Turtle Rescue Forecast Challenge

Machine Learning Project
Lavdar Aliko & Kilian Gedat
15/01/2024

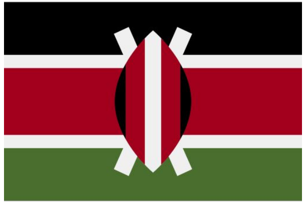
Guideline

1. Project Overview
2. Dataset Overview
3. Data Cleaning and Preprocessing
4. EDA
5. Model Development
6. Results
7. Future Steps

Project Overview

Goal and Objectives

- Help Kenyan non-profit organization Local Ocean Conservation to **anticipate the number of turtles** they **will rescue from each of their rescue sites**. An accurate prediction will enable Local Ocean Conservation to **allocate staff and resources more efficiently**.
- Baseline Model: We will have a higher capture number of turtles in dry season, because hatching season is aligned to the dry season.



Scope and Duration

- is a zindi learning competition without closing date (zindi link here)
- possible to earn zindi points

Collaborators and Stakeholders

- LOC is a private, not-for-profit organisation committed to the protection of Kenya’s marine environment
- they support the communities and coastal areas in Watamu and Diani, Kilifi County with marine conservation and community development projects - centered around a holistic approach to conservation
- LOC has been doing marine conservation for over 20 years

LOCAL OCEAN CONSERVATION

NGO Website

Dataset Overview

Data Source

- historic data on the number of turtles rescued from each site from 1998 until 2018 incl. : Capture Method, Fisher (Capturer), Turtle Tags, Condition, Species, etc.
- based on three datasets: train data, capture site category and sample submission

Key Features

- Capture Site, Capture Number and Capture Date

Size and Structure of the Data

- 3 datasets: 18062 rows and 23 columns
- final data: 23727 rows and 13 columns

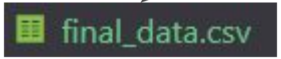
Any Challenges Encountered

- dataset contains empty data for 2019 (prediction year)
- training and testing will be made with data from 1998 to 2018

Data Cleaning and Preprocessing

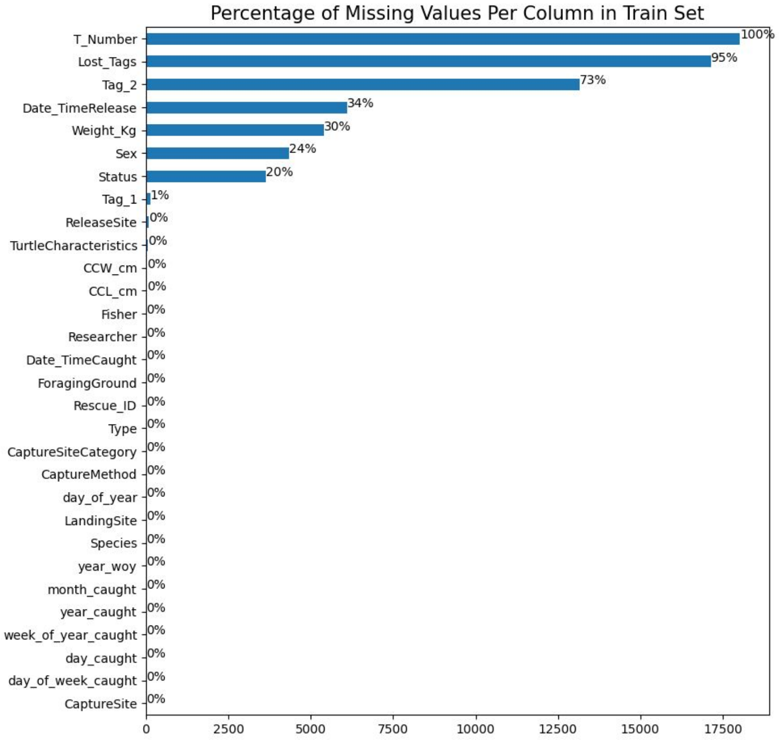
Cleaning Steps Undertaken

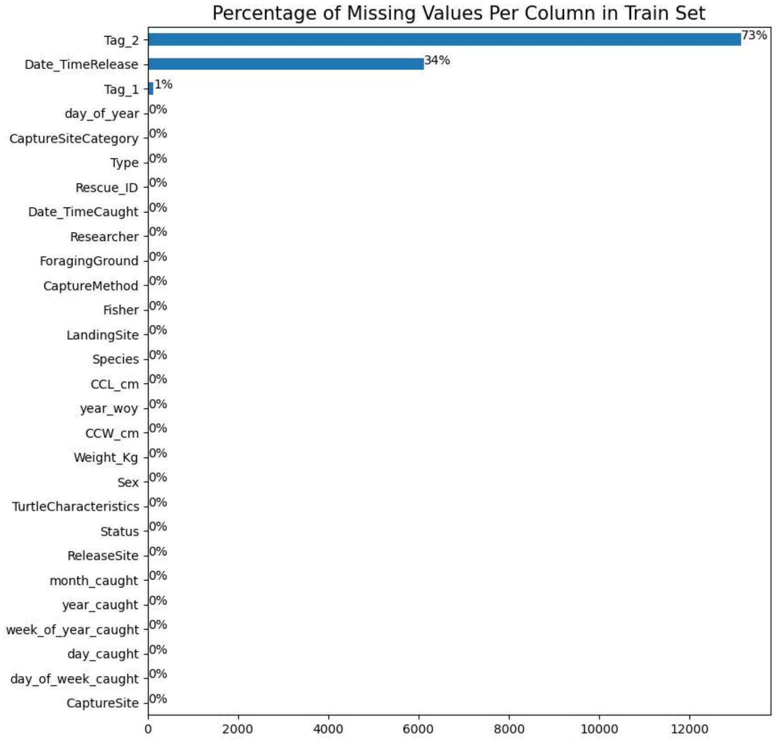
- merged train data and capture category on capture site
- split date column into week of year, etc.
- sample sub: split ID in year of week and capture site
- split into 2019(prediction data) and everything before 2019 (train data)
- create new csv’s


final_data.csv

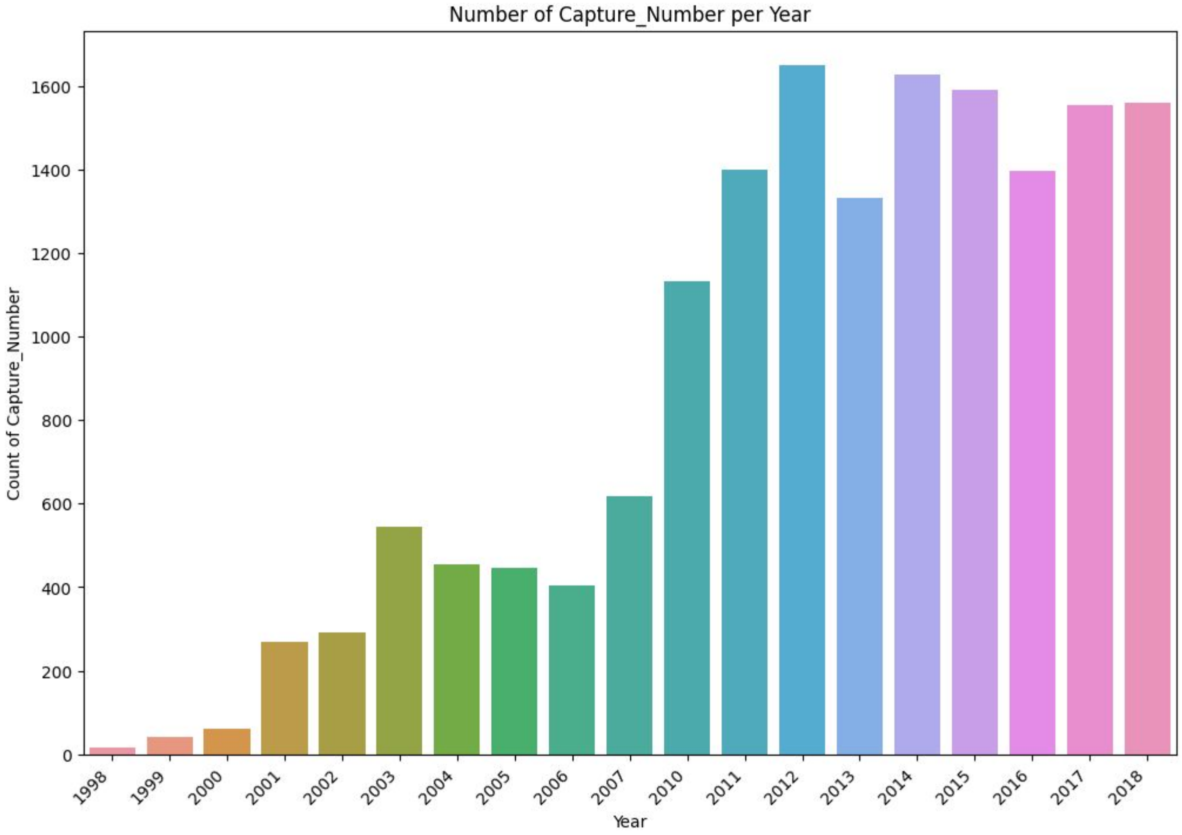
submission_data.csv

Data Cleaning and Preprocessing

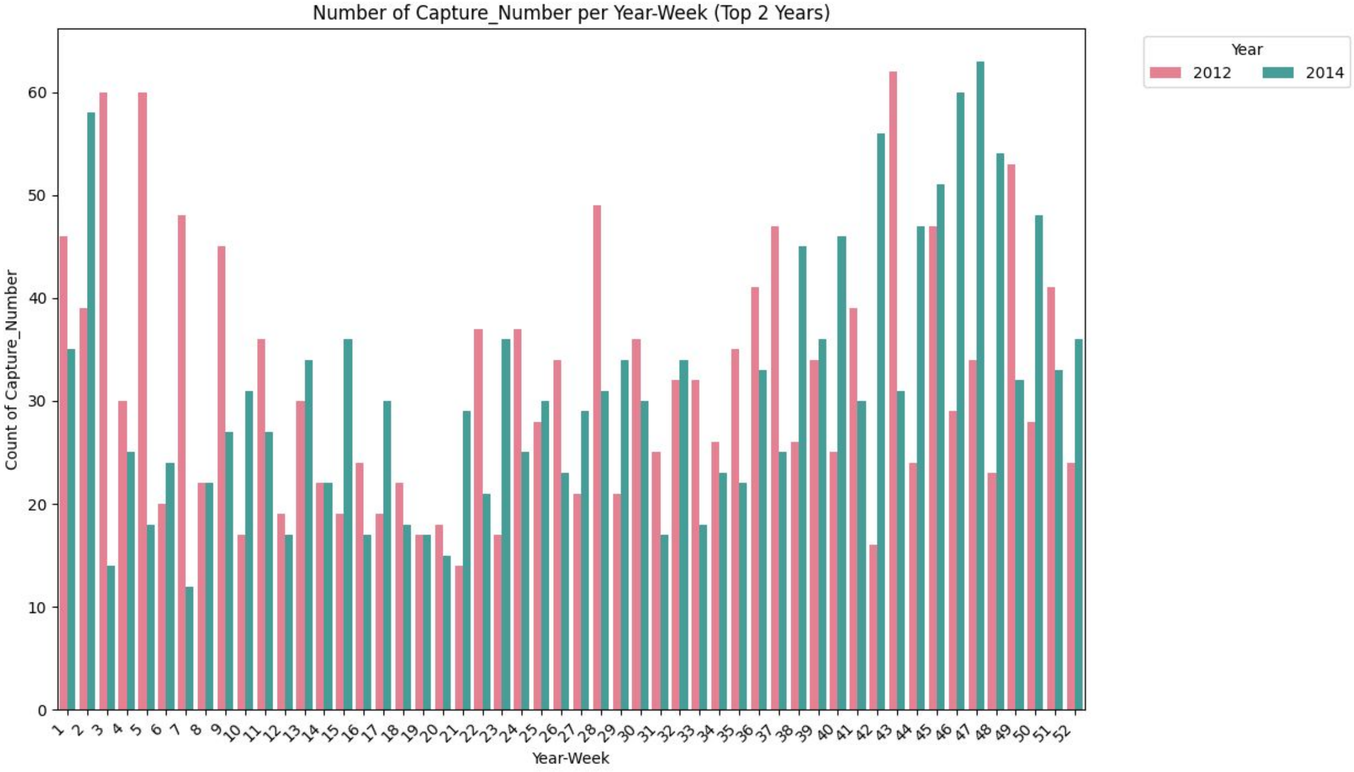




Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)



Model Development

Choice of Models

- Linear Regression
- Random Forest
- XGBRegressor
- KNN

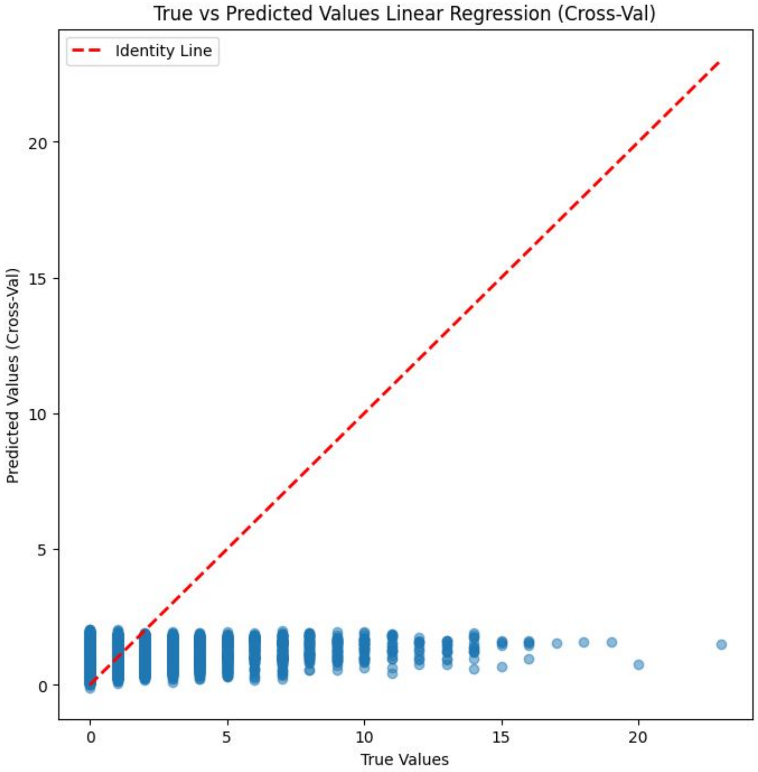
Training and Testing Split

- several different attempts to train, test, split

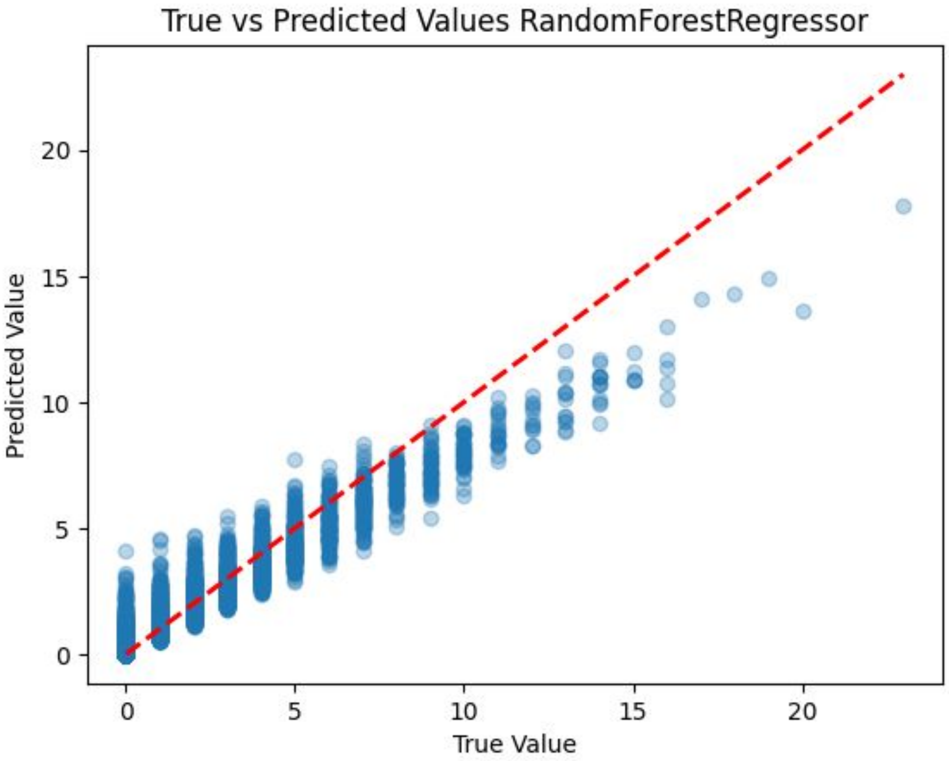
Model Evaluation Metrics

- given metric RMSE

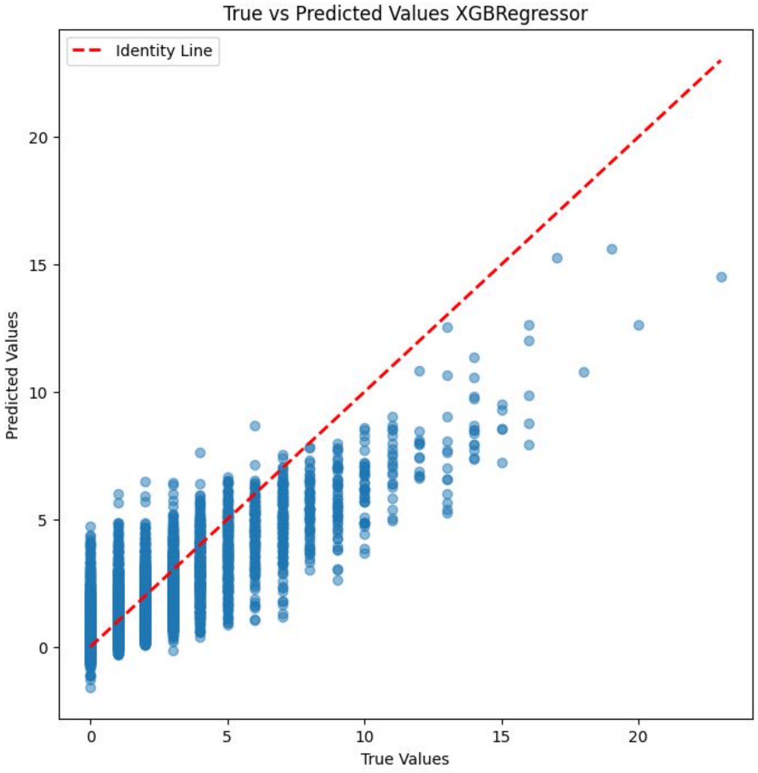
Results - Linear Regression



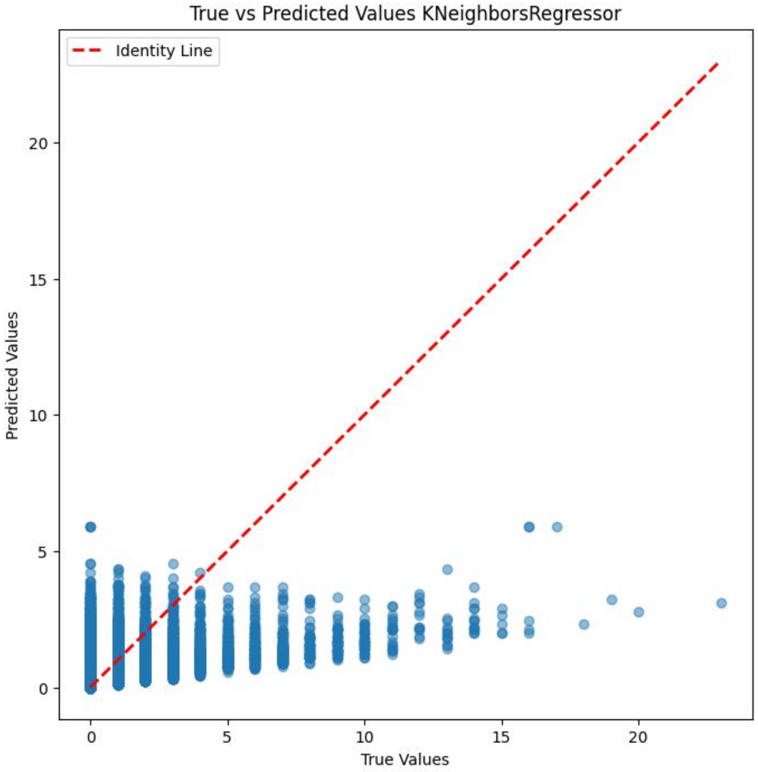
Results - Random Forest



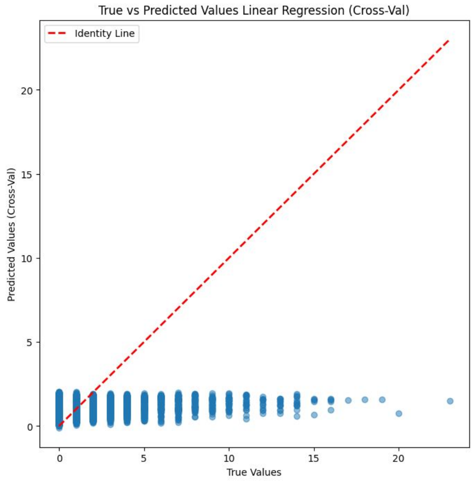
Results - XGBRegressor

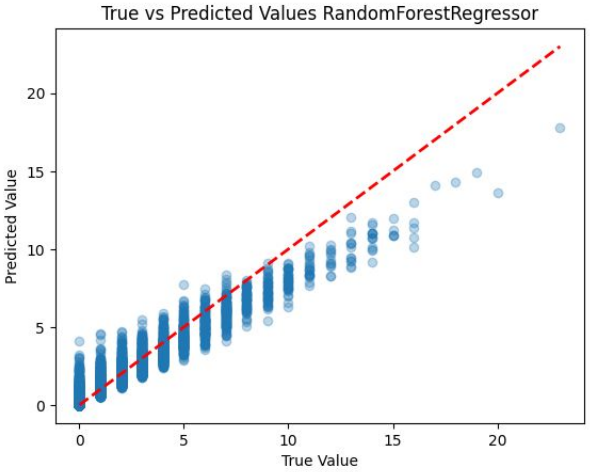


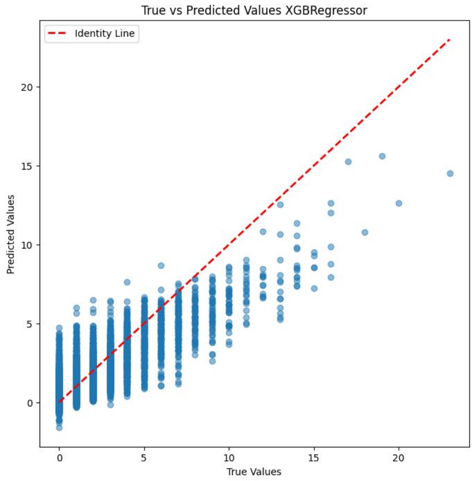
Results - KNeighborsRegressor

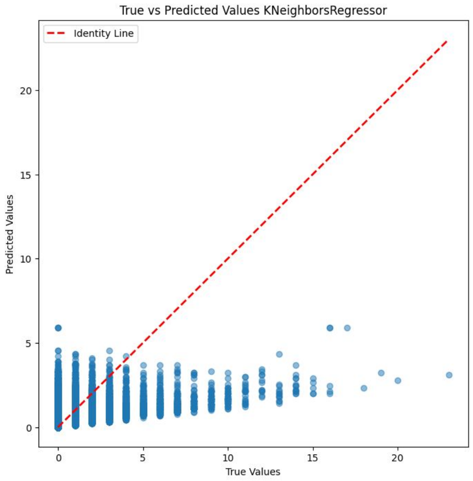


Results


True vs Predicted Values Linear Regression (Cross-Val)


True vs Predicted Values RandomForestRegressor


True vs Predicted Values XGBRegressor

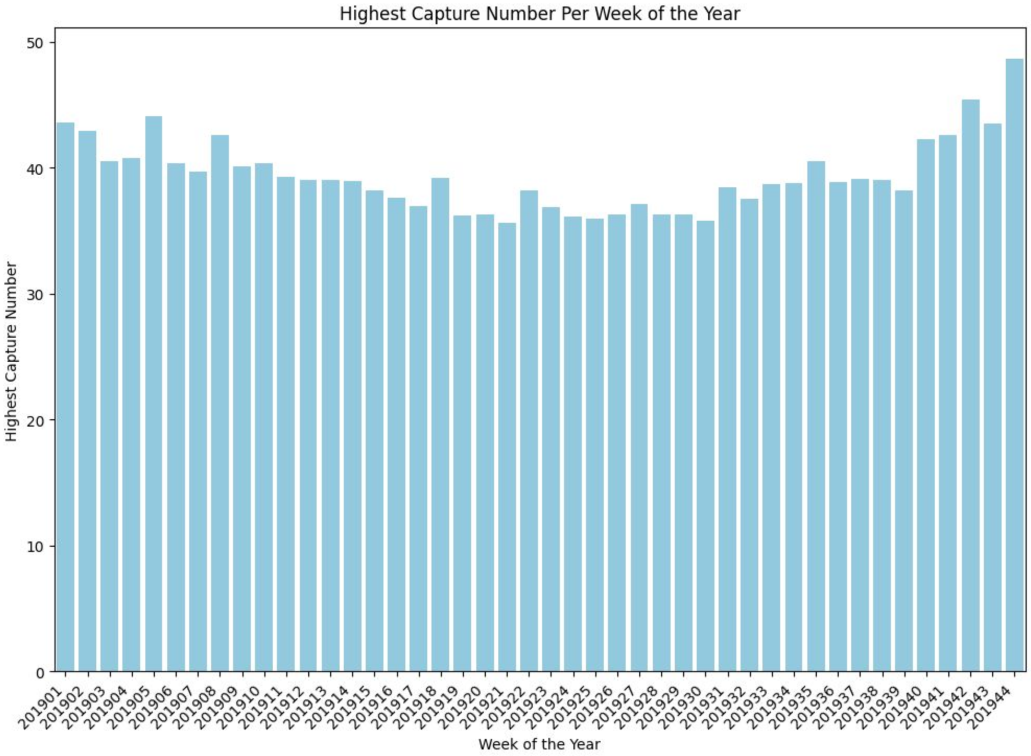

True vs Predicted Values KNeighborsRegressor

Submission Data

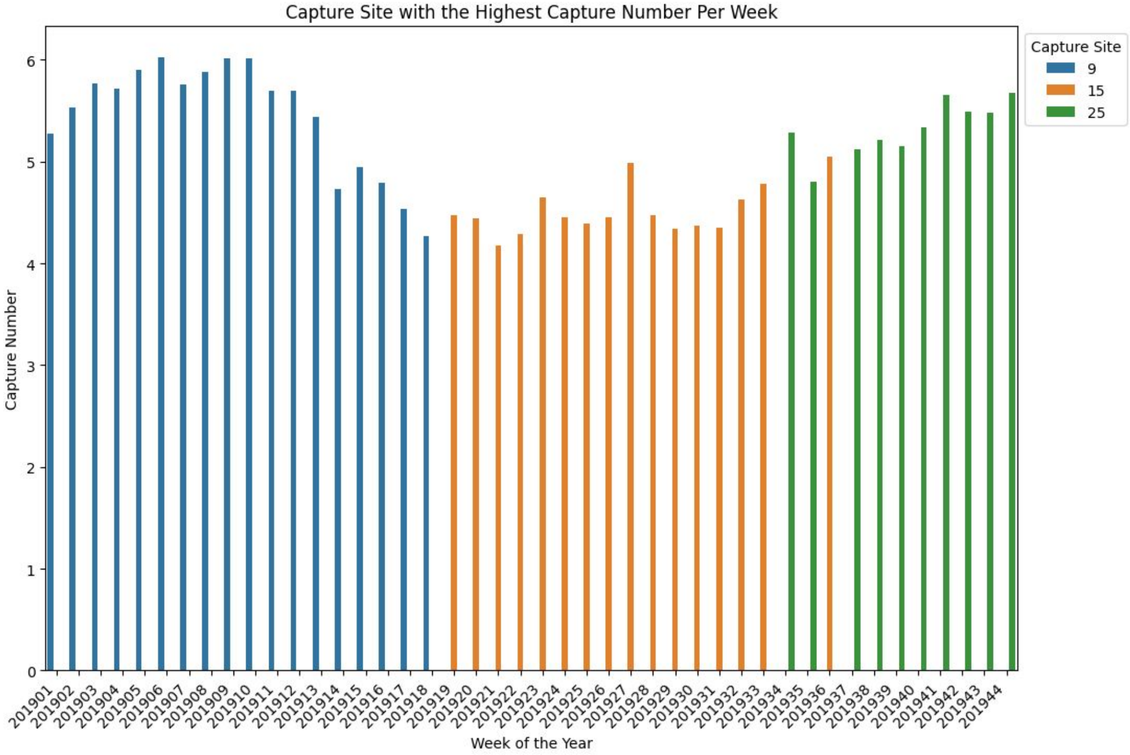
1654

Number of rescued turtles according to our model prediction for 2019.

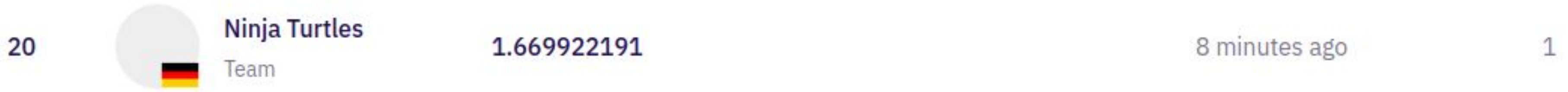
Results - Highest Capture Number per Week (2019)



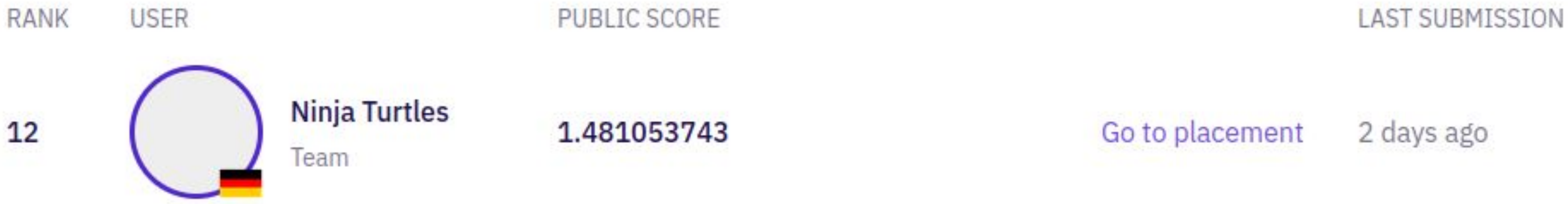
Results - Capture Site with highest Captures p Week



Results - Submission Score



20	Ninja Turtles Team	1.669922191	8 minutes ago	1



RANK	USER	PUBLIC SCORE		LAST SUBMISSION
12	Ninja Turtles Team	1.481053743	Go to placement	2 days ago

Future Steps

Potential Improvements to the Model

- separate Creek and Beach (Foraging Grounds)
- include holidays into our data

Additional Data Sources or Features

- consider the weather conditions
- research the species and their hatching seasons
- connect the lifecycle of the turtles and the weather

Thank you for your attention!