Development of an application based on a Machine Learning model to identify potentially dangerous asteroids.

bartlomiej brzostek

it_depends

AGENDA

it_depends

- **introduction**
  - about the project
  - definitions

- **Exploratory Data Analysis, EDA**
  - dataset
  - study of correlations between variables
  - challenges of model development:

- **models**
  - XGBoost
  - choosing the best model

- **deployment**
  - building a mock-up
  - haas application

INTRODUCTION

ABOUT THE PROJECT

ABOUT THE PROJECT

In the era of rapid technological advancement, monitoring threats from outer space has become crucial. One significant danger comes from asteroids that could potentially collide with Earth, causing catastrophic consequences. Our presentation, titled "Development of an Application Based on a Machine Learning Model to Identify Potentially Dangerous Asteroids," focuses on using advanced machine learning techniques to identify and assess the risk posed by these objects.

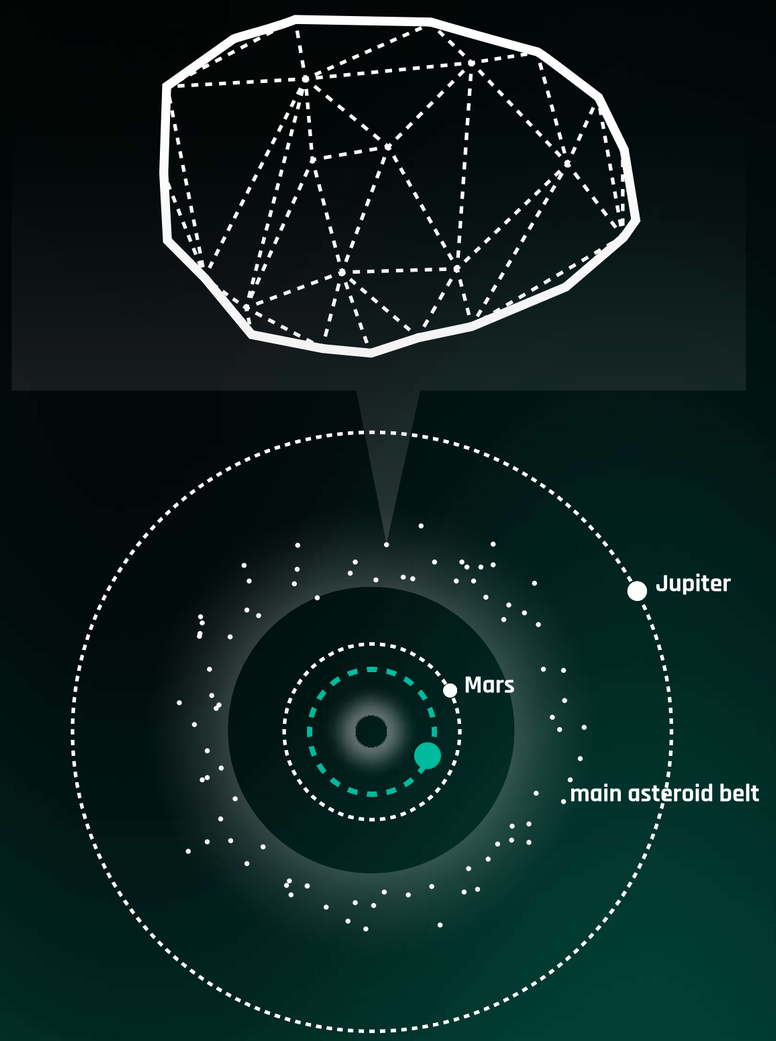
In this presentation, we will discuss the steps to create an effective predictive model, starting with NASA data analysis. This includes data processing, feature selection, and model training. We will present our results, covering evaluation metrics like precision and recall, and the challenges faced. Finally, we will showcase our application, featuring visualizations and interactive elements, enabling users to understand asteroid risks and make informed decisions.

DEFINITIONS

ASTEROID (PLANETOID)

Asteroid (planetoid) is a celestial body, which is not a planet, moon, comet or meteoroid, with a size from 1m to more than 1000 km.

Most of the recorded 1.3 million asteroids are located between the orbits of Mars and Jupiter - in the so-called main asteroid belt.

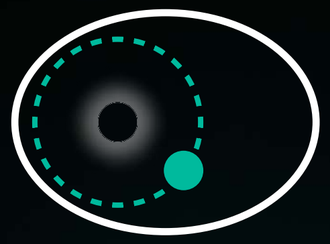


DEFINITIONS

POTENTIALLY HAZARDOUS ASTEROIDS

Potentially Hazardous Asteroids (PHA for short) - asteroids that move around the Sun in orbits that intersect the Earth's orbit or are close to it.

grupa Amora
orbita znajduje się całkowicie poza orbitą Ziemi



grupa Apolla
półoś wielka większa niż orbita Ziemi
orbity przecinają się

grupa Atena
półoś wielka mniejsza niż orbita Ziemi
orbity przecinają się

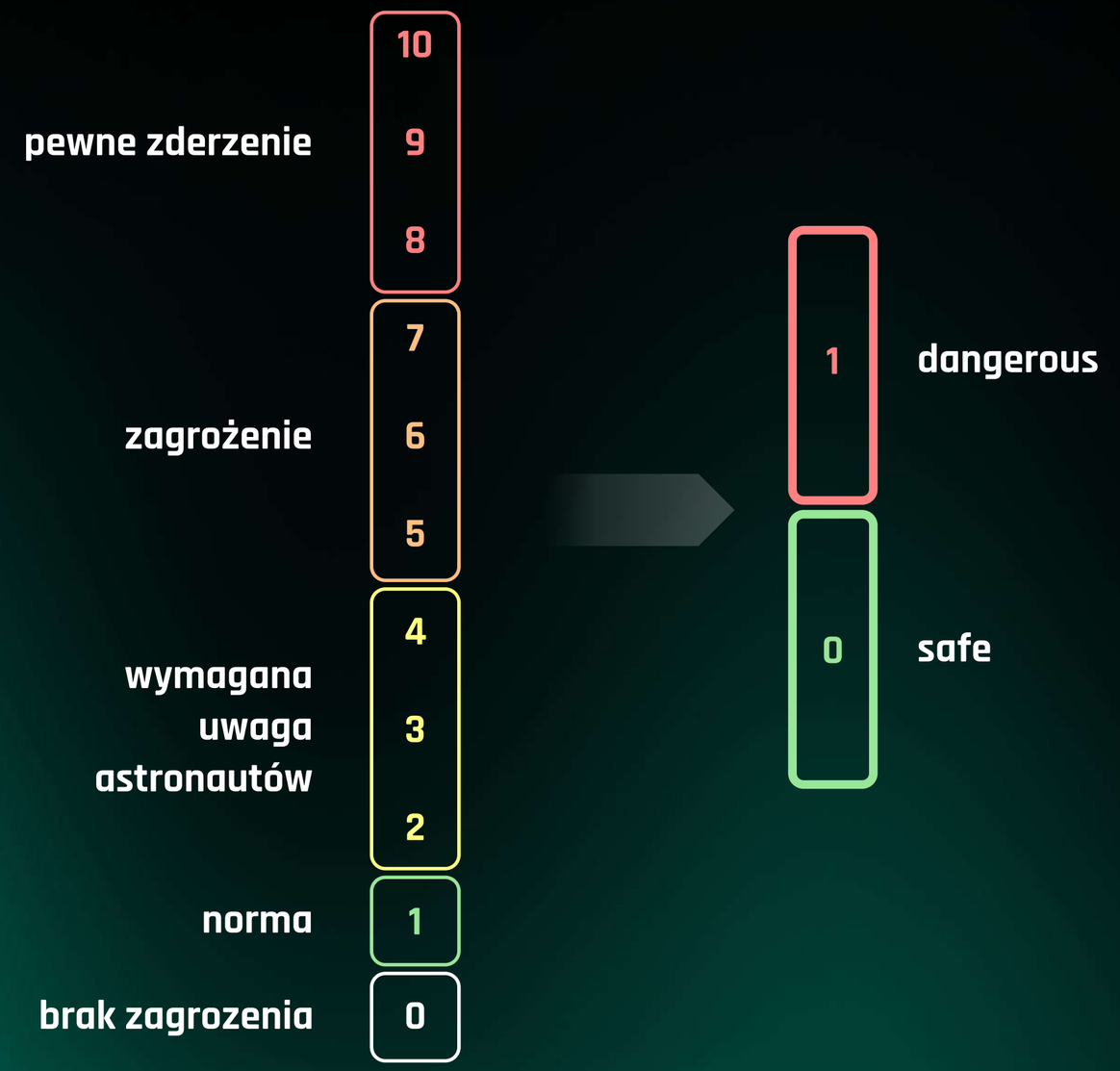
grupa Atiry
orbita zawiera się całkowicie wewnątrz orbity Ziemi

DEFINITIONS

TORINO SCALE

Torino Scale - an 11-point scale (from 0 to 10) by which NASA assesses the potential threat of space catastrophe posed by various objects approaching Earth from space.

For the purposes of the project, the scale has been reduced to a binary classification (0 - safe, 1 - hazardous).

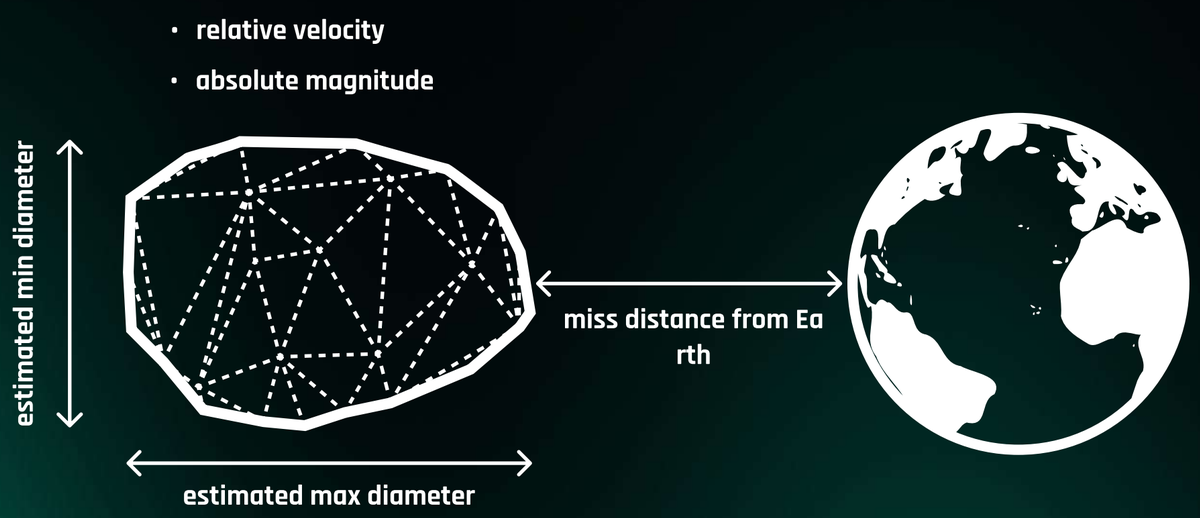


EXPLORATORY DATA ANALYSIS, EDA

DATASET

ANALYSED FEATURES OF THE ASTEROID

estimated min diameter (est_diameter_min)
estimated max diameter (est_diameter_max)
relative velocity (relative_velocity)
miss distance from Earth (miss_distance)
absolute magnitude (absolute_magnitude)



STUDY OF CORRELATIONS BETWEEN VARIABLES

CORRELATION MATRIX

Pearson correlation matrix

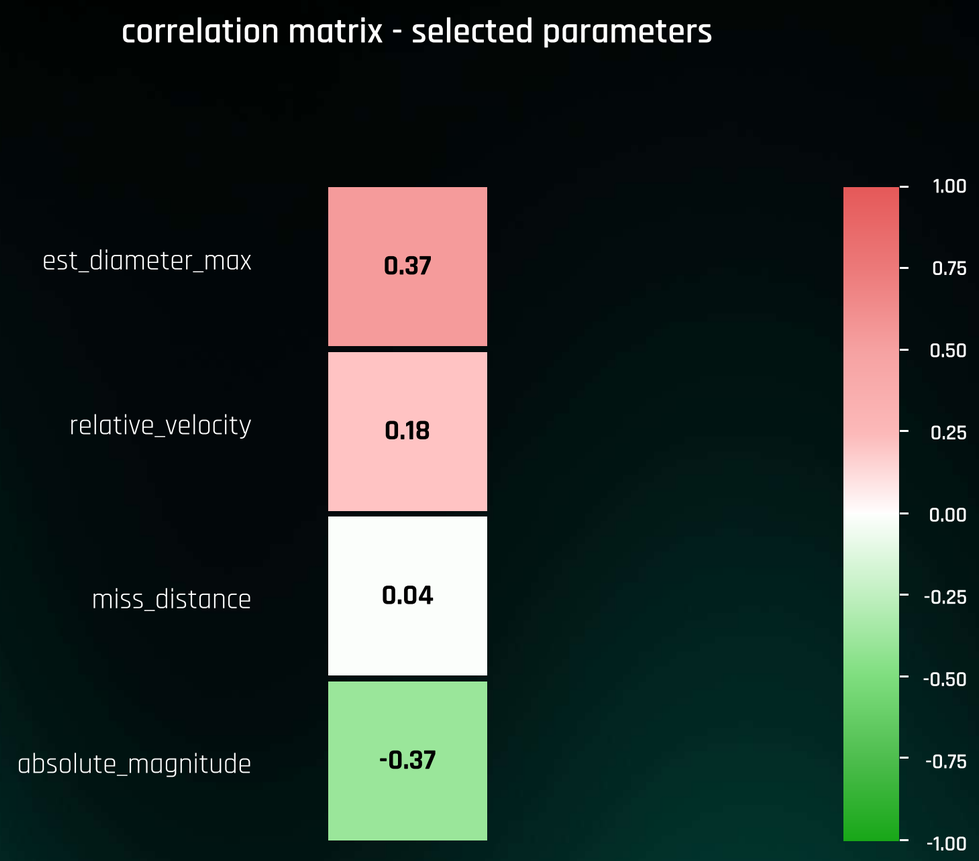
Both diameters (min and max) are correlated, so we will only use one of them to train the model. The max diameter will be used.

STUDY OF CORRELATIONS BETWEEN VARIABLES

CORRELATION MATRIX

Spearman correlation matrix

Correlations between variables and the potential asteroid potentially hazardous to Earth.



CHALLENGES OF MODEL DEVELOPMENT

UNBALANCED DATA SET

In the dataset, the number of records is unbalanced. Hazardous asteroids account for 9.7% of the cases.

This can be a challenge when training machine learning models.

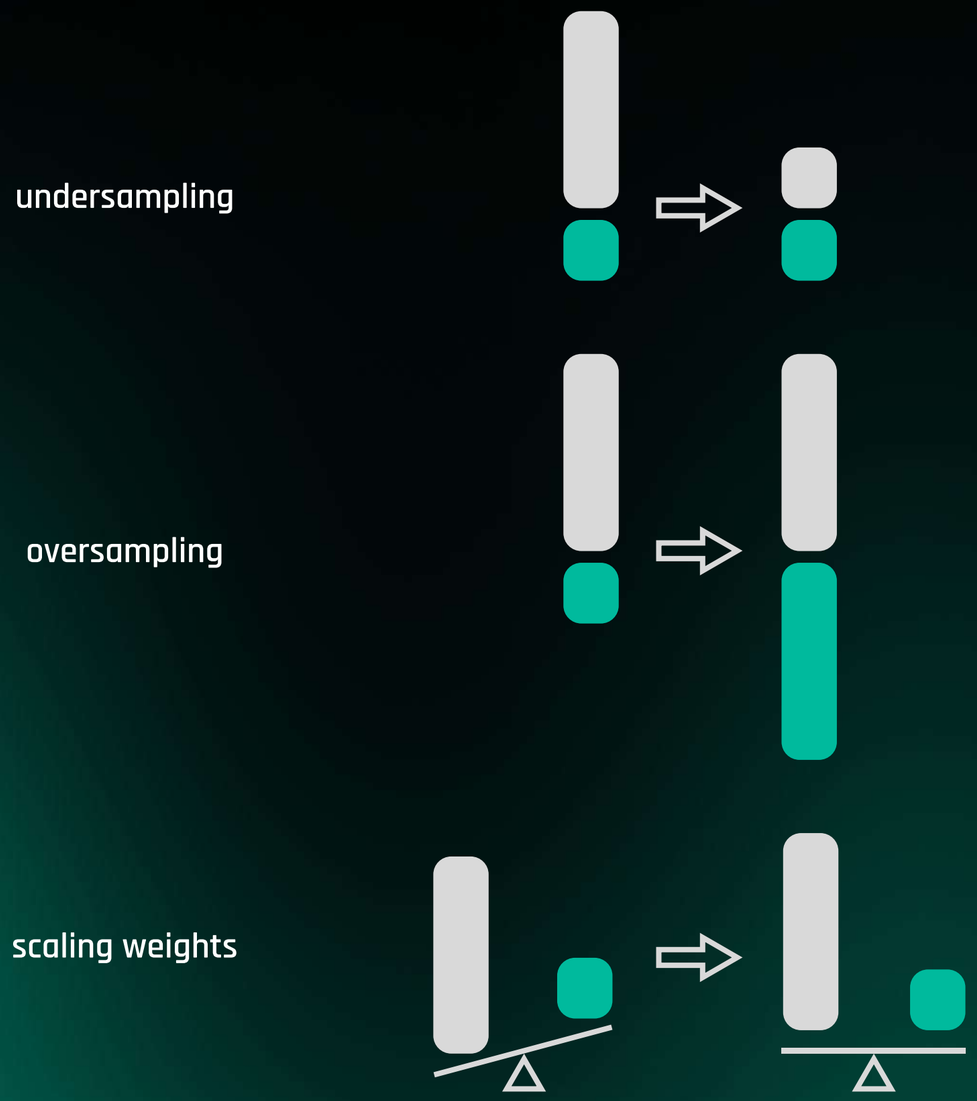
CHALLENGES OF MODEL DEVELOPMENT

BALANCING METHODS

undersampling
- Near miss - deleting data between both groups
- Random sampler

oversampling
- Random sampler
- SMOTE - synthetic data for minority group

increase/decrease weighting



MODELS

XGBOOST

XGBOOST INTRODUCTION

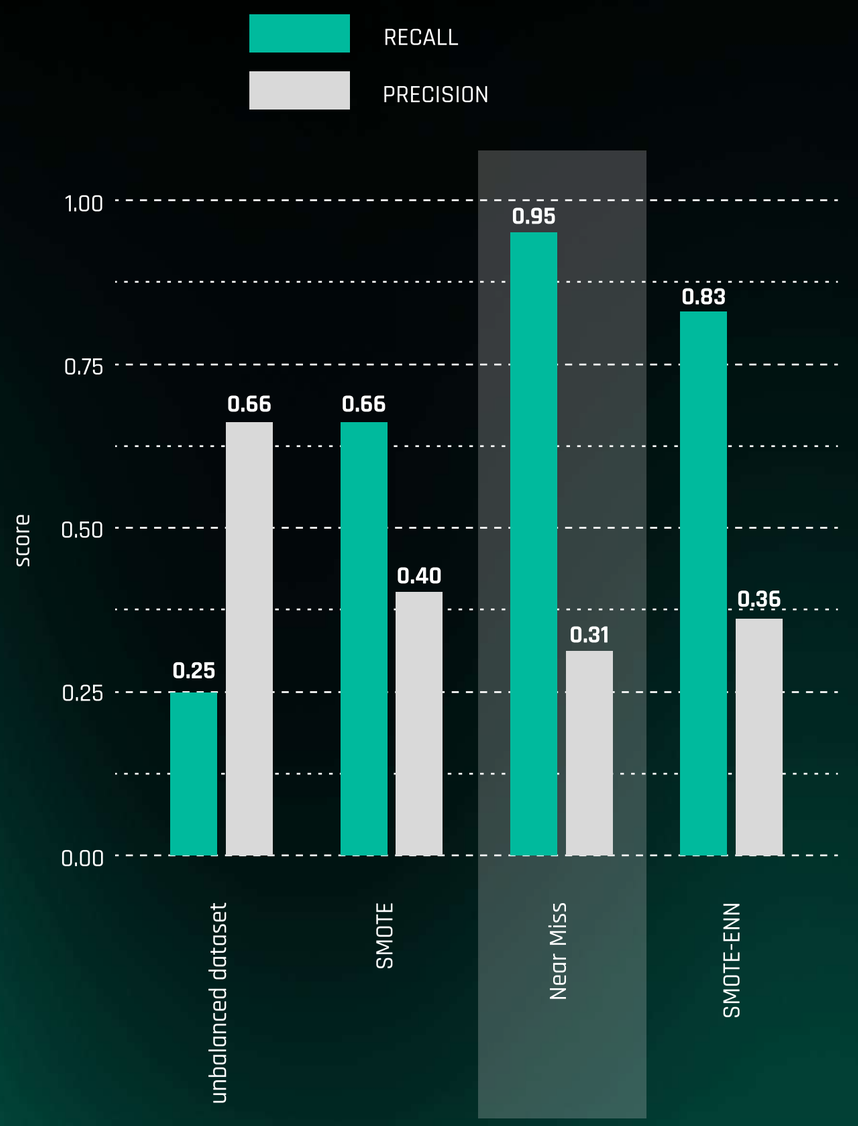
XGBoost is an optimized distributed gradient boosting library designed to be
highly efficient, flexible and portab le. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

XGBOOST

IMPACT OF BALANCING TRAINING DATA ON THE XGBOOST MODEL

The effect of balancing training data on XGBoost model performance.

The model's task is not to miss any potentially hazardous asteroid. In order to meet this criterion, the decisive parameter for model evaluation will be the value of the recall parameter.



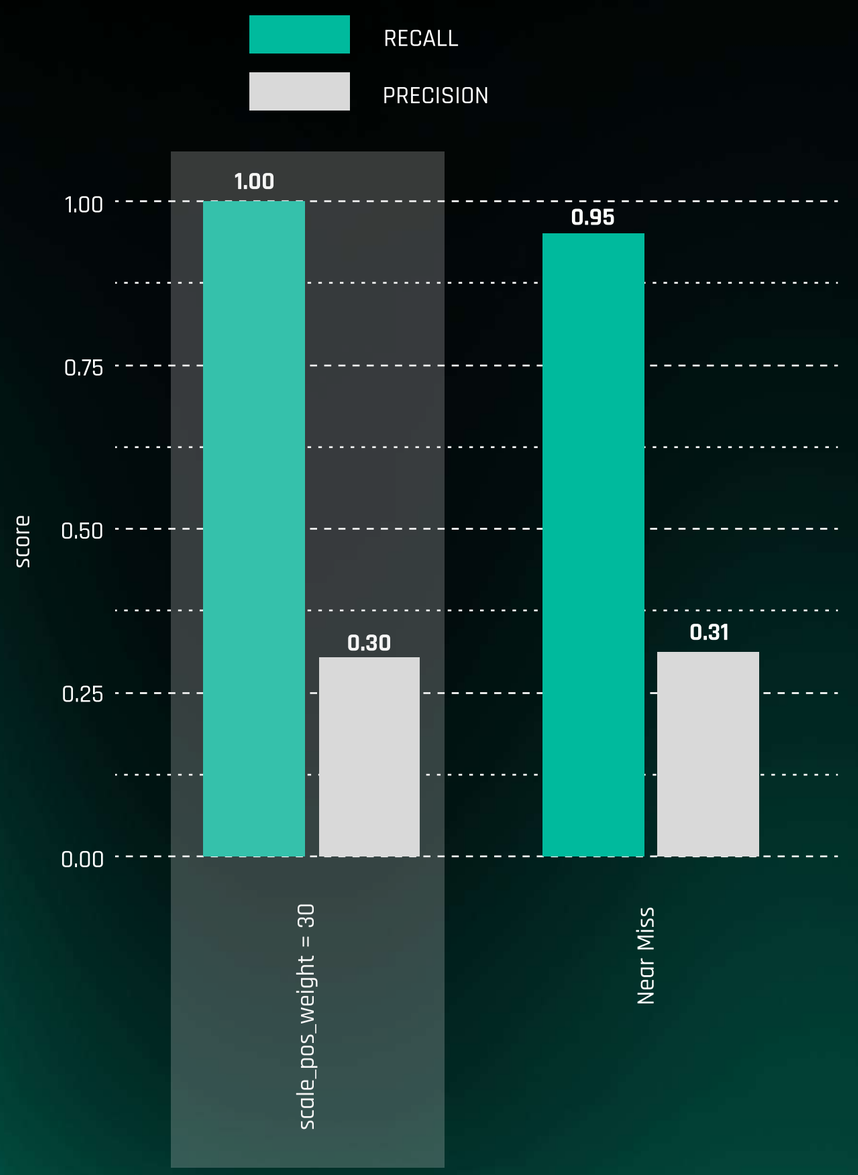
XGBOOST

OPTIMALISATION OF THE SCALE_POS_WEIGHT PARAMETER FOR THE XGBOOST MODEL

Optimisation of the scale_pos_weight parameter for the XGBoost model

The scale_pos_weight parameter: Adjusts the error weights for the positive (minority) class to improve the prediction accuracy of this class.

$$\frac{\text{number of records classified as 0}}{\text{number of records classified as 1}} = 9.3$$



CHOOSING THE BEST MODEL

FINAL MODEL USED IN THE APPLICATION

XGBClassifier

n_estimators = 50,
max_depth = 3,
learning_rate = 0.1,
scale_pos_weight = 30

DEPLOYMENT

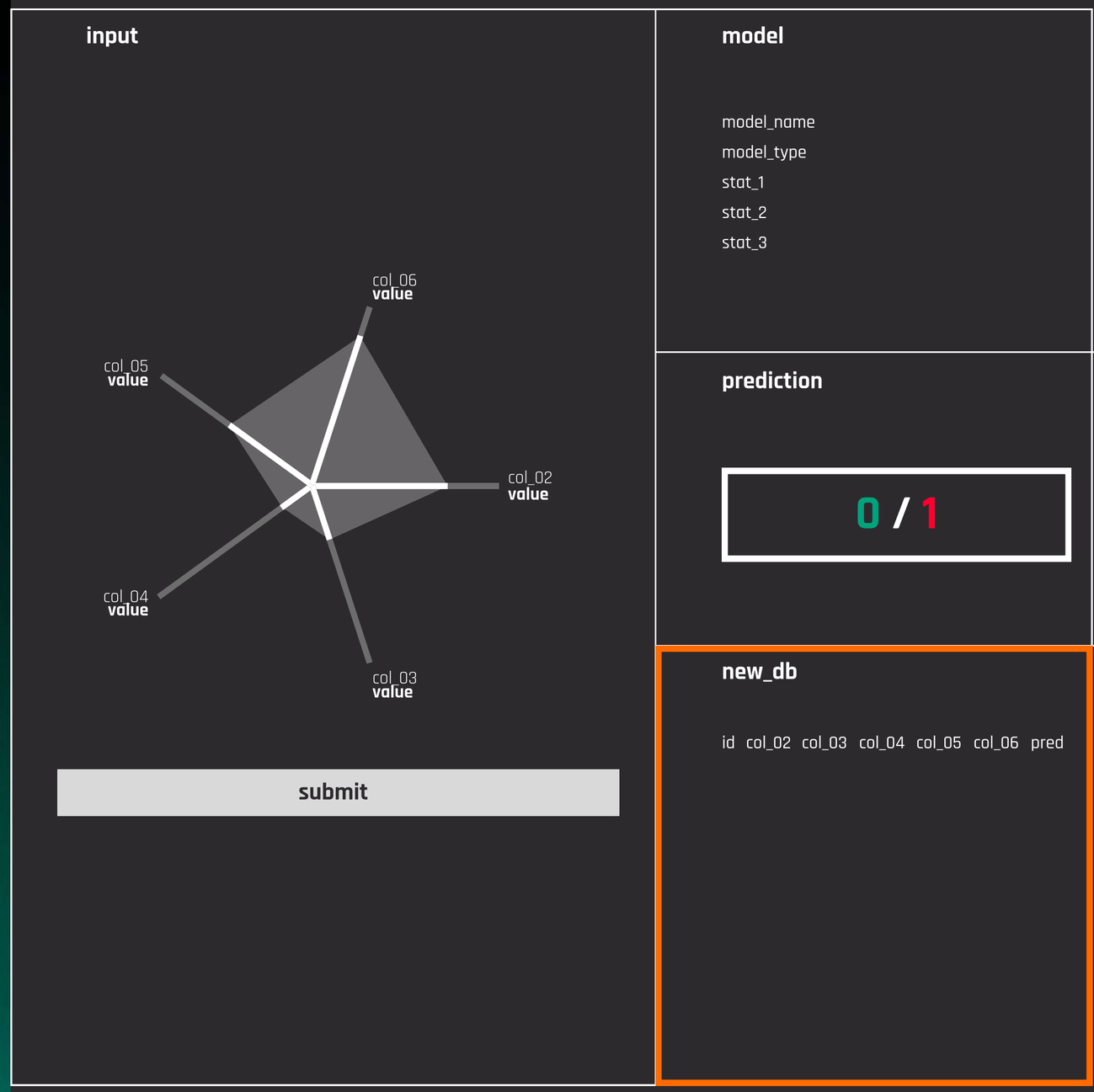
building a mock-up

it_depends

model.py | request.py | **app.py**

```python
    id_letter = string.ascii_uppercase[len(database) % 26]
    asteroid_id = f"2024 O{id_letter}"

    # Create a dictionary with the input data and the prediction
    asteroid = {
        'id': asteroid_id,
        'est_diameter_min': est_diameter_min,
        'est_diameter_max': est_diameter_max,
        'relative_velocity': relative_velocity,
        'miss_distance': miss_distance,
        'absolute_magnitude': absolute_magnitude,
        'hazardous': int(prediction)
    }

    # Add the dictionary to the database list at the first positi
    database.insert(0, asteroid)

    # Keep only the last 5 entries
    if len(database) > 5:
        database.pop()


    return render_template('index.html', prediction_text=formatte
@app.route('/results', methods=['POST'])
def results():
    data = request.get_json(force=True)
    prediction = model.predict([np.array(list(data.values()))])
    output = prediction[0]
    return jsonify(output)
if __name__ == "__main__":
    app.run(debug=True)
```

web application mock-up

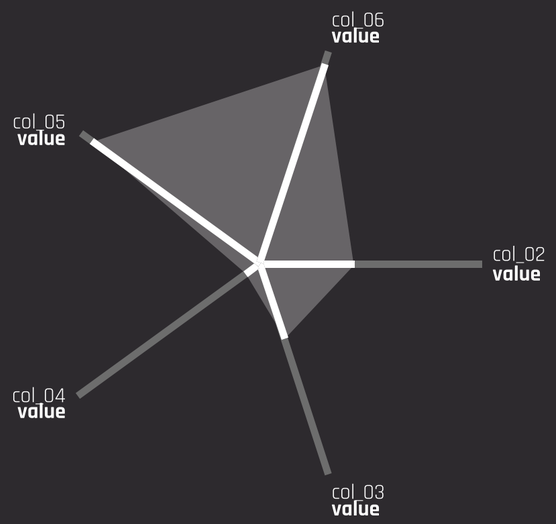


building a mock-up

it_depends

web application mock-up

input



submit

model

model_name
model_type
stat_1
stat_2
stat_3

prediction

0

new_db

id	col_02	col_03	col_04	col_05	col_06	pred
id	value	value	value	value	value	0
id	value	value	value	value	value	1

haas

monitoring the sky and detecting hazardous asteroids

by it_depends

asteroid

absolute_magnitude
31.68

miss_distance
43500000.00

37.60
est_diameter_min

12000.00
relative_velocity

82.80
est_diameter_max

PREDICT

model

model name: **A. J. Frost**
model type: **XGBClassifier**
precision: **0.30**
recall: **1.00**
f-1 score: **0.46**

prediction

new database

ID	est_diameter_min	est_diameter_max	relative_velocity	miss_distance	absolute_magnitude	hazardous

haas application

click the button below
to go to the github repository with the web application

REPO WITH WEB APPLICATION

bartlomiej brzostek

it_depends

Development of an application based on a Machine Learning model
to identify potentially dangerous asteroids.