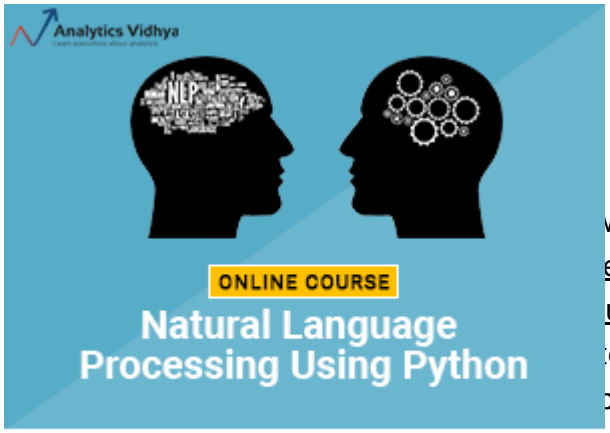
Decoding the Black Box: An Important Introduction to Interpretable Machine Learning Models in Python

ANKIT CHOUDHARY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/AUTHOR/ANKIT2106/), AUGUST 26, 2019 LOGIN TO BOOKMARK THIS ...

Overview

- Interpretable machine learning is a critical concept every data scientist should be aware of

- How can you build interpretable machine learning models? This article will provide a framework
- We will also code these interpretable machine learning models in Python



v about a random forest with 500 trees? Building a complex and
es.analyticsvidhya.com/courses/applied-machine-learning-
utm_medium=decoding-black-box-step-by-step-guide-interpretable-
ential of reaching our desired accuracy, but does it make sense?
plain how it arrived at the final result?

Learn to Solve
Text Classification Problems Using **NLP**

(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp?
utm_source=Sticky_banner1&utm_medium=display&utm_campaign=NLPcourse)

er as data scientists. A wide variety of businesses are relying on
(nlp) to drive their strategy and increase their bottomline. Building a model that we can explain to
and other stakeholders is key.

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

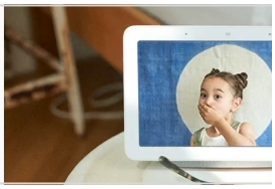
(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/interpretable_machine_learning.jpg)

Can you imagine building a facial recognition software that misclassifies a person? Or a credit card fraud detection model that raises an alarm for a perfectly legal transaction? And then not being able to explain why that's happening – not ideal.

So the question is – how do we build interpretable machine learning models? That's what we will talk about in this article. We'll first understand what interpretable machine learning is and why it's important. Then we will understand a simple framework for interpretable ML and use that to build machine learning models.

(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=blog&utm_medium=decoding-black-box-step-by-step-guide-interpretable-machine-learning-models-python)

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)



ning so strap in for a thrilling learning journey!

*comprehensive **'Applied Machine Learning' course** **s/applied-machine-learning-beginner-to-professional/?utm_source=blog&utm_medium=decoding-black-box-step-by-step-guide-interpretable-machine-learning-models-python)**. The course provides you all the tools and techniques you need to solve business problems using **or beginners as well as intermediate-level professionals!***

isplay&utm_campaign=NLPcourse)

?

Learning?

lity?

earning

Models

- Model Agnostic Techniques for Interpretable Machine Learning
- LIME (Local Interpretable Model Agnostic Explanations)
- Python Implementation of Interpretable Machine Learning Techniques

What is Interpretable Machine Learning?

How do we build trust in machine learning models? That's essentially what this boils down to.

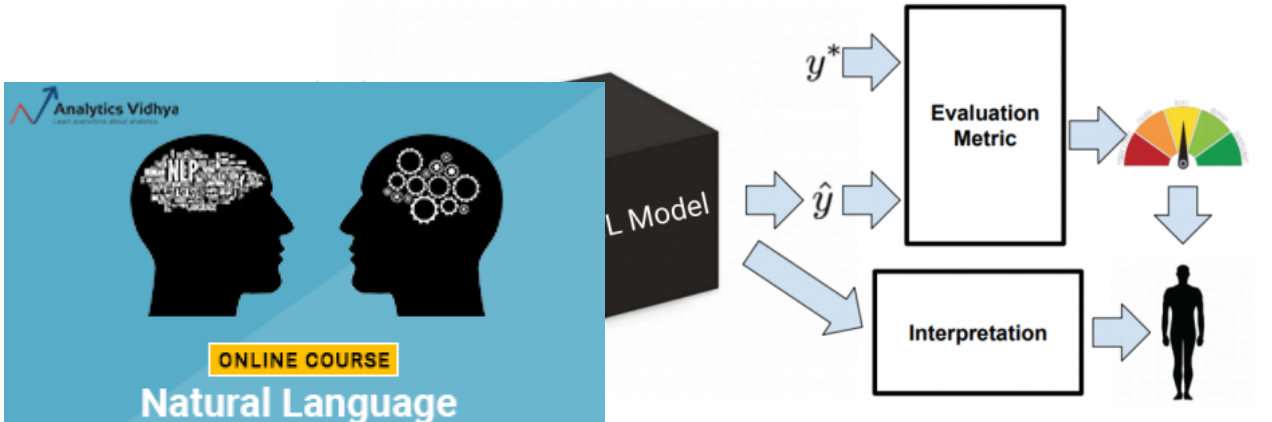
Machine learning-powered applications have become an ever-increasing part of our lives, from image and facial recognition systems to conversational applications, autonomous machines, and personalized systems.

nade by these machine learning-enabled systems are becoming much more profound, and in many cases, critical to life, death, and personal wellness. **The need to trust these AI-based systems is paramount**

estion: What involves a predictive modeling lifecycle?

lems based on structured datasets. However, a key issue that arises especially as model complexity increases is interpretability. Let's start with the formal definition:

> *Interpretation of a machine learning model is the process wherein we try to understand the predictions of a machine learning model.*



/uploads/2019/08/Screenshot-2019-08-16-at-3.07.26-PM.png)

odeling lifecycle is at two important stages:

(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?

- One – where we monitor the evaluation metric and try different ideas of feature engineering, feature
- rove and build more robust models

isplay&utm_campaign=NLP course)

f this article, is to interpret the models using the predictions and

sifier chose a particular class for example

hine Learning?

ues for interpreting machine learning models

s/applied-machine-learning-beginner-to-professional/?

ack-box-step-by-step-guide-interpretable-machine-learning-models-

python), let's look at why this is important.

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

Fairness

Let's take a simple example to understand this. Suppose we are trying to predict employees' performance in a big company to expedite the appraisal process and identify the best employees.

We have data from the last 10 years about the performance reviews of employees. But what if that company tends to promote more men than women?

(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-3.11.25-PM.png)
(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?...isplay&utm_campaign=NLPcourse)

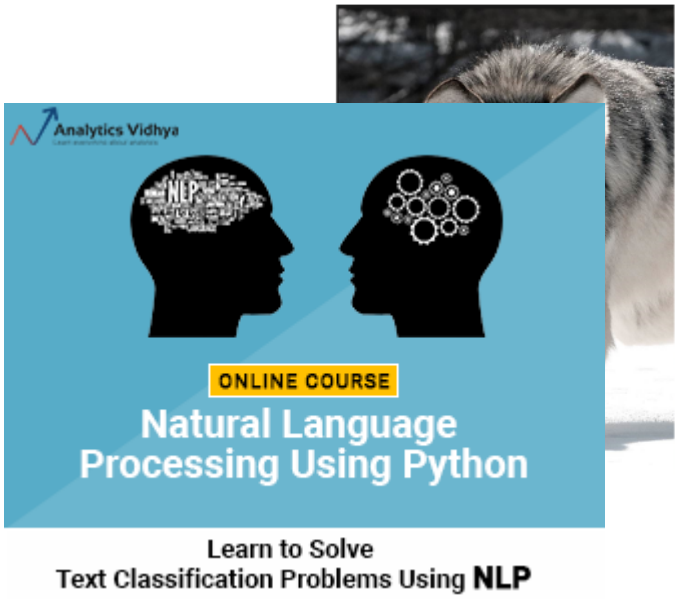
...nd predict that men tend to perform at a higher scale (and this bias ...orld scenarios). Now, if there is no way to interpret our model at this ... insights at the cost of compromising on fairness.

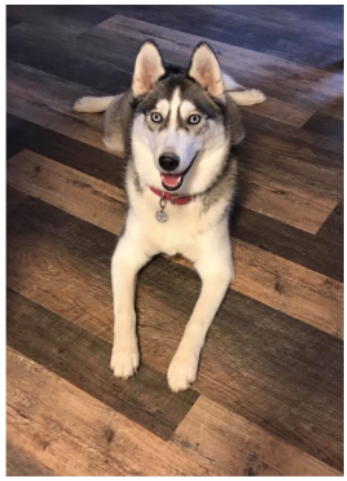
...ctices WATCH NOW

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

Checking causality of features & Debugging models

Let's consider another example. Consider that we are building a model for classifying wolves vs dogs. The data that is available is simply labeled images of dogs and wolves.



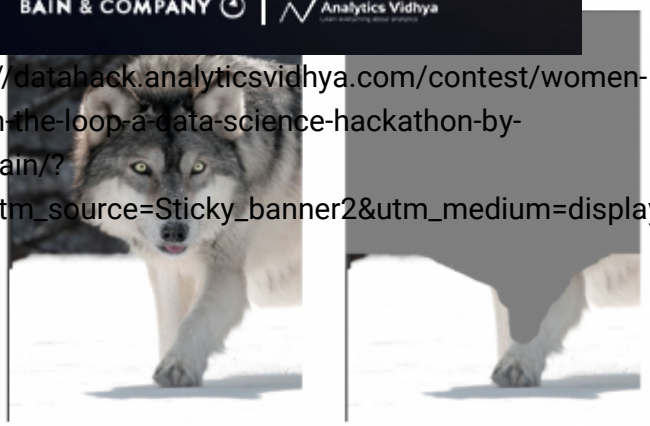


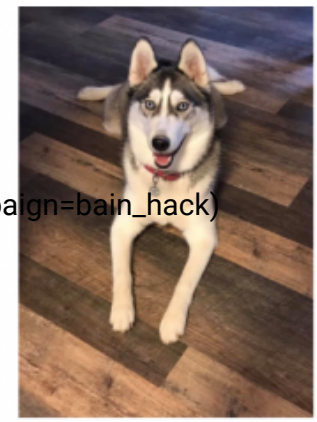
Dog

(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-3.20.23-PM.png)

(https://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)





(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-3.36.47-PM.png)

Having an interpretable model, in this case, enables us to test the causality of
ty and ultimately can help us to debug the model

(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?

ome with
nt Starter Kit.

dding images of the animals in different backgrounds or simply crop the background out of the images to ensure that the right signal is picked up by our machine learning or

isplay&utm_campaign=NLPcourse)

PR's article 12 allows individuals to enquire about algorithmic
ance industry, questions such as the following can come up and

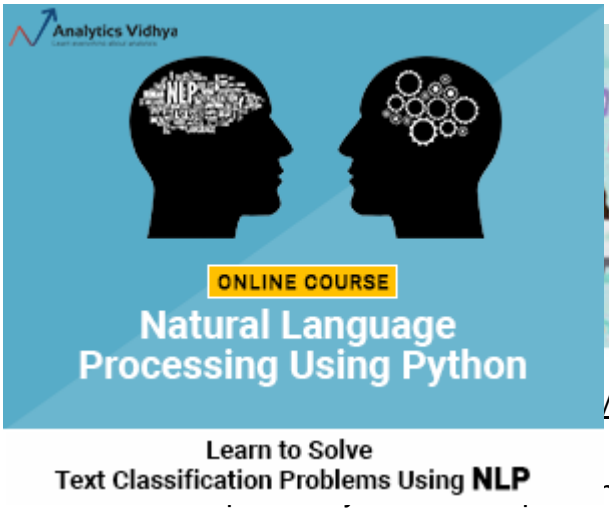
redit card? etc.

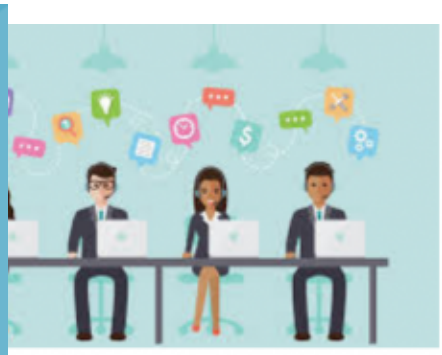
(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-3.40.35-PM.png)

When can we do Away with Interpretability?

Not everything requires interpretability. It is also important to understand when we do *not* need to invest in building interpretable machine learning models.



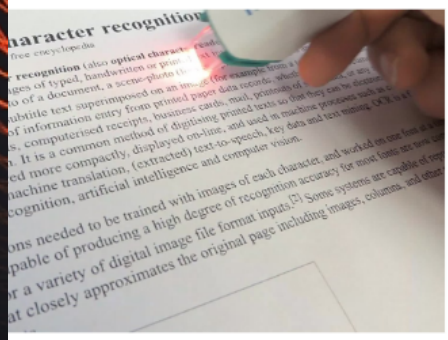


/uploads/2019/08/Screenshot-2019-08-16-at-3.48.28-PM.png)

(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?

ne end customer. For example, a situation where we are looking to refine some internal processes. 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

isplay&utm_campaign=NLPcourse)



(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-3.49.32-PM.png)

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

- If the problem is well studied, we are confident about the results. For example, optical character recognition for which we can get a lot of training data and can rely on a good performance for the task at hand

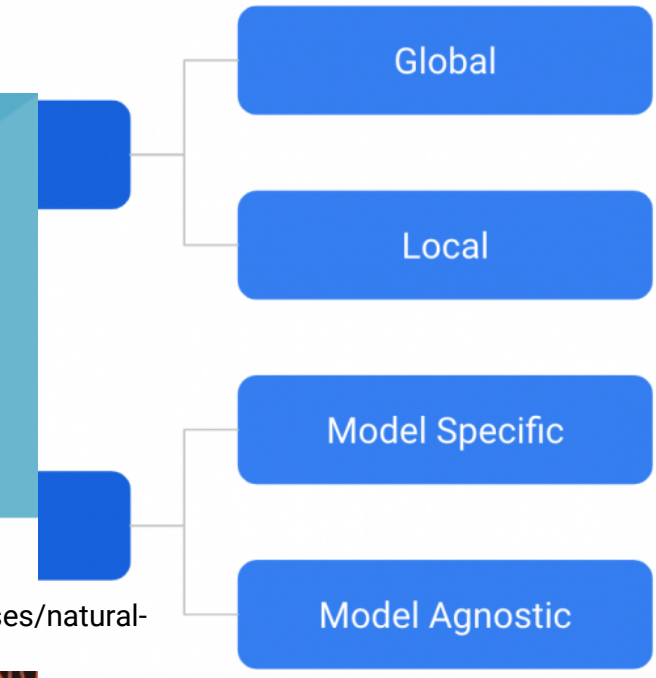
Framework for Interpretable Machine Learning

Now that we have an intuition of what ML interpretability is and why it's important, let's look at the different ways to classify interpretability techniques:

(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=NLPcourse)

(/uploads/2019/08/Screenshot-2019-08-16-at-3.51.47-PM.png)

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)



wo structures:

or all data points, the importance of each variable, or are we looking
al?

Model

The second way of looking at this is whether we are talking about a technique that works across all types of models (model agnostic) or is tailor made for a particular class of algorithms (model specific).

Let's Talk About Inherently Interpretable Models

Linear/Logistic

For linear models such as a linear and logistic regression, we can get the importance from the weights/coefficients of each feature.

Let's revisit that quickly. Suppose we are trying to predict an employee's salary using linear regression. The independent variables are experience in years and a previous rating out of 5.

(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?

1*experience + W2*rating

/uploads/2019/08/Screenshot-2019-08-16-at-4.01.23-PM.png)

ally tell us whether the experience is more important towards salary

ific technique that can be used for both global and local

on in the below articles:

- 7 Regression Techniques you should know!

/2015/08/comprehensive-guide-regression/?

isplay&utm_campaign=NLPcourse)

ling-black-box-step-by-step-guide-interpretable-machine-learning-

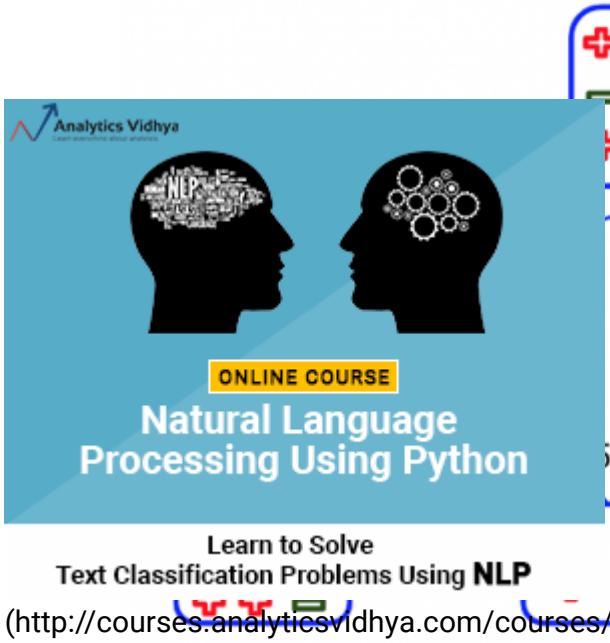
(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

R and Python

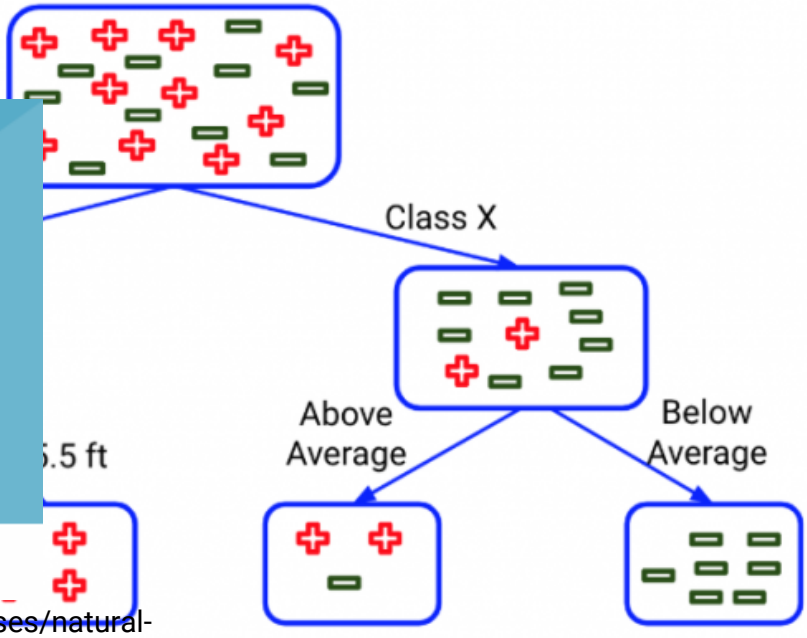
/2015/11/beginners-guide-on-logistic-regression-in-r/?

ling-black-box-step-by-step-guide-interpretable-machine-learning-

A decision tree is another algorithm that is very interpretable as we have access to all the splits for each feature:



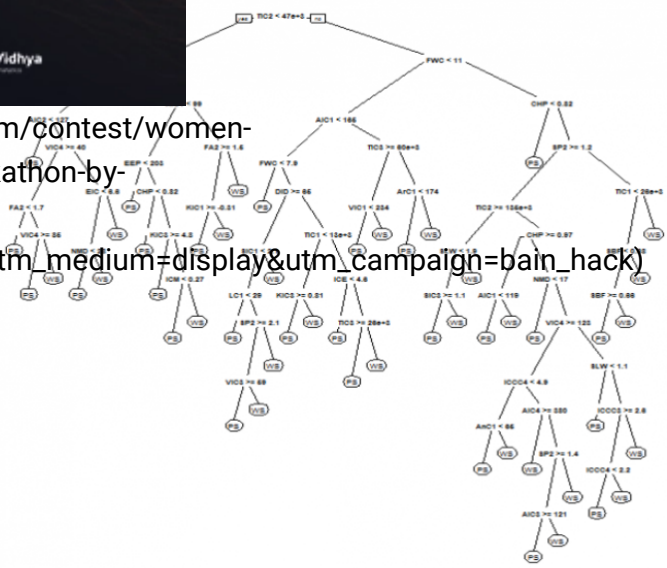
(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?



uploads/2019/08/Screenshot-2019-08-16-at-4.03.14-PM.png)

ng taken starting from the root node to the leaf node. We just have t variables and list them down to explain each prediction. Again, **used for local explanations.**

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)



(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-4.06.13-PM.png)

For a small decision tree, we can use the above diagram. However, if we have a lot of features and we are training a deep decision tree for, let's say, a depth of 8 or 9, there will be too many decision rules to present effectively. In this case, **we can use feature importance to interpret the importance of each feature at a global level.**

importance for a decision tree.

> *maximize the decrease in impurity. We can use this*
> *ntribution of each feature.*

- **Step 1:** Go through all the splits in which the feature was used
- **Step 2:** Calculate the reduction in criterion (Gini/information gain) compared to the parent node weighted by
  - N is the total number of observations
  - N & Gini represents the number of samples & Gini impurity in the parent, left and right node
- **Step 3:** Take the sum for all splits for each feature and compare

are important as they help us incorporate the number of samples
s well into the equation. The formula is:

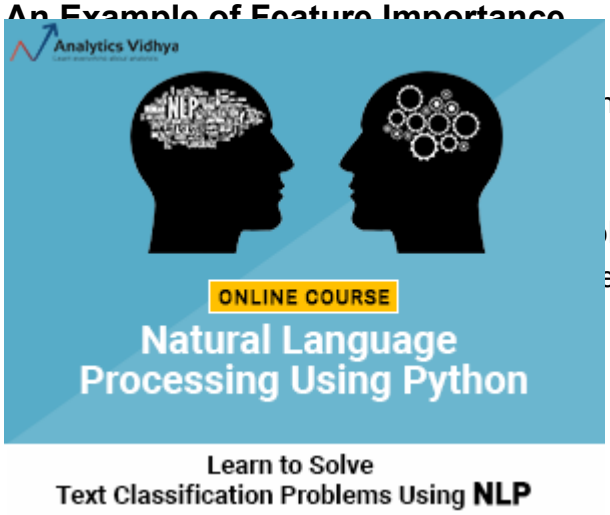
$$\left(\frac{N_{Right}}{N_{parent}} \cdot Gini_{Right} - \frac{N_{Left}}{N_{parent}} \cdot Gini_{Left}\right)$$

/uploads/2019/08/Screenshot-2019-08-16-at-4.12.05-PM.png)

Here, again, **this is a model-specific technique that can be used for only global explanations.** This is because we are looking at the overall importance and not at each prediction.

Learn more about decision trees in this superb tutorial (https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/?utm_source=blog&utm_medium=decoding-black-box-step-by-step-guide-interpretable-machine-learning-models-python).

An Example of Feature Importance

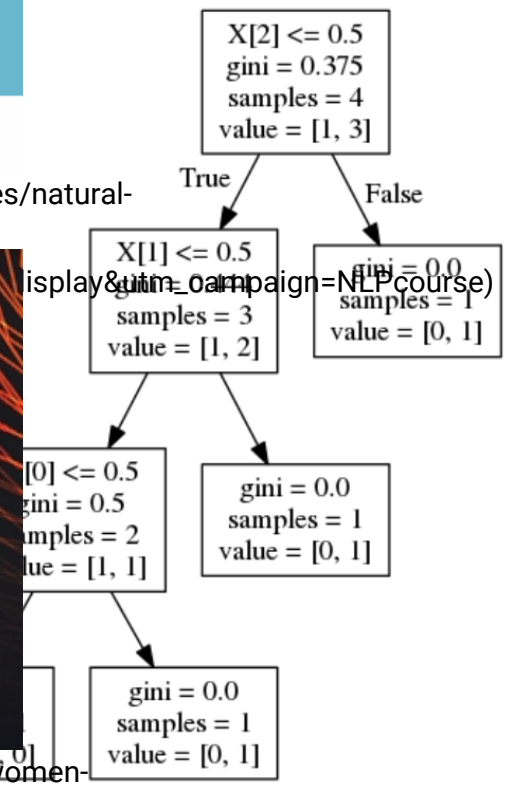


(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=NLPcourse)

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

mple. This will help you visualize what we've covered so far (and

les and Gini impurity values as shown in the below figure. Each ectly use the formula to calculate feature importance:



(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-4.16.06-PM.png)

Since each feature is used once in our case, there is no need to calculate the sum.

- For X[2] :
  - feature_importance = (4 / 4) * (0.375 – (0.75 * 0.444)) = 0.042
- For X[1] :
  - feature_importance = (3 / 4) * (0.444 – (2/3 * 0.5)) = 0.083
- For X[0] :
  - feature_importance = (2 / 4) * (0.5) = 0.25

Take a moment to pause here and calculate this on your own. You will have a much better grasp on the concept once you do it by yourself.

(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=blog&utm_medium=display&utm_campaign=NLP course)

(///datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

st and Gradient Boosting Machines, we can use the same feature
rage across all trees. Let us look at the steps involved:

e technique explained in the above section
nce across all trees using this formula:

$$\text{Feature Imp} = \frac{\text{Sum of Feature Imp of all estimators}}{\text{Number of Trees}}$$

/uploads/2019/08/Screenshot-2019-08-16-at-4.19.29-PM.png)

e to find feature importance for each feature.

andom Forest and Ensembling:

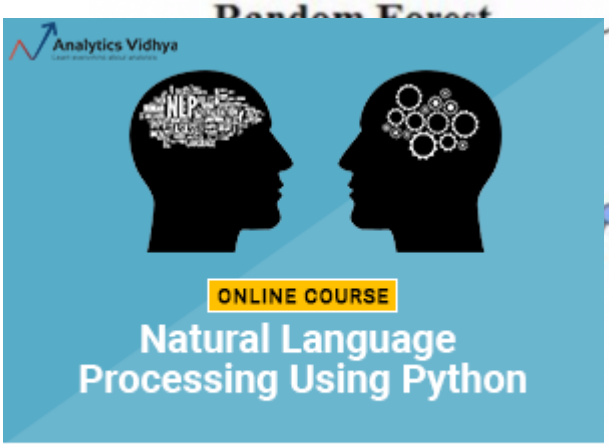
ified (https://www.analyticsvidhya.com/blog/2014/06/introduction-
=blog&utm_medium=decoding-black-box-step-by-step-guide-
python)

- A Comprehensive Guide to Ensemble Learning (with Python codes) (https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/?utm_source=blog&utm_medium=decoding-black-box-step-by-step-guide-interpretable-machine-learning-models-python)

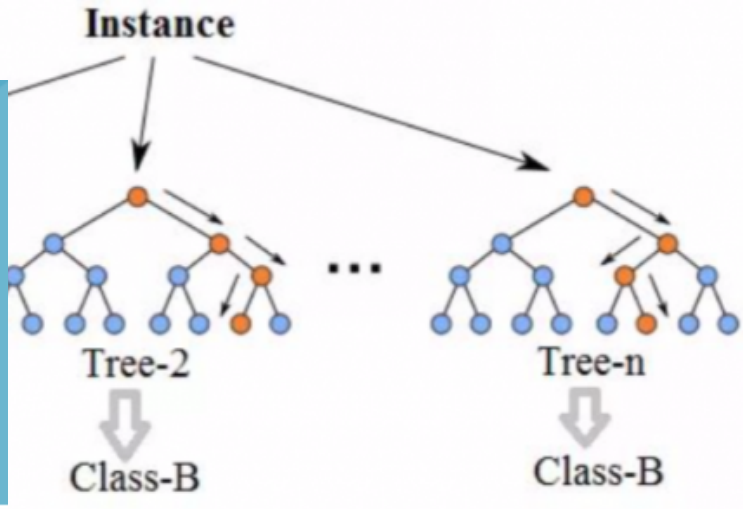
Model Agnostic Techniques

So far, we have discussed model-specific techniques for linear and logistic regression, as well as decision trees. We also spoke about the feature importance methods that are used for ensemble methods. I'm sure you're wondering – what about other models?

Now, we know that some models are hard to interpret, such as random forest and gradient boosting.



(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=NLPcourse)

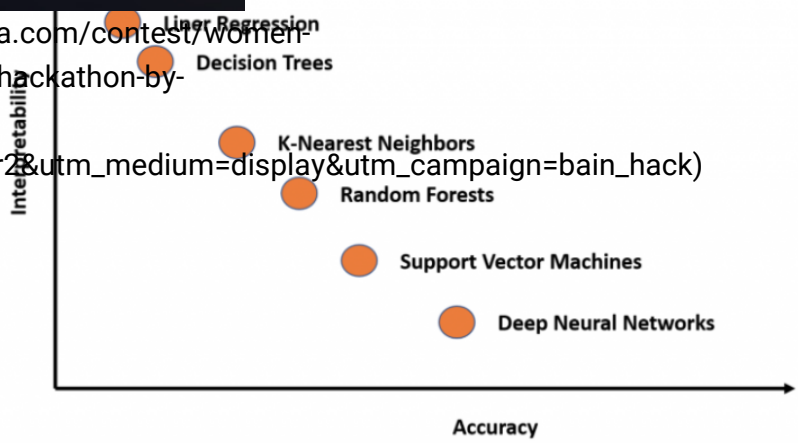


(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-6.14.59-PM.png)We

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

es. But it does not tell us whether a particular feature affects the
Y important in certain cases.

er to interpret. For example, a deep neural network model
/10/introduction-neural-networks-deep-learning/?
ack-box-step-by-step-guide-interpretable-machine-learning-models-
ers and it essentially ends up being an extreme version of a black-



(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-6.20.59-PM.png)

I really like this plot. As the complexity of the machine learning model increases, we get better performance but lose out on interpretability. You should keep this figure handy the next time you're building your own model.

arning models that don't compromise on accuracy?

ou can ensure you have full confidence in the interpretability.
better performance. So is there a way we can have some level of

> *ues allow us to build and use more complex models*
> *bility power.*

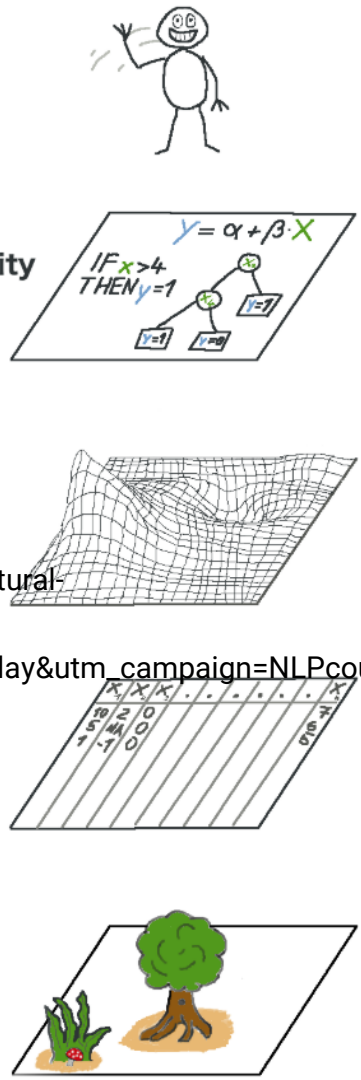
(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=NLPcourse)

interpretability. We capture the world by collecting data, and
ata with a machine learning model. Interpretability is just another
e black box using a simpler, more interpretable model.

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=NLPcourse)

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)



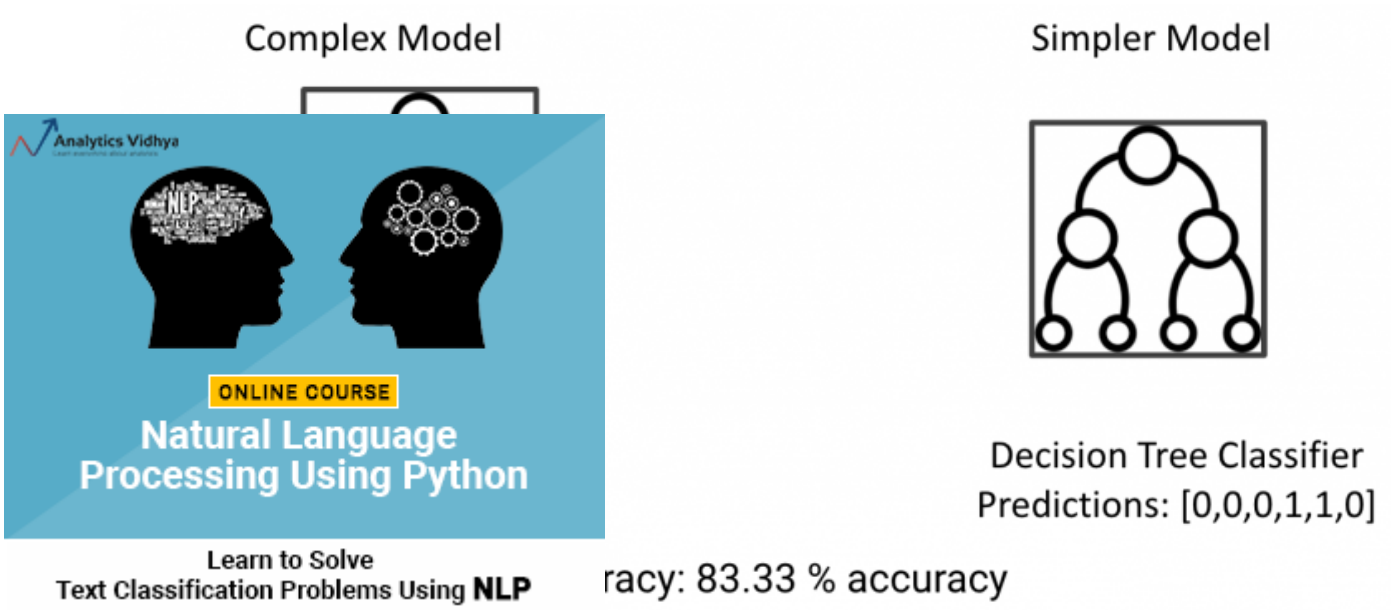
(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-6.24.30-PM.png)

Global Surrogate Method

The first model agnostic method we will discuss here is the global surrogate method. **A global surrogate model is an interpretable model that is trained to approximate the predictions of a black-box model.**

We can draw conclusions about the black-box model by interpreting the surrogate model. So, we are basically solving machine learning interpretability by using more machine learning!

For example, we could interpret a random forest classifier using a simple decision tree to explain its predictions:



(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=NLPcourse)
(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-6.28.49-PM.png)

predictions of the black-box model (which is a random forest in our
uracy, we can use it to explain the random forest classifier.

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

> *etable surrogate model that is trained to*
> *s of a black-box model and draw conclusions.*

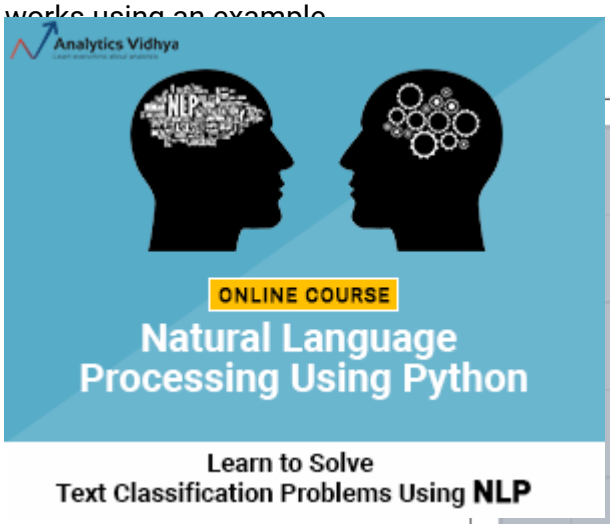
nd how a global surrogate model works:

1. We get predictions from the black-box model
2. Next, we select an interpretable model (Linear, decision tree, etc.)
3. We train an interpretable model on the original dataset and use black box predictions as the target
4. Measure the performance of the surrogate model
5. Finally, we interpret the surrogate model to understand how the black-box model is making its decisions

LIME (Local Interpretable Model agnostic Explanations)

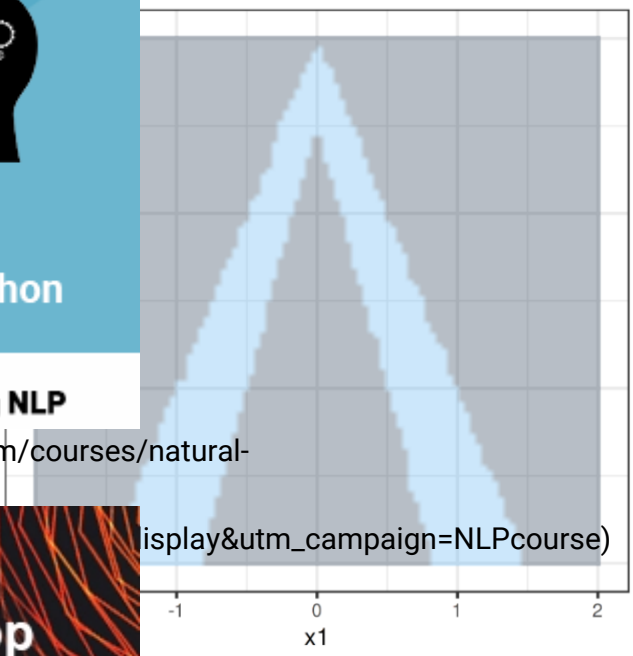
The global surrogate method is good for looking at an interpretable model that can explain predictions for a black-box approach. However, this will not work well if we want to understand how a single prediction was made for a given observation.

This is where we use the LIME technique which stands for local interpretable model agnostic explanations. LIME is based on the work presented in this paper (https://arxiv.org/abs/1602.04938). Let us understand how LIME works using an example



(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?

isplay&utm_campaign=NLPcourse)

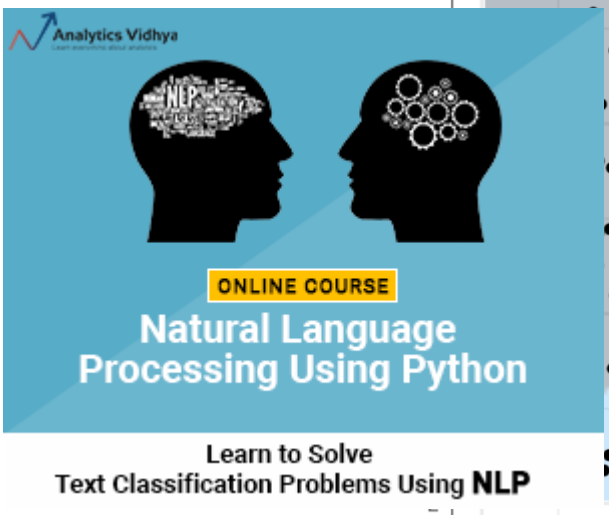


/uploads/2019/08/Screenshot-2019-08-16-at-6.45.28-PM.png)

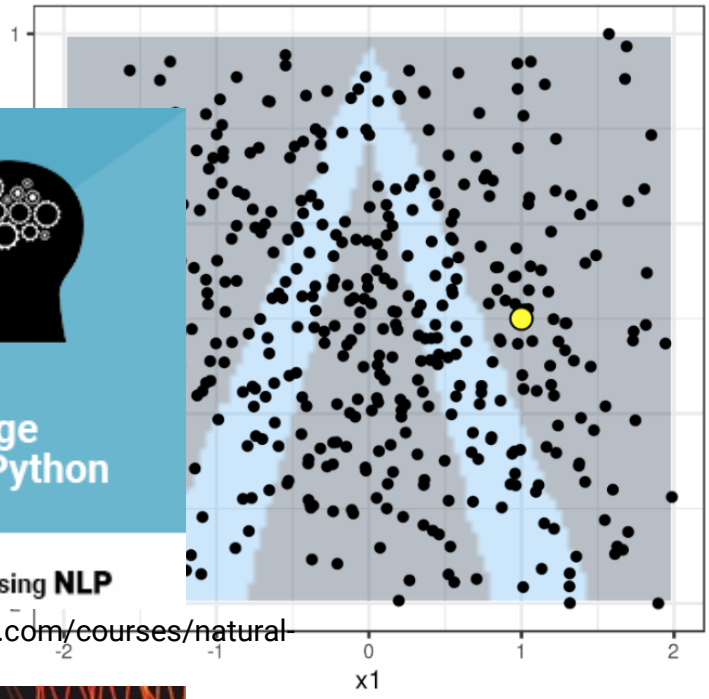
ation problem. As we can see in the above image, we have a two features.

s of x1 and x2 for the observation in yellow (in the below image). ribution to generate fake data around the observation:

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)



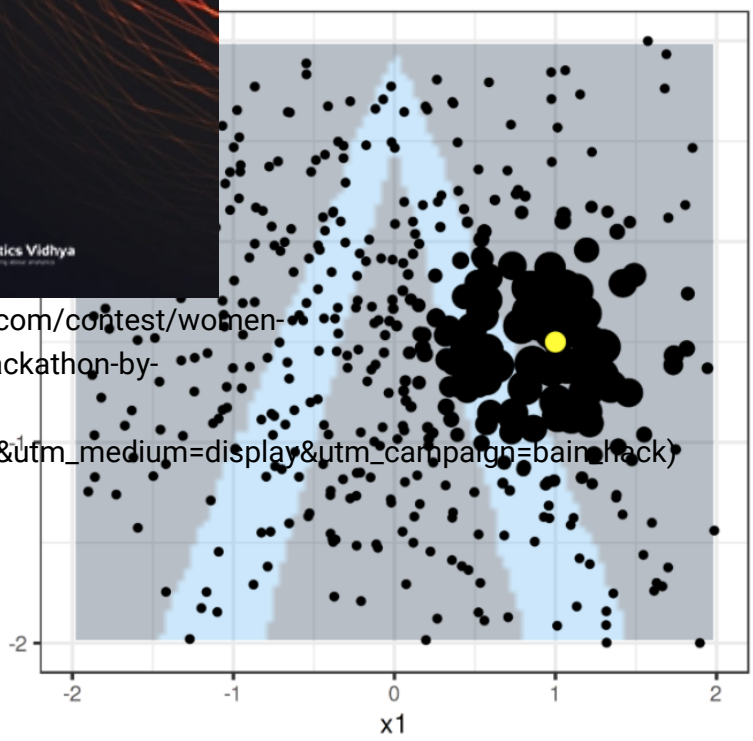
(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=NLPcourse)



(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-6.48.17-PM.png)

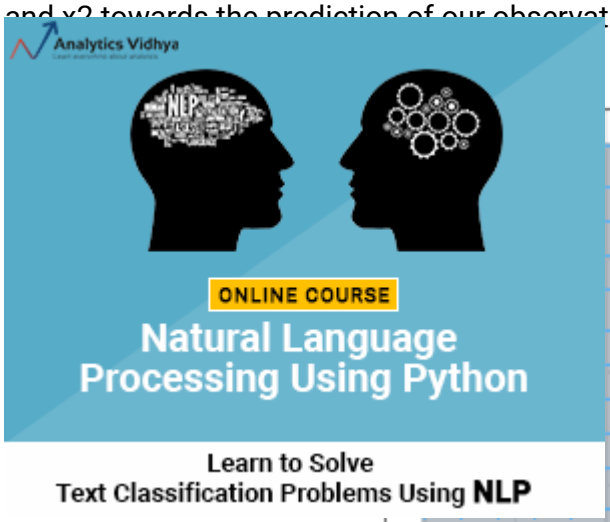
(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

hat are closer to our observation:

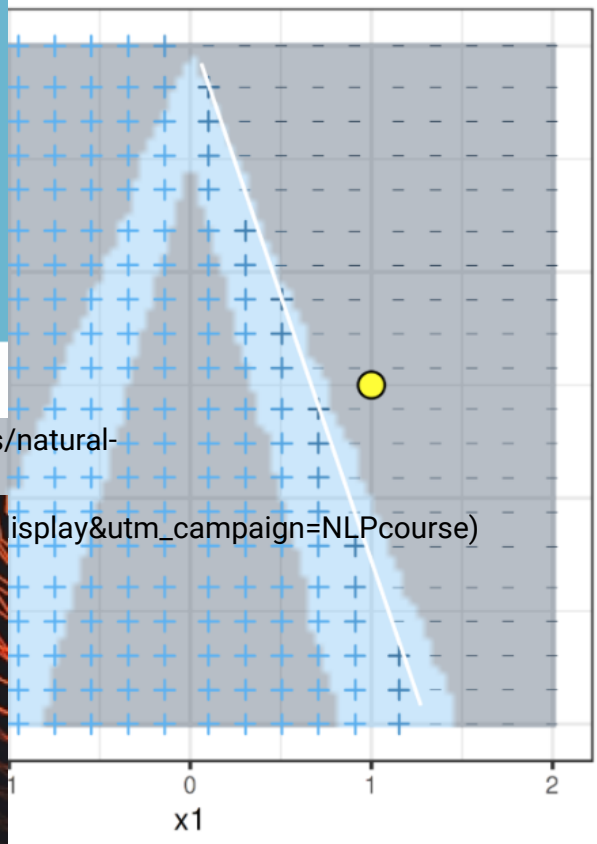


(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-6.50.42-PM.png)

We train an interpretable model over the fake data generated from the distribution. Now, we have a new local decision boundary for the locally learned model (in white) that can be used to understand the contributions of x1 and x2 towards the prediction of our observation:



(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=NLPcourse)



/uploads/2019/08/Screenshot-2019-08-16-at-6.57.00-PM.png)

Summarizing all the steps:

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

- 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 fake data 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

Python Implementation of Interpretable Machine Learning Techniques

My favorite part of the article – building interpretable machine learning models in Python!

Here, we will work on the implementation of both the methods we covered above. We will use the big mart sales problem (https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/?utm_source=blog&utm_medium=decoding-black-box-step-by-step-guide-interpretable-machine-learning-models-
...tps://datahack.analyticsvidhya.com/?
...ack-box-step-by-step-guide-interpretable-machine-learning-models-
...dicting sales for different items being sold at different outlets. You
...k.

*...//courses.analyticsvidhya.com/courses/big-mart-sales-prediction-
...n/?utm_source=blog&utm_medium=decoding-black-box-step-by-
...odels-python) to fully understand how to build models using this
...retability part.*

Building and Understanding Interpretable Machine Learning Models

...or inherently interpretable machine learning models.


```
..._test_split
..._error
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from xgboost.sklearn import XGBRegressor
from sklearn.preprocessing import OneHotEncoder, LabelEncoder
from sklearn import tree

import matplotlib.pyplot as plt
%matplotlib inline
```

Reading Data

Feature Engineering

Data Preprocessing

```
# label encoding the ordinal variables

le = LabelEncoder()
                                                    et_Identifier'])
                                                    cation_Type','Outlet_Size','Item_Type_Combined','Outlet_Typ
```



```
                                                    ical variables
                                                    at_Content','Outlet_Location_Type','Outlet_Size','Outlet_Typ
e',
                                                    'Item_Type_Combined','Outlet'])
```



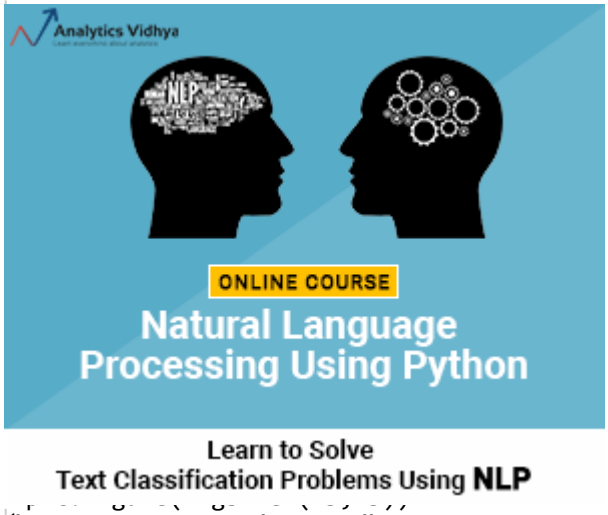
```
                                                    s that have been used to extract new variables
                                                    nt_Year', 'Item_Identifier', 'Outlet_Identifier'],axis=1,inpl
                                                    nt variables
y = df['Item_Outlet_Sales']

# creating the training and validation set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=.25, random_state=42)
```

Training a Decision Tree Model

```
dt = DecisionTreeRegressor(max_depth = 5, random_state=10)


# fitting the decision tree model on the training set
dt.fit(X_train, y_train)
```

Use the Graphviz library to visualize the decision tree

```
out_file='tree.dot', feature_names=X_train.columns, filled=T
```



(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=NLPcourse)

```
plt.imshow(image)
```

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/download.png)

This visualization of our decision tree clearly displays the rules it is using to make a prediction. **Here, Item_MRP & Outlet_Type are the first features that are affecting the sales of various items at each outlet.** If you want to look at the complete decision tree, you can easily do that by changing the *max_depth* parameter using the *export_graphviz* function.

Feature Importance

Now, we will have a look at the feature importance for each feature in case of a random forest.


```
odel
200, max_depth=5, min_samples_leaf=100,n_jobs=-1)

: model

columns
ature_importances_

# Feature_importance values in descending order
tance', ascending=False).head(10)
```

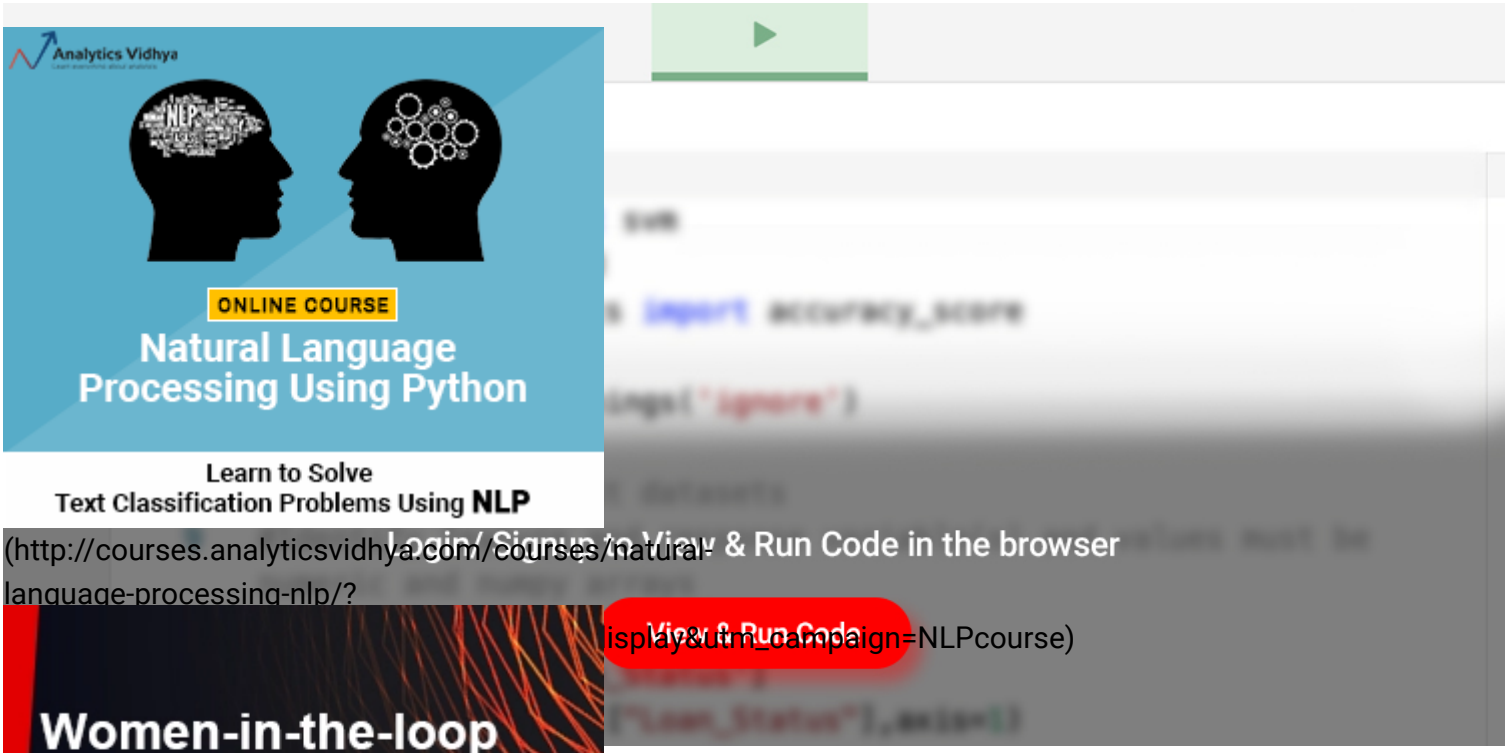
	variable	importance
	Item_MRP	0.552067
	Outlet_Type_0	0.322084
	Outlet_Type_3	0.046778
	Outlet_Years	0.043134
	Outlet_5	0.034728
	Item_Visibility	0.000638
0	Item_Weight	0.000279
13	Outlet_Type_1	0.000113
8	Outlet_Location_Type_2	0.000082
10	Outlet_Size_1	0.000060

(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-2019-08-16-at-7.24.15-PM.png)

The random forest model gives a similar interpretation. Item_MRP still remains the most important feature (exactly as the decision tree model above). **Relative importance also helps us compare each feature.** For example, Outlet_Type_0 is a much more important feature than other outlet types.

Exercise

As an exercise try calculating the feature importance for the decision tree we fit earlier and compare:

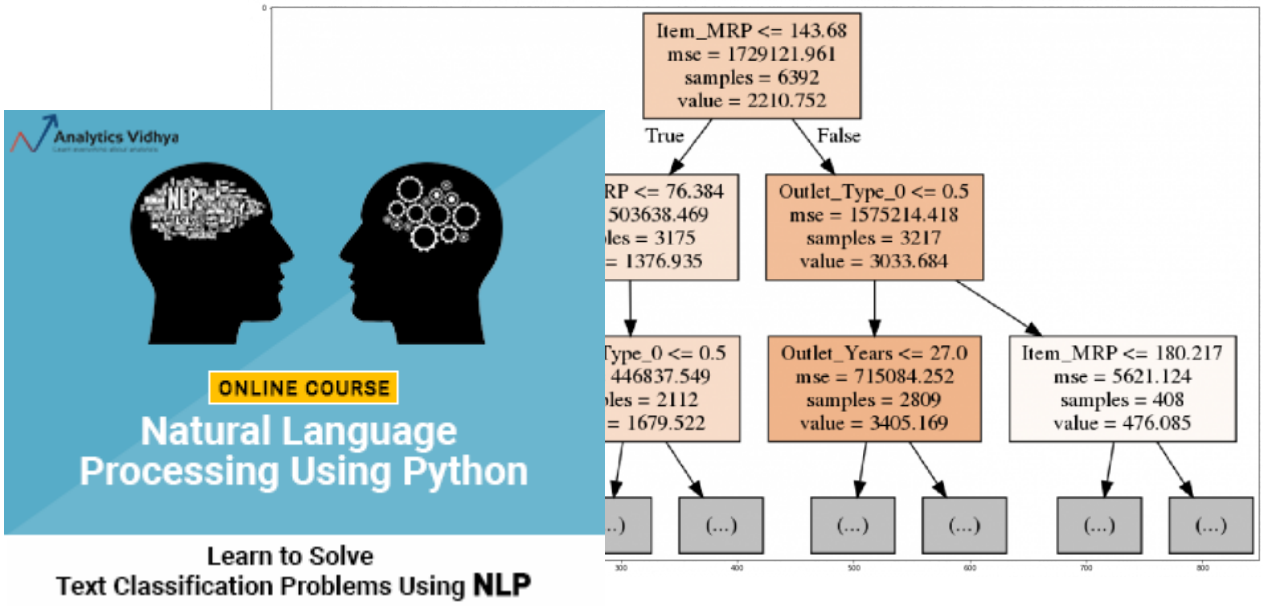


(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=blog&utm_medium=display&utm_campaign=NLPcourse)

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

Next, we will create a surrogate decision tree model for this random forest model and see what we get.

```
# saving the predictions of Random Forest as new target
new_target = rf.predict(X_train)


# defining the interpretable decision tree model
dt_model = DecisionTreeRegressor(max_depth=5, random_state=10)


# fitting the surrogate decision tree model using the training set and new target
dt_model.fit(X_train,new_target)
```



(https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/a.png)
(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=NLP course)

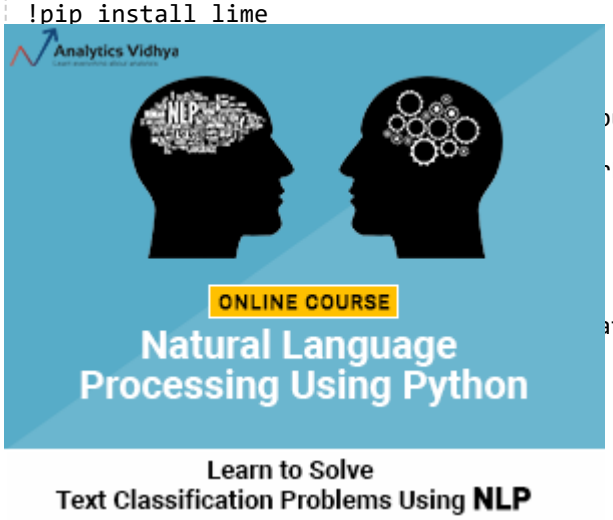
arget and can be used as a surrogate model to explain the
ly, we can use it for any other complex model. Just make sure your
et wrong interpretations (a nightmare!).

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

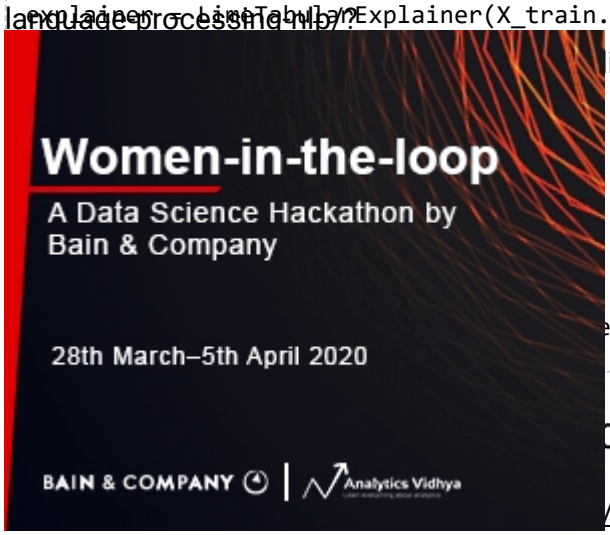
erate local interpretations of black-box models

n R & Python using the LIME package
np into implementation for the same to check the local
E:

```
# installing lime library
!pip install lime
```



(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=Sticky_banner1&utm_medium=display&utm_campaign=NLPcourse)



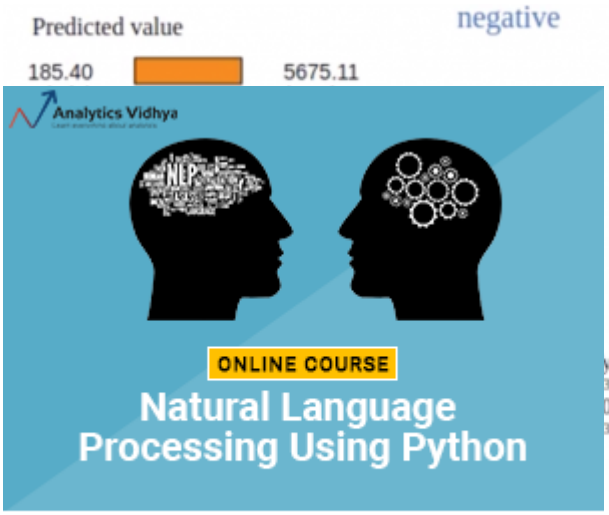
(///datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

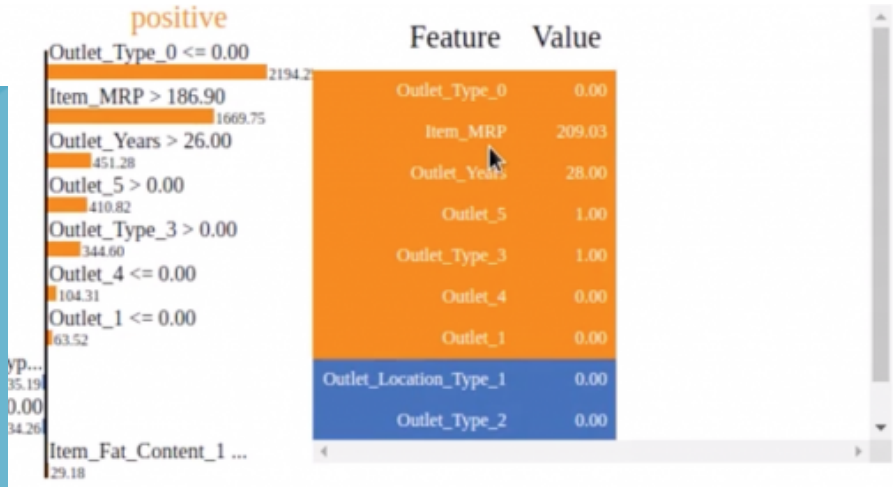
```
ular module of lime library
Explainer

ators=200,max_depth=5, min_samples_leaf=100,n_jobs=-1, random

# creating the explainer function
explainer = LimeTabularExplainer(X_train.values, mode="regression", feature_names=X_train.columns)

ervation)[0]}

0328626019
```

/uploads/2019/08/Screenshot-2019-08-16-at-8.05.48-PM.png)

Generate Explanations using LIME

```
# explanation using the random forest model
explanation = explainer.explain_instance(X_observation.values[0], rf_model.predict)
explanation.show_in_notebook(show_table=True, show_all=False)
print(explanation.score)
```



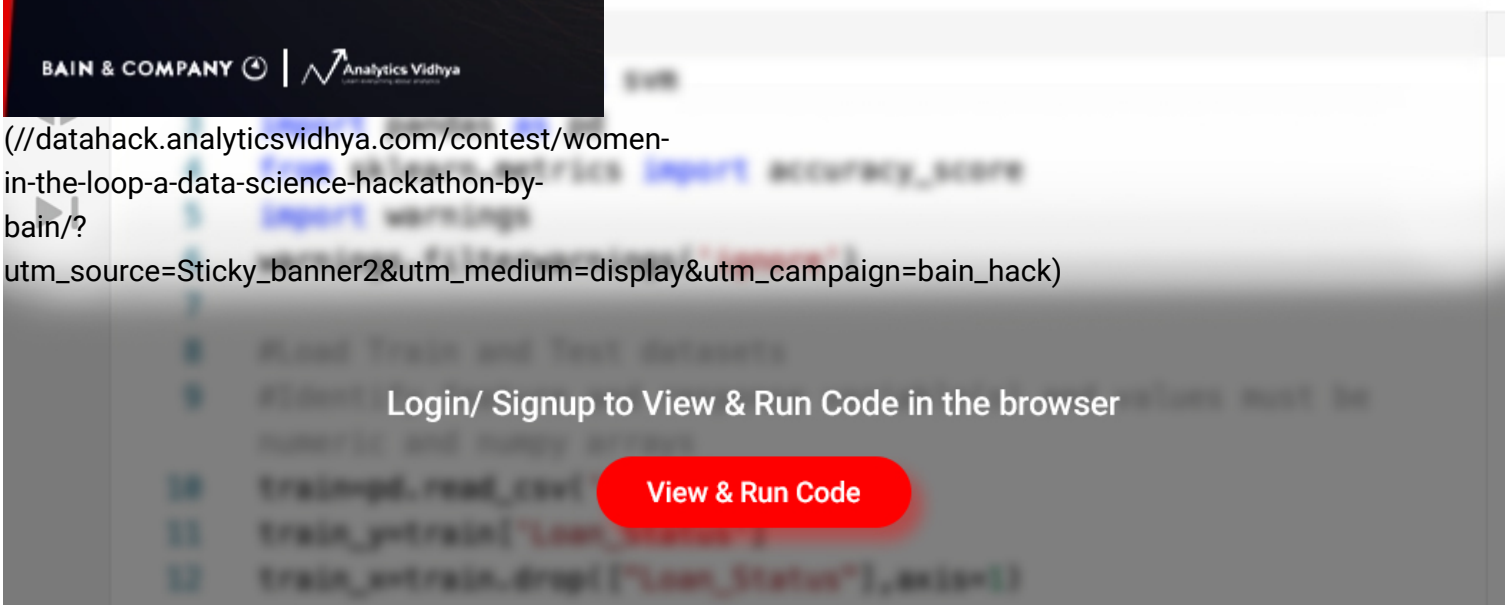


feature's contribution to this prediction is shown in the right bar plot. Orange signifies the positive impact and blue signifies the negative impact of that feature on the target. For example, *Item_MRP* has a positive impact on sales.

(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=blog&utm_medium=display&utm_campaign=NLPcourse)

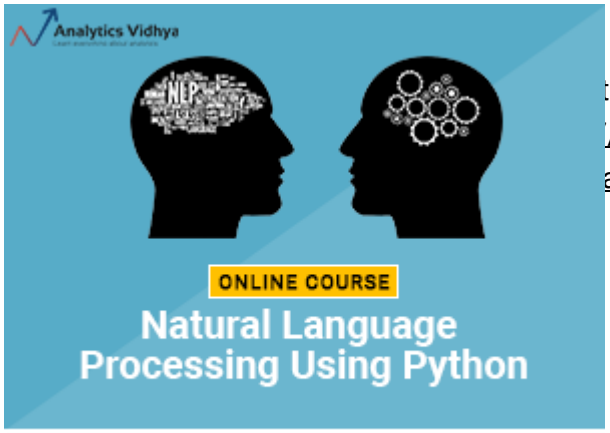
LIME to more complex models such as Xgboost & LightGBM. The already inserted here:

(///datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)



(https://id.analyticsvidhya.com/auth/login/?next=https://www.analyticsvidhya.com/blog/2019/08/decoding-black-box-step-by-step-guide-interpretable-machine-learning-models-python/?utm_source=blog&utm_medium=6-

python-libraries-interpret-machine-learning-models?&utm_source=coding-window-blog&source=coding-window-blog)



(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=blog&utm_medium=decoding-black-box-step-by-step-guide-python)

tion, you can go through this article
/06/building-trust-in-machine-learning-models/?
ack-box-step-by-step-guide-interpretable-machine-learning-models-

dvantages as it relies on locally generated fake data and uses simple linear models to explain predictions. However, it can be used for text and image data as well.

learning&utm_campaign=NLP course)

ehensive end-to-end course:

ses.analyticsvidhya.com/courses/applied-machine-learning-
=blog&utm_medium=decoding-black-box-step-by-step-guide-
python)

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

ny questions or feedback regarding this article, let me know in the
his article on Analytics Vidhya's Android APP

/store/apps/details?
id=com.analyticsvidhya.android&utm_source=blog_article&utm_campaign=blog&pcampaignid=MKT-Other-global-all-co-prtnr-py-PartBadge-Mar2515-1)

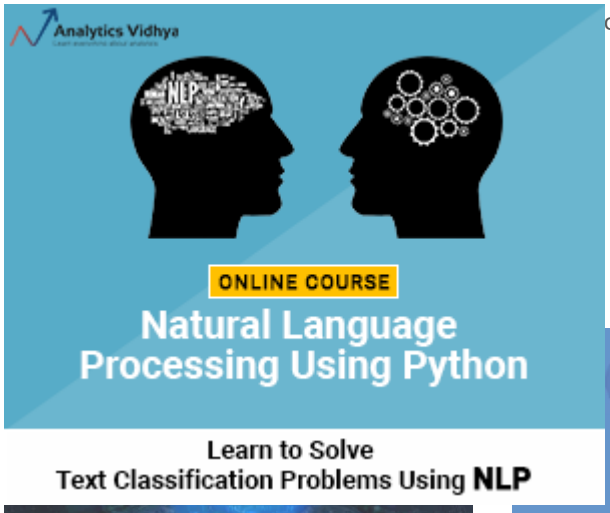
Share this:

(https://www.analyticsvidhya.com/blog/2019/08/decoding-black-box-step-by-step-guide-interpretable-machine-learning-models-python/?share=linkedin&nb=1)

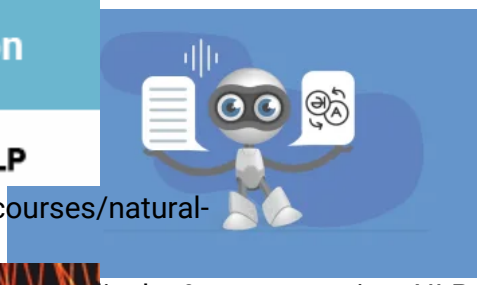
(https://www.analyticsvidhya.com/blog/2019/08/decoding-black-box-step-by-step-guide-interpretable-machine-learning-models-python/?share=facebook&nb=1)

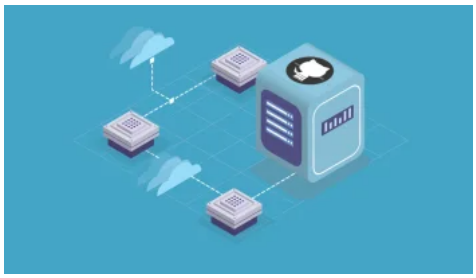
(https://www.analyticsvidhya.com/blog/2019/08/decoding-black-box-step-by-step-guide-interpretable-machine-learning-models-python/?share=twitter&nb=1)

oding-black-box-step-by-step-guide-interpretable-machine-learning-models-



(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?utm_source=blog&utm_medium=display&utm_campaign=NLPcourse)





(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

/03/6 www.analyticsvidhya.com/blog/2019/10/top-
al-language-processing-nlp-
essions-datahack-summit-
engaluru/)
ing NLP Hack Sessions to
ut for at DataHack Summit

/www.analyticsvidhya.com/bl
9/10/top-7-natural-language-
sing-nlp-hack-sessions-
k-summit-2019-bengaluru/)
October 11, 2019
In "Analytics Vidhya"

(https://www.analyticsvidhya.com/blog/2019/09/7-data-science-projects-github-showcase-your-skills/)
Here are 7 Data Science Projects on GitHub to Showcase your Machine Learning Skills!
(https://www.analyticsvidhya.com/blog/2019/09/7-data-science-projects-github-showcase-your-skills/)
September 2, 2019
In "Advanced"

TAGS : INTERPRETABLE MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/INTERPRETABLE-MACHINE-LEARNING/), INTERPRETABLE ML (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/INTERPRETABLE-ML/), LIME (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/LIME/), LIVE CODING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/LIVE-CODING/), MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MACHINE-LEARNING/), MACHINE LEARNING

INTERPRETABILITY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MACHINE-LEARNING-INTERPRETABILITY/), PYTHON (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/PYTHON/)

NEXT ARTICLE

s you Should Learn (in Python, R, Tableau and D3.js)

n/blog/2019/08/11-data-visualizations-python-r-tableau-d3js/)

•••

PREVIOUS ARTICLE

oblem into a Data Problem? A Successful Data Science Leader's Guide

(https://www.analyticsvidhya.com/blog/2019/08/data-science-leader-how-effectively-transition-business-requirements-data-driven-solutions/)

(https://www.analyticsvidhya.com/blog/author/ankit2106/)

Ankit Choudhary (Https://Www.Analyticsvidhya.Com/Blog/Author/Ankit2106/)

IIT Bombay Graduate with a Masters and Bachelors in Electrical Engineering. I have previously worked as a lead decision scientist for Indian National Congress deploying statistical models (Segmentation, K-Nearest Neighbours) to help party leadership/Team make data-driven decisions. My interest lies in putting data in heart of business for data-driven decision making.

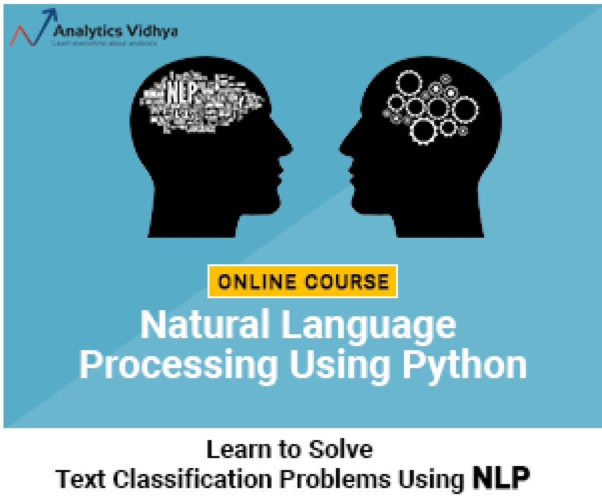
in (https://www.linkedin.com/in/ankit-choudhary-b9360826/)

LEAVE A REPLY

Your email address will not be published.

Comment

POPULAR POSTS



hms (with Python and R Codes)
017/09/common-machine-learning-algorithms/)

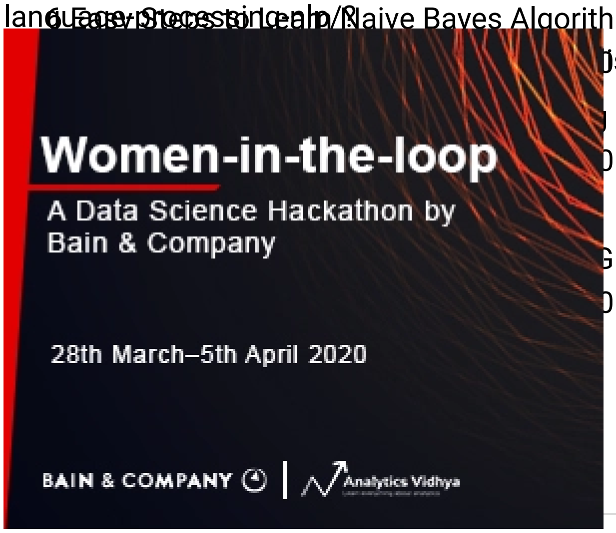
/! (https://www.analyticsvidhya.com/blog/2015/08/comprehensive-

ng) Projects To Boost Your Knowledge and Skills (& can be
dhya.com/blog/2018/05/24-ultimate-data-science-projects-to-

VM) algorithm from examples (along with code)
017/09/understaing-support-vector-machine-example-code/)

Science from Scratch
(https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/)

6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R
017/09/naive-bayes-explained/)



for (aspiring) data scientists
017/02/lintroductory-guide-on-linear-programming-explained-in-

GBoost with codes in Python
016/03/complete-guide-parameter-tuning-xgboost-with-codes-

6 Python Libraries to Interpret Machine Learning Models and Build Trust (https://www.analyticsvidhya.com/blog/2020/03/6-python-libraries-interpret-machine-learning-models/)

MARCH 25, 2020

Using Graphs to Identify Social Media Influencers (https://www.analyticsvidhya.com/blog/2020/03/using-graphs-to-identify-social-media-influencers/)

MARCH 24, 2020

Coronavirus Analysis: Will Social Distancing Help Prevent the Spread? (https://www.analyticsvidhya.com/blog/2020/03/coronavirus-analysis-will-social-distancing-help-prevent-the-spread/)

MARCH 23, 2020

Free GPUs for Everyone! Get Started with Google Colab for Machine Learning and Deep Learning (https://www.analyticsvidhya.com/blog/2020/03/google-colab-machine-learning-deep-learning/)

(http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp/?isplay&utm_campaign=NLPcourse)

(//datahack.analyticsvidhya.com/contest/women-in-the-loop-a-data-science-hackathon-by-bain/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=bain_hack)

(https://www.analyticsvidhya.com/)

Download App

(https://play.google.com/store/apps/details?id=com.analyticsvidhya.android)

(https://apps.apple.com/us/app/analytics-vidhya/id1470025572)

Analytics Vidhya

About Us (https://www.analyticsvidhya.com/about-me/)

Our Team (https://www.analyticsvidhya.com/about-me/team/)

Careers (https://www.analyticsvidhya.com/about-me/career-analytics-vidhya/)

Contact us (https://www.analyticsvidhya.com/contact/)

Data Science

Blog (https://www.analyticsvidhya.com/blog/)

Hackathon (https://datahack.analyticsvidhya.com/)

Discussions (https://discuss.analyticsvidhya.com/)

Apply Jobs (https://www.analyticsvidhya.com/jobs/)

Companies

Post Jobs (https://www.analyticsvidhya.com/corporate/)

Trainings (https://courses.analyticsvidhya.com/)

Hiring Hackathons (https://datahack.analyticsvidhya.com/)

Advertising (https://www.analyticsvidhya.com/contact/)

Visit us

(https://www.linkedin.com/company/analytics-vidhya/) (https://www.facebook.com/AnalyticsVidhya/) (https://twitter.com/analyticsvidhya) (https://www.youtube.com/channel/UCH6gDteHtH4hg3o2343iObA)

Privacy Policy Terms of Use Refund Policy

×

-

(http://play.google.com/store/apps/details?id=com.analyticsvidhya.android)