Interpretability in machine learning

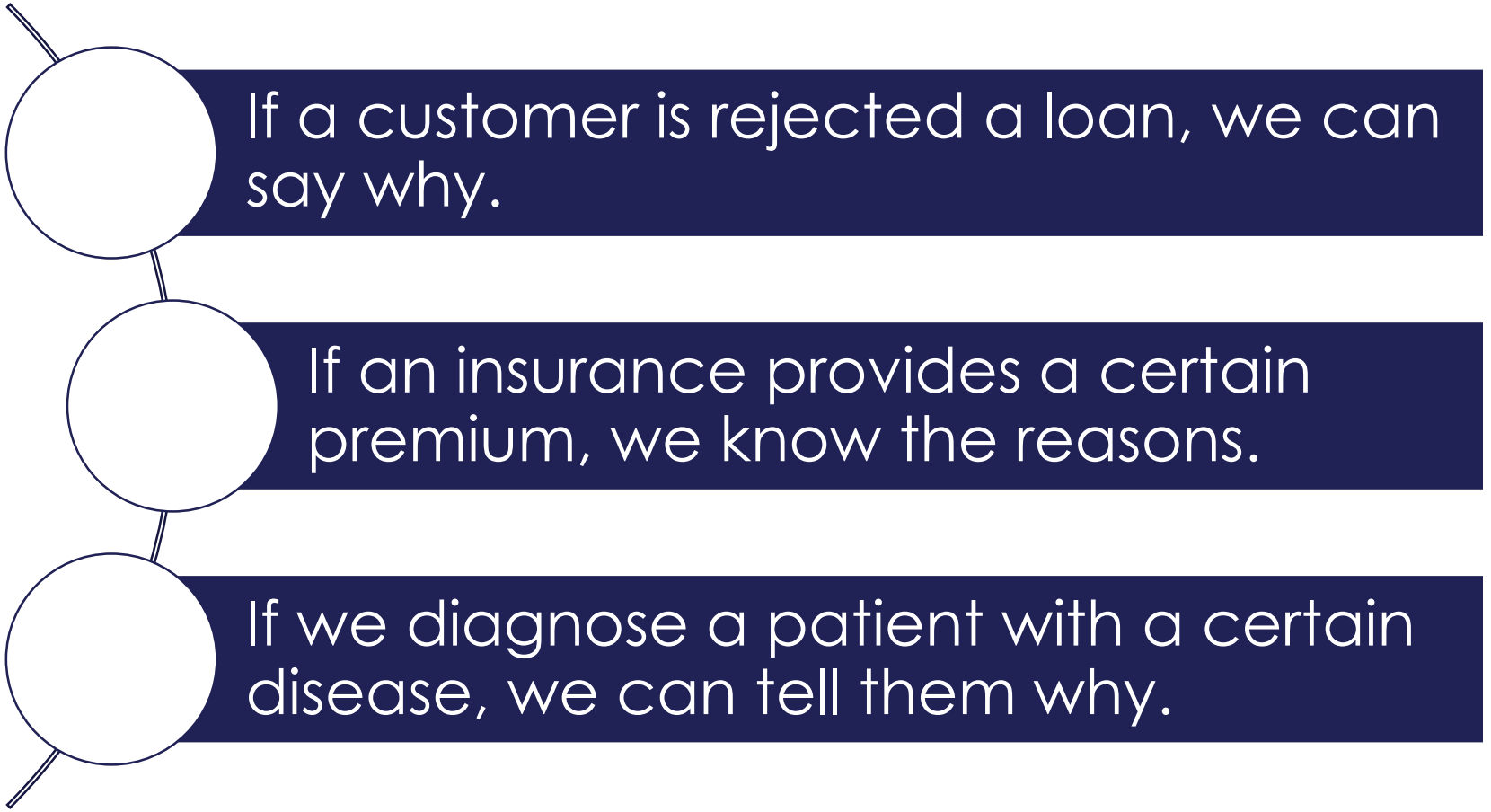
Interpretability

Interpret means to "*explain*" or to present in understandable terms.

Interpretability in machine learning

The ability to express in understandable terms, what the model has learned and the reasons that affect their output.

Interpretability examples



Machine learning models

We commonly talk about machine learning models as tools to automate or optimize decision making.

Machine learning models

Machine learning models produce knowledge about domain relationships.

interpretations

We use these interpretations to learn more about the domain.

ML Interpretability = knowledge

We can use machine learning models to learn:

- ✓ The characteristics of fraudulent motor claims.
- ✓ The main drivers of house prices.
- ✓ The characteristics of malignant tumours in an image.

ML Interpretability = knowledge

By understanding the models' output, we can know if:

- ✓ The model is discriminating or being unfair.
- ✓ The reasons that drive the output make sense.

Machine learning models

Understanding the outputs of machine learning models produce knowledge.

This knowledge has applications in medicine, policy, science, among other industries, and also, in **auditing the model itself** to ensure fairness and that it follows regulations.

Interpretable machine learning

What is it?

Interpretable machine learning

Use of machine learning models to extract knowledge from the relationships in the data.

An interpretable ML model obeys a domain-specific set of rules to allow the model or its predictions to be easily understood by humans.

Train In Data

THANK YOU

www.trainindata.com