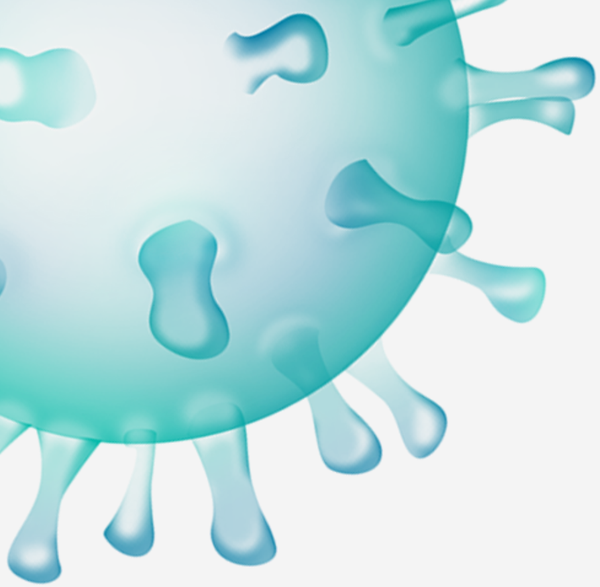
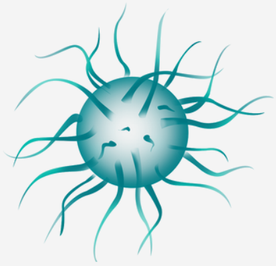
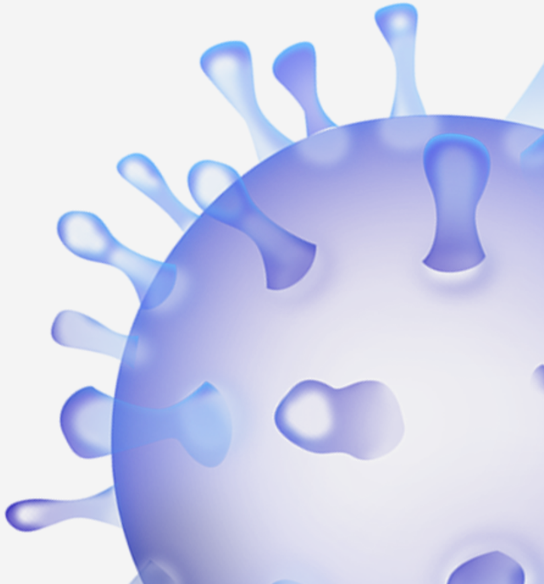



PREDICTING CLINICAL TRIAL TERMINATIONS

By Clement Chan





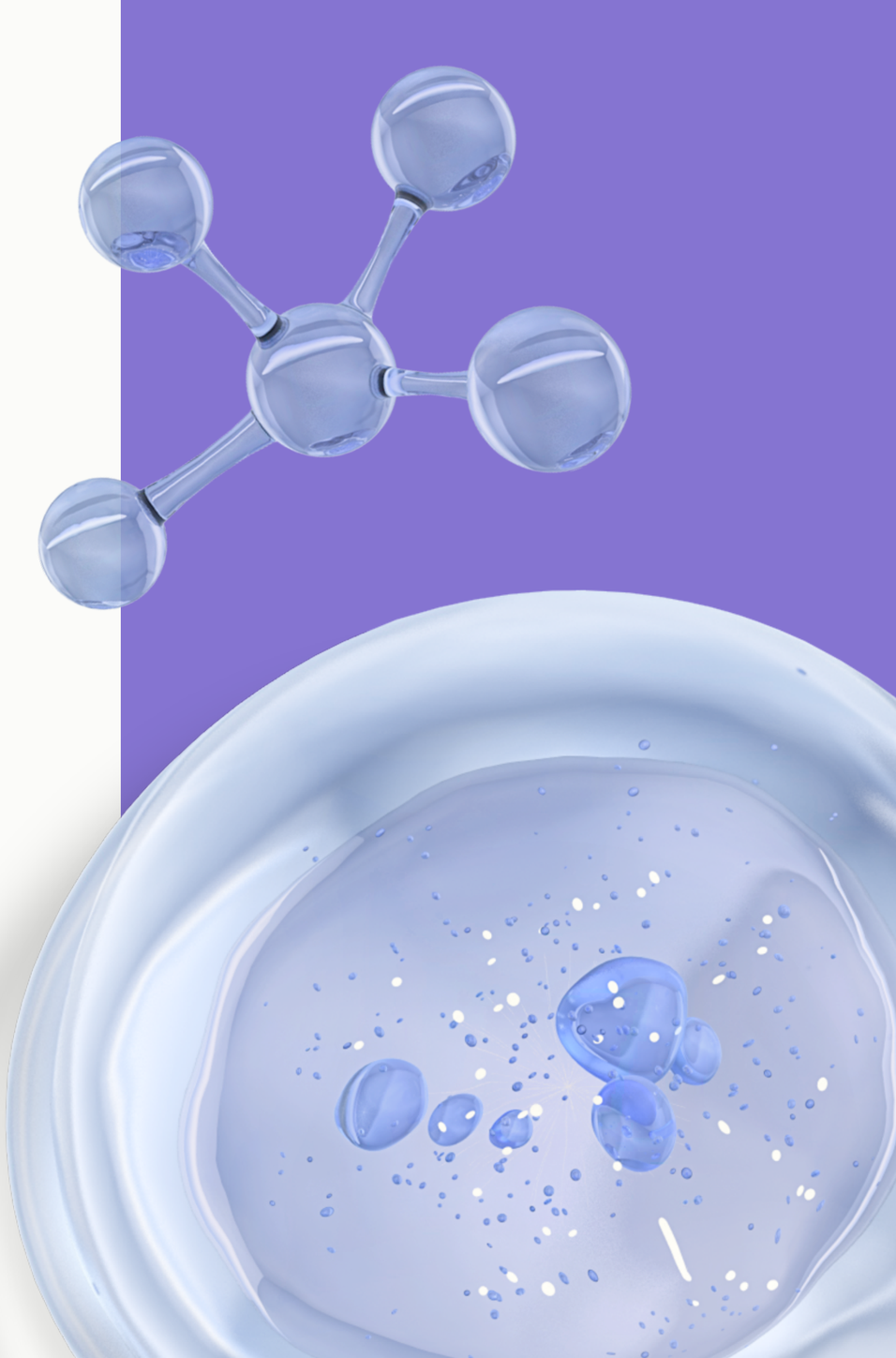
Introduction

What are Clinical Trials?

- Evaluate safety, performance, and effects of drugs
- Ex. Covid Vaccines, diabetes, cancer etc.

What’s the Problem?

- Trials costed 1.4 million - 53 million USD in 2015 and is increasing rapidly.
- Out of 8000 trials, 960 (12%) are terminated.



The Data Solution

1. Determine the main factors or parameters associated to terminated trials.
2. Create a classification model to predict trial terminations.

Results:

1. Save millions of dollars in funding.
2. Design better and more efficient clinical trials.
3. Prevent loss of scientific advancements

Workflow

1

Data Preprocessing

- Analyze data quality, missing values etc.
- Selecting important variables/features

2

Exploratory Data Analysis

- Evaluate distributions of the dataset
- Visualize patterns and explore issues
- Formulate questions and find hidden info

3

Baseline Modelling

- Basic Classification Models
- Logistic Regression, Decision Tree
- Evaluate with Confusion Matrix

4

Future Steps & Advanced Modelling

- Ensemble Learning (Random Forest)
- Word Embedding + neural networks
- Hyperparameter fine tuning

Data Preprocessing

Data Summary

- 482350 rows
- 23 columns

Study Status - Dependent Variable

- 1 = Terminated
- 0 = Completed

Important Features - Independent Variable

- Study duration
- Sponsors, Collaborators, Funder Type
- Age, Sex, Enrollment

Workflow

1

Data Preprocessing

- Analyze data quality, missing values etc.
- Selecting important variables/features

2

Exploratory Data Analysis

- Evaluate distributions of the dataset
- Visualize patterns and explore issues
- Formulate questions and find hidden info

3

Baseline Modelling

- Basic Classification Models
- Logistic Regression, Decision Tree
- Evaluate with Confusion Matrix

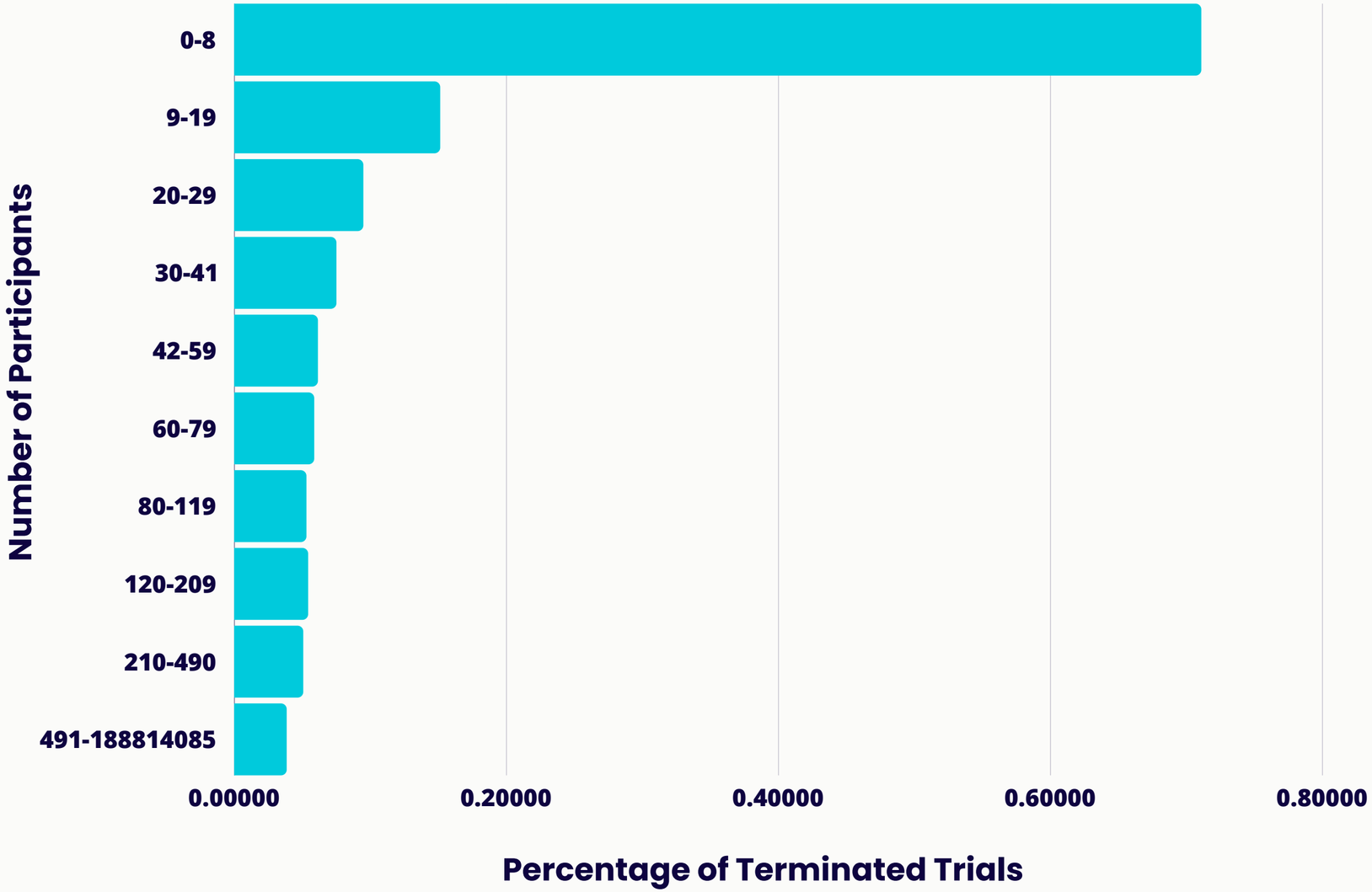
4

Future Steps & Advanced Modelling

- Ensemble Learning (Random Forest)
- Word Embedding + neural networks
- Hyperparameter fine tuning

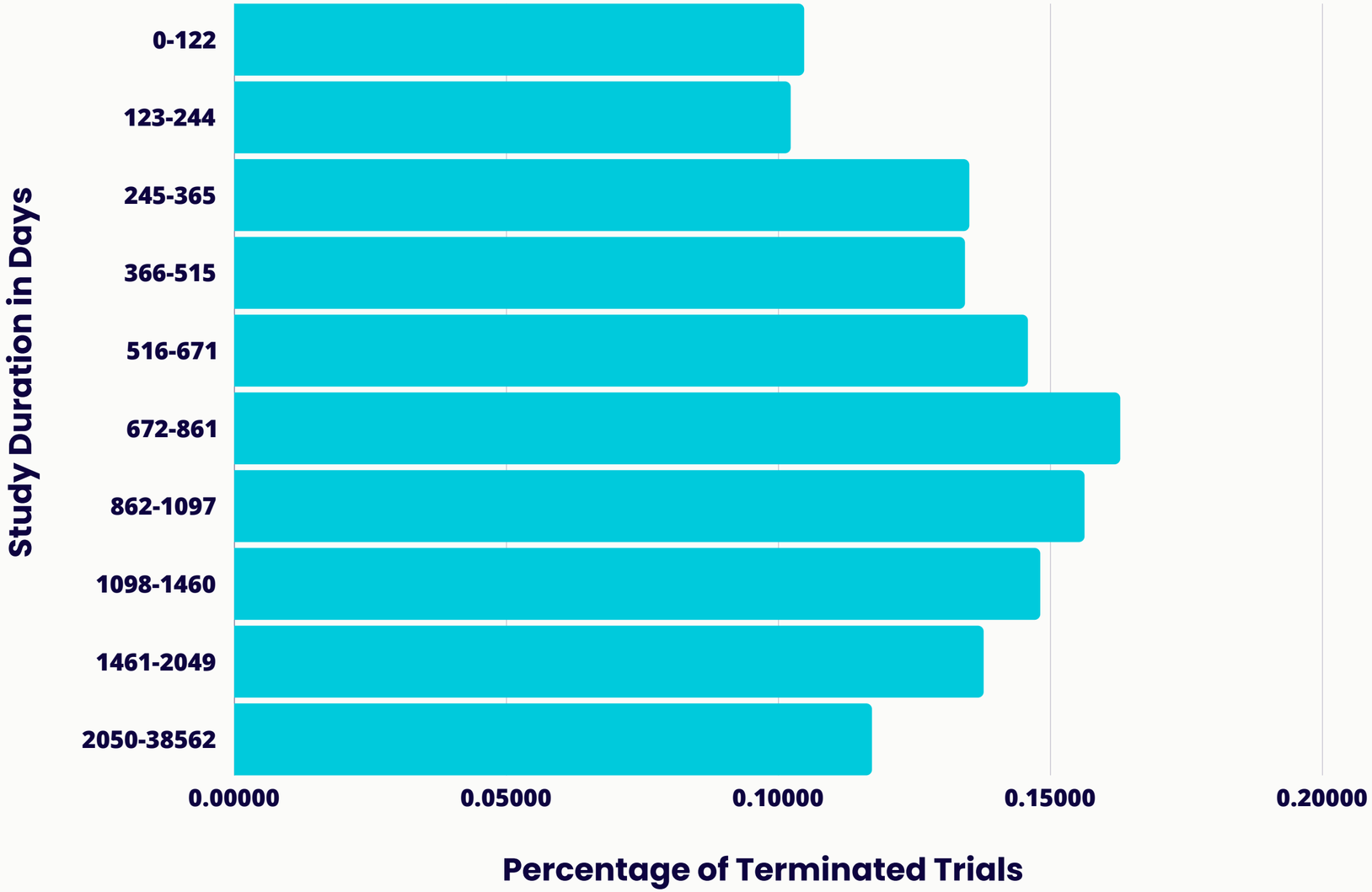
Exploratory Data Analysis

Comparing Terminated Trials With Number of Participants



Exploratory Data Analysis

Comparing Terminated Trials With Study Length



Workflow

1

Data Preprocessing

- Analyze data quality, missing values etc.
- Selecting important variables/features

2

Exploratory Data Analysis

- Evaluate distributions of the dataset
- Visualize patterns and explore issues
- Formulate questions and find hidden info

3

Baseline Modelling

- Basic Classification Models
- Logistic Regression, Decision Tree
- Evaluate with Confusion Matrix

4

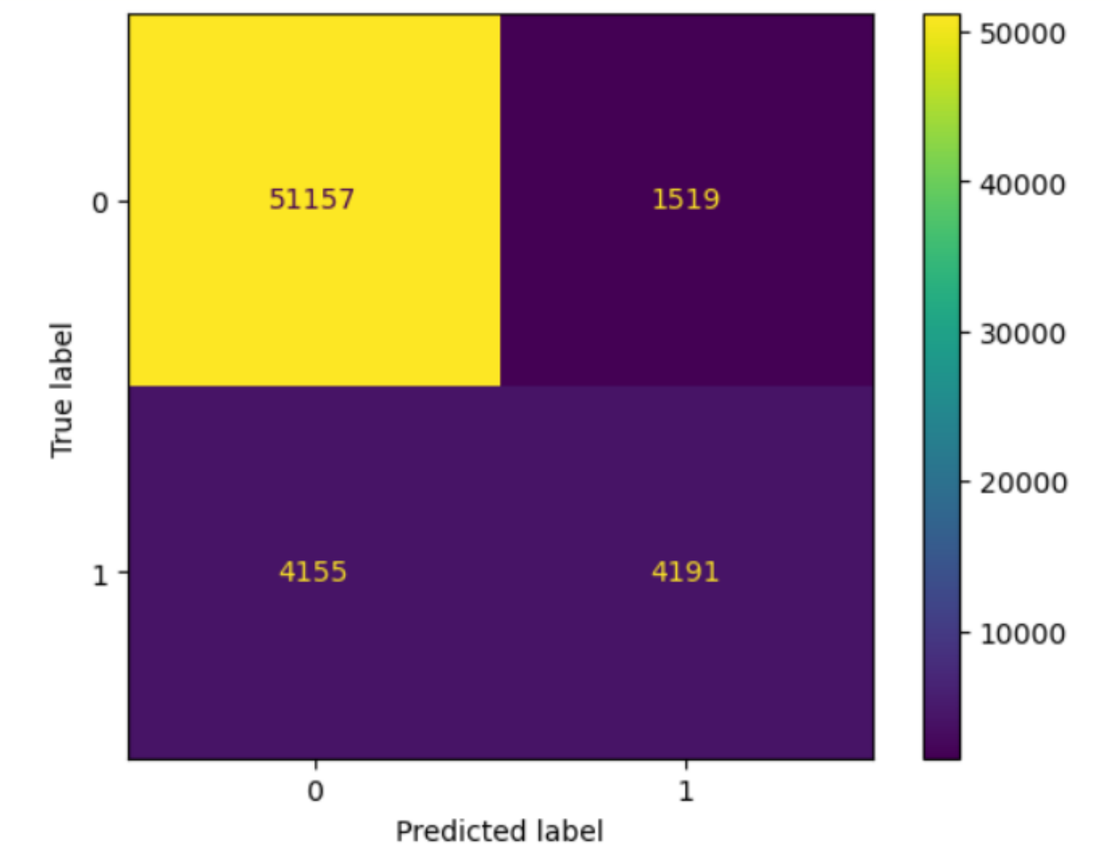
Future Steps & Advanced Modelling

- Ensemble Learning (Random Forest)
- Word Embedding + neural networks
- Hyperparameter fine tuning

Baseline Model - Tabular

Standard Scaler -> Logistic Regression

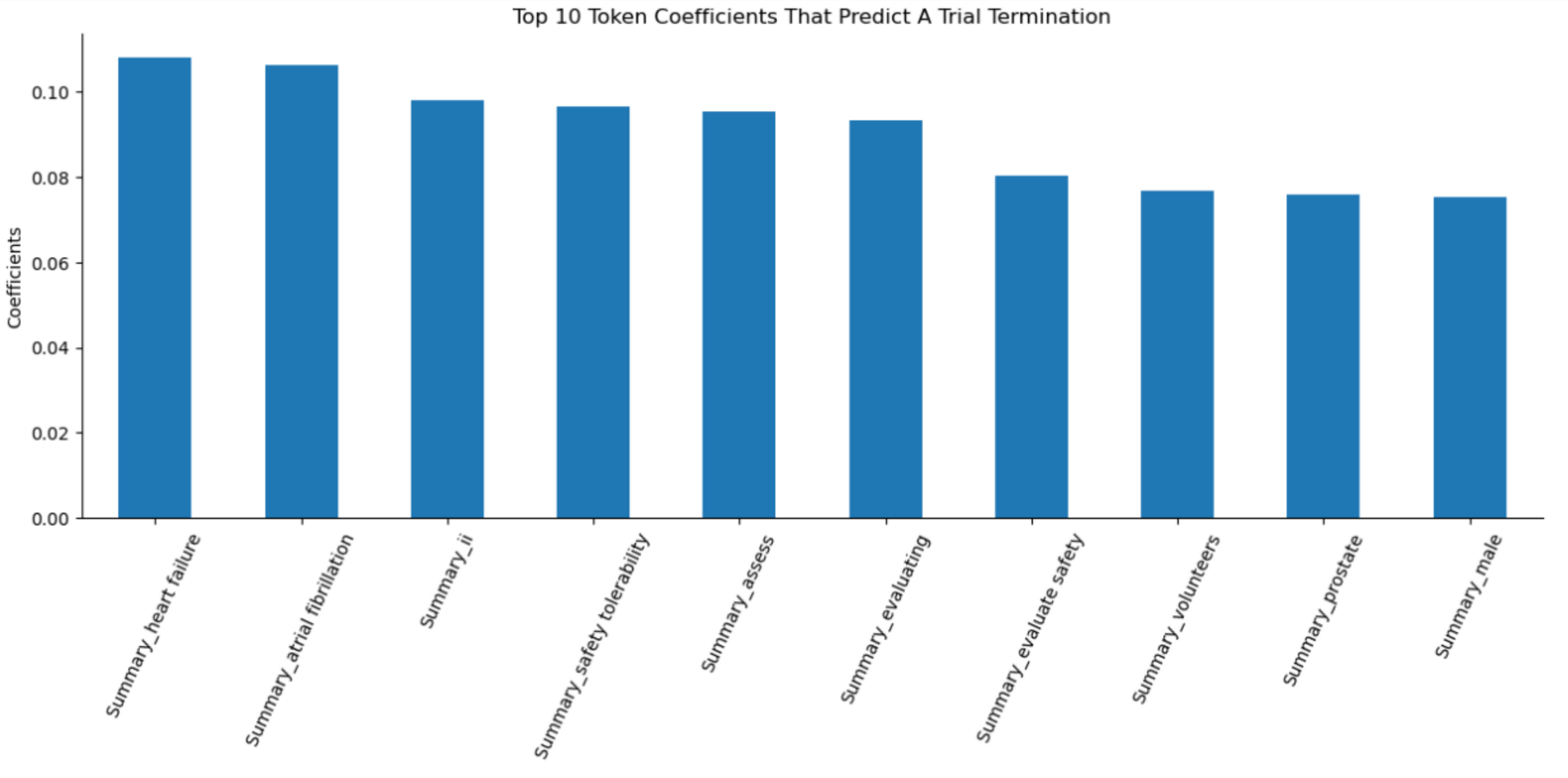
- Training Score: 90.7%
- Test Score: 90.7%



Classification Report

	f1 - score	support
0	0.95	52676
1	0.60	8346

Text Preprocessing



Workflow

1

Data Preprocessing

- Analyze data quality, missing values etc.
- Selecting important variables/features

2

Exploratory Data Analysis

- Evaluate distributions of the dataset
- Visualize patterns and explore issues
- Formulate questions and find hidden info

3

Baseline Modelling

- Basic Classification Models
- Logistic Regression, Decision Tree
- Evaluate with Confusion Matrix

4

Future Steps & Advanced Modelling

- Ensemble Learning (Random Forest)
- Word Embedding + neural networks
- Hyperparameter fine tuning

Next Steps

Data Imbalance

- The recall and f1-scores are bad.
- Perform random desampling and combine with ensemble learning.

Advanced Modelling

- Random Forest Classification
- NLP Word Embedding + Neural Networks
- Hyperparameter finetuning

THANKS FOR LISTENING