Predicting Shark Attacks: A Data Mining Analysis for Prevention

Date: April 2025

Course: ISyE 7406 - Data Mining & Statistical Learning

Team # 90

Abstract

This project investigates shark attack incidents to determine the factors that influence whether an attack is fatal. We utilized a real-world dataset containing information on thousands of shark attacks globally. Our goal was to predict fatality outcomes using machine learning models, perform clustering to identify risk profiles, and apply statistical testing to validate insights. The analysis included extensive preprocessing, exploratory data analysis (EDA), feature engineering, classification modeling (logistic regression, decision trees, random forests, and XGBoost), clustering using KMeans, and hypothesis testing (ANOVA and chi-square). Through model evaluation and optimization techniques such as SMOTE and hyperparameter tuning, we achieved insights into the contributing factors of fatal shark attacks, identifying age, species, activity, and geography as important variables.

Introduction

Shark attacks are becoming more frequent and serious incidents that can lead to fatal consequences. Understanding the patterns and risk factors associated with such attacks has both scientific and public safety implications. With the rise in availability of structured shark attack datasets, data mining techniques can uncover patterns, trends, and predictive insights.

The primary challenge lies in the dataset's imbalance — fatal attacks are far fewer than non-fatal ones — and in handling missing and inconsistent data. Additionally, factors influencing attack outcomes vary across dimensions such as time, location, and human activity. Our approach involved a step-by-step strategy to clean, explore, model, and test the data rigorously.

This report presents a summary of the key steps and findings of the project.

Problem Statement & Data Sources

Our objective was to predict the fatality outcome of a shark attack using available data and uncover meaningful insights regarding the nature and frequency of attacks.

The dataset was sourced from https://www.kaggle.com/datasets/teajay/global-shark-attacks, which compiles global shark attack reports from the Global Shark Attack File. The dataset includes variables such as date, country, activity, species, victim age, sex, and fatality status.

Key challenges with the data included:

- **Missing values**: Many fields had missing or inconsistent entries, especially in free-text columns.
- **Unstructured categorical features**: Fields like species and activity required cleaning and standardization.
- **Time-based inconsistencies**: Some years recorded (e.g., year 250) were unrealistic.

Summary Statistics:

- ~6,000 total records (after cleaning: ~3,800 usable rows)
- Most attacks occurred in the U.S., Australia, and South Africa
- Majority of victims were male, aged between 20–40
- Common species: White shark, Tiger shark, Bull shark

We cleaned and transformed the dataset using Python, converting dates, standardizing text, engineering new variables (e.g., season, weekend, time of day), and handling class imbalance.

Proposed Methodology

Our data mining strategy involved multiple stages:

1. Data Preprocessing & Cleaning

- Removed or corrected inconsistent year values.
- Standardized text fields (e.g., upper/lowercase conversion, strip whitespace).
- Encoded categorical variables.
- Filled missing values using median (numerical) or mode (categorical) where appropriate.

- Created new variables like Is Summer, Time Category, and Risk Level to enrich the dataset.

2. Exploratory Data Analysis (EDA)

- Visualized distributions by year, season, gender, activity, country, and shark species.
- Identified dominant patterns in fatal and non-fatal attacks.
- Revealed country-specific risks and behavior-based patterns.

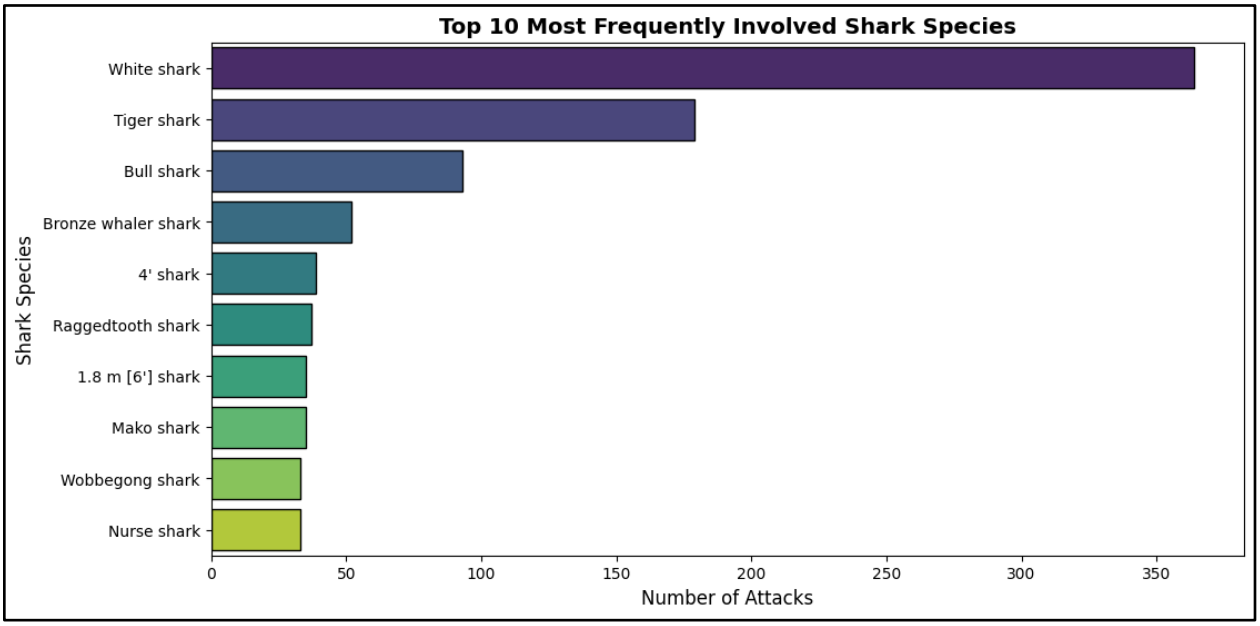


Fig. 1 Top Most Frequently Involved Shark Species

3. Classification Modeling

We framed the problem as a binary classification task (Fatal = Y/N). The models evaluated include:

- **Logistic Regression**: Baseline linear model.
- **Decision Tree**: Tree-based model to identify interpretable splits.
- **Random Forest**: Ensemble of trees for better generalization.
- **XGBoost**: Boosted tree-based model for higher accuracy.

We evaluated model performance using precision, recall, F1-score, and accuracy — with a special focus on the fatal class (minority).

4. Handling Class Imbalance

- **SMOTE**: Synthetic oversampling to balance minority class.
- **Undersampling**: Reduced majority class size.
- **Hybrid**: Combined both approaches.

5. Clustering Analysis

We used KMeans clustering on scaled numerical data to find natural groupings in the attacks. This allowed us to:

- Group attacks are based on characteristics like age, year, and timing.
- Analyze the profile of each cluster.

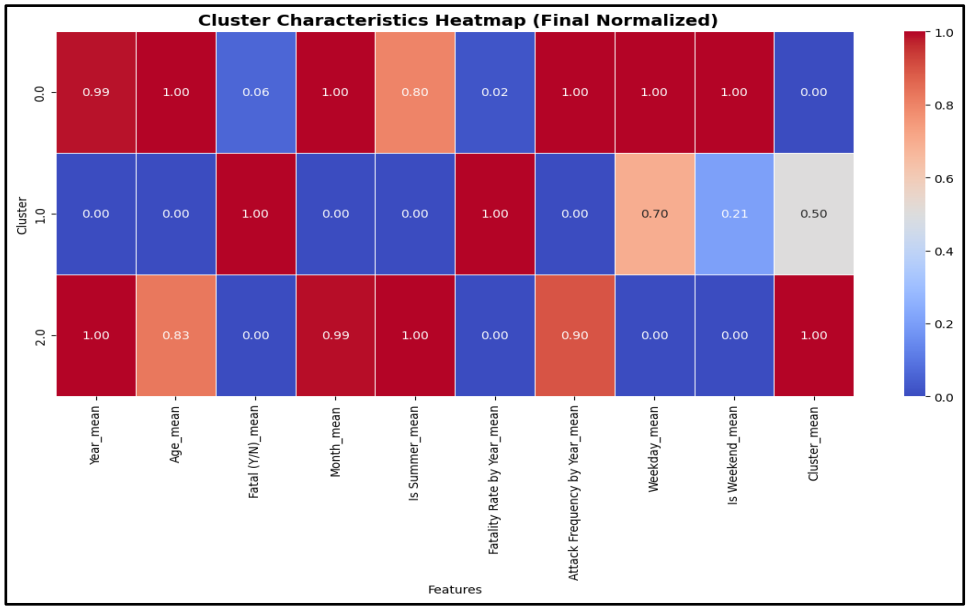


Fig. 2 Cluster Characteristic Heatmap

6. Hypothesis Testing

- **ANOVA**: To check if age and fatality rates differ significantly across clusters.
- **Chi-Square**: To examine associations between clusters and categorical variables (e.g., weekend, season).

7. Feature Engineering & Selection

- Engineered features like weekday/weekend, summer season, activity type.
- Ranked feature importance using Random Forest and XGBoost.
- Removed less relevant features (e.g., Is Summer) to improve performance.

8. Model Optimization

- Tuned hyperparameters using GridSearchCV for both Random Forest and XGBoost.
- Compared optimized models against baselines.

Our methodology was iterative and justified by ongoing performance improvements and interpretability at each step. We combined classification (for prediction), clustering (for segmentation), and testing (for validation), ensuring robustness from multiple perspectives.

Analysis & Results

Model Evaluation

We evaluated multiple models and sampling strategies. Key findings include:

- **Baseline models**: XGBoost and Random Forest performed best, with accuracy around 82%.
- **SMOTE/Undersampling**: Improved recall for fatal attacks but slightly reduced accuracy.
- **Feature Selection**: Removing low-impact variables improved precision without hurting accuracy.
- **Hyperparameter Tuning**: Boosted performance slightly, with optimized Random Forest achieving the highest accuracy.

	Model	Precision (Fatal)	Recall (Fatal)	F1-Score (Fatal)	Accuracy
0	Logistic Regression (Baseline)	0.545455	0.167832	0.256684	0.819481
1	Decision Tree (Baseline)	0.384615	0.419580	0.401338	0.767532
2	Random Forest (Baseline)	0.547945	0.279720	0.370370	0.823377
3	XGBoost (Baseline)	0.486486	0.377622	0.425197	0.810390
4	Random Forest (Feature Selection)	0.566265	0.328671	0.415929	0.828571
5	XGBoost (Feature Selection)	0.533981	0.384615	0.447154	0.823377
6	Random Forest (Optimized)	0.602740	0.307692	0.407407	0.833766
7	XGBoost (Optimized)	0.500000	0.370629	0.425703	0.814286

Clustering Results

- **3 Clusters Identified**: Clustered by time, age, and fatality rates.
- **Cluster 1**: Older attacks with higher fatality.
- **Cluster 2**: Recent, lower-fatality attacks involving younger victims.
- **Cluster 3**: Similar time as Cluster 2 but even lower fatality rate.

Statistical Tests

- **ANOVA** confirmed significant differences in age and fatality rates across clusters.
- **Chi-Square** tests found no strong correlation between cluster membership and season or weekend attacks.

These findings suggest changes in shark attack behavior and reporting across time.

Test Description	Test Statistic	Value	p-value
ANOVA Test for Age	F	58.839	0
ANOVA Test for Fatality Rate	F	119.019	0
Chi-Square Test for Seasonality	Chi^2	2.831	0.243
Chi-Square Test for Weekend Attacks	Chi^2	0.505	0.777

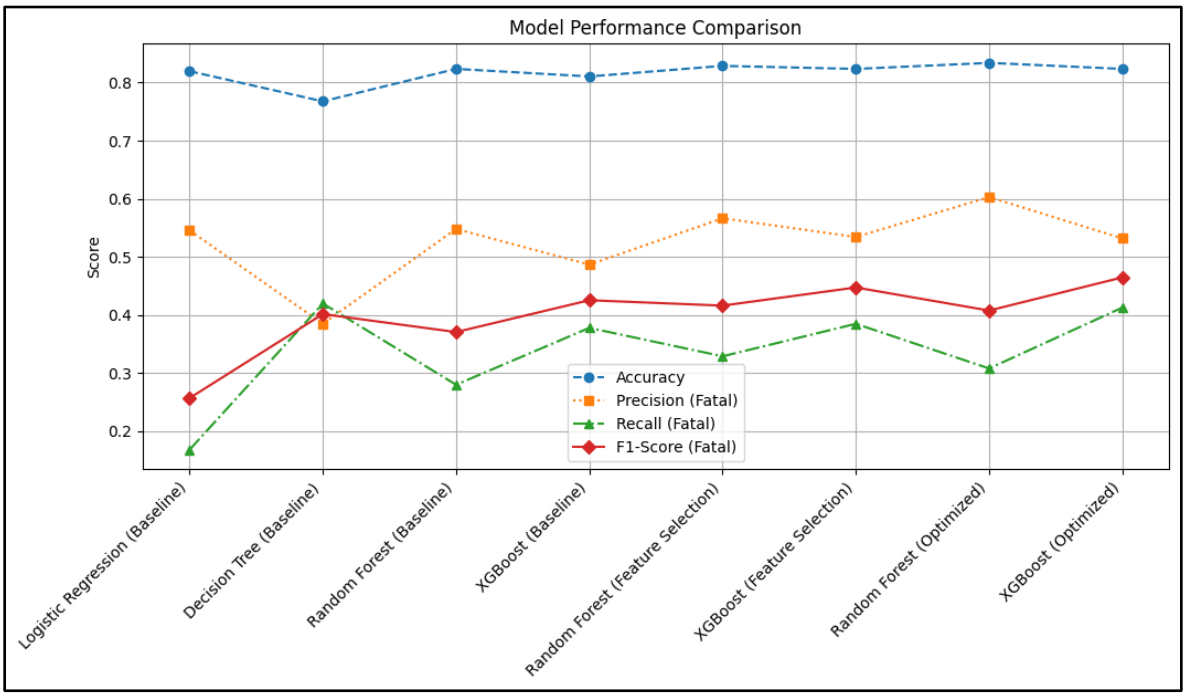


Fig. 3 Model Performance Comparison

Conclusions

We successfully built models to predict shark attack fatality using data mining techniques. Through rigorous EDA, preprocessing, modeling, and validation, we:

- Identified critical predictors like year, activity, country, and species.
- Handled data imbalance effectively.
- Found three natural clusters representing historical and contemporary attack profiles.
- Confirmed statistically significant differences across clusters using ANOVA.

This approach provides a robust framework for analyzing rare events like shark attacks. The biggest takeaways are:

- Fatality risk is decreasing over time, likely due to improved safety measures.
- Activity type & shark species significantly impact risk levels—certain activities (surfing) are more dangerous.
- Clustering revealed distinct attack patterns, useful for safety advisories.
- Machine learning can effectively predict fatality risks, with Random Forest being the best model.

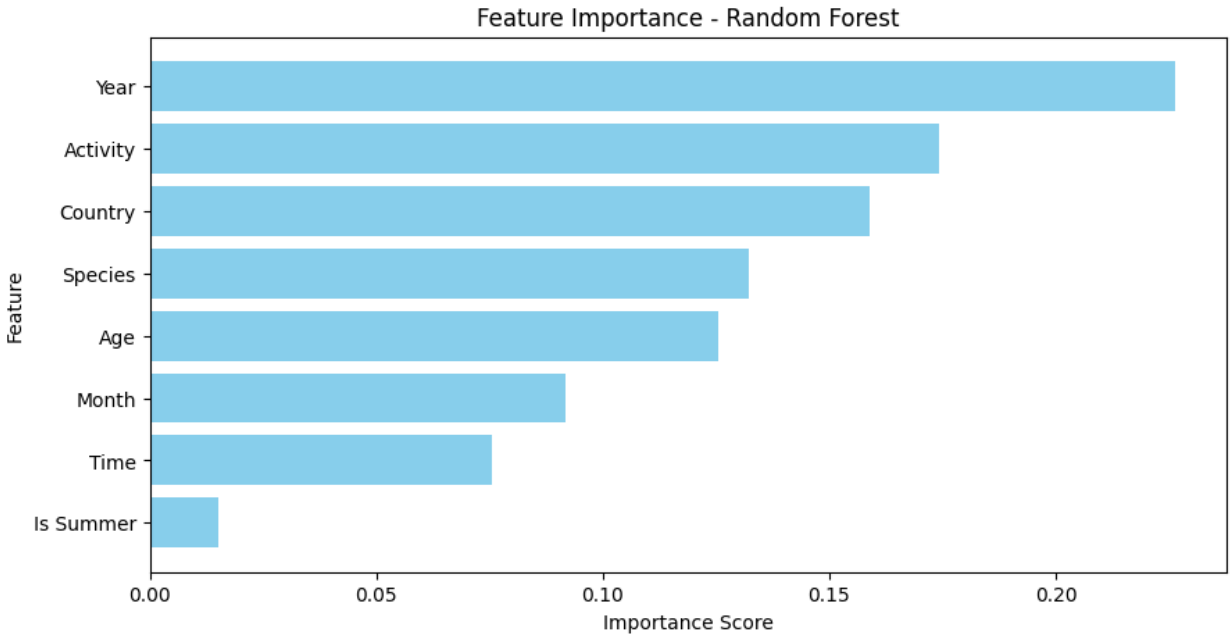
Lessons Learned

- **Real-world data is messy**: We spent significant time cleaning and reshaping the data.
- **Class imbalance is a major challenge**: Without balancing techniques, models underperformed.
- **EDA is critical**: It guided our modeling choices and feature engineering efforts.
- **Experimentation matters**: Iterating on feature selection and tuning led to real performance gains.
- **Visual storytelling is powerful**: Many insights came from well-crafted charts.

Appendix

The appendix includes:

1. Feature Importance Graph for Feature Selection



2. Code in Jupyter Notebook: Attached as a separate file

References

- [1] Felipe Escobar. (2023). *Shark Attack Dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/felipeesc/shark-attack-dataset
- scikit-learn: https://scikit-learn.org/
- imbalanced-learn: https://imbalanced-learn.org/
- matplotlib & seaborn documentation
- Georgia Tech ISyE 7406 Course Materials