Predicting Commercial Success of Fragrances:

A Machine Learning Approach Using Olfactory and Market Features

Open in Colab

Motivation

What drives popularity and consumer approval?

Is therefore a central question for both brands and researchers.

How machine learning can be applied to predict the success?

Analyzing both olfactory characteristics and market signals.

Project Objective

The goal of this study is to predict a fragrance's **Success Score**, a metric that balances consumer-perceived quality (**ratings**) with popularity (**number of reviews**).

To achieve this, the analysis integrates:

- Olfactory features such as **notes, accords, and complexity**
- Market-related factors such as brand **reputation, perfumer expertise, and launch year**

Dataset

The dataset comes from **Fragrantica.com**, one of the largest online fragrance databases, with over **24,000 perfumes**.

It includes information on:

- **Brand**, **country** of origin, and launch **year**
- **Perfumer(s)** responsible for the creation
- **Olfactory details**: top, middle, and base notes, as well as main accords
- **Consumer ratings** and the **number of reviews**

Success Score Definition

Commercial success cannot be captured by a single variable.
For this reason, a **composite Success Score** was created, defined as:

$$Success = 0.8 \times NormalizedRating + 0.2 \times NormalizedLog(ReviewCount)$$

This formulation gives **more weight to quality** (user ratings) while still accounting for visibility (number of reviews).

Feature Engineering

To capture both **fragrance-related** and **context-related** factors of success, several new features were created:

- **Temporal features:** perfume age, whether the fragrance is recent (≤5 years) or vintage (≥20 years).
- **Brand-level features:** portfolio size, average brand rating, popularity, and years active.
- **Perfumer-related features:** portfolio size, historical average rating, and average success score.
- **Olfactory complexity:** number of notes in top, middle, and base layers, plus overall note count.
- **Accords:** one-hot encoding of the 40 most frequent fragrance accords (e.g., woody, citrus, sweet).

Preprocessing

Several data cleaning and transformation steps were required before modeling:

- **Missing years** (~2,000) were imputed using a distribution-based approach, preserving realistic temporal trends.
- **Encoding categorical features**
  - One-Hot Encoding for **Gender** and 40 main **accords**
- **Correlation analysis**
  - Checked **correlation** with target (Success Score)
  - Identified and **dropped highly collinear** features
- **Standardization**
  - Scaled continuous variables

Result: reduced dataset from 62 → 57 features, cleaner and ready for modeling.

Modeling Approach #1

Regression (Success Score)

The prediction task was framed as a **regression problem**, aiming to estimate the continuous **Success Score**.

Several models were tested:

- **Linear Regression** as a baseline, providing interpretability but limited flexibility.
- **Random Forest**, an ensemble of decision trees, robust to heterogeneous data and able to capture non-linear interactions.
- **LightGBM (via AutoML)**, a gradient boosting method that offered the best performance.

Model Evaluation (Regression)

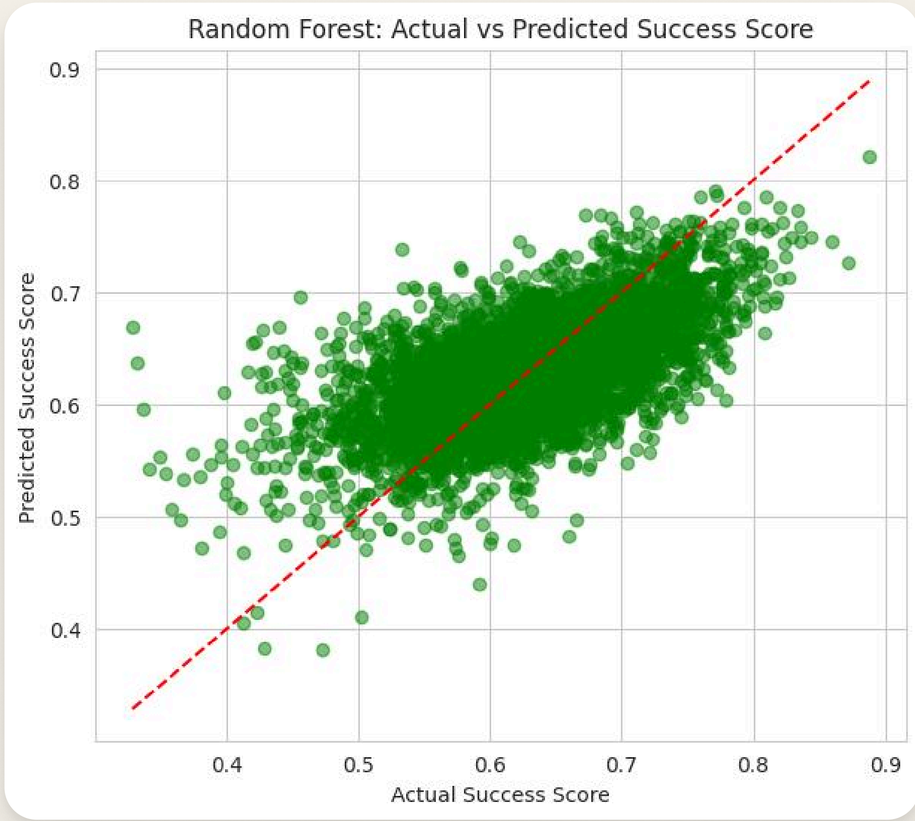
Model	R^2	RMSE	Explanation
Linear Regression	*0.34*	*0.060*	Captures some variance, but too simple for complex relationships.
Random Forest Regressor	*0.378*	*0.059*	Handles non-linear interactions; predictions closer to actual values.
LightGBM (AutoML)	*0.384*	*0.059*	Best performance overall; gradient boosting improves accuracy, though gains are modest.

Model Evaluation (Regression)

R^2 – Coefficient of Determination

Tells us **how much of the variation** in success the **model can explain**.

- In this project: **$R^2 \approx 0.35–0.38$** → the model explains about one-third of success variability.
- **Interpretation:** there is a clear signal, but many external factors (marketing, trends, culture) are not captured.

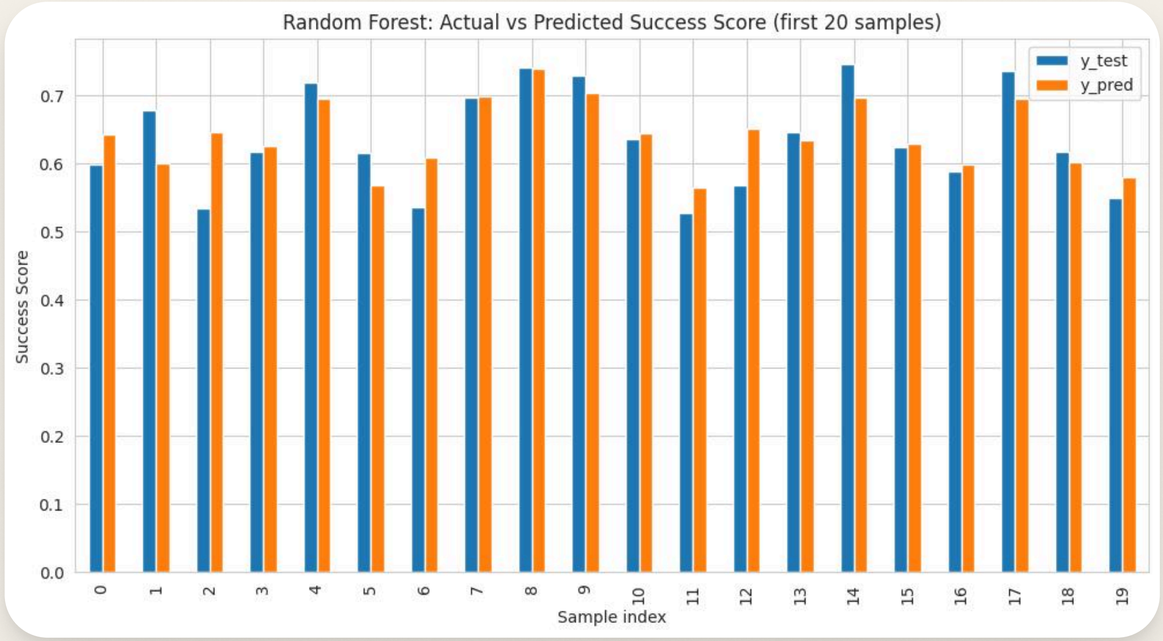


Model Evaluation (Regression)

RMSE - Root Mean Squared Error

Tells us **how far predictions are from the real values**, on average.

- In this project: **RMSE ≈ 0.06** → on average, predictions are off by 6 percentage points.
- **Interpretation:** the error is moderate but reasonable for a complex and subjective outcome like fragrance success.



Modeling Approach #2

A New Prospective With Classification (Rating Value)

The task was also framed as a **classification problem**, using the **Rating Value** as the target for interpretability.

Instead of predicting a continuous score, perfumes were split into two groups.
The threshold was the **median rating** (3.97):

- **High-rated perfumes**: Rating Value above 3.97
- **Low-rated perfumes**: Rating Value below or equal to 3.97

By comparing these groups, we could better understand which factors most strongly influence whether a fragrance is perceived as high or low quality.

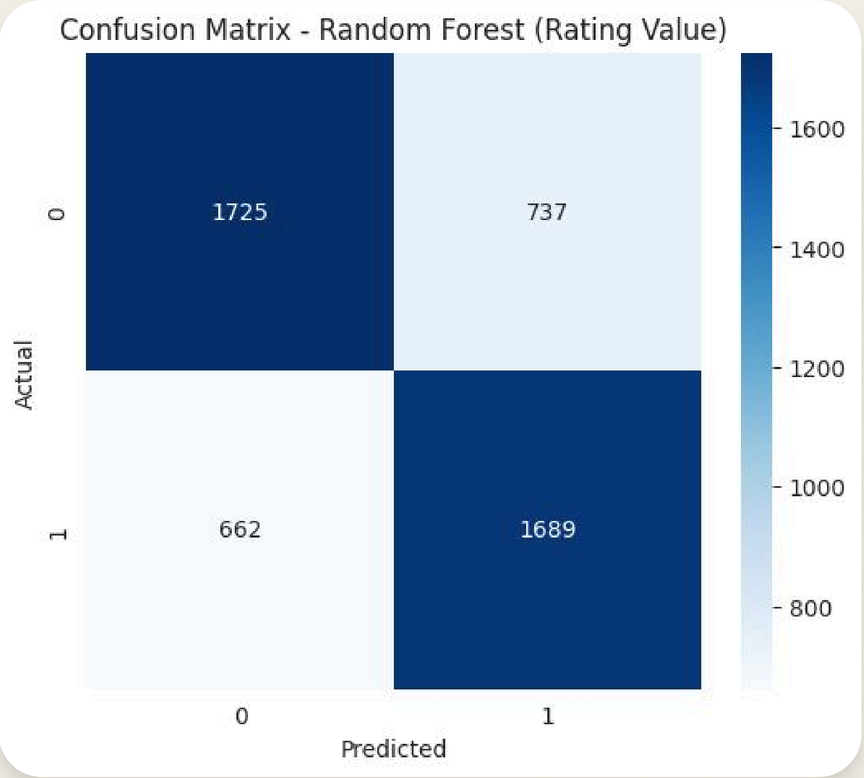
Model Evaluation (Classification)

Results:

- **Accuracy: ~71%** → the model correctly classified 7 out of 10 perfumes.

- **Precision & Recall:** balanced across both classes (≈0.70), showing no major bias.

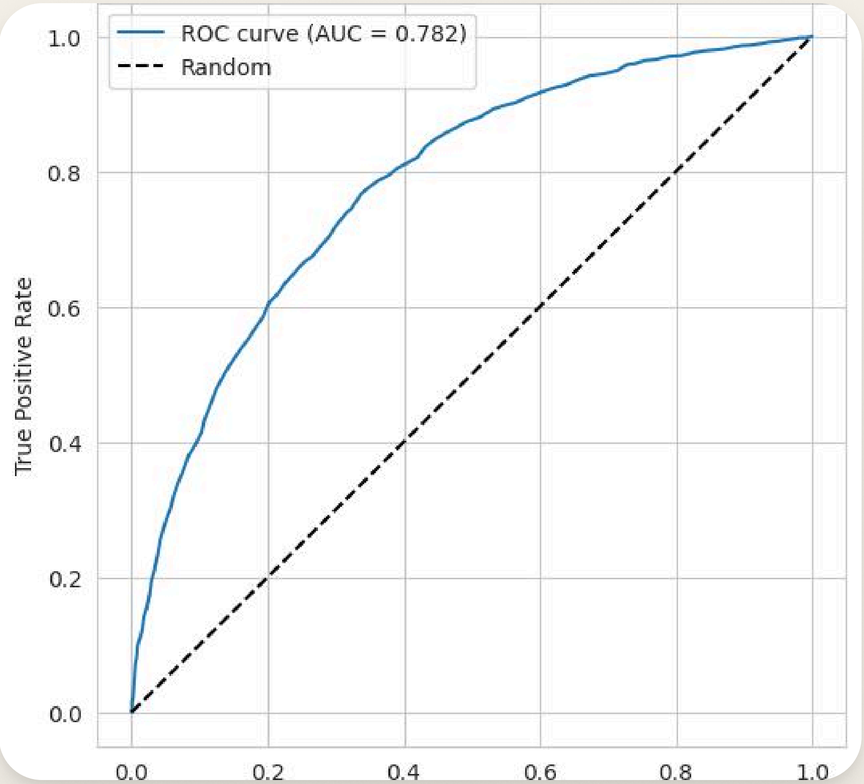
Precision: *How many predicted high-rated perfumes were actually high-rated?*
Recall: *How many of the truly high-rated perfumes did the model manage to find?*



Model Evaluation (Classification)

Results:

- **ROC-AUC: 0.78** → good ability to distinguish between high and low-rated perfumes.



The ROC curve *(Receiver Operating Characteristic) shows how well a classification model can separate two classes.*

Feature Importance

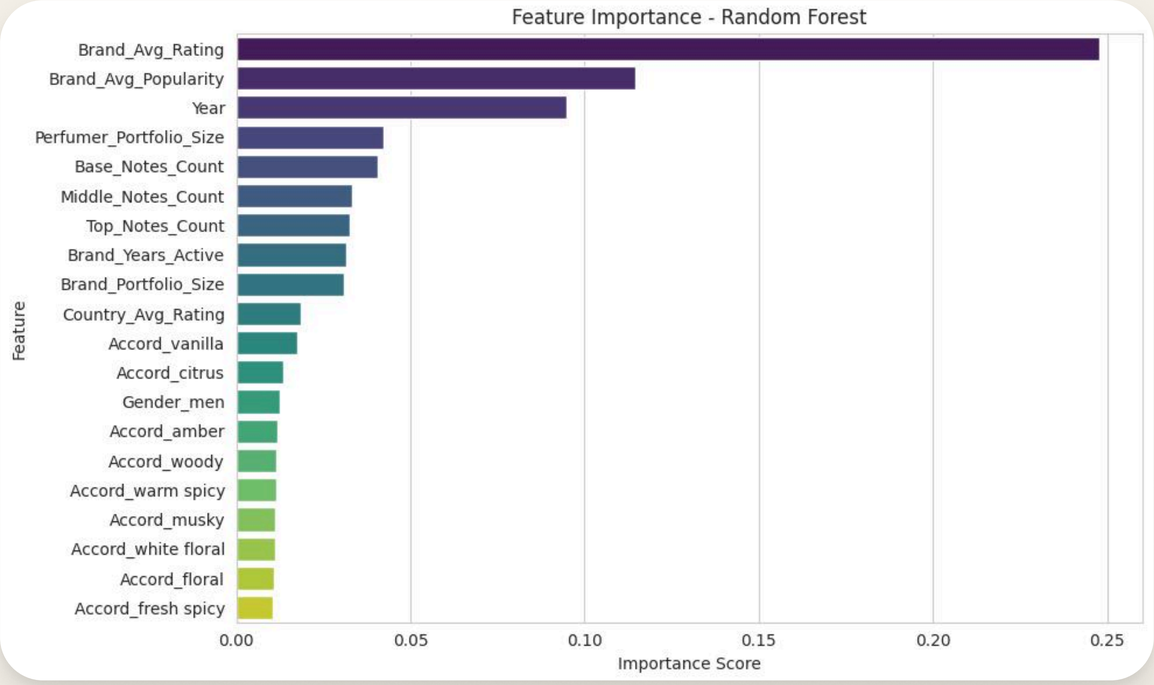
What Drives Success?

*Success is shaped more by **credibility** and **visibility** than by scent composition alone.*

Feature Importance (Regression)

Results:

- **Reputation first** → strong brands & famous perfumers matter most
- **Timing counts** → both new & vintage releases perform well
- **Visibility helps** → larger brand portfolios gain advantage
- **Olfactory families** → woody, citrus, sweet resonate with consumers
- **Gender less relevant** → quality & identity drive success



The two modeling approaches produced similar results in terms of feature importance.

Key Insights

01 **Reputation & visibility dominate** → strong brands and renowned perfumers matter more than scent composition.

02 **Success = mix of factors** → quality, timing, and brand equity combined.

03 **Limitations** → marketing, distribution, and cultural trends remain powerful external drivers not captured by the model.

Takeaway:
Fragrance success is shaped **less by notes alone** and **more by credibility, visibility, and context**.

Conclusion

1. **Machine Learning** can reveal systematic patterns behind fragrance success.
2. **Predictive accuracy is moderate** → consumer choices remain strongly influenced by external factors.
3. **Models highlight clear drivers**: brand reputation, timing, visibility, and certain scent families.
4. Practical value for **recommendation systems** and market intelligence in the fragrance industry.

Final thought:
Success in perfumery **is not random** , it emerges from the interplay of **creativity, credibility, and market forces**.

Thank You

More?

A recommendation system...

Click to Discover