Limits of predicting the Eurovision winner from pre-final data

Anna-Ilona Nikolaeva, Yauheni Zviazdou and Tim Kicker

Technical report, Introduction to Data Science, University of Helsinki, October 2026

Abstract

We predict how many points every country gives every other country in the Eurovision final, using only what is known the day before: earlier votes between the two countries, Google searches for the entry in the voting country, distance, language and the running order. Summed per entry, the predictions give a scoreboard. Gradient boosting, trained on 2008 to 2019 and chosen on where it put the winners of 2021 to 2023, puts the real winner of the test finals 2024, 2025 and 2026 in first, third and third place. When each of the 18 finals is left out once, the winner lands in its top three in 9 of them and on average at place 5.7, where a random order gives 13.3. Three test finals are few and we settled on this measure only after seeing first results. Below the top the scoreboard stays rough at 6.42 places off against 8.43 for a random order. On the validation finals the searches did most to find the winner, while the lyrics and the age and sex of the artist barely mattered.

1 Question and answer

In the Eurovision final 25 to 27 countries compete and every country taking part that year votes, also those knocked out in the semi-finals. A vote gives 12, 10 and 8 down to 1 points to ten other countries, never to the own one. Up to 2015 each country cast one vote. Since 2016 its jury and its public vote separately and both results are published (Adamska and Reiss, 2025). We wanted to know how well the winner of a final can be predicted from what is known the day before. Our canvas named two groups who would care about that answer, people trading on prediction markets and the contestants themselves.

The model we chose on the validation finals put the real winners of 2024, 2025 and 2026 first, third and third, where a random order would put them around 13th. If each of the 18 finals is left out once and predicted from the others, the winner lands in the top three nine times. Further down the scoreboard the prediction is rough, with every entry 6.42 places off on average against 8.43 for a random order.

2 Data

A row is one directed country pair in one final, so the points Greece gave Cyprus are a different row from those Cyprus gave Greece. In 2024 the Greek jury gave Cyprus 10 points and the Greek public 12, so that row has a target of 11. In the 27 finals before, Greece had given Cyprus 10.4 points on average. We predict rows rather than entries because the 18 finals from 2008 to 2026 (2020 was cancelled) have only 462 entries, which is too few for a dozen features, but 18 069 pair rows. Distance or earlier votes also only exist for a pair. Up to 2015 the target is the single vote of a country and from 2016 it is half of jury plus public, so every year is on the same 0 to 12 scale and the points of an entry add up to half of its official total. Viewers in countries without an entry have voted together as “Rest of the World” since 2023. That is a public vote without a jury and we halve it like any other, which at first we did not do. Half of all pair rows (50.3 percent) are zero, since only ten countries score per vote.

Table 1 lists where the data come from. For a pair we use the logarithm of the distance, whether the two share a border, two measures of shared language and three of colonial ties, together with the mean points and the number of earlier meetings of this directed pair in all finals since 1957. For an entry we use the running order, the age and sex of solo artists and the mood, theme and repetition of the lyrics. A language model labelled the lyrics through an OpenAI compatible interface at temperature 0 and could only answer with fixed categories. The model was gpt-oss-120b on the e-INFRA CZ service.

Google Trends gives the search interest for each entry in each voting country, for web search and for YouTube, over the 30 days up to the day before the final. It reports no search volumes. Instead it scales the country with the most interest to 100 and shares that scale between all terms of a query, which is why every entry is queried on its own (Google). A value therefore tells how much a country searched for an entry compared with the country that searched most. One of us suspected that a time zone shift could pull the final itself into the window. The query runs in UTC and stops the day before. Since the final starts at about 19:00 UTC, even a shifted day boundary ends the window at least seven hours before the show. We checked this in the stored windows for 2009, 2015, 2017 and 2024 to 2026 and in the dates of all 17 finals with search data (2010 has none). The years before 2016 were queried with the same script. Their search data are weak, as both values are zero in 98.7 percent of the rows in 2008, against 72.6 percent in 2016 and 38.0 percent in 2024. For 2010 there are neither searches nor lyrics. Age and sex are missing for a third of the entries, mostly bands.

The pipeline is written in Python with pandas and keeps its tables as Parquet files. One command (`uv run esc`

Table 1: Sources. Every download is listed with address, date and checksum in `sources.yml`.

Source	What we take	Licence
Spijkervet (2023)	results and every vote up to 2023	CC BY 4.0
eurovisionworld.com	votes 2024 to 2026, from the data file its result pages load	no licence given, robots.txt allows
CEPII GeoDist (Mayer and Zignago, 2011)	distance, border, language, colonial ties	Etalab 2.0
Google Trends, via pytrends	search interest per entry and voting country	Google terms, cited
Zviazdou (2026)	lyrics labels, trends 2008 to 2015	GPL-3.0
Wikidata	birth date and sex of solo artists	CC0
own tables	country codes, voting rules, final dates	CC0

`all`) downloads, cleans and joins all sources. Only the download step goes online, every response is cached and `sources.yml` lists every source with a checksum wherever the source has a fixed release. The tests run offline. The pipeline table starts in 2008 and a short notebook adds age and sex and writes it out as one CSV file, so the whole group could work on it without running the pipeline. The prediction notebook reads this file. When we rebuilt it from the pipeline on 17 October, all 27 918 rows matched the file we had used. These rows still hold jury and public separately, which the notebook combines into the 18 069 pair rows. A second notebook plots what winners had in common (age, sex, running order, language, lyrics and searches) and provided the charts for the slides. The only people in the data are the artists, who are public figures.

3 Method

Nothing that is measured after the vote goes into the model. The semi-final points stay out because they are published after the final; the points other countries gave the same entry are the result itself. YouTube view counts stay out as well, since the interface only gives today's count, which includes every view a winner got for winning. Adamska and Reiss (2025) name this leak for their own view counts.

The finals are split by year and never at random, because the rows of one final depend on each other. Training is 2008 to 2019, validation 2021 to 2023 and test 2024 to 2026. Missing values are filled with the median of the training years. With scikit-learn (Pedregosa et al., 2011) we fit seven models, all on 2008 to 2019 and none refitted before the test. Linear regression and lasso run on standardised features, with the lasso penalty set by cross validation over groups of years. A random forest gets the best of 24 settings. The forest and the lasso run a second time without the eight features that the lasso set to zero and the forest barely used. A last forest and gradient boosting (best of 8 settings) leave out the features of the artist and the lyrics. Settings and model are chosen by the average place of the real winners in the validation finals; a tie goes to the higher R^2 on the validation rows. Two models found all three winners there, the boosting and the forest without artist and lyrics. The boosting won the tie with R^2 0.206 against 0.183 and runs with learning rate 0.03, depth 3, at least 50 rows per leaf and 300 rounds. We computed the test numbers of the other models only after this choice, for Figure 1.

The predicted points are summed per entry and sorted into a scoreboard. Our main measure is the predicted place of the real winner. Predicting *winner or not* directly would have to learn from the twelve winners of the training years, which is far too few. A random order puts the winner in the middle of the field on average. For the whole scoreboard we also give the rank error (Adamska and Reiss, 2025), the mean distance between predicted and real place, against a random order drawn 1000 times per final. On top of that we ran two checks. The forest with all features was refitted on the training years without one group of features at a time and judged on the validation finals. And each of the 18 finals was left out once and predicted by the chosen model trained on all the others. Those others include later years, whose earlier votes already contain the year being predicted. Trained on earlier years only, the same check gave almost the same places for 2021 to 2026.

4 Results

On the validation finals (Figure 1, left) the trees put the winners at or near the top, while lasso (places 5, 6 and 5) and linear regression (4, 9 and 12) do not. On the test finals the chosen boosting has Switzerland first, Austria third and Bulgaria third, on average place 2.3 against 13.2 for a random order. The forest with all features gets 1, 2 and 8 and lasso 5, 8 and 6, but with three finals a single surprise moves these averages a lot. On the whole scoreboard the boosting is 6.42 places off against 8.43 for a random order (95 percent range 7.18 to 9.61), with R^2 0.156 on the rows. Lasso and linear regression do slightly better there with 5.88 and 5.93 places, so the trees are only better at the top of the scoreboard. The slides of 15 October still showed 1, 4 and 4 for the test winners, from before we counted Rest of the World correctly.

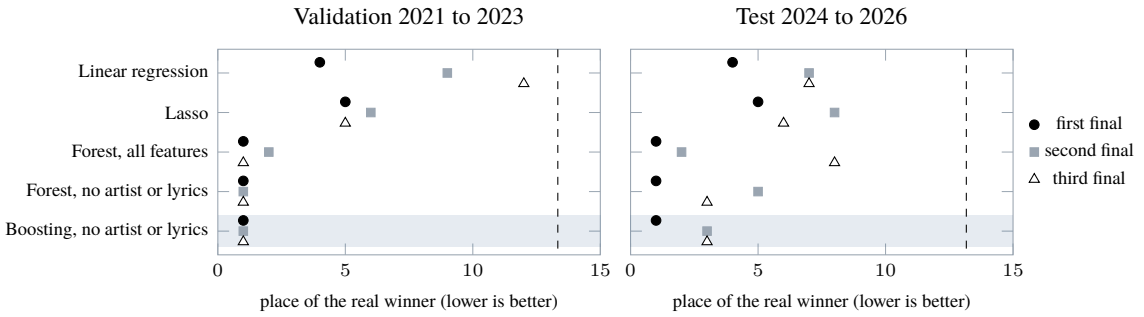


Figure 1: Where five of the seven models put the real winner. The model in the blue band was chosen on the validation finals (left), where it and the forest without artist or lyrics found all three winners. On the test finals (right) it puts the winners first, third and third. Dashed: the middle of the field, where a random order puts the winner on average.

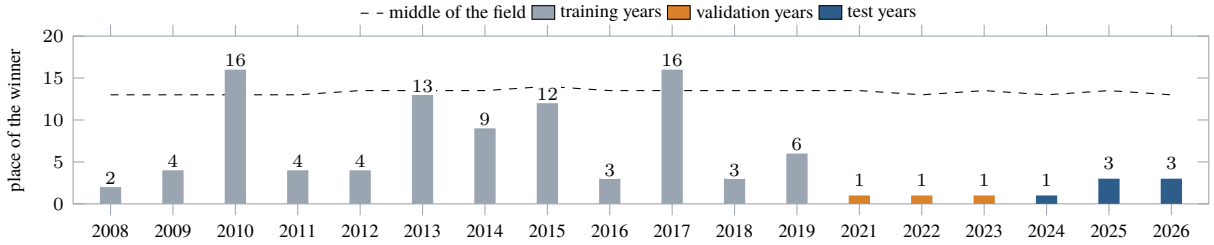


Figure 2: Every final left out once and predicted by the chosen model trained on all 17 others, later years included. The winner lands in the top three in 9 of 18 finals. The settings come from the validation years (orange), so those look better than fresh data would.

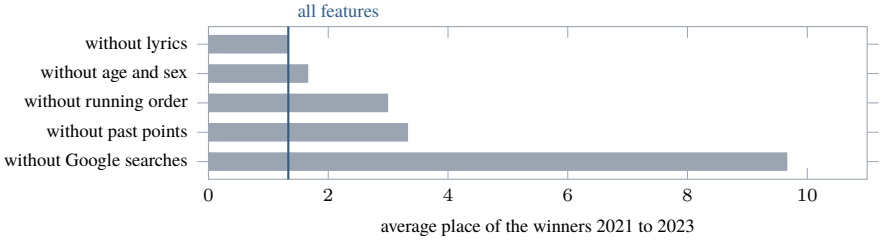


Figure 3: The forest with all features (blue line) against the same forest without one group of features, on the validation finals. A bar ending right of the line means the winners fell without that group. Searches matter most, lyrics and the artist's age and sex barely.

When every final is left out once (Figure 2), the winner comes first in 4 of 18 finals and in the top three in 9, on average at place 5.7 against 13.3 for a random order. The twelve training years on their own average place 7.7 with three winners in the top three. The worst misses are Germany in 2010 (the year without search data) and Portugal in 2017, both 16th, followed by Denmark in 2013 (13th) and Sweden in 2015 (12th). The finals from 2021 on look best. That is partly because their search data are better and partly because the settings were chosen on 2021 to 2023, though we cannot tell how much each contributes.

Figure 3 shows which data find the winner. With all features the forest puts the validation winners at 1, 2 and 1. Without the searches they drop to 2, 9 and 18, without the earlier votes between the two countries to 2, 7 and 1 and without the running order to 7, 1 and 1. Leaving out age and sex costs one winner one place and leaving out the lyrics changes nothing. The years 2008 to 2015 help even with their weak search data, as a forest trained on 2016 to 2019 alone puts the same winners at 7, 12 and 9. How much of the earlier votes is plain neighbourhood would be a question of its own and we leave it aside here.

Our pipeline also predicts jury and public separately for 2016 to 2026, each year left out once with the rest as training. With pair features and searches the jury ranking is 6.40 places off and the public ranking 5.56, against 8.51 and 8.43 for a random order. Without the searches it is 6.77 and 6.41. The searches therefore help the public about twice as much as the jury, whose ranking correlates with the public one at only 0.43 (Spearman). If the earlier votes are computed from the training years alone, these numbers change by at most 0.13 places. They are not directly comparable with Figure 1, as the check is a different one and the feature set was chosen on the same years. For orientation, Adamska and Reiss (2025) also leave out one year at a time (2008 to 2024) and reach 6.07 places in the final with audio, lyrics, YouTube views, running order and voting history, 6.05 with YouTube views alone and 7.59 with running order and voting history. That last set is the closest to ours, while chance in their setup is 8.55. Win probabilities are not in this report. We

Table 2: What the canvas of 14 September planned and what we did.

Canvas	Done	Why
two public Eurovision datasets	Spijkervet up to 2023, eurovisionworld after	the second dataset has no final place, Spijkervet ends in 2023
one row per entry	one row per directed country pair	462 entries are too few, pair features need pairs
features from the songs	video dropped 10 Sep, audio 16 Sep	too much for the course; audio explains little in published work
bookmaker odds	dropped	no longer in the list of data fields of 17 Sep; nobody wrote down why
Wikipedia page views	Google Trends	Trends exists per country, so per pair
migration from Eurostat	dropped 20 Sep	we never confirmed the codes of the table
points normalised per year	half of jury plus public	one 0 to 12 scale in every year
missing values kept as missing	median of the training years	the models need complete rows
all songs since 1956	2008 to 2026 (2016 to 2026 until 15 Oct)	searches start in 2008, more years help

simulated them 2000 times per final from the errors of the public model. Of the eight entries that got more than 50 percent only one won.

5 Limits

Three test finals are few and any difference below one place is within chance. Three validation finals are few as well. Because two models found all three winners there, the one we kept was picked by R^2 and partly by luck. We made the winner's place our main measure only after the first results were in, before that it was the rank error. Along the way the test finals were looked at more than once with earlier versions of the models and features. The final choice was made on the validation finals, but by then the test set was no longer untouched. On the whole scoreboard the chosen model is not the best one either. When we found that Rest of the World counted twice, we fixed it and ran the whole notebook again, choice of settings included. The search data before 2016 are mostly empty and bands have no age or sex. Averaging jury and public also blurs two votes that agree only loosely. People on prediction markets would want win probabilities, which we do not trust enough to report. The code, the notebooks and this report are public at `https://github.com/timkicker/esc-research` and the pipeline rebuilds every table from the sources.

6 Communication and reflection

The spotlight presentation on 15 October explained the prediction to the target group and Anna-Ilona's blog post follows with the submission. What our audience can take from both is how far a forecast from the day before can be trusted and which data matter for it.

Table 2 sets our canvas of 14 September against what we did. The pair as the unit worked well, as did splitting by year and having searches per country. So did the pipeline once it ran on every machine of the group and so did the single CSV file that let all three of us work without it. The lyrics did not work, because a value that is the same for every voter of an entry lets the trees learn the entries by heart. Our first Trends download asked for several terms at once and came back almost all zero, so on 4 October each entry was queried again on its own. Scraping eurovisionworld with a browser failed on a teammate's laptop, which is why the pipeline has downloaded the data file behind the page since 12 October. The others found the pipeline hard to follow, which is how the CSV file came about. Age and sex gave no reliable gain and the win probabilities were badly calibrated.

Besides the workshops, where the TA approved our idea on 10 September, we worked through a group chat. The data sit in one repository and the pipeline in another one that reads them at a fixed commit, with changes going in as pull requests. Looking back we would fix the measure before seeing any results, since the whole scoreboard and the winner tell different stories. We would also keep the test finals closed until the very end, share one simple file early, query Trends one entry at a time from the start and write the main question down in the first week. Our final slide named the next steps we see: a neural network, the video clips themselves, better lyric sentiment and more search sources. We would add cutting the searches two days before the final as Anna-Ilona suggested, a model that learns rankings directly and uncertainty from resampling years. Yauheni asked for a window that ends after the second semi-final, which the current one already does.

We learned that a random split leaks as soon as rows belong together and that anything measured after the event is no forecast. Trends taught us to understand a source before querying it, since it shares its scale between terms. Features that are the same for a whole entry tempt trees to memorise. Reproducible has to mean every machine in the group,

which our first scraper did not manage. And the measure decides how good a result looks, because 6.42 places off and winners in first, third and third place describe the same model.

7 Who did what

Tim built the pipeline with the jury and public models, queried the Trends data again per entry, added the checks, the boosting and the fix for Rest of the World to the prediction notebook and wrote this report. Yauheni collected and labelled the lyrics, added the searches for 2008 to 2015 and the Wikidata fields, extended the pipeline to 2008 and built the CSV file. He also wrote the prediction notebook, switched it to the winner as its measure, wrote the notebook on winners and made the slides. Anna-Ilona ran the first Trends download, worked on the slides and writes the blog post.

References

Katarzyna Adamska and Joshua Reiss. Predicting Eurovision song contest results: A hit song science approach. *Transactions of the International Society for Music Information Retrieval*, 8(1):93–107, 2025. doi: 10.5334/tismir.214.

Google. FAQ about Google Trends data. `https://support.google.com/trends/answer/4365533`.

Thierry Mayer and Soledad Zignago. Notes on CEPII's distances measures: The GeoDist database. Working Paper 2011-25, CEPII, 2011. URL `https://www.cepii.fr/pdf_pub/wp/2011/wp2011-25.pdf`.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Janne Spijkervet. Spijkervet/eurovision-dataset: 2023. Zenodo, version 2023, CC BY 4.0, 2023. URL `https://doi.org/10.5281/zenodo.10088892`.

Yauheni Zviazdou. eurovision-dataset. `https://github.com/ezvezdov/eurovision-dataset`, 2026. Commits e97cefe and e407c54, GPL-3.0.

Use of AI. Parts of this report were drafted with the help of an AI assistant (Claude Opus 5.5 by Anthropic) and revised by the authors.