# AI4D Africa's Anglophone Research Lab Tanzania Tourism Classification Challenge

## By le k-rismatheux

**NDINGUE NYA**
**Fredy Yann**
**Data Scientist**
**University of Yaoundé I**
**yannfredy97@gmail.com**
**yann.ndingue@fac-sciences-uy1.cm**

## Project description

The objective of this project is to develop a machine learning model capable of classifying the range of expenditures that a tourist makes in Tanzania. The model can be used by different travelers and the Tanzania Tourism Board to automatically help tourists around the world to estimate their expenses before visiting Tanzania. For more details, see: https://zindi.africa/competitions/ai4d-lab-tanzania-tourism-classification-challenge.

## PART I : Exploratory Data Analysis

The first step in any data science project is exploratory data analysis. This makes it possible to understand the data in depth and to extract useful information for the modeling/learning phase. In this project, we followed the following steps for data analysis :

### 1. Descriptive univariate analysis

Descriptive data analysis provides an understanding of the characteristics of each attribute in the dataset.

- **Dataset informations**

| **Attributes per type** | |
|---|---|
| Float64 | 02 |
| Int64 | 02 |
| String | 17 |
| **Missing values** | |
| Travel_with | 1075 |
| Totale_female | 06 |
| Total_male | 02 |

- **Attributes distribution**

See notebook eda.ipynb

By looking the distribution of the target attribute ('cost_category'), we observe class imbalance.

### 2. Correlations analysis

Here we have mainly used two (02) methods :

- • The Chi square test to test the dependence between categorical attributes and the target attribute.
- Countplot to study the dependence between the attributes

We eventually collected the following informations:

- ✓ all categorical attributes seem useful for prediction
- ✓ tourists who subscribe to a package pay more (tour_arrangement vs cost_category)
- ✓ attributes package_accomodation and package_food have a same distribution
- ✓ attributes package_transport_int and package_transport_tz have same distribution
- ✓ attributes package_sightseeing et package_guided_tour have same distribution

3. **Feature engineering**

After the observations resulting from the data analysis, the following measures were taken :

| **Attributes** | **To do** |
|---|---|
| Country | Replace each value with the corresponding continent |
| age_group | Concat ranges -18, 18-24 with 25-44, and range 45-64 with 65+ |
| travel_with | Replace missing by ‘Alone’ |
| total_male,<br>total_female | Replace missing by 1 and create attribute ‘total_persons’ |
| main_activity | Rename ‘Widlife Tourism’ by ‘Wildlife Tourism’ (input error when saving data) |
| package_accomodation<br>package_food | Combine both in ‘package_acc_food’ |
| package_transport_int<br>package_transport_tz | Combine both in ‘package_transport’ |
| package_sightseeing<br>package_guided_tour | Combine both in ‘package_sight_tour’ |

Categorical encoding : Mean Estimate Encoding / CatBoost

## PARTIE II : Modeling

For the learning phase, we opted for the XGBoost model, which has proven itself in recent years in numerous competitions on tabular data. We used a search grid with cross-validation (CV=5) to find the optimal parameters of the model :

1. Initialize the model and find the best value of the parameter *n_estimators*
2. Tune *max_depth* and *min_child_weight*
3. Tune *gamma*
4. Recalibrate *n_estimators* using previous values
5. Tune *subsample* et *col_sample_bytree*
6. Recalibrate *n_estimators*

## Conclusion and further work

This interesting project allowed us to apply the pipeline of a classic data science project from start to end. We were ultimately able to measure the impact of good exploratory data analysis on the results of the prediction model. The final score obtained (log_loss = 1.045), although below the best score of the competition (1.032) remains above the score we obtained when the competition closed (1.0497), which proves our progress. To improve, we are considering the option of combining all package_ attributes into one and testing other models or approaches.

Source code :

- EDA : https://colab.research.google.com/drive/1_1oKKLD54RvXWFaBdZc2o0kX4rsKQNDm?usp=sharing
- Modeling : https://colab.research.google.com/drive/1cNZNWqA915ZiI0A2T1yp2ZPjEjPNJu47?usp=sharing