«Human-Centered Data Science»

Next week: Ensuring Transparency in your Data Science Workflow

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

June 16, 2022

Lecture Overview

Recap

The Need for Transparency

Scopes of Transparency (Openess, Interpretability, Direction of Interpretability, Intrinsic Interpretability, Properties of Interpretable Models - Linearity)

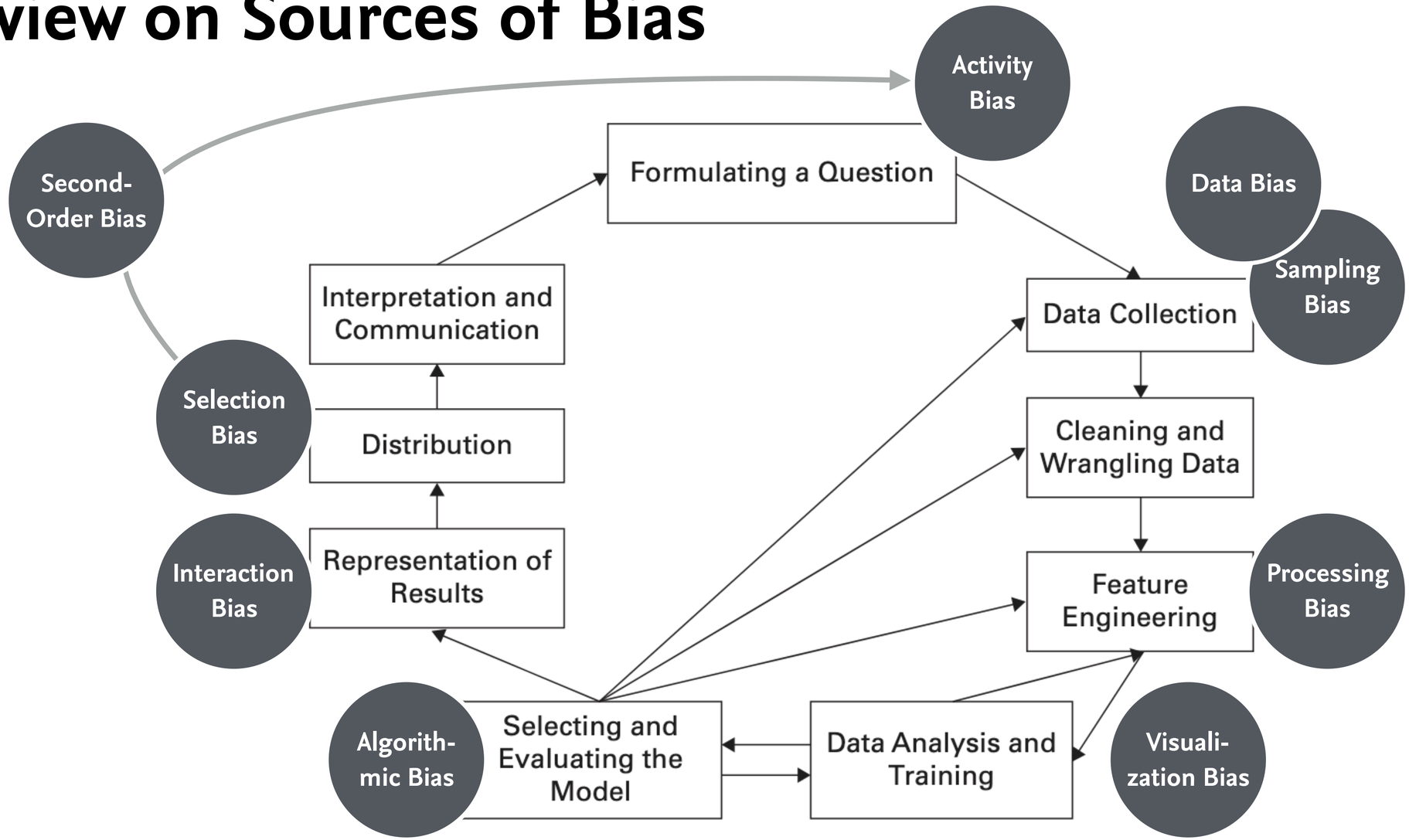
Break

Scopes of Transparency (Properties of Interpretable Models - Interaction, Summary)

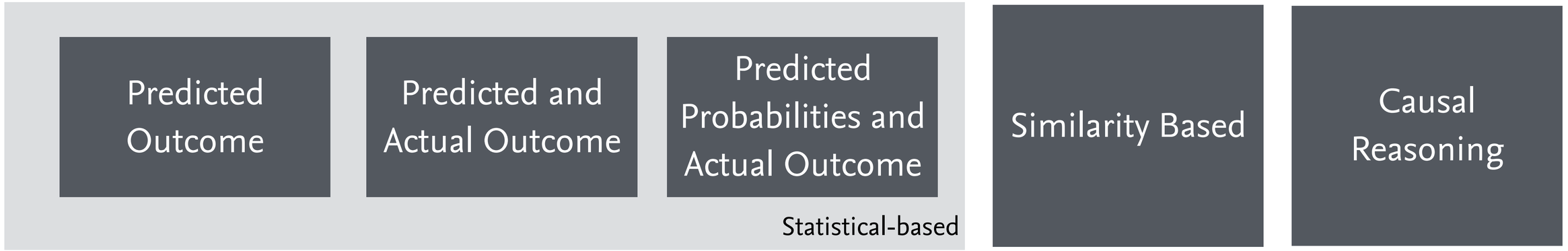
Requirements beyond Transparency

Recap

Overview on Sources of Bias



Overview on Approaches to Define Fairness



Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

What does it mean for our Scenario?

Statistical-based

- Predicted Outcome
- Predicted and Actual Outcome
- Predicted Probabilities and Actual Outcome
- Similarity Based
- Causal Reasoning

Group Fairness ✓

Predictive parity ✓

Test-fairness ✗ ✓

Conditional statistical parity ✗

False positive error rate balance ✗

Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare@ICSE, 1–7. http://doi.org/10.1145/3194770.3194776

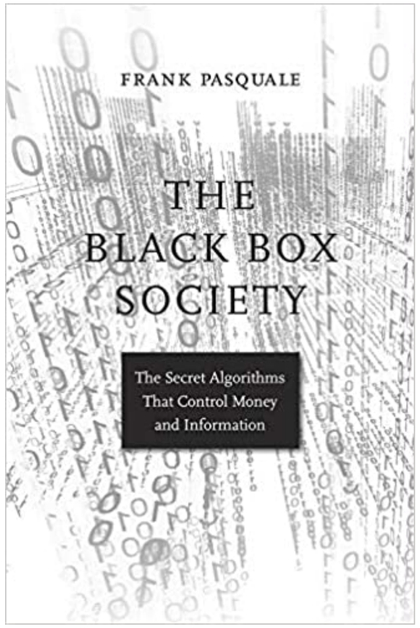
The Need for Transparency

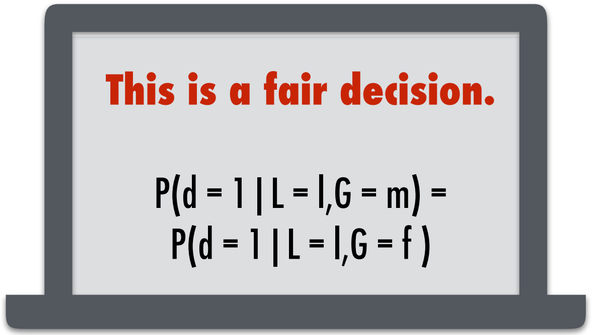
Motivating the Need for Transparency

Why is this fair?

What do I have to do to make it a fair solution?

Why is this fair but not the other approach?





What are the assumptions?

How can I correct an error?

When can I trust this decision?

Questions adapted from D. Gunning, Explainable artificial intelligence (xAI), Tech. rep., Defense Advanced Research Projects Agency (DARPA) (2017).
https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Defining Transparency

> *Transparency refers to the need to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environment, and to the governance of the data used or created.*

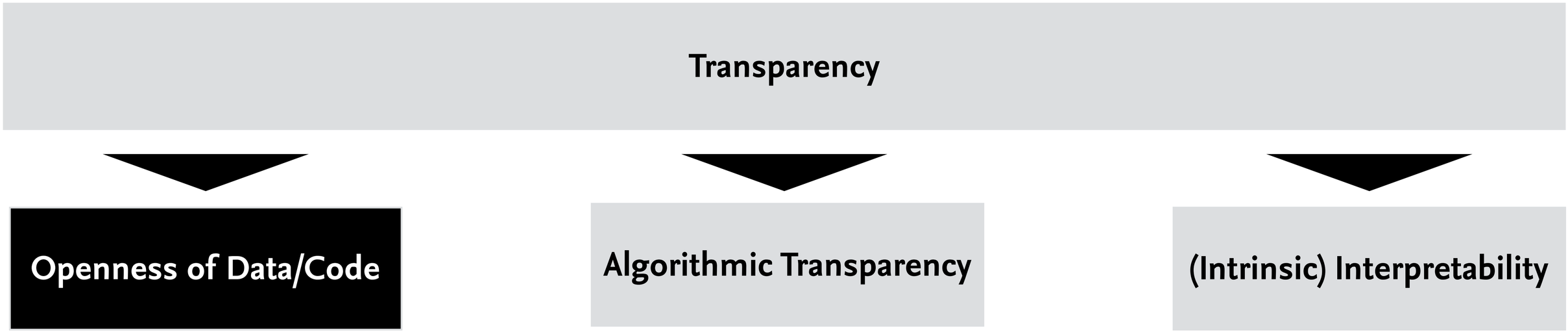
Dignum, Virginia. "Responsible artificial intelligence: designing AI for human values." JTU J., ICT Discoveries, vol. 1, pp. 1–8, Sep. 2017.

Notions of Transparency

Transparency

- **Openness of Data/Code**
- **Algorithmic Transparency**
- **(Intrinsic) Interpretability**

Notions of Transparency: Openness



Transparency means Openness

Transparency is used to describe **openness**.

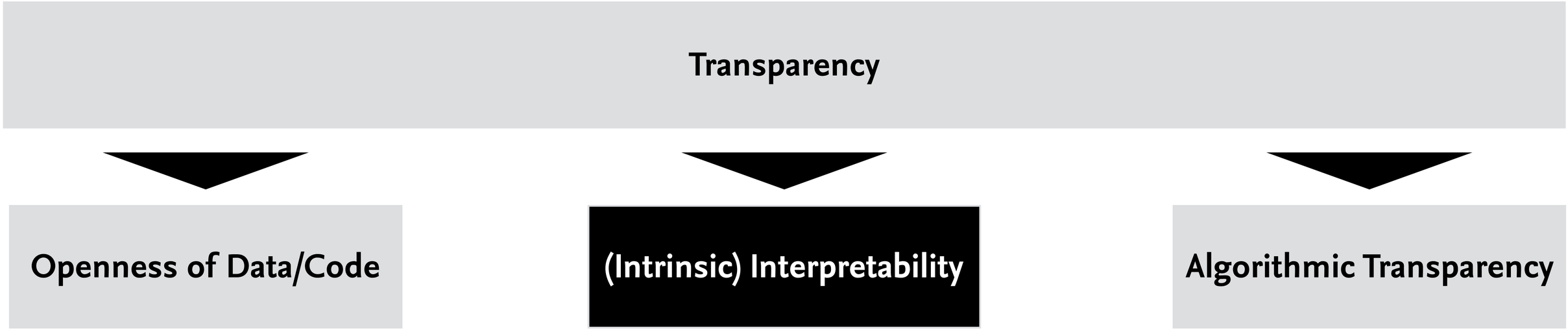
From a technical perspective, openness comprises that

- the model (code) and training/test data are publicly inspectable,
- individual decisions are reproducible, and
- changes are logged and version controlled.

} **Accountability**

Kohli, N., Barreto, R., & Kroll, J. A. (2018). Translation tutorial: a shared lexicon for research and practice in human-centered software systems. In 1st Conference on Fairness, Accountability, and Transparancy. New York, NY, USA.

Notions of Transparency: Interpretability



Defining Interpretability

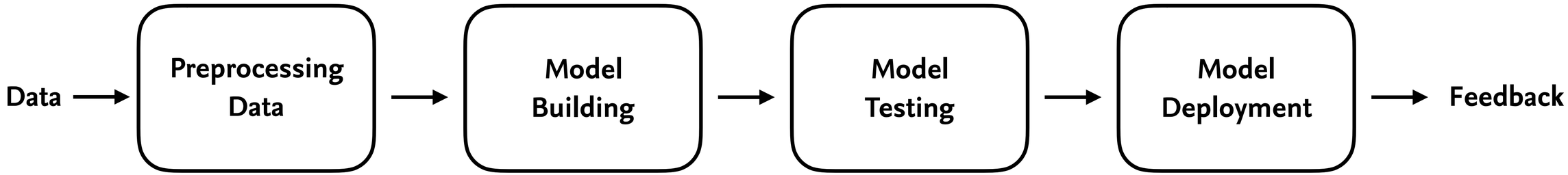
> Kim et al. describe interpretability as
> ***"the degree to which a human can consistently predict the model's result".***

> Molnar notes that
> ***"interpretable machine learning refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans".***

> Doshi-Velez and Kim define interpretability as the
> ***"ability to explain or to present in understandable terms to a human".***

Kim, B.; Khanna, R.; Koyejo, O.O. Examples are not enough, learn to criticize! Criticism for interpretability. In Advances in Neural Information Processing Systems; MIT Press, 2016; pp. 2280–2288.
Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.
Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).

Directions of Interpretability



Pre-Model	In-Model	Post-Model
» Focus on exploratory data analysis techniques and visualization	» Focus on intrinsically interpretable models	» Focus on improving interpretability after building a model (post hoc)

Intrinsic Interpretability

Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso. "Machine learning interpretability: A survey on methods and metrics." Electronics 8.8 (2019): 832.

Intrinsic Interpretability

Intrinsic interpretability answers the question of **how the model works**.

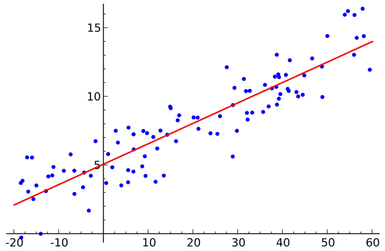
Intrinsic interpretability refers to models that are **interpretable by themselves**. It can be achieved, for example, through the imposition of constraints on the model.

A subset of algorithms create **interpretable models**, such as linear regression, logistic regression and decision trees.

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Selected Properties of Interpretable Models

Linearity: A model is linear if the association between feature values and target values is modelled linearly.



Monotonicity: Enforcing monotonicity constraints on the model guarantees that the relationship between a specific input feature and the target outcome always goes in the same direction over the entire feature domain.

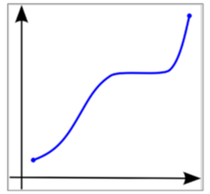


Figure 1 - A monotonically increasing function

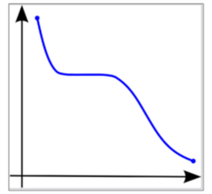


Figure 2 - A monotonically decreasing function

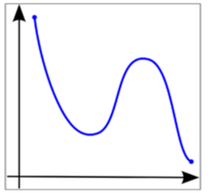


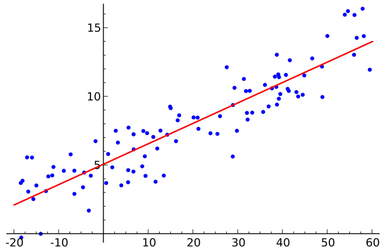
Figure 3 - A function that is not monotonic

Interaction: Some models have the ability to naturally include interactions between features to predict the target outcome.

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.
Graphs Public Domain from https://en.wikipedia.org/wiki/Generalized_linear_model and https://en.wikipedia.org/wiki/Monotonic_function

Selected Properties of Interpretable Models

Linearity: A model is linear if the association between feature values and target values is modelled linearly.



Monotonicity: Enforcing monotonicity constraints on the model guarantees that the relationship between a specific input feature and the target outcome always goes in the same direction over the entire feature domain.

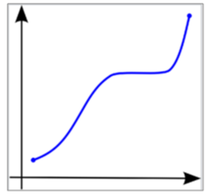


Figure 1 - A monotonically increasing function

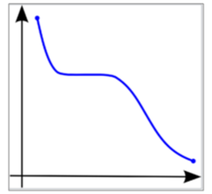


Figure 2 - A monotonically decreasing function

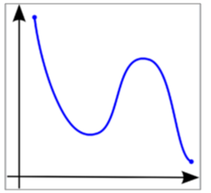


Figure 3 - A function that is not monotonic

Interaction: Some models have the ability to naturally include interactions between features to predict the target outcome.

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.
Graphs Public Domain from https://en.wikipedia.org/wiki/Generalized_linear_model and https://en.wikipedia.org/wiki/Monotonic_function

Bike Sharing Dataset Data Set

- Count of bicycles including both casual and registered users.
- The season (spring, summer, fall or winter).
- Indicator, whether the day was a holiday or not.
- The year, either 2011 or 2012.
- Number of days since the 01.01.2011.
- Indicator whether the day was a working day or weekend.
- The weather situation on that day. One of:
  - clear, few clouds, partly cloudy, cloudy
  - mist + clouds, mist + broken clouds, mist + few clouds, mist
  - light snow, light rain + thunderstorm + scattered clouds, light rain + scattered clouds
  - heavy rain + ice pallets + thunderstorm + mist, snow + mist
- Temperature in degrees Celsius.
- Relative humidity in percent (0 to 100).
- Wind speed in km per hour.

Photo by planimetrica on Unsplash
https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Estimated Regression Weights

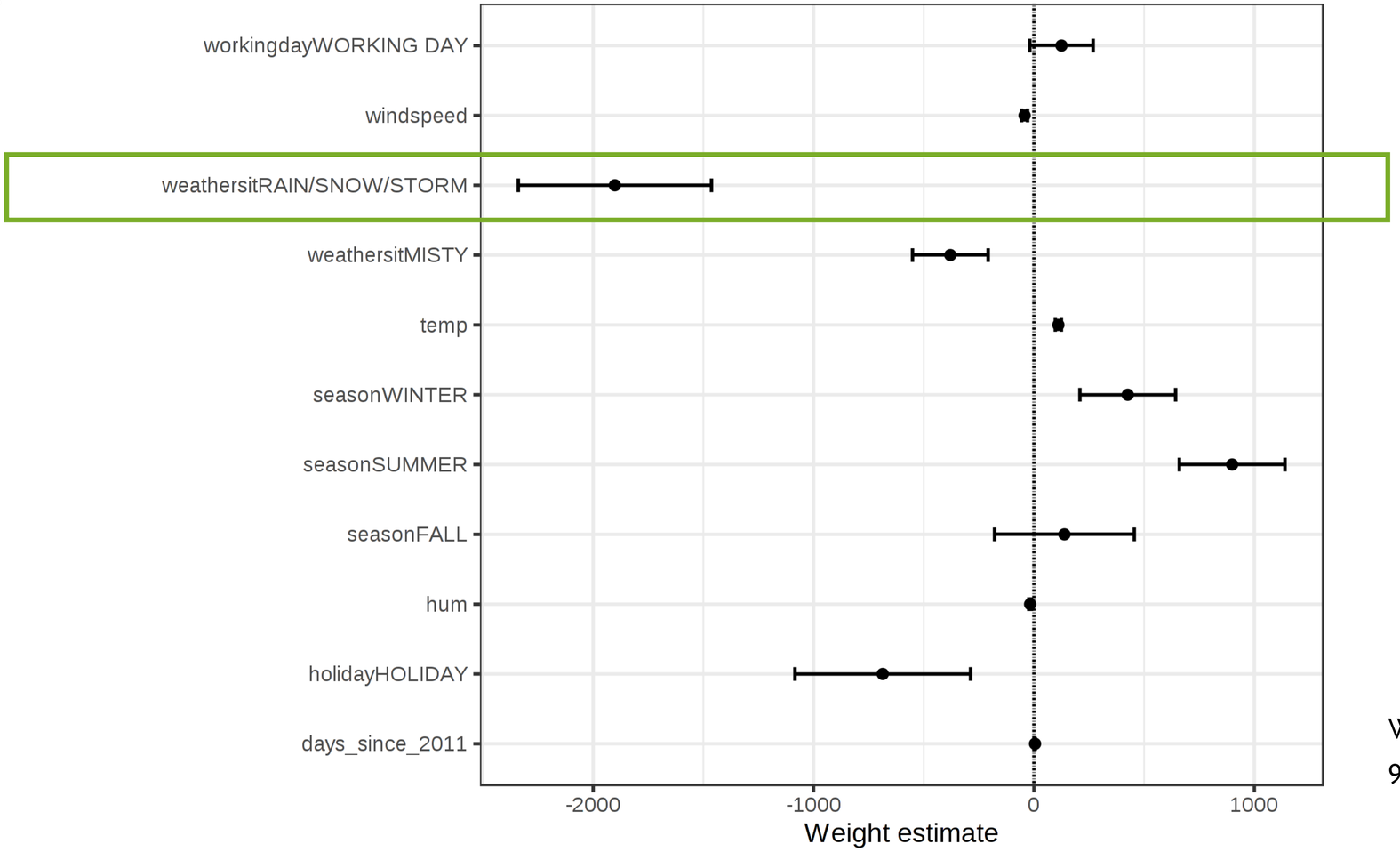
	Weight	SE	|t|
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

Weight: Estimated weight
SE: the standard error of the estimate
|t|: feature importance, i.e., absolute value of the t-statistic

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

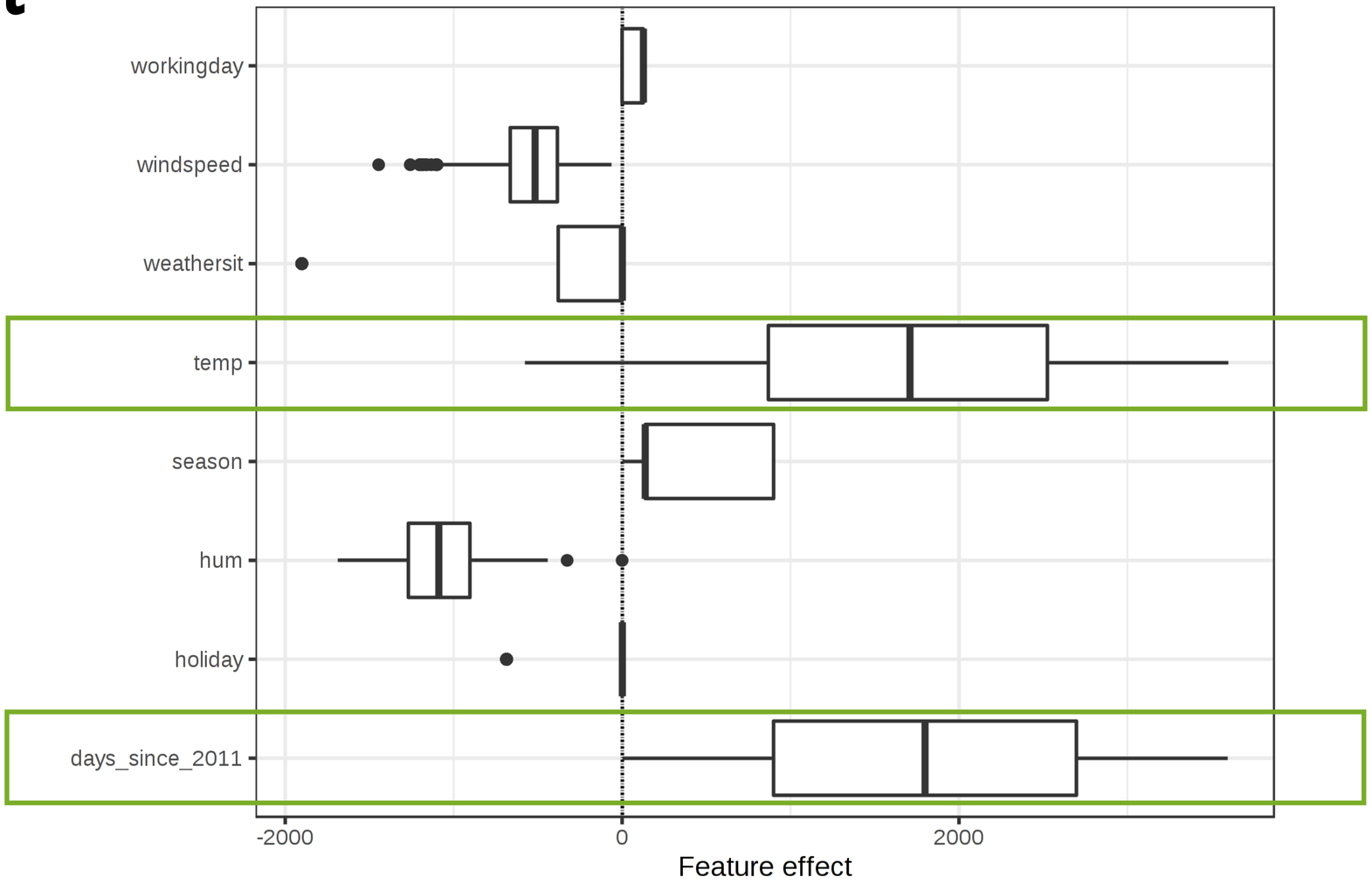
Weight Plot



Weights are displayed as points and the 95% confidence intervals as lines.

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

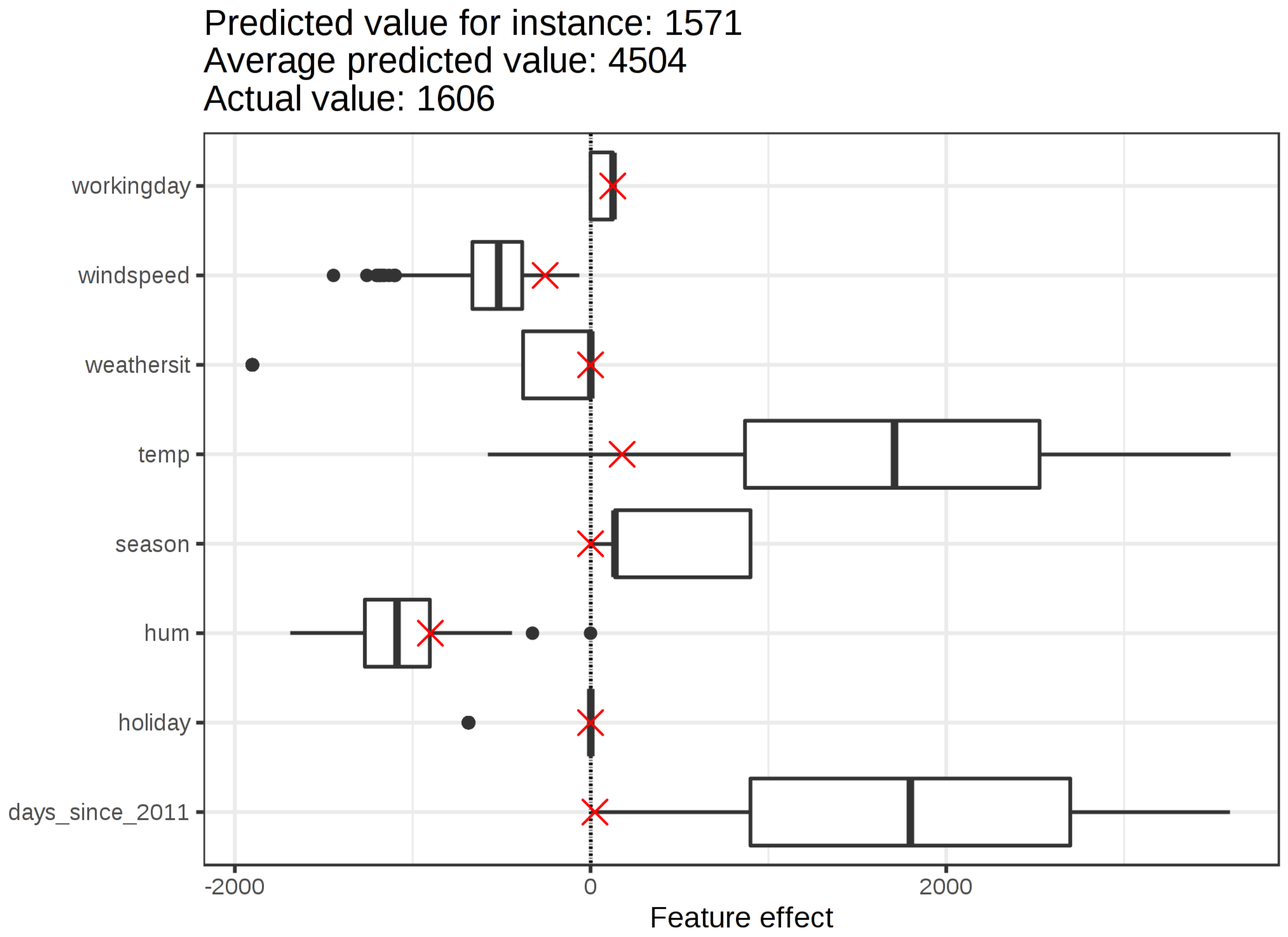
Effect Plot



Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Explain Individual Predictions: Effect Plot for one Instance

Feature	Value
season	SPRING
yr	2011
mnth	JAN
holiday	NO HOLIDAY
weekday	THU
workingday	WORKING DAY
weathersit	GOOD
temp	1.604356
hum	51.8261
windspeed	6.000868
cnt	1606
days_since_2011	5



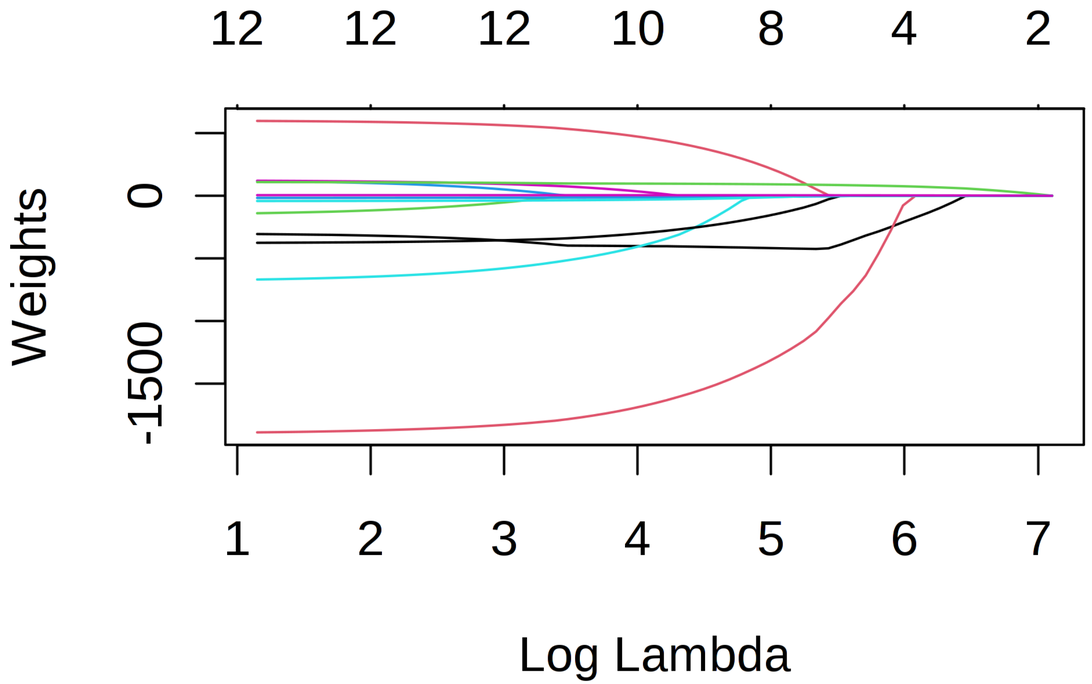
Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Adding Interpretability Constraints by Lasso Regression

Challenge to interpretability is the number featured used in Linear Regression Models.

Lasso ("least absolute shrinkage and selection operator") is an automatic and convenient way to **introduce sparsity** into the linear regression model.

It performs feature selection and regularization of the selected feature weights.



Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Pros and Cons for Interpreting Linear Regression

Advantages

» The modeling of the predictions as a weighted sum makes it transparent how predictions are produced.

» With Lasso you can ensure that the number of features used remains small.

» A lot of experience and expertise, including teaching materials.

» Many software implementations in R, Python, Java, etc.

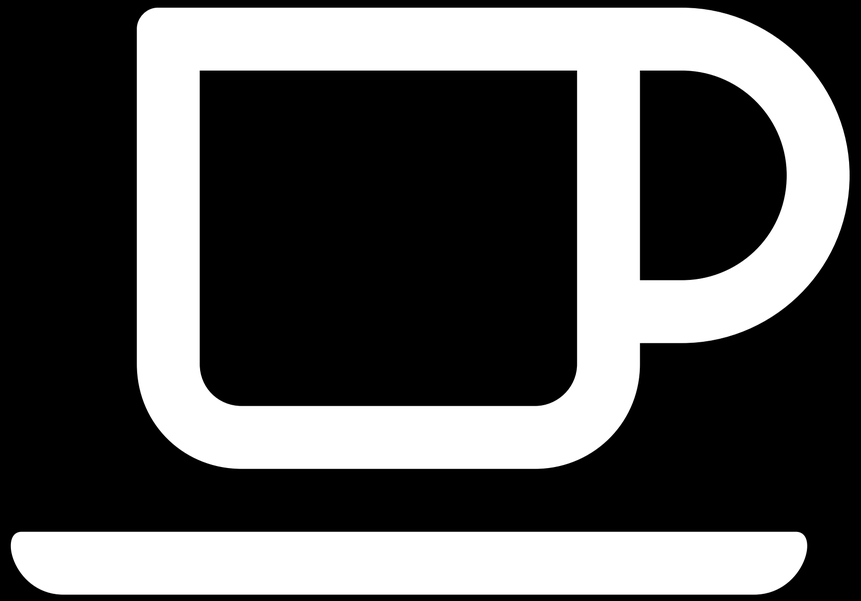
Disadvantages

» Linear regression models can only represent linear relationships.

» Relationships that can be learned are very restricted.

» The interpretation of a weight can be unintuitive because it depends on all other features.

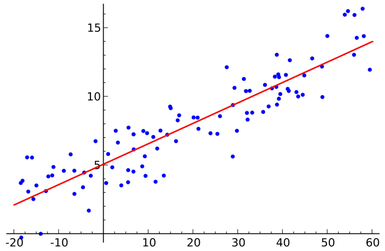
Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.



5 minutes break

Selected Properties of Interpretable Models

Linearity: A model is linear if the association between feature values and target values is modelled linearly.



Monotonicity: Enforcing monotonicity constraints on the model guarantees that the relationship between a specific input feature and the target outcome always goes in the same direction over the entire feature domain.

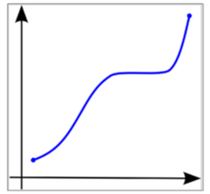


Figure 1 - A monotonically increasing function

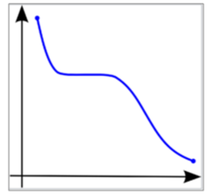


Figure 2 - A monotonically decreasing function

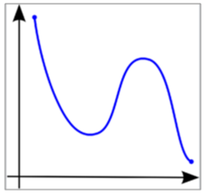
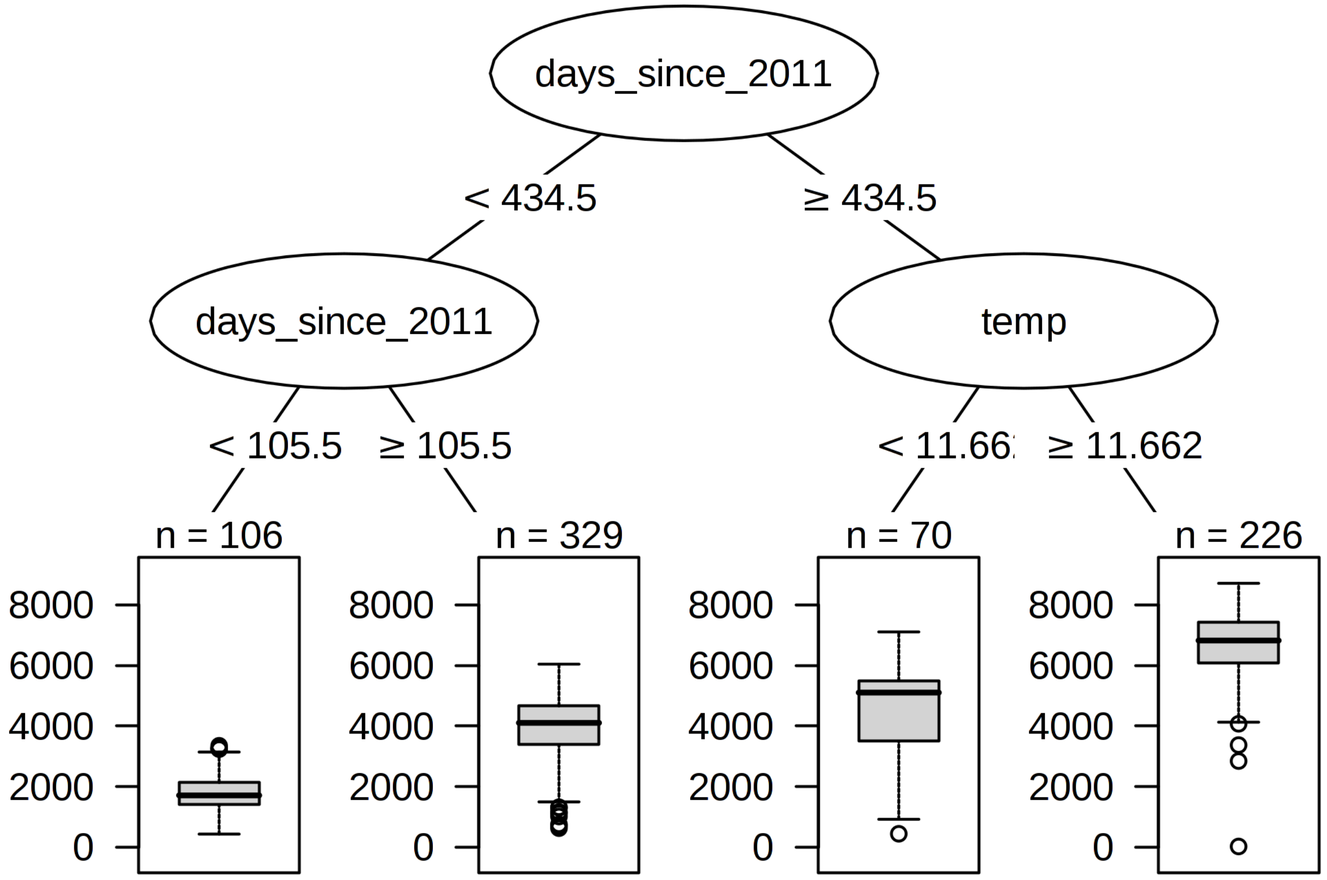


Figure 3 - A function that is not monotonic

Interaction: Some models have the ability to naturally include interactions between features to predict the target outcome.

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.
Graphs Public Domain from https://en.wikipedia.org/wiki/Generalized_linear_model and https://en.wikipedia.org/wiki/Monotonic_function

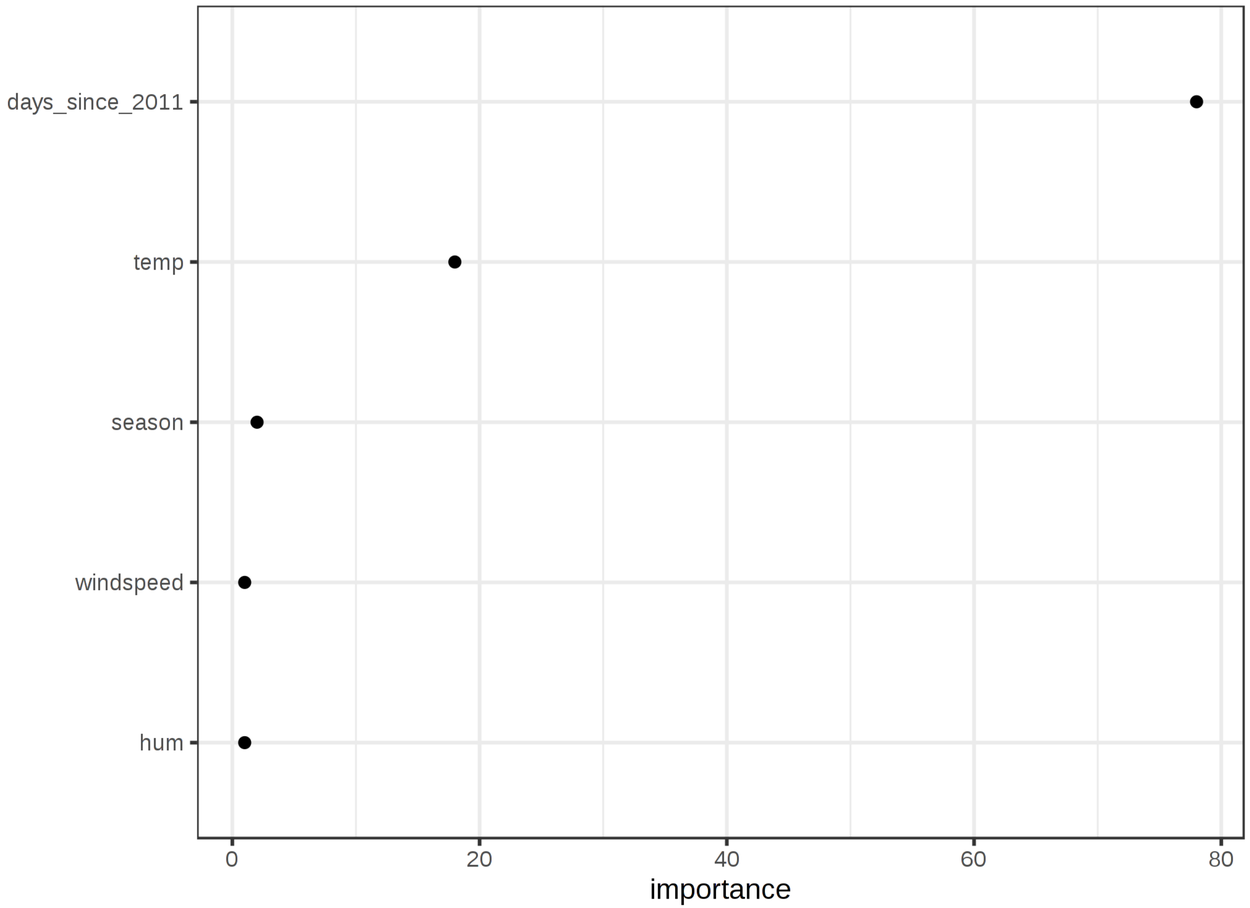
Example of a Decision Tree based on Bike Data



434.5 days after the 01.01.2011, with a temperature above 12 degrees celsius, about 6.600 bikes were requested.

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Importance of the Features



Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.
Check out for an overview https://mlandaiblog.wordpress.com/2019/04/29/feature-importance-and-model-interpretation/

Pros and Cons for Interpreting Decision Trees

Advantages

» It captures interactions between features.

» The data ends up in distinct groups that are often easier to understand than points in linear regression.

» The tree structure also has a natural visualization, with its nodes and edges.

» There is no need to transform features.

Disadvantages

» Trees fail to deal with linear relationships.

» Slight changes in the input feature can have a big impact on the predicted outcome.

» Trees are also quite unstable, i.e, a few changes in the training dataset can create a completely different tree.

» Decision trees are very interpretable - as long as they are short.

» Trees are prone to overfit.

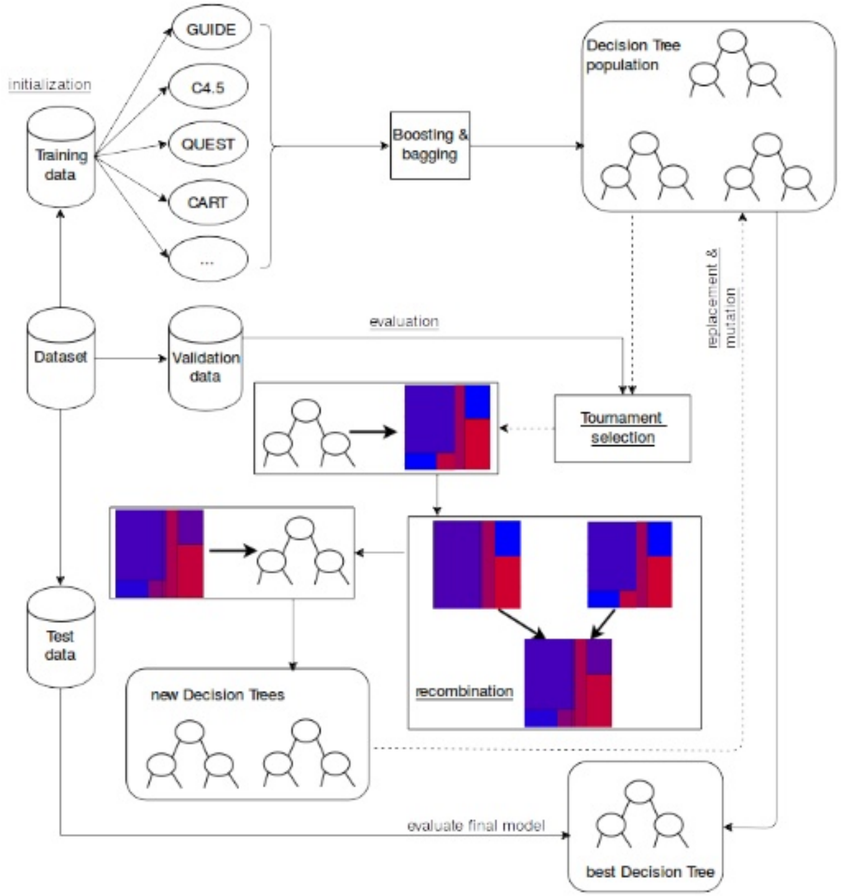
Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Interpretable Model Extraction/Mimic Learning

Ensemble techniques have been proposed to avoid overfitting. However, these ensemble techniques cannot be interpreted.

Mimic learning is to approximate a complex model by using an easily interpretable model. The statistical properties of the complex model will be reflected in the interpretable model.

An example is the GENESIM algorithm.



Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. Commun. ACM, 63(1), 68–77. http://doi.org/10.1145/3359786
Vandewiele, G., Janssens, O., Ongenae, F., De Turck, F., & Van Hoecke, S. (2016, November 17). GENESIM: genetic extraction of a single, interpretable model. arXiv.org.

Overview on Interpretable Model Types and Their Properties

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	Yes	Yes	No	regr
Logistic regression	No	Yes	No	class
Decision trees	No	Some	Yes	class,regr
RuleFit	Yes	No	Yes	class,regr
Naive Bayes	No	Yes	No	class

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Overview on Transparency

Transparency

- **Openness of Data/Code**
- **(Intrinsic) Interpretability**
- **Algorithmic Transparency**

Overview on Transparency: Openness of Data/Code

Transparency

Openness of Data/Code

(Intrinsic) Interpretability

Algorithmic Transparency

Level of the whole pipeline

- Model (code) and training/test data are publicly inspectable.
- Individual decisions are reproducible, and changes are logged and version controlled.

Lipton, Z. C. (2018). The mythos of model interpretability. Commun. ACM, 61(10), 36–43. http://doi.org/10.1145/3233231

Overview on Transparency: (Intrinsic) Interpretability

Transparency

Openness of Data/Code

(Intrinsic) Interpretability

Algorithmic Transparency

Decomposability

Level of individual components

- Each part of the model admits an intuitive explanation.
- Example: Each node in a decision tree might correspond to a plain text description.

Simulatability

Level of the entire model

- A model might be called *transparent* if a person can contemplate the entire model at once.
- Example: Lasso regression is are more interpretable than dense linear models.

Lipton, Z. C. (2018). The mythos of model interpretability. Commun. ACM, 61(10), 36–43. http://doi.org/10.1145/3233231

Overview on Transparency: Algorithmic Transparency

Transparency

Openness of Data/Code

(Intrinsic) Interpretability

Algorithmic Transparency

Level of the learning algorithm

» Transparency applies at the level of the learning algorithm itself.

» Example: Understand the error metrics in linear regression.

Lipton, Z. C. (2018). The mythos of model interpretability. Commun. ACM, 61(10), 36–43. http://doi.org/10.1145/3233231

Take-aways for Your Data Science Practice

Depending on the context, the audience, and the purpose, you can pursue different strategies to improve transparency.

For example:

- Use interpretable models (if possible);
- Use concrete features of the input data, rather than composite features;
- Use fewer features;
- Provide supporting documentation (data, model, etc.) written for (non-)data scientists;
- Make it easy for people to explore various inputs/outputs of your model.

Slide adapted from Jonathan T. Morgan & Oliver Keyes (Course DATA 512) https://wiki.communitydata.science/Human_Centered_Data_Science_(Fall_2018)

Is considering such model perspective enough?

Definitions of Interpretability

> Kim et al. describe interpretability as
> ***"the degree to which a human can consistently predict the model's result".***

> Molnar notes that
> ***"interpretable machine learning refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans".***

> Doshi-Velez and Kim define interpretability as the
> ***"ability to explain or to present in understandable terms to a human".***

Kim, B.; Khanna, R.; Koyejo, O.O. Examples are not enough, learn to criticize! Criticism for interpretability. In Advances in Neural Information Processing Systems; MIT Press, 2016; pp. 2280–2288.
Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.
Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).

Why does Interpretability Matter?

Human understanding improves with the degree to which a system can be interpreted.

There is a strong user **preference for**, and **trust in**, models that exhibit “sound”, i.e., human comprehensible, reasoning and “clear communication” about their decision making.

Users perceive “sound” models as more accurate (which not necessarily correlate with actual or statistical accuracy).

Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. 2007. Toward harnessing user feedback for machine learning. In Proceedings of the 12th international conference on Intelligent user interfaces (IUI '07). Association for Computing Machinery, New York, NY, USA, 82–91. DOI:https://doi.org/10.1145/1216295.1216316

Manipulating and Measuring Model Interpretability

H1. Simulation. A clear model with a small number of features will be easiest for participants to simulate.

H2. Deviation. On typical examples participants will be more likely to follow (that is, less likely to deviate from) the predictions of a clear model with a small number of features than the predictions of a black-box model with a large number of features.

H3. Detection of mistakes. Participants in different conditions will exhibit varying abilities to correct the model's inaccurate predictions on unusual examples.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018, February 21). Manipulating and Measuring Model Interpretability. arXiv.org.

Experimental Design

Task: *Predict apartment prices with the help of a machine learning model.*

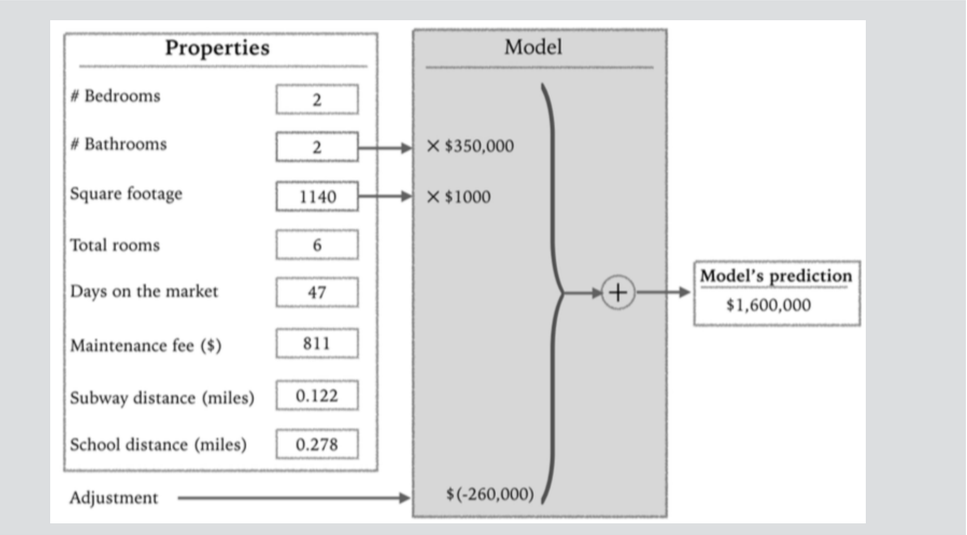
Experimental Conditions:

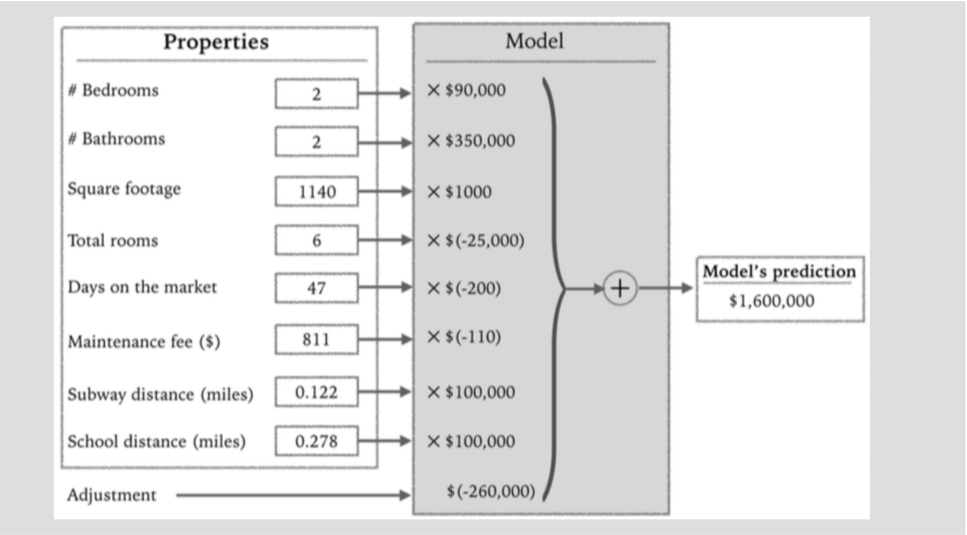
Transparency \ Features		
	Two features	Eight features
White box	White box	White box
Black box	Two features Black box	Eight features Black box

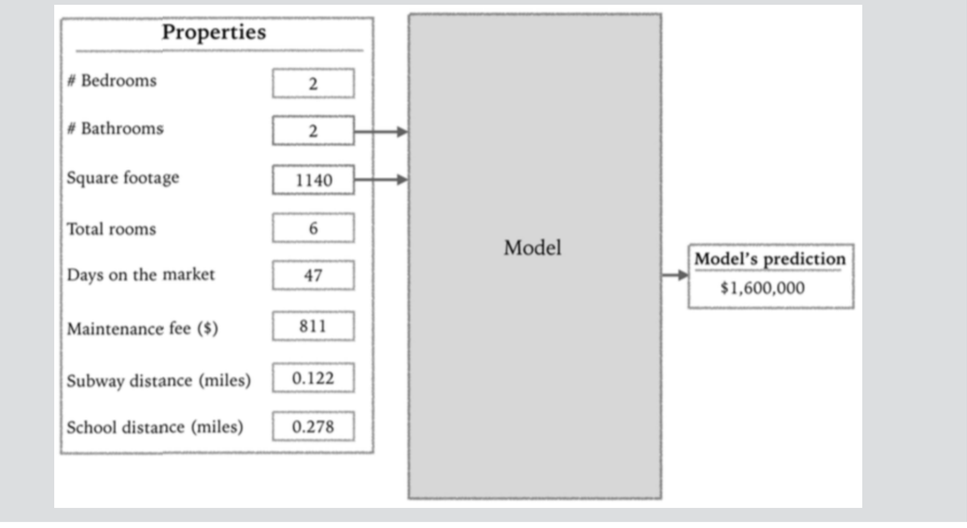
All participants saw the same set of apartments and the same model prediction for each apartment, regardless of their randomly assigned experimental condition.

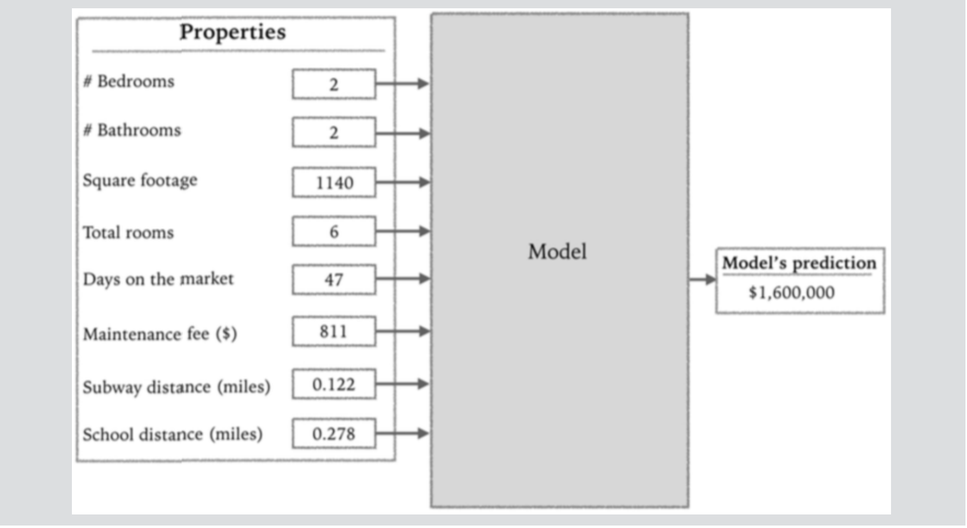
Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018, February 21). Manipulating and Measuring Model Interpretability. arXiv.org.

Experimental Conditions









Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018, February 21). Manipulating and Measuring Model Interpretability. arXiv.org.

Selected Insights

H1. Simulation. Their results indicate that participants who saw a clear model with a small number of features were better able to predict what the model w dict.

H2. Deviation. Their results suggests, no sig participants' deviation from a clear model with a small number of edictions of a black-box model with a large number of features

H3. Detection of mist s. Participants in the clear conditions deviated from the model's prediction less, on average, than participants in the black-box conditions.

Transparency is useless?

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018, February 21). Manipulating and Measuring Model Interpretability. arXiv.org.

Course Evaluation

Check your Insights

How do you differentiate openness, intrinsic interpretability, and algorithmic transparency?

How do interpretability relate to bias and fairness?

What are the three directions of interpretability and what is the focus of each of them?

What are characteristics on interpretable models and what are approaches to make complex models interpretable?

How can you ensure intrinsic interpretability?

Why does interpretability matter from a human-centered design perspective?

«Human-Centered Data Science»

Next week: Post-hoc Interpretability: Approaches and their Limitations

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

June 23, 2022