Understanding Opacity in Machine Learning Models

Ankit Rathi, Yatin Bhatia

AGENDA

- Context & Motivation
- Real-world Use Cases
- Understanding Interpretation
- Types of Interpretation
- Scope of Interpretation
- Demo / Examples

About Speakers

Ankit Rathi

- 14+ years of IT experience, mainly into data & analytics.
- 5 years of experience in Data Science.
- Area of Interest: Data Science, Data Architecture, Data Engineering.
- Education : B.Tech HBTI Kanpur (Electronics)

Yatin Bhatia

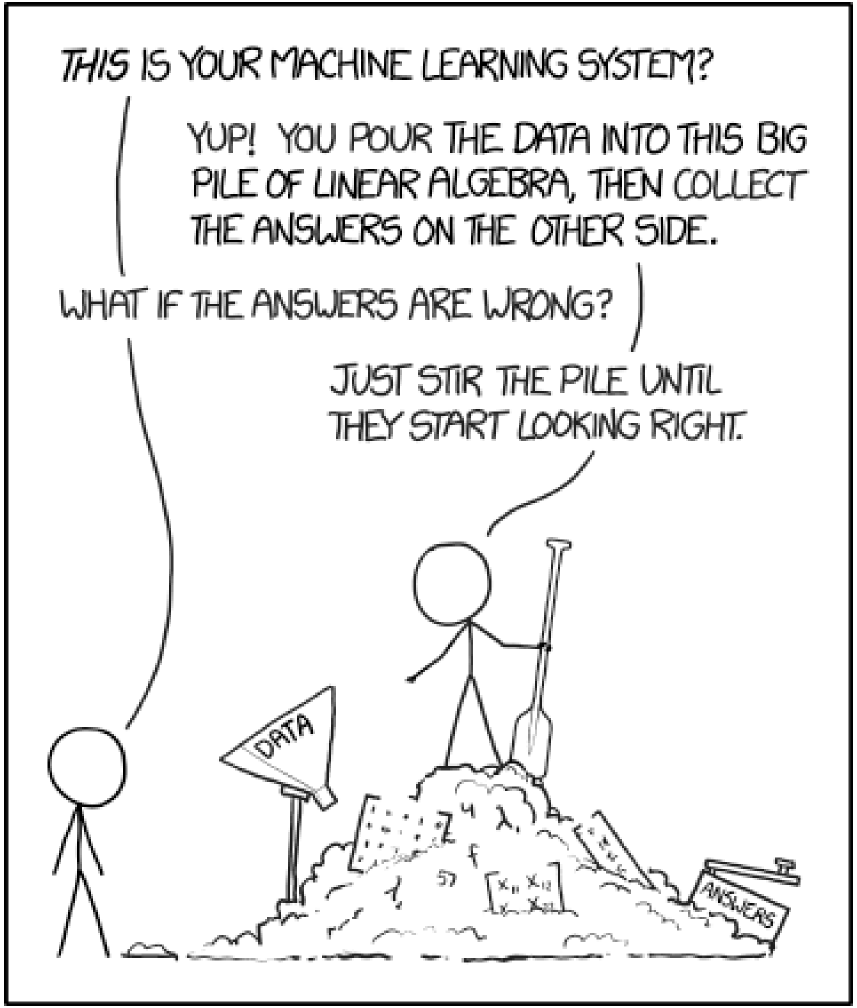
- 13+ years of IT experience.
- 6 years of experience in Data Science.
- Area of Interest: Machine Learning, Data Mining, Deep Learning ,Big Data.
- Education : M.Tech IIT Delhi (Computers)

Let me start with a story…


737 MAX
167
BOEING
A380
737
LEAP

Context

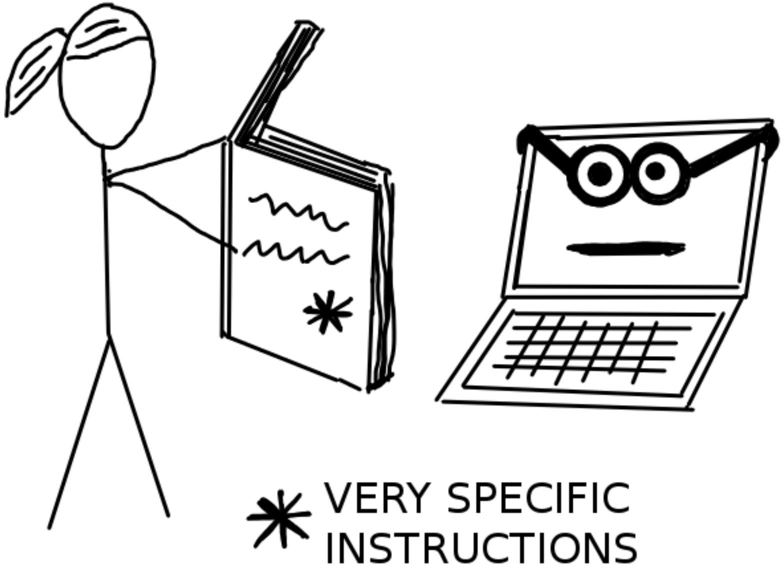
- AI (ML/DL) has evolved a lot in the last decade
- From academia/research to industry adoption
- Industry focus on ‘applied’ AI (ML/DL)
- Effective application is paramount


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Motivation

- ML/DL models learn patterns and relationships
- Challenge is to explain it to business
- Regulatory requirements in some domains
- Certain models have inherent bias

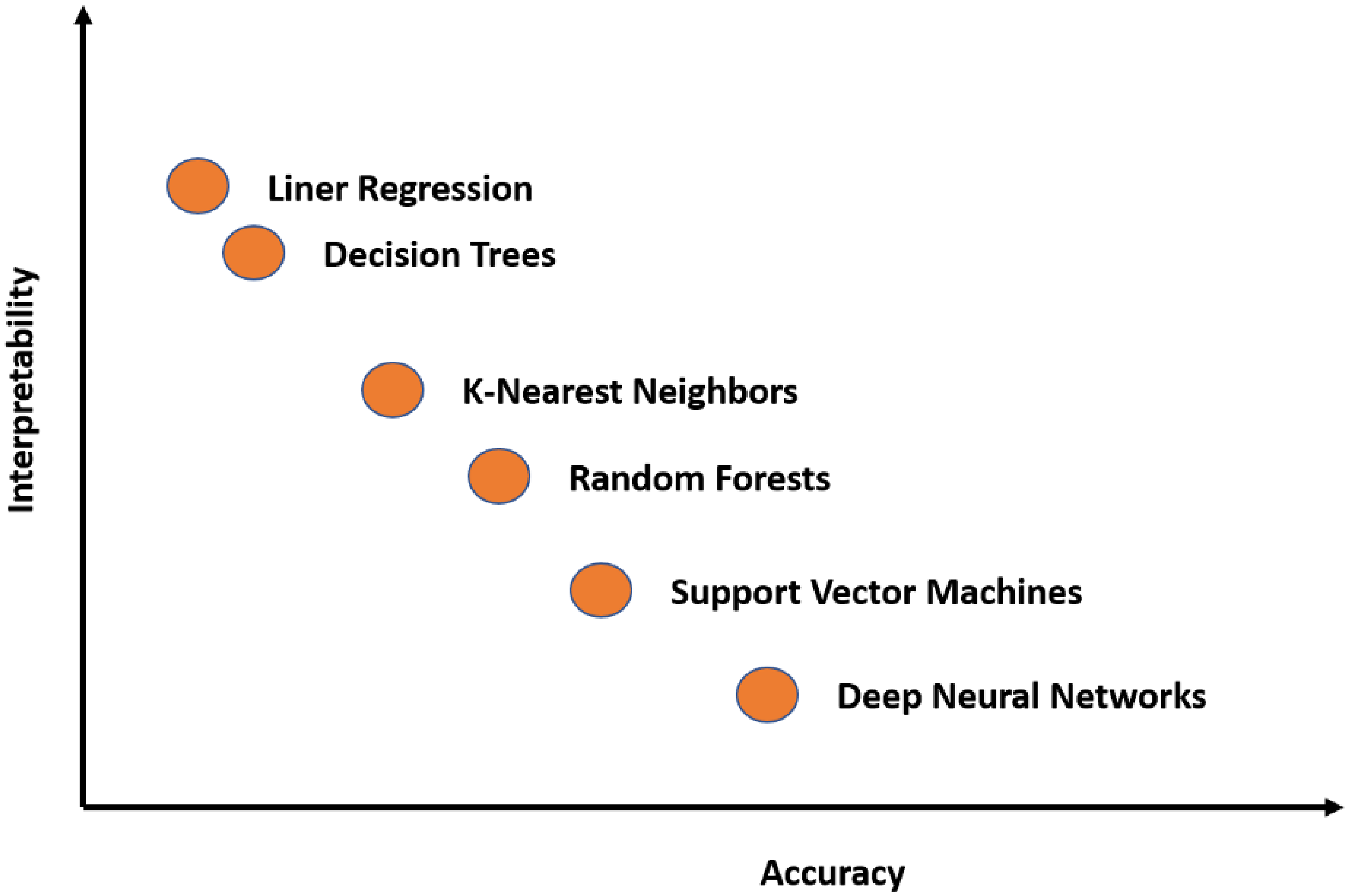
Without Machine Learning



With Machine Learning

Motivation (Contd…)

- Model Interpretability is important
- Interpretable models have inherent problems (i.e. high bias in linear models & high variance in tree models)
- This often leads to a sacrifice in performance
- Interpretability vs Performance


Liner Regression
Decision Trees
K-Nearest Neighbors
Random Forests
Support Vector Machines
Deep Neural Networks
Interpretability
Accuracy

Real-world Use Cases

- Predicting potential criminals
- Credit Scoring
- Fraud Detection
- Health Assessment
- Loan Lending

Understanding Interpretation

- Model is basically a response function
- Understand/Explain response function
- **What** drives model predictions? (ability to question — fairness)
- **Why** did the model take a certain decision? (ability to justify — accountability)

System that performs behaviour but you don’t know how it works

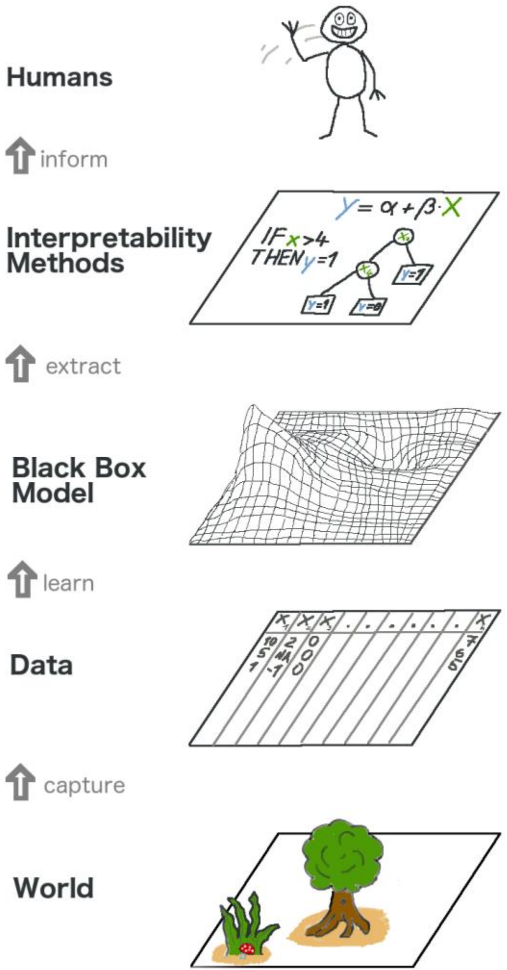
Understanding Interpretation

- **How** can we trust model predictions? (ability to validate — transparency)

- Besides model performance, human understand.ing is important

Types of Interpretation

- Intrinsic Vs Post hoc
- Model-specific Vs Model-agnostic
- Local or Global


Humans
inform
Interpretability Methods
y = α + β·X
IF x>4
THEN y=1
extract
Black Box Model
learn
Data
capture
World

Scope of Interpretation

- Global Interpretation (How?)
- Local Interpretation (Why?)
- Model Transparency (How?)

Model Interpretability - Necessity by Law

GDPR: Right to be Informed

...the controller shall, at the time when personal data are obtained, provide the data subject with the following further information:

- the existence of automated decision-making... meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

Article 13, GDPR

Model Interpretability - Necessity by Bias

Models Can Be(come) Racist & Sexist

```
In [7]: model.most_similar(positive=['computer_programmer', 'woman'], negative=['man'])

Out[7]: [('homemaker', 0.5627118945121765),
         ('housewife', 0.5105047225952148),
         ('graphic_designer', 0.505180299282074),
         ('schoolteacher', 0.497949421405792224),
```

```
In [10]: model.most_similar(positive=['mexicans'], topn=30)

Out[10]: [('hispanics', 0.7345616817474365),
          ('latinos', 0.6618988513946533),
          ('ILLEGALS', 0.6574230194091797),
          ('LEGAL_immigrants', 0.6541558504104614),
          ('mexican', 0.6493428945541382),
          ('thats_ok', 0.6343405246734619),
          ('americans', 0.6324713230133057),
          ('illegals', 0.6298996210098267),
          ('ILLEGAL_aliens', 0.6289116144180298),
```

Model Interpretability - Kaggelification

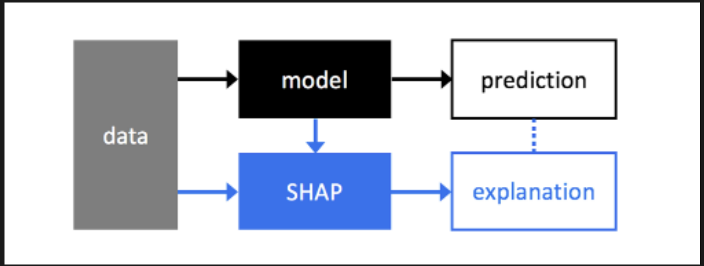
Kagglefication: Death by 1,000 Models

Ensembling of different types of models is part of Kaggle 101. If you don't do it, you're at a disadvantage. Now, should you do it in a business environment? That's a very different question. But in Kaggle you should.

- Quora, Giuliano Janson

Model Interpretability - SHAP VALUES

- SHAP (SHapley Additive exPlanation)




```
SHAP has the following explainers: deep, gradient, kernel, linear, tree, sampling
```

Model Interpretability - LIME

LIME (Local Interpretable Model-agnostic Explanations) builds sparse linear models around each prediction to explain how the black box model works in that local vicinity.

Steps to Calculate LIME Values:

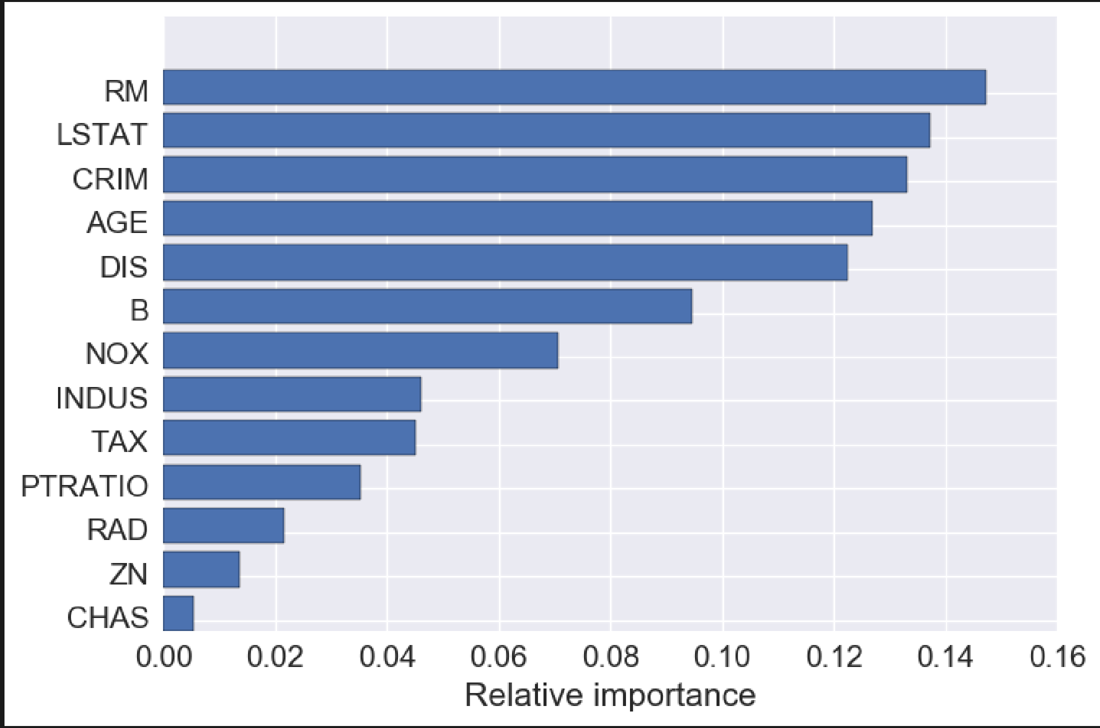
- Select your instance of interest for which you want to have an explanation of its black box prediction.
- Perturb your dataset and get the black box predictions for these new points.
- Weight the new samples according to their proximity to the instance of interest.
- Train a weighted, interpretable model on the dataset with the variations.
- Explain the prediction by interpreting the local model.

Model Interpretability - Boston Housing Data Set

```
Boston house prices dataset
---------------------------

**Data Set Characteristics:**

    :Number of Instances: 506

    :Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.

    :Attribute Information (in order):
        - CRIM     per capita crime rate by town
        - ZN       proportion of residential land zoned for lots over 25,000 sq.ft.
        - INDUS    proportion of non-retail business acres per town
        - CHAS     Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
        - NOX      nitric oxides concentration (parts per 10 million)
        - RM       average number of rooms per dwelling
        - AGE      proportion of owner-occupied units built prior to 1940
        - DIS      weighted distances to five Boston employment centres
        - RAD      index of accessibility to radial highways
        - TAX      full-value property-tax rate per $10,000
        - PTRATIO  pupil-teacher ratio by town
        - B        1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
        - LSTAT    % lower status of the population
        - MEDV     Median value of owner-occupied homes in $1000's
```

Global Interpretability - Feature Importance

(using sklearn- GradientBoostingRegressor)

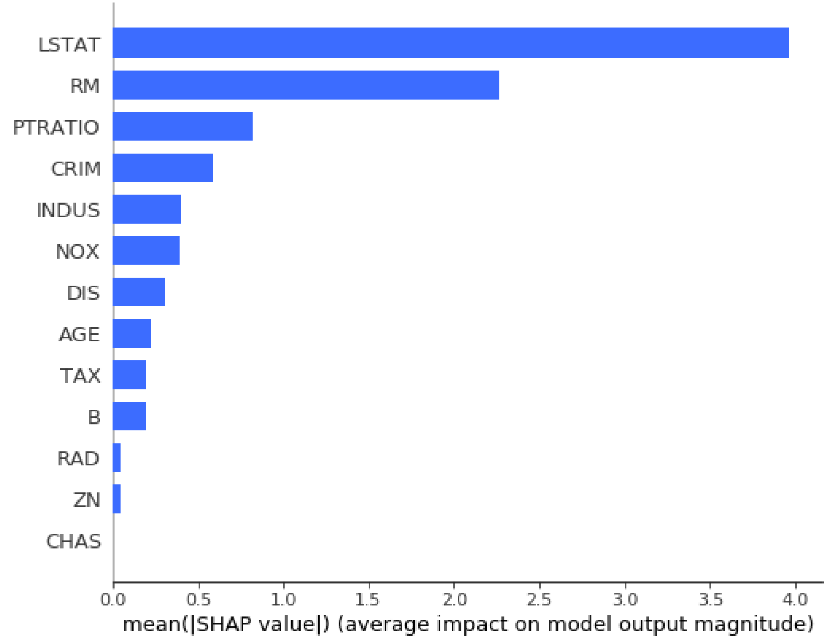


Global Interpretability - Feature Importance

Only SHAP provides global interpretability, LIME does not provide.

Importance plot via SHAP values

```
: shap.summary_plot(shap_values_XGB_train, X_train, plot_type="bar")
```



Local Interpretability - SHAP Values

Local interpretability of models consists of providing detailed explanations for why an individual prediction was made.

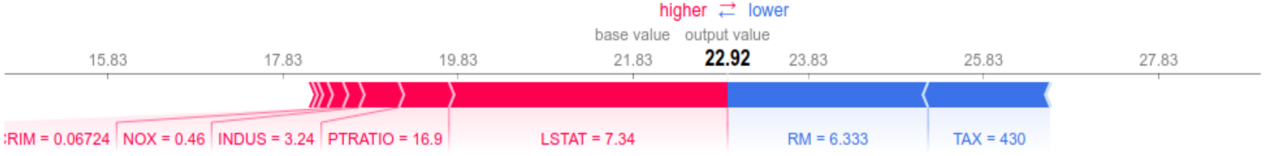
```
In [42]: X_test.iloc[[j]]
```

Out[42]:

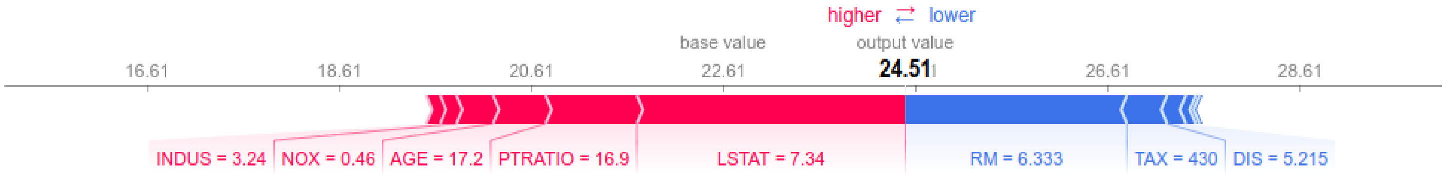
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
329	0.06724	0.0	3.24	0.0	0.46	6.333	17.2	5.2146	4.0	430.0	16.9	375.21	7.34

```
In [37]: shap.force_plot(explainerXGB.expected_value, shap_values_XGB_test[j], X_test.iloc[[j]])
```

Out[37]:




```
shap.force_plot(explainerSKGBT.expected_value, shap_values_SKGBT_test[j], X_test.iloc[[j]])
```

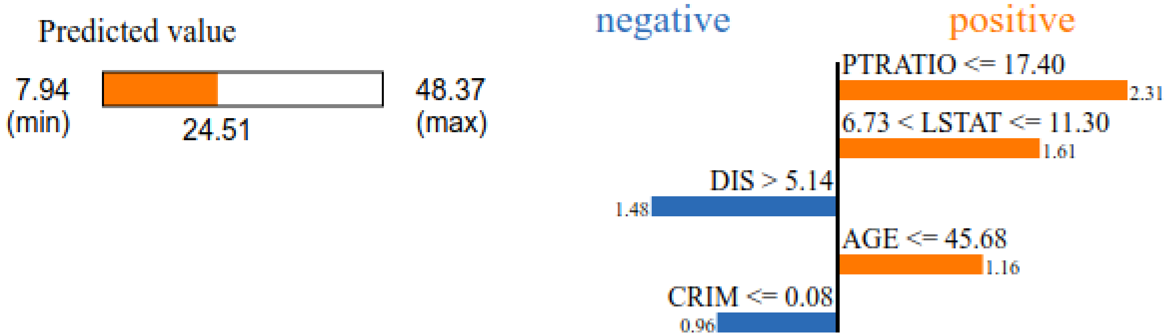


	LSTAT
count	404.000000
mean	12.706188
std	7.299031
min	1.730000
25%	6.727500
50%	11.300000
75%	17.112500
max	36.980000

Local Interpretability - LIME

```
expSKGBT = explainer.explain_instance(X_test.values[j], sk_xgb.predict, num_features=5)
expSKGBT.show_in_notebook(show_table=True)
```

```
Intercept 22.69430655112992
Prediction_local [25.32416513]
Right: 24.509385768795976
```



Feature	Value
PTRATIO	16.90
LSTAT	7.34
DIS	5.21
AGE	17.20
CRIM	0.07

Inculcating Model Interpretability

Teach, Practice, Preach Interpretability

- Include sections on interpretability and introspection in your curriculum, blog posts and talks.
- Work on difficult problems in the interpretability space and share your results.
- Add sample explanations and model or architecture introspection to your daily workflow.
- Talk with your colleagues and peers about how we can *all* work together to improve model accountability.

Inculcating Model Interpretability

Embrace Interpretable Model Engineering

- When feature engineering, ask yourself: am I doing the Kaggle thing again?
- Challenge yourself to find the MVP of models: what is the minimal amount of preprocessing and engineering I can do to make this work in a feasible way?
- Work on an interpretability metric for your team or end user and strive to achieve a high score.

Questions?

Contact Us:

https://ankitrathi.com

https://www.linkedin.com/in/yatin-bhatia-241a996/