CAPSTONE PROJECT4

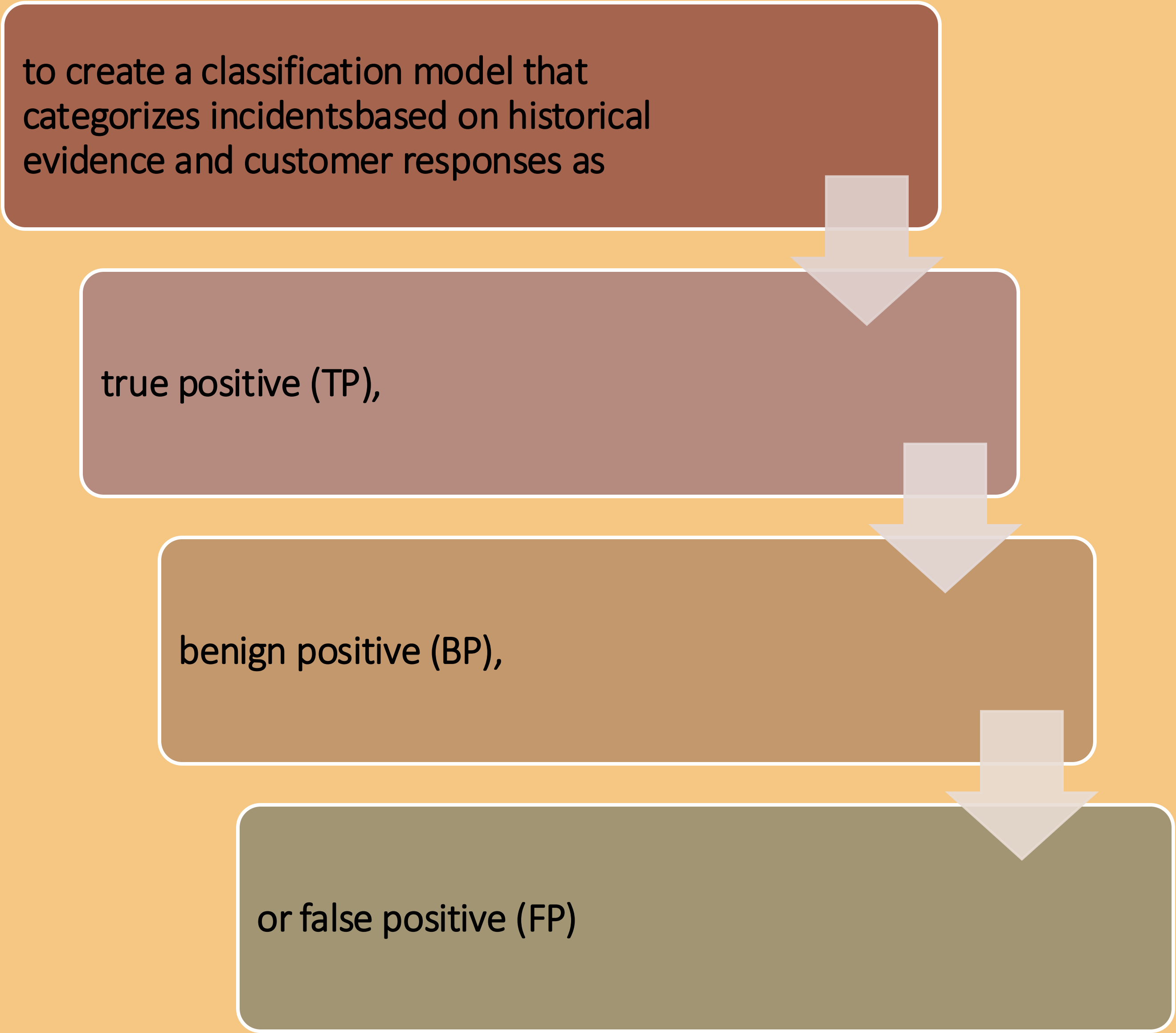
By

S.Vinodha

Microsoft Classifying Cybersecurity Incidents

Machine Learning

Objective



to create a classification model that categorizes incidentsbased on historical evidence and customer responses as

true positive (TP),

benign positive (BP),

or false positive (FP)

BUSINESS USE CASES

:The solution developed in this project can be implemented in various business scenarios, particularly in the field of cybersecurity. Some potential applications include:

- Security Operation Centers (SOCs)
- Incident Response Automation
- Threat Intelligence
- Enterprise Security Management

Data Exploration and Understanding

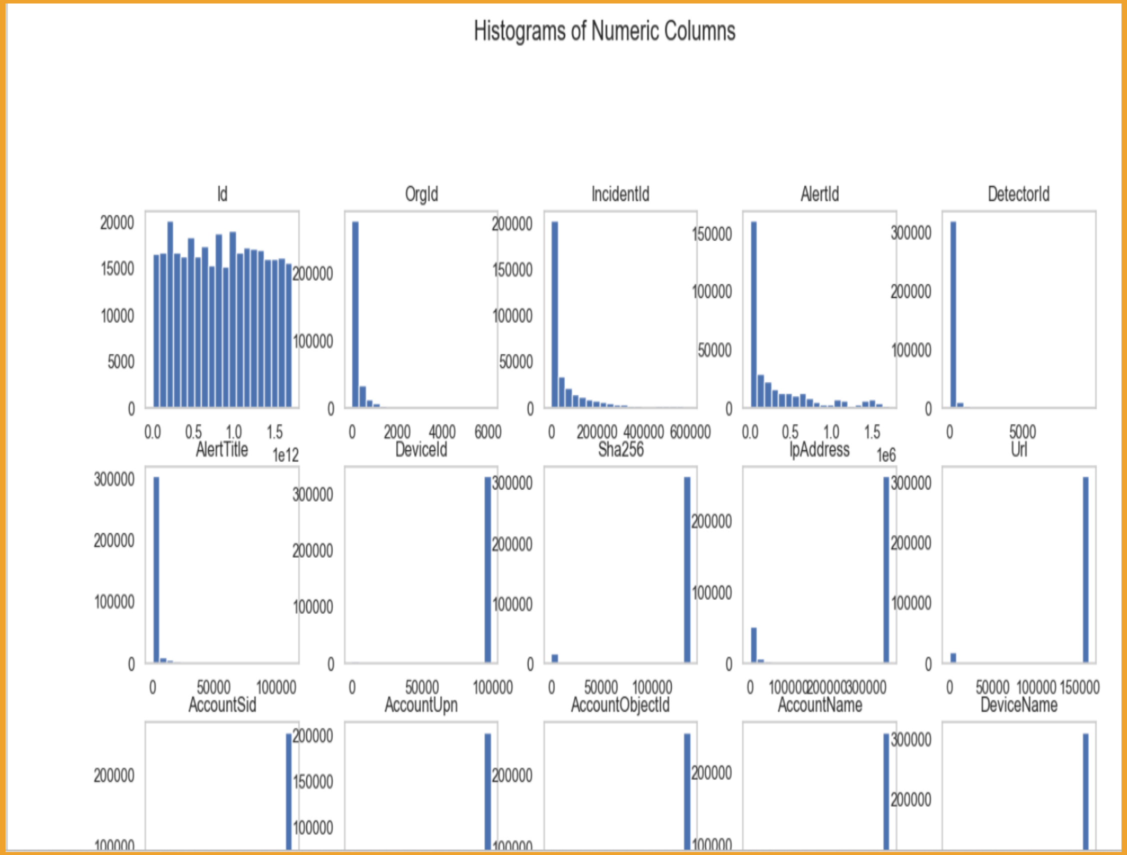
Initial Inspection- Started by loading the train.csv dataset and test.csv and performed an initial inspection to understand the

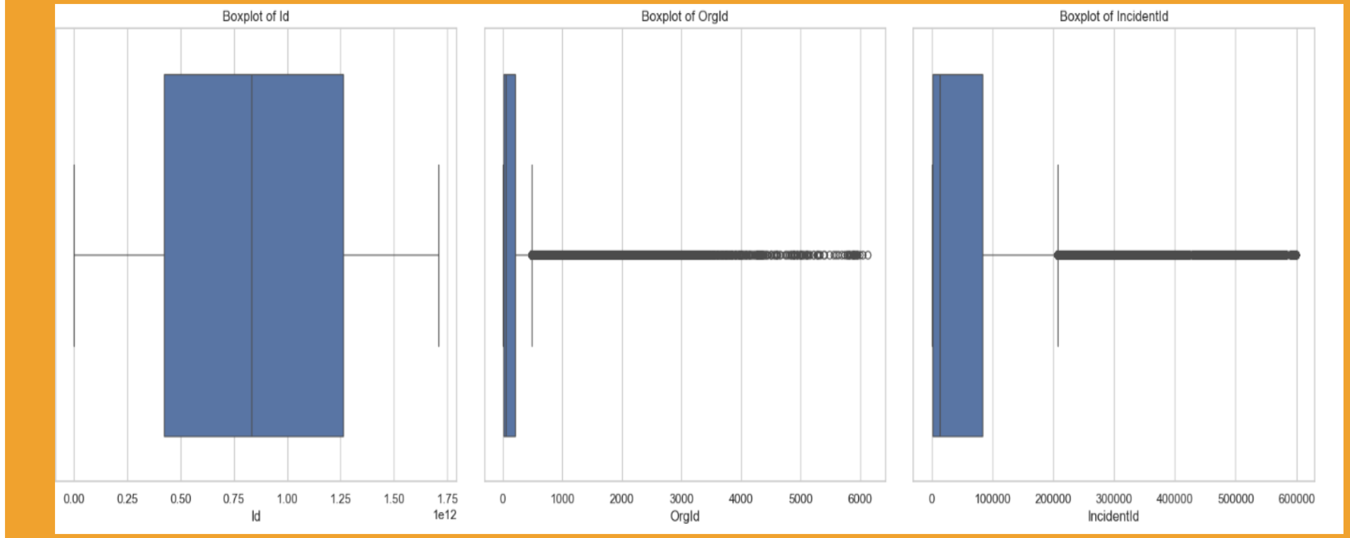
structure of the data,

including the number of features,

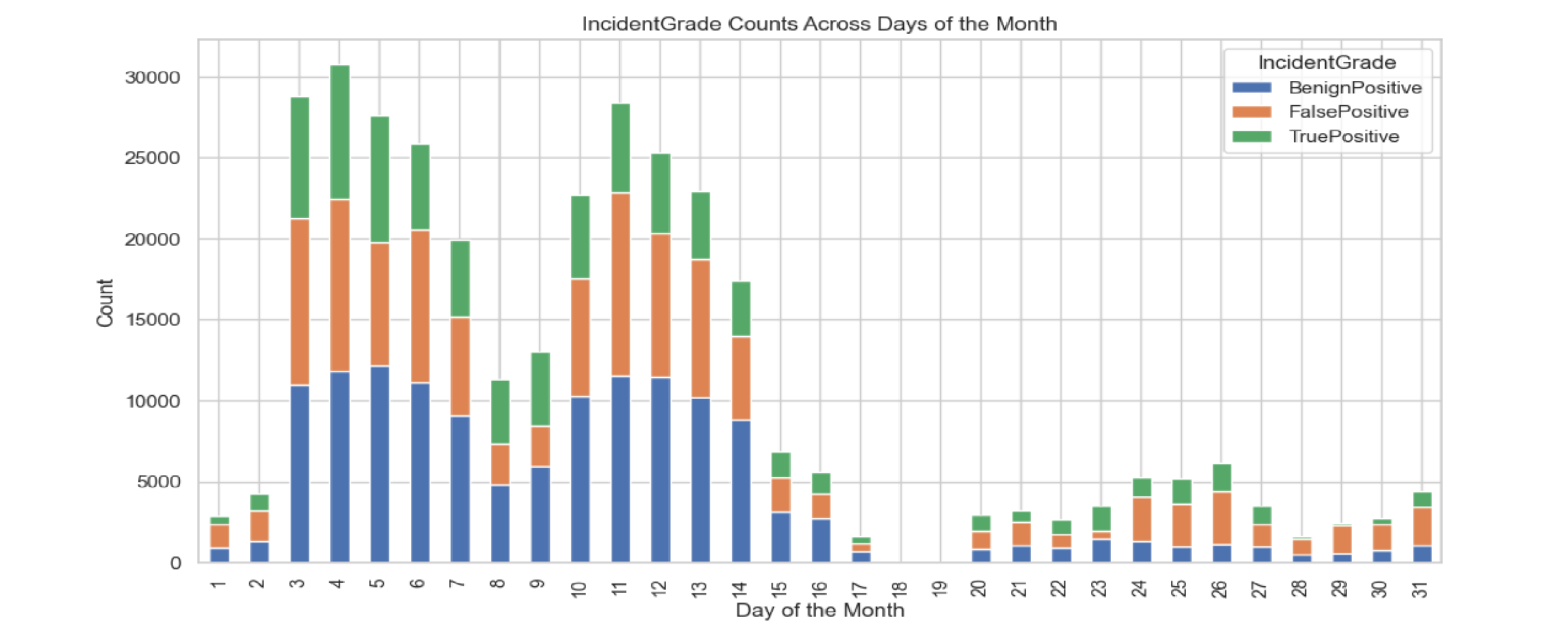
types of variables (categorical, numerical),

and the distribution of the target variable (TP, BP, FP).

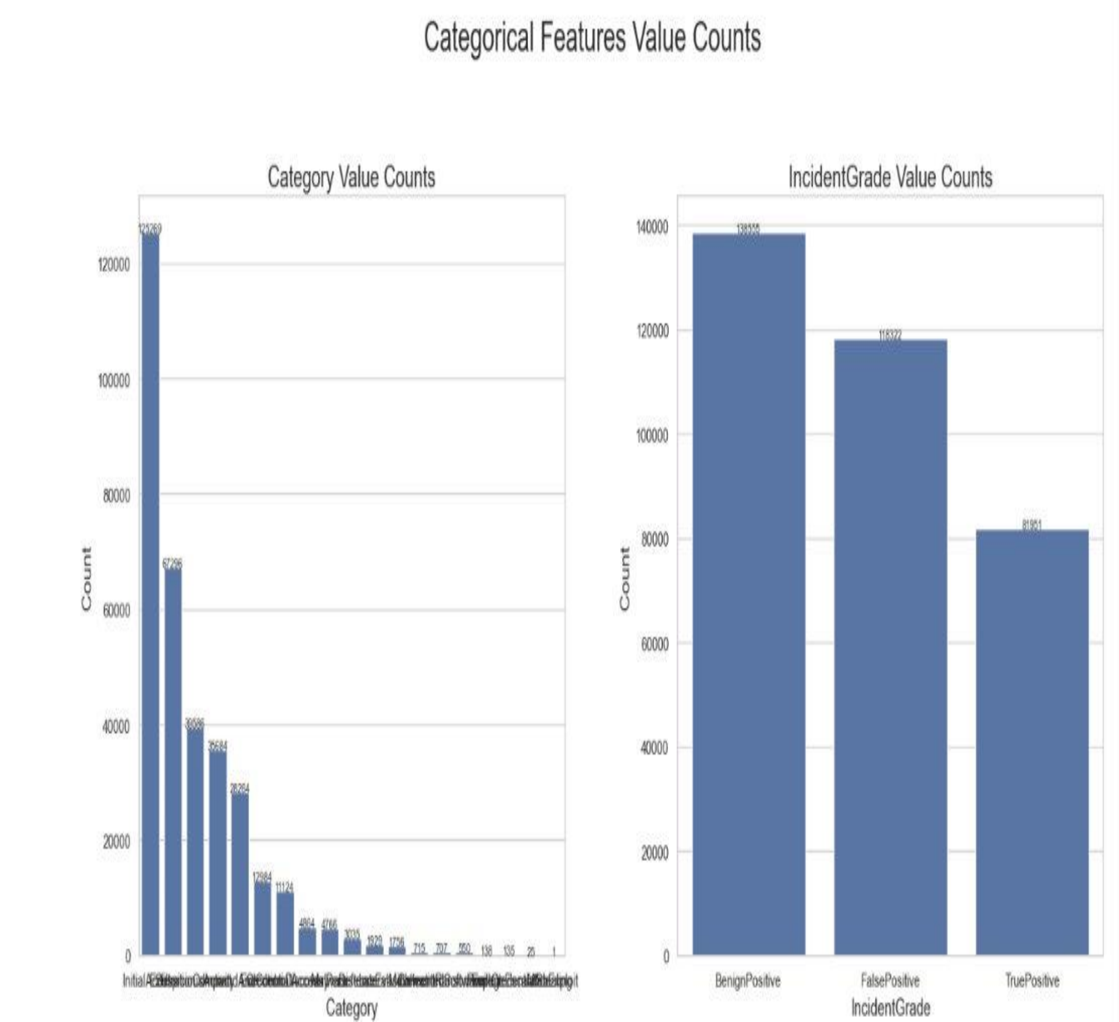


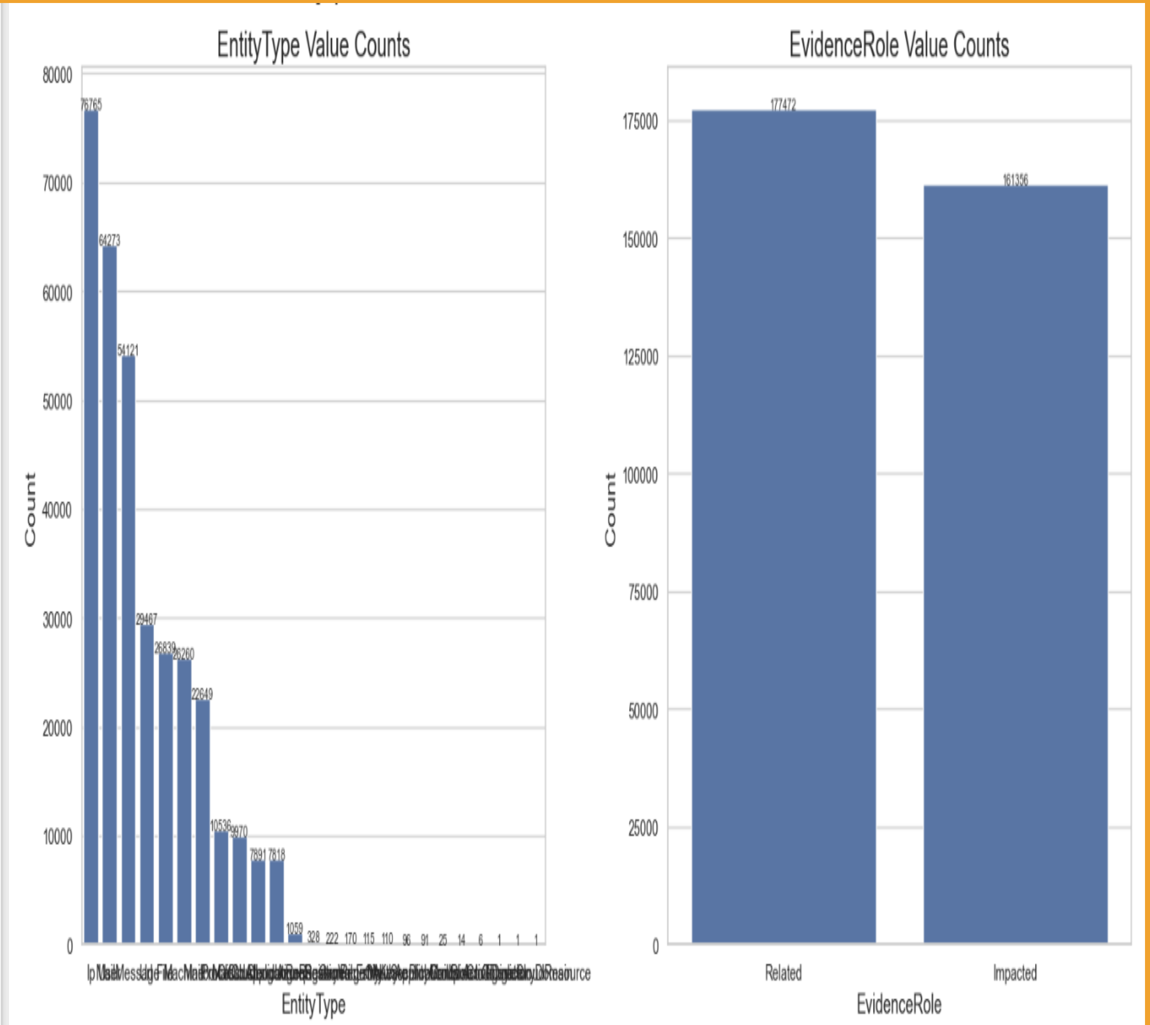


EXPLORATORY DATA ANALYSIS (EDA)



Across day of the month





Visuals on categorical columns

Data Preprocessing

Handling Missing Data

- Identified missing values in the dataset and removed columns which were more than 50% empty and removed affected rows having value o.

Feature Engineering

- Derived new features from timestamps (like hour of the day or day of the week).

Encoding Categorical Variables

- Converted categorical features into numerical representations using label encoding.

Model Selection and Training

Since the target variable is imbalanced, used stratified sampling to ensure that both the training and validation sets have similar class distributions.

Models	Accuracy	F1 Score	Recall	Precision
Logistic Regression	0.495927	0.480229	0.495927	0.496448
Decision Tree Classifier	0.943836	0.943839	0.943836	0.943856
Random Forest Classifier	0.921657	0.921436	0.921657	0.922262

Analysis of the performance

Logistic Regression:

Accuracy: 0.496, which is close to random guessing (around 50%).

F1 Score and Precision Both are low, indicating that this model struggles to correctly classify cases.

Conclusion: This model might not be well-suited for the dataset, possibly due to the complexity of the relationships or non-linear features.

Decision Tree Classifier:

Accuracy: 0.944, significantly higher than Logistic Regression.

F1 Score, Recall, and Precision: All around 0.944, showing consistent performance across metrics.

Conclusion: This model is performing quite well, capturing patterns accurately. The high recall suggests it's sensitive to identifying cases across classes.

Random Forest Classifier:

Accuracy: 0.922, slightly lower than the Decision Tree.

F1 Score, Recall, and Precision: All around 0.922, showing balanced performance but a bit lower than the Decision Tree.

Conclusion: Random Forest provides strong performance and may offer more robustness compared to a single Decision Tree, though slightly lower in this case.

Model Evaluation - Cross validation

Models	Cross-Validated Accuracy	Cross-Validated F1 Score	Cross-Validated Recall	Cross-Validated Precision
Decision Tree Classifier	0.943789	0.943794	0.943789	0.943810
Random Forest Classifier	0.923832	0.923614	0.923832	0.924329

The cross-validated metrics show consistency with the initial evaluation, which is a good sign that the models generalize well. With both models showing high cross-validated performance, the Decision Tree Classifier remains slightly ahead in accuracy and simplicity. However, I prefer a more robust model that may generalize slightly better in varied scenarios, So the Random Forest is also a strong choice.

Hyperparameter Tuning

Best Parameters for Random Forest:

'n_estimators': 200
'min_samples_split':2
'min_samples_leaf':1
'max_depth': None

Best Cross-Validated Accuracy for Random Forest:
0.925372162672301

Journey of model

	Known data				Unknown data			
model	Accuracy	F1 Score	Recall	Precision	Accuracy	F1 Score	Recall	Precision
At first	92.34%	92	92	92	86.93%	87	87	87
After Tuning	92.49%	92	92	93	87.05%	87	87	87
After Feature Engineering	92.61%	93	93	93	87.15%	87	87	87

Thank You