Turtle Rescue Forecast

Spiced Academy ML project, Aug/Sep 2023

David, Don, Kaja & Lara

Team Turtle

Data Scientists with backgrounds in

- Donatello — chemistry
- Leonardo — molecular medicine
- Michelarangelo — geometry
- Raphael — naval architecture

LOCAL OCEAN CONSERVATION

Introduction

- Stakeholder: Local Ocean Conservation (Kenyan non-profit)
- LCO rescues sea turtles with a “by-catch net release” program
- Challenge: predictive model for anticipating the number of rescued turtles
- Model will help LCO to plan staff schedules and budget

LOCAL OCEAN CONSERVATION

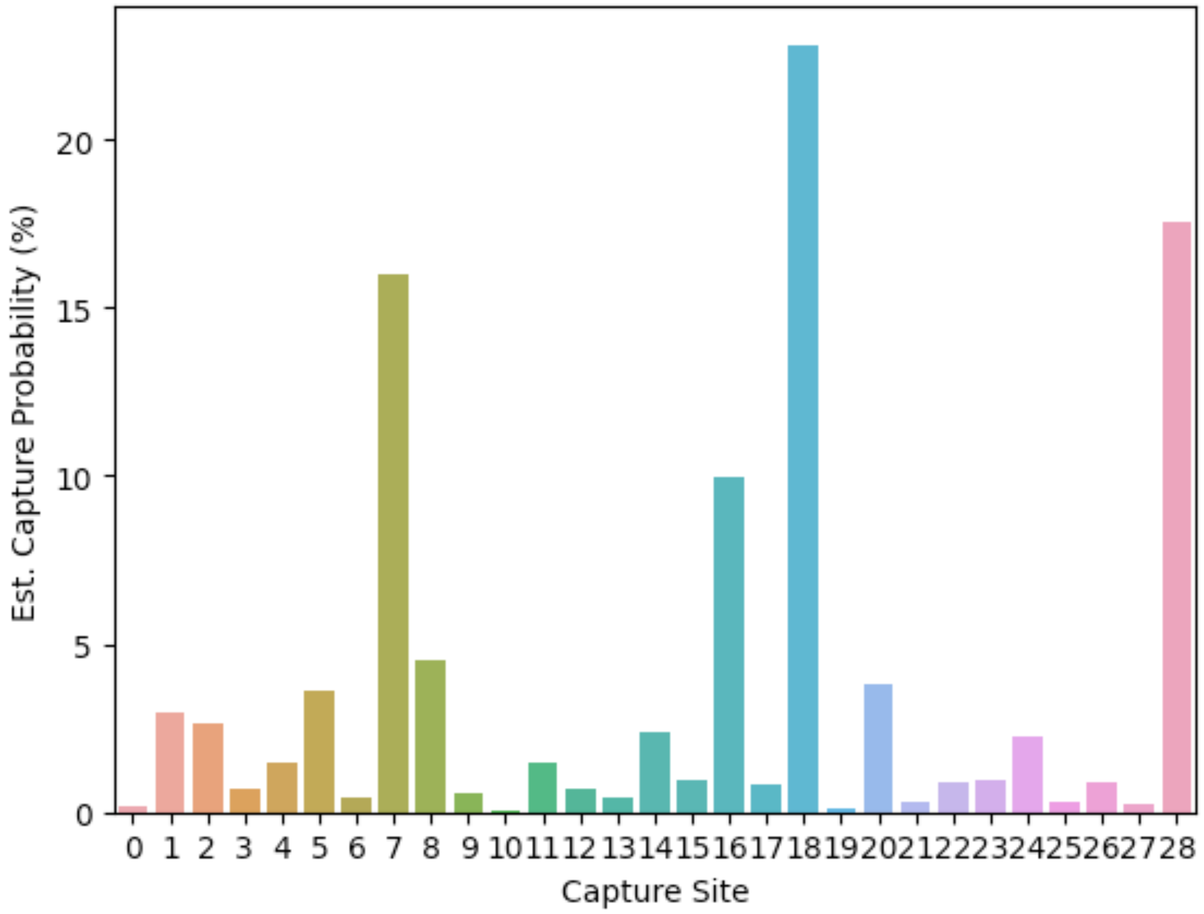
Data and goal

- Records of turtle rescues during the years 1998 — 2018 (Zindi)
- Mainly categorical entries, many missing values & redundancies
- **Goal:** Predict the number of turtles per site per week for 2019 onwards
- Success evaluated by root mean squared error (RMSE)

LOCAL OCEAN CONSERVATION

Data and goal

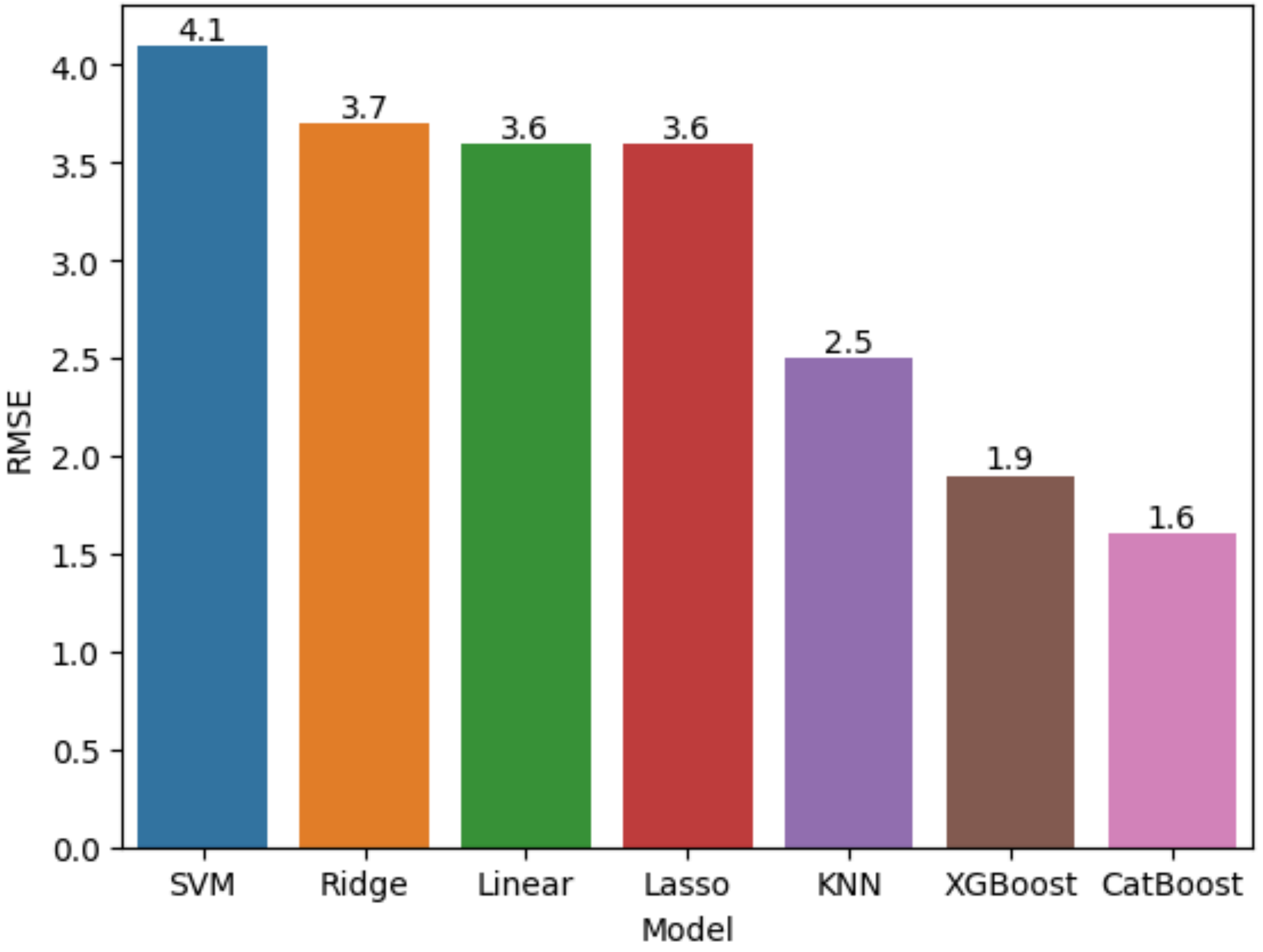
- Records of turtle rescues during the years 1998 — 2018 (Zindi)
- Mainly categorical entries, many missing values & redundancies
- **Goal:** Predict the number of turtles per site per week for 2019 onwards
- Success evaluated by root mean squared error (RMSE)



LOCAL OCEAN CONSERVATION

Models

Linear regression as baseline



LOCAL OCEAN CONSERVATION

Best Model: CatBoost

Competition Leaderboard

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.

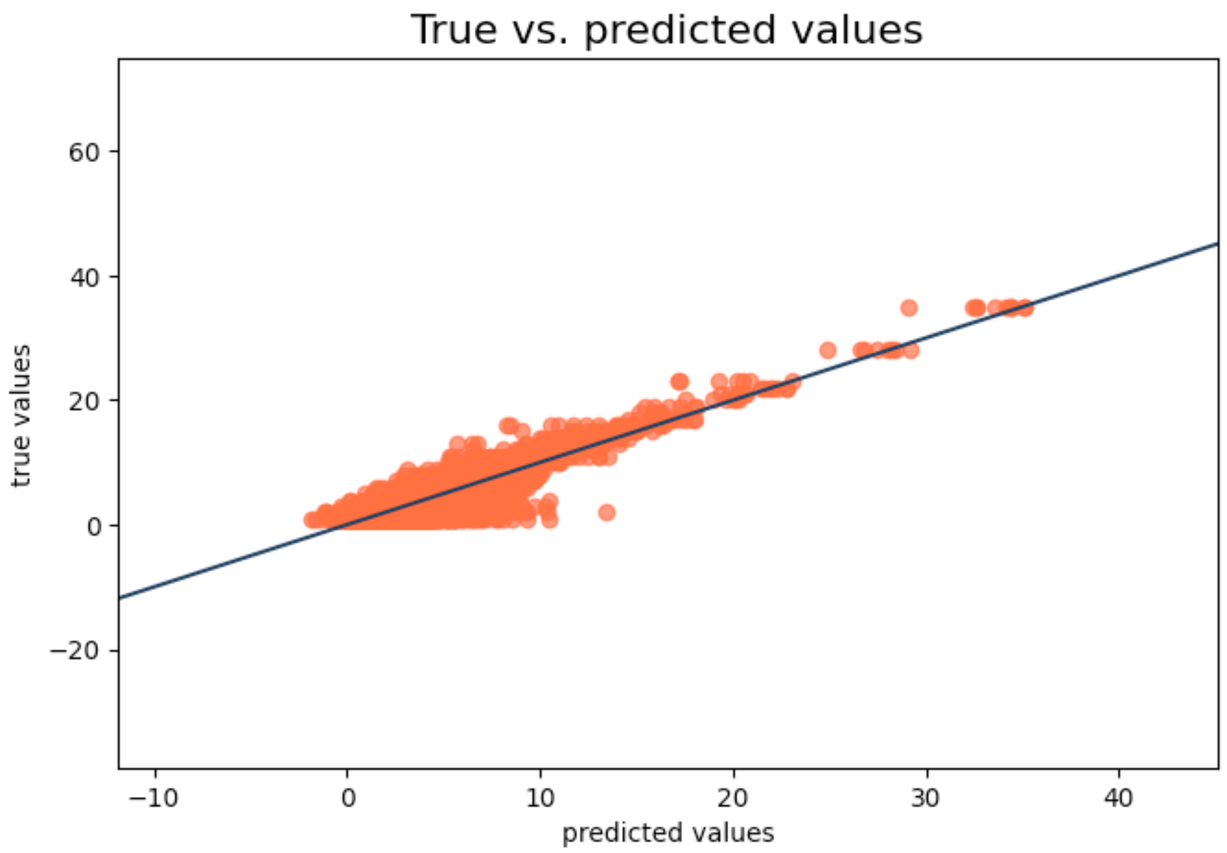
RANK	USER	PUBLIC SCORE	LAST SUBMISSION	# SUBMITTED
1	 MICADEE	1.409399952	over 1 year ago	30
2	PRO_DEL Team	1.413356604	~1 year ago	65

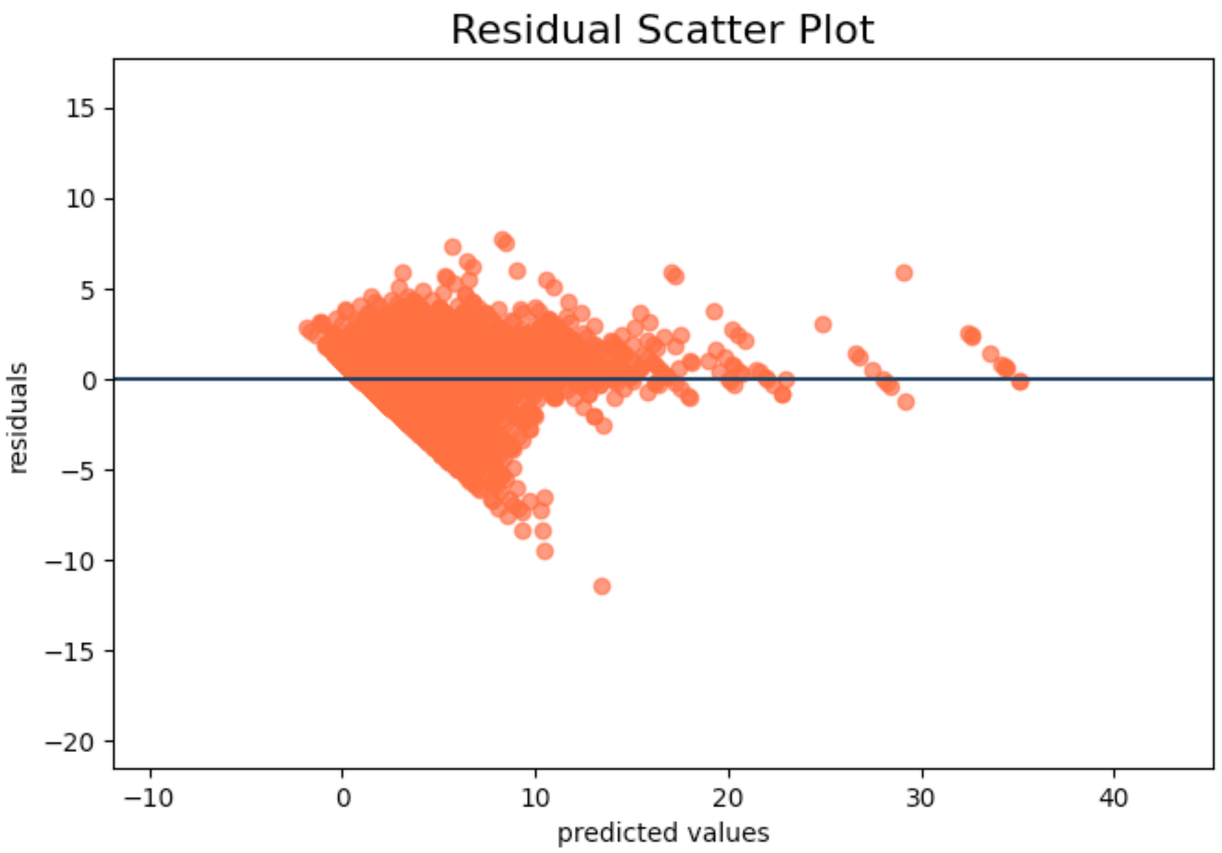
Our best model, CatBoost would put us in position 18 in the Zindi leaderboard (RMSE = 1.6)

LOCAL OCEAN CONSERVATION

Best Model: CatBoost

Error analysis





LOCAL OCEAN CONSERVATION

Our data science product

A. **Shell script**

Our script can

- load and clean raw data,
- perform feature engineering
- train the model and make predictions
- test and evaluate the predictions
- perform error analysis

```
Number of predictions:  5419
RMSE:  1.644
```

LOCAL OCEAN CONSERVATION

Our data science product

A. **Shell script**

Our script can

- load and clean raw data,
- perform feature engineering
- train the model and make predictions
- test and evaluate the predictions
- perform error analysis

```
Number of predictions:  5419
RMSE:  1.644
```

B. **Python script**

Our script runs in terminal and can

- generate and run a model
- Input: model, X_test, y_test
- Output: y_predict, RMSE

```
❯ python3 predict.py models/cat_boost_model.sav data/X_test.csv data/y_test.csv
Number of arguments: 4 arguments.
Argument List: ['predict.py', 'models/cat_boost_model.sav', 'data/X_test.csv', 'data/y_test.csv']
RMSE on test is: 1.6440787361944256
```

LOCAL OCEAN CONSERVATION

Discussion / Conclusion / Outlook

- Built a ML model based on CatBoost algorithm for predicting the number of rescued turtles
- Future work I: perform time series analysis
- Future work II: more feature importance analysis to improve feature selection
- Deeper pattern analysis, e.g., per species
- Give recommendations for best practices for future data collection

LOCAL OCEAN CONSERVATION

Appendix

Our take-home-messages

- Data cleaning is hard work and crucial
- Feature engineering was a necessity
- git collaboration needs practice
- Team work makes the dream work