ORIGINAL RESEARCH

# Analytical Explanations of Beauty Pageant Winning Factors

**Victoria Li '24**
Washington University in St. Louis

**Founded in 1952, the Miss Universe pageant is the most viewed and impactful beauty contest. We sometimes wonder why contestants from certain countries have consistently performed so well and placed in the semifinals and finals of the competition. Are there winning strategies? In this paper, we analyze the personal and country profiles of the semi-finalists and finalists of the Miss Universe pageants from 2010 to 2021, examining essential factors for winning. Based on the list of beauty pageant semi-finalists and finalists, we augmented their basic data of (1) personal profile, including height, age, education, and language spoken; and (2) country profile, including population, male/female ratio, tourism revenue as a percentage of GDP, GDP per capita, economic freedom index, and geographic regions. We fill in missing data via smoothing or manual calculation and build a forecast model with XGBoost. Then, we used SHAP to visualize and analyze the importance of each factor's influence on winning and if it is positive or negative.**

**Our results show that GDP per capita and tourism revenue as a percentage of GDP have a strong negative influence on pageant placement, meaning that countries with lower GDP per capita and lower tourism revenue (as a percentage of GDP) tend to achieve better pageant results. This suggests that contestants from developing countries often perform better in the competition. Regarding personal features, contestant height is a key factor in success, while age, education, and language proficiency have minimal impact – likely due to the homogeneity of these attributes among participants. Overall, this study offers insights into factors influencing pageant success and demonstrates an end-to-end approach to data collection, model selection, and explainable AI-driven analysis.**

In 2020, I participated in the Miss Asian Global pageant [4] to gain confidence, improve my public speaking skills, and volunteer for community events. When I was preparing for the pageant, I got to know a few other pageants, the competition format, and the evaluation criteria. The commonly known Big 4 pageants include Miss Universe, Miss World, Miss International, and Miss Earth, each with a different emphasis and purpose. For instance, the Miss Earth pageant focuses on environmental awareness and advocacy, allowing contestants to champion sustainability causes from their home country while promoting environmental initiatives. The Miss Earth pageant then places more emphasis on the contestant's environmental campaign and intelligence and awareness when scoring. While most pageants generally include segments like interviews, onstage question and answer (Q&A), evening gown, fitness wear, and sometimes talent and cultural attire, each one has nuances in what they value more.

Founded in 1952, the Miss Universe pageant [5] is the most widely-viewed and influential beauty pageant in the world. It celebrates women from diverse cultures and backgrounds while promoting positive change through advocacy and philanthropy.

Observing common trends in Miss Universe, I noticed that the contestants from certain countries have been constantly placing well as semifinalists (i.e., Top 16 or Top 25, varies in different years) and finalists (i.e., Top 5 finishes). I started to wonder, are there any important factors or formulas to winning pageants? In this paper, we analyze the personal and country profiles of the semi-finalists and finalists of the Miss Universe pageants from 2010 to 2021 to find out the critical factors for winning.

The rest of this paper is organized as follows: We first explain our initial research on this subject, review prior work, and describe how we collect valuable data to conduct analysis. Next, we augmented contestants' fundamental data of (1) personal profile, including height, age, education, and language spoken; and (2) their country profile, including population, male/female ratio, tourism revenue as a percentage of GDP, GDP per capita, economic freedom index, and geographic regions. We then built a forecast model with the XGBoost algorithm, a powerful and efficient Machine Learning algorithm that works well in handling structured data and identifying complex patterns, making it ideal for predicting pageant winning factors drawn from a mix of demographic, economic, social, and personal attributes. Finally, we used SHapley Additive exPlanations (SHAP), a tool for XAI (Explainable AI), to visualize and analyze the importance of each factor and whether or not it has a positive or negative influence on the factors winning the pageant.

The results are interesting, as they show that GDP per capita and tourism revenue (as percentages of GDP) have the strongest negative impact on winning – suggesting that wealthier countries experience less social impact from pageants compared to developing countries, where such events and exposure may have greater cultural or societal significance. Moreover, we found that the height of the contestants is vital for winning, while the contestants' age, education, and mother language are not. Taller contestants are more likely to win compared to others due to the variance in contestants' heights, but the latter attributes are not as significant because most contestants have at least a bachelor's degree and speak English well so it is hard to draw insights from the minor differences. Also, the contestants' ages are within a small range with an average of 23-24 years old, even though the age requirement is 18 to 28 years old.

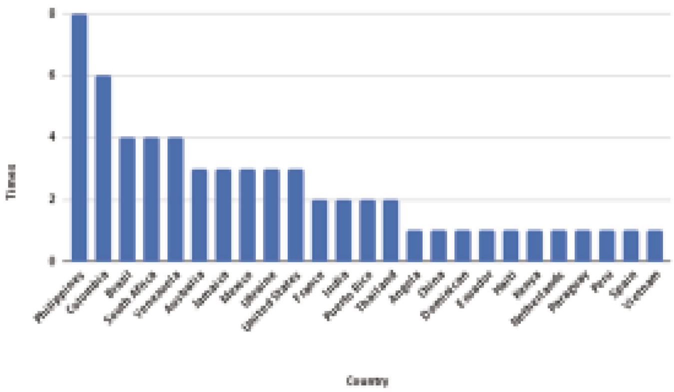


**Fig. 1** | Number of finalists by country, 2010-2021

In addition to providing insightful analysis of beauty pageant winning factors, this paper makes several key contributions. It details the selection of the proper algorithms for this work and the end-to-end practice — from initial research on this topic, to collecting and augmenting data for training the forecast model, and finally, interpreting the results with XAI tools.

## Methods

### Research and Collecting Data

To research this topic, we collected data from the Miss Universe pageant results from 2010 to 2021 and concluded that the Philippines, USA, Brazil, Columbia, France, Venezuela, Australia, Indonesia, Puerto Rico, and South Africa are the top ten countries with the most semi-finalists. The Philippines, Columbia, Brazil, South Africa, and Venezuela are among the top five countries with the most finalists, as shown in Figure 1. We looked for influential factors that could affect pageant placements, such as the contestant's education level, talent training and resources, and language abilities. We also analyzed the country's profile to identify why certain countries or geographic regions cultivate semi-finalists and finalists year after year.

Figure 1: Number of Finalist by Countries 2010-2021

In the early phase of the research, we faced several challenges including insufficient data points and incomplete features. The Miss Universe pageant has been running for 80 years, however, the data from earlier years is hard to find and often incomplete. Many features are unavailable from events before 2010, and the 2020 Miss Universe event was held in 2021. Due to the Coronavirus pandemic, some features and data in 2020 are incomplete. There are less than 200 data points on the semifinalists from 2010 to 2021. We must first fill in missing data items via smoothing or manual calculation. Additionally, the competition format has evolved significantly in recent years. Considering these two obstacles, we focused on pageant results from 2010 to 2021.

To solve the issue of “small data” in contrast to big data, we looked at a research paper by Robert Lawson and Justin Ross [2] that analyzes the role of market liberalism, measured by economic freedom [8], on beauty pageant success. While economic freedom provides valuable insights, such as access to resources, pageant training, and platforms for self-expression, it alone may not fully explain variations in pageant placement. Beauty pageants are complex, multidimensional competitions influenced by various sociocultural, financial, and demographic factors. To capture a more holistic view, we looked for more data sources and augmented basic data to include additional factors to create more extra points. These factors, direct and indirect, include the height and age of the contestant, the country represented, the region or continent represented, GDP per capita of the country represented, Economic Freedom Index, population, tourism percentage of GDP, the language of the country represented, the male and female ratio of the country represented, and the language spoken by the contestant. By integrating these additional factors, we aim to enhance the granularity of our analysis and uncover deeper patterns that could contribute to a more comprehensive understanding of what drives success in international beauty pageants. Most of the required data for country profiles can be found in The World Fact Book [9] and the U.S. Bureau of Economic Analysis [10].

### Data and Feature Augmentation

From the list of beauty pageant semi-finalists (i.e., top 16 or top 21 finishes depending on the year), and finalists (i.e., top 5 finishes), we first collect the pageants' basic data: country, year of the contest, and position, which are available on the Miss Universe website. We augment the fundamental data with:

- Personal profile, including height, age, education, and language the contestant speaks.
- Country and regional profile, including population, male/female ratio, tourism revenue as a percentage of GDP, GDP per capita, economic freedom index, and geographic region (i.e., Asia, North America, South America, Europe, and Africa).

These additional data points are features that help identify factors that play essential roles in winning the pageants as semi-finalists or finalists. We make countries and regions features so that we can analyze if certain countries or regions have a higher chance of winning. This entails labeling each contestant by their country and broader geographical region, then examining the trends in pageant placement outcomes.

The paper by Robert Lawson and Justin Ross [2] is the first attempt to analyze the correlation between beauty pageant winners and the economic freedom index. Due to the lack of available data points and features, they augmented the data with derived statistical features, such as average, maximum, minimum, variance, etc.

Compared to their work, we extend the feature space by adding personal and country profiles. Our approach is more advanced and produces more accurate analysis results, as it considers additional factors that potentially impact pageant success. We can future augment our training data set with statistical features as well if they prove useful in improving forecast accuracy.

The format of the datasets used for training is shown in Figure 2. Using smoothing techniques or manual calculation, we fill in missing data items, such as the Economic Freedom Index for certain countries in certain years. We then build a forecast model with the XGBoost algorithm [11] to train on various features, like contestant demographics or countries' economic indicators. Finally, we used SHapley Additive exPlanations (SHAP) [3] to analyze and visualize the importance of each factor and if these factors positively or negatively influence winning the pageant.

| Country Represented | Semi-Finalist Name | Population (millio | Per Cap GDP ($t | Economic Freedom | Male/Female Ratio | Country Region | Tourism (% of GDI | Year | Position | Height | Born | Age | Top 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| India | Harnaaz Sandhu | 1399.716988 | 1.75 | 6.66 | 1.02 | Asia | 6.8 | 2021 | 1 | 171 | 2000 | 21 | 1 |
| Paraguay | Nadia Ferreira | 7.259565 | 5.2 | 7.22 | 1.017 | Latin America | 14.9 | 2021 | 2 | 175 | 1999 | 22 | 1 |
| South Africa | Lalela Mswane | 60.394854 | 7.08 | 6.97 | 0.982 | Africa | 8.7 | 2021 | 3 | 175 | 1997 | 24 | 1 |
| Colombia | Valeria Ayos | 51.670809 | 7.7 | 6.82 | 0.968 | Latin America | 5.6 | 2021 | 4 | 176 | 1994 | 27 | 1 |
| Philippines | Beatrice Gomez | 111.699463 | 3.16 | 7.42 | 1.006 | Asia | 12.7 | 2021 | 5 | 175 | 1995 | 26 | 1 |
| Aruba | Thessaly Zimmerman | 0.107424 | 30.253 | 0 | 0.90285 | Africa | 63 | 2021 | 6 | 170 | 1995 | 26 | 0 |
| France | Clémence Botino | 65.483478 | 42.6 | 7.55 | 0.948 | Europe | 8.5 | 2021 | 7 | 175 | 1997 | 24 | 0 |
| Puerto Rico | Michelle Colón | 2.752463 | 34.14327 | 8.24 | 0.92 | North America | 6.9 | 2021 | 8 | 183 | 2000 | 21 | 0 |
| The Bahamas | Chantel O'Brian | 0.398662 | 25.7 | 7.56 | 0.9455 | North America | 40.3 | 2021 | 9 | 170 | 1994 | 27 | 0 |
| United States | Elle Smith | 333.823623 | 56.2 | 8.24 | 0.97 | North America | 7.8 | 2021 | 10 | 175 | 1998 | 23 | 0 |
| Great Britain | Emma Collingridge | 68.404314 | 41.1 | 8.15 | 0.97 | Europe | 10.9 | 2021 | 11 | 168 | 1998 | 23 | 0 |
| Japan | Juri Watanabe | 125.914522 | 35.2 | 7.98 | 0.949 | Asia | 7.5 | 2021 | 12 | 170 | 1994 | 27 | 0 |
| Panama | Brenda Smith | 4.413024 | 11.8 | 7.79 | 1 | North America | 14.9 | 2021 | 13 | 170 | 1994 | 27 | 0 |
| Singapore | Nandita Banna | 5.917077 | 59.5 | 8.81 | 1.017 | Asia | 10 | 2021 | 14 | 176 | 2000 | 21 | 0 |
| Venezuela | Luiseth Materán | 28.318786 | 9.5 | 2.83 | 0.98 | Latin America | 9.4 | 2021 | 15 | 178 | 1996 | 25 | 0 |
| Vietnam | Nguyen Huynh Kim D | 98.32 | 2.03 | 6.26 | 0.978 | Asia | 9.2 | 2021 | 16 | 173 | 1995 | 26 | 0 |
| Mexico | Andrea Meza | 130.26 | 10.18 | 7.21 | 0.96 | North America | 17.3 | 2020 | 1 | 180 | 1994 | 26 | 1 |
| Brazil | Julia Gama | 214.13 | 8.72 | 6.56 | 0.97 | Latin America | 8.1 | 2020 | 2 | 177 | 1993 | 27 | 1 |
| Peru | Janick Maceta | 33.36 | 6.3 | 7.76 | 0.96 | Latin America | 9.4 | 2020 | 3 | 177 | 1994 | 26 | 1 |
| India | Adline Castellino | 1394.69 | 2.1 | 6.56 | 1.08 | Asia | 6.8 | 2020 | 4 | 168 | 1998 | 22 | 1 |
| Dominican Republic | Kimberly Jimenez | 10.95 | 7.7 | 7.58 | 1.02 | Latin America | 17 | 2020 | 5 | 178 | 1997 | 23 | 1 |
| Australia | Maria Thattil | 25.79 | 49 | 8.23 | 0.99 | Austoceania | 3.1 | 2020 | 6 | 160 | 1993 | 27 | 0 |
| Costa Rica | Ivonne Cerdas | 5.15 | 12.24 | 7.62 | 1 | Latin America | 13.5 | 2020 | 7 | 180 | 1992 | 28 | 0 |
| Jamaica | Miqueal-Symone Willi | 2.82 | 4.65 | 7.69 | 0.98 | Latin America | 34.7 | 2020 | 8 | 178 | 1997 | 23 | 0 |
| Puerto Rico | Estefania Soto | 3.14 | 33.4 | 8.22 | 0.9 | Latin America | 6.9 | 2020 | 9 | 180 | 1992 | 28 | 0 |
| Thailand | Amanda Obdam | 69.48 | 7.81 | 6.75 | 0.96 | Asia | 21.9 | 2020 | 10 | 170 | 1993 | 27 | 0 |
| Argentina | Alina Luz Akselrad | 45.65 | 9.2 | 5.78 | 0.98 | Latin America | 9.8 | 2020 | 11 | 178 | 1998 | 22 | 0 |
| Colombia | Laura Olascuaga | 50.35 | 5.3 | 6.71 | 0.98 | Latin America | 5.6 | 2020 | 12 | 180 | 1995 | 25 | 0 |
| Curacao | Chantal Wiertz | 0.015 | 24.5 | 0 | 0.92 | Europe | 5.69 | 2020 | 13 | 183 | 1999 | 21 | 0 |
| France | Amandine Petite | 68.08 | 42.6 | 7.4 | 0.96 | Europe | 8.5 | 2020 | 14 | 175 | 1997 | 23 | 0 |
| Great Britain | Jeanette Akua | 68.21 | 42.3 | 8.08 | 0.99 | Europe | 10.9 | 2020 | 15 | 170 | 1991 | 29 | 0 |
| Indonesia | Ayu Maulida | 276.36 | 4.45 | 7.39 | 1 | Asia | 6.23 | 2020 | 16 | 180 | 1997 | 23 | 0 |
| Myanmar | Thuzar Wint Lwin | 54.81 | 1.63 | 5.81 | 0.97 | Asia | 6.387 | 2020 | 17 | 170 | 1998 | 22 | 0 |
| Nicaragua | Ana Marcelo | 6.7 | 1.91 | 7.05 | 0.95 | Latin America | 11.5 | 2020 | 18 | 173 | 1997 | 23 | 0 |
| Philippines | Rabiya Mateo | 111.19 | 3.3 | 7.43 | 1.01 | Asia | 12.7 | 2020 | 19 | 170 | 1996 | 24 | 0 |
| United States | Asya Branch | 332.92 | 56.2 | 8.22 | 0.97 | North America | 7.8 | 2020 | 20 | 171 | 1998 | 22 | 0 |
| Vietnam | Khanh Van Nguyen | 102.79 | 2.72 | 6.2 | 1.01 | Asia | 9.2 | 2020 | 21 | 175 | 1996 | 24 | 0 |
| South Africa | Zozibini Tunzi | 58.4 | 6 | 6.97 | 0.97 | Africa | 8.7 | 2019 | 1 | 178 | 1993 | 26 | 1 |
| Puerto Rico | Madison Anderson | 3.65 | 32.85 | 8.19 | 0.9 | Latin America | 6.9 | 2019 | 2 | 175 | 1995 | 24 | 1 |
| Mexico | Sofia Aragón | 127.58 | 9.95 | 7.2 | 0.96 | North America | 17.3 | 2019 | 3 | 174 | 1994 | 25 | 1 |
| Colombia | Gabriela Tafur | 50.34 | 6.42 | 6.82 | 0.98 | Latin America | 5.6 | 2019 | 4 | 179 | 1995 | 24 | 1 |
| Thailand | Paweensuda Drouin | 66.56 | 7.82 | 6.87 | 0.96 | Asia | 21.9 | 2019 | 5 | 181 | 1993 | 26 | 1 |
| France | Maëva Coucke | 65.13 | 40.38 | 7.55 | 0.96 | Europe | 9.6 | 2019 | 6 | 176 | 1994 | 25 | 0 |
| Iceland | Birta Abiba Þórhallsdc | 0.34 | 68.88 | 7.9 | 1 | Europe | 33.8 | 2019 | 7 | 177 | 1999 | 20 | 0 |
| Indonesia | Frederika Alexis Cull | 270.63 | 4.14 | 7.26 | 1 | Asia | 6.1 | 2019 | 8 | 171 | 1999 | 20 | 0 |
| Peru | Kelin Rivera | 32.51 | 7.03 | 7.78 | 0.96 | Latin America | 9.4 | 2019 | 9 | 177 | 1993 | 26 | 0 |
| United States | Cheslie Kryst | 330.15 | 65.28 | 8.24 | 0.97 | North America | 7.8 | 2019 | 10 | 168 | 1991 | 28 | 0 |
| Albania | Cindy Marina | 2.88 | 5.36 | 7.81 | 0.98 | Europe | 27 | 2019 | 11 | 180 | 1998 | 21 | 0 |
| Brazil | Julia Horta | 211.05 | 8.9 | 6.63 | 0.97 | Latin America | 8.1 | 2019 | 12 | 172 | 1994 | 25 | 0 |
| Croatia | Mia Rkman | 4.13 | 14.94 | 7.36 | 0.9297 | Europe | 25.1 | 2019 | 13 | 169 | 1997 | 22 | 0 |

**Fig. 2** | Miss Universe partial data set after augmentation with additional personal and country profile

### Select Algorithms to Build a Forecast Model

To measure the winning factors and if they have a positive or negative influence on pageant placements, we first need to build a forecast model and then analyze the model's behavior per principles of XAI [6]. In the past few decades, machine learning has shown a high capability of capturing complex logic embedded in data [1]. The more powerful a machine learning model is, the more accurate its predictability is. For example, a neural network or deep neural network can perform complex tasks like image recognition; however, it requires a significantly large training data set to achieve high predictability accuracy. High-capability models also have limited explainability. There is a trade-off between prediction accuracy and explainability [12].

With these pros and cons of various algorithms to build a prediction model, we considered the following seven algorithms listed in [12]:

1. Linear regression
2. Logistic regression
3. Time series algorithm, such as ARIMA
4. Decision tree
5. Random forest
6. Gradient Boost Machine
7. Deep neural network

Given our small dataset and the multivariate nature of the problem, we chose this pool of algorithms because they represent a variety of complexity, interpretability, and suitability for structured tabular data, allowing us to find a balance between explainability and prediction capability. The beauty pageant finalist forecast is more complex than what linear regression algorithms and logistic algorithms can handle as it involves complicated, non-linear relationships between various personal, demographic, and socioeconomic factors and their interactions cannot be captured by these simple algorithms. It is not a time series problem as the data points in consecutive years have no temporal relationship since contestants in semi-finals and finals cannot participate again. Therefore, a time-series forecasting algorithm, such as ARIMA, cannot be applied here. The deep neural network is the most powerful algorithm for capturing logic embedded in the training data set. However, deep neural networks may cause overfitting as we only have a few hundred data points and the algorithm cannot support explainability as the relationship between features and the output is incorporated in the weights of the neural network matrix. Also, using deep neural networks to build a model is less desirable due to its low explainability.

Decision Tree, Random Forest, and Gradient Boost Machine are all tree-based algorithms with good prediction accuracy yet high explainability. We picked the Gradient Boost Machine as it is the most potent algorithm among the three and can be automatically explained by SHAP [3]. SHAP values uses a game theory approach to measure each factor's contribution to the final outcome and is a way to explain the outcome of machine learning models. Random Forest can be slow to train on large datasets, even if it improves on Decision Trees by consolidating many trees to reduce volatility and maximize accuracy. In the Gradient Boost Machine algorithm family, we selected XGBoost [11] to build the forecast model for the Miss Universe pageant placements because of its highly optimized implementation of gradient boosting that provides faster training, better scalability, and consistently higher accuracy.

XGBoost, short for eXtreme Gradient Boosting, was developed by Tianqi Chen, having been widely used among Kaggle competitors and data scientists in the industry. As a tree-based ensemble machine learning algorithm, XGBoost constructs several decision trees and ensembles their results to increase prediction accuracy overall. It builds an ensemble, a collection of many trees, rather than relying on a single tree (which could overfit or overlook patterns). Using gradient boosting, XGBoost gradually adds new trees that correct the errors of previous ones, giving the algorithm high predictive power and flexibility that are well-suited for a wide range of tasks including most regression, classification, ranking problems and user-built objective functions.

We used XGBoost, which is available in Scikit-learn [7]. Scikit-learn is a free software machine-learning library for the Python programming language. It features various classification, regression, and clustering algorithms, including support vector machines, random forests, k-means, and gradient boosting. It has been viewed as a standard tool in building machine learning models.

After building the forecast model, we can now use SHAP to analyze and visualize the logic and behavior embedded in the trained model.

### Explaining the Forecast Model with SHAP

SHapley Additive exPlanations (SHAP) [3] are frequently used for local explanations for individual predictions. SHAP is based on the theoretically optimal Shapley values from game theory. Shapley values are a widely used approach from cooperative game theory that comes with desirable properties. The feature values of a data instance act as players in a coalition. The Shapley value is the average marginal contribution of a feature value across all possible coalitions. We selected SHAP from a variety of model interpretation approaches because it offers reliable, mathematically supported explanations and allows both local and global interpretability, which makes it particularly useful for complex models like XGBoost. We used SHAP to analyze the forecast model built with XGBoost and visualize the feature's importance and whether the feature has positive or negative impacts on winning.

The first figure we plotted is the Bar plot. The Bar plot in Figure 3 visualizes the importance of each feature ranked by their SHAP values regardless of whether the impact is positive or negative to winning the pageant.

The second figure we plotted is a Bee Swarm plot, as shown in Figure 4, which shows SHAP value of each data point (i.e., observation). The figure shows SHAP values for the most important features of the model while predicting Miss Universe Finalists. Features in the summary plots (y-axis) are organized by their mean absolute SHAP values. Each point corresponds to a person in the study. The position of each point on the x-axis shows the impact that feature has on the prediction for a given individual.

Values of those features (i.e., height and age of the contestants, GDP, country of origin, etc.) are represented by their color. The red color indicates a positive impact, while the blue color indicates a negative impact. We can summarize how to interpret the SHAP summary plots in the following steps:

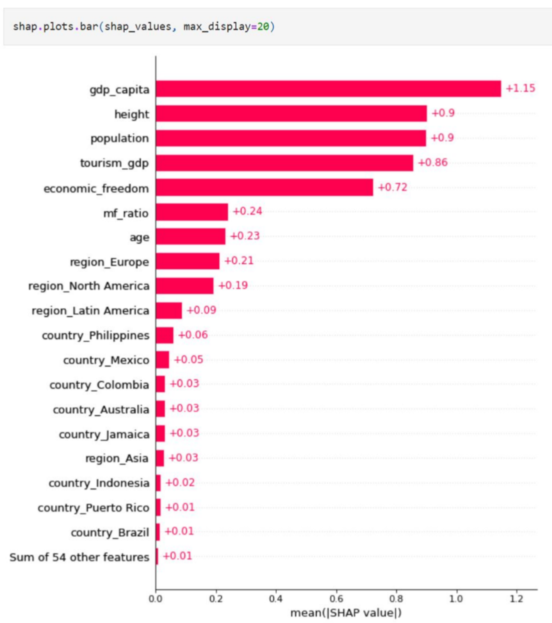


**Fig. 3** | Factor importance

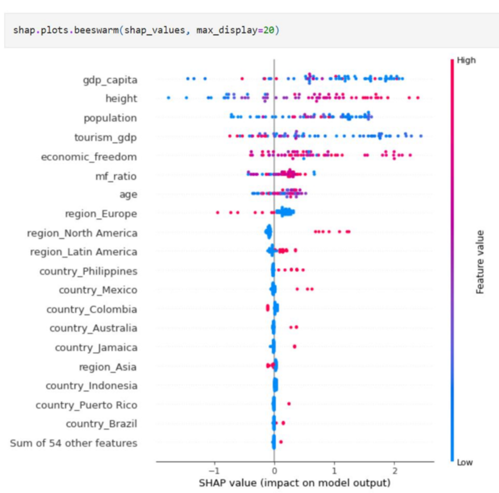


**Fig. 4** | Feature importance and impacts (positive or negative)

1. The y-axis indicates the variable name in order of importance from top to bottom. The value next to them is the mean SHAP value of the data points collected. Positive value means that the feature has a positive influence to the outcome. The magnitude of the value is relative to other features. For instance, GDP per capita, with the SHAP value of +1.15, has a stronger impact on the outcome compared to Age, with the SHAP value of +0.23, although both have a positive influence.

2. On the x-axis is the SHAP value.

3. Gradient color indicates the original value for that variable.

4. Each point represents a row from the original dataset.

## Results

We can look into the SHAP value of each feature used in the forecast model in Figure 3:

- *gdp_capita*: A lower value indicates a higher chance of winning. Therefore, Miss Universe candidates from countries with lower GDP per capita are eager to win and have a positive social impact on their countries.
- *Height*: A taller contestant has a higher chance of winning.
- *tourism_gdp*: A lower value indicates a better chance of winning. Such a trend signifies that a country with a low percentage of tourism contribution to GDP may try to utilize the Miss Universe to advertise its tourism.
- *Population*: A lower population has a larger impact, but if the population has a positive or negative impact, it is inconclusive as the SHAP values scatter in a wide range.
- *economic_freedom*: A country with a higher economic freedom has a high chance of winning.
- *mf_ratio* (male/female in population): As the average of mf_ratio is 0.97 and the standard deviation is 0.04, the data shows all countries' male/female ratios are similar, but the analysis shows the country with more women than men tend to have a slightly higher chance of winning.
- *Age*: The age requirement for contestants is between 18 to 28. The average age of finalists is 23.5, and the standard deviation is 2.2. Such analysis indicates that age has little impact on the pageant results.
- Origin of candidates
    - *region_Europe* (1 or 0): a candidate from Europe has a lower chance of winning than other regions.
    - *region_North America* (1 or 0): a candidate from North America, has a high chance of winning. This is probably because Miss Universe is held in the US every year, attracting many contestants to participate in the pageant.
    - *region_Latin America* (1 or 0): a candidate from Latin America, has a higher chance of winning. This further confirms that GDP per capita and tourism revenue as a percentage of GDP play an impactful role in winning the pageant for countries with lower values of these two factors.
    - *region_Asia* (1 or 0): a candidate from Asian countries. It has little impact on winning.
    - *country_Philippines*, *country_Mexico*, *country_Australia*, *country_Puerto Rico* (1 or 0): a candidate is from one of these countries, has a higher chance to win.

The results are interesting as they indicate that GDP per capita and tourism revenue as a percentage of GDP have the most substantial negative influence on winning, meaning wealthier countries, compared to developing countries, do not see the social impact or importance of the pageant events to their countries. Secondly, we found out that the height of the contestants is vital for winning, while the age, education, and the native language of contestants are not; most contestants have at least a bachelor's degree and speak English well. Consequently, the contestants' ages are within a small range with an average of 23.4 years old, even though the required age for the contestants is 18 to 28. These results demonstrate how both structural and individual characteristics—especially those that differ more between countries and individuals—have a more significant impact on pageant results.

## Concluding Remarks and Future Work

The Miss Universe pageant is the most viewed and impactful beauty pageant in the world, celebrating women of all cultures and backgrounds and providing positive changes through advocacy and philanthropy. In this paper, we analyze the personal and country profiles of semi-finalists and finalists of the Miss Universe pageant from 2010 to 2021 to research the critical factors for winning the beauty pageants.

Based on the list of beauty pageant semi-finalists and finalists, we augmented the contestants' basic data with a personal and country profile, including male/female ratio, tourism revenue as a percentage of GDP, GDP per capita, and the economic freedom index. Following the principle of XAI, we built a forecast model with XGBoost and used SHAP to visualize the importance of each factor. Gathering our data, we determined whether these factors have a positive or negative influence on winning the pageant. The results are interesting and provide insightful information on winning factors and the broader significance of beauty pageants, revealing how both individual attributes and national context influence outcomes and reflect deeper social and cultural values.

The technical contributions of this paper include:

- How we select proper algorithms and augment the basic data to overcome the issue of "small data", as even a beauty pageant series like Miss Universe with 70 years of history have only a few hundred data points for us to analyze. It is far less than the required training data sizes in a typical machine-learning problem.
- How we work across the whole life cycle of data science from data collection, data analysis, model selection, training, and model explanations.

We can consider a few places to improve the current results further. For example, the forecast model would be more accurate if we could collect data on candidates who did not make it into the semi-finals. We can use their personal and country profiles as negative examples of pageant-winning in contrast to the semi-finalists, which were used as positive examples for model training. Furthermore, SHAP assumes all factors are independent, however, such assumption may not be perfectly accurate. Additional work can be done to analyze the correlations between each pair of factors to reveal further insights on the complexity of beauty pageants and its greater social impact.

Finally, this work can be used as a foundation to study social reasons for why countries with a lower GDP per capita and higher tourism revenue generated more Miss Universe. To do so, we may utilize the Large Language Model (LLM) to analyze some social opinions of the public from these countries toward the beauty pageant event and the motivation of contestants for attending these events.

## References

[1] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(832). https://doi.org/10.3390/electronics8080832

[2] Lawson, R., & Ross, J. (2010). Economic freedom and beauty pageant success. Journal des Economistes et des Etudes Humaines, 16(1), 1–17.

[3] Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[4] Miss Asian Global & Miss Asian America Pageant. (n.d.). Miss Asian Global & Miss Asian America Pageant.https://www.missasianglobal.com/

[5] Miss Universe. (n.d.). Miss Universe. https://www.missuniverse.com/

[6] Murdoch, J., Singh, C., Kumbier, K., et al. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1900654116

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf

[8] The Fraser Institute. (2021). Economic freedom of the world 2021: 2021 annual report.https://www.fraserinstitute.org/studies/economic-freedom

[9] Central Intelligence Agency. (2021). The world factbook 2021. https://www.cia.gov/the-world-factbook/

[10] U.S. Bureau of Economic Analysis. (n.d.). Gross domestic product per capita. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/A939RC0Q052SBEA

[11] XGBoost. (2021). XGBoost documentation—xgboost 1.5.1 documentation. https://xgboost.readthedocs.io/en/stable/

[12] Zhang, A. (n.d.). Introduction to statistical machine learning.https://ajzhanghk.github.io/Stat3612/Lecture1_Introduction.pdf