**Background Information on Predicting Eurovision Finalists**

**Context and Needs of Data Product**

The data product is based off predictive modeling for the Eurovision Song Contest. During the contest, the contestants are placed into two different semi-finals, and ten acts from each semi-final will advance to the grand final. The models I am building will predict which countries pass on to the grand final. There is use for these predictions beyond simply being able to predict the acts. The predictions can be used to inform bets placed on gambling websites. There are dozens of websites on which a person can place bets, and some of them allow you to place bets on which proceed to the grand final. The data product is a dashboard that displays the historical data of the contest as well as the predictions for the current year’s contest, when data is mature enough to be run through the model.

Gambling consumers are constantly looking for tools that will help them get an edge on the betting market, and there are many good existing programs that can help do the job. Some functionality of a software that can be important for a consumer is the ability to see the true odds of a bet in order to make an informed decision and be able to see the returns of bets over time. If my models are effective, they will essentially be showing the correct way to bet—eliminating the need for true odds—and will also show them how much money they could have historically won if relying on my models. Currently when googling for “Eurovision Predictive Models”, there are zero predictive models available for predicting which semi-finalists will qualify to the final; instead, the predictive models available are for predicting the winner of Eurovision overall. Additionally, very few of the models’ code have been made public, and I have found none which involve GUI or dashboard-like displays.

**Purpose and Goals of Data Product**

This data product will serve a several purposes: it will show all gathered data for the Eurovision Song Contest, show which acts are predicted to win, and will provide information on how much money could be won if bets were placed on historical predictions. The justification for including the gathered data is that there are no public datasets with the same information that I have compiled for this project. The purpose of showing which acts are predicted to win is self-explanatory; though there are several algorithms made public, they are generally only using information that is within the scope of the contest itself, like running order and voting history. I have come upon no algorithms that use social media aspects like YouTube and Spotify as my models do. The purpose of showing historical information about betting is key to this project as a big goal is to be able to use the predictions for betting. If information about theoretical historic gains could be shown, it will give more weight to the decision to make bets based on the data.

**Gains from Using Data Product**

Clearly the ultimate gain for using this data product is to be able to have data points for choosing which songs to bet on if a person intends to gamble on the Eurovision Song Contest. These gains are very easy to understand as they are monetary in nature and can be directly measured as profit or losses. There are two different audiences that I can think to identify: gamblers who will make bets on any event and Eurovision fans that will place bets for fun because they enjoy Eurovision. I believe my data product would be of more use to the former audience. This is because gamblers who do not know much about Eurovision would need to gain insight into which acts are more likely to win. Eurovision fans, on the other hand, are more likely to have intuition on which acts will prevail or they may only be interested in betting on their favorite acts. Regardless of which type of gambler uses the tool, the gains will be the same if they trust the model and place bets based on the predictions.

**Data Types and Sources**

My project will source data from multiple places and will mainly be string and numeric. Below is a table that details each variable, what the data type is, where it comes from, and what it represents.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Data Type | Data Source | Variable Description |
| Country | String | Wikipedia | The country represented |
| Artist | String | Wikipedia | The artist singing the song |
| Song | String | Wikipedia | The song title |
| Year | Date | Wikipedia | The year of the ESC contest |
| Semi | Integer | Wikipedia | The semi-final the song is performed in |
| Qualified | Binary | Wikipedia | Indicator of whether the song qualified |
| Place | Integer | Wikipedia | The vote ranking of the song |
| Running Order | Integer | Wikipedia | The order of participation in the semi-final |
| Returning Artist | Binary | Wikipedia | Indicator of if the artist has competed in a previous ESC contest |
| Youtube Views | Integer | YouTube and VidIQ | The rank of number of YouTube views |
| Youtube Upvotes | Integer | YouTube | The rank of the proportion of YouTube upvotes to downvotes |
| Spotify Listens | Integer | Spotify | The rank of the number of Spotify plays |
| Sung in English | Binary | Wikipedia | Indicator of if the song is sung in English |
| Pop | Binary | Chosic | Indicator of if the song is Pop versus another genre |
| Wiwibloggs Rating | Float | Wiwibloggs | Wiwiblogg’s Wiwi Jury score |
| OGAE Rating | Integer | OGAE | OGAE’s international voting score |
| Dance | Integer | Chosic | Chosic’s Dance Score |
| Energy | Integer | Chosic | Chosic’s Energy Score |
| Acoustic | Integer | Chosic | Chosic’s Acoustic Score |
| Happy | Integer | Chosic | Chosic’s Happy Score |
| Speech | Integer | Chosic | Chosic’s Speech Score |
| Live | Integer | Chosic | Chosic’s Live Score |
| Loud | Integer | Chosic | Song Decibels |
| Tempo | Integer | Chosic | Song Tempo |
| Odds | Float | Eurovision World | The odds of a song making it into a semi-final as determined by betting websites |

**Processing and Analysis**

The processing for this project is simply importing the data into Python. The data needs to be split into two different data frames: “x” and “y”. “X” contains all of the predictor variables plus the intercept constant, and “y” contains the response variables. Different models will have different “x” and “y” data frames. The “x” data frames may be reduced in the case of linear and logistic regression models due to variable selection based on significance. The “x” data frame will also be different for the naïve Bayes model because it cannot use negative data. The linear regression “y” data frame will contain the “place” variable, which is continuous, whereas all other models will use the “qualified” variable, which is binary.

There are several different models which are used: linear regression, logistic regression, decision tree, random forest, naïve Bayes, and ensemble. Though none of these models are necessary, it is good practice to try out as many models as possible to see which one provides the best results. The end result of the models it to produce a prediction of whether or not a song qualifies for the Eurovision final from the semi-finals. These results are validated by splitting the data into training and testing and comparing the results between the two; if they have similar performance then it shows that the model can be trusted and is not overfitted.

**Analytical Objectives**

The analytical objectives for this project are two-fold: correctly predicting those who will progress to the finals from the semi-finals as well as maximizing gains from betting. The analytical process will ensure that the expected results will be achieved because I will implement multiple models and see which one has the best accuracy for predicting the progression from the semi-finals to the finals. Additionally, the F1-score will be able to tell us which model will get the best gains from betting; the reason why we want to focus on the F1-score rather than either precision or recall alone is because there are benefits from both having high precision and high recall (Brownlee, 2014). High precision allows us to not waste money on betting for false positives. High recall allows us to place bets on contestants that will actually see returns from bets. As mentioned, the metrics used to assess the extent of the usefulness of this data product is to see how much money can be gained from betting on the predictions of the models; for this, I have gathered the historic odds for betting on various acts and can calculate how much can be gained if the results of the predictive models are used to place bets.