akaigraham / capstone_project Public

0 stars 0 forks

Star Unwatch

Code Issues Pull requests Actions Projects Wiki Security

main

akaigraham added notebook pdf ... 3 minutes ago 51

View code

README.md

Capstone Project - Predicting Fishing Habits Using AIS and World Ocean Database Database

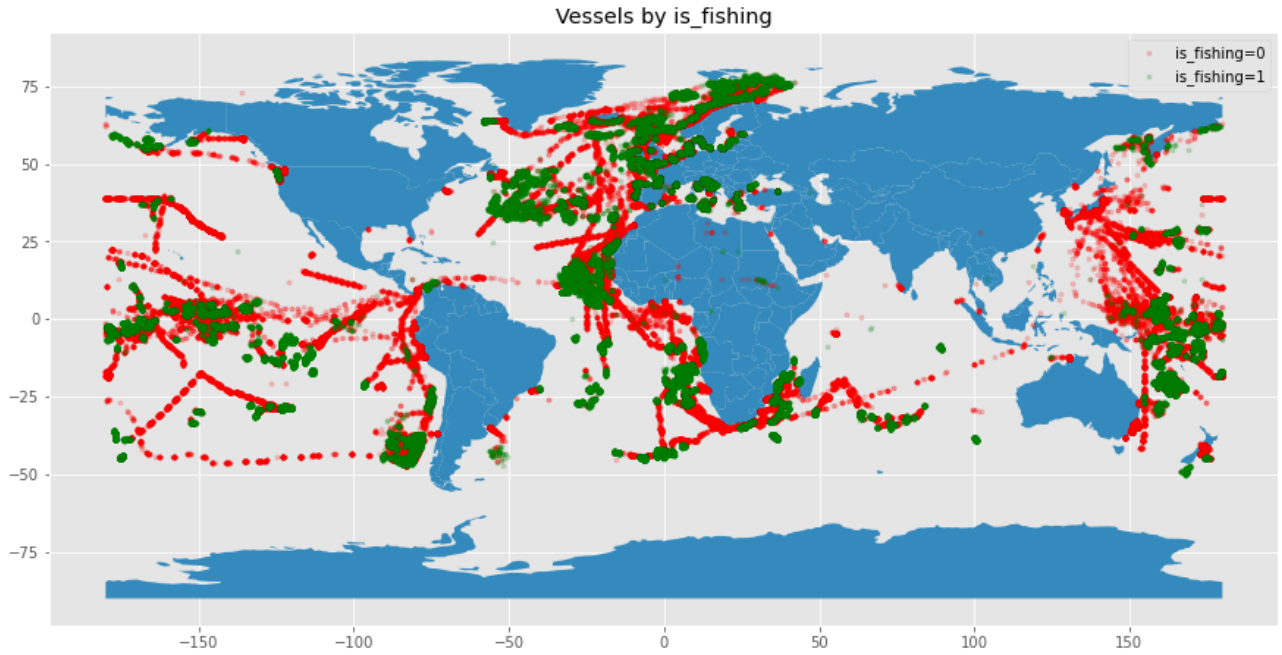
README Outline

Within this `README.md` you will find:

1. Introduction
2. Overview of Repository Contents
3. Project Objectives
4. Overview of the Process
5. Findings & Recommendations
6. Conclusion / Summary

Introduction

Build a classifier to identify whether a vessel is fishing. Ultimate goal is to create a classifier that can be used by policymakers, regulators, and other stakeholders involved with preservation of ocean resources to identify vessels that are fishing.



Repository Contents

1. `README.md`
2. `notebook.ipynb`
3. `/datasets`
4. `/tide_api_call.ipynb`

Project Objectives

Build a classifier to predict whether a vessel is engaged in fishing activity in context of providing policymakers, regulators, and other ocean resource stakeholders a tool to identify vessels that are fishing. Follow CRISP-DM machine learning process to explore dataset, prepare data for modeling, modeling, and post-model evaluation. Main performance metrics focused on were accuracy and recall as our dataset is not very sensitive to producing false positives.

Overview of the Process:

Following CRISP-DM, the process outlined within `notebook.ipynb` follows 5 key steps, including:

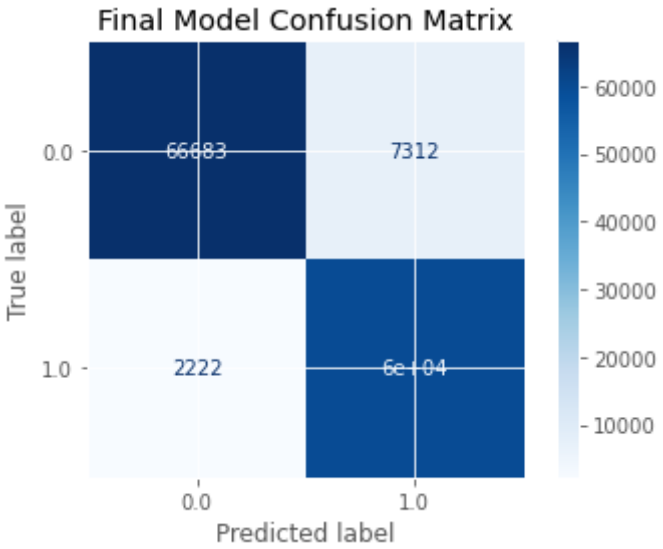
1. Business Understanding: Outlines facts and requirements of the project.
2. Data Understanding: focused on unpacking data available and leveraged throughout classification tasks. Section will focus on the distribution of our data, and highlighting relationships between target and predictors.
3. Data Preparation: further preprocessing of our data to prepare for modeling. Includes separating validation sets, handling missing values, and encoding certain columns
4. Modeling: this section iteratively trains a number of machine learning models, specifically using Decision Trees, Random Forests, XGBoost, and Neural Networks.
5. Evaluation: Final / optimal model is selected and final performance metrics of final model are discussed and evaluated. Focused primarily on accuracy and recall as performance metrics.

Findings & Recommendations

The best performing model identified was a tuned random forest model. Final model scores:

```
Accuracy:
Training Set: 0.9380431102584359
Testing Set: 0.9298294680905873
-----------------------------
Precision:
Training Set: 0.8991085516328288
Testing Set: 0.8908070007765366
-----------------------------
Recall:
Training Set: 0.973149879055936
Testing Set: 0.9640883084979152
-----------------------------
F1 Score:
Training Set: 0.9346651764925026
Testing Set: 0.9260000931402226
-----------------------------
ROC AUC:
Training Set: 0.9409182901449071
Test Set: 0.9326354104149147
```

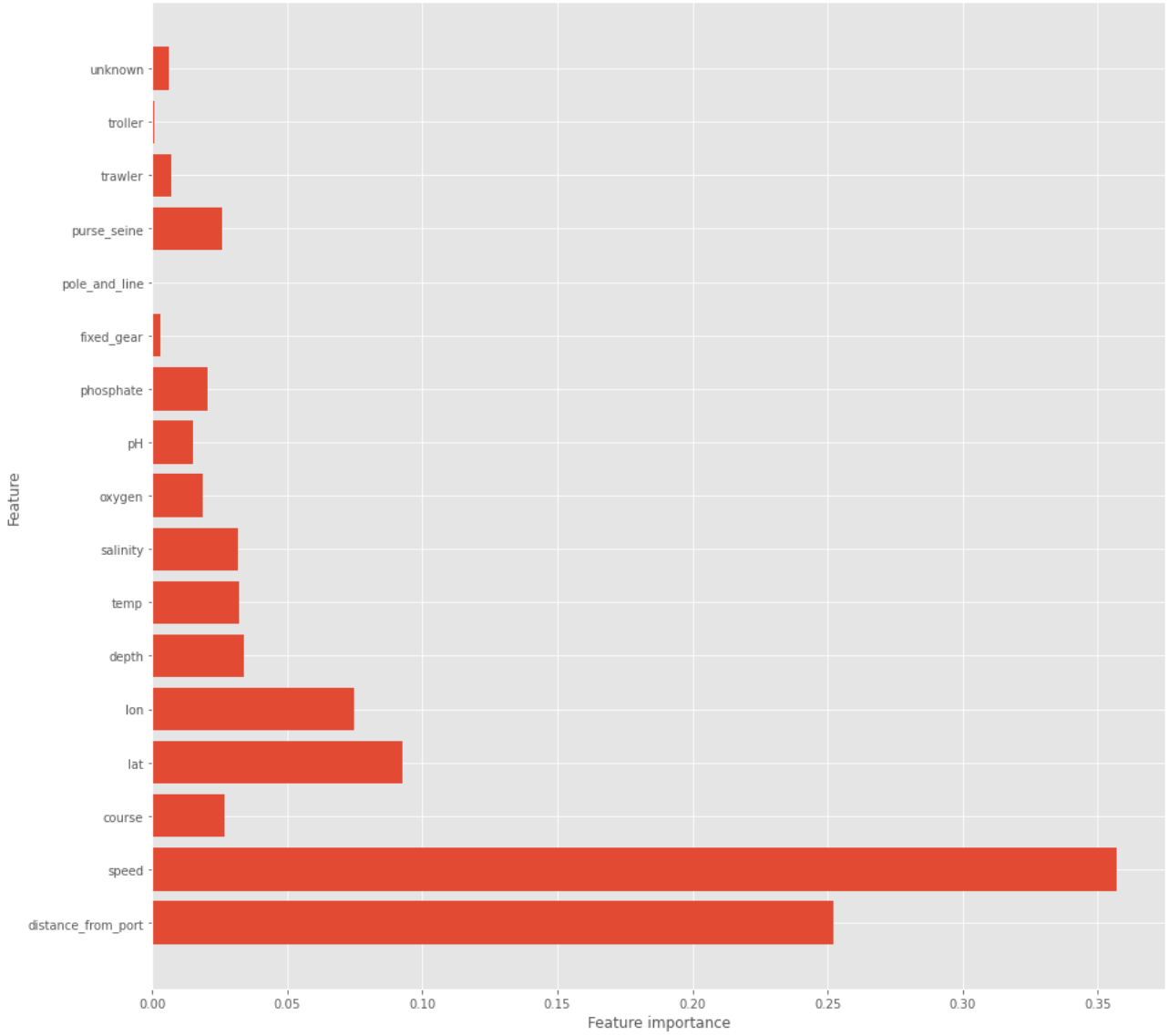
Looking at the confusion matrix for the final model, we can see that there are far more false positives than false negatives, as indicated by the number in the upper right quadrant vs. lower left quadrant. We are more interested in a strong recall, and therefore are not too worried with the level of false positives generated given the end applications.



The top 6 most important features to our model were:

1. `speed`
2. `distance_from_port`
3. `lat`
4. `lon`
5. `depth`
6. `temp`

When looking to identify a vessel that is fishing or not, these features will be most important to have as inputs.



Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%