Title: Wildlife Trade Risk Prediction

Hamoye Internship Fall' 23 Premiere Project

Group: Code-Analytics

Group Members: Bamidele Tella, Bharat Kumar Jhawar, Okerinde Peculiar Temilola, Olorunleke White, Priscila Waihiga Kamiri, George Israel Opeyemi, Oladimeji Williams, Halimah Oyebanji, Lukman Aliyu, Duke Effiom, Martha Edet.

INTRODUCTION

Wildlife trading, often referred to as the wildlife trade or wildlife trafficking, involves the illegal or legal buying, selling, and exchanging of animals and plants from the wild, including their body parts, derivatives, or products. Legal wildlife trade occurs when governments or organizations permit the controlled exchange of certain wildlife species and their products. This is often for purposes such as conservation, research, or the pet trade. Protected species are organized into three appendixes. Appendix I species are those whose trade threatens them with extinction. Appendix II species are those not threatened with extinction, but whose trade is nevertheless detrimental. Finally, Appendix III animals are those submitted to CITES by member states as a control mechanism. Their export or import requires permits from the submitting member state(s).

PROBLEM STATEMENT

The international trade in endangered species is a great concern for biodiversity preservation. This project seeks to utilize the CITES dataset of 2016 and 2017 in order to tackle the following significant challenges:

1. Common Species Trading: Explore the extent of export and import activity for iconic and endangered species and other high-profile targets for poaching and trade.
2. Animal Products Trade: Determine what percentage of the trades are plant species as opposed to animal products. How does this vary across different species and regions?
3. Conservation Impact Assessment: Evaluate the impact of CITES regulations and trade restrictions on the conservation of species in different CITES appendices (I, II, III).

By addressing these challenges, our project seeks to contribute to a better understanding of international wildlife trade, its implications for biodiversity conservation, and the effectiveness of CITES regulations. This knowledge can inform conservation efforts and policy-making at both national and international levels.

AIM

To analyze the international wildlife trade data from CITES in 2016-2017 and derive actionable insights that contribute to biodiversity preservation and conservation efforts.

OBJECTIVE

- Identify key species, trade routes, and countries involved in the trade.
- Map the regions with the highest trade volumes and assess their impact on local ecosystems.
- Evaluate the effectiveness of CITES regulations in curbing trade in species listed in different appendices (I, II, III).
- Investigate the correlation between the regulatory status of species and their trade activities.

DATASET

The dataset was obtained from Kaggle via the link below:
https://www.kaggle.com/datasets/cites/cites-wildlife-trade-database

DATA PREPROCESSING

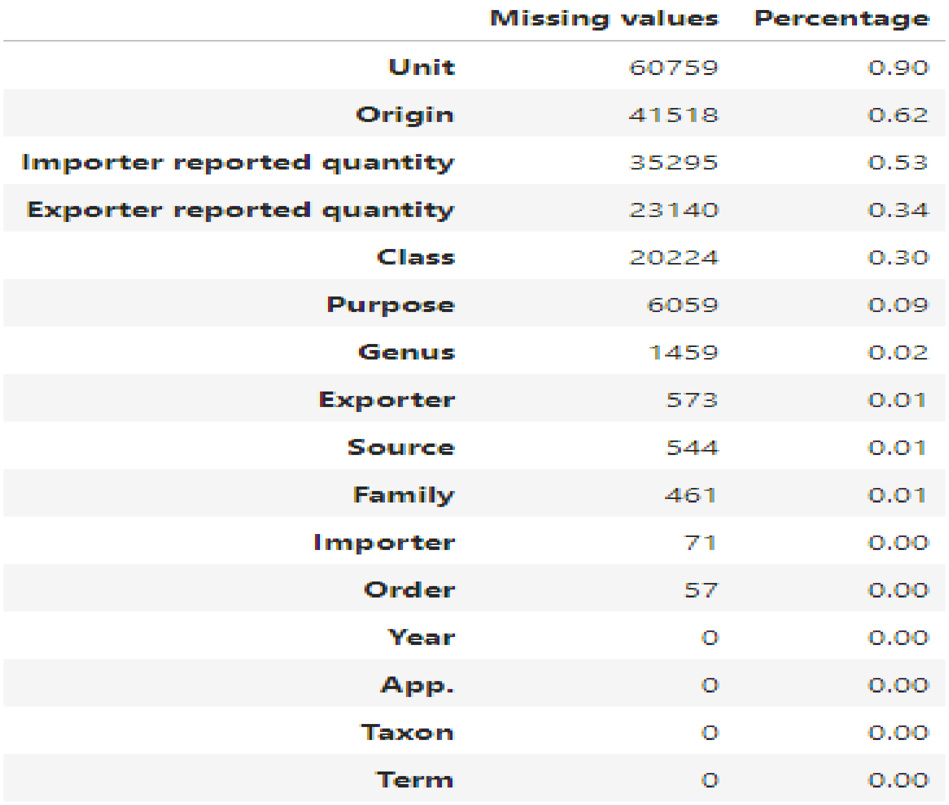
The following procedures were used to prepare the data:

1. Features which had more than 30000 non-null values were dropped and the rest of the features were dropped.
2. Null values in Import and export features were filled with a zero.
3. Removes all features with null values beyond a threshold of 30000.
4. Feature “Year” was dropped due to lack of impact on our analysis.
5. Feature “Origin” and “Unit” are dropped due to null values above threshold (30000).
6. Dropped leaky features i.e. features that won’t be available to the model in production and provide way too much initial context than is required.

EXPLORATORY DATA ANALYSIS

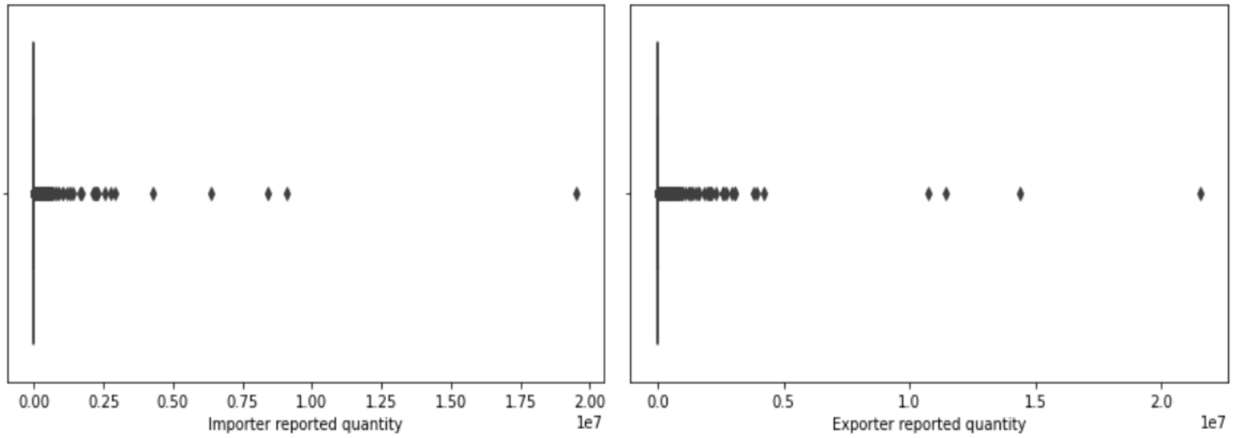
For exploratory data analysis (EDA), we perform univariate exploration and also perform multivariate exploration so that we can understand all the features and also the relation between features.

Firstly, we print a table that shows missing values in the data, 90% of Units, 62% of origin, 53% of Imported reported quantity, 34% of Exported reported quantity, and 30% of class are all missing values.



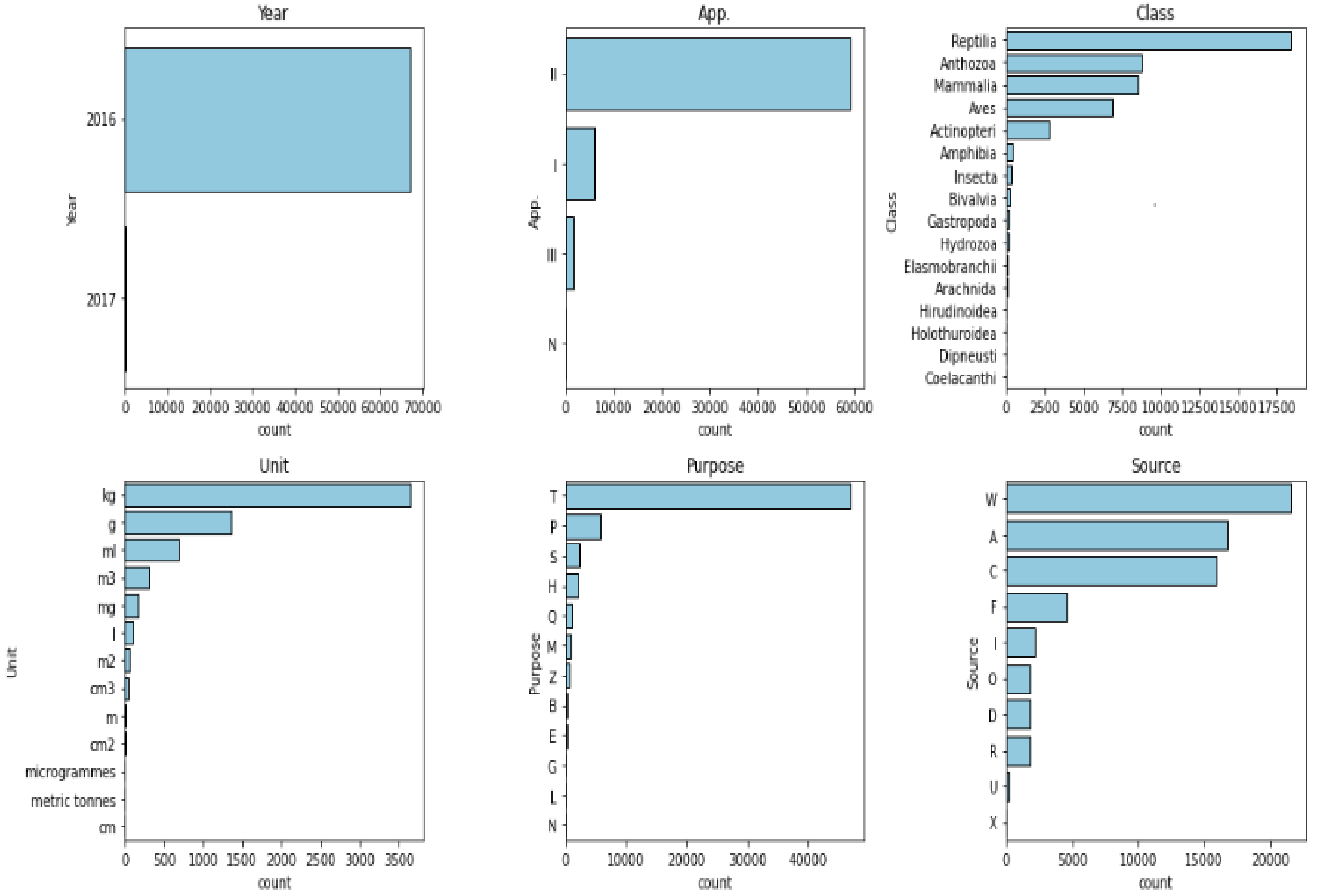
	Missing values	Percentage
Unit	60759	0.90
Origin	41518	0.62
Importer reported quantity	35295	0.53
Exporter reported quantity	23140	0.34
Class	20224	0.30
Purpose	6059	0.09
Genus	1459	0.02
Exporter	573	0.01
Source	544	0.01
Family	461	0.01
Importer	71	0.00
Order	57	0.00
Year	0	0.00
App.	0	0.00
Taxon	0	0.00
Term	0	0.00

```
Duplicated rows are 3 which constitute of 0.00447 % of our dataset
==================================================================================
```



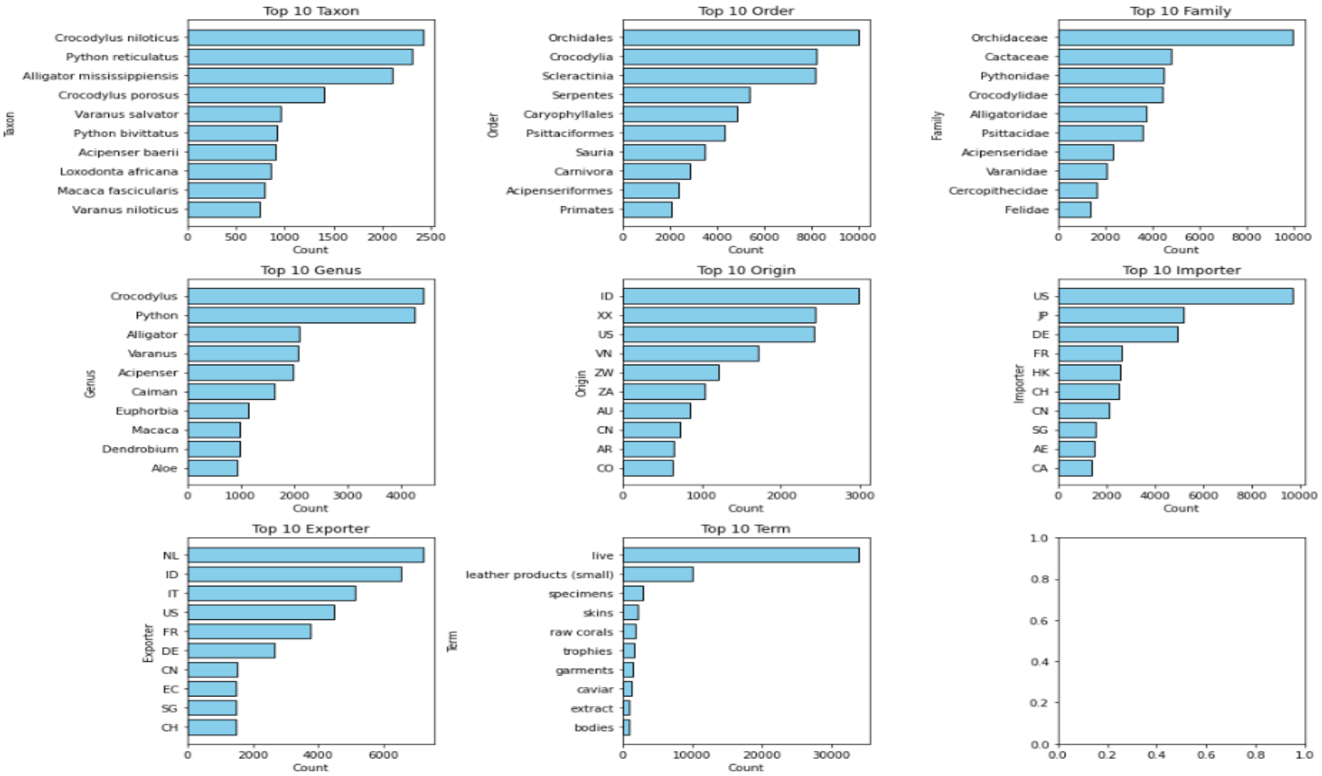
Above, we confirm that there are 3 duplicate rows which constitute 0.00447% of our data. The plot shows that the importer-reported quantity and Exporter recorded quantity both contain outliers.

Categorical Variables



We then use a histogram to show the different categorical classes present in some features of the dataset in ascending order.

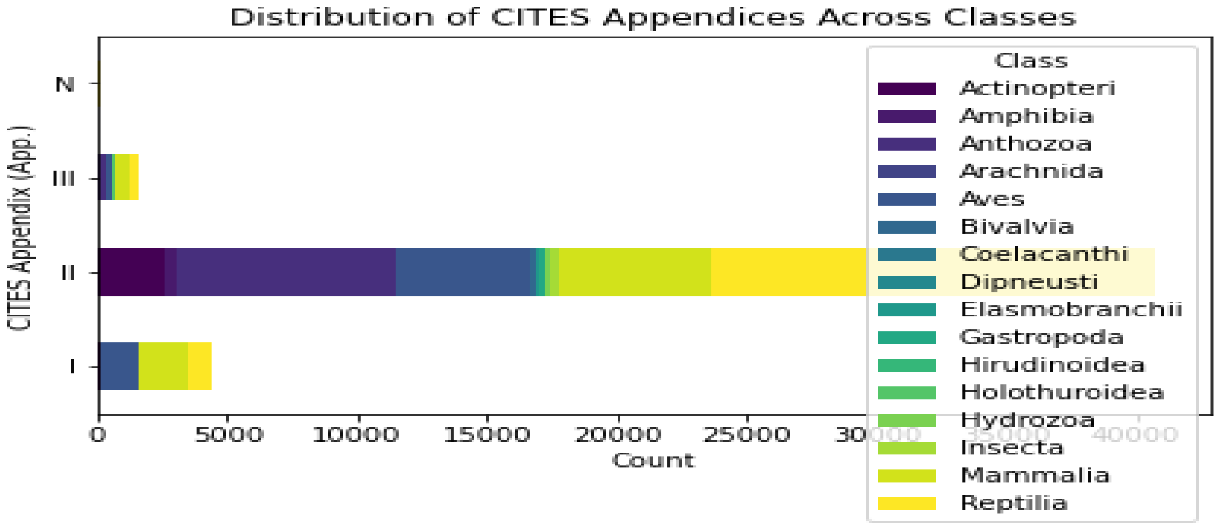
Categorical Features Top Count



Next, we plot the top 10 categories for some features in ascending order: Taxon, Order, Family, Genus, Origin, Importer, Exporter, Term.

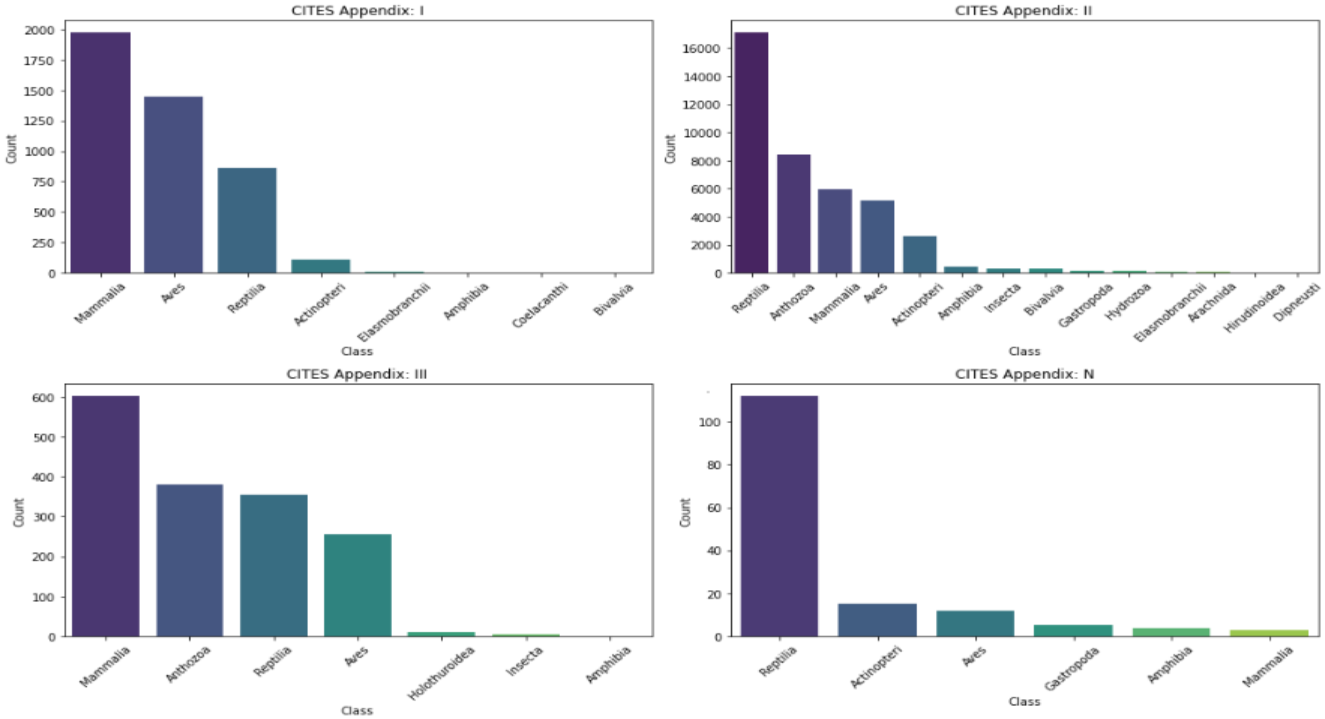
We then used **Multivariate Analysis** to understand the relation between the features of the dataset more clearly.

Classes vs App.



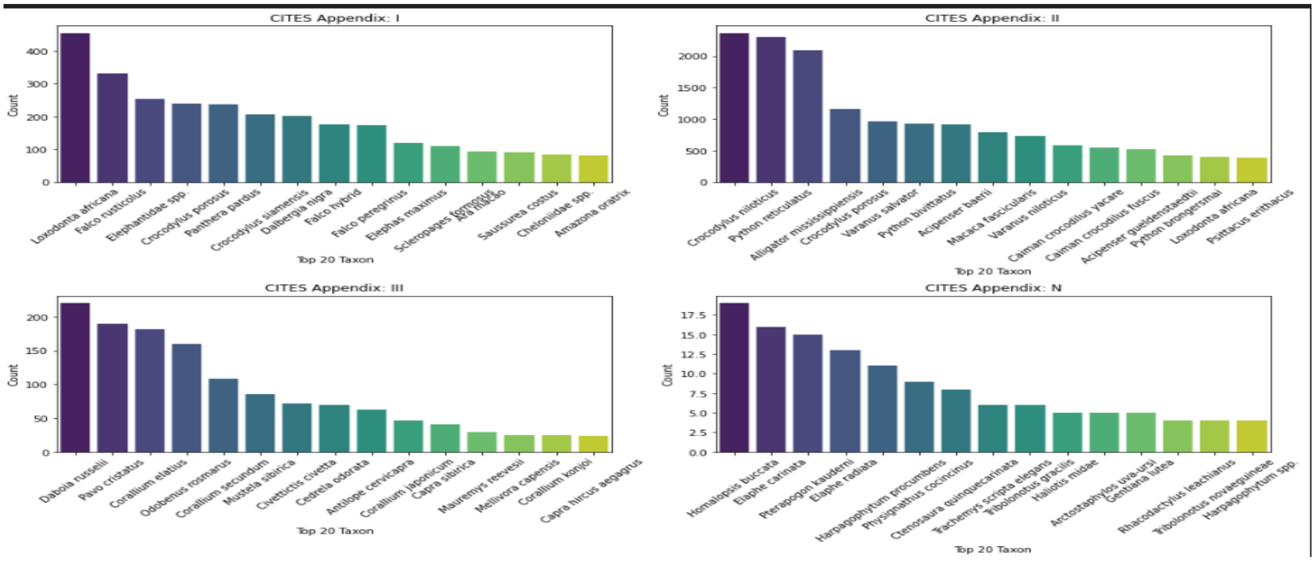
The horizontal bar plot above shows the relation between a traded species class and the different levels of wildlife trade risk using the Appendix feature and it is clearly seen that "Reptilia" is the Species Class with trade risk II.

To get a better idea of the amount of species classes that belong to each category of Risk, we use a vertical bar plot for the top species class of each Risk level.



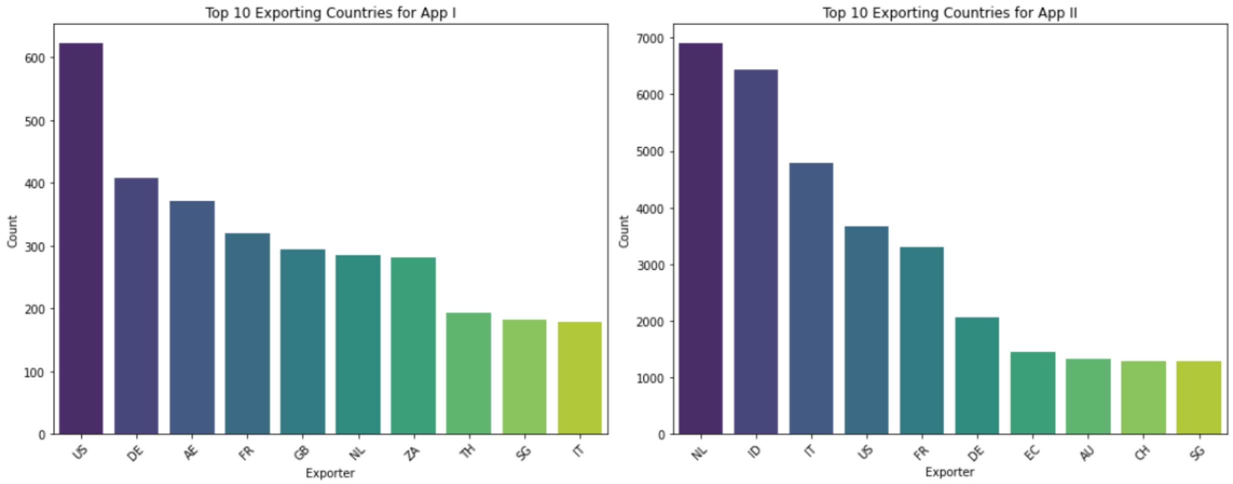
We see that class Mammalian has the highest trade risk with the threat of extinction. As seen earlier, class Reptilian has the highest trade for Appendix level II which means species whose trade is detrimental. We also see that class Mammalian has the most protection from CITES member states in Appendix III.

Top 15 Taxon being traded across the App. categories



We also take a look at the Taxon which is the Scientific name of an animal or plant traded and the CITES level of regulation in the Appendix feature.

Top 10 Exporting Countries for App I and App II(Endangered categories)



Among the three (3) levels of risk, the above plot shows the countries with the most export trade of Endangered Species (I) and Species which trade is detrimental to (II).

BASELINE MODEL

For our first model, we played around with the different features of the dataset. We dropped the ‘Class’, ‘Origin’, and ‘Unit’ feature due to the presence of a large number of null values. The null values in the quantity imported and exported were filled with zero (0) to indicate no quantity imported nor exported. Next, we use Scikit Learn’s simple imputer to fill in the null values in the categorical features with the most frequent value each feature contains. Afterwards, we dropped the rows with the ‘N’ value in the Appendix feature as it was unexplained and assumed as null values. Next, we encode the categories to make it easy to build a model as classifiers cannot convert string values to float. Features like Taxon, Order, Family, Genus, Importer, Exporter, and Term were encoded with the frequency of each unique value using the map function while features like Year, App, Purpose, and Source were label encoded. Next, we split the dataset into train and test datasets. We trained models using Logistic Regression, Random Forest and Decision Tree Classifiers to train models and evaluate them with the test dataset. With the Logistic Regression model, we had an accuracy of 88% and an F1 score of 31%. The Decision Tree model gave us an accuracy of 97% and an F1 score of 94%. The Random Forest model had an accuracy of 98% and an F1 score of 95%.

FINAL MODEL DEVELOPMENT

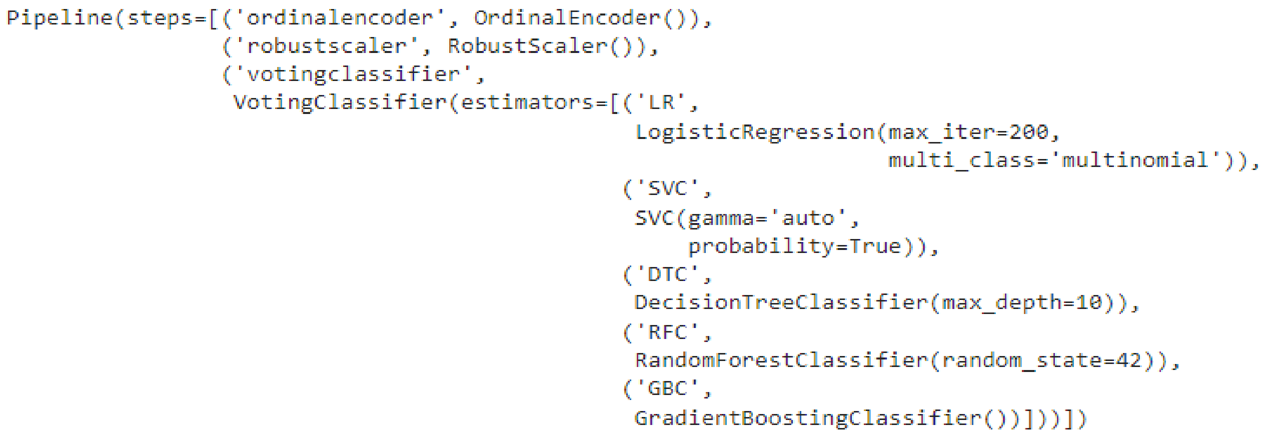
Model Pipeline

The model is a hard voting classifier comprising of:

- Multinomial Logistic Regression
- Support Vector Classifier
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier

The main pipeline uses an ordinal encoder to handle categorical features and a RobustScaler to curb the effect of outliers in the dataset before fitting the Voting Classifier.

In our final model, we first defined a function named ‘data_wrangler’ that cleaned the data in a single call. It had an argument ‘threshold’ which was set to the value 30000, this was used to drop any column that had up to or more than 30000 missing values. Next, we filled in Null values in the Quantity Imported and Quantity Exported columns with ‘0’ to imply that no quantity was either imported or exported in such an instance. Afterwards, we dropped 193 rows that had the value ‘N’ in the Appendix column. Next, we dropped the columns with the least importance to the output of the model and also dropped the rows with null values. Next, we built a pipeline that encoded, robust-scaled and used a voting classifier with hard voting to combine the use of Logistic Regression, Support Vector Classifier(SVC), Random Forest Classifier and Gradient Boosting Classifier to build the final model. Finally, we trained the model with our dataset.

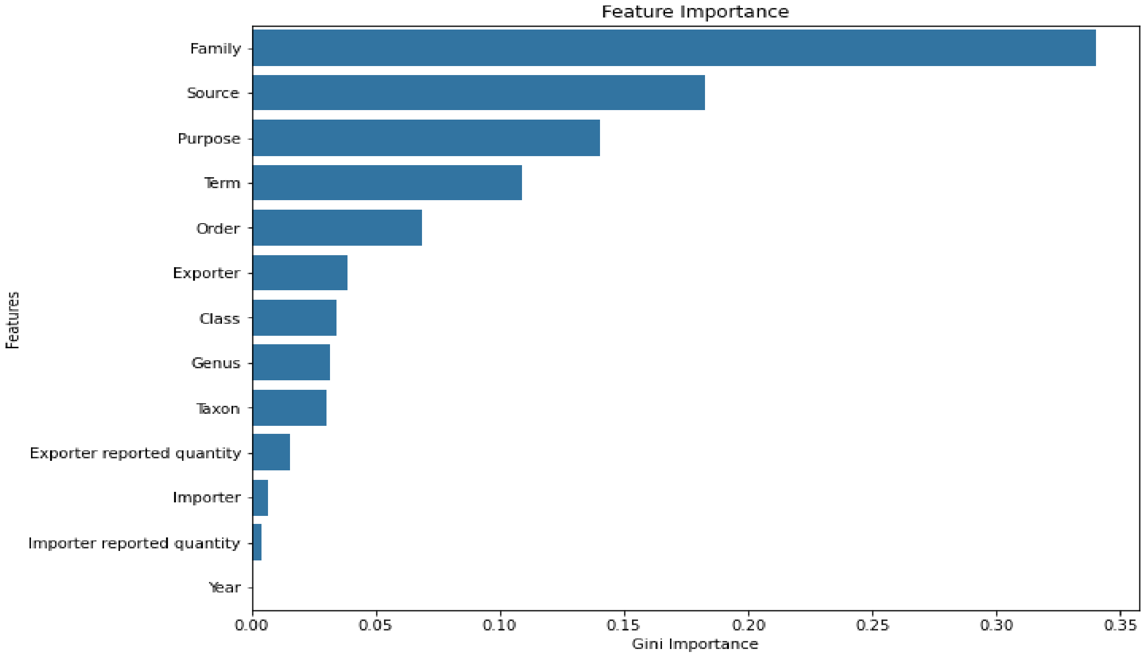


```
Pipeline(steps=[('ordinalencoder', OrdinalEncoder()),
                ('robustscaler', RobustScaler()),
                ('votingclassifier',
                 VotingClassifier(estimators=[('LR',
                                               LogisticRegression(max_iter=200,
                                                                  multi_class='multinomial')),
                                              ('SVC',
                                               SVC(gamma='auto',
                                                   probability=True)),
                                              ('DTC',
                                               DecisionTreeClassifier(max_depth=10)),
                                              ('RFC',
                                               RandomForestClassifier(random_state=42)),
                                              ('GBC',
                                               GradientBoostingClassifier())]))])
```

Final Model

EVALUATION

Our final model had an accuracy of 91% on the test dataset and 93% on the training dataset. For the Appendix class I, we got a precision score of 90%, a recall score of 89% and an F1 Score of 90%. In Appendix Class II, we got a precision score of 99%, a recall score of 98% and an F1 Score of 99%. In Appendix Class III, we got a precision score of 54%, a recall score of 62% and an F1 Score of 58%. We were also able to use the Feature Importance Plot to the features in descending order of importance to the model.



DEPLOYMENT

The model was deployed as a streamlit application where values to predict are entered and the model's predictions are displayed on the same page.

The app has two modes:

- Single prediction
- Batch prediction

Single prediction requires a user to enter the individual features in the text box provided and click the on-screen button to return a prediction.

Batch prediction requires a file upload. The upload should be a CSV file with columns bearing all required features. The final predictions are then appended to the dataframe which can be downloaded for further use outside of the application. After building the application and testing locally, the app code and underlying dependencies were pushed to a GitHub repo and used to create an accessible web app on the Streamlit cloud. The streamlit cloud feeds from the GitHub repository and takes in the repository path where the app script sits. The libraries specified in the requirements.txt file are first installed on a container in the cloud and the app engine serves our application via the specified URL at the time of creation.

CONCLUSION

The efforts of CITES regulations in preserving biodiversity cannot be overlooked and are indeed commendable. However, from our analysis, we can see that a lot of wildlife Specie trade are detrimental to Biodiversity Preservation and increases the rate at which a lot of specie can be lost to extinction due to human activities. We have built a web application that predicts the level of trade risk as a result of a user's input on the features of the trade, These features of the Species to be traded like Taxon, Class, Genus, Order, Family, Genus, Term, Purpose, and Source. This is intended to easily get the risk level of a wildlife Species trade and hasten security measures if required.

LINKS

Our web application link can be found at **https://wildlifeclassifier.streamlit.app/**.
GitHub Repository: **https://github.com/deletella01/Code-Analytics**
Documentation:
https://docs.google.com/document/d/1Ak53g3AA9o-YYTPoXKzEBbbwnai4qF1Rq6nWQLj9PyA/edit?usp=sharing
Google Slide:
https://docs.google.com/presentation/d/1_7C4MoAj_ckispGnMnSw6q5wn49efoeSTIAtQFhI-co/edit?usp=sharing