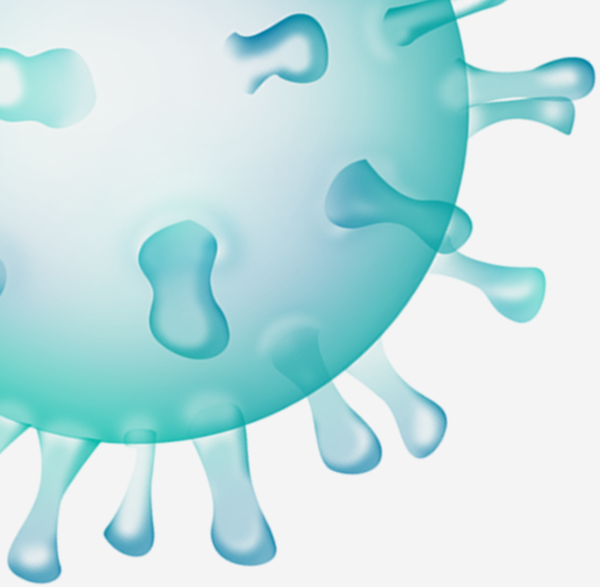
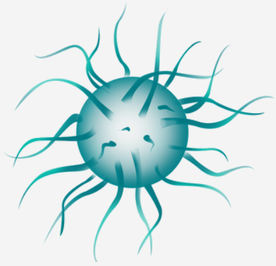
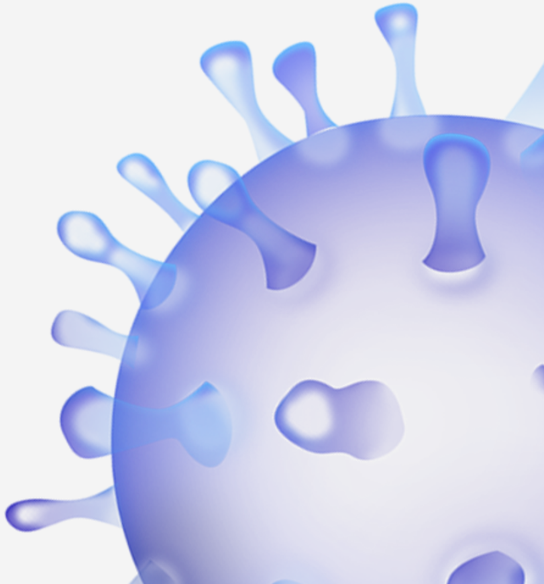




# PREDICTING CLINICAL TRIAL TERMINATIONS

By Clement Chan



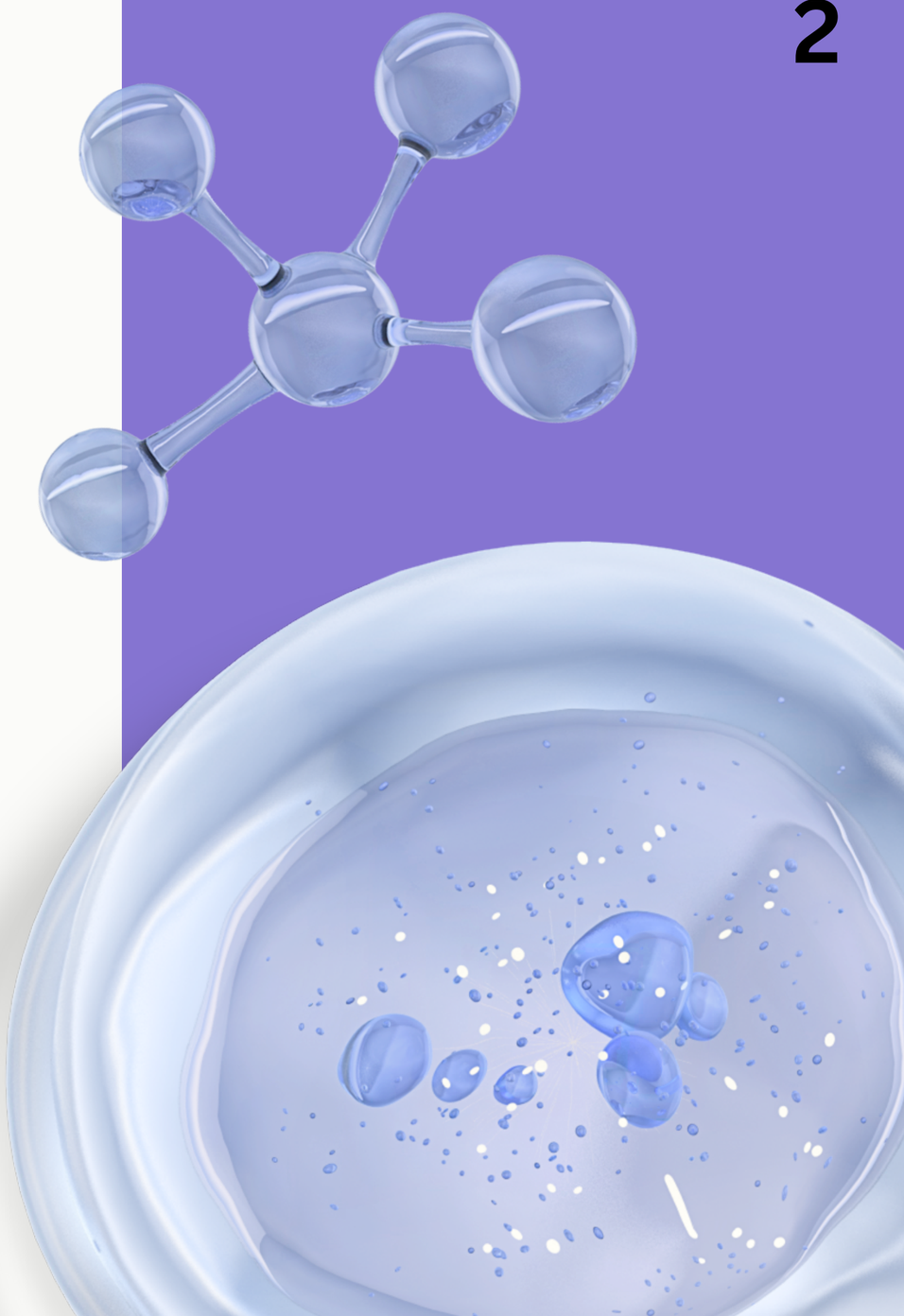
# Introduction

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 up to 50 million dollars per trial 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

**Recap Baseline Model & EDA**

- Logistic Regression Classifier
- Evaluate Confusion Matrix

3

**Advanced Modelling**

- Ensemble Learning
- RandomForest Classification
- Word Embedding Bio-ClinicalBERT

4

**Product Demo**

- Streamlit Application
- Drop Down Menus + Visualizations
- Textbox Classification

# Data Preprocessing

**Data Summary**

- ~480000 rows (ClinicalTrials.gov)
- 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

**Recap Baseline Model & EDA**

- Logistic Regression Classifier
- Evaluate Confusion Matrix

3

**Advanced Modelling**

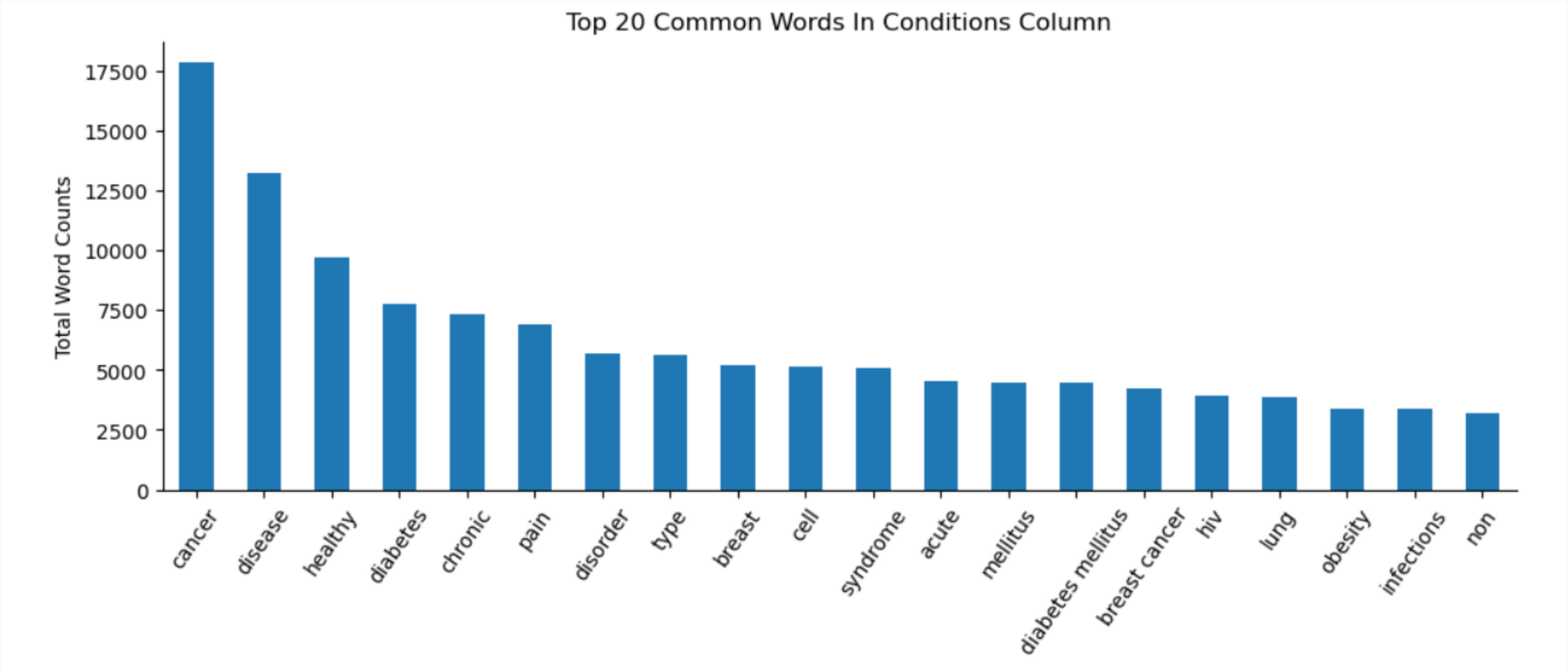
- Ensemble Learning
- RandomForest Classification
- Word Embedding Bio-ClinicalBERT

4

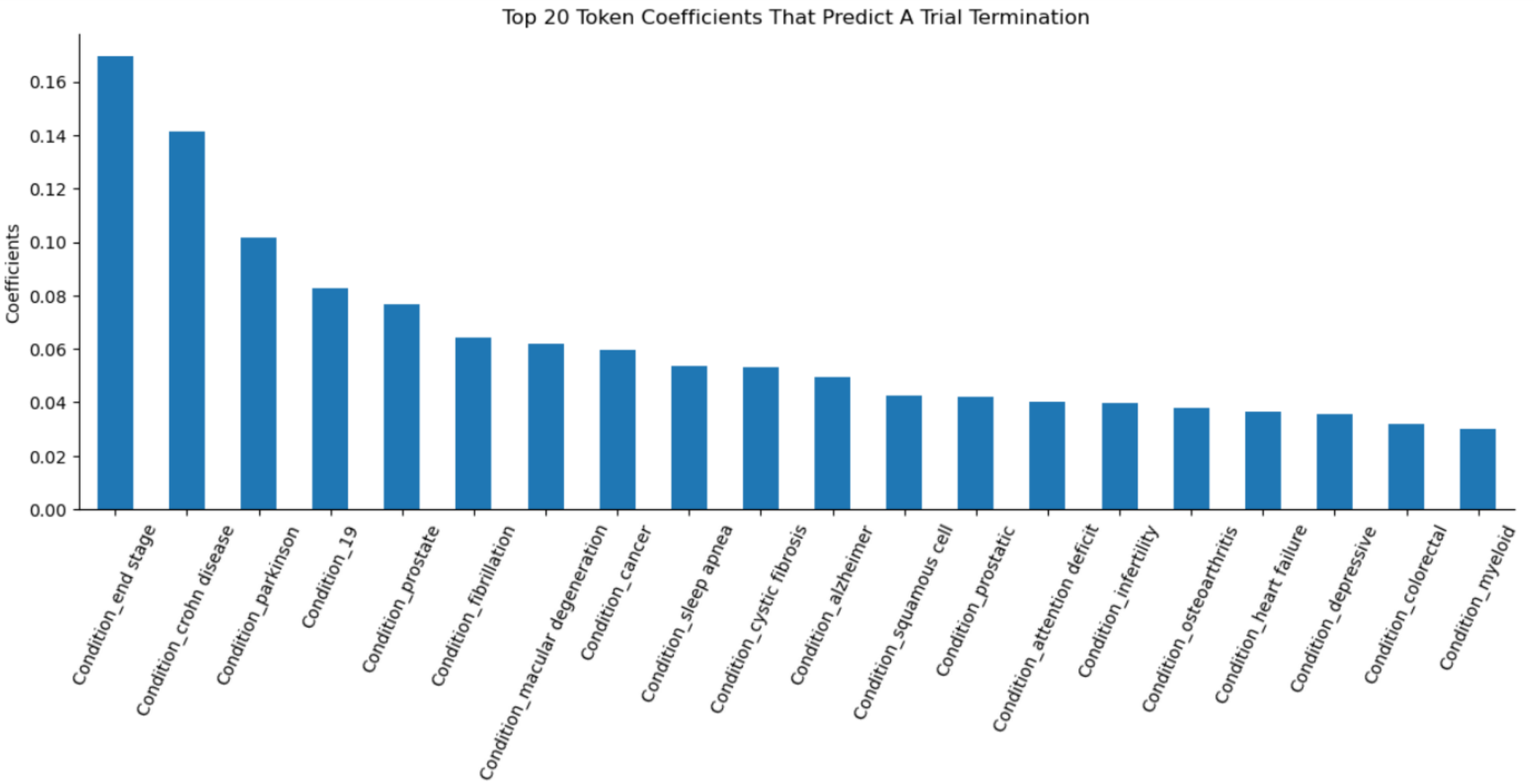
**Product Demo**

- Streamlit Application
- Drop Down Menus + Visualizations
- Textbox Classification

# Exploratory Data Analysis



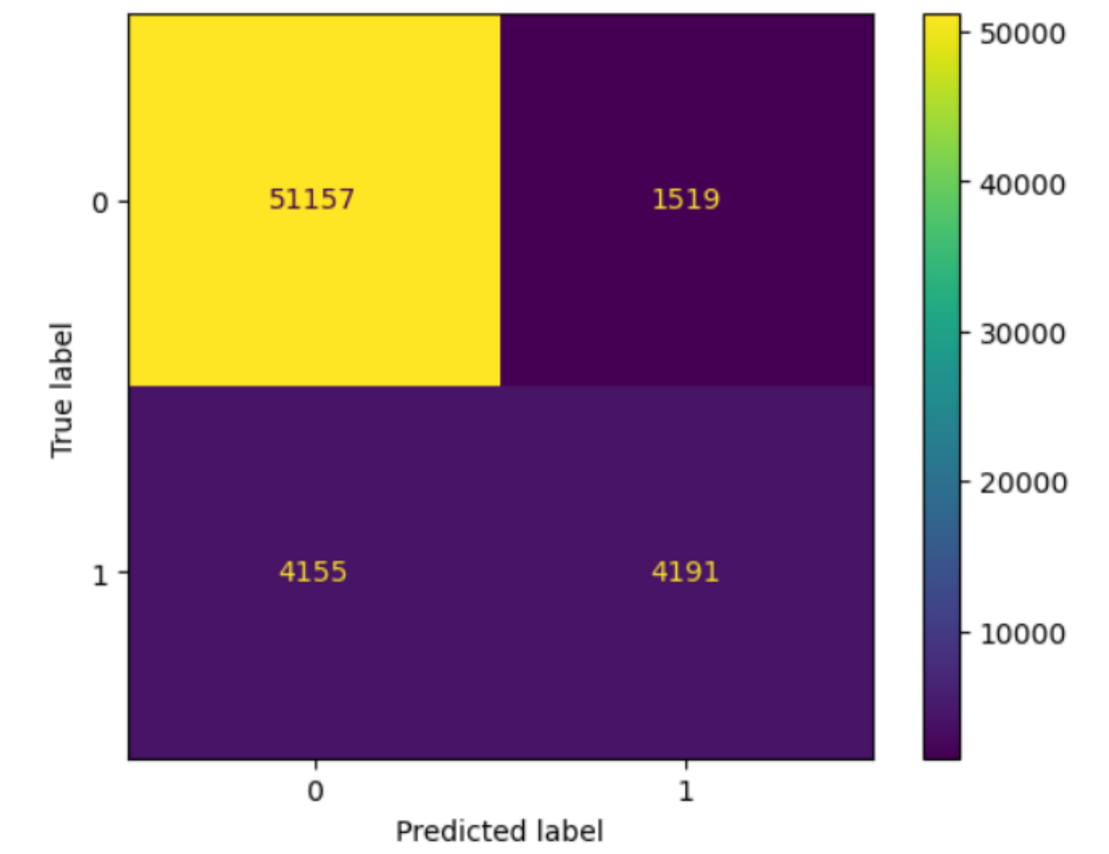
# Exploratory Data Analysis



# Baseline Model - Recap

**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 |

# Workflow

1

**Data Preprocessing**

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

2

**Recap Baseline Model & EDA**

- Logistic Regression Classifier
- Evaluate Confusion Matrix

3

**Advanced Modelling**

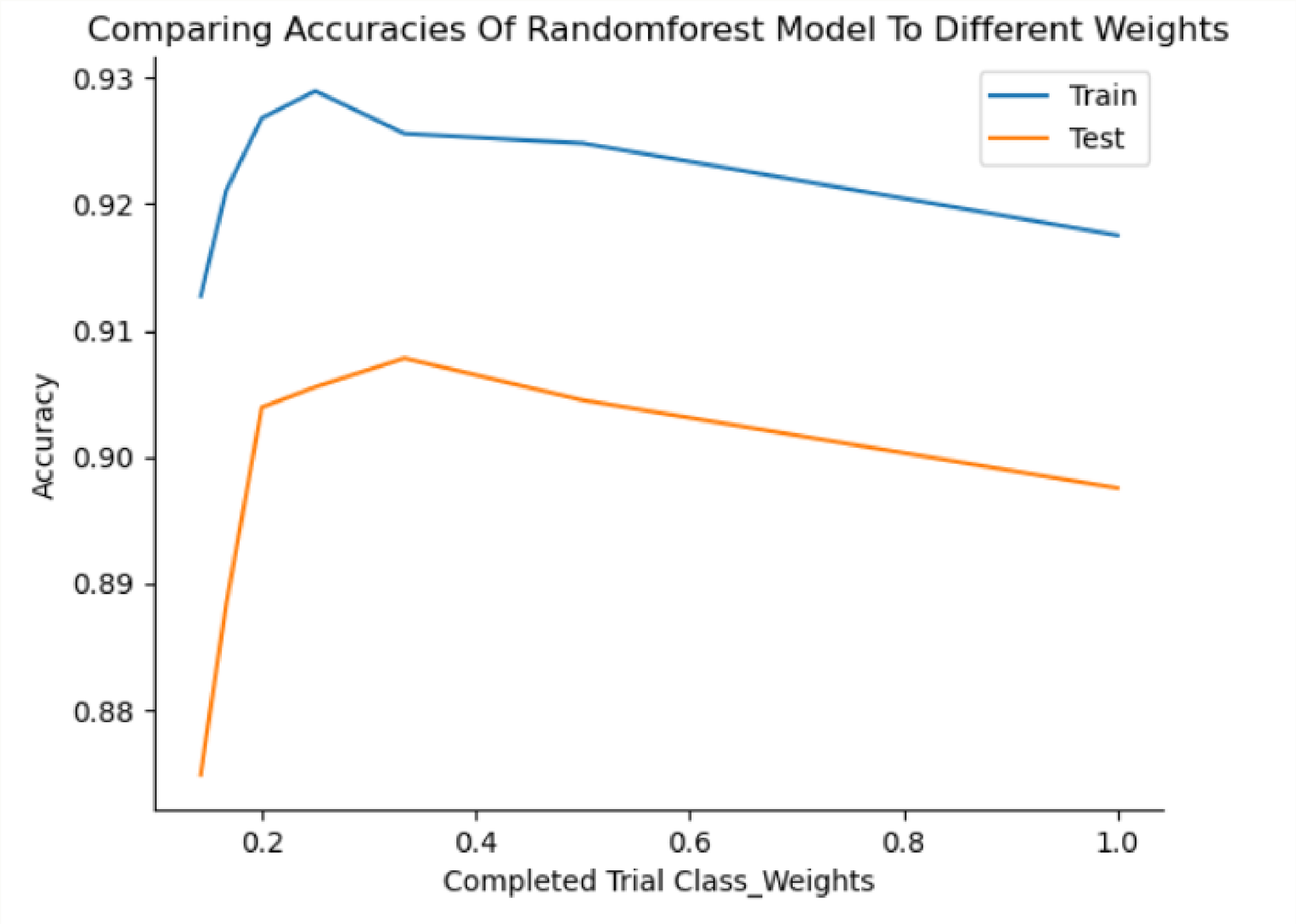
- Ensemble Learning
- RandomForest Classification
- Word Embedding Bio-ClinicalBERT

4

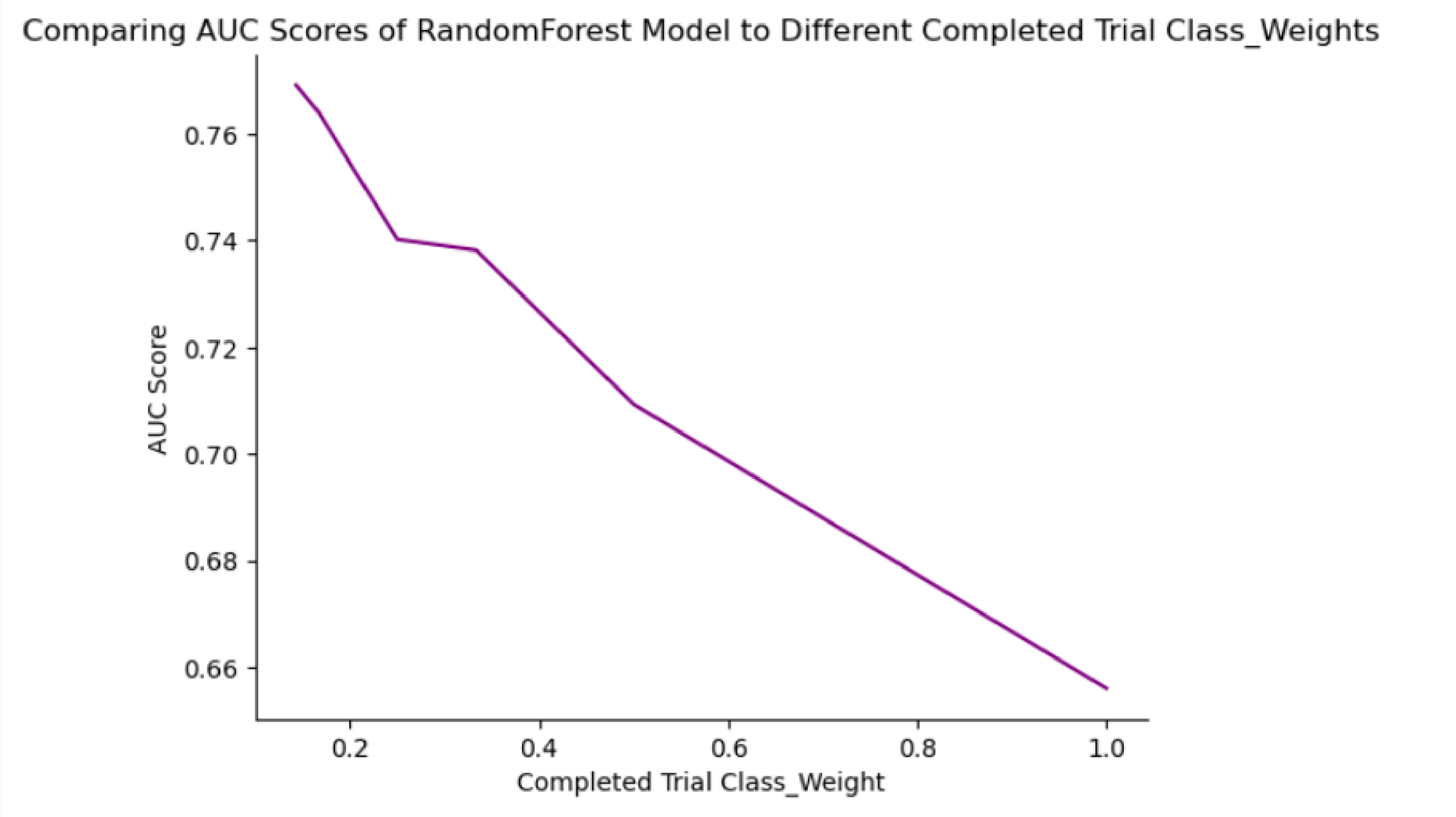
**Product Demo**

- Streamlit Application
- Drop Down Menus + Visualizations
- Textbox Classification

# Advanced Modelling - Ensemble



# Advanced Modelling - Ensemble



# Model Tuning

| Model | f1-Score | AUC |
|---|---|---|
| Decision Tree | 0.49 | 0.70 |
| RandomForest | 0.61 | 0.75 |
| Hyper Parameter RandomForest | 0.61 | 0.77 |

# Advanced Modelling
# Word Embedding

**Bio_ClinicalBERT**

- Bidirectional Encoder Representations from Transformers
- Trained on MIMIC III

# Workflow

1

**Data Preprocessing**

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

2

**Recap Baseline Model & EDA**

- Logistic Regression Classifier
- Evaluate Confusion Matrix

3

**Advanced Modelling**

- Ensemble Learning
- RandomForest Classification
- Word Embedding Bio-ClinicalBERT

4

**Product Demo**

- Streamlit Application
- Drop Down Menus + Visualizations
- Textbox Classification

# Streamlit App

## *Drop Down Menu*

Example:

- Age
- Phases
- Enrollment
- Country

## *Text Box Classification*

Example:

- Study Title
- Brief Description
- Condition

**1 = Terminated, 0 = Completed**

# THANKS FOR LISTENING