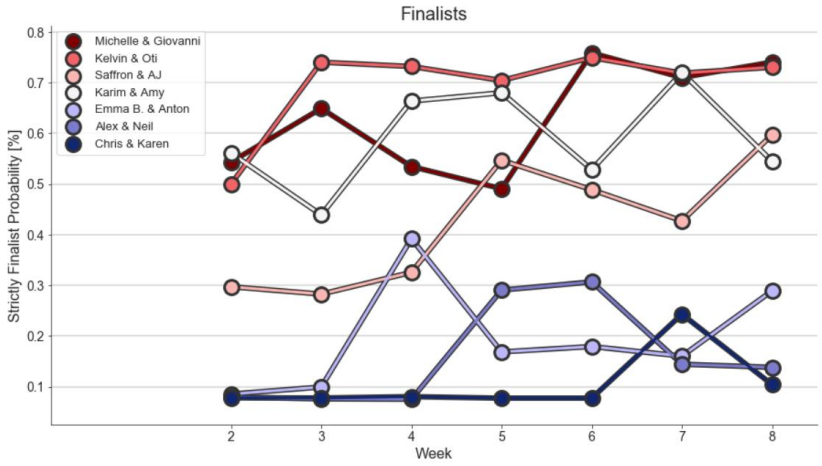
Predictly Come Dancing

This article describes how Intechnica used data science to predict the results of this year's Strictly Come Dancing. For this article we focused on three different predictions: who would make it to the final, who would be in the bottom two each week and who would be the winner. This article outlines how we produced these predictions, from gathering the data to constructing the machine learning models.

Before we get into the details, here are a couple of the results for this season so far. We've been able to correctly predict which couples end up in the bottom two 72% of the time. The current predictions for who is most likely to end up in the final and how this has changed over the course of the competition can be seen in the graph below.



A full explanation of this graph can be found in the results sections below. According to our model, the three couples most likely to make it to the final at this stage are Michelle & Giovanni, Kelvin & Oti and Saffron & AJ.

All the data that is referenced in this article can be found on Wikipedia. The foundation of the training data is taken from results tables across seasons 1-16; the current season on TV is 17. An example of a table that was scraped can be seen below.

Couple	Score	Dance[10]	Music	Result
Ore & Joanne	27 (6,7,7,7)	Cha-Cha-Cha	"Hot Stuff"—Donna Summer	Safe
Claudia & AJ	30 (6,8,8,8)	Waltz	"You Light Up My Life"—Whitney Houston	Safe
Will & Karen	27 (5,7,8,7)	Jive	"Rock Around the Clock"—Bill Haley & His Comets	Safe
Lesley & Anton	26 (6,7,7,6)	Cha-Cha-Cha	"Perhaps, Perhaps, Perhaps"—The Pussycat Dolls	Safe
Greg & Natalie	26 (6,7,6,7)	Tango	"Jump"—Van Halen	Safe
Tameka & Gorka	29 (7,7,7,8)	Charleston	"Yes Sir, That's My Baby"—Firehouse Five Plus Two	Safe
Laura & Giovanni	32 (8,8,8,8)	Waltz	"If I Ain't Got You"—Alicia Keys	Safe
Melvin & Janette	23 (5,6,6,6)	Tango	"Moving on Up"—M People	Eliminated
Louise & Kevin	32 (8,8,8,8)	Viennese Waltz	"Hallelujah"—k.d. lang	Safe
Anastacia & Brendan	22 (4,6,6,6)	Salsa	"Sax"—Fleur East	Bottom two
Ed & Katya	23 (3,7,6,7)	Charleston	"The Banjo's Back in Town"—Alma Cogan	Safe
Naga & Pasha	23 (4,6,6,7)	Cha-Cha-Cha	"A Fool in Love"—Ike & Tina Turner	Safe
Judge Rinder & Oksana	27 (6,7,7,7)	American Smooth	"Marvin Gaye"—Charlie Puth feat. Meghan Trainor	Safe
Daisy & Aljaž	30 (7,8,7,8)	Cha-Cha-Cha	"Forget You"—CeeLo Green	Safe
Danny & Oti	32 (8,8,8,8)	Viennese Waltz	"Never Tear Us Apart"—INXS	Safe

This table is taken from the first week of results from season 14. It shows the couple, the score they received, the chosen dance, music and the result.

In addition to tables containing results, we also extracted all tables that contained a score in weeks where no one was eliminated; in most seasons no one is eliminated in the first week of the show. For each season, all tables on the Wikipedia page are then stacked on top of each other to form a results table that appears like the following.

Couple	Score	Dance	Music	Result	Week
Saffron & AJ	23 (5,5,6,7)	Cha-Cha-Cha	"One Touch"—Jess Glynne & Jax Jones	Safe	2
Anneka & Kevin	19 (4,5,5,5)	Waltz	"Run to You"—Whitney Houston	Safe	2
Dev & Dianne	27 (6,7,7,7)	Jive	"Dance with Me Tonight"—Olly Murs	Safe	2
Emma W. & Aljaž	22 (4,6,6,6)	Tango	"Sucker"—Jonas Brothers	Safe	2
Chris & Karen	26 (5,7,7,7)	Charleston	"Out of Our Heads"—Take That	Safe	2
Emma B. & Anton	24 (6,6,6,6)	Foxtrot	"Sunshine of Your Love"—Ella Fitzgerald & Tomm...	Safe	2
James & Luba	13 (3,3,3,4)	Jive	"Tutti Frutti"—Little Richard	Eliminated	2
Catherine & Johannes	19 (4,5,5,5)	Samba	"Let the Groove Get In"—Justin Timberlake	Safe	2
Michelle & Giovanni	32 (8,8,8,8)	Viennese Waltz	"That's Amore"—Dean Martin	Safe	2
David & Nadiya	10 (2,3,2,3)	Paso Doble	"España cañí"—Pascual Marquina Narro	Bottom two	2
Karim & Amy	32 (8,8,8,8)	Foxtrot	"The Way You Look Tonight"—Frank Sinatra	Safe	2

This table shows 10 rows from the combined results table for the current season.

Using the results table above we extracted information such as the judges scores, total score and the type of dance. We then transformed the table to produce a running aggregate results table. The first 10 columns and top two rows of this aggregate results table can be seen below.

Couple	Week	Total_dances	Bottom_2_sum	Total_score_max	Total_score_min	Total_score_mean	Total_score_weeks_mean	Score_vs_average	Gender	Age
Alex & Neil	2	2	0	22	21	21.50	21.5	-0.866667	1	34
Alex & Neil	3	3	0	23	21	22.00	23.0	-2.928571	1	34

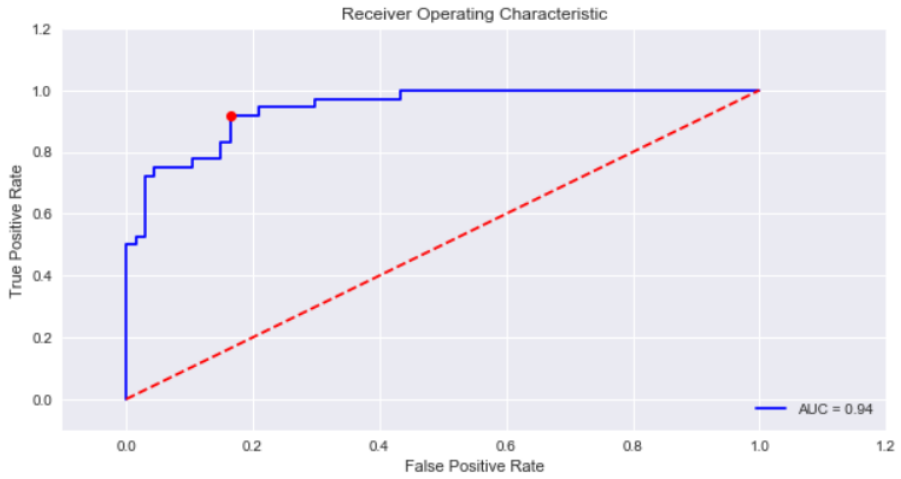
These rows are taken from the latest season and show the results for Alex & Neil from weeks two and three. It is this data, combined with a corresponding label, that is used to train each model.

Due to the relatively small size of the dataset (1508 rows), we were slightly restricted in the number of algorithms that we could consider; however, in many ways this made our job easier. The two algorithms that consistently performed the best were, unsurprisingly, Random Forest and XGBoost.

To tune the parameters for each model we used grid search with customised cross validation. Each season (1-16) had a turn as the validation set with the final model chosen based on its performance on the validation sets for a selection of performance metrics, discussed below.

The first metric we used to assess the models was the average weighted accuracy across each validation set. The reason we used the weighted average is because the size of the validation set varies, depending on the chosen season; season 1 had fewer episodes and fewer contestants than season 16, for example.

Aside from conventional weighted accuracy, we also used the weighted AUC (area under curve) score across the validation sets to get an insight into how well the models separated examples. See the figure below for the ROC (receiver operating characteristic) curve from season 14. The model that generated this ROC curve is used to predict which couples will make it to the final. This curve plots the false positive rate against the true positive rate. The area under this curve corresponds to the AUC score, displayed in the bottom right corner of the graph.



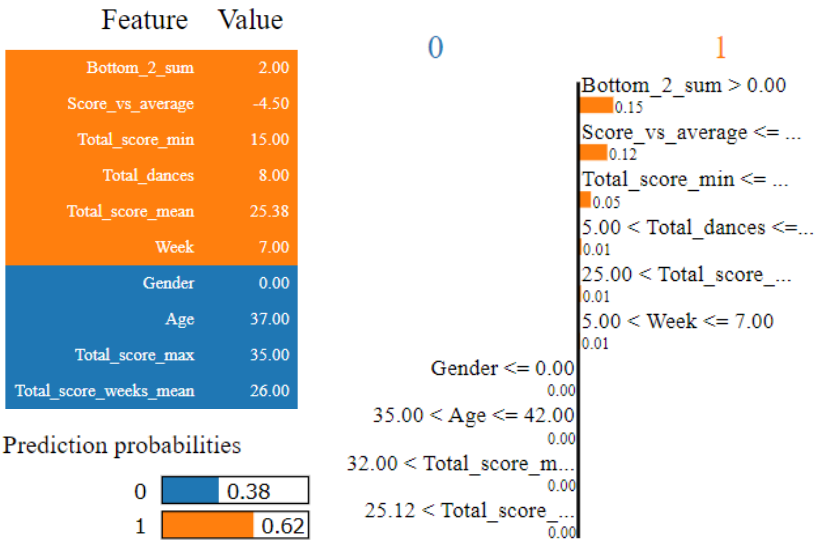
This graph shows the ROC curve and corresponding AUC score for season 14. The model that produced these results was trained on seasons 1-16, excluding season 14. Season 14 was then used as the validation set.

In addition to the two metrics described above, we created our own custom metric that we believe better encapsulates the way the show is run. The custom metric assesses the performance of each relevant model using individual weeks of the validation season. For example, for the model that predicts the bottom two contestants, each week the model orders the contestants based on the predictions and compares the top two contestants (the most likely to be in the bottom two according to the model) to the actual contestants that appear in the bottom two, irrespective of the prediction probabilities. The average accuracy in terms of the top two predicted vs top two actual is then computed across the different weeks. We found this custom accuracy metric useful because it informs us how well the model can differentiate between contestants in the same week. We used the combination of all these metrics to find the optimal parameters for model training.

As well as using the performance metrics described above, we also experimented with SMOTE (synthetic minority oversampling technique) for oversampling. For all the different models we trained the corresponding dataset was imbalanced e.g. when predicting the bottom two there are only two examples per week and some weeks contain up to 15 couples. We found that using SMOTE had mixed results, dependant on the target variable. For the model that predicts finalists, using SMOTE improved the recall of the model and, although there were small decreases in the weighted accuracy, both the average weighted top two validation accuracy and the average weighted AUC score were better. However, for the all the other target variables (bottom two and winner) using SMOTE decreased the score for all metrics. Based on these results, we decided to only use SMOTE for the model that predicts finalists.

For our final test, to ensure the models were producing predictions that intuitively made sense, we used LIME (locally interpretable machine learning - https://arxiv.org/pdf/1602.04938.pdf) to study a selection of predictions.

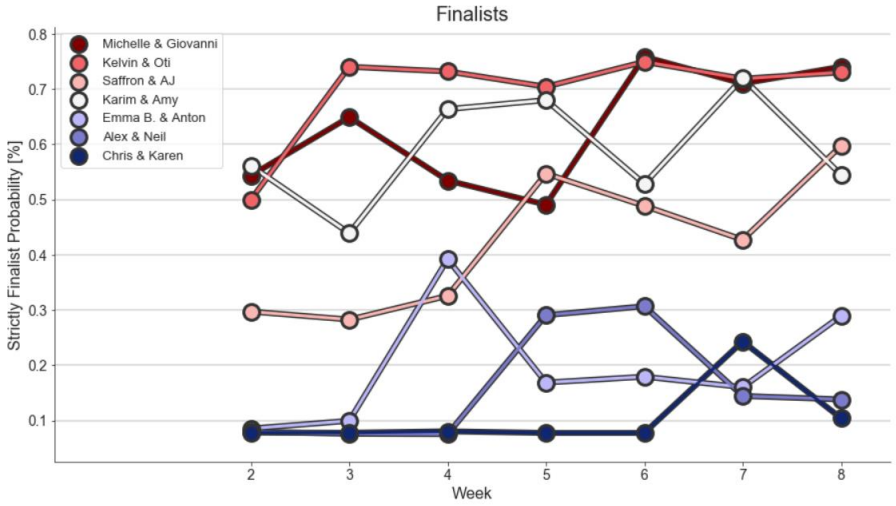
LIME can be useful when trying to explain why any given model has made a certain prediction. LIME gives you a numerical value for each feature that explains how much that feature impacted the specific prediction. There are also handy built in visualisations that make it easy to use. The image below depicts the LIME results for a specific prediction.



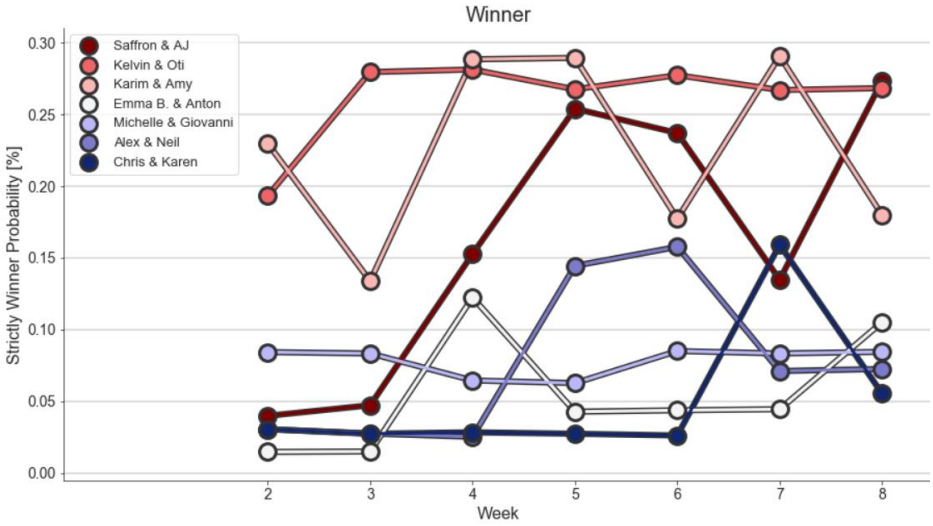
This visualisation gives an example of the results produced by LIME for a specific example. The table on the top left shows the top 10 features and their corresponding value. The colour indicates whether each features contribution was positive or negative. The visualisation on the right shows the impact of the top 10 features and the order indicates the level of impact each feature had. Finally, the display on the bottom left shows what the model's prediction was for the specific example.

These LIME visualisations were produced using the model that predicts which couples will be in the bottom two next week. For this specific example, the couple in question have a 62% chance of being in the bottom two. The main features that contribute to this prediction are the Bottom_2_sum (the number of times the couple have historically been in the bottom two, which in this case is 2) and the Score_vs_average (how the couples average score for the week compares to the average across all couples for the same week. In this case the average score for the week in question was lower than the average score for that week across all couples). The reason we found LIME useful is because it gives us confidence that our models are making predictions based on feature values that make sense e.g. for the example above, the features that contributed to a positive prediction (that the couple will be in the bottom two next week) make sense intuitively.

Below are some of the predicted results at the half way stage of the competition. Note, the inspiration for the graphs below comes from a similar piece of work that predicted the results of Great British Bake Off - https://github.com/dantaki/DeepBake

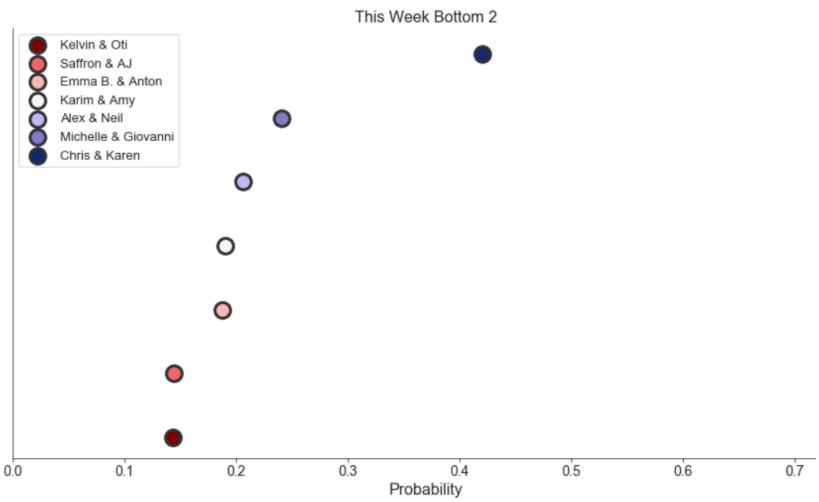


This graph shows the probability (on the vertical axis) of each couple making it to the final and how that has changed over the weeks (on the horizontal axis). For example, Karim & Amy have fluctuated in their likelihood of making it to the final; after week 3 their chance of making it to the final was around 45%, after week 7 they were up to around 70% and now they are hovering just above 50%. Of the couples remaining, the couple least likely to make it to the final are Chris & Karen



This graph is very similar to the one above except it shows the probability of each couple winning the competition (instead of just making it to the final) and how that has changed over the weeks. At this stage of the competition the winners model predicts that Saffron & AJ are most likely to win (with a chance of just above 25%) and Chris & Karen are least likely (with a chance of around 5%).

An interesting observation about the difference between the finalist's model and the winner's model is the difference in predictions for Michelle & Giovanni. Michelle & Giovanni are the front runners when it comes to making it to the final but near the bottom when it comes to winning the competition. Using LIME to study the results for Michelle & Giovanni from the most recent week, we can see that the main thing that contributed to this prediction was Michelle's age (51 according to Wikipedia). This makes sense when you consider the historical data; no one over the age of 41 has ever won the competition. To provide a bit more context, the average age of the contestants across all seasons is 40 but the average age of the winners is 30. In other words, there is a correlation between age and the likelihood to win.



This graph shows the probability of each couple being in the bottom two next week. According to our model, the two most likely couples to be in the bottom two next week are Michelle & Giovanni & Chris & Karen.

Although our models do reasonably well at capturing the general trend, there are some issues, the main one being the fact that we can't directly take into consideration public opinion and therefore public votes. This can have a large impact on the predictions for both popular couples that perform poorly and unpopular couples that perform well. Although this kind of public sentiment analysis would be possible for the current series, it was not feasible for us to gather this data for historic seasons and the models are affected somewhat as a result.