Tanzania Tourism Prediction

How much money will a tourist will spend when visiting Tanzania?

About this project

Motivation: The Tourism sector is important for the Tanzanian economy

Goal:

- Support tourists estimating their expenditure before visiting Tanzania
- Give advice for designing travel offers and targeting advertisements

Stakeholders: Different tour operators and the Tanzania Tourism Board

Overview - What data do we have?

Observations: 4,809 records (3,847 train data, 962 test data, 80/20-split)

Outliers removed: 116 records (based on numerical features, 3% of train data)

Features: 21 in total (10 categorical, 4 numerical, 7 dependent)

Target value to be predicted: Total Cost of a trip to Tanzania (in TZS)

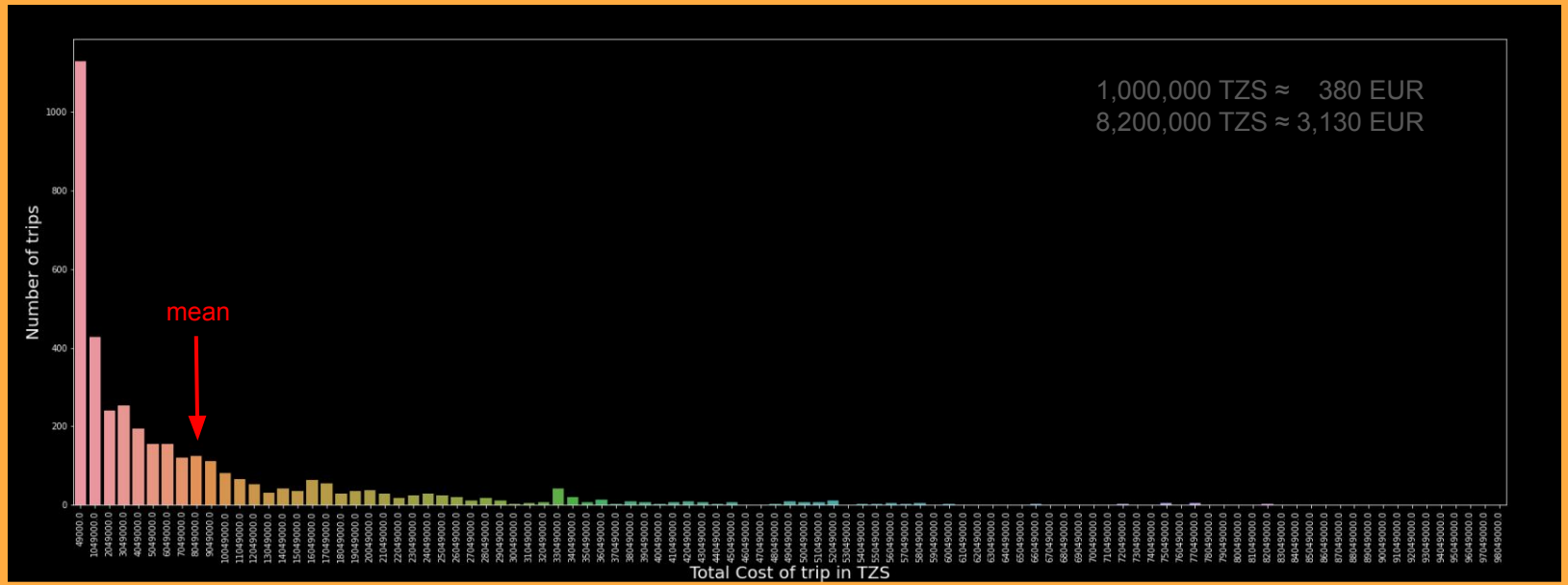
Evaluation Metric: Root Mean Square Error (RMSE)

- Average distance between the actual Total Cost and the predicted value
- Metric measures error in the target currency Tansania-Schilling(TZS)
- Big errors contribute more to the metric than small errors

Trends in the Data

What can be seen at first glimpse?

Average Total Cost of a trip ≈ 8,200,000 TZS

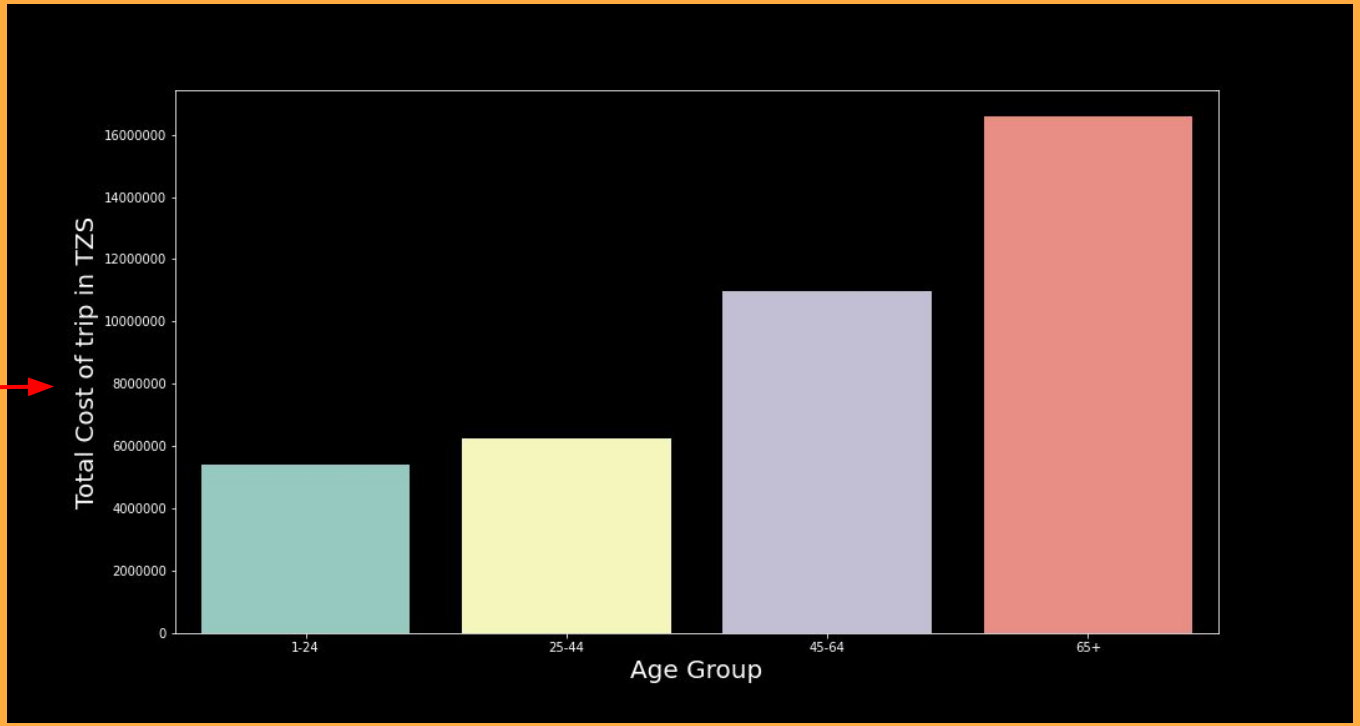


Possible Issue:
Value of Total Cost is right-skewed

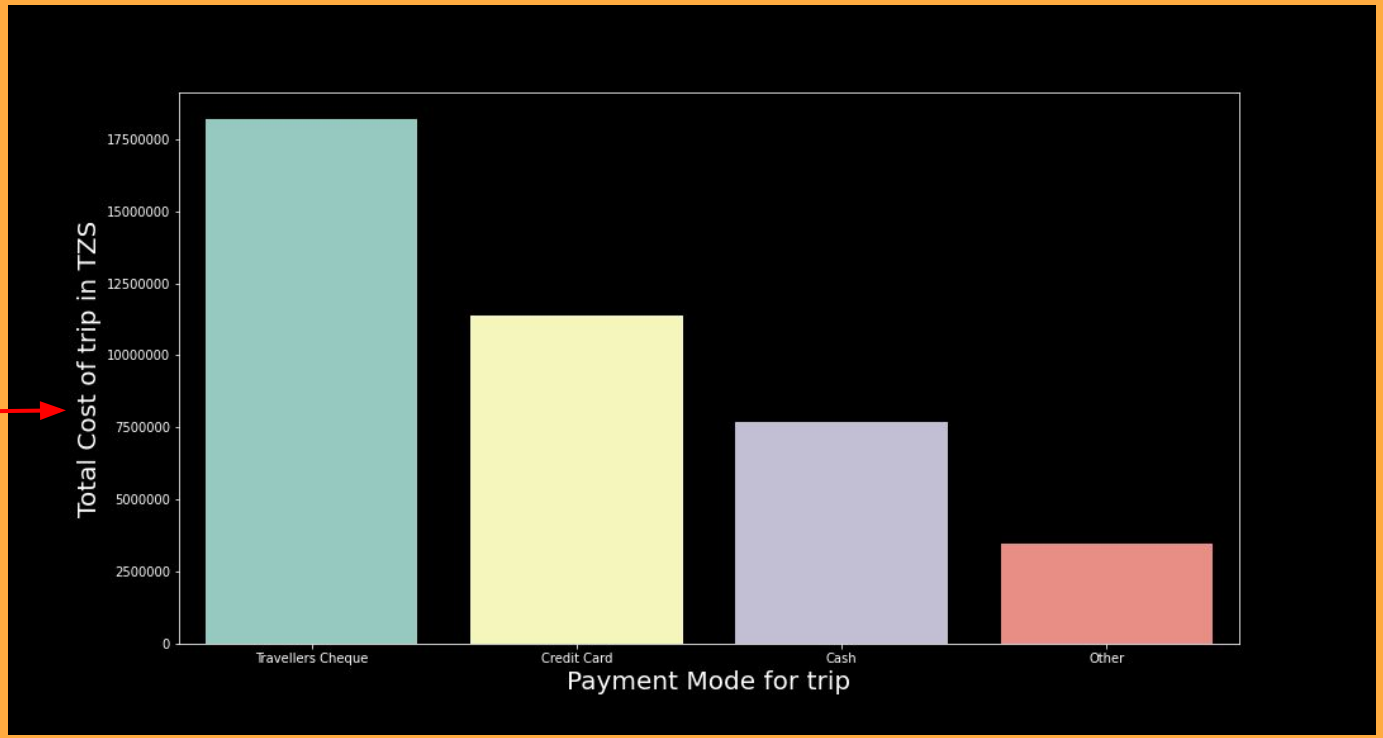
Possible Solution:
Log-transformation of the target value

Influence of Age Group on Total Cost of trip

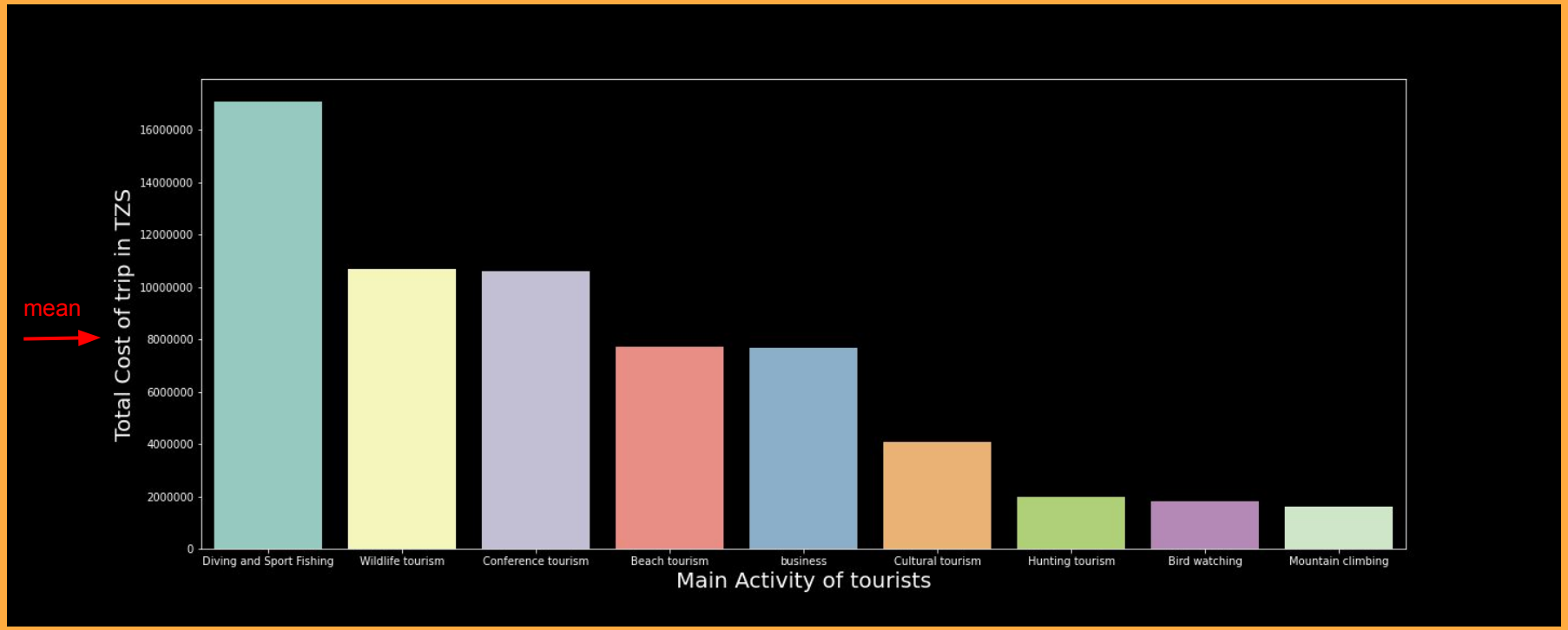
mean



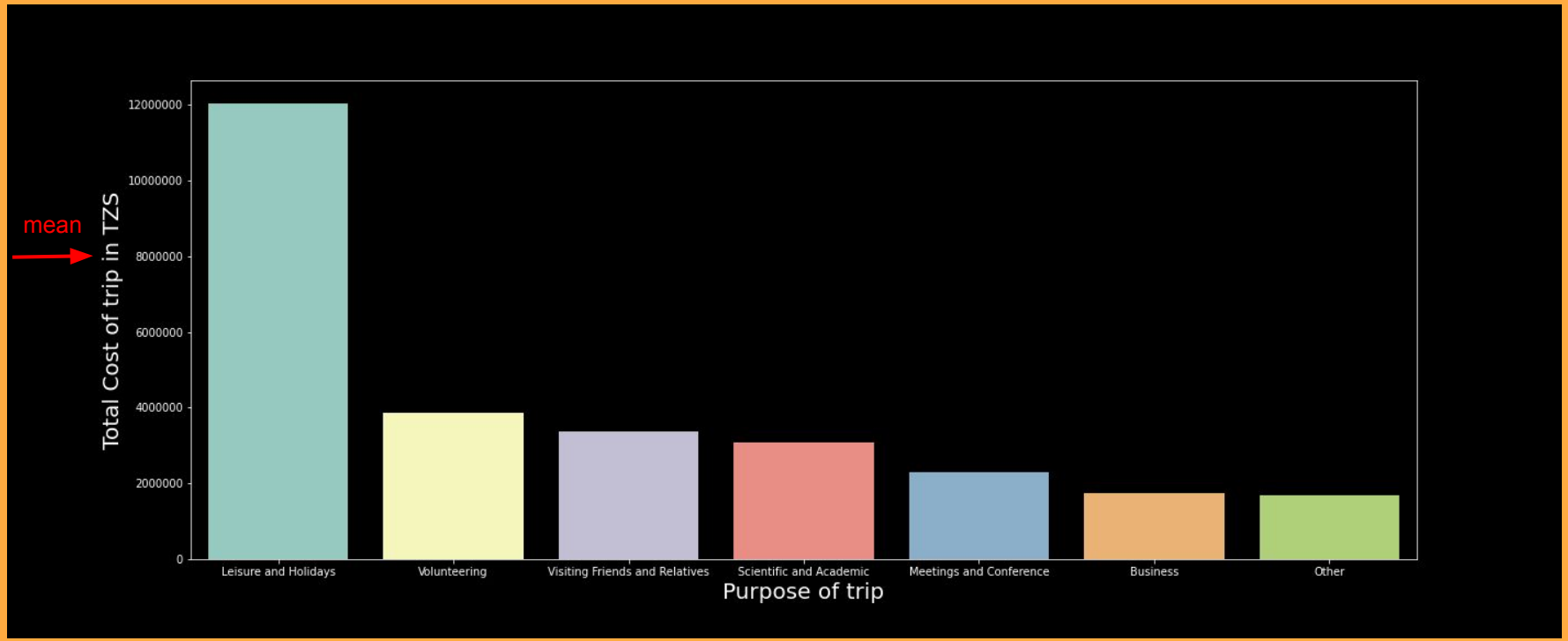
Influence of Payment mode on Total Cost of trip



Influence of Main Activity on Total Cost of trip

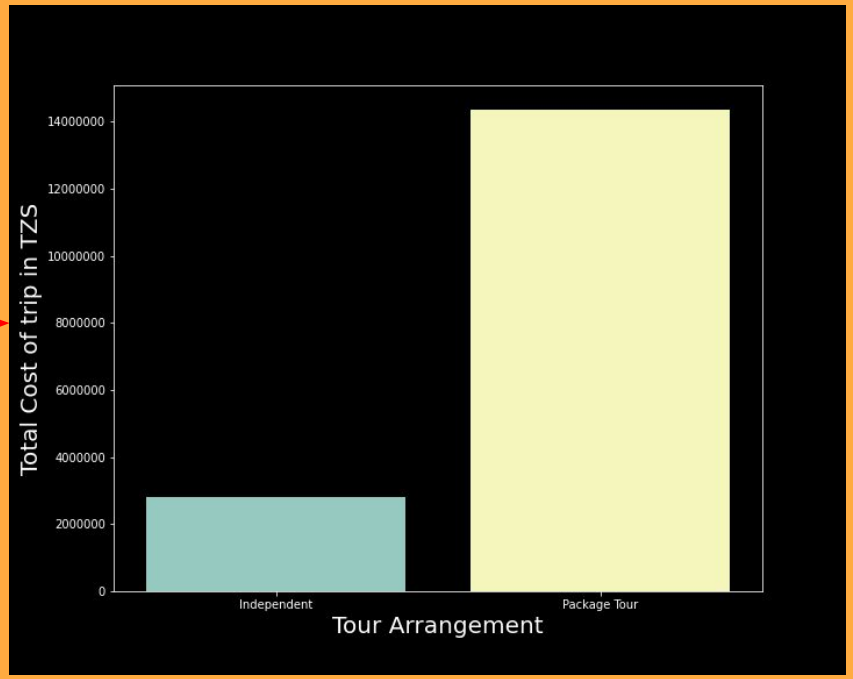


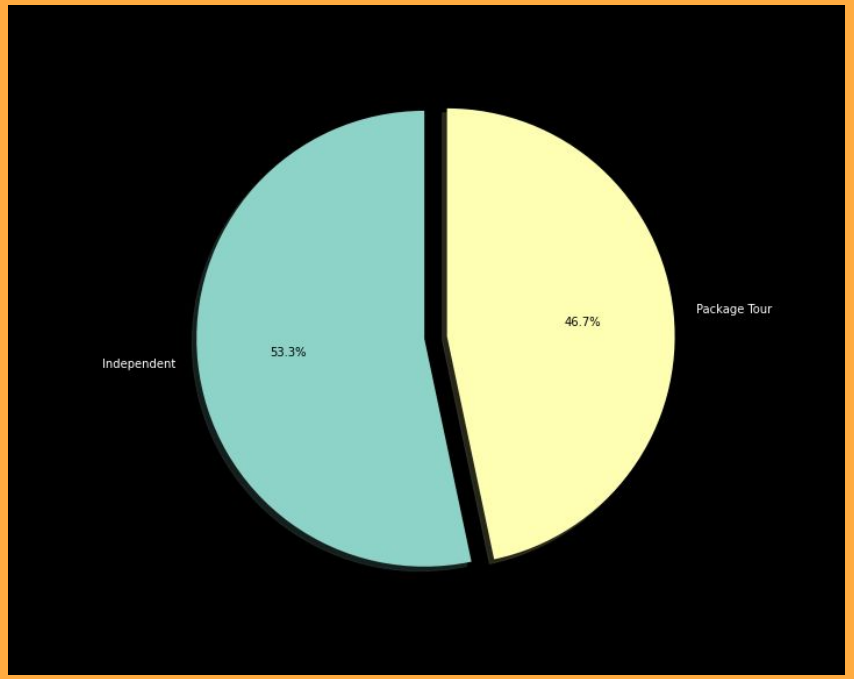
Influence of Purpose on Total Cost of trip



Influence of Tour Arrangement on Total Cost of trip

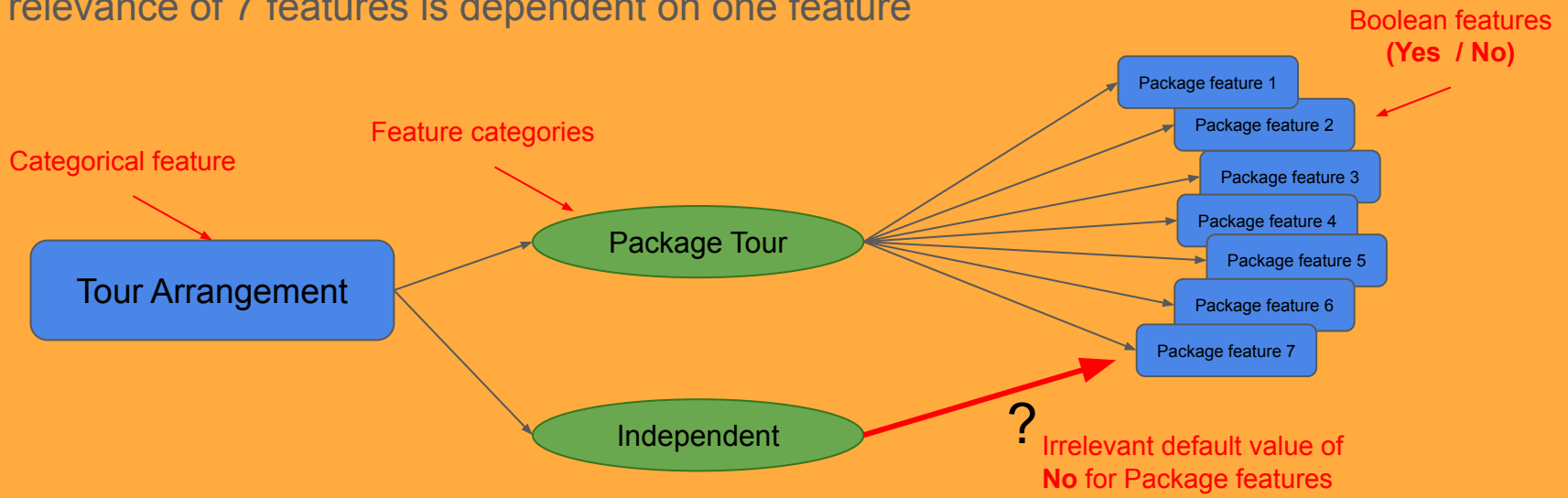
mean





Issue - Dependent features

The relevance of 7 features is dependent on one feature



Three strategies:

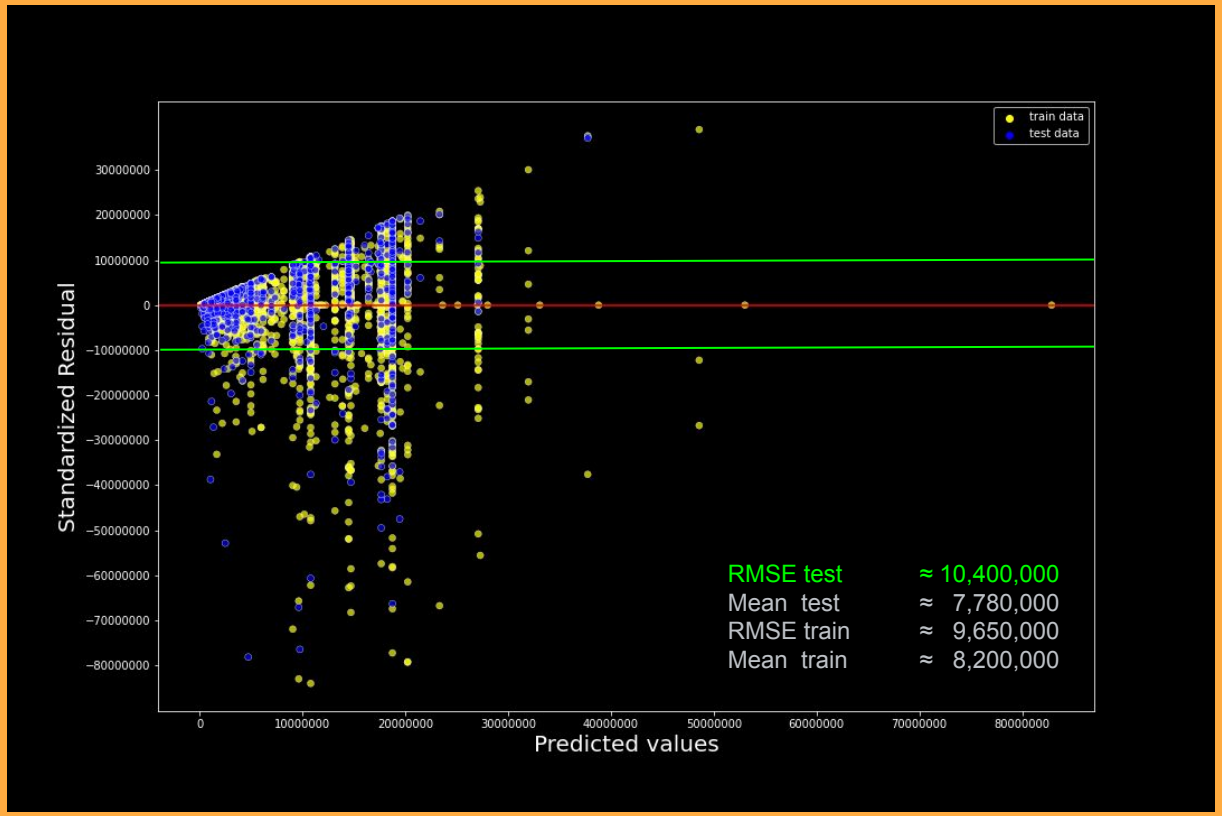
1. Leave them as they are
2. Replace with dummy value for Independent travellers
3. Remove the 7 Package features from the dataset

Baseline Model

What's an educated guess about the Total Cost?

Baseline Model strategy

Error - Residual Plot of the Baseline Model

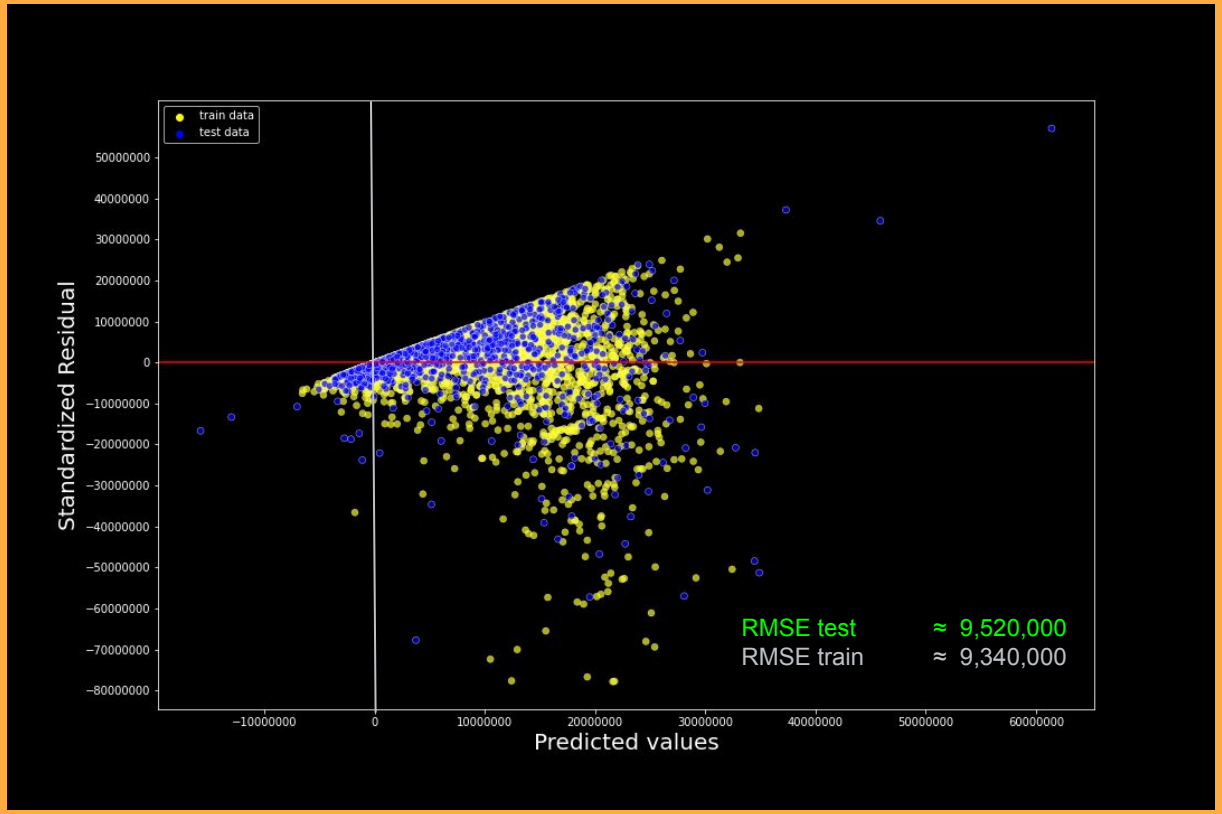


Target value original
Package features original

Machine Learning Model

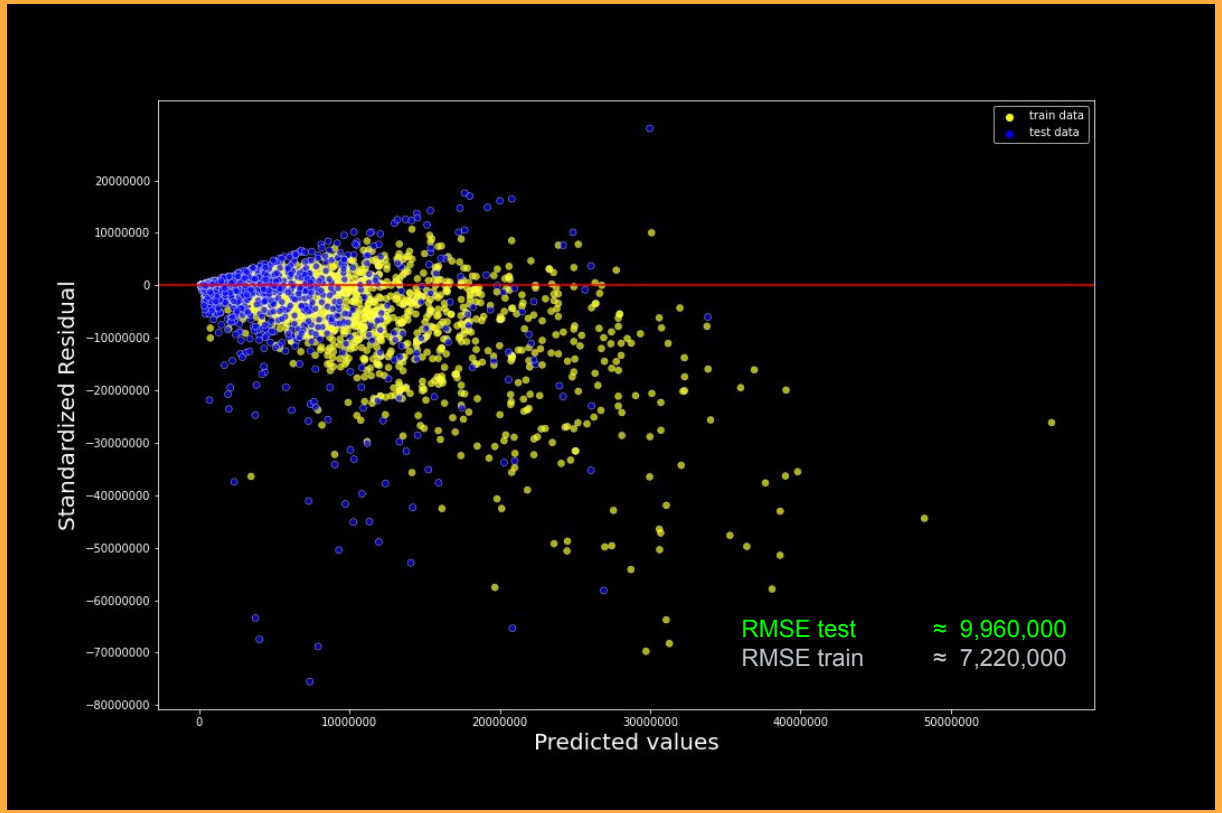
Can automation improve the prediction?

Simple Linear Regression



Target value original
Package features original

Optimized AdaBoost Regressor with Decision Tree



Target log-transformed
Without Package features

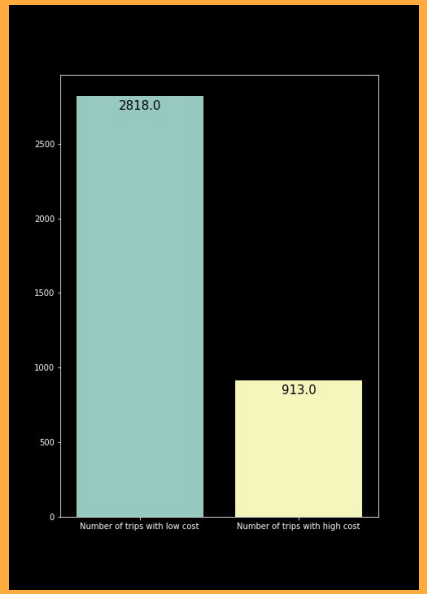
Error Analysis

What could make the prediction better?

What's causing problems for the prediction?

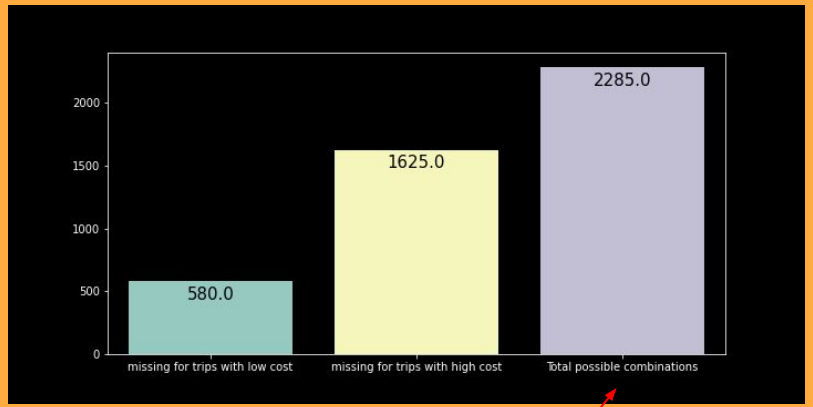
1) The data is highly "imbalanced"

→ Lack of data for high Total Cost



2) Most features are categorical

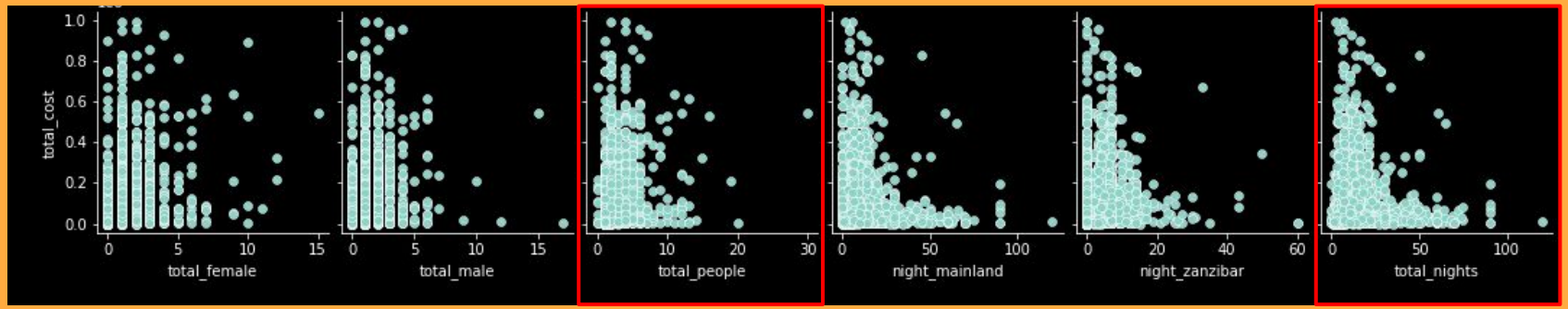
→ Not enough data to cover all occurring feature-category combinations



Definition of high Total Cost:
Total Cost > 10,000,000

What more causes problems for the prediction?

3) There seem to be no correlations between the numerical features



Solution:

More data is needed! (trips with high Total Cost)

Thank you