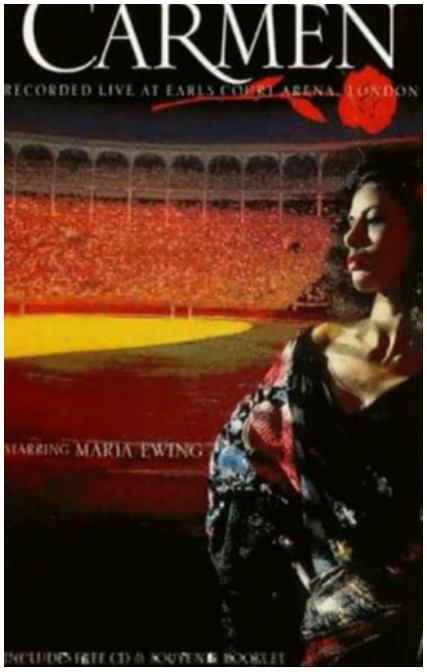
BEST LOCATIONS FOR LAUNCHING A NEW OPERA COMPANY

BY ANGELIQUE ALEXANDER

INTRODUCTION

WHY?

If you think you don't know opera, trust me—you do. The music from Carmen is some of the most recognizable in the world, and it's one of the most performed operas of all time. Maria Ewing's portrayal of the iconic Carmen is, in my opinion, one of the greatest performances ever captured. In her opening scene, she throws a rose at Don José, claiming it has a spell that will make him fall in love with her. As a singer, Ewing is passionate, playful, and stunningly talented. Her performance left an indelible mark on me, and it was in large part due to this version that I decided in high school to pursue my first degree as an opera singer. You could say I was enchanted by her rose—maybe even more so than Don José!



Despite my deep love for opera, I realized early on that I wasn't in the ideal environment to pursue a full opera career. While I cherished classical music, I was also drawn to mathematics and teaching. I spent over ten years as a high school math teacher before taking on the challenge of running a non-profit math tutoring company that serves low-income students. This unconventional path led me to data science, and now, here we are.

Throughout this journey, I've always wondered where the best opportunities for opera exist. While many told me that Europe, and especially Germany, is a great place for singers, I've always been more than just a singer—I'm also a data enthusiast. Show me the numbers! That's what makes this study so exciting: it's the perfect opportunity to merge my love of opera with my passion for data. Exploring where the best markets for a new opera company exist has been both a rewarding and challenging experience.

AUDIENCE

The audience for this study includes both professionals within the opera field and opera enthusiasts without professional experience. Professionals may include singers, voice coaches, teachers, stage managers, theater investors, orchestra players, conductors, and current opera company managers. While some familiarity with opera is helpful, it is not essential to understanding the study's findings. Additionally, the study is designed to engage and inform opera lovers, regardless of their level of experience, making it accessible to a wider audience interested in the art form's growth and future.

DATA SOURCES

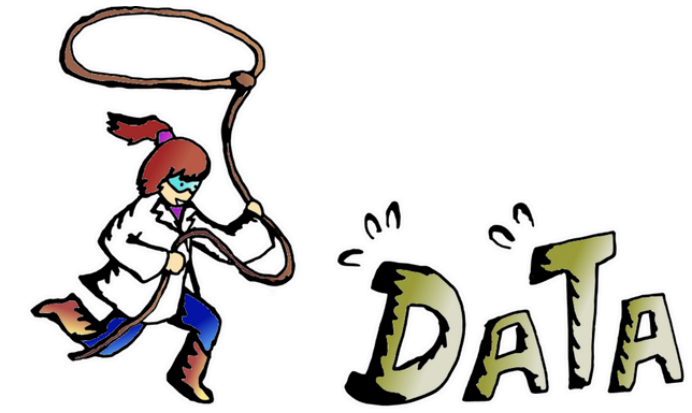
This dataset, sourced from Kaggle, contains information on opera performances worldwide. It includes data on the season, city, composer, work,

start date, and number of performances. To connect this information with population data, there were also incorporated datasets from sources such as the World Bank Group.

The metrics for success for this study are based on the following objectives for a traveling opera company:

1. **Informed Scheduling Decisions:**
- Metric: Identifying optimal performance schedules by analyzing the number of opera performances per capita across cities and countries in different seasons.
2. **Targeted Repertoire Strategy:**
- Metric: Analyzing top-performing operas in countries with the highest growth to determine repertoire selection.
3. **Regional Expansion Strategy:**
- Metric: Measuring audience growth and brand awareness in emerging markets over time, focusing on cities and countries with consistent growth.
4. **Optimized Market Entry During Off-Peak Seasons:**
- Metric: Leveraging off-peak seasons to strengthen existing markets while expanding into emerging markets during peak periods.

DATA WRANGLING



Data Wrangling, Yee Ha!

Merging the population data for cities and countries presented several challenges. The opera dataset included city names and ISO codes but lacked direct population data. To merge the population data, I used city names along with ISO codes to match cities to their corresponding countries and merge the populations listed for each city.

Variations in city name spellings sometimes prevented successful merges. Additionally, large cities like New York were broken down into smaller entities such as Brooklyn and Queens, each with population data for individual boroughs, rather than the city as a whole. Similar issues arose with country names, where slight spelling differences caused mismatches in merging population data.

To address these issues, I used the groupby method to isolate the first 10-15 entries for each city or country with missing population values. I manually verified the correct population from multiple sources and used .loc to update all of the rows in the data set for “city population” or “country population” fields with accurate values, matching them by ISO code and city or country name.

During exploratory data analysis (EDA), I discovered that some cities had multiple population entries, particularly in the U.S., where multiple cities share the same name across different states. To resolve this, I researched which city had an existing opera company and how many performances it had hosted. Once the correct city was identified, I ensured that its population was applied consistently across all relevant entries.

EXPLORATORY DATA ANALYSIS

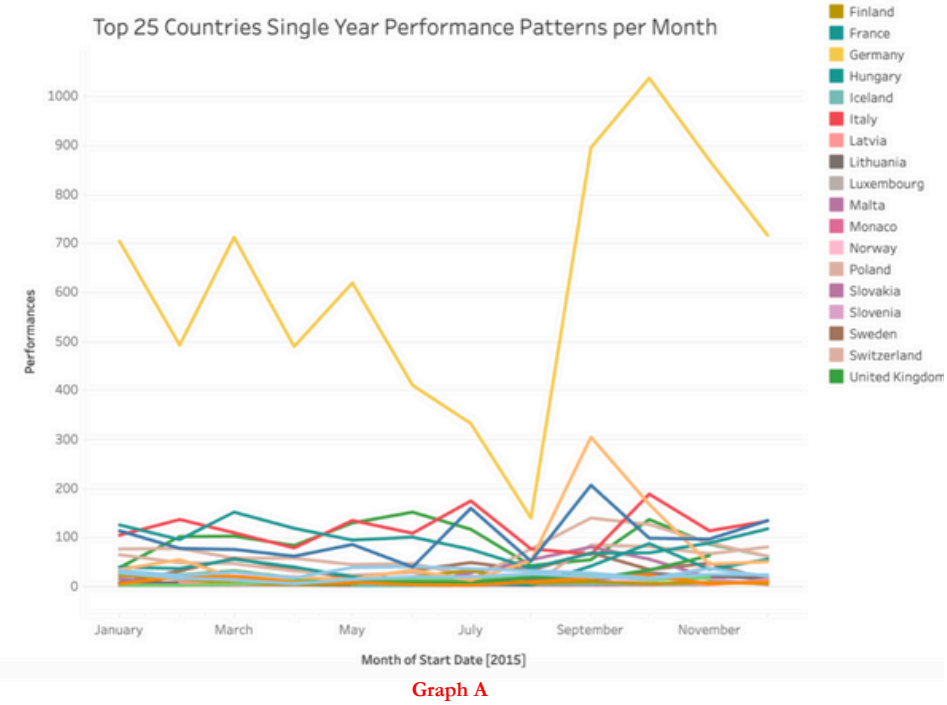
The Exploratory Data Analysis (EDA) was a crucial part of this project, as it allowed me to visually unpack the data and uncover insights. I grouped countries by continent and sub-region to create meaningful regional comparisons and derived additional metrics using the performance, season, and population data. The following columns were added to the dataset:

- Performances per country by season
- Performances per 10k population by country
- Performances per city by season
- Performances per 1k population by city
- Opera work and composer names
- Performance totals by country and city across years
- Breakdown of each season into separate years
- Change in performances from the previous season by country and city

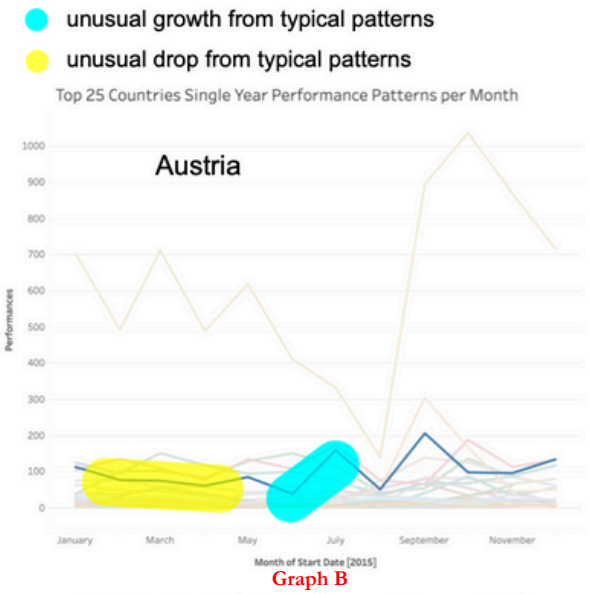
Using these new columns alongside the original data, I was able to generate unique visualizations in Tableau, revealing patterns and potential emerging markets. These insights also served as a foundation for my models, helping to assess alignment with predictive analysis.

Seasonality Graphs:

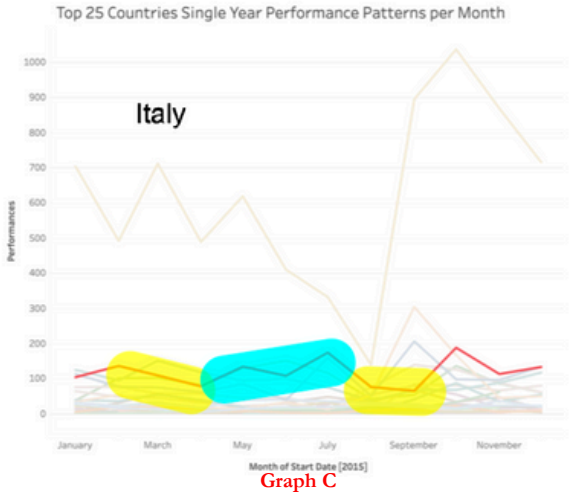
Looking a little closer at this graph, we can see something pretty interesting. Austria and Italy both have different seasonality to the majority of other countries. In Italy they have operas during the first couple of months of summer and actually start to have lower numbers of performances in August, September and the start of October. That makes sense since there are a lot of Summer Opera programs in these countries, so it wouldn't be a good idea to compete with those for audience members. In Austria it looks like the drop happens during the winter from February through the end of April. Both of these still leave an opening between the end of summer programs and when the bigger opera houses resume their regular seasons to be worth further investigation since they already have a large demand for opera in both countries.



Graph A



Graph B

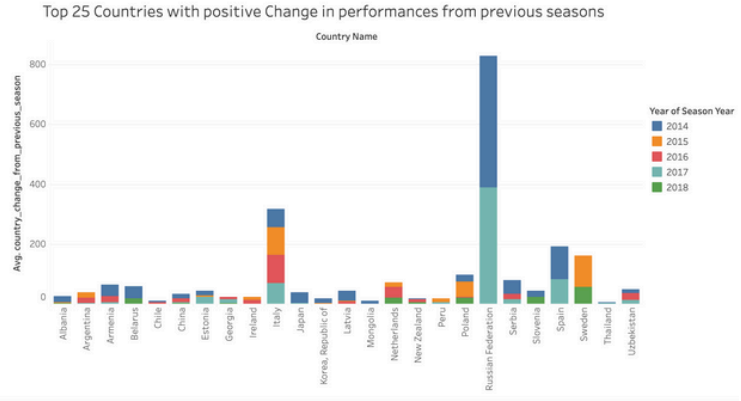


Graph C

Growth Graphs:

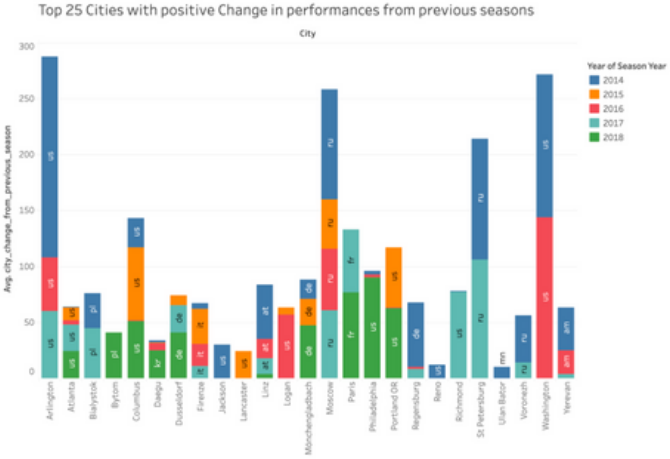
Looking comparatively at the cities with the highest average performances per 1000 citizens we can see some patterns that correlate with graphs based on positive change from previous season along by country and city.

In the graph below, each bar is segmented to show the number of performances in the country that it has grown by (meaning only positive differences are shown). There are countries with general positive trends that can be further explored such as: the Netherlands, Italy, Estonia, Poland, Serbia, Uzbekistan and possibly Sweden.



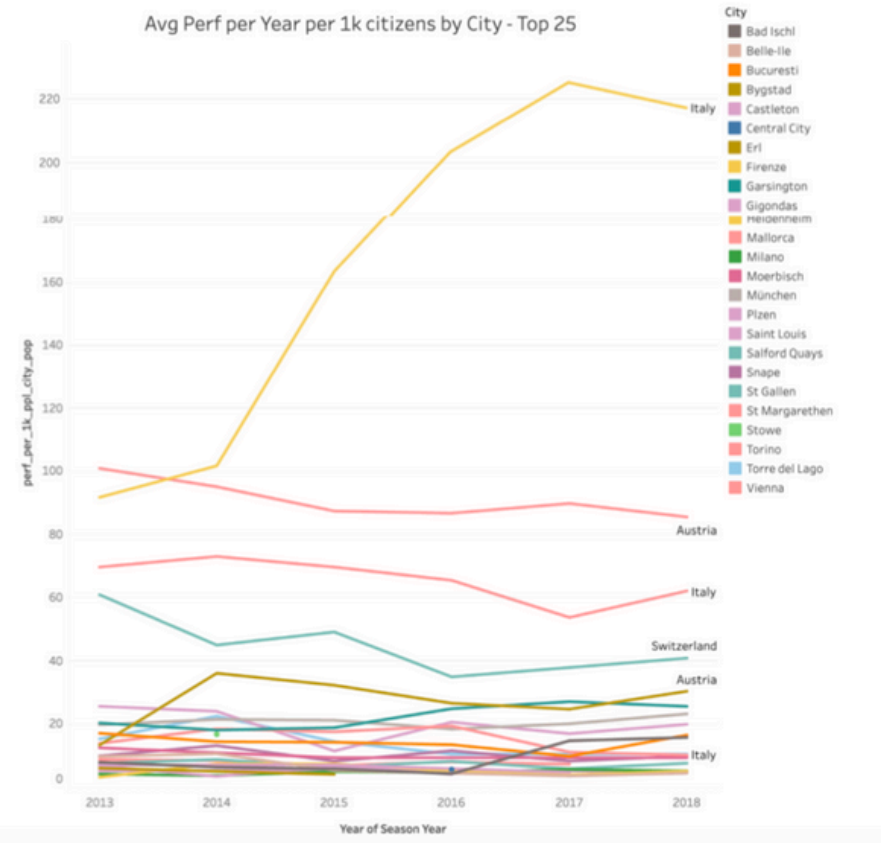
There are cities with general positive trends that match with our previous graph of countries such as: Firenze IT, Bialystok PL, Byton PL, Yerevan AM (all in identified countries with growth in performances). In addition to these cities, there has been consistent growth in Moscow Ru, Atlanta US, Columbus US, Philadelphia US, Dusseldorf DE, Mönchengladbach DE, and Linz AT.

Graph D



Graph E

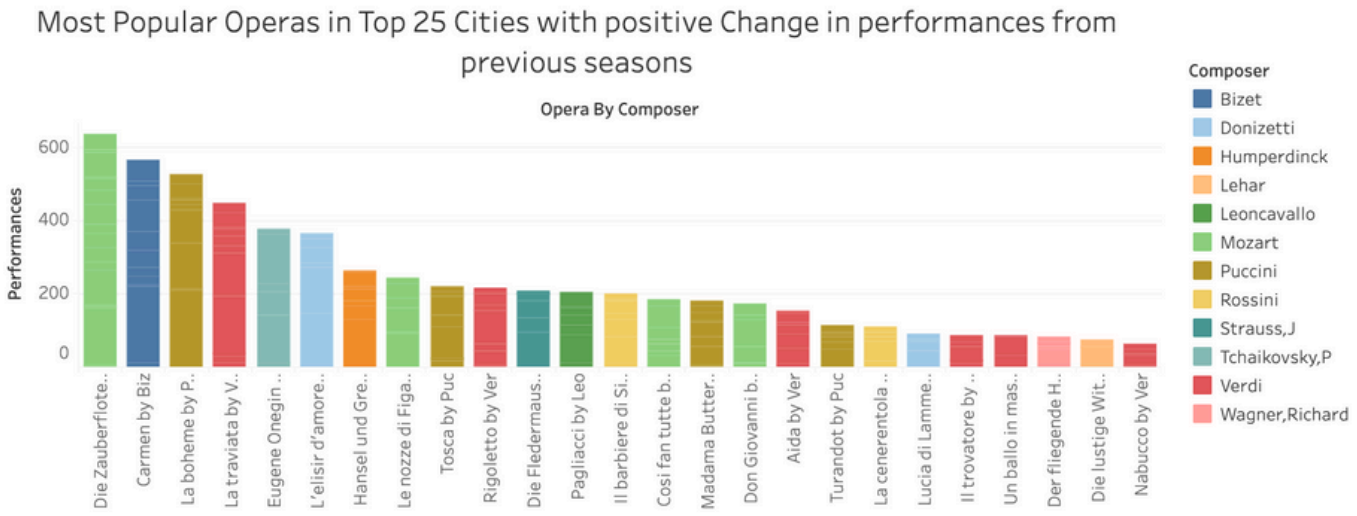
This data could help to see which cities have the highest number of performances per year per 1k citizens. Some of them are expected, like cities in countries where opera is popular, but there are some cities in countries that are more unexpected and could be a potential opening for a larger market.



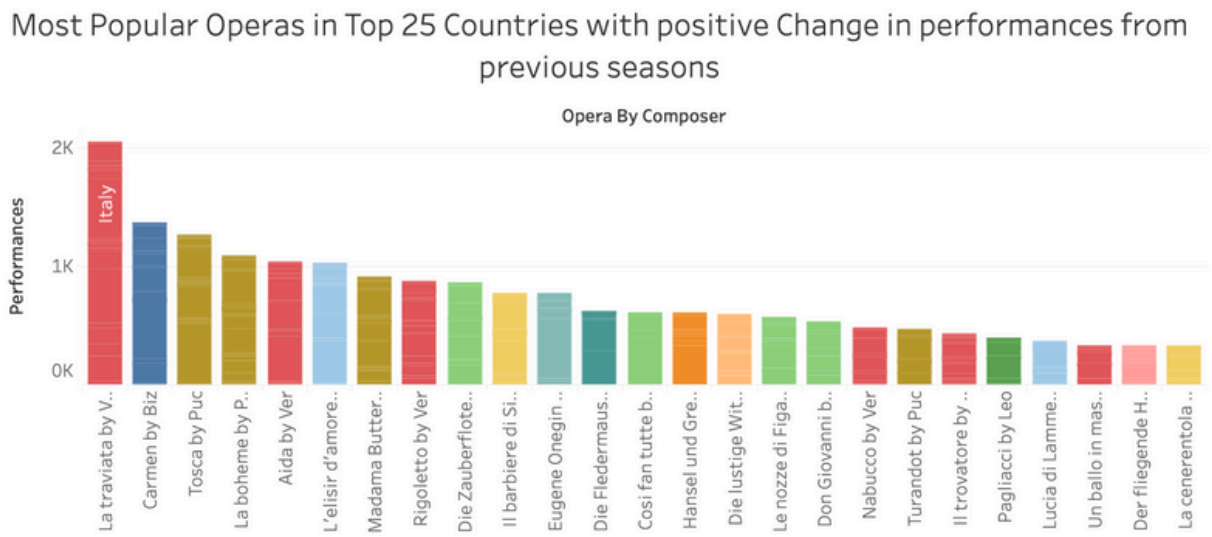
Graph F

Countries that stand out in this comparison include Italy and Austria, which exhibit unique seasonality patterns, as shown in Graph B and Graph C. Poland shows overall growth in the number of performances compared to the previous season (shown in Graph D), with several cities experiencing growth over multiple years (shown in Graph E). Other countries of interest from Graph D include Uzbekistan, Armenia, and the Netherlands. Although these countries show smaller changes in performance numbers (12-15 performances per year), they have demonstrated consistent positive growth over several years. Firenze, in particular, appears in Graph E and Graph F so it should be highlighted as a growing and promising market. I plan to use these insights as guidelines in my models to explore whether incorporating features beyond the year-to-year delta in performances can help identify countries and cities with predicted growth.

Regarding the repertoire performed in these countries and cities, we see a clear overlap in the most popular operas. While the ranking differs between Graph G and Graph H, the operas themselves remain consistent. I'm pleased to share that Carmen, the opera that sparked my own operatic journey, ranks second in popularity among both the Top 25 cities (from Graph D) and Top 25 countries (from Graph E). Given this overlap, a new opera company should focus on familiar repertoire from these categories to attract audiences in both established and emerging markets. Starting with this narrowed selection will allow the company to build its programming for the first few seasons, ensuring flexibility and variety while avoiding excessive repetition. This approach will help engage audiences and create a stronger foundation for the company's growth.



Graph G



Graph H

PRE-PROCESSING AND TRAINING

Image of Singers in Training

In this section, I removed features that didn't contribute to prediction. Here's a summary of the columns I removed and their meaning:

- 'composer' (was combined into a new column: 'opera_by_composer')
- 'db' (composer's date of birth)
- 'dd' (composer's date of death)
- 'nat' (composer's nationality)
- 'mf' (composer's gender)
- 'worknat' (opera's nationality)
- 'type' (performance type represented by symbols) — which could have been useful if deciphered.

I transformed the start date into usable features, creating datetime objects like 'Year', 'Month', 'Weekday', 'Week_of_year', and calculated 'Days_since_start' based on the first date in the dataset. This transformation made the start date more useful for predictive modeling, as the raw date itself was not in a format suitable for analysis. Some of these new columns, such as 'Month', were categorical, despite being represented by numbers (e.g., 1 = January, 2 = February, etc.). Similarly, 'Weekday' and 'Season Year' represented categories. I applied one-hot encoding to these columns— for example, creating 12 new columns for 'Month', labeled 'Month_1', 'Month_2', etc., where a 1 indicates the presence of that month, and a 0 indicates its absence. This ensures that for each date, only one month column has a value of 1, and the others have 0.

Since the data begins with the 2012-2013 season, the 'city_change_from_previous_season' and 'country_change_from_previous_season' columns had missing values for the first appearance of each city and country. To fill these gaps, I used a linear model to estimate the missing values, ensuring every row had a value for these columns.

MODELING

Performance metrics:

- Accuracy – This measures how well the model performs by keeping some known data aside for testing and comparing the predictions to the actual outcomes. It shows the overall correctness of the model.
- Precision – The percentage of true positives out of all predicted positives. It essentially measures how many of the positive predictions are correct (i.e., how few false positives there are).
- Recall – The percentage of true positives out of all actual positives. It measures how well the model captures all the relevant positive cases, i.e., how few false negatives there are.
- F1-Score – This is the harmonic mean of Precision and Recall, providing a balance between the two. It's useful when you need to balance both false positives and false negatives in the model's performance.

1. Logistic Regression: Regression models are typically effective when the data is continuous, meaning there is a consistent numerical pattern across the data. While this was not a binary classification task and the results were not continuous, it was still a classification problem. Logistic Regression performed well in terms of accuracy for identifying countries with high growth potential but struggled with cities. One possible reason for this is that there are more data points per country than per city, which gives the model stronger predictive power when classifying countries.

How Logistic Regression works

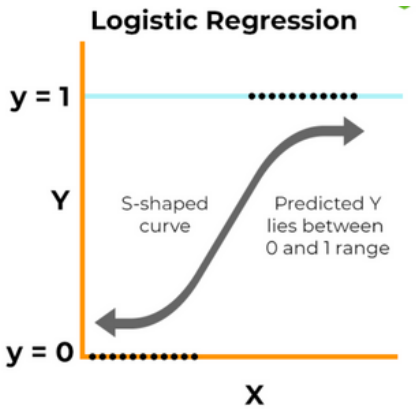


Image Source: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

The Performance Metrics

Country_growth_performance:

- Accuracy: 88.24%
- Precision: 83.10%
- Recall: 84.24%
- f1: 86.33%

City_growth_performance:

- Accuracy: 62.48%
- Precision: 60.75%
- Recall: 62.48%
- f1: 44.95%

The Predicted Countries with Growth

Country Name	predicted_country_growth
Germany	3
France	3
Greece	3
Italy	3
Ireland	3
Hungary	3
United States	3
Croatia	3
Spain	3
Estonia	3

The Predicted Cities with Growth

iso	city	predicted_city_growth
us	Houston	Very High
de	Wiesbaden	Very High
ru	St Petersburg	Very High
hu	Pécs	Very High
ca	Toronto	Very High
hu	Szeged	Very High
de	Regensburg	Very High
dk	København	Very High
ee	Tallinn	Very High
ca	Vancouver	Very High

Summary:

While logistic regression is not the strongest performing model—especially for cities—the results for countries align well with Graph D (Top Growth Countries). This gives me confidence in the model's accuracy for predicting country growth, making it a solid option for model selection.

2**. Random Forest Classifier:** Due to the way I set up the target variable for this dataset, a classifier seems more appropriate for determining the growth predictions for countries and cities. The Random Forest Classifier is an ensemble learning method that combines the predictions of multiple decision trees. Each tree in the forest makes an independent prediction, and the final model's prediction is determined by the majority vote from all the individual trees. Overall, it performed much better across all metric areas for both countries and cities compared to the regression model.

How Random Forest Classifier works

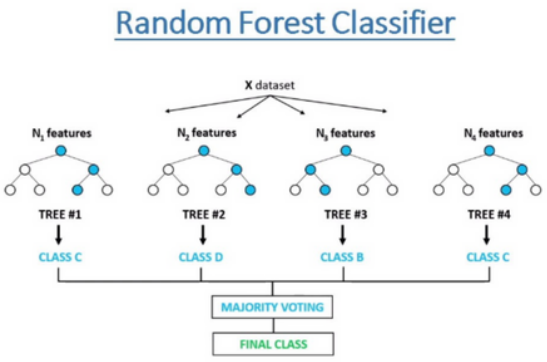


Image Source: https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/

The Performance Metrics

Country_growth_performance:

- Accuracy: 97.22%
- Precision: 97.90%
- Recall: 97.22%
- f1: 99.97%

City_growth_performance:

- Accuracy: 93.87%
- Precision: 93.75%
- Recall: 93.87%
- f1: 88.60%

The Predicted Countries with Growth

Country Name	predicted_country_growth
Germany	Very High
Russian Federation	Very High
Slovenia	Very High
Poland	Very High
Serbia	Very High
Sweden	Very High
United States	Very High
Ukraine	Very High
United Kingdom	Very High
Mexico	Very High

The Predicted Cities with Growth

iso	city	predicted_city_growth
us	Lancaster	Very High
us	Denver	Very High
us	Dayton	Very High
de	Landshut	Very High
de	Lubeck	Very High
de	Halle	Very High
us	Houston	Very High
de	Hannover	Very High
de	Heidelberg	Very High
de	Heidenheim	Very High

Summary:

Given the relatively high F1 score, I am concerned about potential overfitting in the model, particularly when predicting growth for countries. Overfitting could mean that the model is too closely tuned to the training data, which may not generalize well to new, unseen data. To assess this, I would want to perform cross-validation and test the model with new data to ensure its robustness and reliability.

Additionally, the top countries and cities identified in the results are predominantly those with the most available data. This raises concerns that the data imbalance might be skewing the model's predictions, favoring these well-represented regions. While Random Forest seems to perform better overall, it may not be identifying emerging markets as effectively as the Logistic Regression model did, especially for countries with less data.

3. XGBoost Classifier: Similar to the Random Forest Classifier, XGBoost is an appropriate model for predicting growth for both countries and cities. XGBoost also uses decision tree models but has the added advantage of handling class imbalance issues, which is why I decided to try it. In practice, it achieved an extremely high F1 score of 99.90% for country growth, which suggests significant overfitting of the data. Cities also had a high F1 score of 97.23%, which further indicates potential overfitting. It should be noted that in using XGBoost, a score of 3 correlates with "Very High", 2 = "High", 1 = "Medium", 0 = "Low".

How Random Forest Classifier works

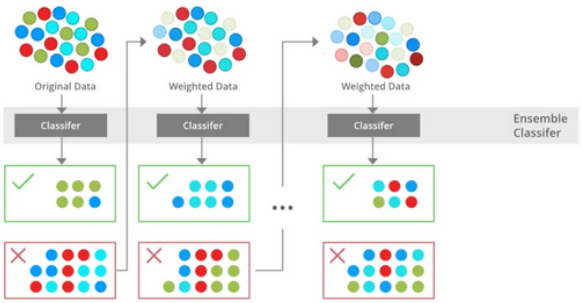


Image Source: https://www.geeksforgeeks.org/xgboost/

The Performance Metrics

Country_growth_performance:

- Accuracy: 90.89%
- Precision: 88.00%
- Recall: 90.89%
- f1: 99.90%

City_growth_performance:

- Accuracy: 92.90%
- Precision: 92.13%
- Recall: 92.90%
- f1: 97.23%

The Predicted Countries with Growth

Country Name	predicted_country_growth
Germany	3
Italy	3
Japan	3
Korea, Republic of	3
United States	3
United Kingdom	3
Poland	3
Latvia	3
Mexico	3
Hungary	3

The Predicted Cities with Growth

iso	city	predicted_city_growth
us	Lancaster	3
de	Greifswald	3
ru	Moscow	3
de	Gera	3
de	Goerlitz	3
se	Stockholm	3
de	Flensburg	3
rs	Beograd	3
ru	Krasnoyarsk	3
ru	Ekaterinburg	3

Summary:

There are similar concerns with XGBoost as there were with the Random Forest Classifier, particularly regarding overfitting, as indicated by the high F1 score. Cross-validation is needed to better assess the model's performance and ensure it generalizes well. The other metrics (accuracy, precision, recall) are more reasonable, but the fact that accuracy and recall are identical for both countries and cities suggests class imbalance may still be influencing the results.

The predicted top countries from the XGBoost model seem more balanced compared to the Random Forest model, with greater alignment with Graph D (Top Growth Countries). This overlap provides more confidence in the model's predictions for countries. However, for cities, there still isn't a model that aligns well with the findings from the exploratory data analysis (EDA). This is likely due to the limited data for each specific city, particularly outside major opera regions, which makes accurate growth predictions challenging.

CONCLUSION:

By merging the top countries identified by Logistic Regression and XGBoost, I found a solid overlap of 18 countries that align with those in Graph D (Top Growth Countries) and the cities listed in Graph E (Top Growth Cities). This overlap gives us a reliable list of high-growth markets. Again, a score of 3 indicates "Very High" for the growth and the final list is not given in growth order, but in alphabetical order.

In Top Countries (Graph D)

In both Top Countries and Top cities (Graph D and Graph E)

Has Top Cities in Graph E

Country Name	predicted_country_growth	Model
Argentina	3.0	XGB - Country
Australia	3.0	Both
Belgium	3.0	XGB - Country
Brazil	3.0	LR - Country
China	3.0	Both
Croatia	3.0	LR - Country
Czech Republic	3.0	Both
Denmark	3.0	Both
Estonia	3.0	Both
Finland	3.0	LR - Country
France	3.0	Both
Germany	3.0	Both
Greece	3.0	LR - Country
Hong Kong, China	3.0	LR - Country
Hungary	3.0	Both
Ireland	3.0	Both
Italy	3.0	Both
Japan	3.0	Both
Korea, Republic of	3.0	Both

Country Name	predicted_country_growth	Model
Latvia	3.0	Both
Lithuania	3.0	LR - Country
Mexico	3.0	Both
Netherlands	3.0	Both
New Zealand	3.0	XGB - Country
Norway	3.0	Both
Peru	3.0	XGB - Country
Poland	3.0	Both
Portugal	3.0	LR - Country
Romania	3.0	LR - Country
Russian Federation	3.0	Both
Serbia	3.0	Both
Slovakia	3.0	XGB - Country
Slovenia	3.0	XGB - Country
Spain	3.0	Both
Sweden	3.0	Both
Ukraine	3.0	XGB - Country
United Kingdom	3.0	XGB - Country
United States	3.0	Both

A strategic approach for a traveling opera company could involve using well-established opera hubs, such as Florence and other Italian cities or various Austrian cities, as anchor locations during their off-seasons. These cities have unique seasonality trends that differ from other European markets. The company could then take the same productions to emerging markets during peak seasons in their anchor cities, reaching new audiences in countries like Sweden, the Netherlands, Poland, Estonia, and Latvia—markets that are slowly expanding with lower investment risks. With the right investors and a compelling repertoire, there is significant potential for success in these markets with relatively modest investment.

As mentioned at the beginning of this report, this exploration is intended for anyone passionate about opera. It not only identifies potential markets but also offers travel recommendations for opera lovers seeking captivating performances. Ultimat ı much more to explore and develop.

La Bohème