## **DS 3000 - Fall 2020**

### **DS Report**

### **Project Title**

#### **Glenn Billman, Shreya Yalamanchili, Sam Zlota**

#### 

#### **Executive Summary: - Shreya**

Add your summary here (100-150 words)

Provide a brief summary of your project. After reading this executive summary, your readers should have a rough understanding of what you did in this project. You can think of this summary in terms of the four sections of the report and write 1-2 sentences describing each section.

## **1. INTRODUCTION - Shreya**

In this section, orient your readers to your project. You've already written some of these in previous deliverables. Based on your final analysis, revise your problem statement and write a concise introduction section. This section should touch upon the following points, but should be written in full paragraphs. Your writing should incorporate all of these points (and more if you like) in a coherent way. Remember that you are trying to convince your readers that this is an important problem to tackle.

Problem Statement

Knowing what kind of music to release can be difficult for artists and producers, so this project can help users explore how popular a given song would be today. Artists would then know what type of music is the most popular, and can release songs accordingly. More specifically, our project is focused on predicting a song’s popularity based on features of the track. In this project, we sought to compare several supervised machine learning algorithms to determine the popularity of a given song. Through our analysis, we would like to learn more about what factors are important in determining how popular a song is, what types of songs are more popular, and why certain songs are more popular than others.

Significance of the Problem

* Why is it important to tackle this problem in your project?
* In what ways could the insights from this project be useful?
* **(optional)** Has there been previous work on your topic in terms of applying ML techniques to analyze similar datasets? Do some research into your topic. What algorithms were used? What was the performance of those algorithms? Cite your sources appropriately. You can use the numbered reference format or APA (if you are more comfortable with it).

Questions/Hypothesis

* End this section with a list of questions and hypotheses
* You should tie these questions/hypotheses to the problem statement and its significance
  + e.g. Given the aforementioned problem and its importance, we set out to tackle the following questions:

**Requirement:**

* You should have at least one question tapping into the comparison of various machine learning algorithms (at least three) in predicting/classifying your target variable from your features variables.
* You should have at least one hypothesis regarding the relationship between two variables, which may be different from your ML problem.

While this section provides bullet points as prompts, your Introduction should consist of coherent paragraphs. As noted in the rubric, you are required to delete all the prompts from this template, except for headings. Failing to do so will result in substantial point deductions.

## **2. METHOD**

### **2.1. Data Acquisition - Glenn**

* Describe where you obtained your data. Provide a link to the original source.
* If you scraped your data, include your code as a **separate** script file.
* Your data should be stored in an online repository (e.g., GitHub) and your code should retrieve your data from that online resource. You can read csv files from the Web in the same way that you read files from your local drive.
* Describe the dataset (i.e., what it is about) and the number of variables/rows included.

### **2.2. Data Analysis - Shreya**

* For your hypotheses, what are your IVs and DVs?
* Specifically describe your predictive model. What outcome variable are you going to predict from what feature variables? Why do you think those are important predictors?
* Describe why this is a supervised ML problem and identify the sub-category of the learning task (e.g. classification).
* What machine learning algorithms are you going to use? **Why?** You should compare at least three algorithms.

## **3. RESULTS**

### **3.1. Data Wrangling - Glenn**

In this section, you should do the following and explain why you are doing what you are doing. For each, you should include your code in a cell, followed by a sample output. For instance, if you are one-hot encoding one of your variables, you should first describe what it is and why you are doing it. You should then include your code in a cell, and the sample output should be available as well.

* Perform simple data cleaning (delete extra columns, deal with NA values, etc.)
* Perform data wrangling to get your features and target values (e.g., grouping your dataframe by columns, applying functions to format dataframes, etc.)
* Preprocess your variables (e.g., scaling/transforming feature variables to normalize them)
* Perform feature extraction (dummy variables, new features from existing features, etc.)
* Use one feature selection technique to select a subset of your original features

### **3.2. Data Exploration - Sam**

* Generate appropriate data visualizations for your key variables identified in the previous section
* You should have at least three visualizations (and at least two different visualization types)
* For each visualization provide an explanation regarding the variables involved and an interpretation of the graph.
* If you are using Plotly, insert your visualizations as images as well (upload the graph images to an online source, e.g. github, and embed those into the cells in Jupyter Notebook). This is a requirement.
* Scatter plots with lines of best fit for each feature variable with target variable
* Bubble chart with artist popularity (magnitude), days since release (x axis), and song popularity (y-axis)
* Histogram of distribution for each variable

### **3.3. Model Construction**

* Conduct your hypothesis test(s) here.
* For your machine learning question(s), use the Training, Validation, and Testing approach through GridSearch
* Apply machine learning algorithms (apply at least three different algorithms)
* Train your algorithms

### **3.4. Model Evaluation**

* Evaluate the performance of your algorithms on appropriate evaluation metrics, using your **validation set**
  + Use at least two different metrics
* Evaluate your results from multiple ML models and hypothesis tests
  + What was the performance of each algorithm in plain English? Is there any indication of overfitting/underfitting?
  + Was there a significant difference? Use the template from lecture slides when reporting the results of your hypothesis testing.

### **3.5. Model Optimization**

* Tune your models using appropriate hyperparameters
* Explain why you are doing this (e.g., to avoid overfitting, etc.)

### **3.6. Model Testing**

* Test your tuned algorithms using your **testing set**

## **4. DISCUSSION**

* Interpret your findings from 3.4., 3.5, and 3.6
  + Which algorithms did you compare?
  + Which algorithm(s) revealed best performance?
  + Which algorithm(s) should be used for your predictive model?
  + Based on your findings, can we use the features in your dataset to predict the outcome variable you identified using the algorithms you've applied? (It is okay if the answer is no. We're interested in the process, not the performance of the model.)
* For your hypotheses, interpret the results. What does it mean to have significant/non-significant differences with regards to your data?
* End this section with a conclusion paragraph containing some pointers for future work
  + (e.g., get more data/features, perform another analysis, etc.)

### **CONTRIBUTIONS**

* Describe each team member's contributions to the report (who did what in each section)
* Remember this is a team effort!
* Each member of your team will provide peer evaluation of other team members. Your final grade on the project will be based on those peer evaluations. A survey will be shared after the deadline for this deliverable.