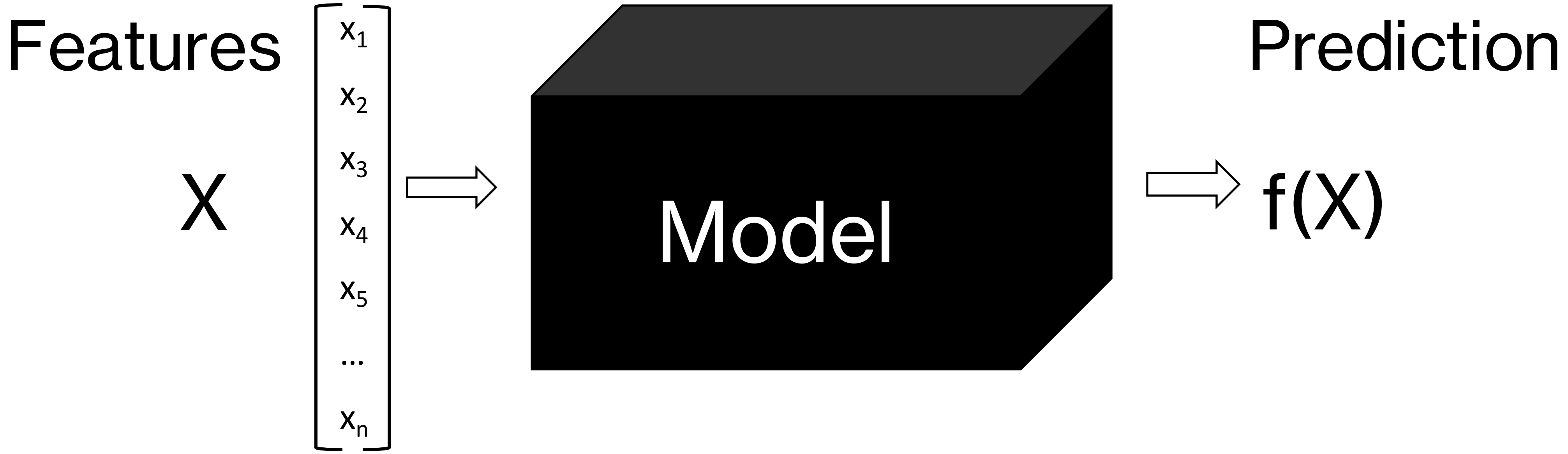
# Proving Reliability of Machine Learning Systems using Explainable AI

Dene Brown


What actions can we take to change the outcome and secure the loan?

- Earn more money
- Pay off other loans
- Pay off credit cards
- Move house

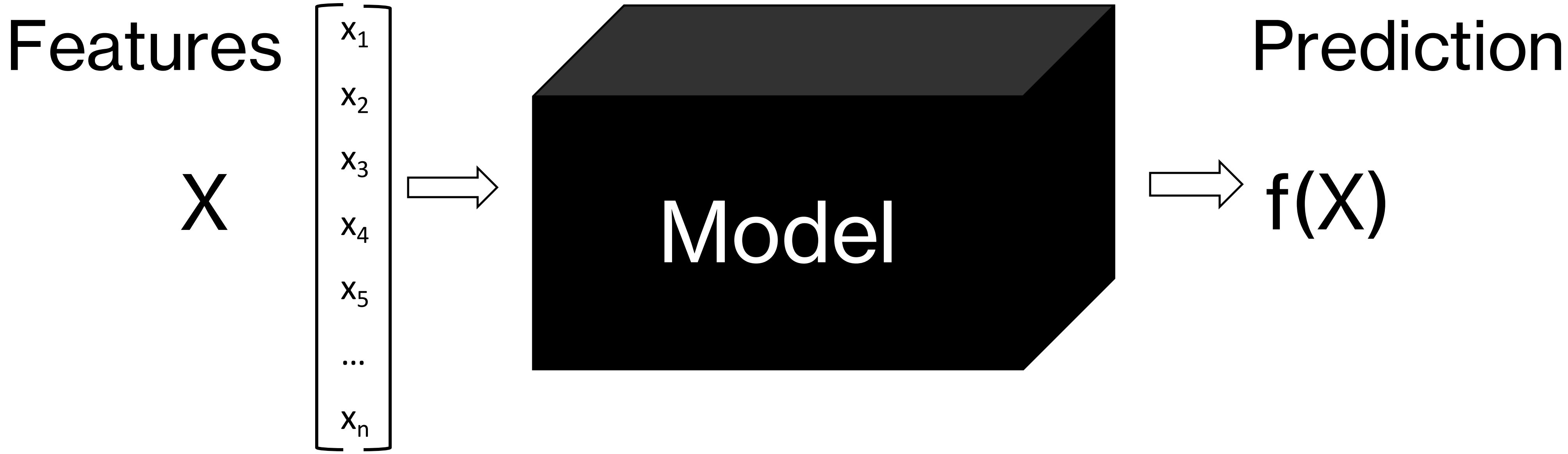
# AI Black Box Model




- We have no knowledge of the black box model reasoning.
- How do we determine model reliably for critical systems?

# Wolf or Husky?

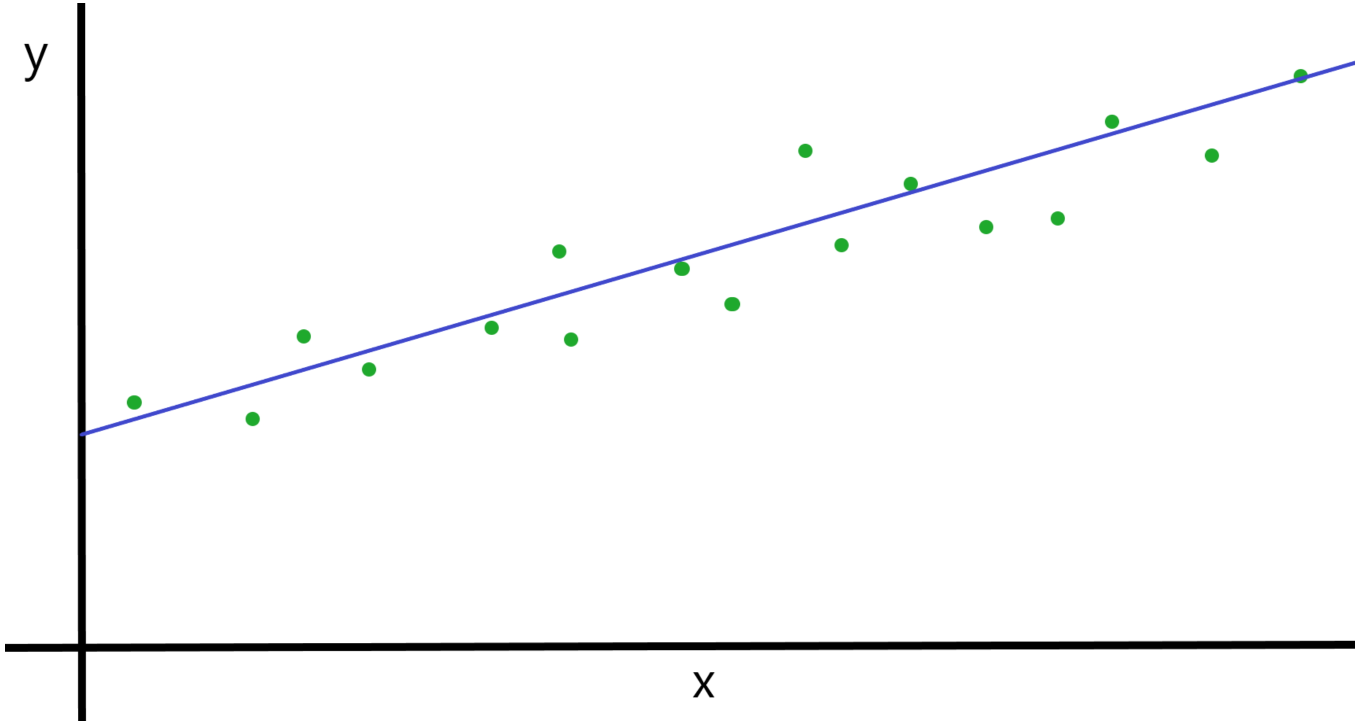
# AI Black Box Model




- We have no knowledge of the black box model reasoning.
- How do we determine model reliably for critical systems?
- Is testing the model predictions sufficient?

# LIME

# LIME is based on a Linear Regression Model




$$y = w_0 + w_1 x$$

$$y = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$$

# LIME

Local Interpretable Model-agnostic Explanations

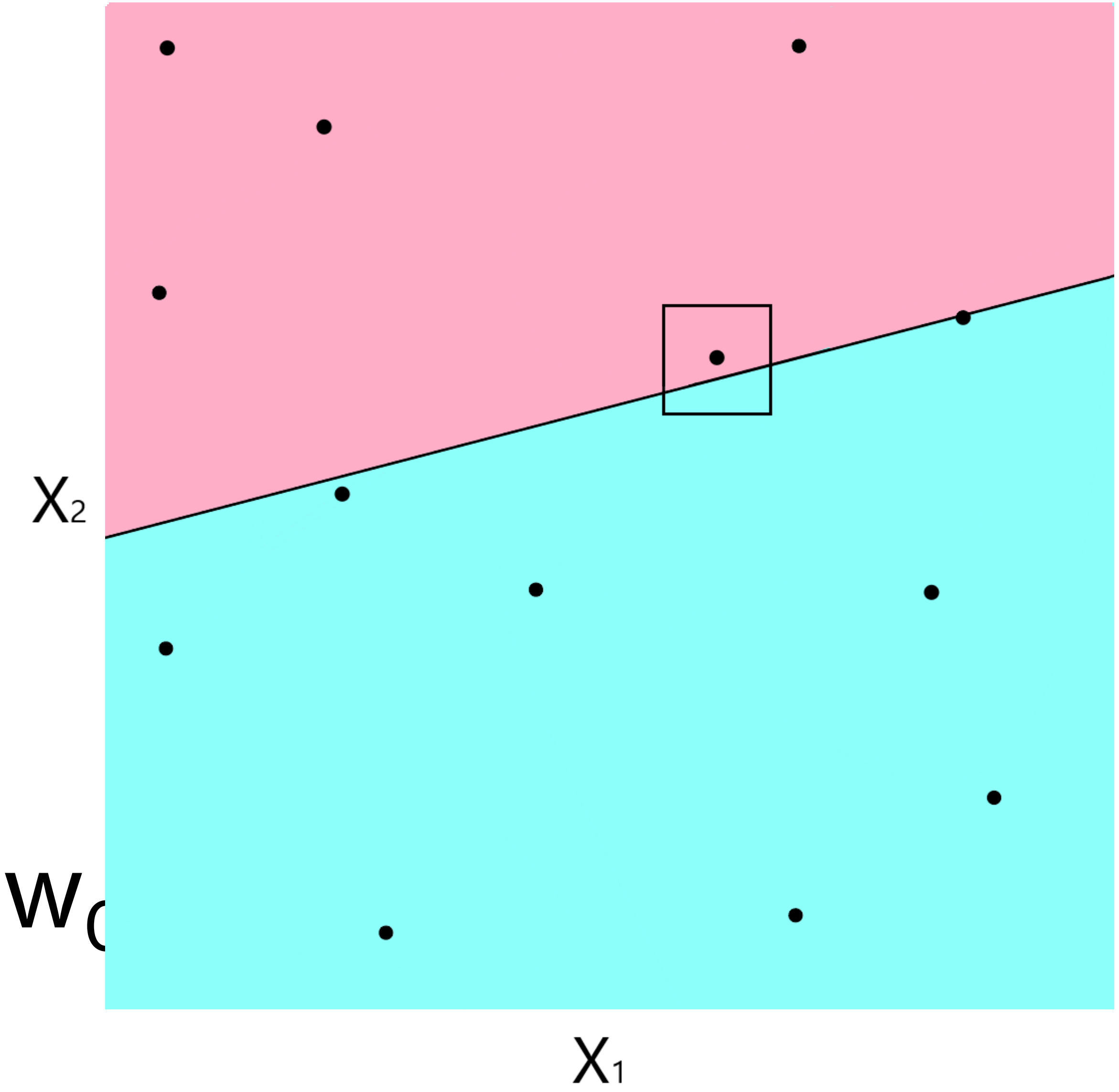
1. Choose point to explain locally.

2. Create surrogate data around point using perturbed  points in feature space.

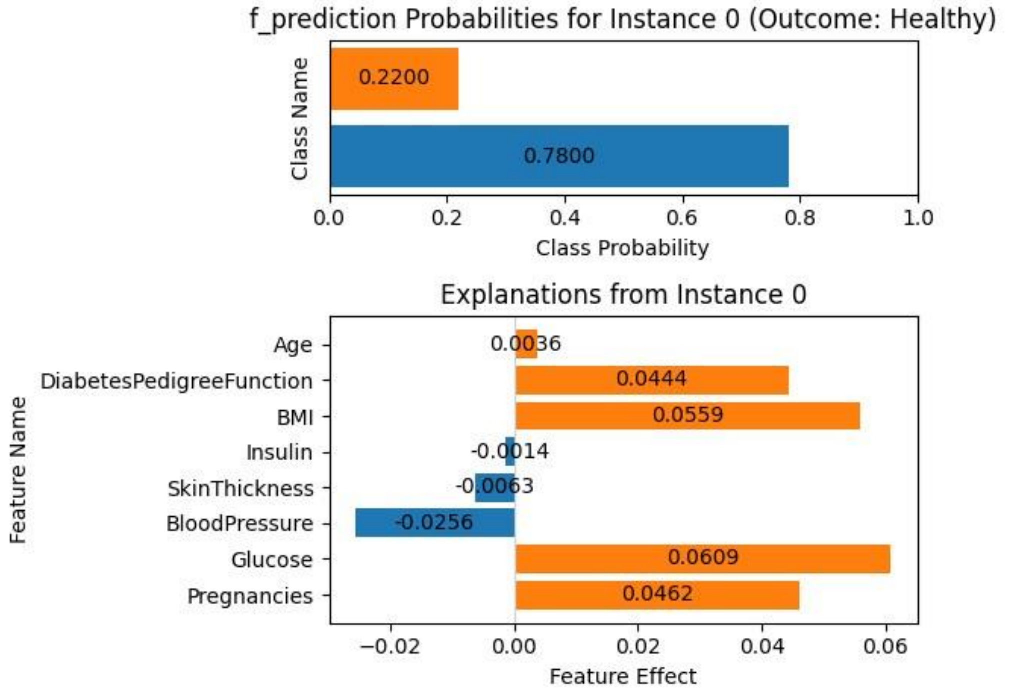
3. Train weighted linear model using surrogate data.

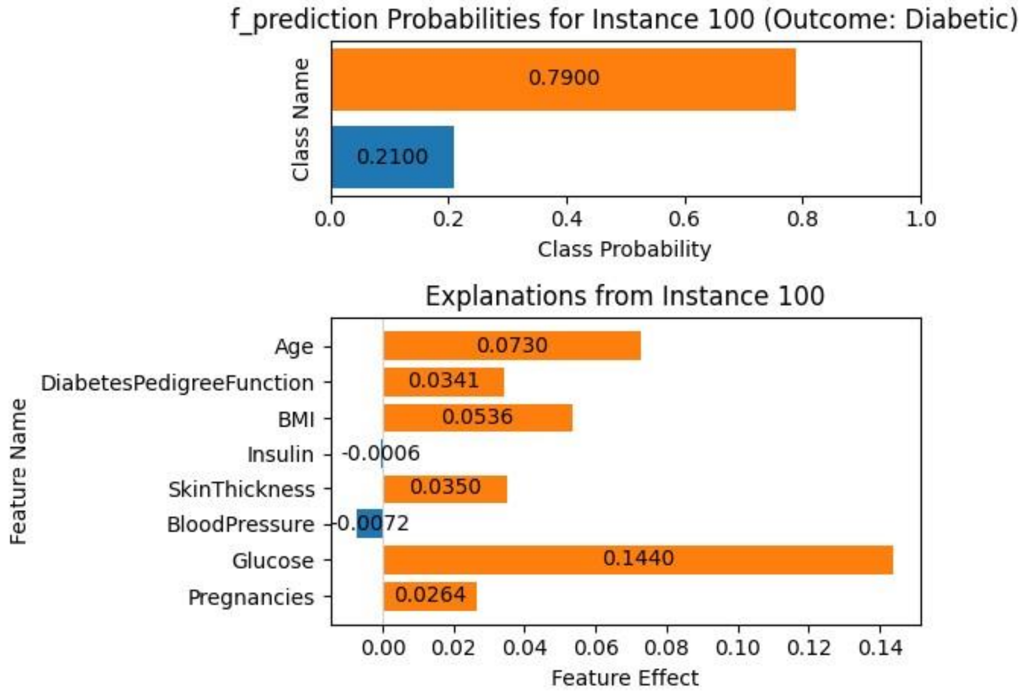
$y = w_0$

4. Feature importance from linear model coefficients $w_1$ and $w_2$.

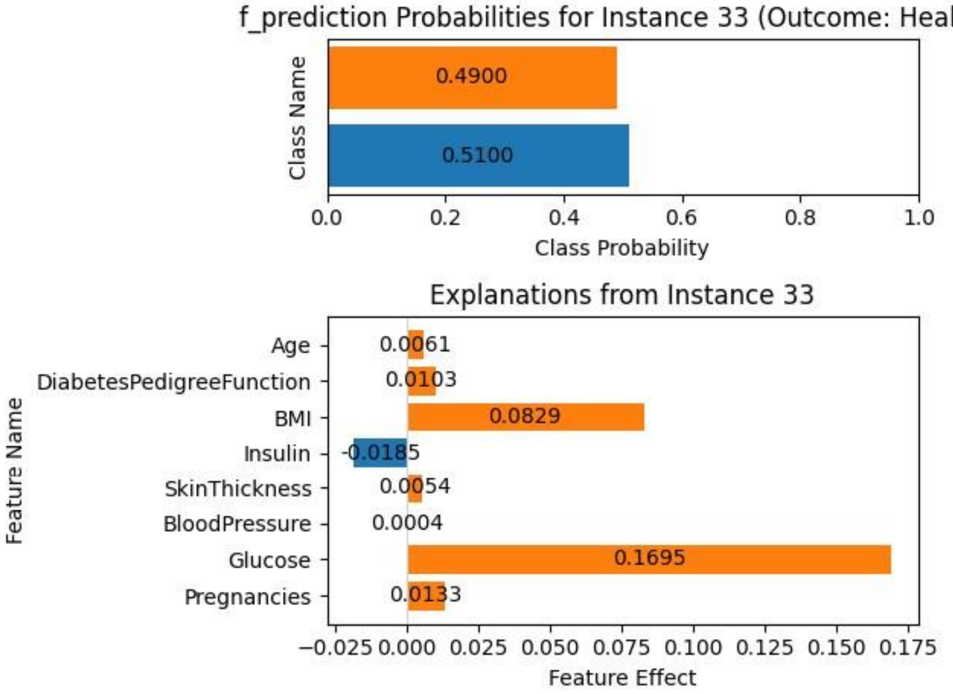


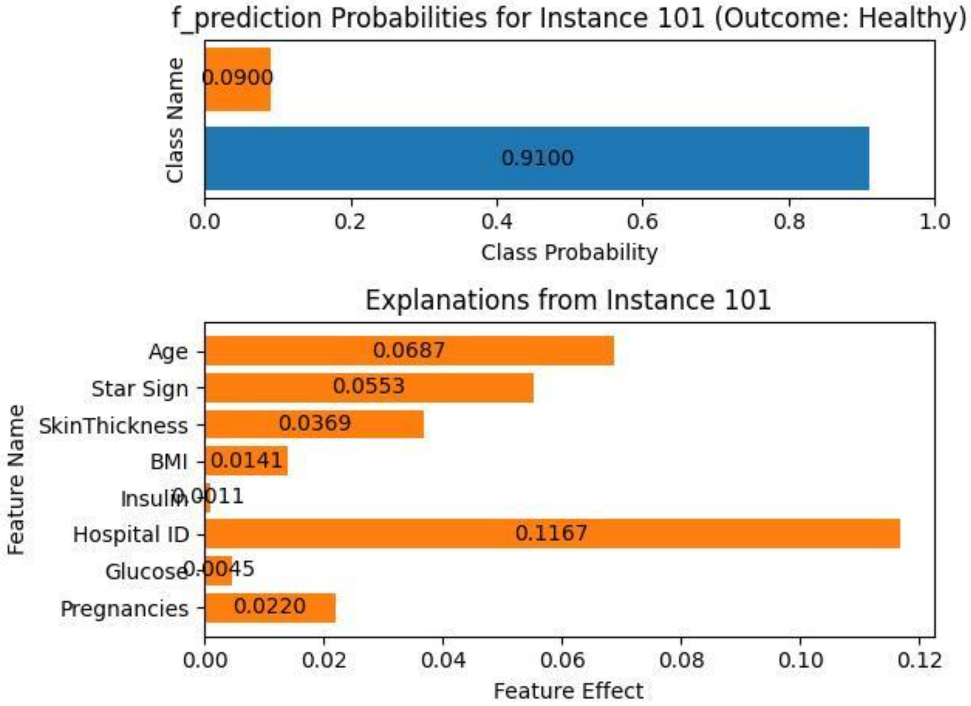
# LIME: Diabetes Results 1/3



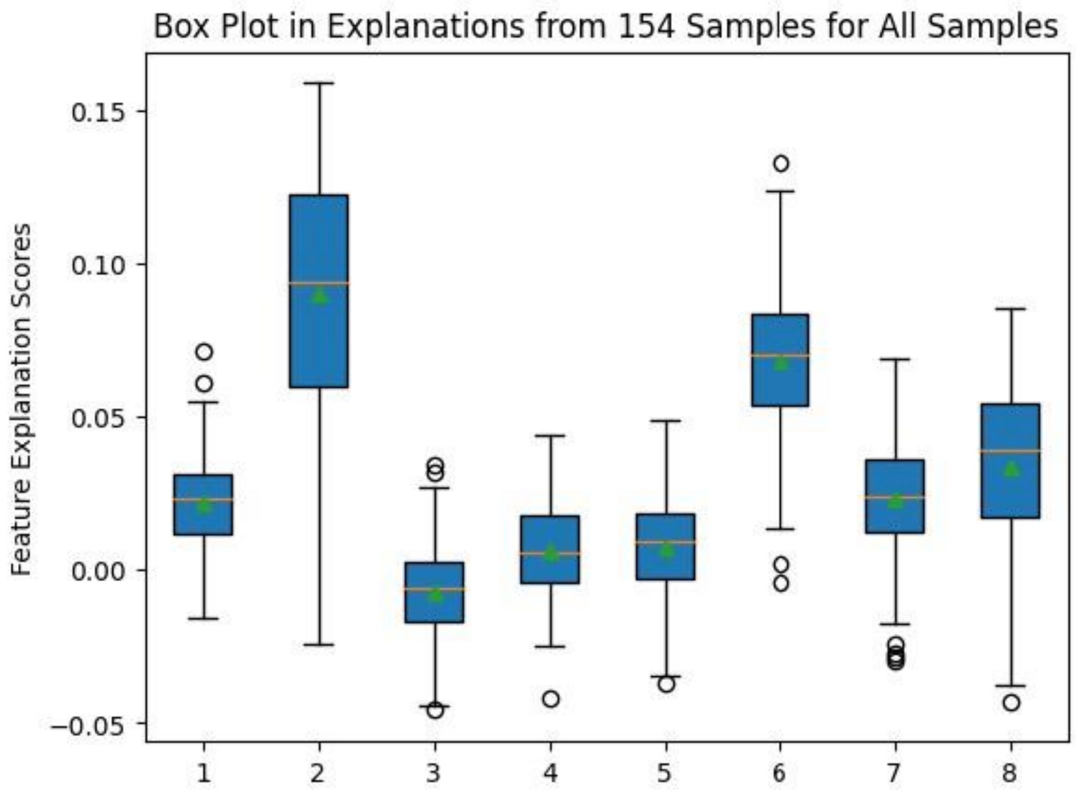



# LIME: Diabetes Results 2/3

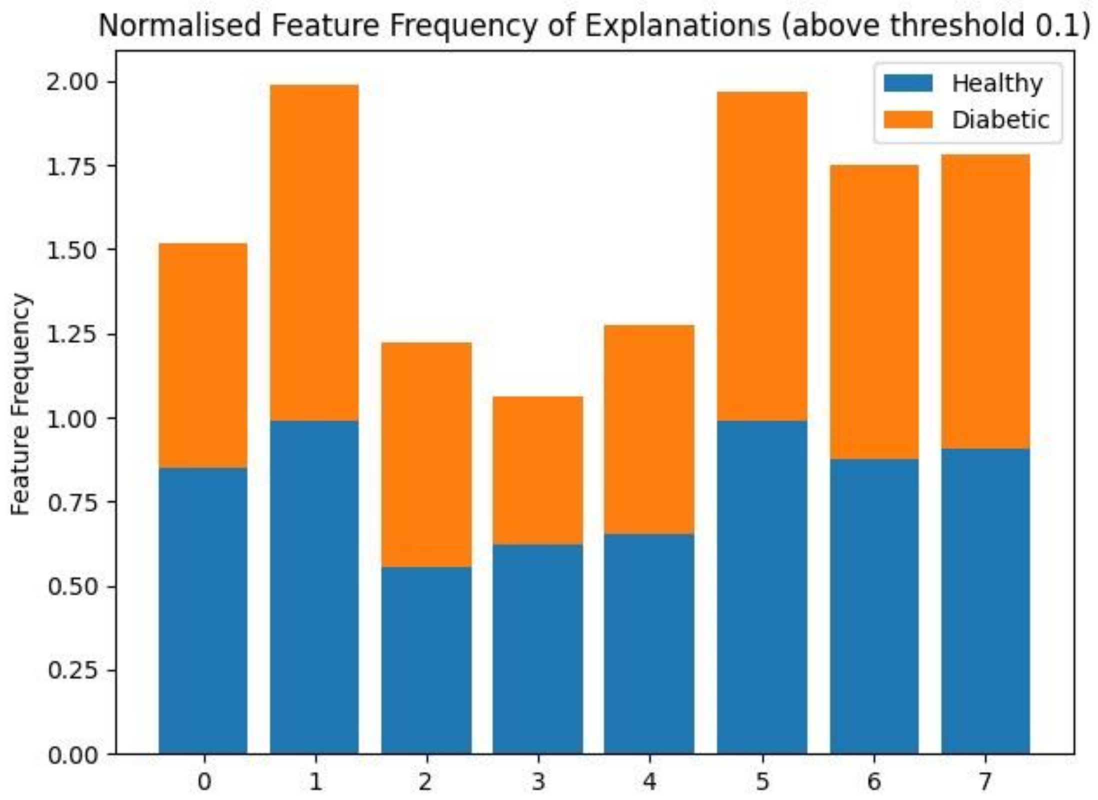


# LIME: Diabetes Results 3/3

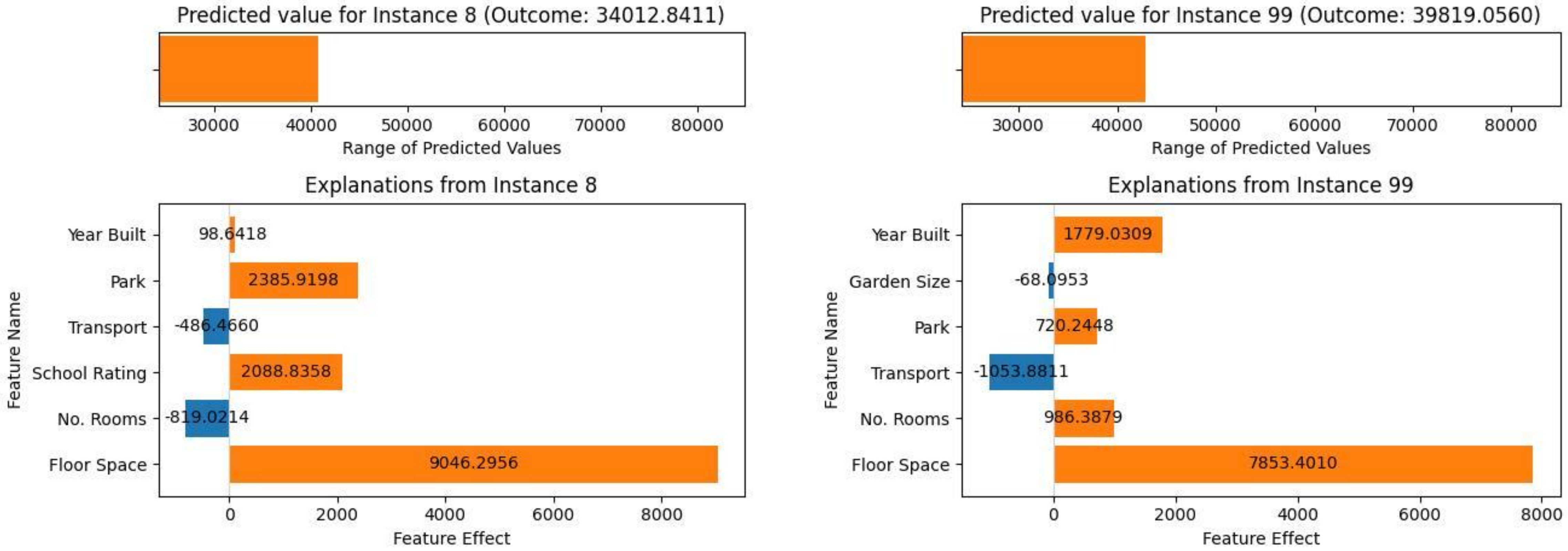



1 - Pregnancies
2 - Glucose
3 - BloodPressure
4 - SkinThickness
5 - Insulin
6 - BMI
7 - DiabetesPedigreeFunction
8 - Age

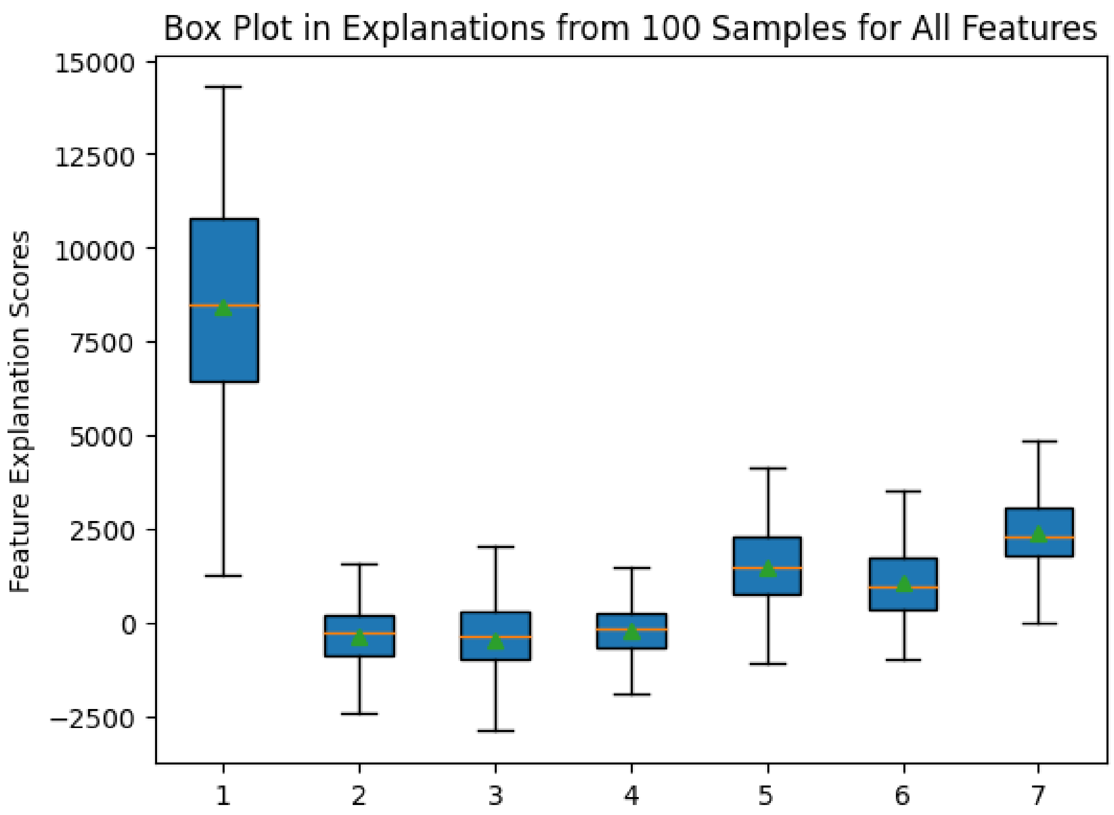



0 - Pregnancies
1 - Glucose
2 - BloodPressure
3 - SkinThickness
4 - Insulin
5 - BMI
6 - DiabetesPedigreeFunction
7 - Age

# LIME: House Price Results 1/2



# LIME: House Price Results 2/2




1 - Floor Space
2 - No. Rooms
3 - School Rating
4 - Transport
5 - Park
6 - Garden Size
7 - Year Built

# Multiple Flavours  of LIME

- DLIME: Clustering of Real Data Points
- ALIME: Auto Encoder to create Synthetic Dataset
- UnRAvEL: Gaussian Process
- BayesLIME: Baysian Approach

All methods are intended to improve the selection of surrogate data as a way to improve model performance.

# Scoped Rules

# Scoped Rules (Anchors)

- Anchors are given their name because they 'anchor' the prediction to that of the instance being explained.
- We create perturbations around the instance that could have their outcome predicted by the rule.
- The rules have the format *'if – then'* that are easy to interpret.
- Each rule comes with a coverage and precision.

# Anchors Example: Diabetes

- A Model has been trained to predict if patients will develop diabetes or not (Unhealthy or Healthy).
- Use anchors to explain why one instance is predicted to be Unhealthy.

The following rule with two predicates is generated from the instance:

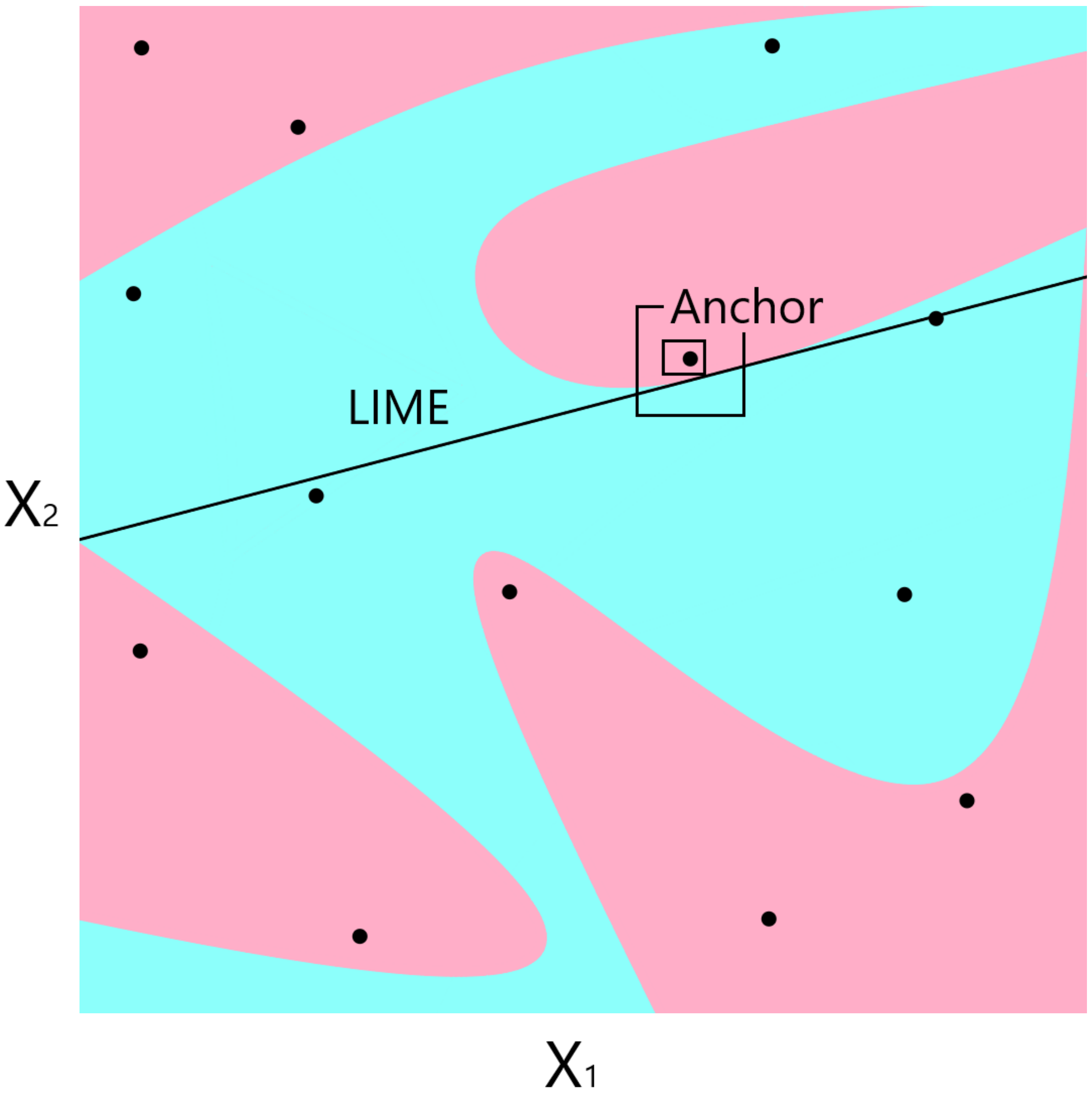
```
if BMI     > 29  and
   Glucose > 120 then
   Predict UNHEALTHY with Precision 90% and Coverage 32%
```

# Anchors Visualisation

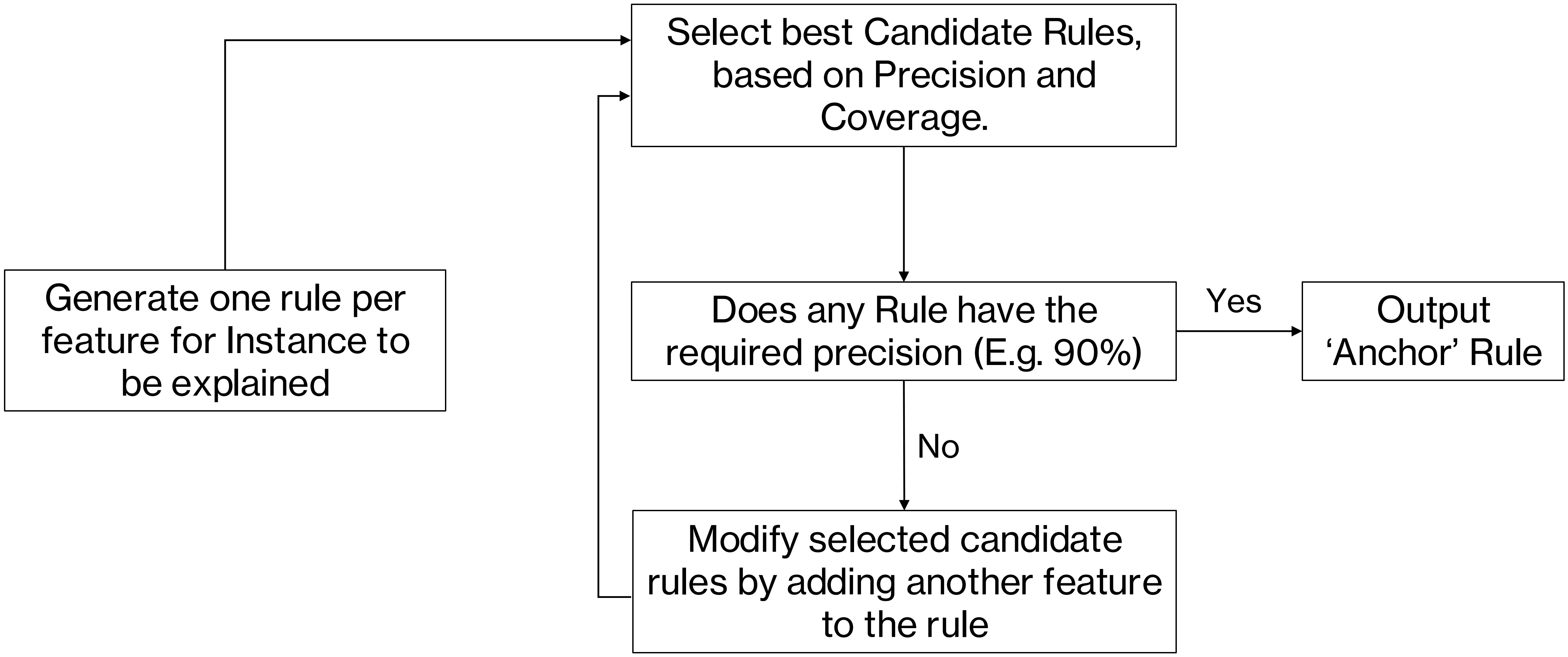
```
Coverage:  32%
Precision: 90%
```

A coverage of 32% mean that the anchor rule can apply to 32% of perturbation instances within the perturbation space .

A precision of 90% means of those perturbation instances that the anchor rule can apply to, 90% have their outcome correctly predicted by the anchor rule.



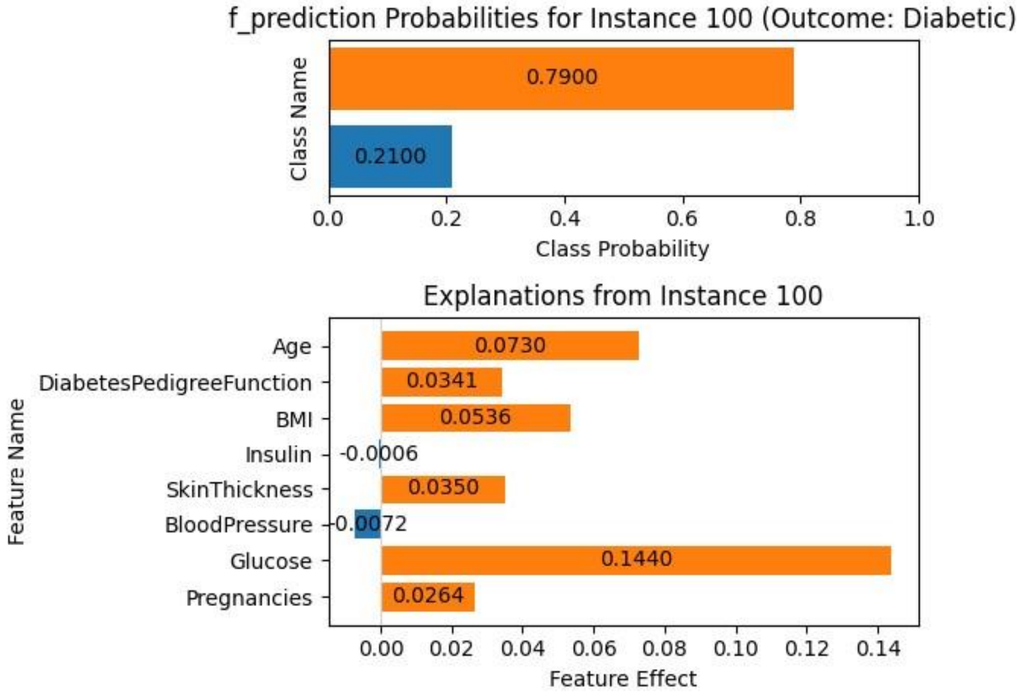
# Generating  Anchors (Simplified Model)



# Generating Anchors Simple Example

| | | | | |
|---|---|---|---|---|
| Generate one rule per feature | `if BMI > 29`<br>`then UNHEALTHY`<br>`Precision = 42%` | `if Preg > 4`<br>`then UNHEALTHY`<br>`Precision = 15%` | `If Gluc < 100`<br>`then HEALTHY`<br>`Precision = 28%` | `If DPF < 0.5`<br>`then HEALTHY`<br>`Precision = 54%` |
| Select best Candidate Rules | `if BMI > 29`<br>`then UNHEALTHY`<br>`Precision = 42%` | | | `If DPF < 0.5`<br>`then HEALTHY`<br>`Precision = 54%` |
| Precision > 90%? | No | | | No |
| Modify selected candidate, by adding feature | `if BMI > 29 and`<br>`Gluc > 120`<br>`then UNHEALTHY`<br>`Precision =` **93**`%` | `If BMI > 29 and`<br>`Preg > 3`<br>`then UNHEALTHY`<br>`Precision = 53%` | `if DPF < 0.5 and`<br>`Gluc < 100`<br>`then HEALTHY`<br>`Precision =` **91**`%` | `if DPF < 0.5 and`<br>`BMI < 27`<br>`then HEALTHY`<br>`Precision = 72%` |
| Precision > 90%? | Yes: Rule 1 | | Yes: Rule 2 | |

# Anchors Results Compared to LIME 1/2



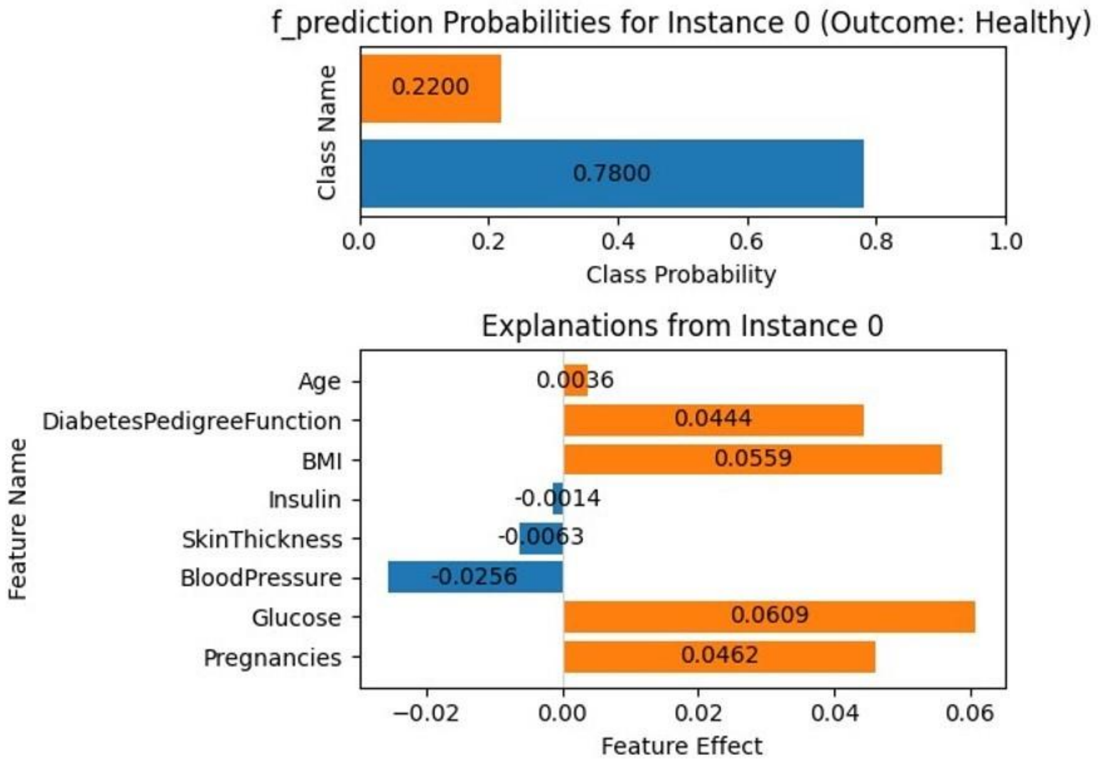

## Instance 100

| Age | DPF | BMI | Insulin | Skin | BP | Glucose | Preg |
|---|---|---|---|---|---|---|---|
| 50 | 0.62 | **34** | 0 | 35 | 72 | **148** | 3 |

## Rule 1

```
if BMI     > 29 and
   Glucose > 120 then
   Predict UNHEALTHY with Precision 90%
     and Coverage 32%
```

# Anchors Results Compared to LIME 2/2




## Instance 0

| Age | DPF | BMI | Insulin | Skin | BP | Glucose | Preg |
|---|---|---|---|---|---|---|---|
| 31 | **0.35** | 24 | 0 | 29 | 66 | **85** | 1 |

## Rule 2

If DPF < 0.50 and
Glucose < 100 then
Predict HEALTHY with Precision 90%
and Coverage 15%

# Precision and Coverage Trade Off with Anchors

- Anchors with 2 to 4 predicates can be easily understood, above this rules are difficult to interpret.

- Increasing the number of predicates improves precision, the accuracy of the rule.

- A high number of predicates reduces the coverage, the scope of the rule.

# Accumulated Local Effects (ALE)

# Accumulated Local Effects

- ALE is a Global Modeling Method, despite its name.
- ALE is a method that can used with correlated features.
- It gives the effect of a feature has on the outcome, over the full range of the features values.

# ALE: Theory

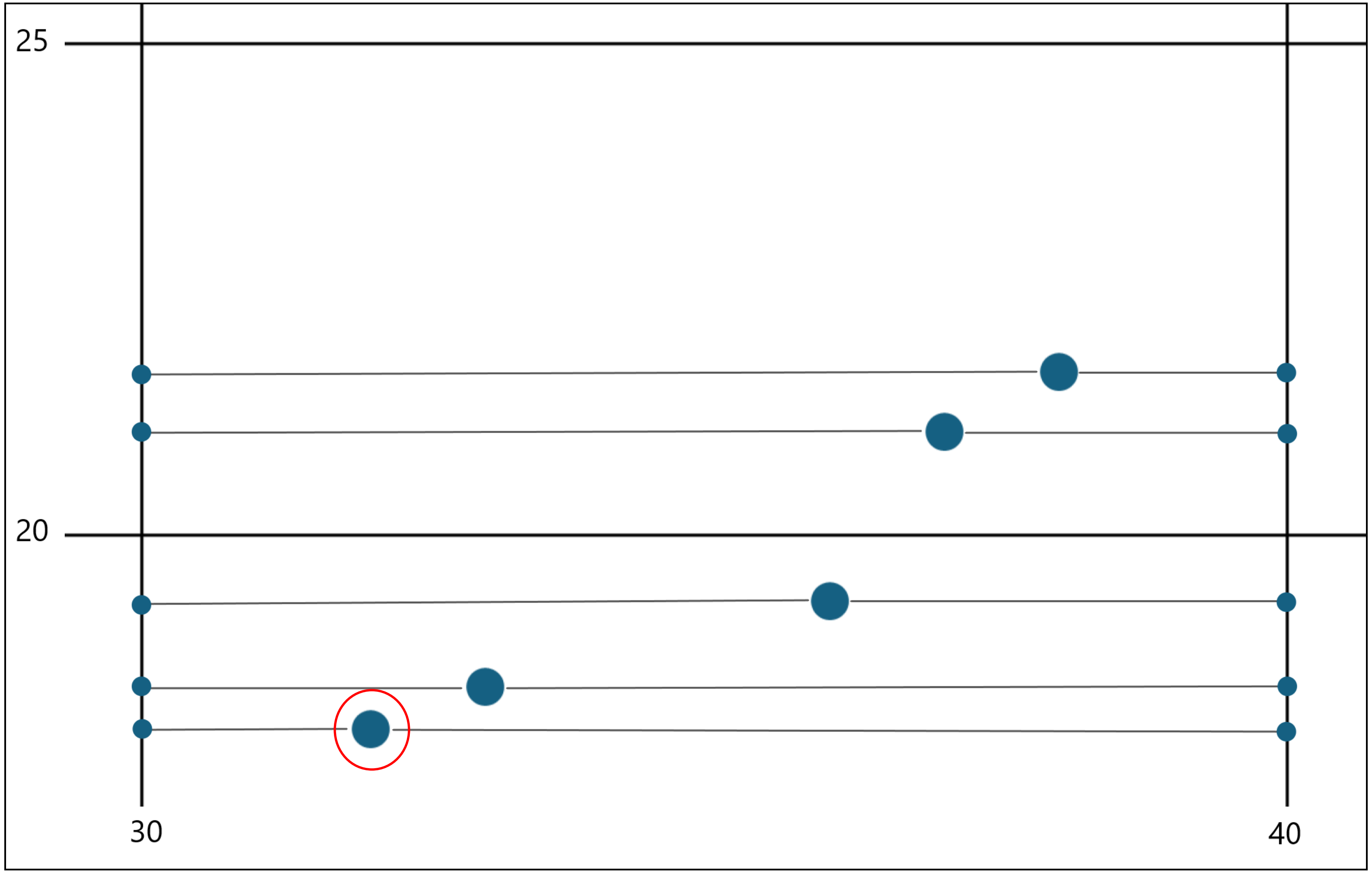
1. Choose Feature to be Explained

2. Divide Feature into Ranges

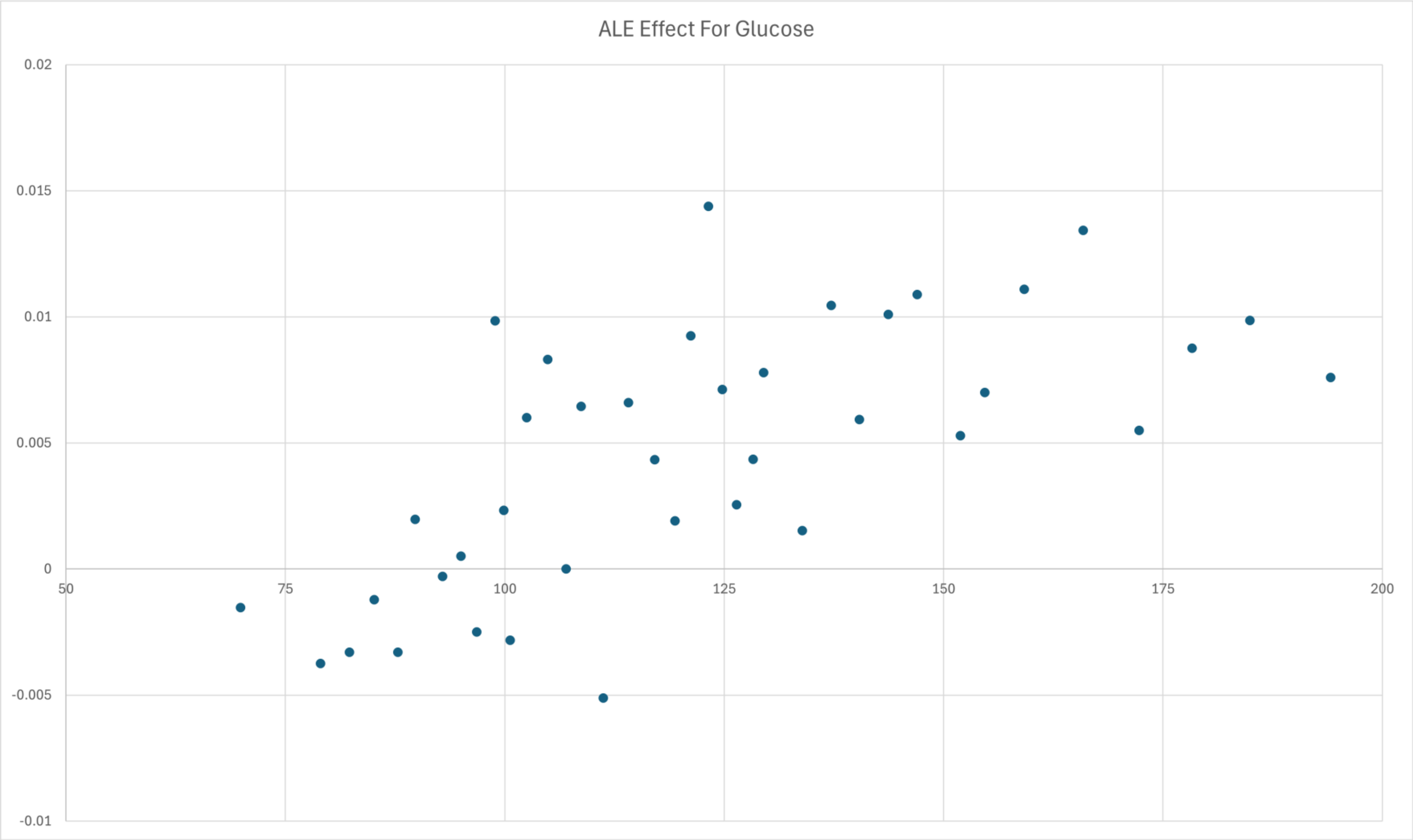
3. Select an instance:

- Replace value of feature of interest with upper and lower bounds of the range.
- Calculate difference in outcomes from new points at the bounds (Local Effect).

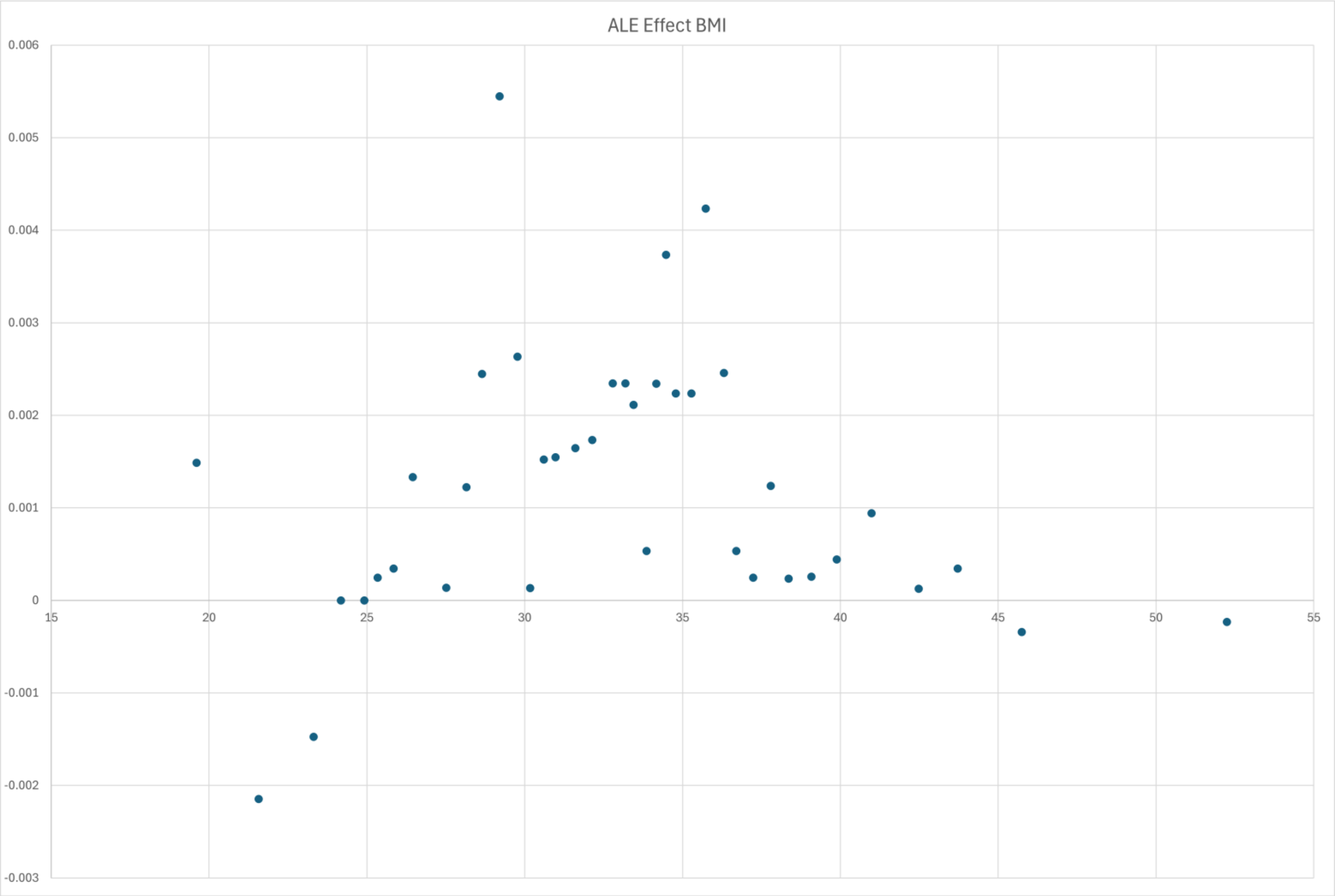
4. Repeat for all instances in the range and take average (Accumulated Local Effect).



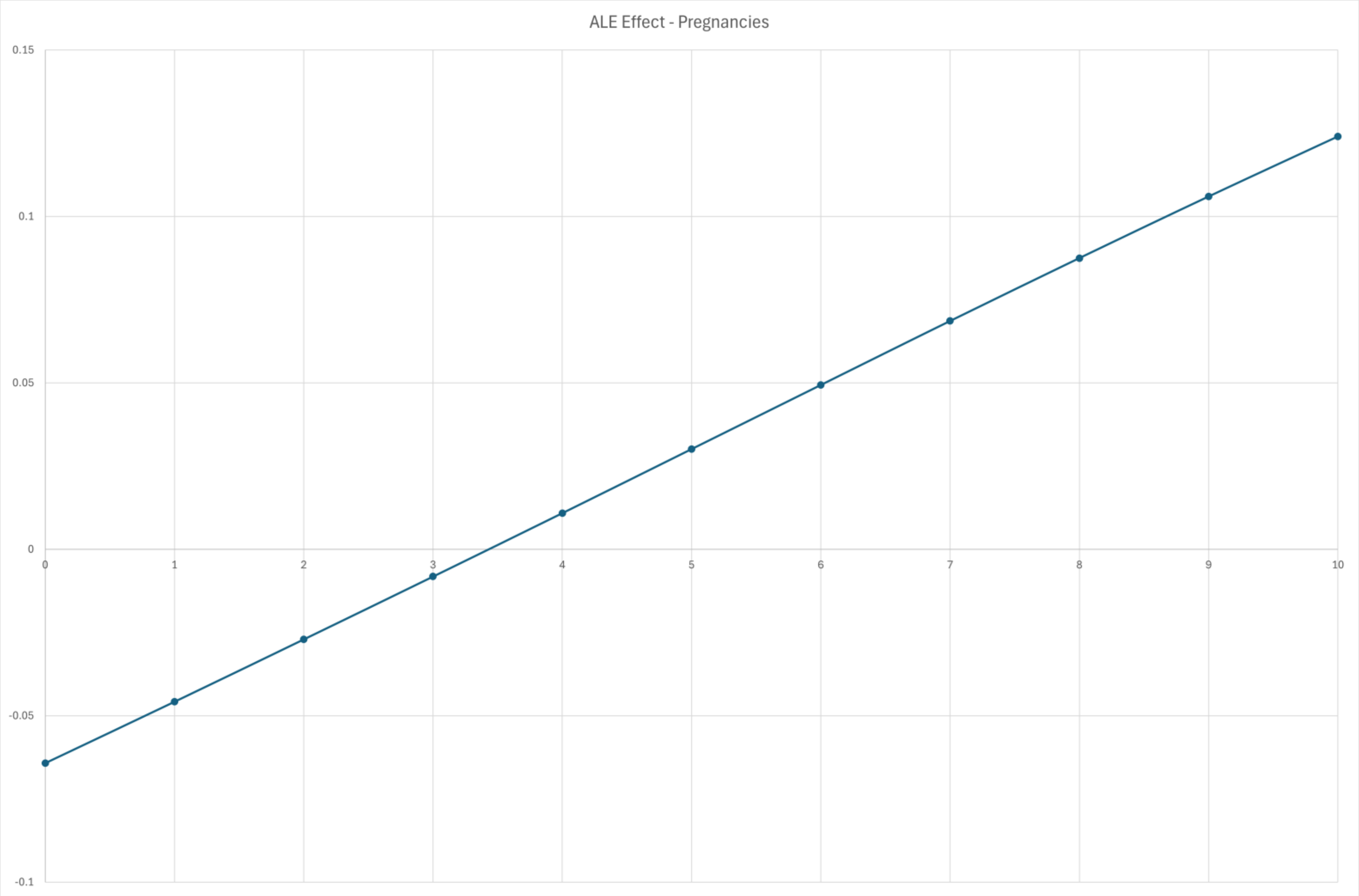
# ALE Results: Diabetes Risk - Glucose



# ALE Results: Diabetes Risk - BMI



# ALE Results: Diabetes Risk - No. Pregnancies



# Current Research

# Current Academic Research in Explainable AI

- Explainable AI is currently an important topic in academic research.
- Many models from research are accompanied by their own explainability models, many using the methods explained here.
- Model specific explainability methods are also used, these can provide greater insight than model agnostic methods.
- Many of these explainability methods are for Neural Network models.

# Summary and Questions

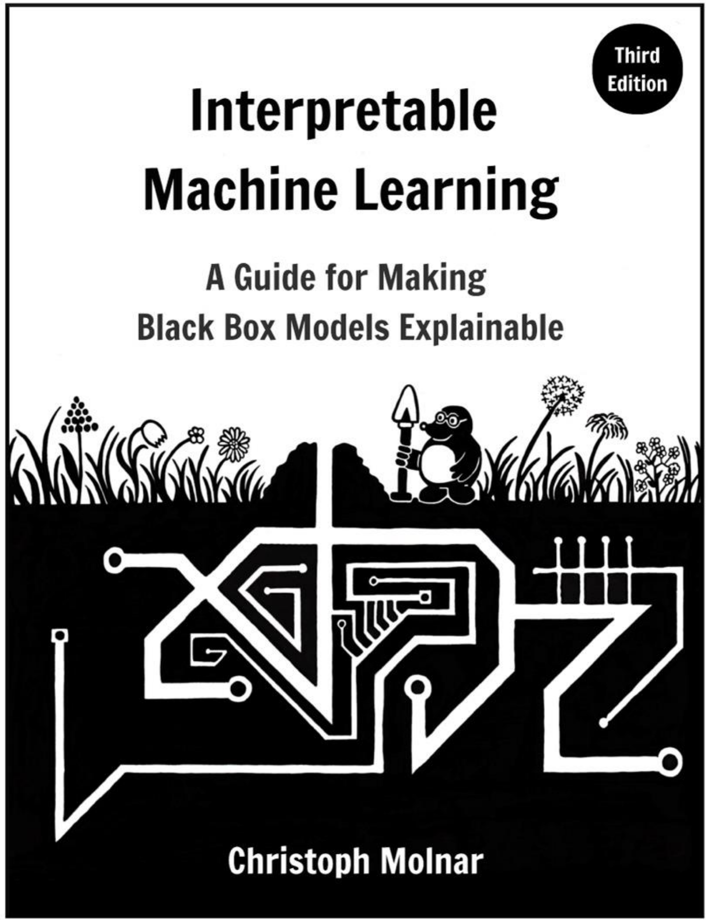
# Summary 1/2

- Explainable or Interpretable Machine learning has been around for over a decade.
- Explainable AI is relatively easy to implement with lots of available open-source software.
- Model agnostic methods can interpret any model, with having to understand the workings of the model.
- There are both Global (ALE) and Local methods (LIME).
- Model specific methods can perhaps provide greater insight.

# Summary 2/2

- Machine learning models are being implemented in High Integrity Systems.
- Explainable AI can provide insights to how these models are reasoning.
- Can Explainable AI improve the quality and reliability of the models?
- Should Explainable AI become a requirement in standards for critical systems based on Machine Learning?
- Yes, and Yes (unless other methods better appear).

# Questions etc.


christophm.github.io/interpretable-ml-book/