

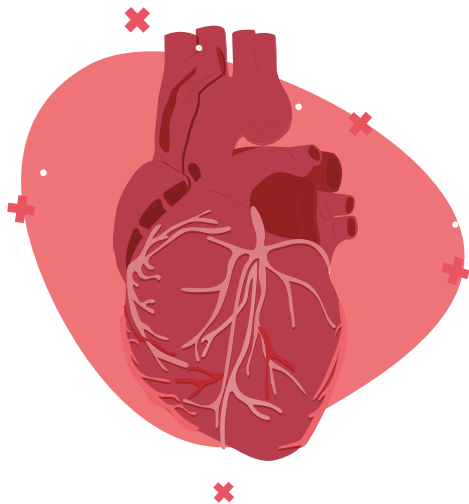
Classifying Heart Disease Status using Clinical and Demographic Data



Pranit Yadav

Context and Background

Introduction



Heart disease is the **leading cause of death worldwide**

Responsible for about **18 million** deaths annually



The Question:

Based on a patient's demographics and clinical data, can we reliably predict if they have heart disease?

Dataset Overview

Why This Dataset?



**Dynamic
Demographic**



**Reputable &
Rich Clinical
Data**



**Access to Multi
Factor Analysis**

Source

This dataset represents the largest available resource for heart disease research, created by integrating five well-known datasets.

Dataset Composition

The final dataset was compiled from the following sources:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Statlog (Heart): 270 observations

Pre-Cleaned Data

- Initial Combined Records: 1,190
- Identified & Eliminated Duplicates: 272
- Final Unique Records: 918
- Missing Values: 0

Variables

Explanatory Variables (Predictors)

- **Age** – Patient age (years)
- **Sex** – Male (M) / Female (F)
- **ChestPainType** – TA (Typical Angina), ATA (Atypical Angina), NAP (Non-Anginal Pain), ASY (Asymptomatic)
- **RestingBP** – Resting blood pressure (mm Hg)
- **Cholesterol** – Serum cholesterol (mg/dl)
- **FastingBS** – Fasting blood sugar > 120 mg/dl (1 = Yes, 0 = No)
- **RestingECG** – Normal, ST (ST-T wave abnormality), LVH (ventricular hypertrophy)
- **MaxHR** – Maximum heart rate achieved (60–202 bpm)
- **ExerciseAngina** – Exercise-induced angina (Y = Yes, N = No)
- **Oldpeak** – ST depression (numeric value)
- **ST_Slope** – Up (upsloping), Flat, Down (downsloping)



Response Variable (Target)

HeartDisease

[1: Heart Disease, 0: Normal]

Dataframe Info & Summary Statistics

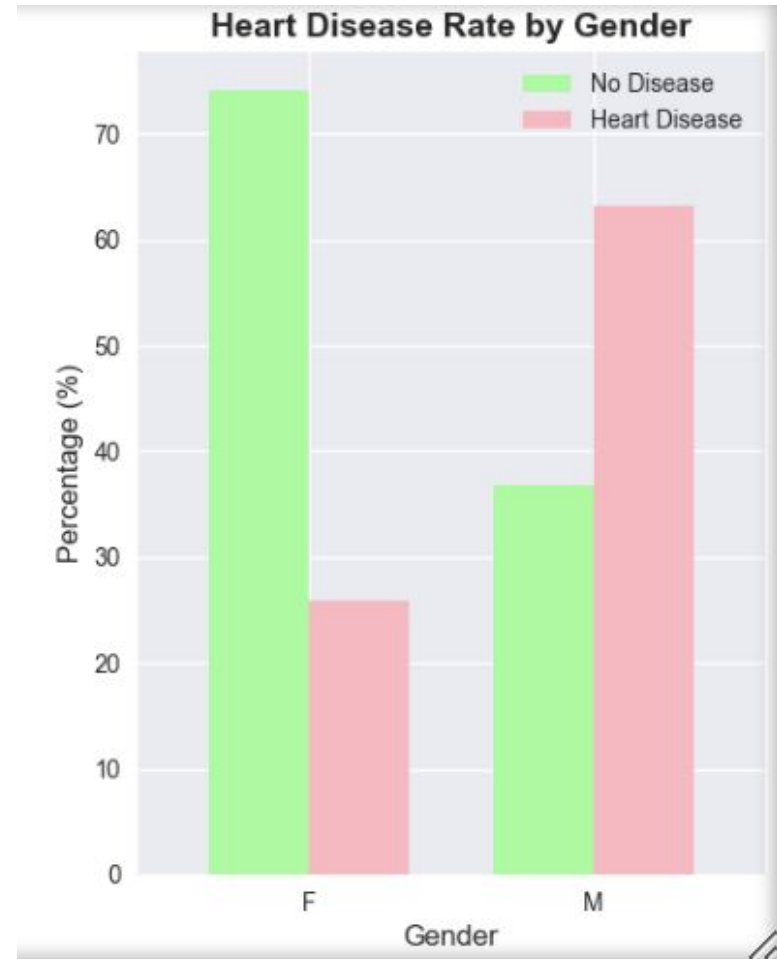
```
----- Dataframe Info -----
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype  
---  --
 0   Age             918 non-null   int64  
 1   Sex             918 non-null   object  
 2   ChestPainType   918 non-null   object  
 3   RestingBP       918 non-null   int64  
 4   Cholesterol     918 non-null   int64  
 5   FastingBS       918 non-null   int64  
 6   RestingECG      918 non-null   object  
 7   MaxHR           918 non-null   int64  
 8   ExerciseAngina  918 non-null   object  
 9   Oldpeak         918 non-null   float64 
10   ST_Slope        918 non-null   object  
11   HeartDisease    918 non-null   int64  
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
None

----- Dimensions (rows, columns) -----
(918, 12)
```

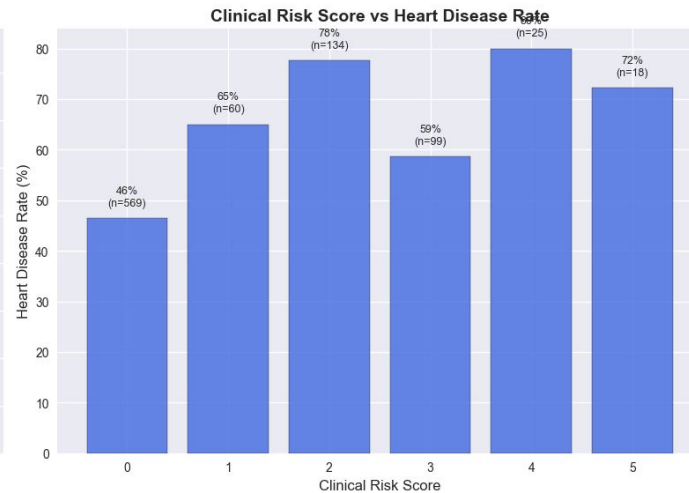
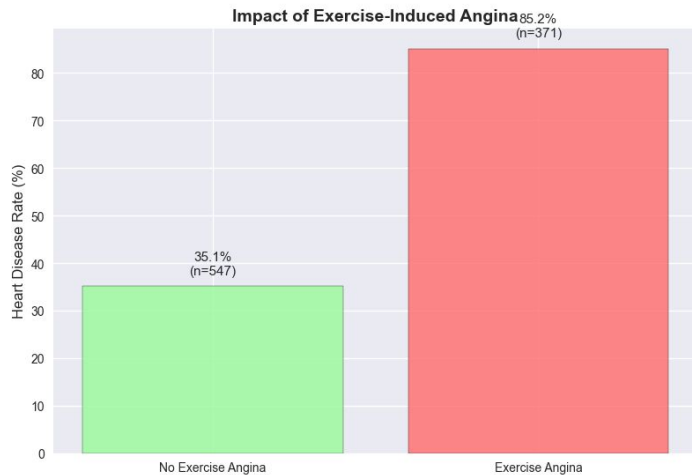
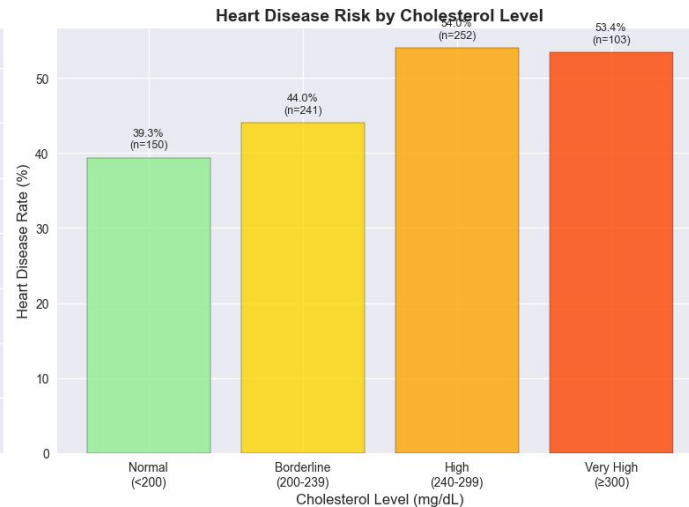
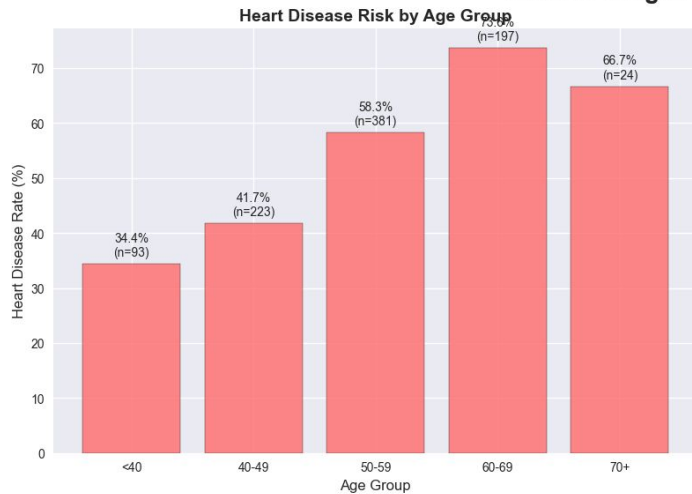
	Heart Disease	No Heart Disease
Age	55.90	Age 50.55
RestingBP	134.19	RestingBP 130.18
Cholesterol	175.94	Cholesterol 227.12
FastingBS	0.33	FastingBS 0.11
MaxHR	127.66	MaxHR 148.15
Oldpeak	1.27	Oldpeak 0.41
HeartDisease	1.00	HeartDisease 0.00
	mean	mean

Categorical Variables

Variable	Category	Percentage
Sex	M	79.0%
Sex	F	21.0%
ChestPainType	ASY	54.0%
ChestPainType	NAP	22.1%
ChestPainType	ATA	18.8%
ChestPainType	TA	5.0%
RestingECG	Normal	60.1%
RestingECG	LVH	20.5%
RestingECG	ST	19.4%
ExerciseAngina	N	59.6%
ExerciseAngina	Y	40.4%
ST_Slope	Flat	50.1%
ST_Slope	Up	43.0%
ST_Slope	Down	6.9%



Clinical Insights and Risk Factors



Dataset Preparation for ML Models

Feature Engineering & Preprocessing

Step 1 — Feature Engineering

- Created **new, more informative columns** by converting raw values (e.g., cholesterol, age) into clinical flags and descriptive categories
- Purpose to **boost model accuracy** by making key patterns and relationships easier for the models to learn

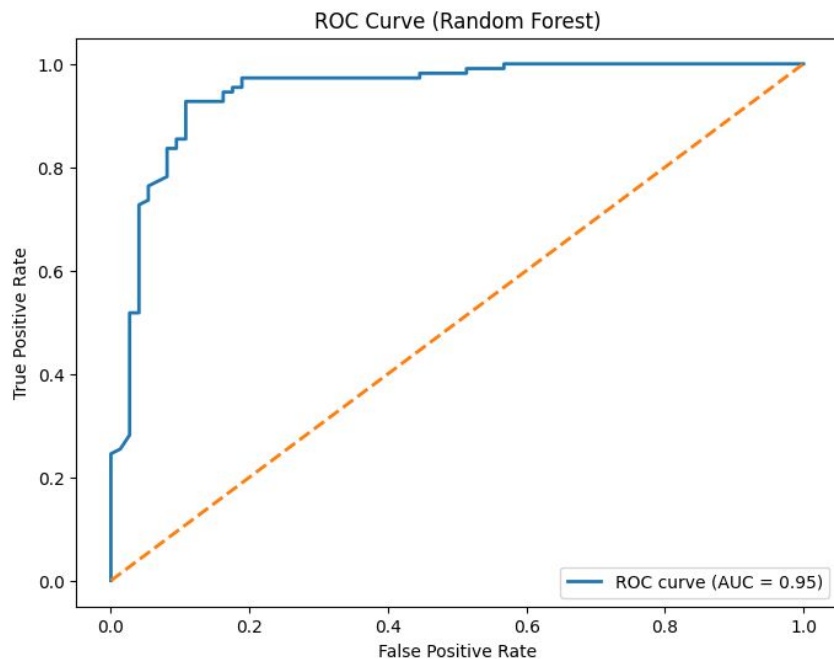
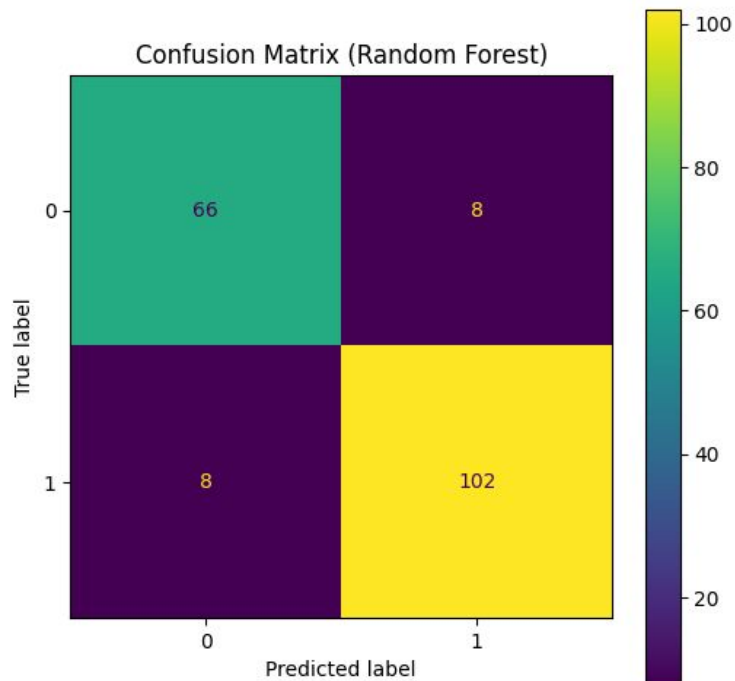
Step 2 — Preprocessing

- Built a pipeline using scikit-learn which encoded categorical variables
- Purpose to ensure data was **totally clean and model-ready**, improving consistency and models' ability to learn

Our Models

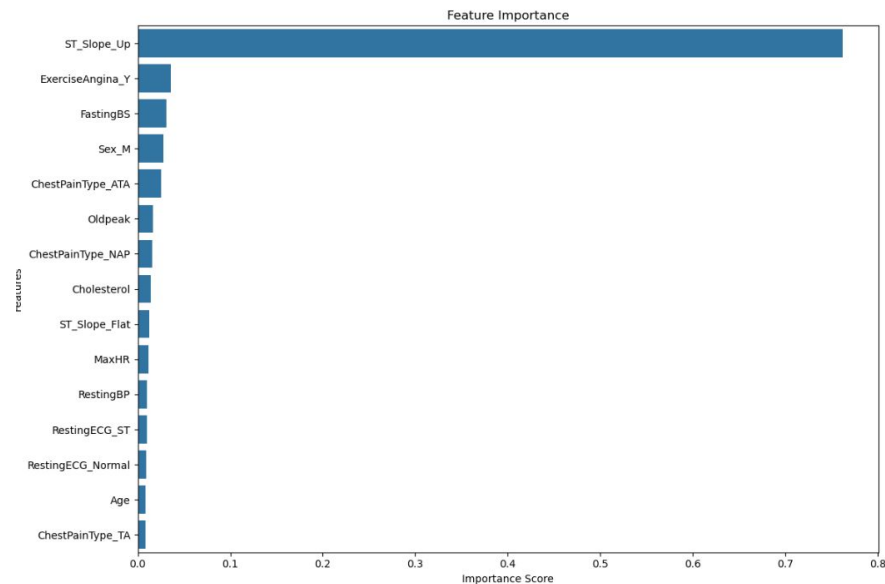
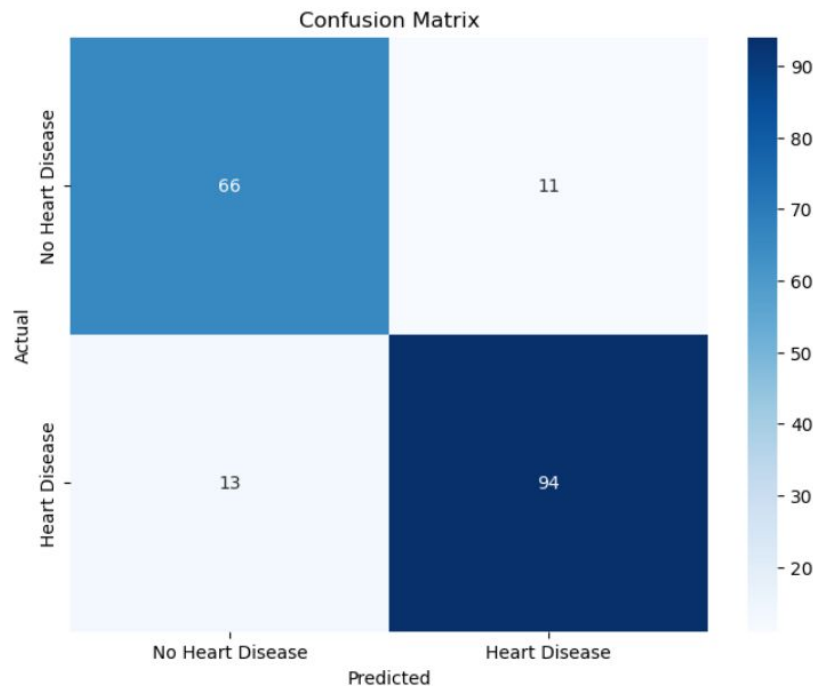
Random Forest Classifier

- **AUC: 0.95**



- **Accuracy: 91.3%**
- **Recall (True Positive Rate): 92.7%**
- **Precision: 92.7%**

Gradient Boosting



- **Accuracy: 87%**
- **Recall (True Positive Rate): 88.04%**
- **Precision: 88.04%**

Ensemble Method

Accuracy: 90.22%

Recall: 90%

AUC: 0.935

Individual Model Test Accuracies:

RandomForest: 84.78%

GradientBoosting: 88.04%

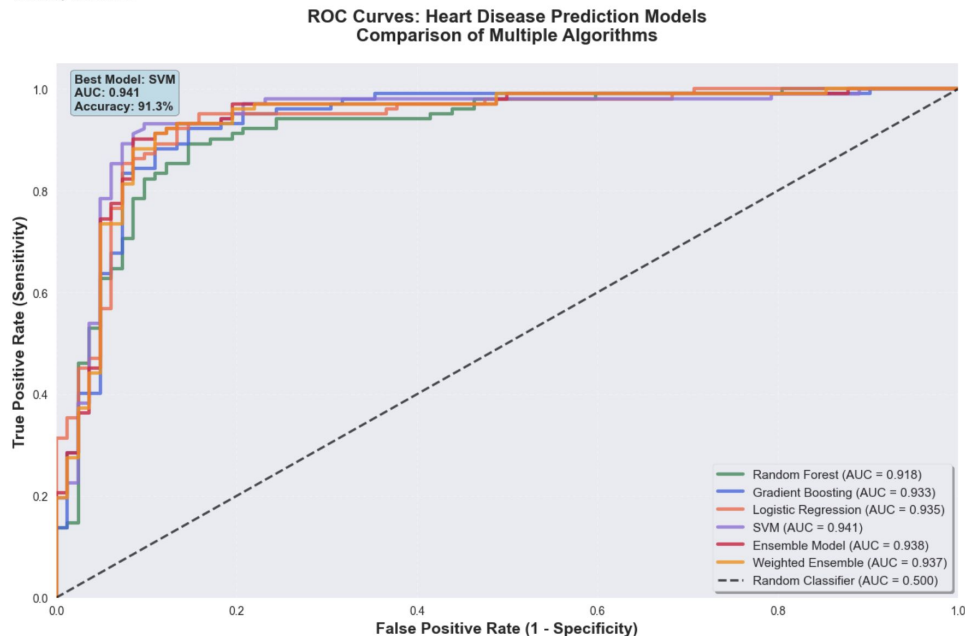
SVM: 89.13%

LogisticRegression: 88.59%

Neural Network: 84.78%

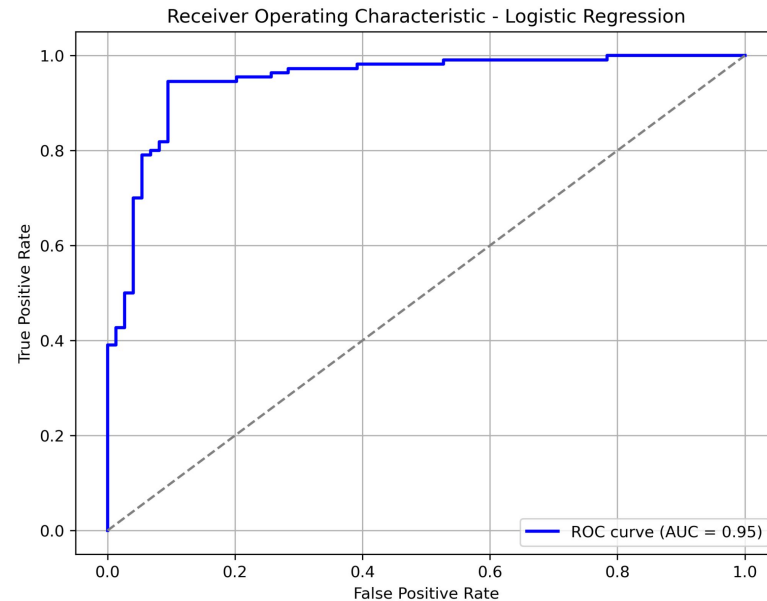
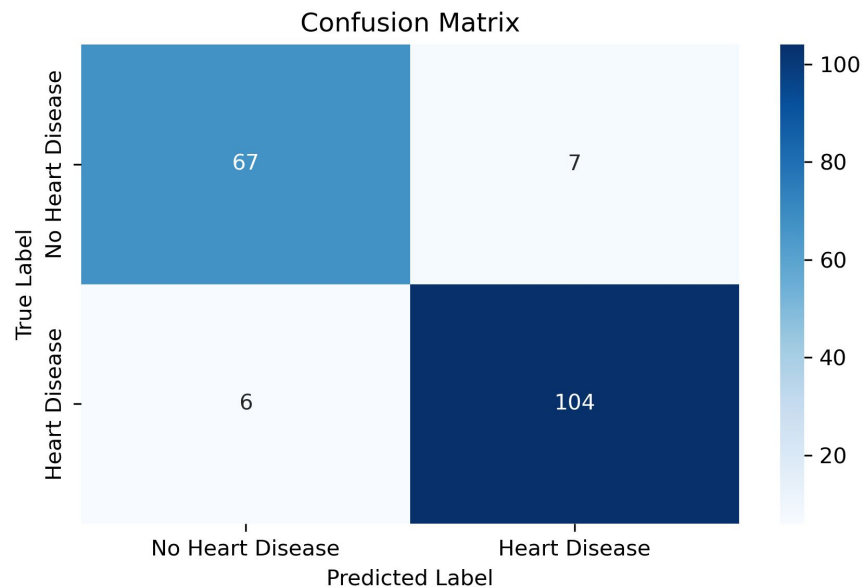
Voting Ensemble: 89.67%

Final Weighted Ensemble: 90.22%



Our Best Model: Logistic Regression

- **AUC: 0.95**



- **Accuracy: 92.9%**
- **Recall (True Positive Rate): 94.5%**
- **Precision: 93.7%**

Conclusion

Real Time Impact



Empowers Preventive Healthcare



Reduces Healthcare Costs



Lowers Hospitalization Rates



Promotes Health Equity

Improvements / Future Work

Hyperparameter Tuning



Cost - Sensitive Learning



External Validation



Deployment Readiness - Real Time



Continuous Learning Pipeline



Thank You
