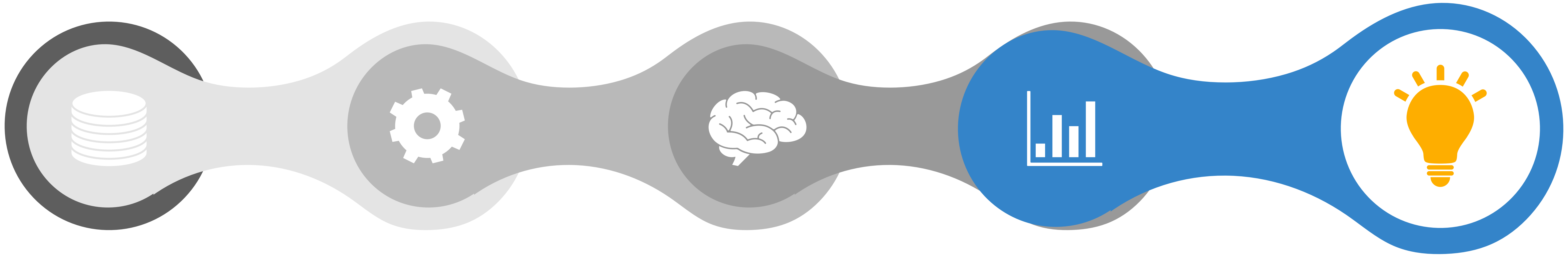
Cancer Prediction

Feature Importance

PRESENTATION OVERVIEW



DATA COLLECTION

Data ingestion and analysis

DATA PROCESSING

Clean, classify, and preprocess data, followed by splitting the data into training and testing to prepare it for model development.

MODEL TRAINING

Train models and fine-tune hyper parameters to optimize performance.

MODEL EVALUATION

Evaluate models based on accuracy, confusion matrix, and AUC score to select the best model

FEATURE IMPORTANCE

Identify the most important features using various feature importance methods.

ML OPS

Building an MLOps solution to ensure continuous training, evaluation, deployment, and monitoring of ML models.

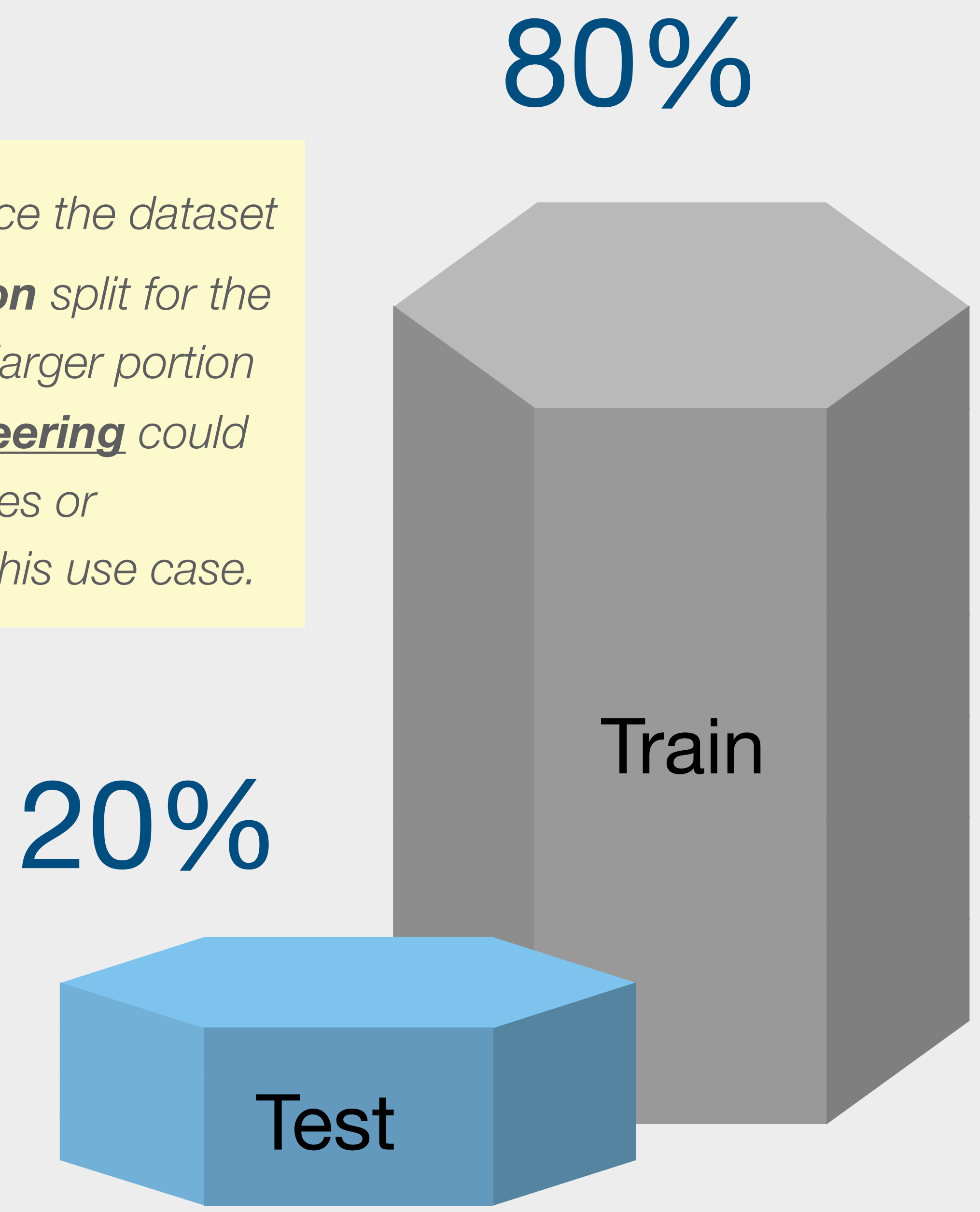
ℹ *As we acquire new data, model results continue to evolve, highlighting the need for an ongoing ML operationalization process. I'll add an extra slide to explain how we can effectively operationalize this to maintain the usefulness of our ML models.*

DATA COLLECTION + PROCESSING

Variable	Description
age	age in years
sex	(1 = male; 0 = female)
cp	chest pain type
trestbps	resting blood pressure (in mm Hg on admission to the hospital)
chol	serum cholestoral in mg/dl
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	resting electrocardiographic results
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment
ca	number of major vessels (0-3) colored by flourosopy
thal	3 = normal; 2 = fixed defect; 1 = reversable defect
target	Flag if the patient has heart disease (1) or not (0)

- ✓ Check Missing Values
- ✓ One Hot Encoding for Categorical Variables
- ✓ Train / Test Split

ℹ *For this case study, I used an 80/20 split since the dataset is relatively small. I'm also skipping the* ***validation*** *split for the same reason, allowing the models to train on a larger portion of the data. (random_state: 42)* ***Feature engineering*** *could also be applied (e.g., creating age bucket features or combining two features), but I'm skipping it for this use case.*



ℹ *For this case study, I experimented with four ML models and fine-tuned their hyperparameters individually to balance performance and prevent overfitting or underfitting.*

LOGISTIC REGRESSION

Logistic Regression predicts probabilities for binary outcomes, making it useful for cancer prediction by assessing risk levels. Its simplicity and interpretability help identify key factors in medical decision-making.

HYPERPARAMETERS

Penalty = l2

C parameter = 0.1

DECISION TREE

Decision Trees are intuitive classification models that split data based on feature values, making them useful for cancer prediction. They provide clear decision rules, helping identify key factors in diagnosis and risk assessment.

HYPERPARAMETERS

Max Depth = 5

Min Samples Leaf = 10

RANDOM FOREST

Random Forest is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. It is useful for cancer prediction as it enhances stability and identifies important features for diagnosis.

HYPERPARAMETERS

Max Depth = 5

Min Samples Leaf = 10

N Estimators = 100

GRADIENT BOOSTING

Gradient Boosting builds decision trees sequentially, correcting errors from previous iterations to improve binary classification. Its ability to capture complex patterns makes it effective for cancer risk assessment and diagnosis.

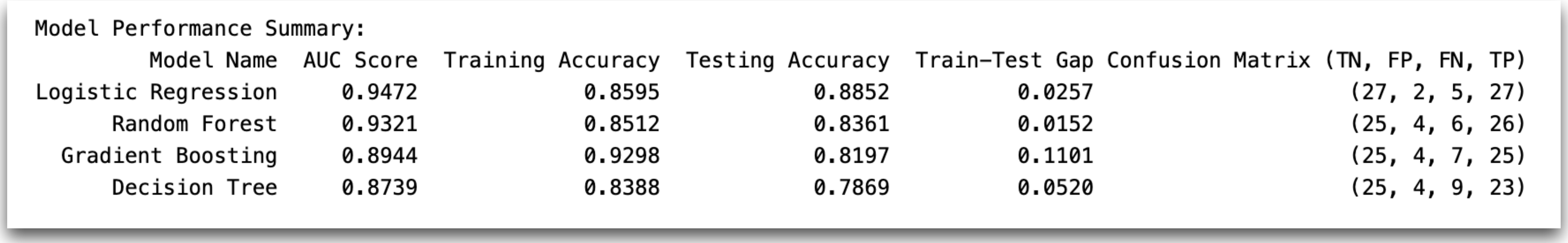
HYPERPARAMETERS

N Estimators = 100

Learning Rate = 0.05

Min Samples Leaf = 5

MODEL **EVALUATION**



```
Model Performance Summary:
         Model Name  AUC Score  Training Accuracy  Testing Accuracy  Train–Test Gap Confusion Matrix (TN, FP, FN, TP)
Logistic Regression     0.9472             0.8595            0.8852          0.0257                   (27, 2, 5, 27)
      Random Forest     0.9321             0.8512            0.8361          0.0152                   (25, 4, 6, 26)
  Gradient Boosting     0.8944             0.9298            0.8197          0.1101                   (25, 4, 7, 25)
      Decision Tree     0.8739             0.8388            0.7869          0.0520                   (25, 4, 9, 23)
```

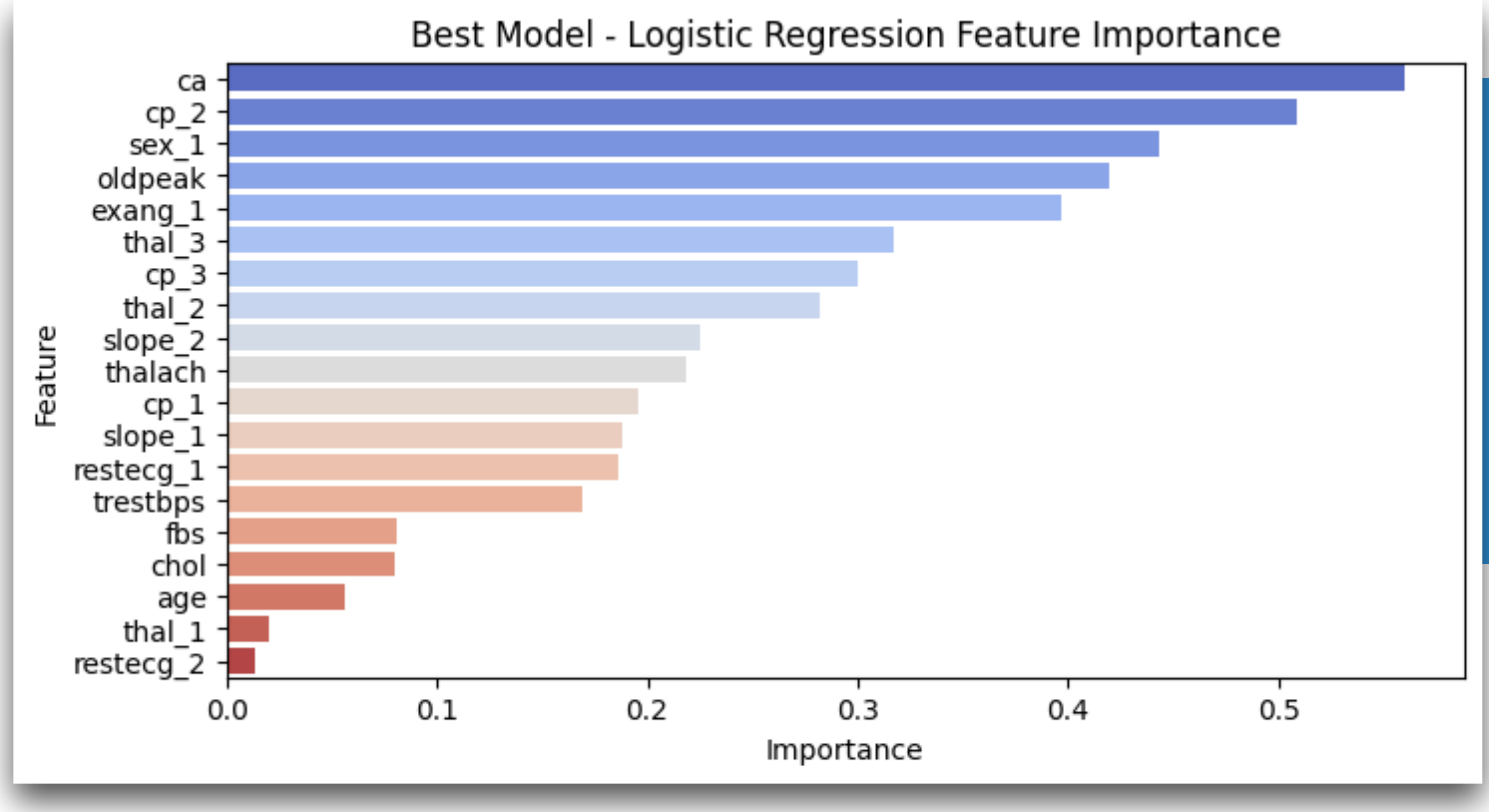
- **Logistic Regression (Chosen Model):** Highest AUC (0.9472) and strong testing accuracy (0.8852) with a low train-test gap (0.0257), ensuring good generalization.
- **Random Forest:** Moderate AUC (0.9321) and slightly lower testing accuracy (0.8361). Lowest train-test gap (0.0152), indicating stability.
- **Gradient Boosting:** Higher training accuracy (0.9298) but lowest testing accuracy (0.8197), suggesting overfitting (train-test gap: 0.1101).
- **Decision Tree:** Lowest AUC (0.8739) and testing accuracy (0.7869). Moderate overfitting with a train-test gap of 0.0520.

ℹ️ ***Chosen Model:*** *Logistic Regression was chosen for its best AUC, strong accuracy, and balanced generalization.*

Model Comparison: *I ensured fairness during model training by applying consistent preprocessing steps and tuning hyperparameters for each model to optimize performance while preventing overfitting for majority of the models but to ensure* ***fair model comparison*** *further, I can test more hyperparameters adjustments. Additionally, adjusting feature engineering and using cross-validation can improve reliability.*

FEATURE IMPORTANCE

ℹ️ *Importance here is the absolute value of the coefficients*



These numbers represent the ***absolute magnitude*** *of the model's coefficients to rank features by impact rather than effect direction. I will be sharing the raw coefficients to show the directional impact in the next slide.*

ℹ️ *The most impactful features are vascular health, chest pain type, exercise response, and ECG abnormalities, while age, cholesterol, and blood pressure have comparatively less impact in this model. and using cross-validation can improve reliability.*

Top Features by Absolute Magnitude:

- **ca** (Major Vessels Colored by Fluoroscopy)
- **cp_2** (Chest Pain Type - Atypical Angina)
- **sex_1** (Male Gender)
- **oldpeak** (ST Depression Induced by Exercise)
- **exang_1** (Exercise-Induced Angina)

MODEL FEATURE COEFFICIENTS

What are the top features that contributes to Heart Disease?

cp_2 (0.509)

- If a person has chest pain type 2, their log-odds of having heart disease increase by 0.51

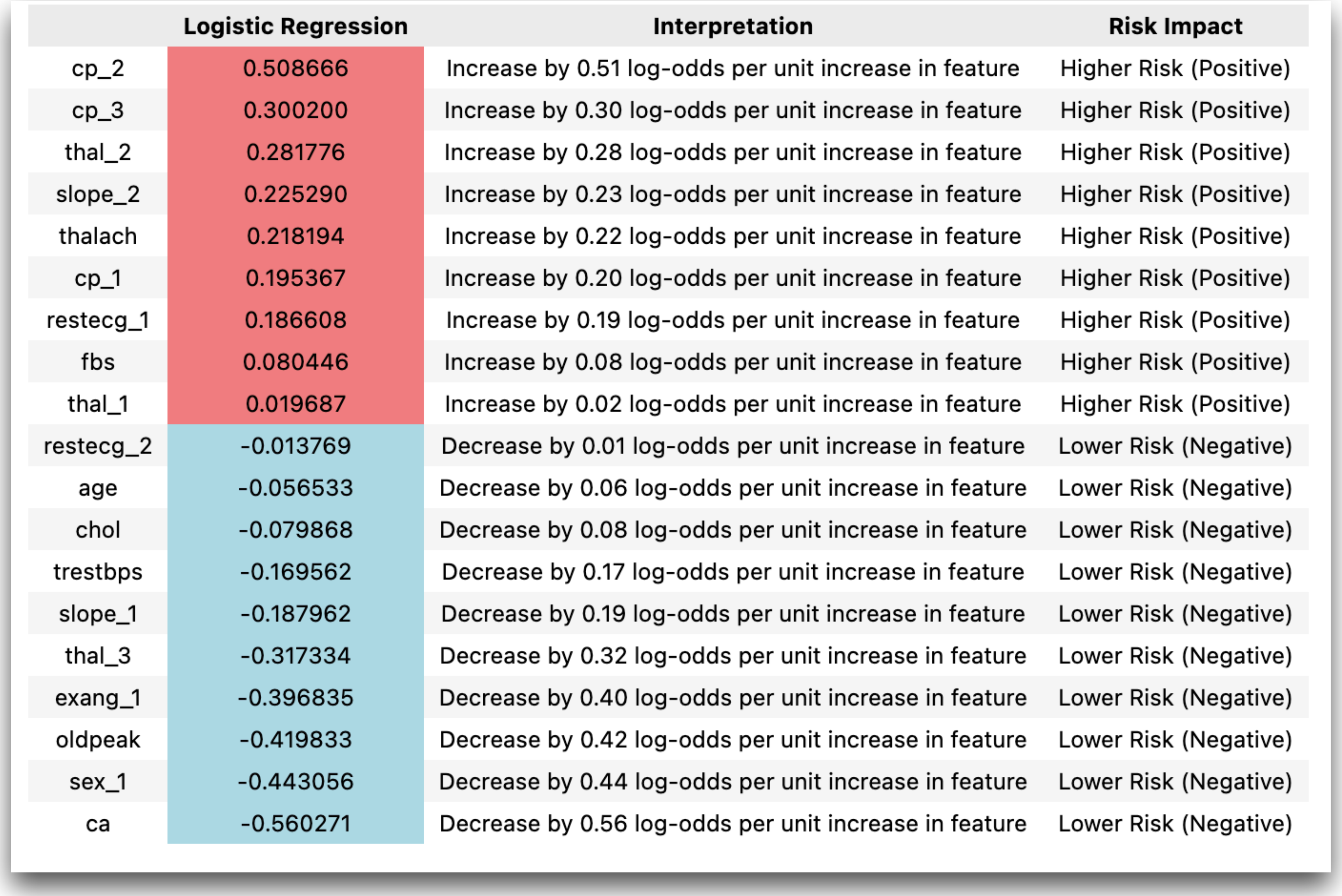
cp_3 (0.3002)

- If a person has chest pain type 3 increases the log-odds of heart disease by 0.30.

thal_2 (0.2818)

- Having a fixed defect in thalassemia test increases the log-odds of heart disease by 0.28.

ℹ️ *cp_3 (Non-anginal pain), even if chest pain is not traditionally heart-related, it still raises heart disease risk. Thal_2 (Fixed defect in thalassemia test) indicates possible permanent heart damage, making it a strong predictor of heart disease.*



	Logistic Regression	Interpretation	Risk Impact
cp_2	0.508666	Increase by 0.51 log-odds per unit increase in feature	Higher Risk (Positive)
cp_3	0.300200	Increase by 0.30 log-odds per unit increase in feature	Higher Risk (Positive)
thal_2	0.281776	Increase by 0.28 log-odds per unit increase in feature	Higher Risk (Positive)
slope_2	0.225290	Increase by 0.23 log-odds per unit increase in feature	Higher Risk (Positive)
thalach	0.218194	Increase by 0.22 log-odds per unit increase in feature	Higher Risk (Positive)
cp_1	0.195367	Increase by 0.20 log-odds per unit increase in feature	Higher Risk (Positive)
restecg_1	0.186608	Increase by 0.19 log-odds per unit increase in feature	Higher Risk (Positive)
fbs	0.080446	Increase by 0.08 log-odds per unit increase in feature	Higher Risk (Positive)
thal_1	0.019687	Increase by 0.02 log-odds per unit increase in feature	Higher Risk (Positive)
restecg_2	-0.013769	Decrease by 0.01 log-odds per unit increase in feature	Lower Risk (Negative)
age	-0.056533	Decrease by 0.06 log-odds per unit increase in feature	Lower Risk (Negative)
chol	-0.079868	Decrease by 0.08 log-odds per unit increase in feature	Lower Risk (Negative)
trestbps	-0.169562	Decrease by 0.17 log-odds per unit increase in feature	Lower Risk (Negative)
slope_1	-0.187962	Decrease by 0.19 log-odds per unit increase in feature	Lower Risk (Negative)
thal_3	-0.317334	Decrease by 0.32 log-odds per unit increase in feature	Lower Risk (Negative)
exang_1	-0.396835	Decrease by 0.40 log-odds per unit increase in feature	Lower Risk (Negative)
oldpeak	-0.419833	Decrease by 0.42 log-odds per unit increase in feature	Lower Risk (Negative)
sex_1	-0.443056	Decrease by 0.44 log-odds per unit increase in feature	Lower Risk (Negative)
ca	-0.560271	Decrease by 0.56 log-odds per unit increase in feature	Lower Risk (Negative)

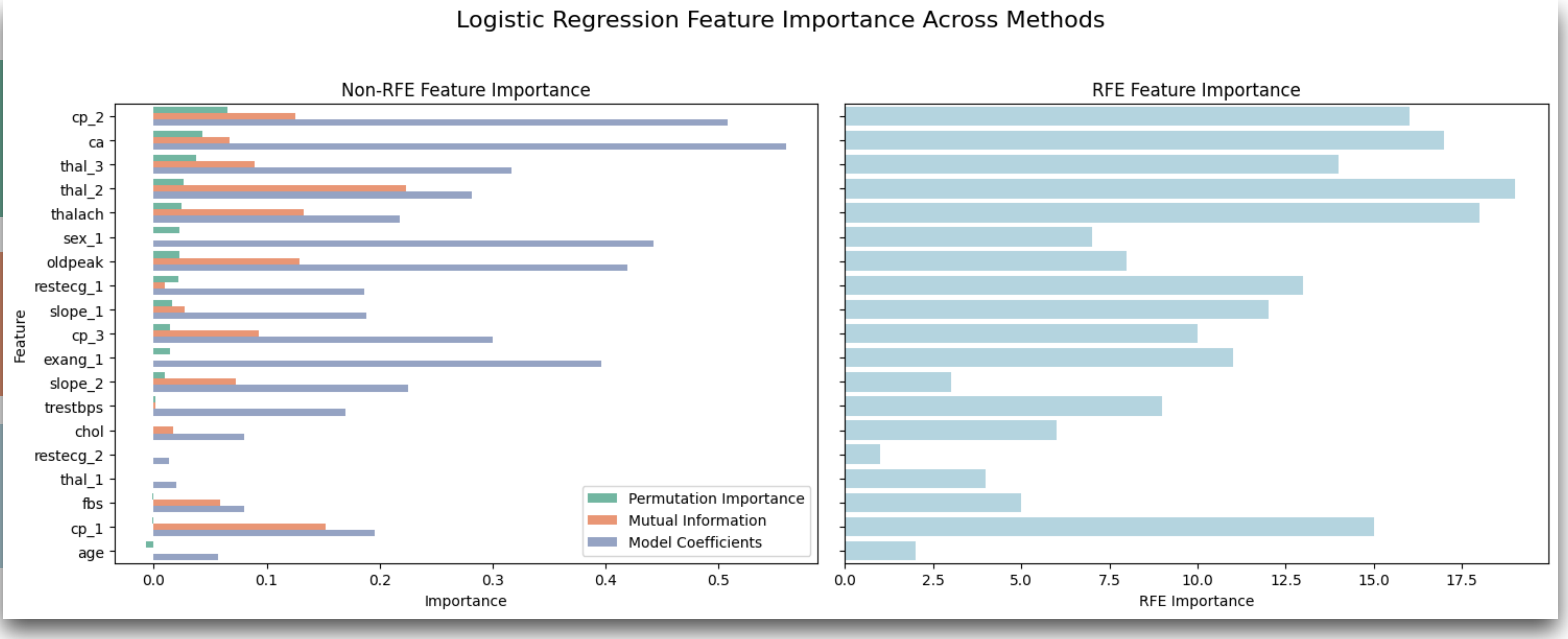
EXTRA ANALYSIS

FEATURE IMPORTANCE: OTHER METHODS

Permutation Importance: Measures how much a feature contributes to model performance by shuffling its values and observing the change in accuracy.

Mutual Information: Measures the dependency between a feature and the target variable. Higher values mean stronger relationships.

Recursive Feature Elimination (RFE): Selects the most relevant features by iteratively removing the least important ones.



Ranking Across All Methods

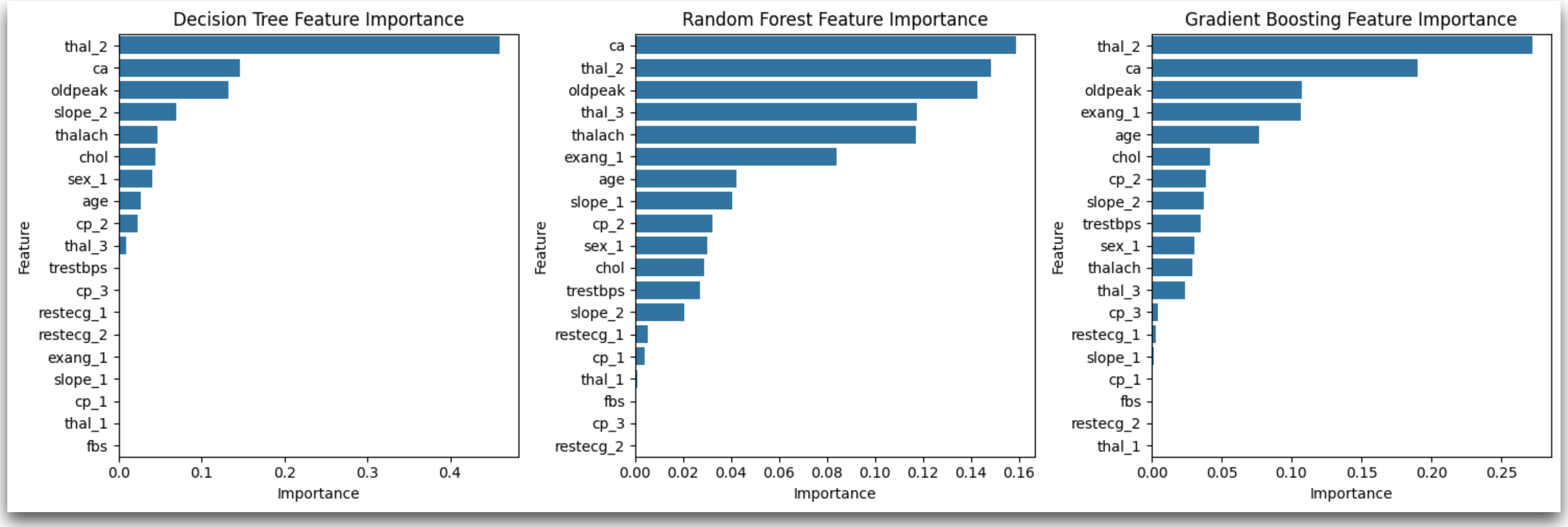
By combining the rankings from all methods, here's the strongest contributors in terms of importance to the model:

- thal_2
- thal_3
- ca
- thalach
- oldpeak

ℹ *While it's important to look at the model coefficients to show how much an individual feature affect prediction, there are also other feature importance methods that can be used if we want to understand how a specific feature behaves when real world data changes or other features are added. Another example of feature importance technique that can be used is SHAP.*

EXTRA ANALYSIS

FEATURE IMPORTANCE: OTHER MODELS



Decision Tree Feature Importance

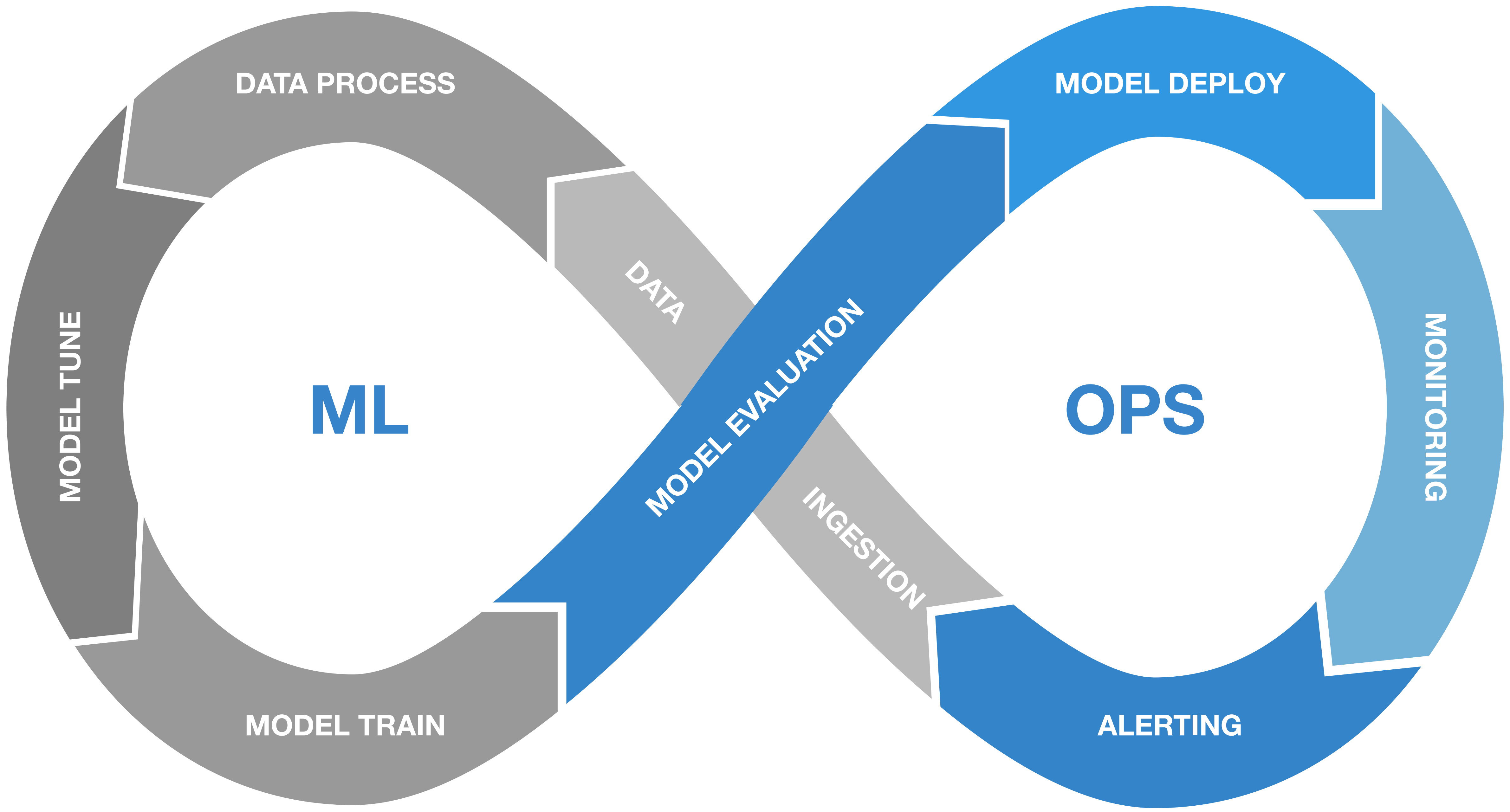
Random Forest Feature Importance

Gradient Boosting Feature Importance

ℹ️ *While not the chosen best model, I also looked at the feature importance for the 3 other models, these are also the magnitude impact* ***(not directional impact).*** *As you can see, the ca (which is the top most impactful by magnitude in the chosen model is also one of the tops in the other models. This chart can be used as future improvement or to do further analysis to answer questions like (why is cp_2 and sex_1 not as impactful here compared to Logistics?)*

EXTRA SLIDE

MLOPS



ℹ️ *This is just an example MLOps process that I would setup to operationalize ML model runs.*

A unified framework is needed to streamline ML operations, improve visibility, and enhance model performance while addressing resource constraints.

Below are a few tools / services that can help with ML Ops **integration,** Operational **Visibility**, as well as improve computational speed.

RAY

Ray is a distributed computing framework for building and running ML models. You can use Ray tune your models, **Ray Serve** for Model serving which exposes the ML model through API to make it accessible by several applications at the same time without each application having to maintain their own copy.

mlflow™

MLFlow is a platform you can use to track model runs, compare model performance runs and register model.

aws

Amazon SageMaker

AWS offers different services that will help with MLOps. Other cloud platforms also offer some similar services