Machine Learning Project:

Tourist Expenditure Prediction Tanzania

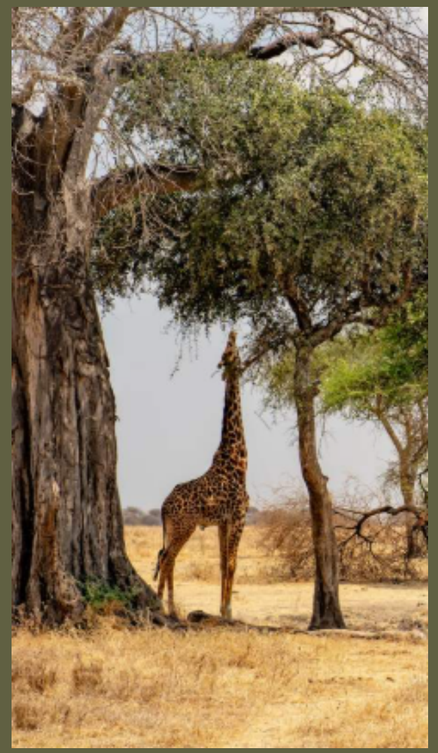
(Zindi)

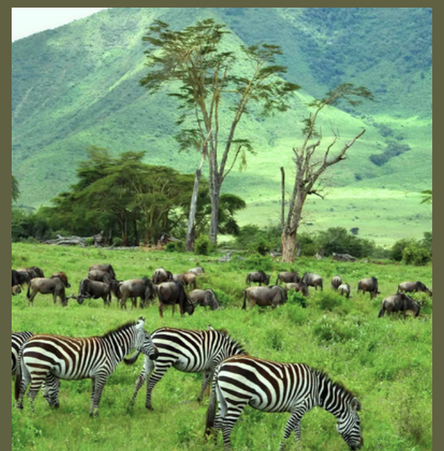
by Karl-Johann Jäkel & Mohamed Elwadiny

Source

Structure

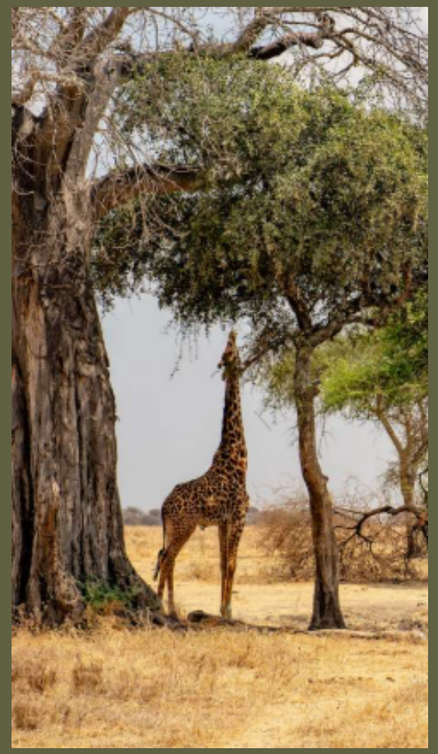
1. Introduction
2. Data Overview
3. Project Explanation (Non-Technical)
4. Product Value
5. Modeling Approach
6. Results Summary
8. Next Steps & Zindi Challenge

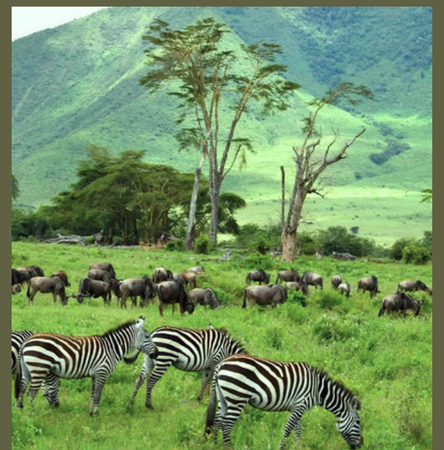




Main Goal:

- Use machine learning to help the Tanzania Tourism Board and tour operators predict what a tourist will spend before visiting Tanzania.





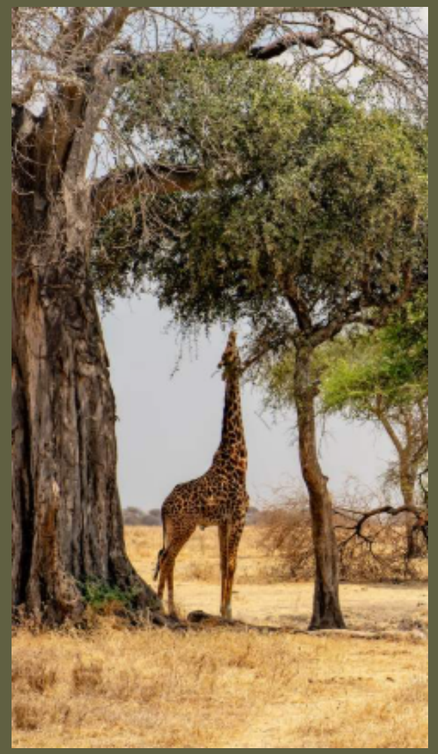
1. Introduction

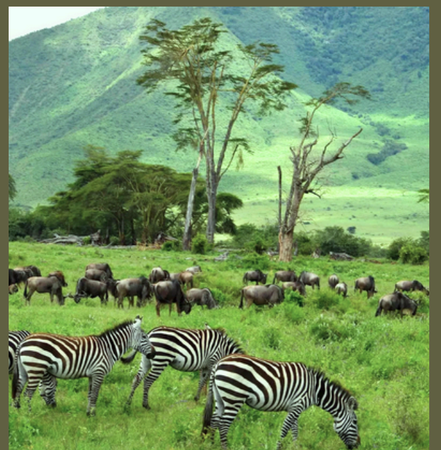
Context:

- Tanzania welcomes thousands of tourists every year
- Predicting how much they'll spend helps budget planning, pricing, and tourism strategy
- Currently, this prediction is done manually or with rough averages

Business Question:

- "Can we predict a tourist's total spending in Tanzania based on their demographics, trip purpose, and travel behavior?" → Yes!!





2. Data Overview 🔍

Size: ~4,8k records (train) + test dataset (1,6k) with 22 features

Key Features:

Demographics:

- Country, age group, travel companions

Trip details:

- Nights in Mainland/Zanzibar, purpose, main activity

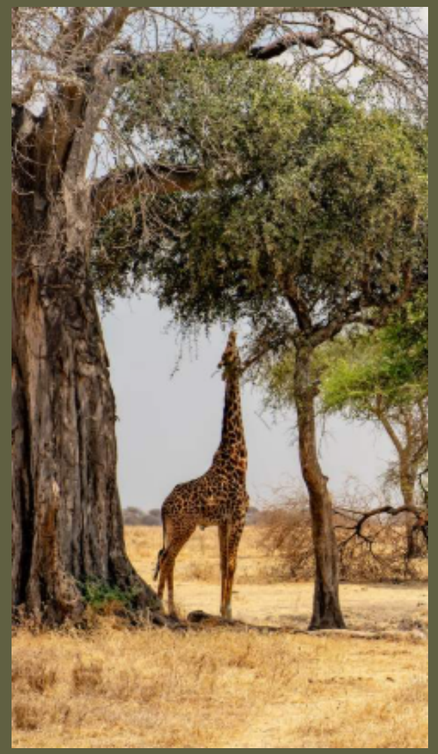
Tour package details:

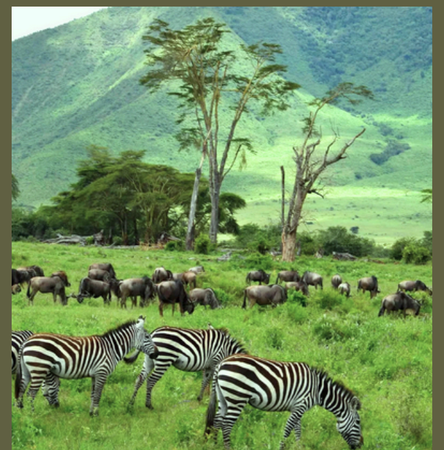
- Accommodation, transport, food, insurance

Target:

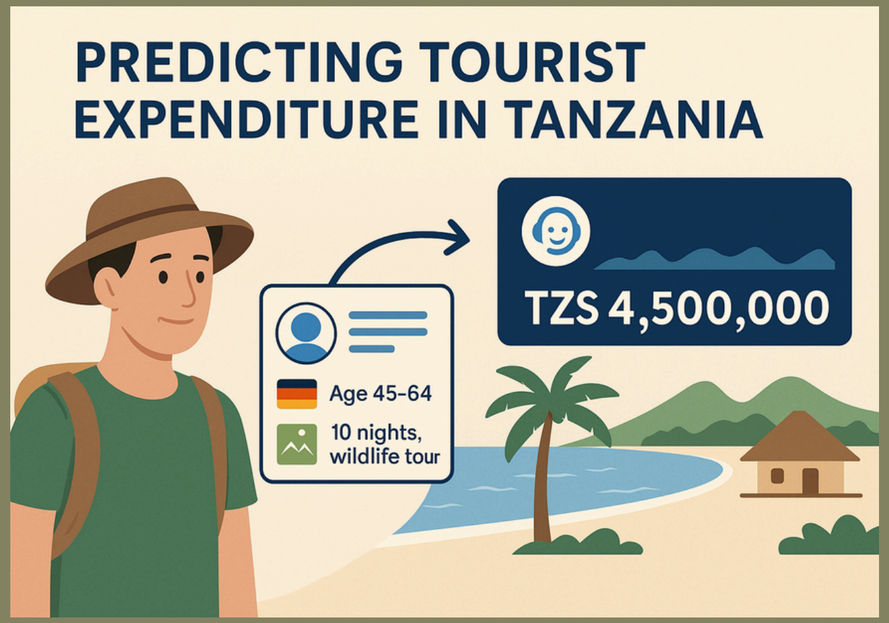
- total_cost (tourist expenditure in Tanzanian Shillings)

🧭 EDA Goal: Find relationships between trip patterns and total spending.





3. Project explanation



Imagine a travel agency or tourism board that wants to answer: "If a tourist from Germany, age 45-64, visits Tanzania for 10 nights and does a wildlife tour – how much will he likely spend?"

This Project provides a data-driven prediction tool that can:

- Estimate tourist expenditure before they arrive
- Help plan travel budgets and marketing campaigns
- Support policy decisions (e.g., which markets bring more revenue)

It's like a financial weather forecast – not 100% exact, but directionally accurate and very useful 📈.

PRODUCT VALUE

REDUCE COSTS

OPTIMIZE PLANNING

IMPROVE NEGOTIATIONS

DRIVE GROWTH

Who benefits?

- 🧭 Tanzania Tourism Board: forecast national tourism revenue
- 🚙 Tour Operators: price packages more accurately
- 🧳 Travelers: estimate trip cost in advance
- 💼 Policy Makers: identify high-value market segments

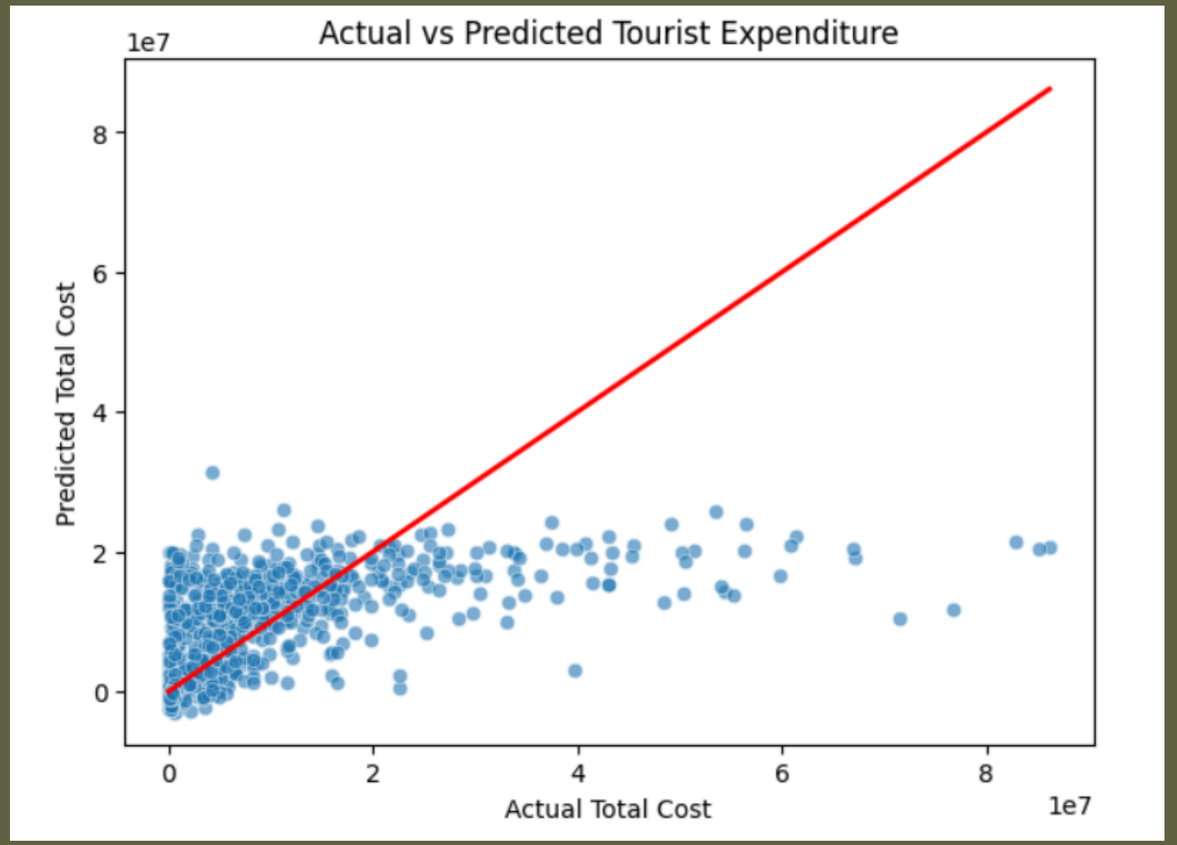
Business Value:

- Improves budget accuracy
- Supports data-driven marketing
- Enhances tourist experience via better recommendations

5. Modeling Approach

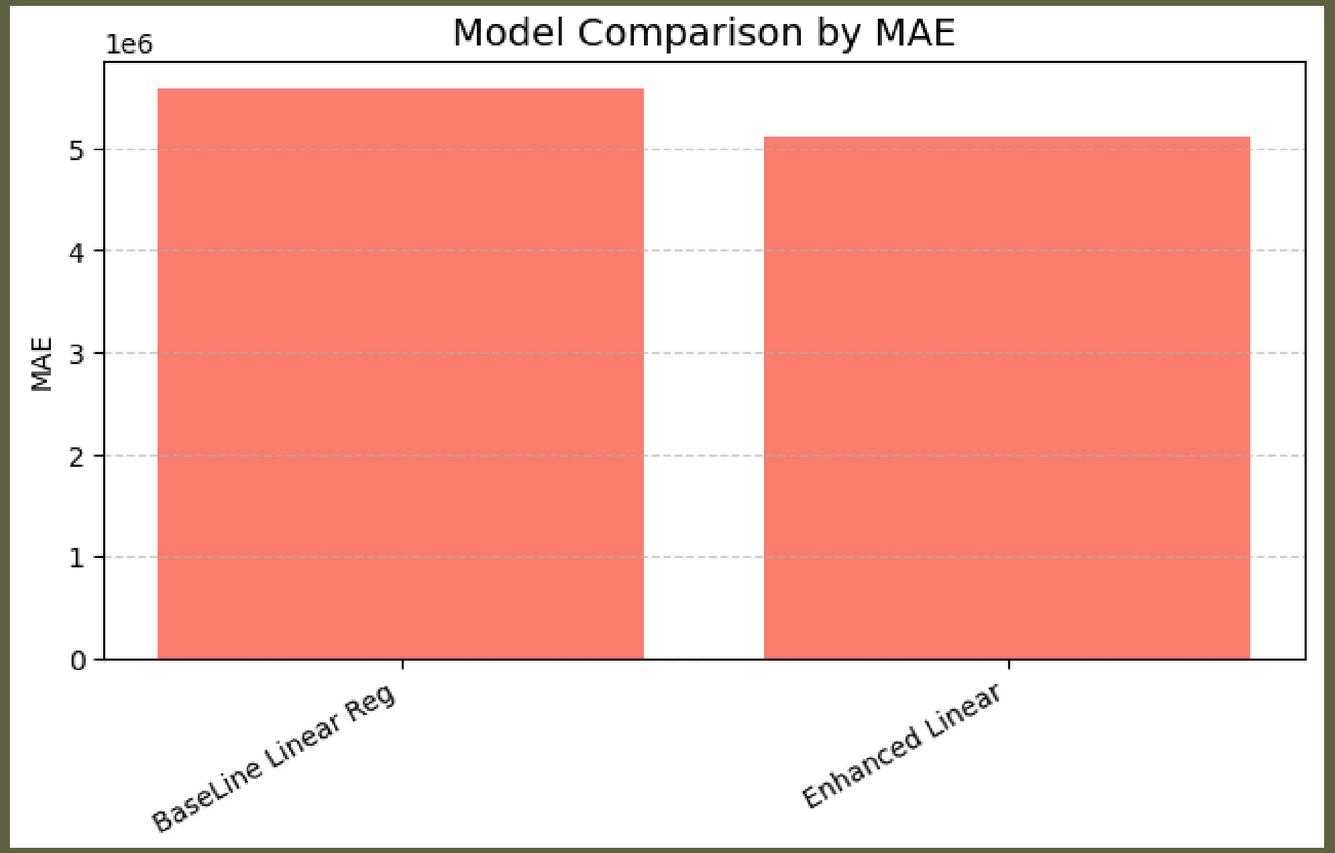
1) Baseline Model

- Model type: Linear Regression
- Features: 22 original dataset features (no modification)
- Feature Engineering: None
- Feature Elimination: None
- Transformation: None
- MAE: 5,584,619 TZS



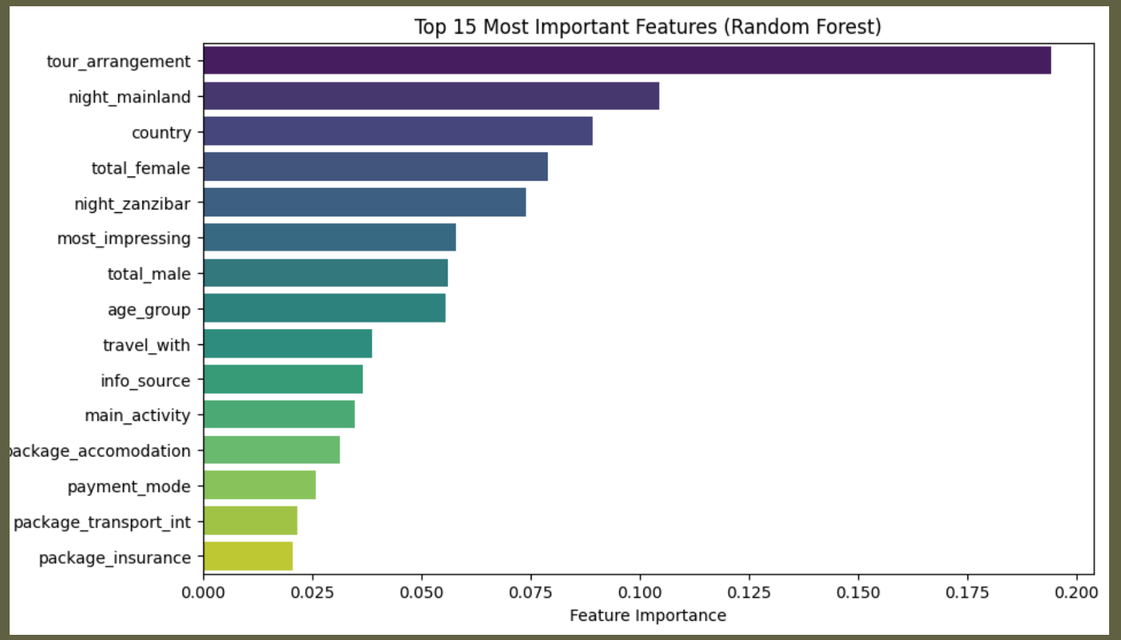
Enhanced Linear Model

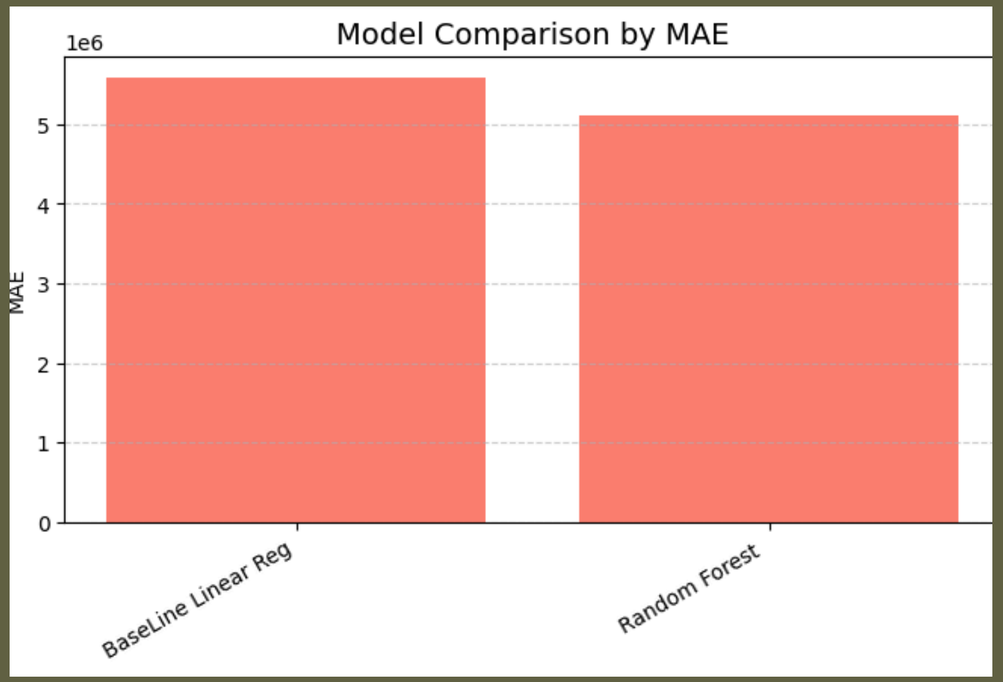
- Model type: Linear Regression
- Feature Engineering:
  - Outlier handling: Clipped numeric features at 1st and 99th percentiles
  - Standardisation: Applied StandardScaler to all numeric features
- Feature Elimination:
  - NON
- MAE: 5,105,699 TZS
- Effect on MAE: 8.6 % improvement over baseline



RandomForest Regressor

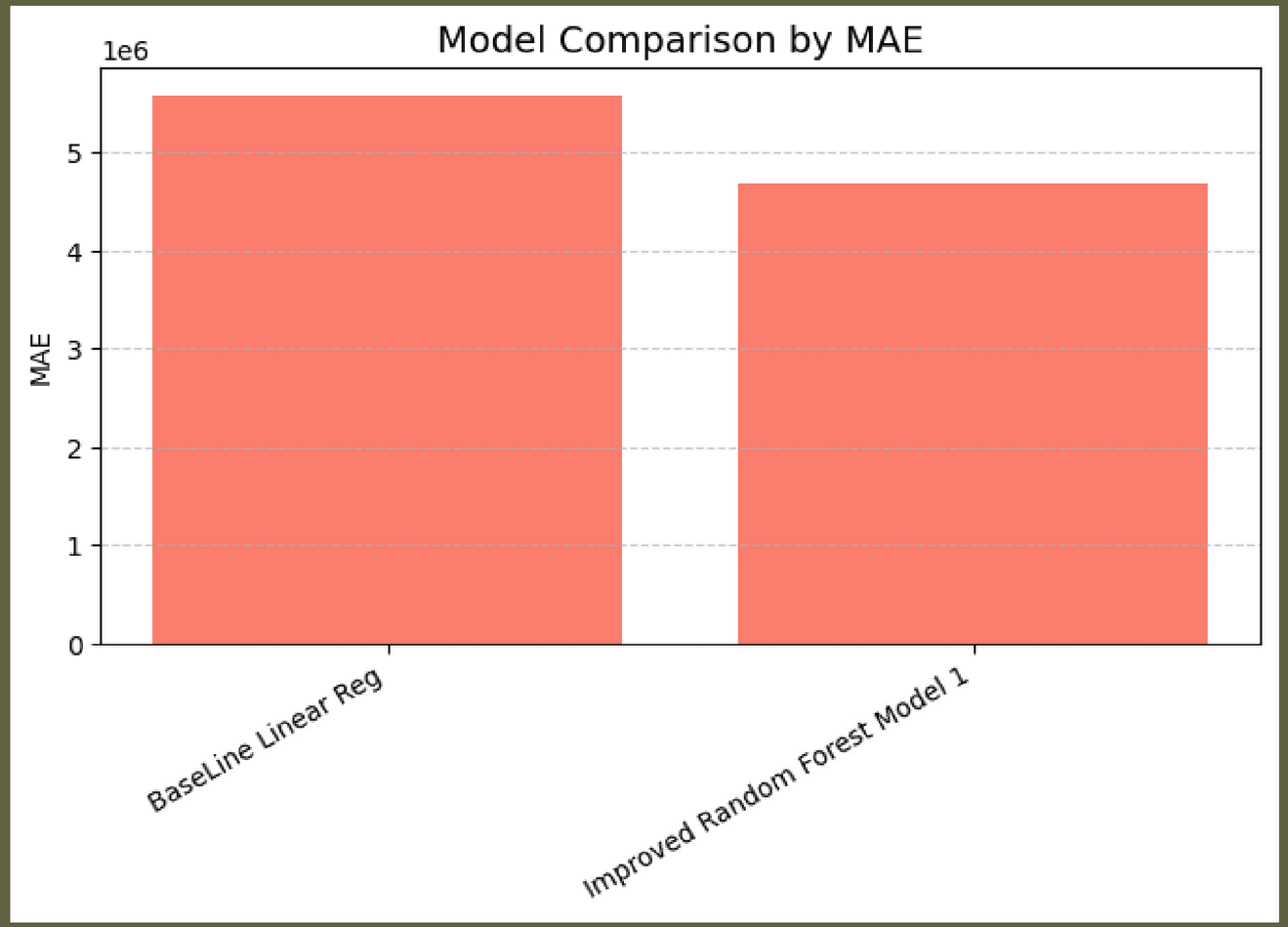
- Feature Engineering: Same preprocessed features as Enhanced Linear
- Select 15 Most important Features
- MAE: 5,120,772 TZS
- Effect on MAE: 8.3 % improvement over baseline





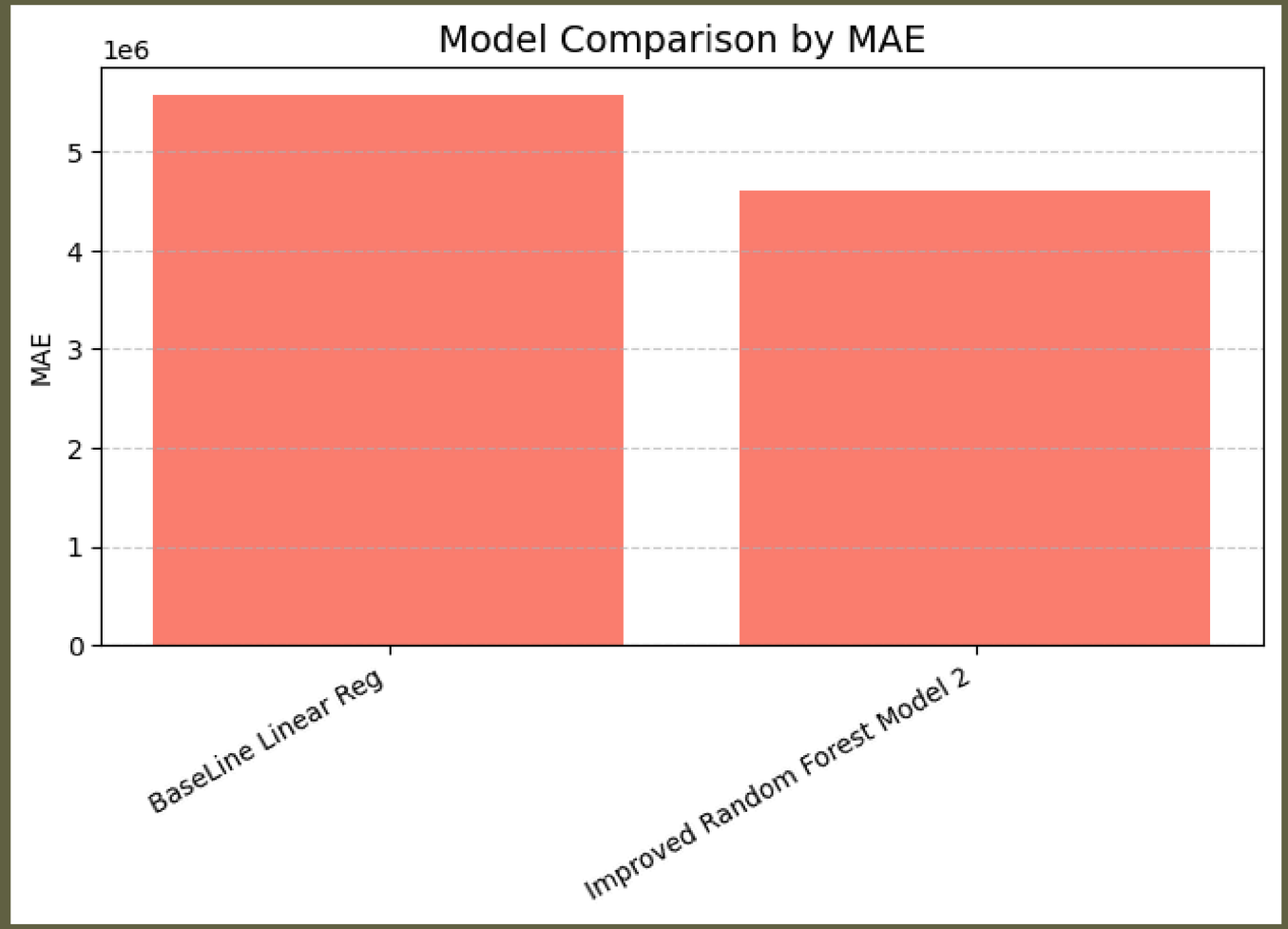
Improved RandomForest 1

- Model type: RandomForestRegressor with GridSearchCV
- Hyperparameter Tuning with GridSearchCV
- MAE: 4,690,796 TZS
- Effect on MAE: 16.0 % improvement over baseline



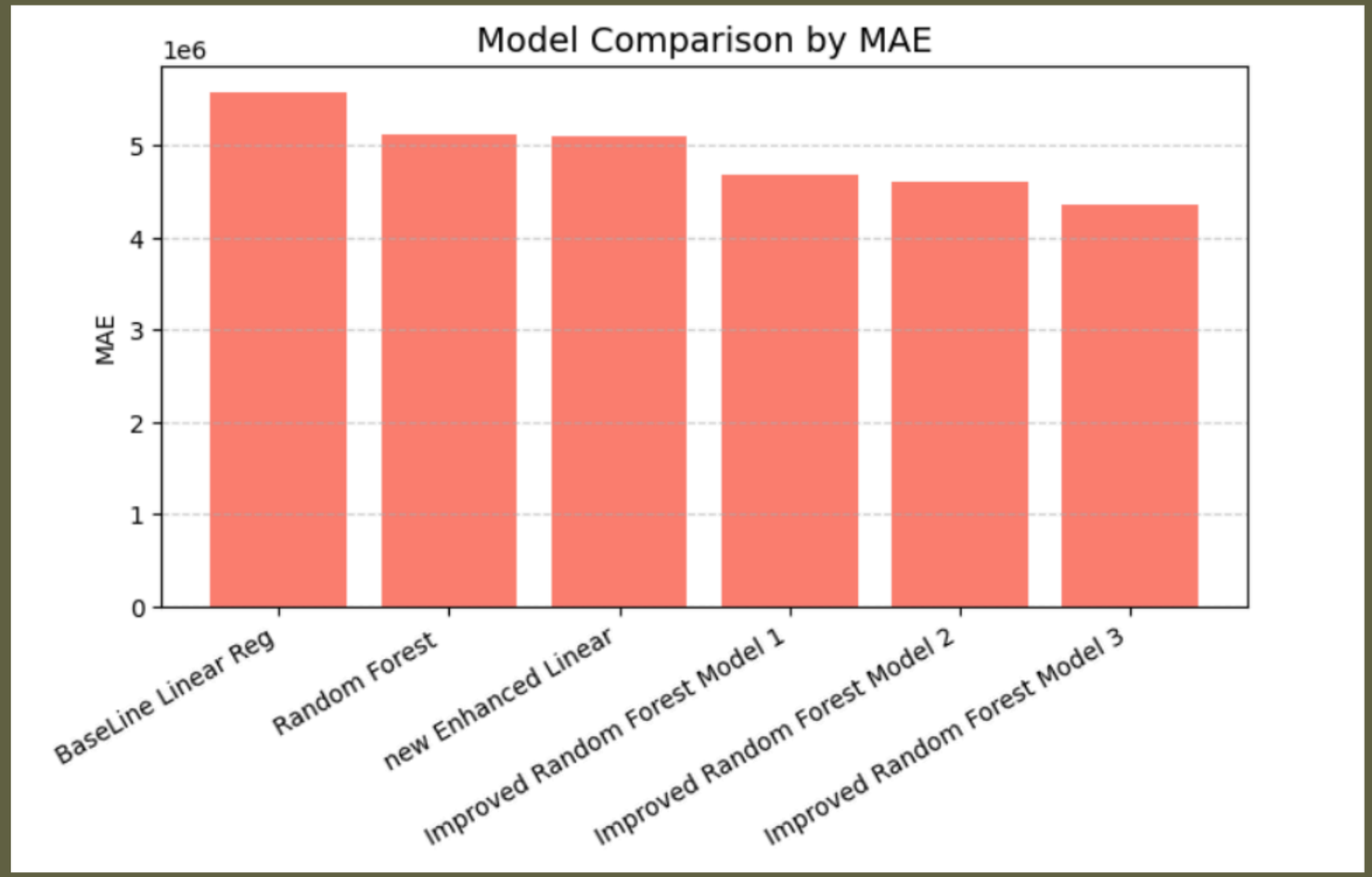
Improved RandomForest 2

- Model type: Tuned RandomForestRegressor
- Feature Engineering:
  - removed weak correlated columns those showing low feature importance.
  - Adjusted split ratio to 85/15 for stability.
- MAE: 4,612,386 TZS
- Effect on MAE: 17.4 % improvement over baseline



Improved RandomForest 3

- Model type: Tuned RandomForestRegressor with log-transformed target
- Transformation:
  - Applied log transform to target: y_log = log1p(y)
  - Inverse-transformed predictions with expm1()
- MAE: 4,384,462 TZS
- Effect on MAE: 21.5 % improvement over baseline



6. Results Summary

- Baseline Linear Regression had the highest prediction error.
- Enhanced Linear model improved accuracy by ~9%.
- Random Forest models captured complex non-linear patterns.
- Improved RF 3 achieved the best result – ≈21% better than baseline.

Model	MAE	RMSE	R^2	Enhancement
Improved Random Forest Model 2	4,612,386.00	8,422,643.30	0.44	17.41%
Improved Random Forest Model 1	4,690,795.57	8,549,329.48	0.42	16.01%
Enhanced Linear	5,105,698.56	8,178,935.54	0.38	8.58%
Random Forest	5,120,771.67	9,372,241.09	0.38	8.31%
BaseLine Linear Reg	5,584,619.38	9,633,228.47	0.34	0.00%
Improved Random Forest Model 3	4,384,461.70	9,194,255.38	0.33	21.49%

8. Next Steps & Zindi Challenge

Public Score on Zindi Leaderboard:
→ 5,155,118.159

Rank 145

Leaderboard Ranking of Zindi Challenge: https://zindi.africa/competitions/tanzania-tourism-prediction/leaderboard

Tanzania Tourism Prediction

Helping Tanzania, United Republic of

Knowledge

Active

Skills you will learn Prediction

1266 joined 317 active

Info Data Chat **Leaderboard** Team Submissions

Submit

Unless stated otherwise in the Info Page, this leaderboard reflects scores based on only a portion of the total test set until the competition closes. See competition Info for more information.

RANK	USER	PUBLIC SCORE	LAST SUBMISSION	# SUBM
145	Karl-Johann Go to placement	5155118.159	32 minutes ago	1
1	MICADEE LAHASCOM	4753976.019	over 4 years ago	4
2	FTIMS Team 9 Moduł Sumatywny Team	4763303.691	4 months ago	178
3	AIFan	4771095.882	over 1 year ago	73
4	segni Adama Science And Technology University	4780025.896	14 days ago	24
5	HaythemAbid	4780938.956	over 3 years ago	1

8. Next Steps & Zindi Challenge

Public Score on Zindi Leaderboard: → 5,021,768.158 Rank 84

Final Model - Ensemble of 3 models:

- **Random Forest**
- **Log-Transformed Random Forest**
- **Gradient Boosting**

(MAE: 4,300,647.31 TZS)

Leaderboard Ranking of Zindi Challenge: https://zindi.africa/competitions/tanzania-tourism-prediction/leaderboard

Tanzania Tourism Prediction

Helping Tanzania, United Republic of

Knowledge

Active

Skills you will learn Prediction

1266 joined 317 active

Info Data Chat **Leaderboard** Team Submissions **Submit**

Unless stated otherwise in the Info Page, this leaderboard reflects scores based on only a portion of the total test set until the competition closes. See competition Info for more information.

RANK	USER	PUBLIC SCORE	LAST SUBMISSION	# SUBM
84	Karl-Johann Go to placement	5021768.158	less than a minute ago	2
1	MICADEE LAHASCOM	4753976.019	over 4 years ago	4
2	FTIMS Team 9 Moduł Sumatywny Team	4763303.691	4 months ago	178
3	AIFan	4771095.882	over 1 year ago	73
4	segni Adama Science And Technology University	4780025.896	14 days ago	24
5	HaythemAbid	4780938.956	over 3 years ago	1

Questions?