Predicting Missile and Drone Destruction:

A Data-Driven Approach Using Machine Learning

Bohdan Hukovych

Hanbee Yoo

Lucia Žídeková

Abstract

This project explores the predictive power of machine learning models in forecasting the outcomes of missile and drone interceptions using a comprehensive dataset detailing missile attacks in Ukraine, combined with meteorological data. The study utilizes various machine learning techniques, including Elastic Net, Random Forest, Gradient Boosting, with a particular focus on ensemble methods to enhance predictive accuracy and robustness. Key findings indicate that while weather conditions have a minimal impact on interception success, the volume of missiles launched is a significant predictor, suggesting that higher launch numbers increase interception probabilities. The research underscores the effectiveness of ensemble models, which outperform individual predictive models in accuracy and stability. This project not only contributes to the theoretical advancements in applying machine learning to military strategy but also provides practical insights that can aid defence planners and strategists in understanding and predicting missile defence outcomes.

Introduction

This project aims to develop a predictive model that estimates how many missiles and drones will be destroyed based on historical attack patterns, weapon characteristics, and weather conditions. By leveraging machine learning algorithms and time-series analysis, we seek to provide insights into the key drivers of successful destructions.

Recent research highlights the growing role of machine learning and environmental factors in missile and drone interception. Studies on reinforcement learning for UAV interception demonstrate how AI-driven algorithms can optimize response times and improve interception accuracy in dynamic aerial environments (Cai et al., 2024). Similarly, deep reinforcement learning has been applied to missile guidance systems, allowing for improved trajectory optimization against manoeuvring targets (He et al., 2023). Moreover, the effect of weather conditions, with factors like rain, fog, and strong winds, which might in theory significantly affect the probability of destruction remains rather unexplored in academic literature.

To develop the predictive model, "Massive Missile Attacks on Ukraine" dataset, available on Kaggle, was utilized. This dataset provides a detailed historical record of missile and drone attacks during the ongoing conflict in Ukraine, covering key variables such as the number of missiles and drones launched, interception success rates, weapon types, and launch locations. In addition, it includes information allowing for the analysis of temporal trends, attack intensities, and defence effectiveness over time.

To enhance predictive accuracy, this dataset is supplemented with weather data, capturing key meteorological factors (e.g., temperature, wind speed, precipitation) that may influence missile interception rates. By integrating these datasets, the project aims to quantify the impact of different attack patterns, weapon systems, and environmental conditions on the probability of successful missile and drone destruction.

The dataset is first cleaned and pre-processed, addressing missing values, one-hot encoding categorical variables (e.g., missile types and launch locations), and creating temporal features such as seasonal trends or moving averages. Weather data is merged with the attack dataset to assess the potential influence of meteorological conditions on interception success.

Following preprocessing, exploratory data analysis is conducted to identify patterns, correlations, and potential biases in the data. The modelling phase begins with a baseline prediction using the average historical destruction rate. First Ordinary Least Squares regression is implemented for interpretability. Then a more advanced models, like Elastic Net and Random Forest, is utilized to capture non-linear relationships. Next XGBoost – a gradient boosting algorithm – is used to optimize and improve predictive accuracy. Finally, Multilayer Perceptron, kNN and SVM Regression are trained to compare with previous models.

The project is structured as follows. The first chapter discusses the two datasets used and the major transformation implemented. The second chapter provides an overview of the methodological approaches that were employed. The third chapter discusses and evaluates the models´ performance. Finally, the last chapter provides a general overview of the work.

Data Description

The primary dataset, **"Massive Missile Attacks on Ukraine"**, publicly available on Kaggle, contains records of missile and drone attacks from 2022 to 2023, capturing key characteristics such as weapon type, launch and target locations, and interception success. To enhance predictive accuracy, this dataset is merged with meteorological data from Meteostat website, enabling an assessment of how weather conditions—such as temperature, wind speed, and precipitation—may influence interception success. Given the potential impact of adverse weather on missile trajectory, drone stability, and detection efficiency, weather variables are expected to contribute to the explanatory power of the model. This data is taken as daily weather in Kyiv, which is assumed to be an approximation of the weather in the locations the targets were flying through. We acknowledge, that this may give too little information for our task since the weather can be significantly different in different locations of Ukraine and change throughout the day.

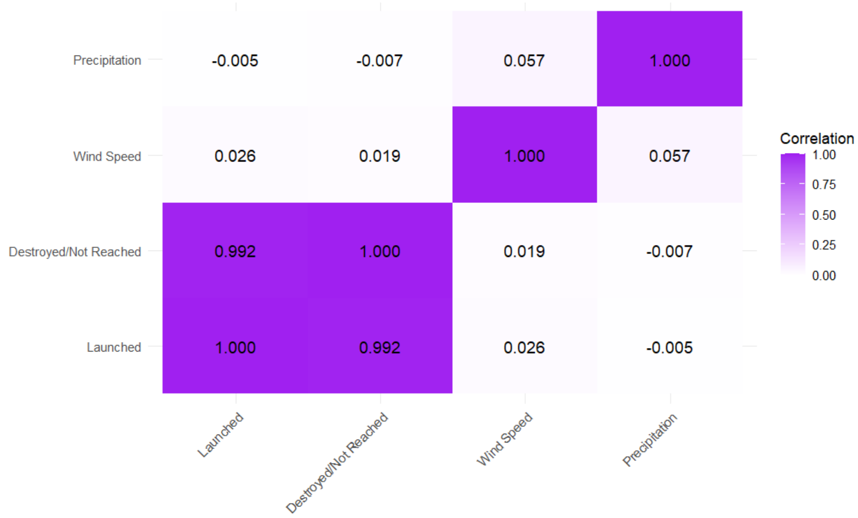
Before modelling, extensive data cleaning and preprocessing steps were undertaken. Missing values were converted to NA to ensure consistency, and redundant or irrelevant variables such as destroyed_details, launched_details, or launch_place_details were removed to reduce dimensionality. Categorical variables, including carrier, model, and launch_place, were transformed using one-hot encoding to enable processing by machine learning algorithms.

Temporal features were extracted from time_start, classifying each observation by weekday and season to account for potential seasonal effects on missile performance and air defence. Additionally, moving averages of destruction rates were computed over seven-day, thirty-day, and one-hundred-day windows to smooth fluctuations and highlight broader trends in interception effectiveness.

A correlation analysis revealed a **strong positive correlation (0.992)** between *launched* (number of launched rockets) and *destroyed_not_reached* (number of destroyed targets and those that did not reach its goal), indicating that the number of missiles launched is a primary determinant of interception outcomes. However, weather variables exhibited **weak correlations** with both *launched* and *destroyed_not_reached*, with wind speed (*wspd*) showing only **0.026** correlation with *launched* and **0.019** with *destroyed_not_reached*, while precipitation (*prcp*) had negligible correlations of **-0.005** and **-0.007**, respectively. This suggests that weather conditions may have a limited direct effect on missile interception success within the dataset.

The final dataset, structured with key temporal, geographical, and weather-related predictors, ensures that the model captures both attack patterns and environmental influences while minimizing redundancy. This structured approach enhances the dataset's suitability for predictive modelling and statistical analysis.

Figure 1: Correlation between the main variables:



Methodology

We employed a variety of machine learning and econometric techniques to develop and assess the best-performing predictive models. Key methodologies include Ordinary Least Squares (OLS) Regression, Elastic Net Regression, Random Forest, Gradient Boosting, Multilayer Perceptron (MLP), kNN, and SVM Regression. The models were tuned and evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) as benchmarks to identify the most effective approach. The following sections provide a brief introduction to each methodology, grounded in relevant literature

For the **OLS regression model**, we utilized all available predictors to establish relationships with the target variable. OLS provides a useful baseline for performance comparison against more complex models, thanks to its simplicity and interpretability.

Elastic Net effectively addresses multicollinearity by combining the L1 and L2 regularization of Lasso and Ridge regression techniques by shrinking some coefficients toward zero (like Lasso) and others toward each other (like Ridge). The tuning of hyperparameters such as the mix ratio (alpha), lambda, and polynomial transformations of features was guided by cross-validation using RMSE. Zou and Hastie (2005) in their seminal paper provide comprehensive insights into the advantages of Elastic Net in handling various data anomalies.

The **Random Forest,** introduced in a seminal paper by Breiman (2001), is a technique that employs multiple decision trees to mitigate the overfitting risk associated with single decision trees and enhances generalization across diverse datasets. Parameters such as the number of trees, the number of features per split, and the minimum size of leaf nodes are crucial for optimizing performance.

Gradient Boosting constructs an ensemble of models sequentially, with each new model correcting errors from the previous ones. We specifically employed **XGBoost**, renowned for its execution speed and model performance, and adjusted parameters like the number of trees, maximum depth, regularization, number of features per split, share of data subsampling, weighting of child trees and learning rate. The improvements in prediction accuracy are supported by the detailed examination in Chen and Guestrin's (2016) paper on XGBoost.

The **MLP** is a simple neural network with one hidden layer. The tuning of parameters, including the number of neurons in the hidden layer and learning rate decay, was aimed at capturing the non-linear relationships in the data. Goodfellow, Bengio, and Courville (2016) provide an extensive discussion on the capabilities and applications of MLPs in their book on deep learning. In addition, we tested kNN and SVM Regression algorithms also tuning their hyperparameters.

The data was randomly split into training set (80% of the data, 1552 observations) and the test set (20% of the data, 388 observations). The models' performance was rigorously tested on the independent test set, focusing on minimizing RMSE and MAE while maximizing R^2. The final evaluation involved an ensemble technique using weighted average by RMSE of predictive outputs of the individual models to leverage their collective strength and reduce prediction variance. It is based on the idea that models with independent errors will cancel those error with each other as more models are used.

Finally, we perform permutation tests to check if some models statistically outperform others. This test is based on the idea that under the null hypothesis, both models have equal performance, which means that it is not important from which model prediction outcome is taken. Using this idea we sample observations with equal probability from both models and subsequently calculate the evaluation metric (RMSE). This procedure is repeated many times to obtain a distribution of evaluation metrics under the null. The approximate p-value that the better model is significantly better is calculated as a percentage of distribution's RMSEs which are higher than the model's RMSE.

Results

The goal of this project was to investigate the impact of attack patterns, weather conditions, and temporal factors on the probability of missile and drone destruction. The chapter highlights the procedure and results from the models implemented.

Model Performance

First, the performance of the models was tested using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). The baseline models, which used the average destruction rate from the training data, had RMSE of 4.93, and MAE of 2.27, and the Ordinary Least Squares (OLS) regression with only *launched* variable provided an improvement with RMSE of 2.37, and MAE of 1.51.

OLS which included all variables performed slightly better than the baseline OLS with RMSE of 2.1, and MAE of 1.22. Elastic Net Regression outperformed other models with RMSE of 1.967 and MAE of 1.15, demonstrating a significant improved in predictive accuracy. Random Forest did slightly worse in terms of RMSEs of 2.07 but much better in terms of MAE of 0.989. Boosted tree performed the best among individual models in terms of RMSE of 1.914 and decent MAE of 1.06. The Multilayer Perceptron (MLP) model underperformed slightly, with an RMSE of 2.18, and MAE 1.22. SVM Regression performed comparably with RMSE of 2.08, and MAE of 1.25. It is likely that either tuning of parameters was not very successful for MLP and SVM or their package functions were not the most effective. The K-Nearest Neighbors (KNN) model had the weakest performance, with an RMSE of 5.15 and an R^2 of only 0.93, indicating that it struggled to generalize patterns from training data compared to other models.

The most effective models were ensemble models that combined predictions from several different methods; a weighted ensemble of all non-baseline models except kNN had the lowest RMSE of 1.85 and very decent MAE of 1.01. Unweighted ensemble performed comparably. From the table, it is evident that ensemble models provided the best predictive performance, reducing RMSE and improving R^2 compared to individual models.

Table 1: Model Performance

Model	RMSE	MAE	R^2
Baseline: Avg. Dest. Rate	4.93	2.3	0.938
Baseline: OLS (*launched*)	2.3	1.5	0.985
OLS (all variables)	2.1	1.22	0.989
Elastic Net Regression	1.97	1.15	0.989
Random Forest	2.07	0.989	0.990
XGBoost	1.914	1.06	0.990
Multilayer Perceptron	2.18	1.22	0.987
k-Nearest Neighbors	5.15	3.02	0.93
Unweighted Ensemble	1.847	1.02	0.991
Weighted Ensemble	1.846	1.01	0.991
Ensemble (Elastic net, Forest and XGBoost)	1.856	1.008	0.991

Table 1 presents the overall RMSE, MAE and R^2 values for each model, highlighting that ensemble models outperformed individual models with lower RMSE, MAE and higher R^2.

The results of permutation test are in Table 2 (only models with higher RMSE than OLS with all variables are tested). We see that most tested models are significantly outperforming OLS with all variables except Random Forest and SVM Regression. Moreover, if we compare best models with each other, we see that we cannot claim that XGBoost is actually better than Elastic Net. If we compare those 2 models with Unweighted Ensemble we see that it is only significantly better than Elastic Net, while not significantly better than XGBoost.

Table 2: Permutation Tests

Model comparison	p-value
Elastic net vs. OLS with all variables	*0.001*
Random Forest vs. OLS with all variables	0.441
XGBoost vs. OLS with all variables	*0.015*
SVM vs. OLS with all variables	0.480
Unweighted Ensemble vs. OLS with all variables	*0*
XGBoost vs. Elastic net	0.245
Unweighted Ensemble vs. Elastic net	*0.002*
Unweighted Ensemble vs. XGBoost	0.141

Interpretation of Key Predictors

The most significant predictor turned out to be the number of missiles launched, which showed a strong correlation of 0.992 with the number of missiles destroyed or failing to reach their target. This suggests that the higher the volume of launches, the greater the number of interceptions. Missile type also played a significant role. Ballistic missiles showed a lower destruction rate compared to cruise missiles and drones, likely due to their speed and evasion techniques. The missiles from anti-aircraft missile systems (e.g., C-300, C-400) also had a notable effect on destruction probabilities.

Moreover, the method of missile deployment, especially the launch carrier type also significantly impacts interception success. Missiles launched from jet-based and missile-system-based platforms are associated with higher destruction rates, while naval-launched missiles have lower interception probabilities.

Geographical factors further influence missile interception. Intuitively, the missiles launched from the East and South have slightly higher destruction rates than those from the North. The variable indicating launches deep inside Russia negatively impacted destruction rates, suggesting that missiles from long-range launch points may be harder to intercept. However, weather conditions, including wind speed, precipitation, and air pressure, showed weak correlations with interception success. This suggests that it is likely that modern missile guidance systems and technologies are relatively not affected by moderate weather conditions.

Finally, temporal trends, particularly the 30-day moving average of destruction rates, demonstrated the highest predictive power. This indicates that medium-term trends offer valuable insights for forecasting missile interception effectiveness.

Validation and Robustness

To ensure robustness, models were validated using an independent test set. By using an ensemble approach, which combined the predictions from the best-performing models and excluded those that underperformed, we were able to reduce variance and enhance overall accuracy. The feature importance analysis showed consistent results across different models, further supporting the robustness of our findings.

Conclusion

This study aimed to evaluate the capability of machine learning models to predict missile and drone interception outcomes based on extensive historical data and variable weather conditions. Through a rigorous comparison of different modelling techniques, our findings reveal that ensemble models, which combine multiple predictive approaches, provide the highest accuracy and robustness, aligning with current advances in machine learning research that advocate for the power of collective prediction methodologies.

Contrary to our initial hypothesis, weather conditions played a minimal role in influencing interception success rates. This outcome indicates that contemporary missile and drone technologies are likely to withstand varied environmental factors. Importantly, the quantity of missiles launched proved to be the most impactful predictor, underscoring the principle that a higher number of launches increases the probability of interception successes, which could be due to the saturation of defensive systems.

Despite achieving high model accuracy, this study acknowledges limitations such as the exclusion of potential influential factors like evolving defence technologies and strategic modifications which may affect interception dynamics. These limitations suggest space for further research, possibly incorporating data on specific missile defence systems and their technological advancements to refine predictive accuracy.

The practical implications of our research extend to military strategy and defence planning, offering a data-driven basis for anticipating outcomes of missile defence engagements, which can support more strategic allocation of defence resources in real-world conflict scenarios.

Ultimately, this project not only advances our understanding of key determinants in missile and drone defence efficacy but also exemplifies the application of advanced machine learning techniques in strategic military contexts, thereby contributing valuable insights to the field of defence analytics.

References

Breiman, L. Random Forests. *Machine Learning* **2001**, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cai, H., Li, X., Zhang, Y., Gao, H. Interception of a Single Intruding Unmanned Aerial Vehicle by Multiple Missiles Using the Novel EA-MADDPG Training Algorithm. *Drones* **2024**, *8*, 524. https://doi.org/10.3390/drones8100524

Chen, T., Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* **2016**. ACM, New York, NY, USA, 785-794. https://doi.org/10.1145/2939672.2939785

Goodfellow, I., Bengio, Y., Courville, A. Deep Learning. *MIT Press* **2016**.

He, S., Shin, H.-S., Tsourdos, A. Computational Missile Guidance: A Deep Reinforcement Learning Approach. *J. Aerosp. Inf. Syst.* **2021**, *18*, 571–582. https://doi.org/10.2514/1.I010970

Zou, H., Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2005**, 67(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x