMcLennan/music_lyric_analysis>
* <https://www.tidytextmining.com/sentiment.html>
* <http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010> AFINN reference
* <http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm> nrc reference
* <https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/> target imbalance methods in R