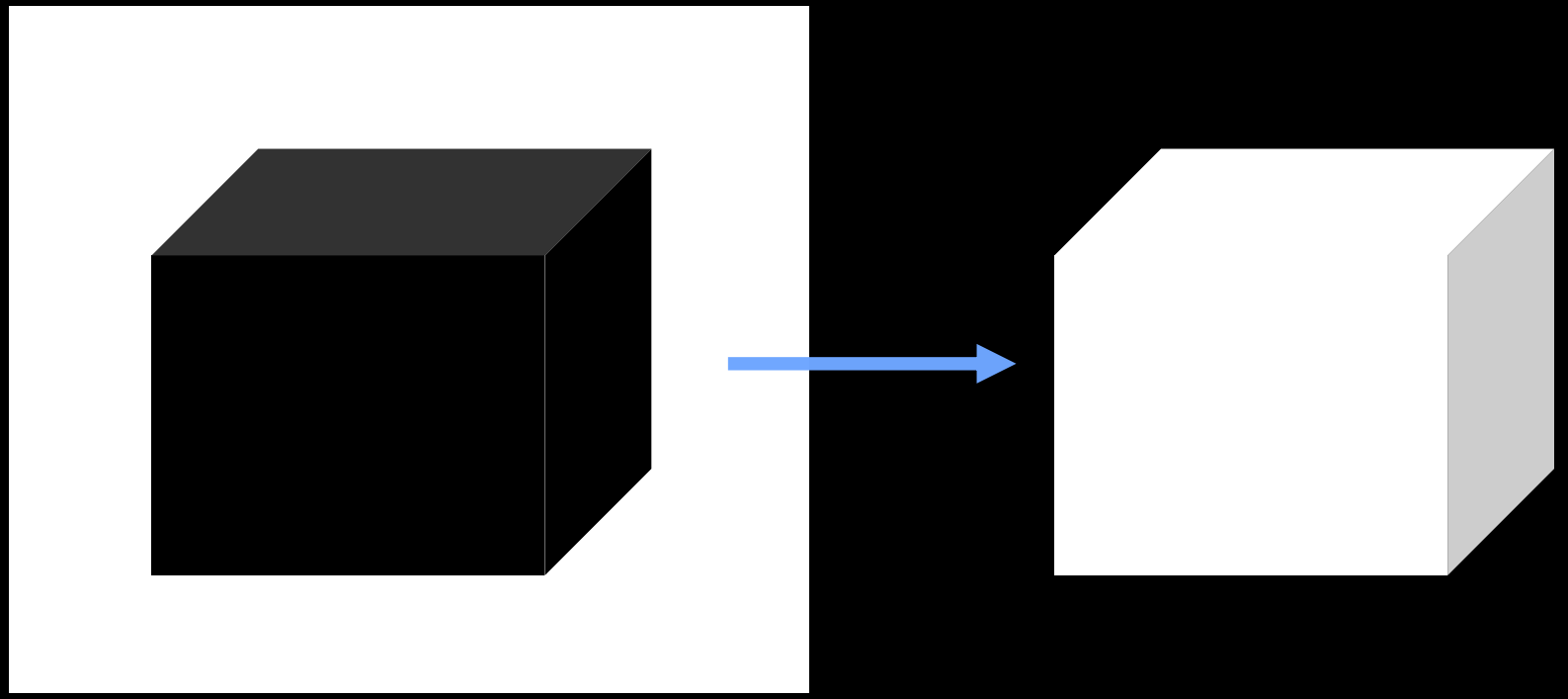
Interpretability in Machine Learning

Kamal Mishra
25-Jan-2020

Agenda

- Background and Context setting
- What is Interpretable ML?
- Why do we need it?
  - Model Bias, Ethics, Fairness
  - Causality of features
  - Ability to debug models & know details
  - Regulatory requirements
  - Trust
  - Critical domain specific need
- When do we NOT need interpretability?
- Framework
- Different flavors of Interpretability
- References

Getting from Black-Box to White-Box ?

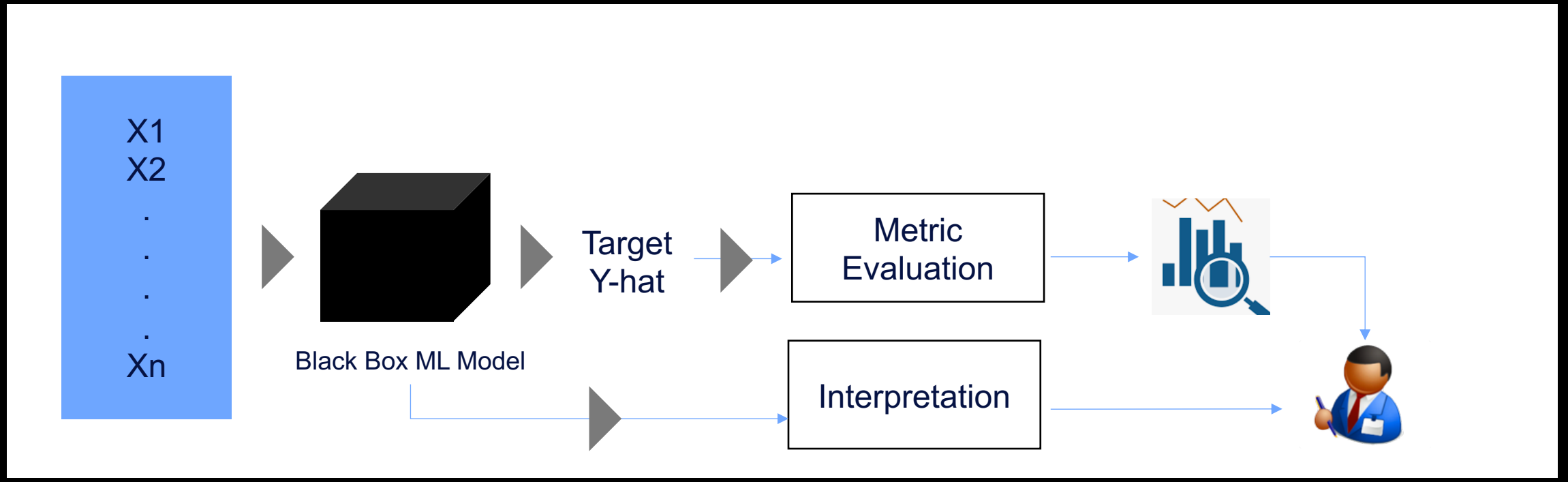


What is Interpretable MLs?

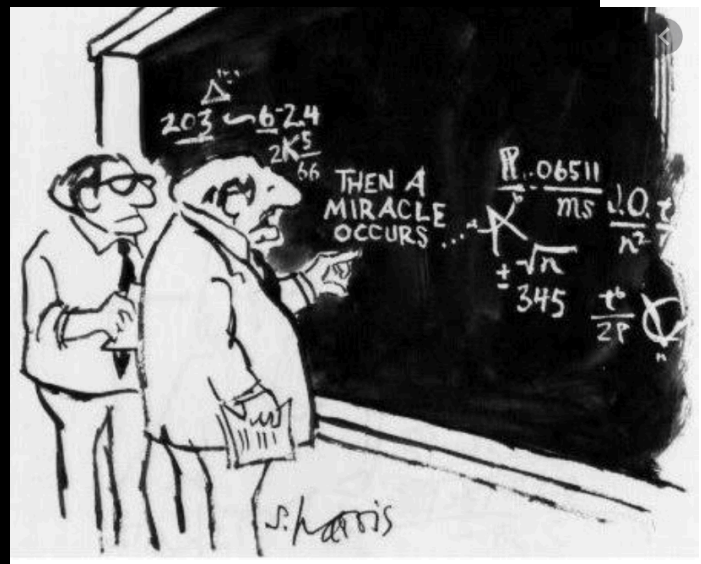
How do we build TRUST in ML models?

How do we know what is happening as part of the ML process?

EXPLAINABILITY? INTERPRETABILITY? MONITORING models?



Why Model Interpretability ?



"I think you should be more explicit here in step two."

from *What's so Funny about Science?* by Sidney Harris (1977)

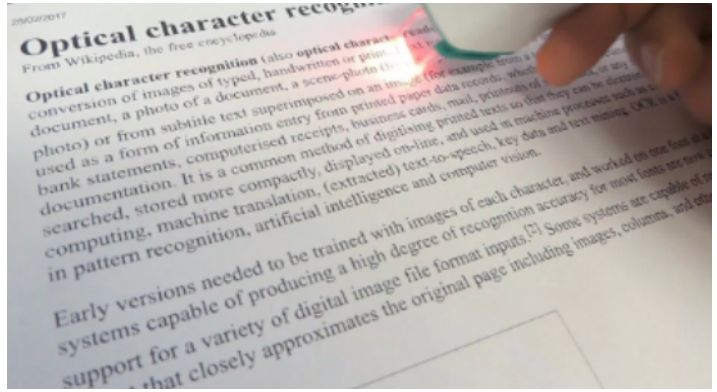
- Model bias, ethics, fairness
  - Does the model discriminate?
  - Promotion of employee example based on last few years of historical data, to promote more men than women?
  - Loan default problem
  - Criminology by geo locations, based on nature of skin color? Or based on dangerous or weird appeal and look?
- Checking causality of features
  - Example of Wolf vs dog classifier on images
  - More data?
  - Different background? And check with validation set
- Ability to debug and know some details
  - Why the model is behaving like this or making this mistake?
- Regulatory requirements
  - Does model satisfy regulatory needs?
  - EU regulation, GDPR Art. 12 – customers can enquire about algorithmic decisions: why loan was rejected?, why low credit limit allowed on CC? etc.
- Trust and understanding
  - How can I understand and trust model's decisions?
- Critical domain and industry related need
  - Finance, Healthcare, Risk, Judicial

Do we need Model Interpretability All the time ?

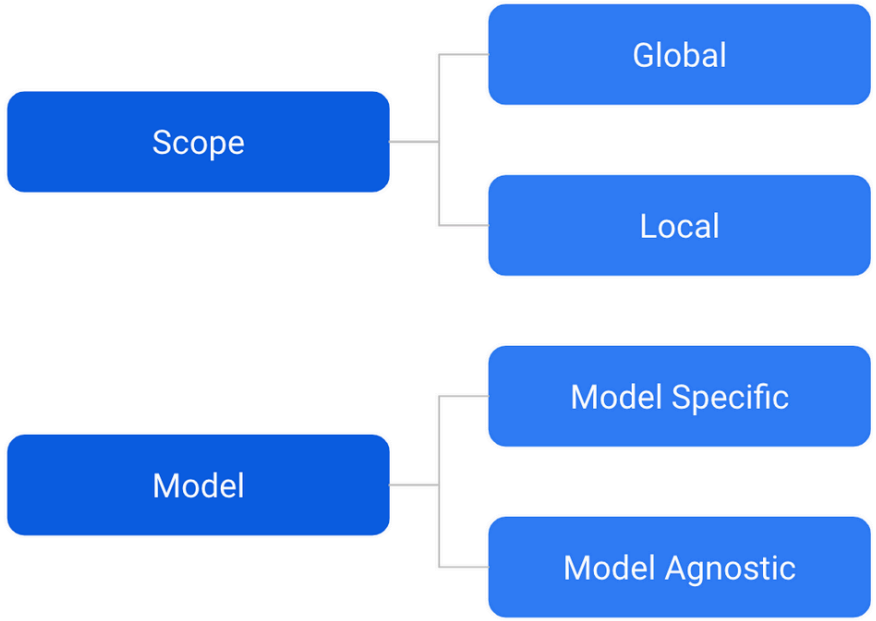
- May not be required every time and a call can be taken depending on context…

- When it does not impact end customer
  - For example, a situation where we are looking to refine some internal processes. Let's say we want to classify the call recordings since a call resulted in a dissatisfied customer. We are good to go as long as we are getting a good performance and it is solving our purpose.

- Another scenario where it does not impact end outcome as automation due to AI solves majority of problem
  - For example, if the problem is well studied, we are confident about the results. For optical character recognition cases where we can get a lot of training data and can rely on a good performance for the task at hand.



Framework On Interpretable ML

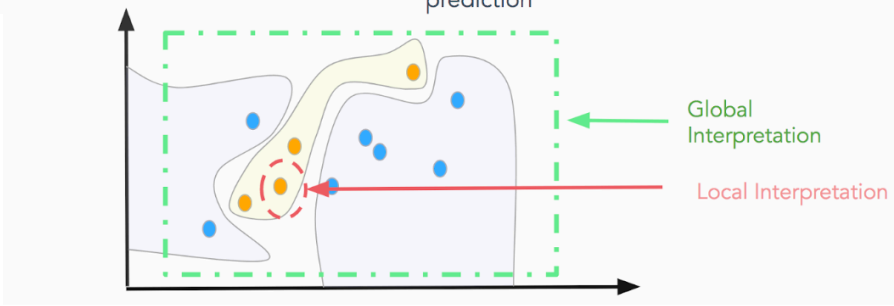


Global Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables based on the complete dataset

Local Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables wrt to a single prediction

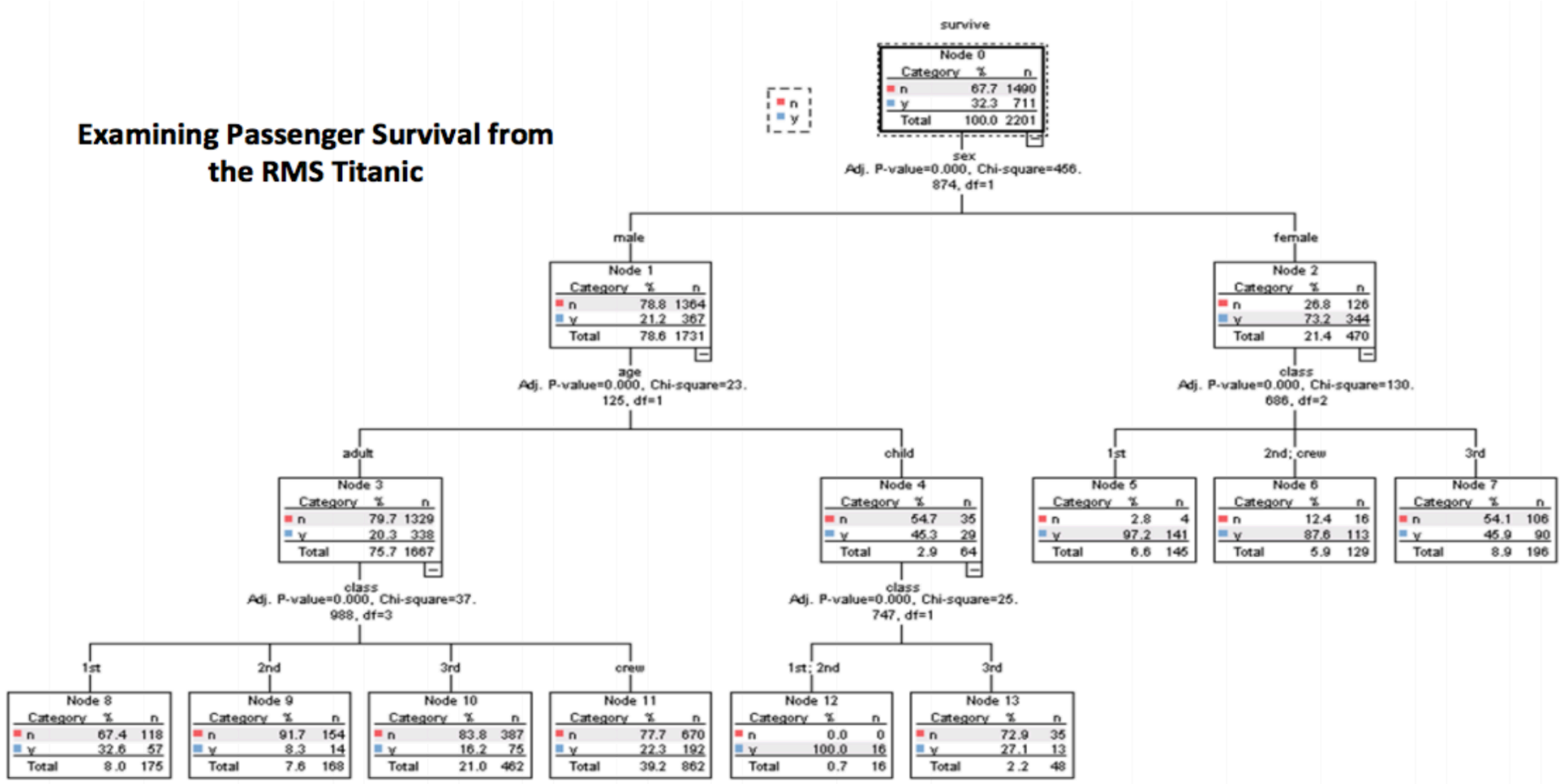


Summarizing Global and Local Interpretation (Source: DataScience.com)

Decision
Tree
Nodes

Local
And
Global

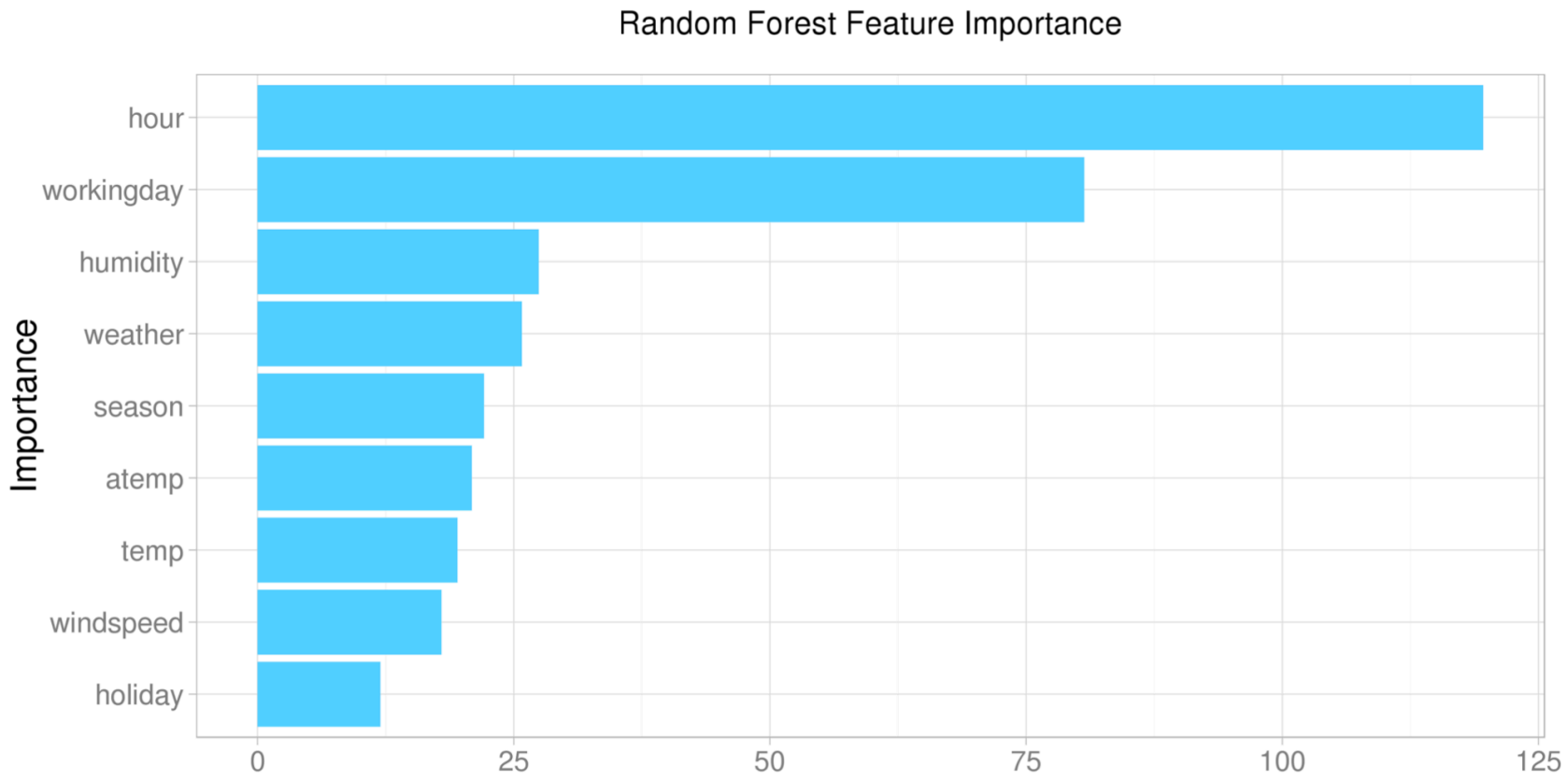
Titanic dataset
For illustration



RF

Variable Importance

Global Interpretability



LIME

Local Interpretable Model-agnostic Explanations

Local - explains why a single data point was classified as a specific class.
Model-agnostic - Treats the model as black-box. Does not need to know how it makes predictions.

It supports different type of data sources:
1. Tabular Explainer
2. Recurrent Tabular Explainer
3. Image Explainer
4. Text Explainer

How does it work:
1. Choose an observation to explain
2. Create new dataset around observation by sampling from distribution learnt on training data
3. Compute distances between new points and observation, that's the measure of similarity.
4. Use model to predict class of the new points
5. Find subset of m features that has strongest relationship with our target class.
6. Fit a linear model on fake data in m dimensions weighted by similarity.
7. Weights of linear model are used as explanation of decision.

Drawbacks:
1. Depends on random sampling of new points, so it can be unstable.
2. Fit of linear model can be inaccurate - R-squared score can be checked though
3. Relatively slow for a single observation, in particular with image datasets

Format:
1. Create a new explainer: *my_explainer = Explainer()*
2. Select an observation (obs) and create an explanation for it
obs = np.array([...])
my_explanation = my_explainer.explain_instance(obs, pred_f)
3. Use methods on explanation to visualize results
my_explanation.show_in_notebook()
my_explanation.get_image_and_mask()

LIME

Demo files

Pls refer to GitHub

Out[19]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
421	0	3	male	21.0	0	0	7.7333	Q

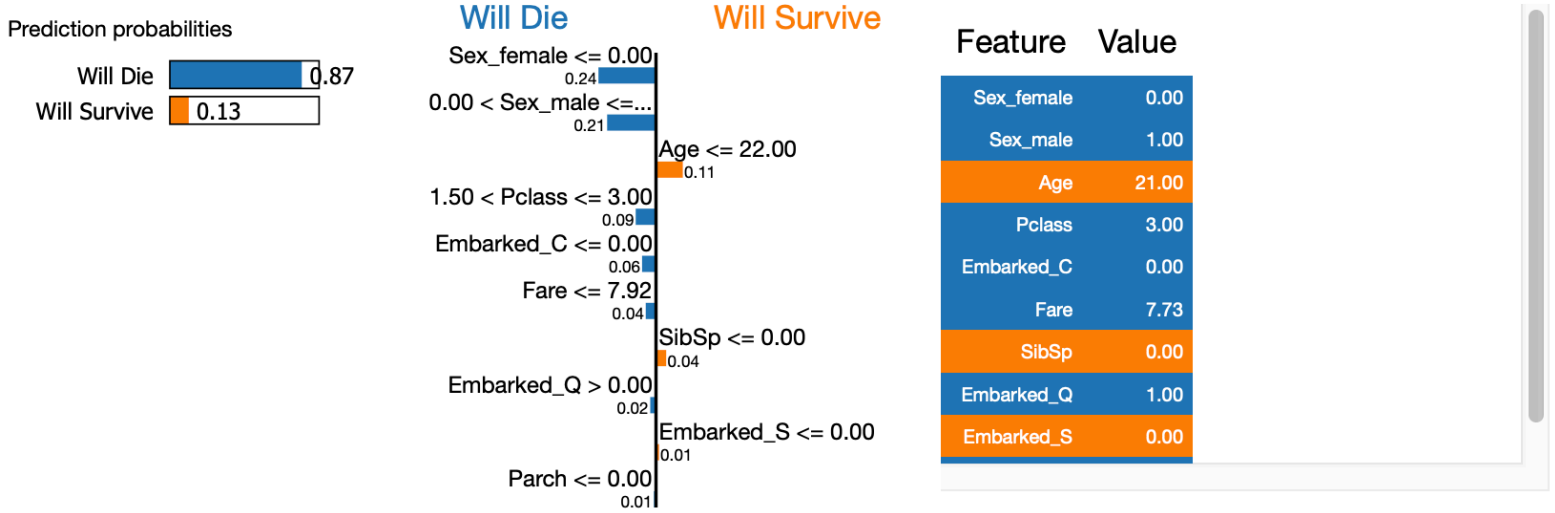
In [20]:
```
# Scenario 1:

# Above interpretation refers to the following:
# The Male passenger of age 21 travelling in passenger class 3, embarked from Q, who is not survived.
# Let's see what and how our model predicts his survival.
```

In [22]:
```
choosen_instance = X_test.loc[[421]].values[0]

my_explainer1 = explainer.explain_instance(choosen_instance, predict_fn_rf,num_features=10)

my_explainer1.show_in_notebook(show_all=False)
```



In [30]:
```
# Interpretation 1:

# Model predicted Will Die (Not survived). Biggest effect is person being a male;
# This has decreased his chances of survival significantly.
# Next, passenger class 3 also decreases his chances of survival while being 21 increases his chances of survival.
```

LIME

More
Demo files

Pls refer to
GitHub

Out[25]:

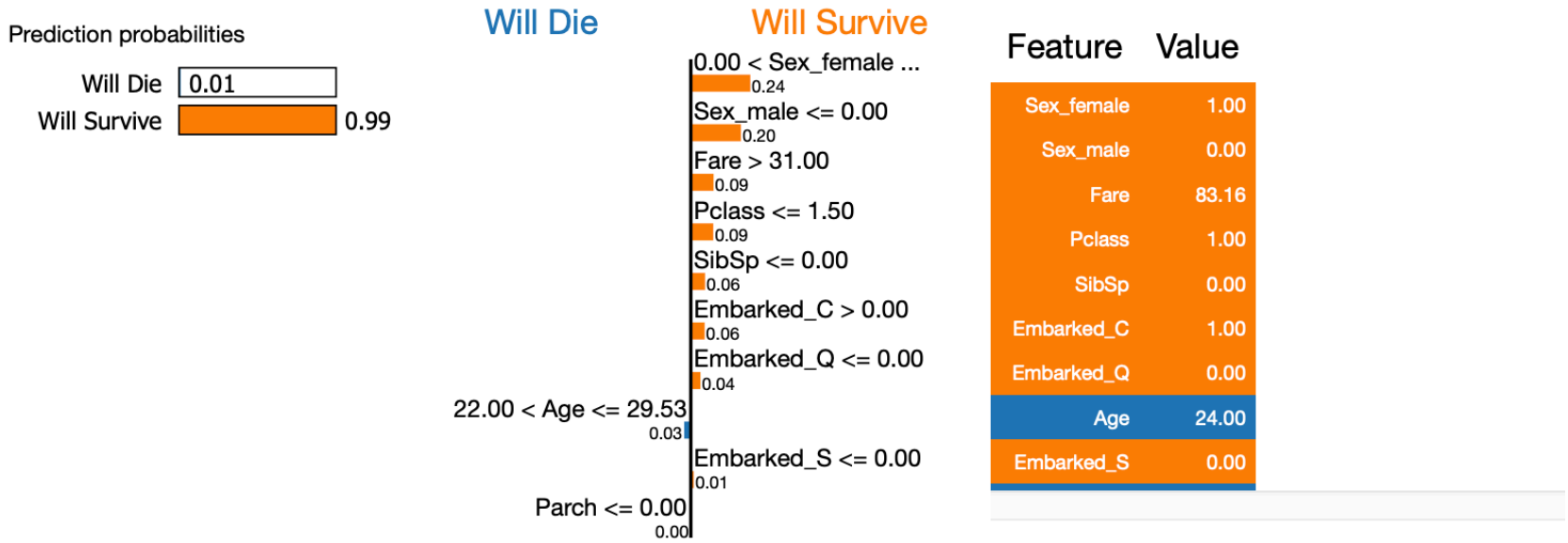
	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
310	1	1	female	24.0	0	0	83.1583	C

In [26]:
```
# Above interpretation refers to the following:
# The Female passenger of age 24 travelling in passenger class 1, embarked from C, who is survived.
# Let's see what and how our model predicts his survival.
```

In [28]:
```
choosen_instance = X_test.loc[[310]].values[0]

my_explainer2 = explainer.explain_instance(choosen_instance, predict_fn_rf,num_features=10)

my_explainer2.show_in_notebook(show_all=False)
```



In [29]:
```
# Interpretation 2:
```

In [30]:
```
# Model predicted 1 (Fully confident that passenger survives).
# Biggest effect is person being a female; This has increased her chances of survival significantly.
# Next, passenger class 1 and Fare>31 has also increases her chances of survival.
```

SHAP

SHapely Additive exPlanations

Replacement of a complex model, by substituting with a simpler and explainer model.

Uses "additive feature attribution methods" - the local explanation is a linear function of the features

Weight of each feature is computed using the shapely values method from game theory.

a) TreeExplainer
 Only for tree based models
 Works with scikit-learn, xgboost, lightgbm, catboost
b) KernelExplainer
 Model agnostic explainer (kNN and others)

Format:
1. Create a new explainer, with model as argument
explainer = TreeExplainer(my_tree_model)

2. Compute shap_values from model using some observations (obs)
shap_values = explainer.shap_values(obs)

3. Use SHAP visualization functions with shap_values
shap.force_plot(base_value, shap_value[0]) #local explain.
shap.summary_plot(shap_values) #global feature import.

Eli5

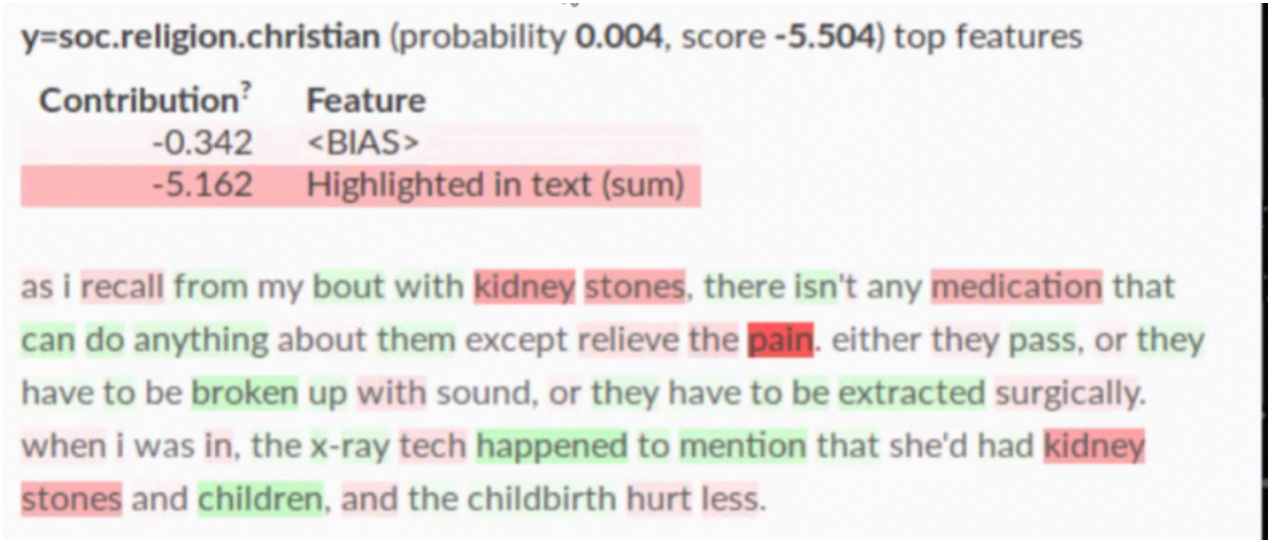
Explain me Like I'm 5

Eli5 - a Python library which allows to visualize and debug various Machine Learning models using unified API. It has built-in support for several ML frameworks and provides a way to explain black-box models.

1. Useful to debug scikit-learn models and communicate with domain experts
2. Provides global interpretation of "white-box" models with a consistent API
3. Provides local explanations of predictions

Format:
1. Explain model globally (features importance)
eli5.show_weights(model)
2. Explain a single prediction
eli5.show_prediction(model, obs)

Github Ref: https://github.com/teamhg-memex/eli5 (A library for debugging/inspecting machine learning classifiers and explaining their predictions)


y=soc.religion.christian (probability 0.004, score -5.504) top features

Contribution?	Feature
-0.342	<BIAS>
-5.162	Highlighted in text (sum)

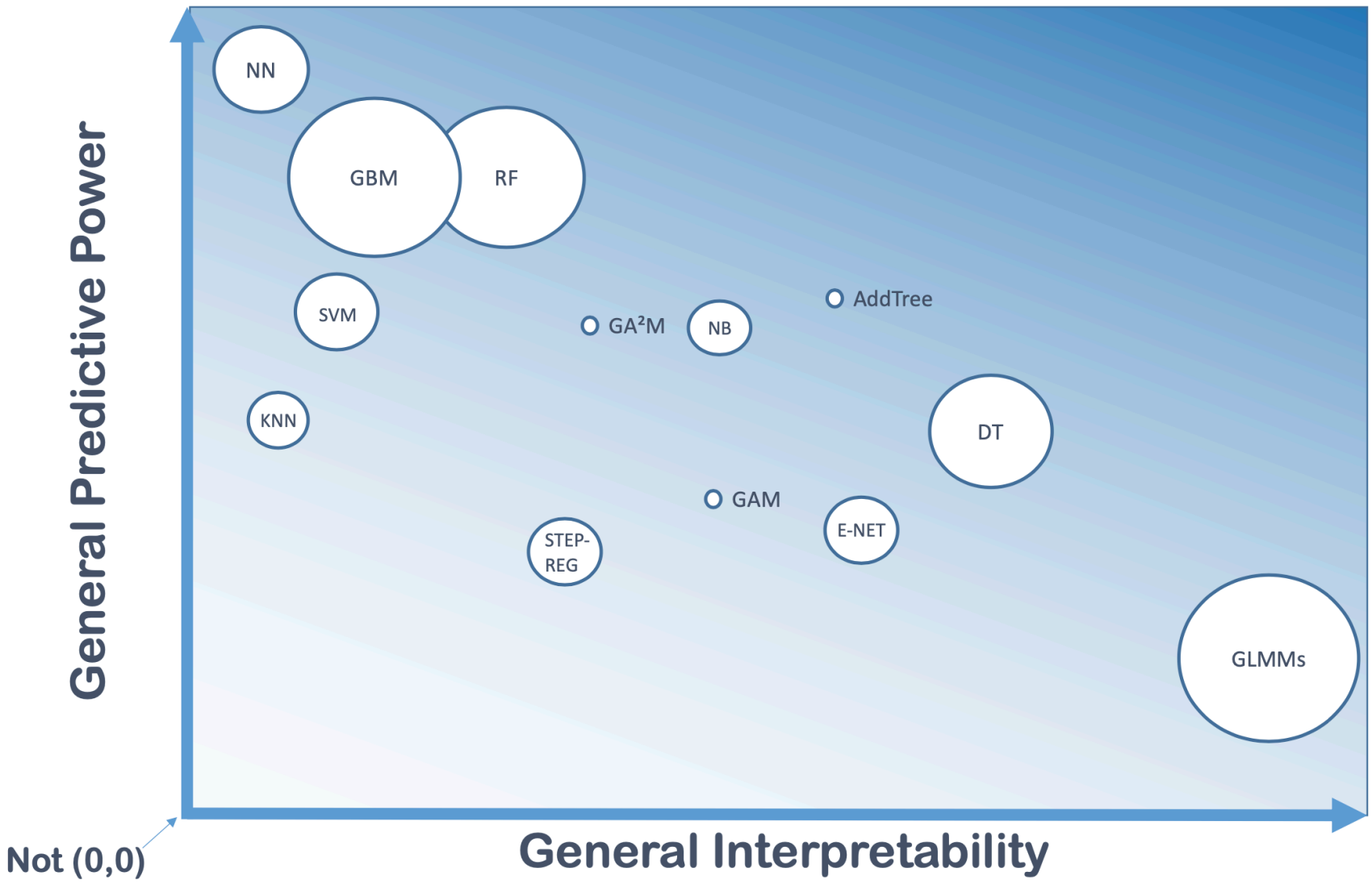
as i recall from my bout with kidney stones, there isn't any medication that can do anything about them except relieve the pain. either they pass, or they have to be broken up with sound, or they have to be extracted surgically. when i was in, the x-ray tech happened to mention that she'd had kidney stones and children, and the childbirth hurt less.

References

- LIME:
  - Paper by Marco et al (2016): "Why Should I Trust You?": Explaining the Predictions of Any Classifier - https://arxiv.org/abs/1602.04938
  - GitHub reference from Marco - https://github.com/marcotcr/lime
  - A reference to LIME example usage with iris dataset by Marco - https://marcotcr.github.io/lime/tutorials/Tutorial%20-%20continuous%20and%20categorical%20features.html
- SHAP –
  - Paper: A unified approach to interpreting model predictions by Lundberg and Lee - http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
  - Paper: Consistent individualized feature attribution for tree ensembles by Lundberg et al 2019 - https://arxiv.org/pdf/1802.03888.pdf
  - Paper: Explainable ML predictions for the prevention of hypoxaemia during surgery - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6467492/pdf/nihms-1505578.pdf
- Sensitivity Analysis:
  - SALib: An open-source Python library for Sensitivity Analysis - https://www.researchgate.net/publication/312204236_SALib_An_open-source_Python_library_for_Sensitivity_Analysis
  - Factorial sampling plans for Preliminary Computational Experiments - https://abe.ufl.edu/Faculty/jjones/ABE_5646/2010/Morris.1991%20SA%20paper.pdf
- InterpretML: A unified framework for ML interpretability by MSFT team - https://arxiv.org/pdf/1909.09223.pdf

General Overview Of models

General Predictive Power Vs General Interpretability



Size = frequency of use in the industry

General Overview Of models

What is ideal Advisory ?

- Interpretable and Predictive need
  - Observational
    - Local
      - Go mostly with LIME, SHAP or AILense for predictive
      - Tree based methods or AddTree as a compromise
      - Baseline with GLMM and co-efficient analysis
    - Global
      - Baseline with GLMM and co-efficient analysis
      - Tree based methods or AddTree as a compromise
      - Go with SHAP for structured content and AILense for deep nets
  - Can collect
    - GLMMs
    - Baseline predictive with modern algorithms

- Interpretable need only
  - Structured data sources
    - Observational
      - GLMMs
      - Tree based models
    - Can collect
      - GLMMs
  - Image content sources
    - Local
      - AILense, SHAP, LIME
    - Global
      - TCAV, AILense

- Predictive need only
  - You may need accuracy at all cost which is very important
  - Go for – Neural Net, GBMs, RF or RRF

Questions??

Thank you

kkm_007

https://www.linkedin.com/in/kamalmishra07/

https://github.com/kkm24132