# Resume Screening Using Machine Learning

Final Project Presentation

**Ahsan Habib**

**Date: 07 May, 2025**

# Problem Statement & Objectives

Manual screening risks unconscious bias affecting fairness.

HR spends 23 hours weekly on resume screening (SHRM, 2023).

Poor candidate fit leads to high employee turnover.

Objectives: Increase fairness, accuracy, speed, and reduce costs.

| Bias in Screening | Time Inefficiency | Turnover Impact |
|---|---|---|
| Human judgment can unintentionally favor certain groups. | Manual processes slow down recruitment cycles and waste effort. | Unfit hires cause disruptions and increased costs for companies. |

# Dataset and Report Summary

On the left side we can see Horizontal bar chart, which shows the Frequency of Job Categories, On the Right we can see have create a histogram with Kernel density Estimate (to smoothen Distribution Curve) using Seaborn, which reveals skewness, central tendency and outliers.
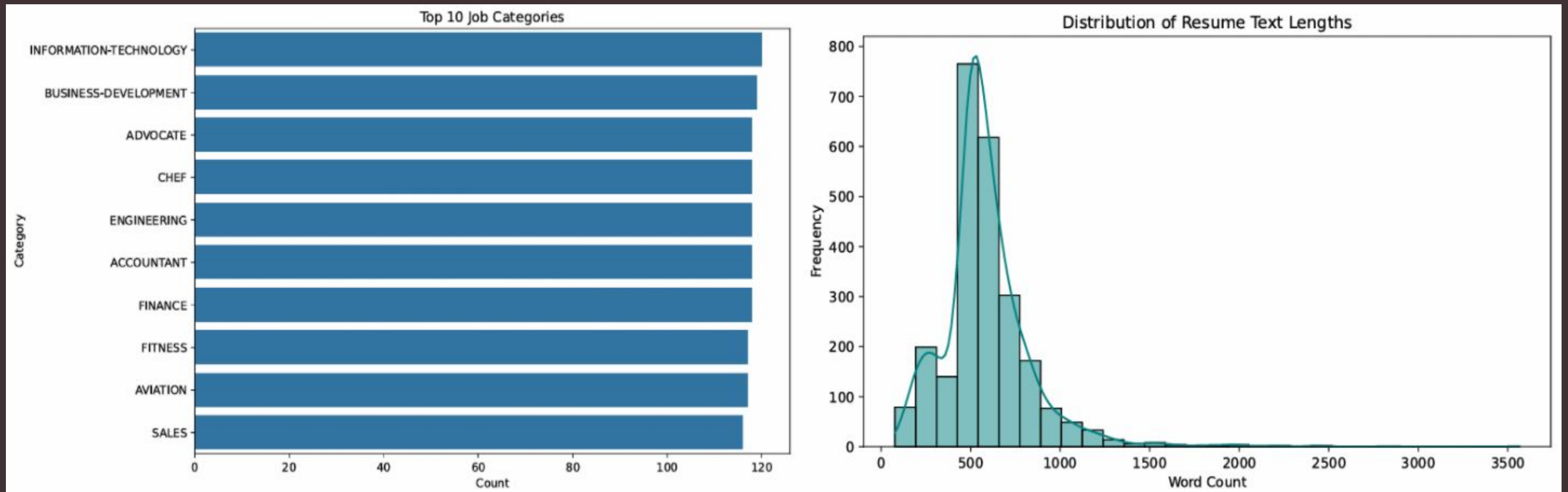
# Dataset and Report Summary

We have also added a box plot and a missing value bar plot, by using Seaborn, to help visualize the dataset better, The Box plot on the left side, helps to detect the median, outliers and quartiles of the length of Long-texts. The graph on the right shows us the frequency distribution of Missing values, which Helps us to Understand the missing Columns across all variables.
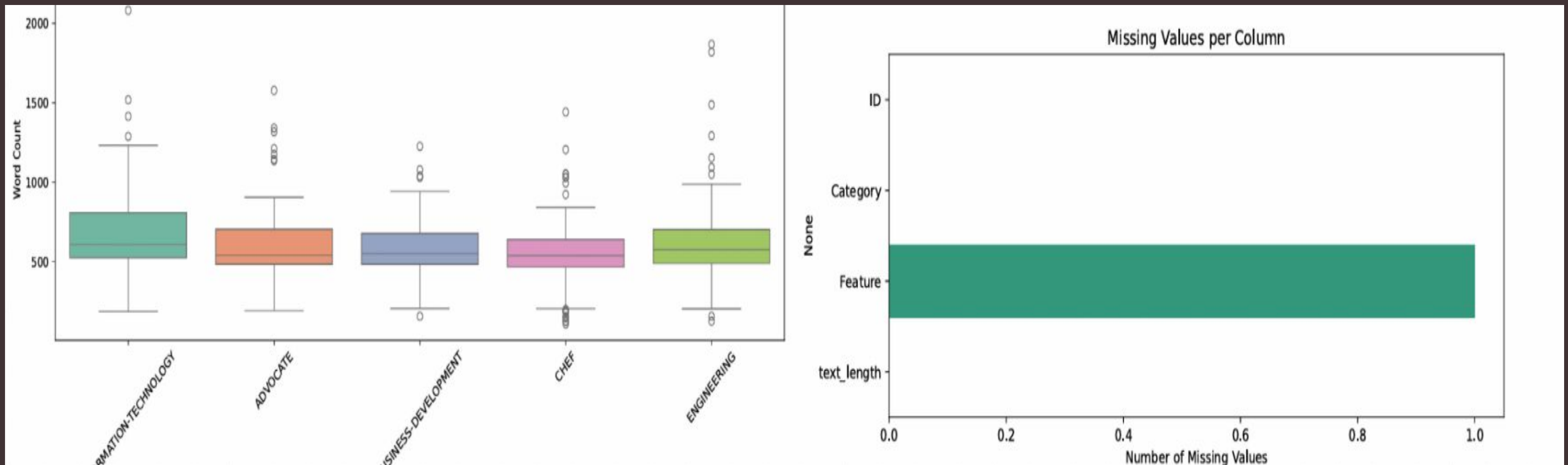
# Dataset and Resport Summary

EDA Summary

```
Dataset Summary:
----------------
Total Records: 2484
Missing Values:
ID                0
Category          0
Feature           1
text_length       0
Resume Text Length - Mean: 587.01, Median: 549.0, Mode: 482
Top Categories:
Category
INFORMATION-TECHNOLOGY        120
BUSINESS-DEVELOPMENT          120
ADVOCATE                      118
CHEF                          118
ENGINEERING                   118
```

# Methodology

**Data Collection & Augmentation**

Data Pre-processing and Feature Engineering

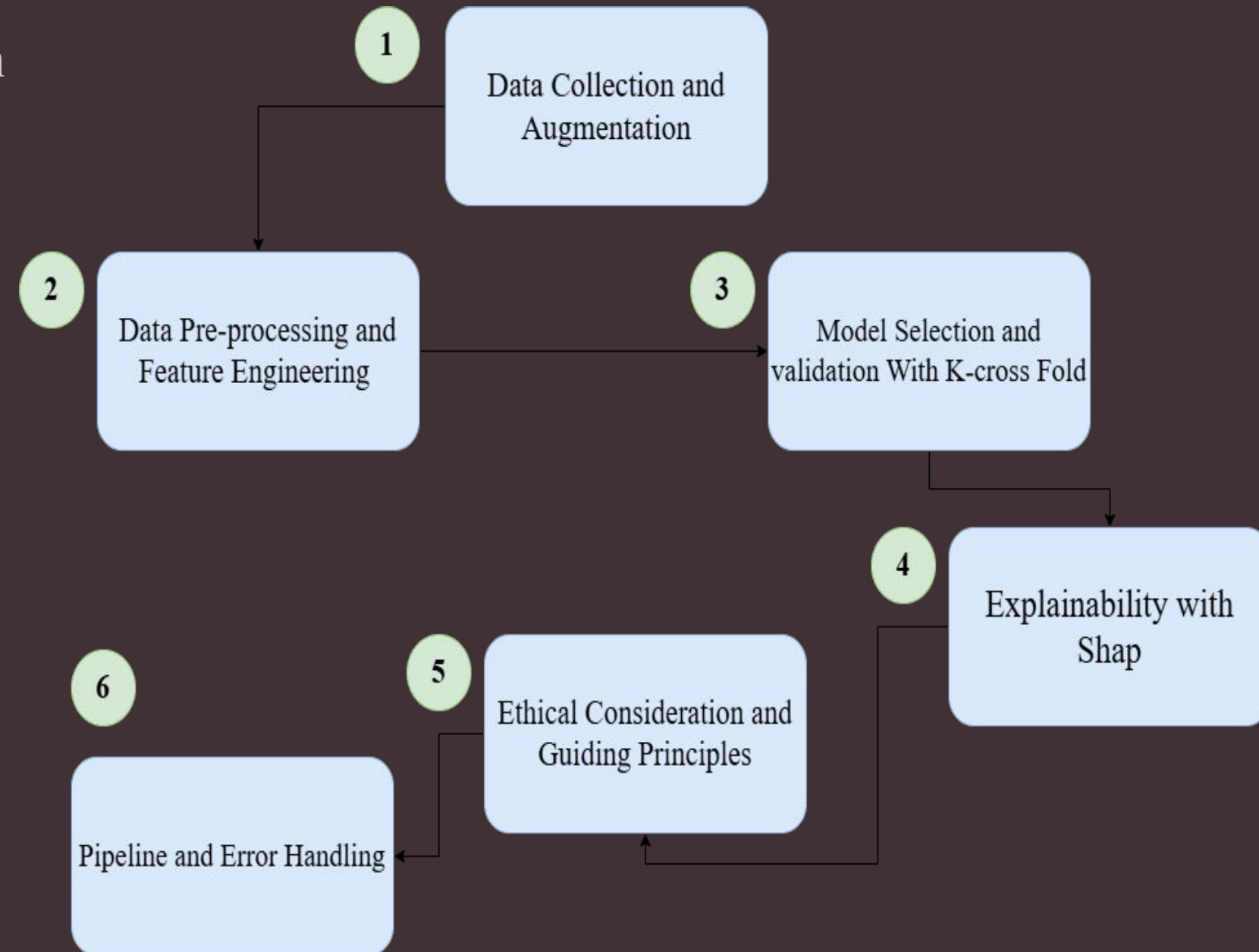Model Selection and validation With K-cross Fold
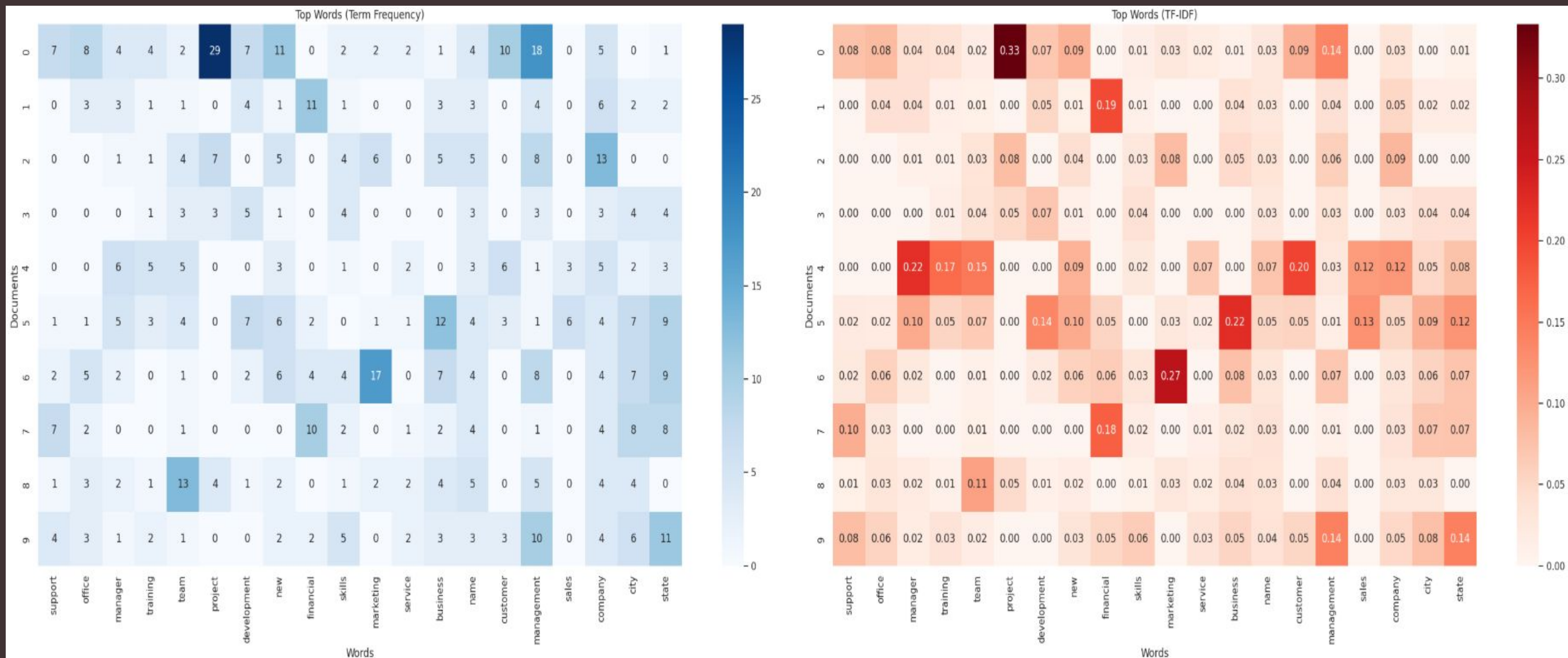
Explainability with Shap

Ethical Consideration and Guiding Principles
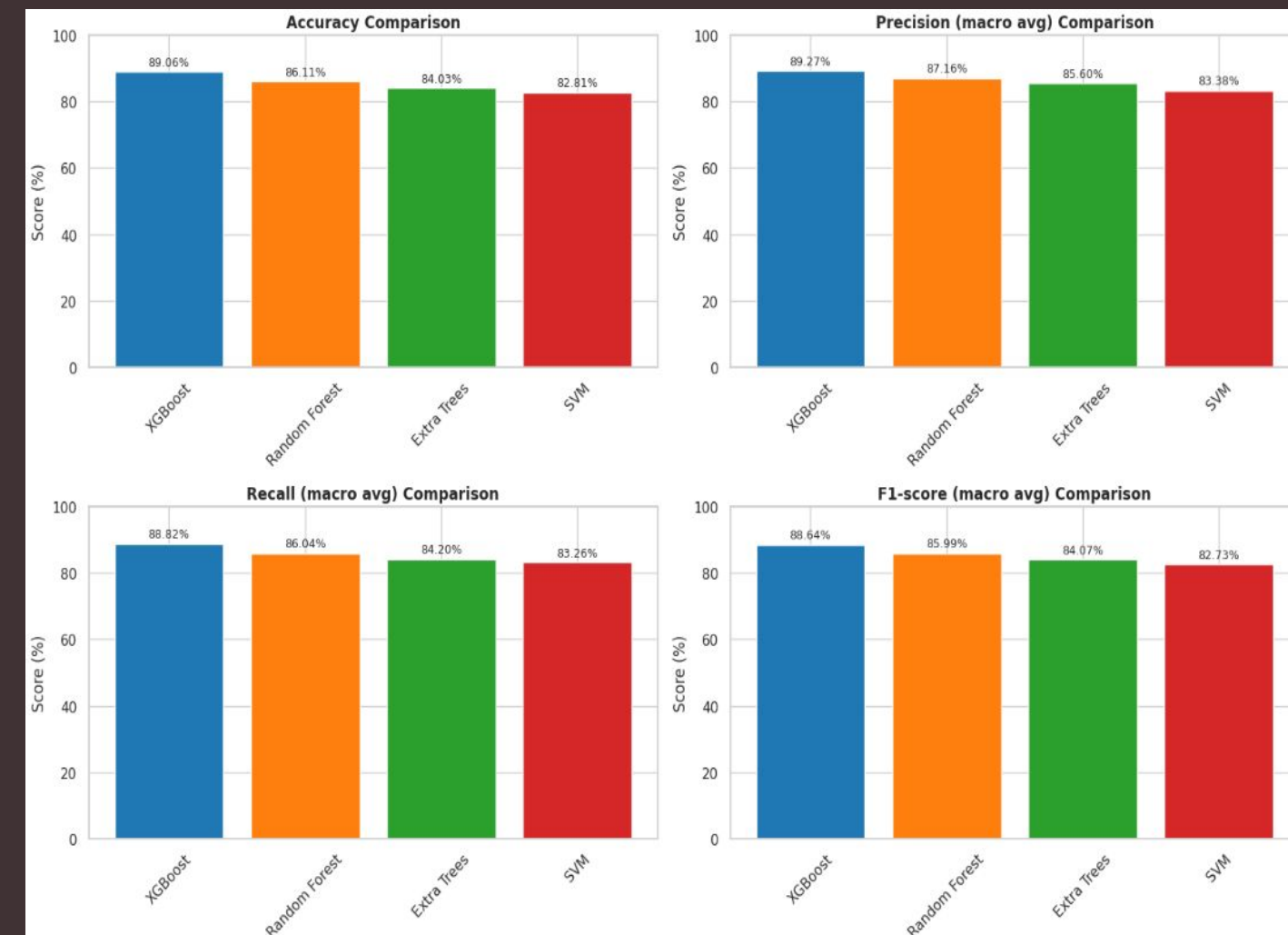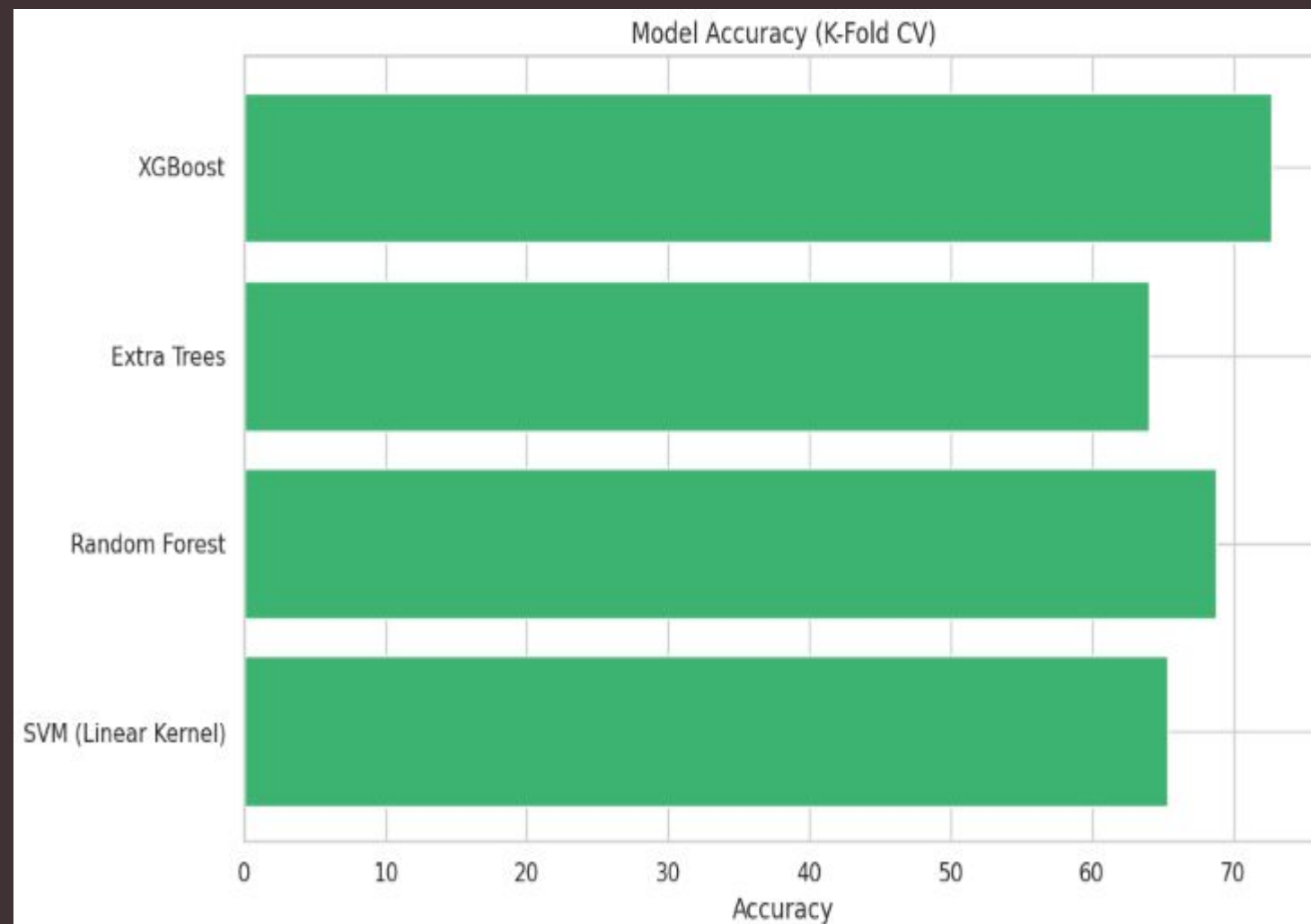
Pipeline and Error Handling

# Correlation Heatmap of Primary Data Vs Pre-processed Data

# Model Selection And Evaluation

Selected Models : By using **Xgboost, SVM, Extra Trees, Random Forest Classifier,** we can compare the performance of these models and Evaluate them using K Cross Fold, based on Accuray, Precision, Recall, F1-Score

# Confusion Matrix Based on the best model XGboost



XGBoost Confusion Matrix
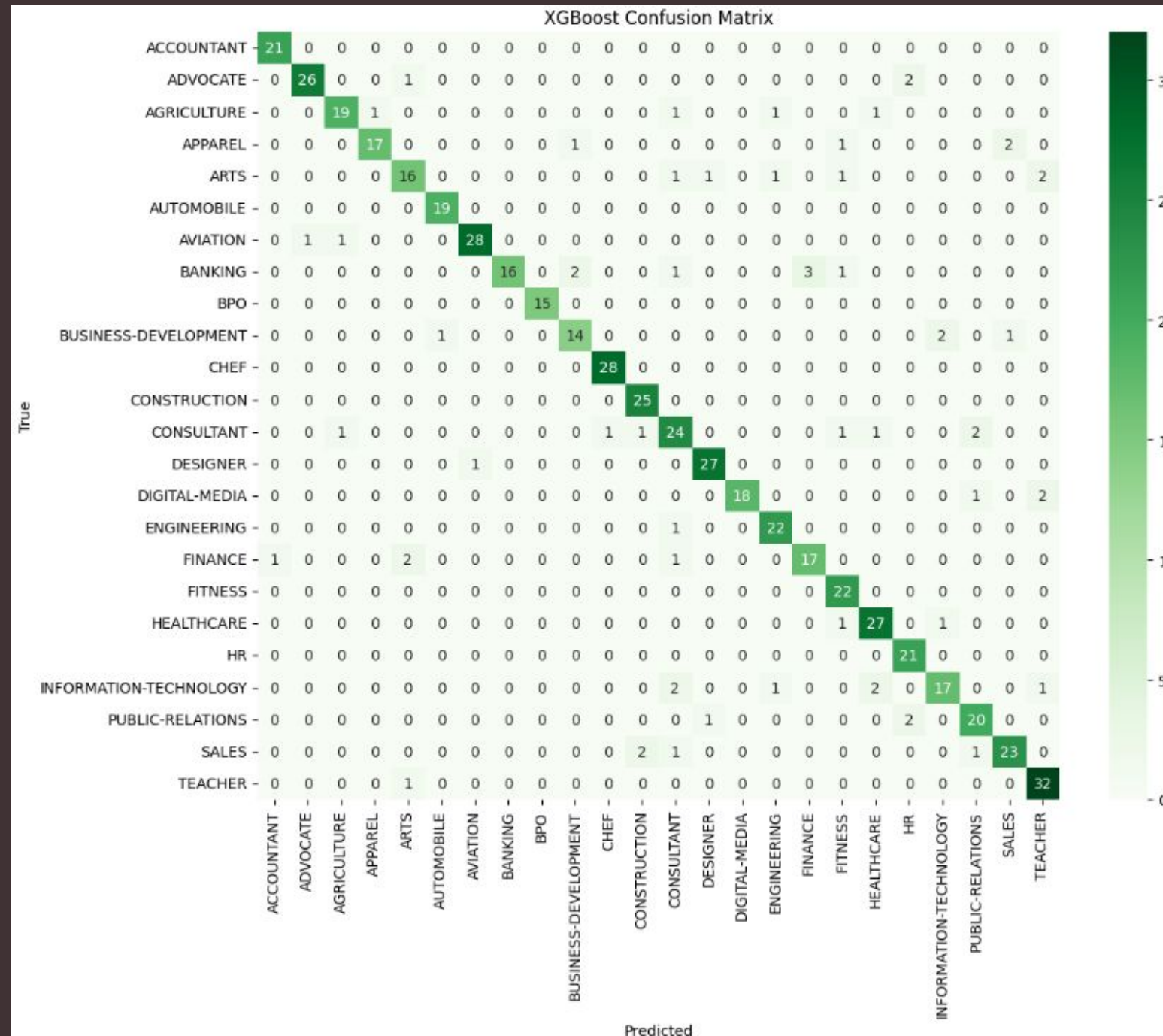
# Explainable AI with Shap (Prediction and Classification result)
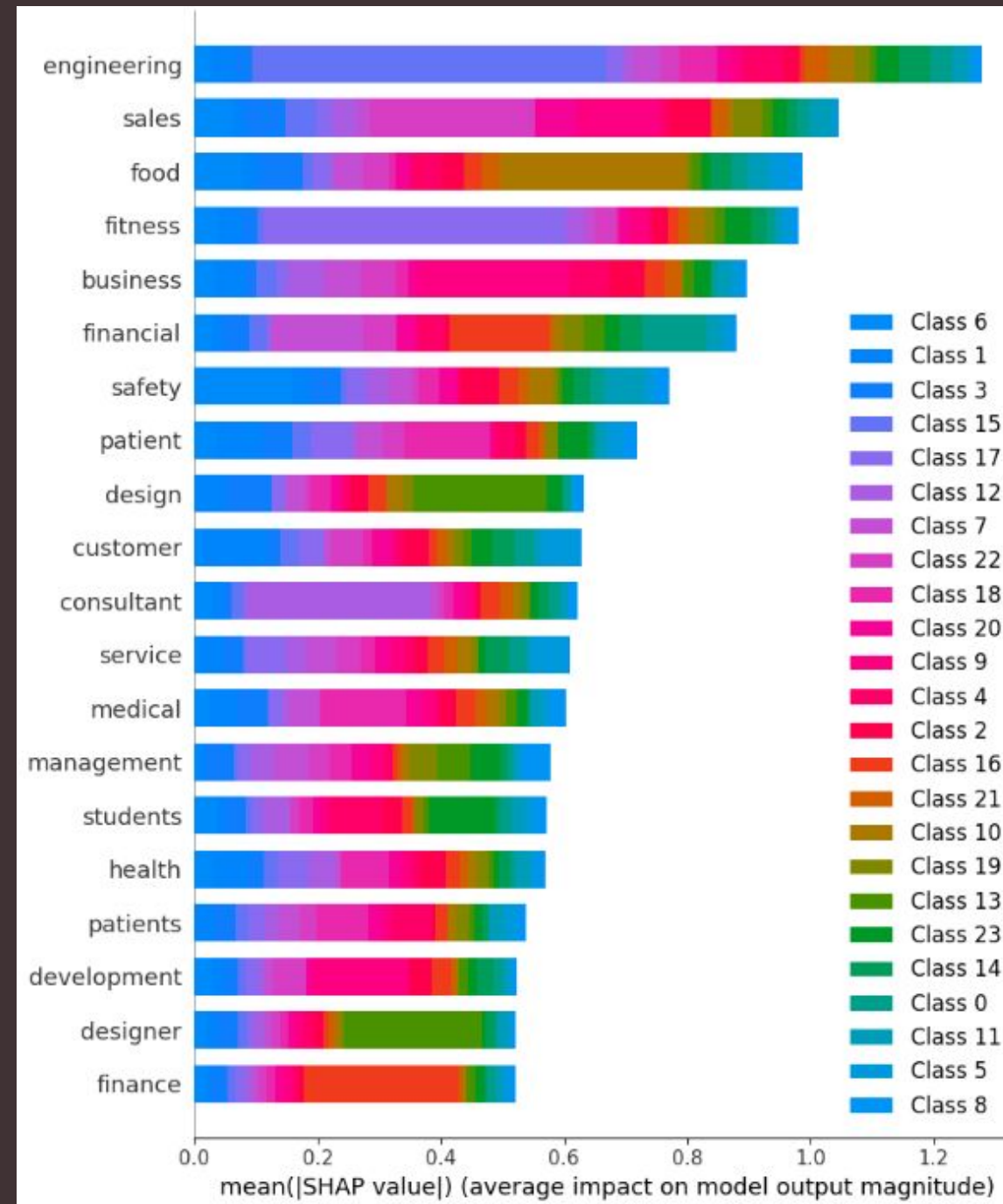
# Future work

To enhance the model's accuracy and fairness, future work may involve several directions:

- **Integrating Deep Learning**: Exploring the use of transformer-based models such as BERT or RoBERTa to capture contextual understanding of resume content.

- **Bias Detection and Mitigation**: Implementing fairness-aware algorithms and conducting regular bias audits to ensure equity across different demographic groups.

- **User Feedback Loop**: Including recruiter feedback in the learning pipeline to continuously refine model predictions based on human evaluation.

- **Multi-label Classification**: Extending the task to predict multiple suitability categories (e.g., job fit, cultural fit) rather than a binary outcome.

- **Deployment and Usability**: Developing an interactive dashboard or API that enables recruiters to upload and screen resumes with interpretable results.

# THANK YOU EVERYONE