

world_model_v7_Final

May 14, 2025

1 World Model

1.1 Authors

- Alberto Landi Cortiñas
- Ruben Gonzalez Braña
- Duarte Novas Álvarez

1.2 1.1 Imports

```
[ ]: import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
import xgboost as xgb
from sklearn.metrics import mean_absolute_error, mean_squared_error
from sklearn.preprocessing import StandardScaler
import joblib
import os
import random
from lazypredict.Supervised import LazyRegressor
import time
```

```
[ ]: # Set random seed for reproducibility
random.seed(42)
np.random.seed(42)
```

1.3 1.2 Load Dataset

```
[3]: def load_data(filepath="../dataset/dataset_v5.txt"):
    """Loads the dataset using pandas."""
    try:
        df = pd.read_csv(filepath)
        print(f"Dataset loaded successfully. Shape: {df.shape}")
        df = df.dropna()
        print(f"Shape after dropping NaNs: {df.shape}")
        return df
    except FileNotFoundError:
        print(f"Error: Dataset file not found at {filepath}")
```

```

        return None
    except Exception as e:
        print(f"Error loading dataset: {e}")
        return None

```

```

[4]: # 1. Load the dataset
dataframe = load_data()

```

Dataset loaded successfully. Shape: (3276, 14)
Shape after dropping NaNs: (3276, 14)

```

[ ]: dataframe.info()

```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 14 columns):
#   Column                Non-Null Count  Dtype
---  -
0   distance_red_init      3276 non-null   float64
1   angle_red_init         3276 non-null   float64
2   distance_green_init    3276 non-null   float64
3   angle_green_init       3276 non-null   float64
4   distance_blue_init     3276 non-null   float64
5   angle_blue_init        3276 non-null   float64
6   rSpeed                 3276 non-null   int64
7   lSpeed                 3276 non-null   int64
8   distance_red_final     3276 non-null   float64
9   angle_red_final        3276 non-null   float64
10  distance_green_final    3276 non-null   float64
11  angle_green_final       3276 non-null   float64
12  distance_blue_final     3276 non-null   float64
13  angle_blue_final       3276 non-null   float64
dtypes: float64(12), int64(2)
memory usage: 358.4 KB

```

1.4 1.3 Preprocess Dataset

```

[6]: def add_lagged_features(df, lag=1, columns_to_lag=None):
    """
    Adds lagged features to specific columns in the dataset.
    :param df: Original DataFrame
    :param lag: Number of lagged steps to include
    :param columns_to_lag: List of columns to apply lagging to
    :return: DataFrame with lagged features
    """
    if df is None:
        print("DataFrame is None, skipping lagged feature creation.")
        return None

```

```

if columns_to_lag is None:
    columns_to_lag = df.columns # Default to all columns if none specified

lagged_df = df.copy()
for i in range(1, lag + 1):
    lagged_columns = {col: f"{col}_lag{i}" for col in columns_to_lag}
    lagged_df = pd.concat([lagged_df, df[columns_to_lag].shift(i).
↪rename(columns=lagged_columns)], axis=1)

# Drop rows with NaN values introduced by lagging
lagged_df = lagged_df.dropna().reset_index(drop=True)
print(f"Lagged features added for columns {columns_to_lag} with lag={lag}. ↵
↪New shape: {lagged_df.shape}")
return lagged_df

```

```

[ ]: def prepare_data(df):
    """
    Separates features (X) and target variables (Y).
    :param df: DataFrame with lagged features
    :return: Features (X) and targets (Y)
    """
    if df is None:
        return None, None

    target_columns = [
        "distance_red_final", "angle_red_final",
        "distance_green_final", "angle_green_final",
        "distance_blue_final", "angle_blue_final"
    ]

    feature_columns = [col for col in df.columns if col not in target_columns]

    # Extract features and targets
    X = df[feature_columns].values
    Y = df[target_columns].values

    print(f"Features (X) shape: {X.shape}")
    print(f"Targets (Y) shape: {Y.shape}")
    return X, Y

```

```

[8]: def split_data(X, Y, test_size=0.2, random_state=42):
    """Splits data into training and testing sets."""
    if X is None or Y is None:
        return None, None, None, None
    X_train, X_test, Y_train, Y_test = train_test_split(
        X, Y, test_size=test_size, random_state=random_state
    )

```

```
)
print(f"Training set size: {X_train.shape[0]} samples")
print(f"Testing set size: {X_test.shape[0]} samples")
return X_train, X_test, Y_train, Y_test
```

```
[ ]: # 2. Prepare Data
columns_to_lag = [
    "distance_red_init", "angle_red_init",
    "distance_green_init", "angle_green_init",
    "distance_blue_init", "angle_blue_init",
    "rSpeed", "lSpeed"
]

dataframe = add_lagged_features(dataframe, lag=1,
    ↪ columns_to_lag=columns_to_lag) # Add lagged features
display(dataframe.info())
X, Y = prepare_data(dataframe)
```

Lagged features added for columns ['distance_red_init', 'angle_red_init', 'distance_green_init', 'angle_green_init', 'distance_blue_init', 'angle_blue_init', 'rSpeed', 'lSpeed'] with lag=1. New shape: (3275, 22)

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 3275 entries, 0 to 3274

Data columns (total 22 columns):

#	Column	Non-Null Count	Dtype
0	distance_red_init	3275 non-null	float64
1	angle_red_init	3275 non-null	float64
2	distance_green_init	3275 non-null	float64
3	angle_green_init	3275 non-null	float64
4	distance_blue_init	3275 non-null	float64
5	angle_blue_init	3275 non-null	float64
6	rSpeed	3275 non-null	int64
7	lSpeed	3275 non-null	int64
8	distance_red_final	3275 non-null	float64
9	angle_red_final	3275 non-null	float64
10	distance_green_final	3275 non-null	float64
11	angle_green_final	3275 non-null	float64
12	distance_blue_final	3275 non-null	float64
13	angle_blue_final	3275 non-null	float64
14	distance_red_init_lag1	3275 non-null	float64
15	angle_red_init_lag1	3275 non-null	float64
16	distance_green_init_lag1	3275 non-null	float64
17	angle_green_init_lag1	3275 non-null	float64
18	distance_blue_init_lag1	3275 non-null	float64
19	angle_blue_init_lag1	3275 non-null	float64
20	rSpeed_lag1	3275 non-null	float64
21	lSpeed_lag1	3275 non-null	float64

dtypes: float64(20), int64(2)
memory usage: 563.0 KB

None

Features (X) shape: (3275, 16)
Targets (Y) shape: (3275, 6)

```
[ ]: def scale_features(X_train, X_test):  
    """Scales input features using StandardScaler."""  
    if X_train is None or X_test is None:  
        return None, None, None  
    scaler = StandardScaler()  
  
    # Fit scaler only on training data  
    X_train_scaled = scaler.fit_transform(X_train)  
  
    # Transform both train and test data  
    X_test_scaled = scaler.transform(X_test)  
  
    print("Features scaled.")  
  
    return X_train_scaled, X_test_scaled, scaler # Return scaler to save it
```

```
[ ]: # 3. Split Data  
X_train, X_test, Y_train, Y_test = split_data(X, Y)  
  
# 4. Scale Features  
X_train_scaled, X_test_scaled, scaler = scale_features(X_train, X_test)
```

Training set size: 2620 samples
Testing set size: 655 samples
Features scaled.

1.5 1.4 Train model

1.5.1 Lazy Predict

```
[ ]: print("\n--- Running LazyPredict ---")  
  
# Initialize LazyRegressor for regression tasks  
reg = LazyRegressor(  
    verbose=0,  
    ignore_warnings=True,  
    custom_metric=None,  
    predictions=False,  
    random_state=42  
)
```

```

start_time_lazy = time.time()

# Fit and evaluate all models using the scaled data
try:
    models, predictions = reg.fit(X_train_scaled, X_test_scaled, Y_train,
    ↪Y_test)
    end_time_lazy = time.time()
    print(f"LazyPredict finished in {end_time_lazy - start_time_lazy:.2f}
    ↪seconds.")

    # Display the results
    print("\nLazyPredict Results (Sorted by RMSE):")
    print(models.sort_values(by='RMSE', ascending=True))

except Exception as e:
    print(f"\nAn error occurred during LazyPredict execution: {e}")
    print("Ensure input data shapes and types are correct.")

```

--- Running LazyPredict ---

0%| | 0/42 [00:00<?, ?it/s]

LazyPredict finished in 5.70 seconds.

LazyPredict Results (Sorted by RMSE):

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
XGBRegressor	0.80	0.80	50.39	1.02
LassoLars	0.58	0.59	74.45	0.01
Lasso	0.58	0.59	74.45	0.01
RidgeCV	0.58	0.59	74.49	0.01
Ridge	0.58	0.59	74.49	0.01
LinearRegression	0.58	0.59	74.50	0.01
TransformedTargetRegressor	0.58	0.59	74.50	0.00
Lars	0.58	0.59	74.50	0.01
OrthogonalMatchingPursuit	0.58	0.59	74.66	0.00
ExtraTreesRegressor	0.54	0.55	80.57	0.96
RandomForestRegressor	0.54	0.55	80.84	2.02
BaggingRegressor	0.49	0.50	86.03	0.23
ElasticNet	0.54	0.55	95.80	0.01
KNeighborsRegressor	0.58	0.59	107.86	0.05
DecisionTreeRegressor	0.16	0.18	114.02	0.04
ExtraTreeRegressor	0.09	0.11	121.60	0.01
MLPRegressor	0.40	0.42	147.56	0.68
RANSACRegressor	-1.04	-0.99	159.51	0.05
DummyRegressor	-0.03	-0.00	261.27	0.00
GaussianProcessRegressor	-0.92	-0.87	372.33	0.31

KernelRidge

-3.35

-3.25 561.39

0.16

1.5.2 XGBoost Regressor

```
[ ]: def train_xgboost_regression(X_train, Y_train):  
    """Trains an XGBoost Regressor model with hyperparameter optimization using  
    ↪GridSearchCV."""  
    if X_train is None or Y_train is None:  
        print("Skipping XGBoost training as data is None.")  
        return None  
  
    print("Training XGBoost Regressor model with GridSearchCV...")  
  
    param_grid_xgb = {  
        'n_estimators': [100, 200, 300],  
        'learning_rate': [0.01, 0.05, 0.1],  
        'max_depth': [3, 5, 7],  
        'subsample': [0.7, 0.8, 1.0],  
        'colsample_bytree': [0.7, 0.8, 1.0],  
        'gamma': [0, 0.1],  
        # 'reg_alpha': [0, 0.01, 0.1],  
        'reg_lambda': [0.1, 1, 10]  
    }  
  
    # Create the XGBoost regressor model  
    xgb_model = xgb.XGBRegressor(objective='reg:squarederror', random_state=42,  
    ↪n_jobs=-1)  
  
    # GridSearchCV  
    grid_search_xgb = GridSearchCV(  
        estimator=xgb_model,  
        param_grid=param_grid_xgb,  
        scoring='neg_mean_squared_error',  
        cv=5,  
        n_jobs=-1,  
        verbose=2  
    )  
  
    # Fit GridSearchCV on the training data  
    grid_search_xgb.fit(X_train, Y_train)  
  
    # Print the best parameters and corresponding score  
    print("\nGridSearchCV Complete for XGBoost.")  
    print(f"Best parameters found: {grid_search_xgb.best_params_}")  
    print(f"Best cross-validation score (negative MSE): {grid_search_xgb.  
    ↪best_score_:.4f}")
```

```
# Return the best model found by GridSearchCV
return grid_search_xgb.best_estimator_
```

```
[ ]: # Train the XGBoost model
print("\n--- Training XGBoost Model ---")
world_model_xgb = train_xgboost_regression(X_train_scaled, Y_train)
```

```
--- Training XGBoost Model ---
Training XGBoost Regressor model with GridSearchCV...
Fitting 5 folds for each of 1458 candidates, totalling 7290 fits
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=0.7; total time= 0.2s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=0.8; total time= 0.2s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=0.7; total time= 0.3s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=0.7; total time= 0.3s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=0.7; total time= 0.2s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=0.7; total time= 0.3s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=0.8; total time= 0.3s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=0.8; total time= 0.3s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=0.8; total time= 0.2s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=1.0; total time= 0.3s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=1.0; total time= 0.2s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=1.0; total time= 0.3s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=1.0; total time= 0.3s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=0.1, subsample=1.0; total time= 0.2s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=1, subsample=0.7; total time= 0.2s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=1, subsample=0.7; total time= 0.3s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
n_estimators=100, reg_lambda=1, subsample=0.7; total time= 0.3s
[CV] END colsample_bytree=0.7, gamma=0, learning_rate=0.01, max_depth=3,
```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[CV] END colsample_bytree=1.0, gamma=0.1, learning_rate=0.1, max_depth=7,  
n_estimators=300, reg_lambda=10, subsample=1.0; total time= 7.2s
```

GridSearchCV Complete for XGBoost.

Best parameters found: {'colsample_bytree': 1.0, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 300, 'reg_lambda': 10, 'subsample': 0.8}
Best cross-validation score (negative MSE): -2314.2500

1.6 1.5 Evaluate

```
[ ]: def evaluate_model(model, X_test, Y_test, model_name="Model"):  
    """Evaluates the model using MAE, MSE, and RMSE."""  
    if model is None or X_test is None or Y_test is None:  
        print(f"Skipping evaluation for {model_name} as model or data is None.")  
        return None, None  
  
    Y_pred = model.predict(X_test)  
  
    mae = mean_absolute_error(Y_test, Y_pred)  
    mse = mean_squared_error(Y_test, Y_pred)  
    rmse = np.sqrt(mse) # Root Mean Squared Error  
  
    print(f"\n--- {model_name} Evaluation ---")  
    print(f"Mean Absolute Error (MAE): {mae:.4f}")  
    print(f"Mean Squared Error (MSE): {mse:.4f}")  
    print(f"Root Mean Squared Error (RMSE): {rmse:.4f}")  
  
    print("\nMAE per output feature:")  
    output_features = [  
        'dist_red_final', 'angle_red_final', 'dist_green_final',  
        'angle_green_final', 'dist_blue_final', 'angle_blue_final'  
    ]  
    for i, name in enumerate(output_features):  
        # Ensure Y_test and Y_pred are 2D  
        y_test_col = Y_test[:, i] if Y_test.ndim > 1 else Y_test  
        y_pred_col = Y_pred[:, i] if Y_pred.ndim > 1 else Y_pred  
  
        mae_feature = mean_absolute_error(y_test_col, y_pred_col)  
        print(f" {name}: {mae_feature:.4f}")  
  
    return mae, mse  
  
[ ]: # Evaluate the XGBoost model  
print("\n--- Evaluating XGBoost Model ---")  
xgb_mae, xgb_mse = evaluate_model(world_model_xgb, X_test_scaled, Y_test,   
    ↪ model_name="XGBoost")  
if xgb_mae is not None:
```



```
print(f"XGBoost Test RMSE: {np.sqrt(xgb_mse):.4f}")
```

--- Evaluating XGBoost Model ---

--- XGBoost Evaluation ---

Mean Absolute Error (MAE): 26.5363

Mean Squared Error (MSE): 2226.2277

Root Mean Squared Error (RMSE): 47.1829

MAE per output feature:

dist_red_final: 14.8164

angle_red_final: 33.4181

dist_green_final: 15.9752

angle_green_final: 40.8843

dist_blue_final: 16.0813

angle_blue_final: 38.0424

XGBoost Test RMSE: 47.1829

1.7 1.6 Example Prediction

```
[ ]: def example_prediction(model, X_test_original, Y_test_original, scaler_obj,
    ↪model_name="Model"):
    if model is None or X_test_original is None or Y_test_original is None or
    ↪scaler_obj is None:
        print(f"Skipping example prediction for {model_name} as essential
    ↪components are missing.")
        return

    print(f"\n--- Example Prediction ({model_name}) ---")

    # Take the first sample from the original unscaled test set
    sample_X_orig = X_test_original[0].reshape(1, -1)
    sample_Y_actual = Y_test_original[0]

    # Scale the sample using the scaler
    sample_X_scaled = scaler_obj.transform(sample_X_orig)

    # Predict using the trained model
    sample_Y_pred = model.predict(sample_X_scaled)

    print(f"Input State + Action (Original): {sample_X_orig[0]}")
    print(f"Input State + Action (Scaled): {sample_X_scaled[0]}")
    print(f"Actual Final State: {sample_Y_actual}")
    print(f"Predicted Final State: {sample_Y_pred[0]}")
```

```
[19]: # Example prediction with XGBoost model
if world_model_xgb and X_test is not None and Y_test is not None and scaler is_
↳not None:
    example_prediction(world_model_xgb, X_test, Y_test, scaler,
↳model_name="XGBoost")
else:
    print("Skipping XGBoost example prediction due to missing components.")
```

```
--- Example Prediction (XGBoost) ---
Input State + Action (Original): [ 2.40447212e+02  2.24474152e+01
1.22459073e+03  8.72418891e+01
    1.08392337e+03  3.70584400e+01 -1.20000000e+01 -1.60000000e+01
    2.53098419e+02  1.39212699e+01  1.22599230e+03  7.91048203e+01
    1.09502619e+03  2.91509349e+01 -1.00000000e+00  5.00000000e+00]
Input State + Action (Scaled):  [-1.4440972 -1.55527371  1.36453177
-0.87776177  0.80226429 -1.41680449
    -0.67912385 -0.9435609 -1.40695855 -1.64067227  1.36368323 -0.93364881
    0.83231492 -1.47058995 -0.04940402  0.28918407]
Actual Final State:             [ 323.71655664    9.16729444 1237.80996378
74.82542519 1157.44558751
    26.28745497]
Predicted Final State:          [ 284.73975    246.66615  1232.974
62.054188 1135.8422    69.61528 ]
```

1.8 1.7 Save model

```
[19]: def save_model_and_scaler(model, scaler, model_filename="world_model_v5_XGB.
↳joblib", scaler_filename="scaler_v5_XGB.joblib"):
    """Saves the trained model and scaler to disk."""
    if model is None or scaler is None:
        print("Skipping saving model/scaler as one of them is None.")
        return

    try:
        model_dir = "../content/" # Relative to script location
        os.makedirs(model_dir, exist_ok=True)

        model_path = os.path.join(model_dir, model_filename)
        scaler_path = os.path.join(model_dir, scaler_filename)

        joblib.dump(model, model_path)
        print(f"Model saved to {model_path}")
        joblib.dump(scaler, scaler_path)
        print(f"Scaler saved to {scaler_path}")
    except Exception as e:
        print(f"Error saving model/scaler: {e}")
```

```
[ ]: # Save the XGBoost model and scaler
print("\n--- Saving XGBoost Model and Scaler ---")
save_model_and_scaler(world_model_xgb, scaler,
                       model_filename="xgb_world_model_v7.joblib",
                       scaler_filename="xgb_scaler_v7.joblib")
```

--- Saving XGBoost Model and Scaler ---

Model saved to ../content/xgb_world_model_v7.joblib

Scaler saved to ../content/xgb_scaler_v7.joblib