Predicting Eurovision Finalists

Read Me Document

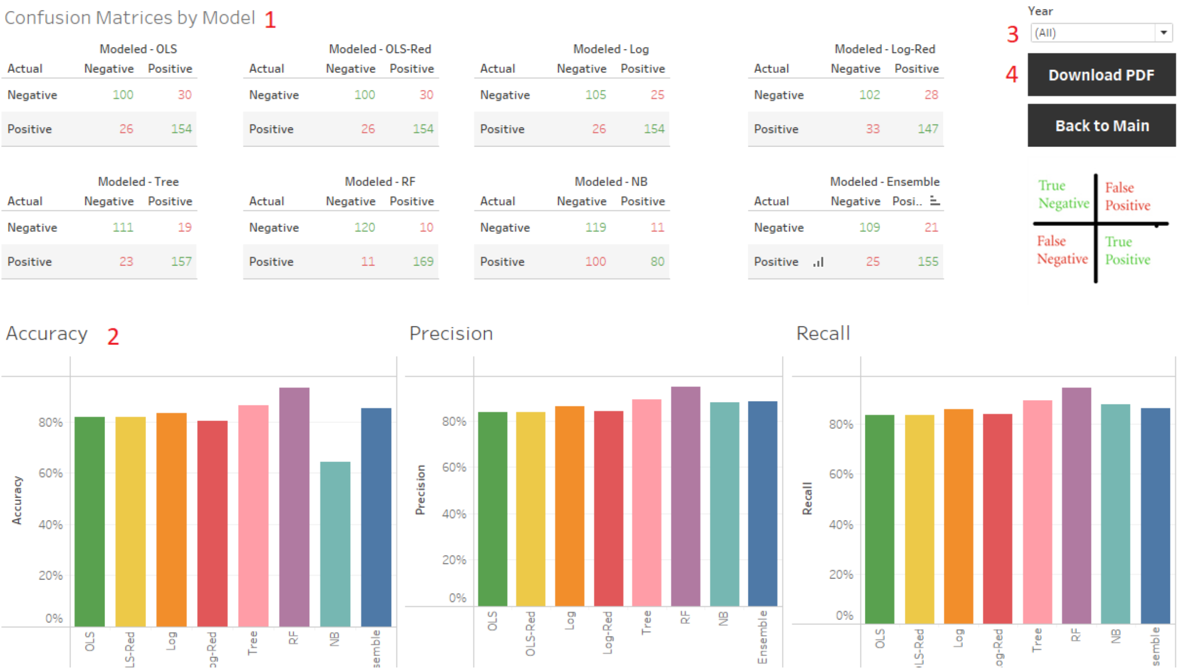
This is my capstone project for my Data Science MS degree at Grand Canyon University. The goal of this project is to build models to predict which Eurovision semi-finalists will qualify for the finals. Eight different models have been built using data from the years 2012-2021 with the variables Wiwi Jury Score, OGAE Score, YouTube Views, YouTube Like Percentage, Spotify Listens, English vs Non-English Language, Pop vs Not Pop, Dance, Energy, Live, Acoustic, Happy, Speech, Loud, and Tempo.

The models that were built are linear regression (OLS), reduced variable linear regression (OLS-Red), logistic regression (Log), reduced variable logistic regression (Log-Red), decision tree (Tree), random forest (RF), naive Bayes (NB), and ensemble. Models that output a probability rather than a binary value use .5 as the cut-off point between classification as qualifying versus non-qualifying.

Tableau Dashboard

The Tableau Dashboard for this project gives information about the performance of the models. It is available on Tableau Public and can be downloaded for anyone to manipulate as they see fit. Below are the views that are built in to the pre-defined view.

The **Model Performance** page will give information about each individual models' performance. You can view these statistics by selecting a given year from the dropdown menu or you may view it in aggregate. A confusion matrix (1) is included to show true positives, false positives, false negatives, and true negatives. From these values we can derive the other metrics: accuracy, precision, and recall (2). You can read more about these metrics and how they are calculated here. You can filter the data by a given year (3) and download the visualization as a PDF (4).



Confusion Matrices by Model 1

	Modeled - OLS	
Actual	Negative	Positive
Negative	100	30
Positive	26	154

	Modeled - OLS-Red	
Actual	Negative	Positive
Negative	100	30
Positive	26	154

	Modeled - Log	
Actual	Negative	Positive
Negative	105	25
Positive	26	154

	Modeled - Log-Red	
Actual	Negative	Positive
Negative	102	28
Positive	33	147

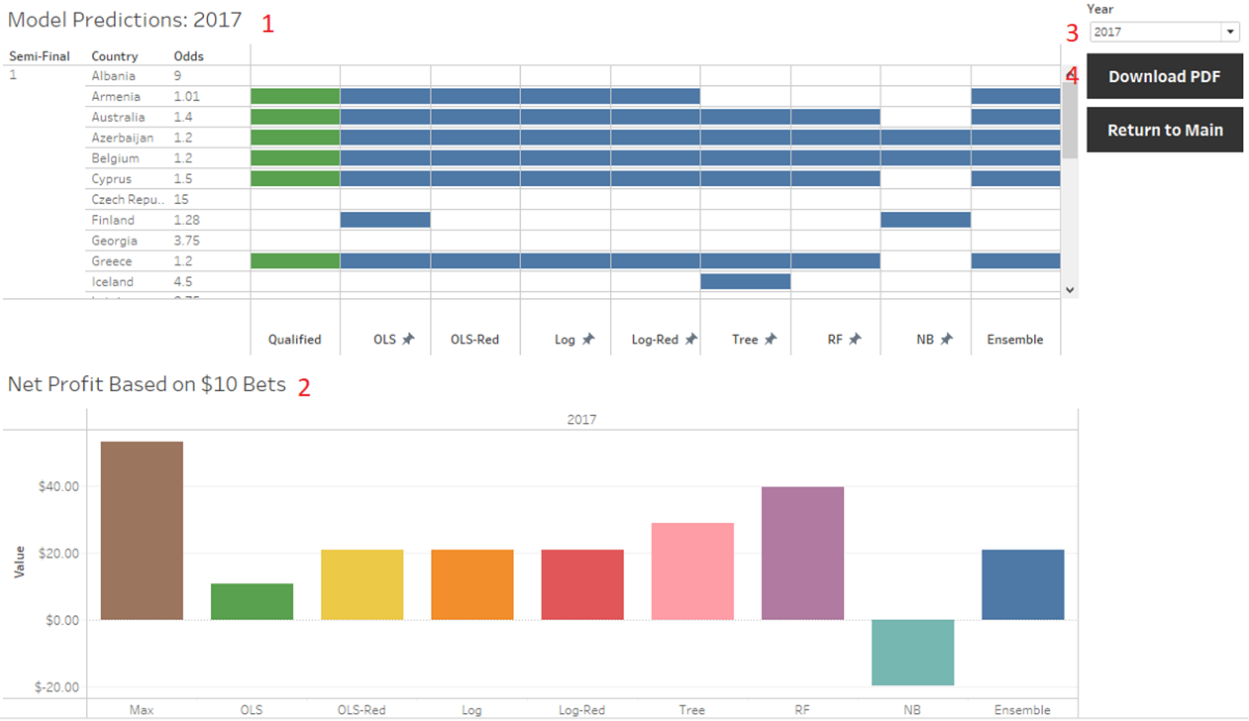
	Modeled - Tree	
Actual	Negative	Positive
Negative	111	19
Positive	23	157

	Modeled - RF	
Actual	Negative	Positive
Negative	120	10
Positive	11	169

	Modeled - NB	
Actual	Negative	Positive
Negative	119	11
Positive	100	80

	Modeled - Ensemble	
Actual	Negative	Posi..
Negative	109	21
Positive	25	155

The **Model Predictions and Betting Profits** page will show additional information about the models' predictions. The top graph (1) shows the contestant qualification as a green bar and predicted qualification as a blue bar by model; if a bar is not present, that means it was not predicted to qualify. You will also notice that the betting odds are shown on this graph as well; betting odds are available from years 2015 onward and were collected from Eurovision World. The bottom graph (2) uses these odds in conjunction with model prediction to determine net profit if a $10 bet was placed on each song the model predicted to qualify. The calculation for a true positive is [$10*(Odds-1)], a false positive is [-$10], and [$0] for true negatives and false negatives. The value "Max" has the theoretical returns if you placed bets on all qualifying songs with no incorrect bets. This view can be filtered by year (3) and downloaded as a PDF (4).

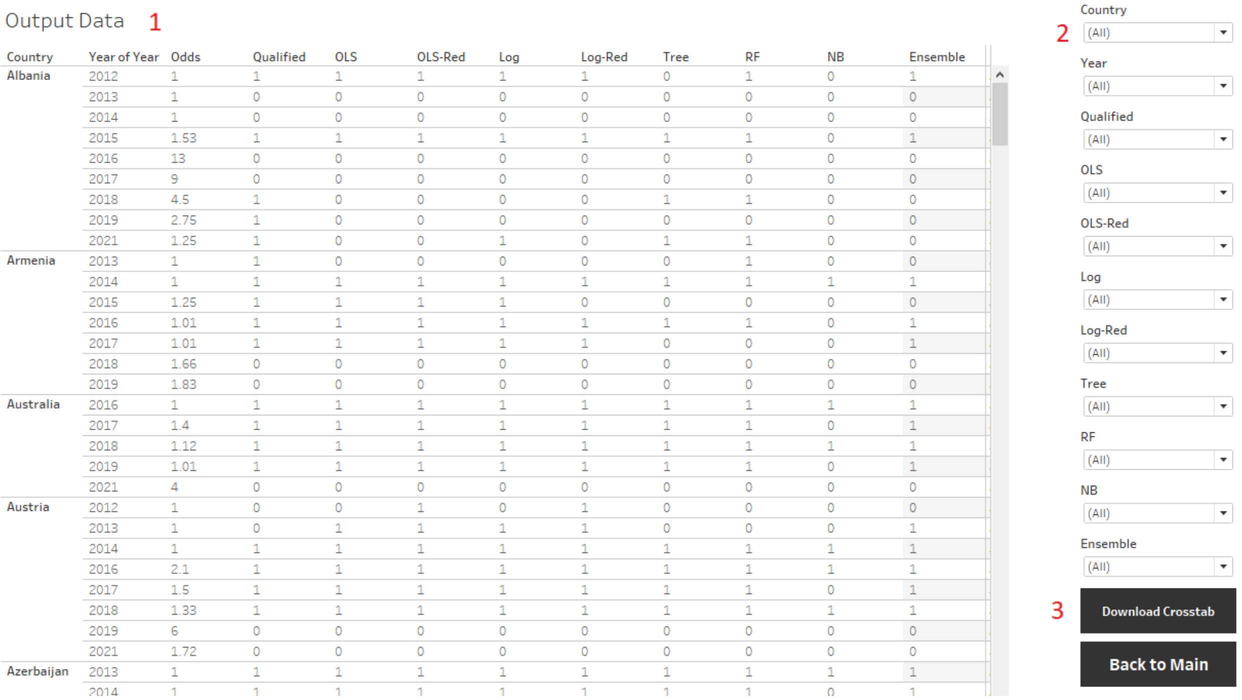


Model Predictions: 2017

Semi-Final	Country	Odds	Qualified	OLS	OLS-Red	Log	Log-Red	Tree	RF	NB	Ensemble
1	Albania	9									
	Armenia	1.01	🟩	🟦	🟦	🟦	🟦				🟦
	Australia	1.4	🟩	🟦	🟦	🟦	🟦	🟦	🟦		🟦
	Azerbaijan	1.2	🟩	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
	Belgium	1.2	🟩	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
	Cyprus	1.5	🟩	🟦	🟦	🟦	🟦	🟦	🟦		🟦
	Czech Repu..	15									
	Finland	1.28		🟦						🟦	
	Georgia	3.75									
	Greece	1.2	🟩	🟦	🟦	🟦	🟦	🟦	🟦		🟦
	Iceland	4.5						🟦			

Year: 2017

The **Model Output** (1) tab contains the information that feeds into this dashboard in a tabular format if you prefer text tables over graphs. You can filter this data by several different variables (2) and can download in a crosstab format (3).

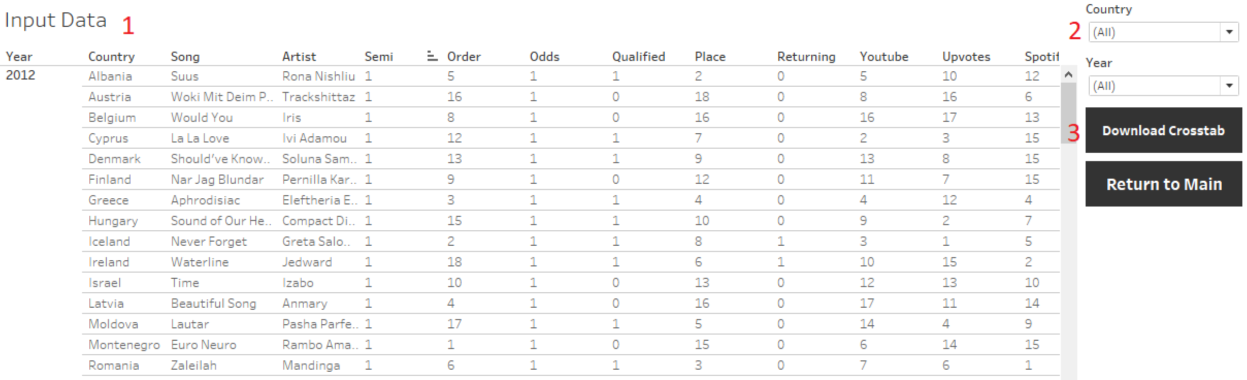


Output Data

Country	Year of Year	Odds	Qualified	OLS	OLS-Red	Log	Log-Red	Tree	RF	NB	Ensemble
Albania	2012	1	1	1	1	1	1	0	1	0	1
	2013	1	0	0	0	0	0	0	0	0	0
	2014	1	0	0	0	0	0	0	0	0	0
	2015	1.53	1	1	1	1	1	1	1	0	1
	2016	13	0	0	0	0	0	0	0	0	0
	2017	9	0	0	0	0	0	0	0	0	0
	2018	4.5	1	0	0	0	0	1	1	0	0
	2019	2.75	1	0	0	0	0	0	0	0	0
	2021	1.25	1	0	0	1	0	1	1	0	0
Armenia	2013	1	1	0	0	0	0	0	1	0	0
	2014	1	1	1	1	1	1	1	1	1	1
	2015	1.25	1	1	1	1	0	0	0	0	0
	2016	1.01	1	1	1	1	1	1	1	0	1
	2017	1.01	1	1	1	1	1	0	0	0	1
	2018	1.66	0	0	0	0	0	0	0	0	0
	2019	1.83	0	0	0	0	0	0	0	0	0
Australia	2016	1	1	1	1	1	1	1	1	1	1
	2017	1.4	1	1	1	1	1	1	1	0	1
	2018	1.12	1	1	1	1	1	1	1	1	1
	2019	1.01	1	1	1	1	1	1	1	0	1
	2021	4	0	0	0	0	0	0	0	0	0
Austria	2012	1	0	0	1	0	1	0	0	0	0
	2013	1	0	1	1	1	1	0	0	0	1
	2014	1	1	1	1	1	1	1	1	1	1
	2016	2.1	1	1	1	1	1	1	1	1	1
	2017	1.5	1	1	1	1	1	1	1	0	1
	2018	1.33	1	1	1	1	1	1	1	1	1
	2019	6	0	0	0	0	0	0	0	0	0
	2021	1.72	0	0	0	0	0	0	0	0	0
Azerbaijan	2013	1	1	1	1	1	1	1	1	1	1
	2014	1	1	1	1	1	1	1	1	0	1

Country: (All)
Year: (All)
Qualified: (All)
OLS: (All)
OLS-Red: (All)
Log: (All)
Log-Red: (All)
Tree: (All)
RF: (All)
NB: (All)
Ensemble: (All)

I have also included a **Model Input** (1) tab so that you can see which variables went into the predictive models. This can be filtered by country or year (2) and can be downloaded in crosstab format (3).



Input Data

Year	Country	Song	Artist	Semi	Order	Odds	Qualified	Place	Returning	Youtube	Upvotes	Spotif
2012	Albania	Suus	Rona Nishliu	1	5	1	1	2	0	5	10	12
	Austria	Woki Mit Deim P..	Trackshittaz	1	16	1	0	18	0	8	16	6
	Belgium	Would You	Iris	1	8	1	0	16	0	16	17	13
	Cyprus	La La Love	Ivi Adamou	1	12	1	1	7	0	2	3	15
	Denmark	Should've Know..	Soluna Sam..	1	13	1	1	9	0	13	8	15
	Finland	Nar Jag Blundar	Pernilla Kar..	1	9	1	0	12	0	11	7	15
	Greece	Aphrodisiac	Eleftheria E..	1	3	1	1	4	0	4	12	4
	Hungary	Sound of Our He..	Compact Di..	1	15	1	1	10	0	9	2	7
	Iceland	Never Forget	Greta Salo..	1	2	1	1	8	1	3	1	5
	Ireland	Waterline	Jedward	1	18	1	1	6	1	10	15	2
	Israel	Time	Izabo	1	10	1	0	13	0	12	13	10
	Latvia	Beautiful Song	Anmary	1	4	1	0	16	0	17	11	14
	Moldova	Lautar	Pasha Parfe..	1	17	1	1	5	0	14	4	9
	Montenegro	Euro Neuro	Rambo Ama..	1	1	1	0	15	0	6	14	15
	Romania	Zaleilah	Mandinga	1	6	1	1	3	0	7	6	1

Country: (All)
Year: (All)

Python Code

The Python code for the models can be found on GitHub along with the original data set for import. You can feel free to download the code and data to run the model as an exercise or even add new predictor variables to the underlying data set to see if you can create a better model.

The modeling code is under Capstone Code.ipynb This code will create the following types of models:
-- Linear Regression
-- Linear Regression, reduced variable model
-- Logistic Regression
-- Logistic Regression reduced variable model
-- Decision Tree
-- Random Forest
-- Naive Bayes
-- Ensemble

The reduced variable models are based on including only variables that are significant from the full-variable models. The decision tree model is limited to 5 levels. The random forest model uses 1000 trees.

If you run this model you will need to change the path to where you put the downloaded data. If you add additional variables to the underlying data or remove them, you will also need to add or remove them to the definitions of *x*, *x_test*, *x_dec*, *x_dec_test*, *x_nb*, *x_nb_test*, *x_year*, *x_2_year*, *x_3_year*, *x_dec_year*, and *x_nb_year*. You will also want to review your significant variables from the linear regression and logistic regression models to input the correct variable selection for *x_2*, *x_2_test*, *x_3*, and *x_3*_test. If you add years to the data, make sure to add running the predictions at the end of the code using the *runmodels()* function. If exporting the data, you will also have to add the year to *pd.concat* function.

Other

For any comments, questions, or concerns, please email **kcgrace89@gmail.com**.

Disclaimer: Predictions shown in this dashboard and produced by the models are not guarantees. The user accepts all risks when using these predictions. Creator is not to be held liable for any monetary gain or loss based on the use of predictions in this dashboard or any that come from the use and manipulation of the Python code.