1. Name, Contact info (e.g. email/phone).

Ryan M. Allen

719-239-0607

[Rmallen09@yahoo.com](mailto:Rmallen09@yahoo.com)

[Linkedin](https://www.linkedin.com/in/ryan-allen-b66a8069/)

[Twitter](https://twitter.com/AllenRyan91)

[Github](https://github.com/RyanMAllen)

1. Title of the project

Regis University Practicum II – Capstone Project: A Prediction of Best Original Song

1. High level description of the project: what question or problem are you addressing?

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.

1. What type of data science task is it? (some example answers but not limited to)
   * Logistic Regression
   * Text Analytics
   * Data Visualization
   * Kmeans clustering (I will check to see how this method performs)
   * Support Vector Machine
2. Data: Brief description of data. How big do you expect the data will be? Is amount of your data too big or too small? If you're web-scraping or collecting data, how long do you expect to collect the data?

The dataset I will be using will be a handmade dataset of a songs, artists, year, and the Spotify features of a song which include: duration, key, acousticness, danceability, valence, energy, instrumentalness, liveliness, speechiness, loudness, tempo, tatums, segments, sections, beats, lyrics, lyrically created analytic features and the target variable winning or losing the Best Original Song. I have a list of the songs, the artists and the URI (the unique Spotify identifier) and will use the API to collect the rest of the data. I do not anticipate the data being too large. The only concern I have is that after around 1999 the songs become increasingly less likely to be on Spotify. Because of this I might have to consider limiting my data to greater than 1999.

1. How will you analyze the data? What machine learning methods do you plan to use, and/or what business intelligence aspect do you plan on incorporating?

I will perform text analytics, sentiment analysis, word counts, ngrams, to create new numeric variables. After this I will then use SVM, logistic regression, and Kmeans clustering on the data to create three models. Once I have a model for each, I will compare their results from the test data set on ultimately choose a model. I will also perform exploratory data analysis techniques to explain and visualize the data.

1. Describe any anticipated difficulties and problems. Discuss how you may overcome the problems.

Anticipated problems

* + - Songs not being found on Spotify, or intermittently being found on Spotify and limiting my dataset to around 1999-2019.
    - Text analytics I am familiar with elementary concepts (word counts, sentiment analysis) but I would like to try and find some more intermediate level statistics to use with the lyrics.
    - One person had mentioned that there is a political motivation to the voting of the Best Original Song which could prove to be an unmeasurable weight on the target variable.
    - I am uncertain how the models will handle a 1:5 ration of winning to losing target variables. I have done spam prediction in the past where I had an overwhelming amount of non-spam emails and just a few spam emails and my model had difficulty predicting new emails to be spam. I will have to watch out to this, but I do not think that ratio is too small.

1. Suggest a timeline for the project.  This should be a weekly breakdown of what you plan on doing each week.

Week 1 – Proposal, data collection

Week 2 – Text Analytics

Week 3 – Exploratory Data Analysis

Week 4 – Kmeans clustering, SVM, Logistic Models creation

Week 5 – SVM, Logistic Models creation

Week 6 – Github account organization

Week 7 – Write up and general organization of the project and video presentation

Week 8 – Finish write up, presentation, and submit.

Some of the weeks have less “work” then others and can be combined to accommodate for other weeks needing more time and “work”.