

Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

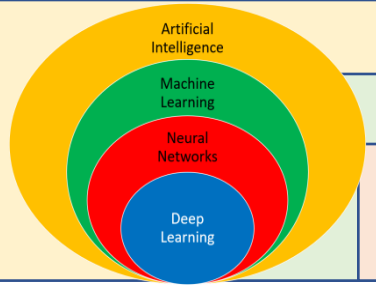
Advanced Deep Learning

Dr. Rastgoo









Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

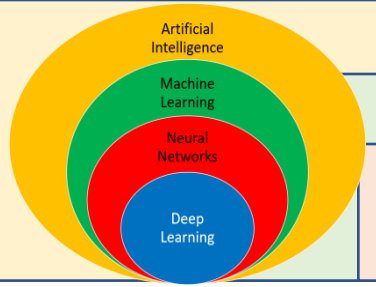
Deep Learning (DL)

Introduction

- Machine learning models are powerful but hard to interpret. However, SHAP values can help you understand how model features impact predictions.

- Machine learning models are becoming increasingly complex, powerful, and able to make accurate predictions. However, as these models become "**black boxes**," it's even harder to understand how they arrived at those predictions. This has led to a growing focus on machine learning interpretability and explainability.

- For example, you applied for a loan at a bank but were rejected. You want to know the reason for the rejection, but the customer service agent responds that an algorithm dismissed the application, and they cannot determine the reason why. This is frustrating, right? You deserve an explanation for the decision that affects you. That's why companies try to make their machine learning models more transparent and understandable.



Artificial Intelligence (AI)

Machine Learning (ML)

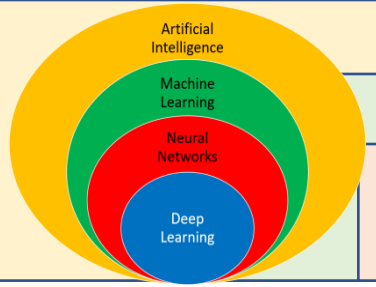
Neural Networks (NNs)

Deep Learning (DL)

Introduction

- One of the most promising tools for this process is SHAP values, which measure how much each feature (such as income, age, credit score, etc.) contributes to the model's prediction.

- SHAP values can help you see which features are most important for the model and how they affect the outcome.



Artificial Intelligence (AI)

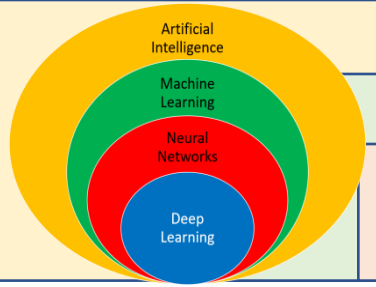
Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

What are SHAP Values?

- SHAP (SHapley Additive exPlanations) values are a way to explain the output of any machine learning model.
- It uses a game theoretic approach that measures each player's contribution to the final outcome.
- In machine learning, each feature is assigned an importance value representing its contribution to the model's output.
- SHAP values show how each feature affects each final prediction, the significance of each feature compared to others, and the model's reliance on the interaction between features.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

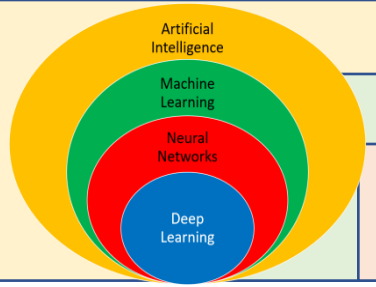
Deep Learning (DL)

SHAP Values in Machine Learning

- SHAP values are a common way of getting a consistent and objective explanation of how each feature impacts the model's prediction.

- SHAP values are based on game theory and assign an importance value to each feature in a model. Features with positive SHAP values positively impact the prediction, while those with negative values have a negative impact. The magnitude is a measure of how strong the effect is.

- SHAP values are model-agnostic, meaning they can be used to interpret any machine learning model, including:
  - ✓ Linear regression
  - ✓ Decision trees
  - ✓ Random forests
  - ✓ Gradient boosting models
  - ✓ Neural networks.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

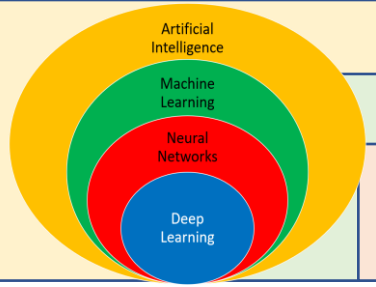
Deep Learning (DL)

The Properties of SHAP Values

- SHAP values have several useful properties that make them effective for interpreting models:

- **Additivity**: SHAP values are additive, which means that the contribution of each feature to the final prediction can be computed independently and then summed up. This property allows for efficient computation of SHAP values, even for high-dimensional datasets.

- **Local accuracy**: SHAP values add up to the difference between the expected model output and the actual output for a given input. This means that SHAP values provide an accurate and local interpretation of the model's prediction for a given input.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

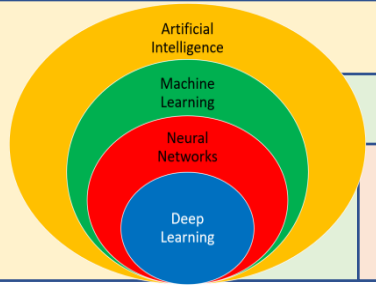
The Properties of SHAP Values

- SHAP values have several useful properties that make them effective for interpreting models:

- **Missingness:** SHAP values are zero for missing or irrelevant features for a prediction. This makes SHAP values robust to missing data and ensures that irrelevant features do not distort the interpretation.

- **Consistency:** SHAP values do not change when the model changes unless the contribution of a feature changes. This means that SHAP values provide a consistent interpretation of the model's behavior, even when the model architecture or parameters change.

- Overall, SHAP values provide a consistent and objective way to gain insights into how a machine learning model makes predictions and which features have the greatest influence.



How to Implement SHAP Values in Python?

- Here, we will calculate SHAP values and visualize feature importance, feature dependence, force, and decision plot..

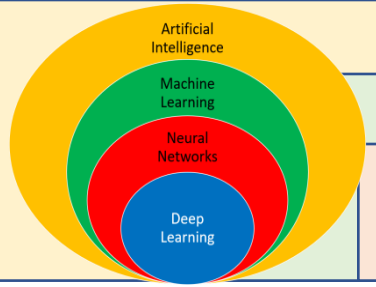
	Call Failure	Complaints	Subscription Length	Charge Amount	Seconds of Use	Frequency of use	Frequency of SMS
0	8	0	38	0	4370	71	
1	0	0	39	0	318	5	
2	10	0	37	0	2453	60	3
3	10	0	38	0	4198	66	
4	3	0	38	0	2393	58	

5 rows

- Load the Telecom Customer Churn. The dataset looks clean, and the target column is "Churn."

Model Training and Evaluation

- Create X and y using a target column and split the dataset into train and test.
- Train Random Forest Classifier on the training set.
- Make predictions using a testing set.
- Display classification report.



Model Training and Evaluation

```
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split

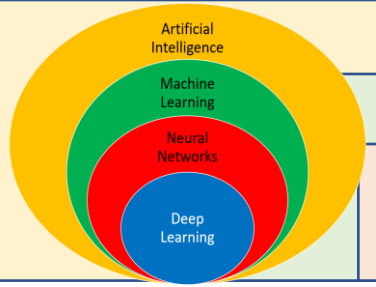
X = customer.drop("Churn", axis=1) # Independent variables
y = customer.Churn # Dependent variable

# Split into train and test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)

# Train a machine learning model
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier()
clf.fit(X_train, y_train)

# Make prediction on the testing data
y_pred = clf.predict(X_test)

# Classification Report
print(classification_report(y_pred, y_test))
```



Artificial Intelligence (AI)

Machine Learning (ML)

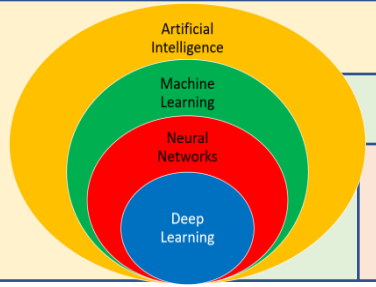
Neural Networks (NNs)

Deep Learning (DL)

Model Training and Evaluation

- The model has shown better performance for "0" label than "1" due to an unbalanced dataset. Overall, it is an acceptable result with 94% accuracy.

	precision	recall	f1-score	support
0	0.97	0.96	0.97	815
1	0.79	0.82	0.80	130
accuracy			0.94	945
macro avg	0.88	0.89	0.88	945
weighted avg	0.94	0.94	0.94	945

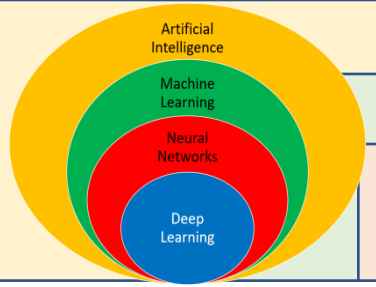


Setting up SHAP Explainer

- Now comes the model explainer part.
- We will first create an explainer object by providing a random forest classification model, then calculate SHAP value using a testing set.

```
explainer = shap.Explainer(clf)
shap_values = explainer.shap_values(X_test)
```

- `shap.Explainer(clf)` : This line creates an explainer object for the classifier `clf`. The `Explainer` class in SHAP is used to interpret the model's predictions.
- `explainer.shap_values(X_test)` : This line computes the SHAP values for the test dataset `X_test`. SHAP values help in understanding the contribution of each feature to the predictions made by the model.



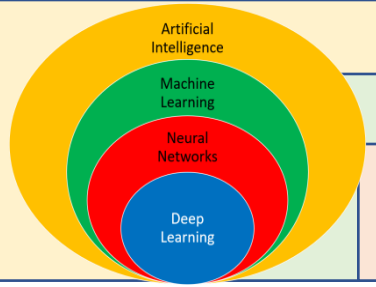
Setting up SHAP Explainer

- The SHAP value of feature \(i\) is calculated using the following equation:

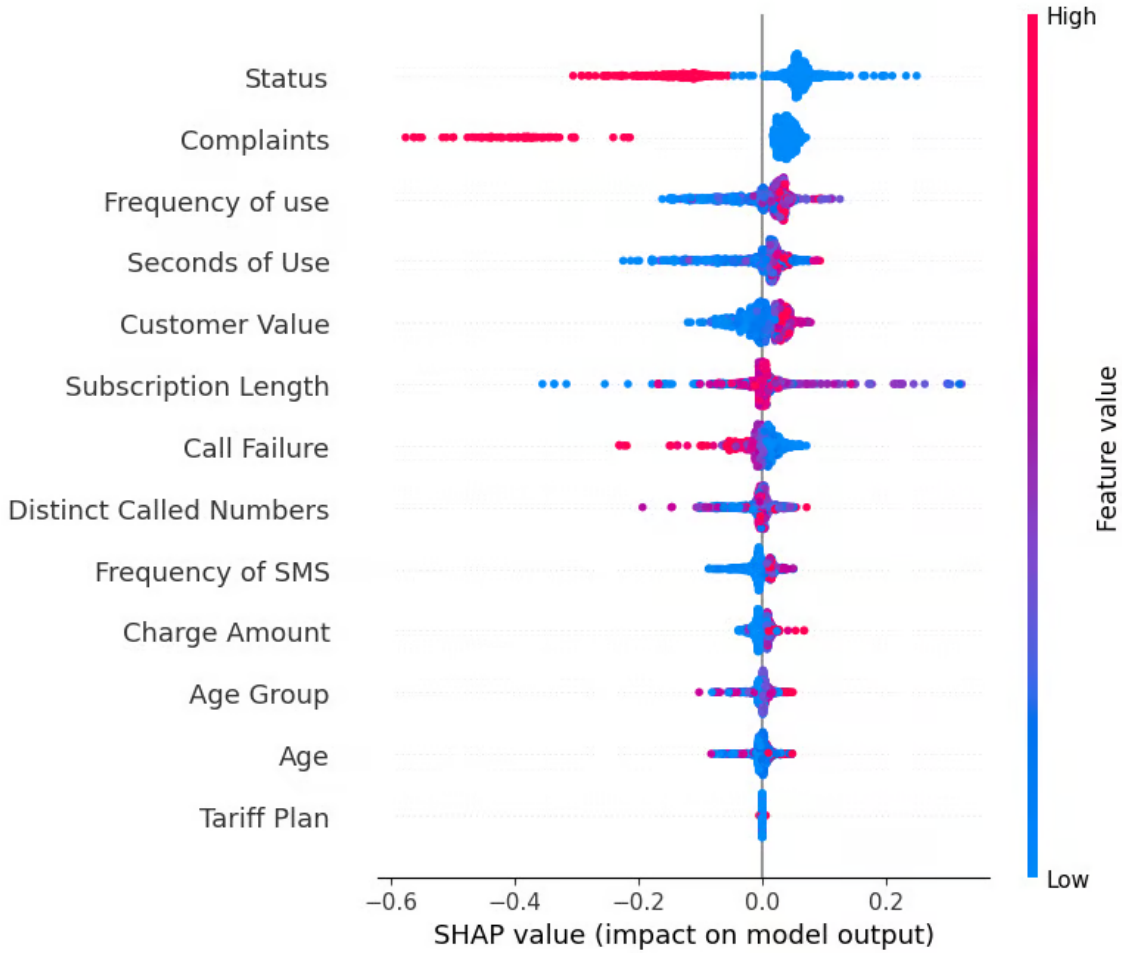
$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f(S \cup \{i\}) - f(S)]$$

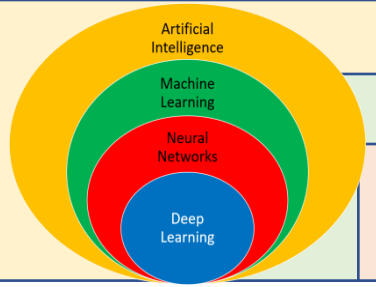
Where:

- F: Set of all input features.
- S: A subset of features excluding feature i.
- $f(S)$: Model prediction using only features in subset S.
- $f(S \cup \{i\})$: Model prediction after adding feature i.
- The factorial term represents the weighting of each feature combination.



Setting up SHAP Explainer





Artificial Intelligence (AI)

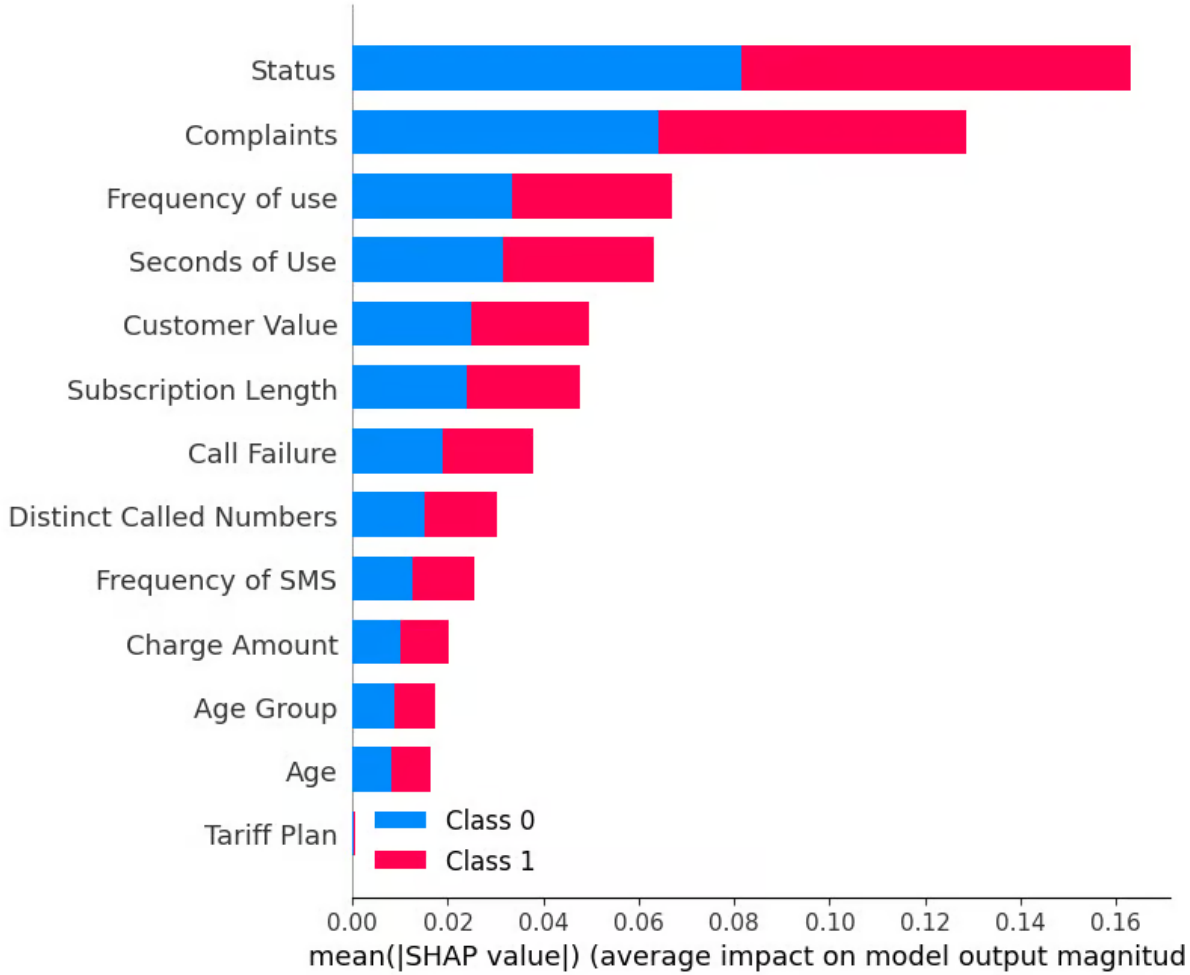
Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

Setting up SHAP Explainer

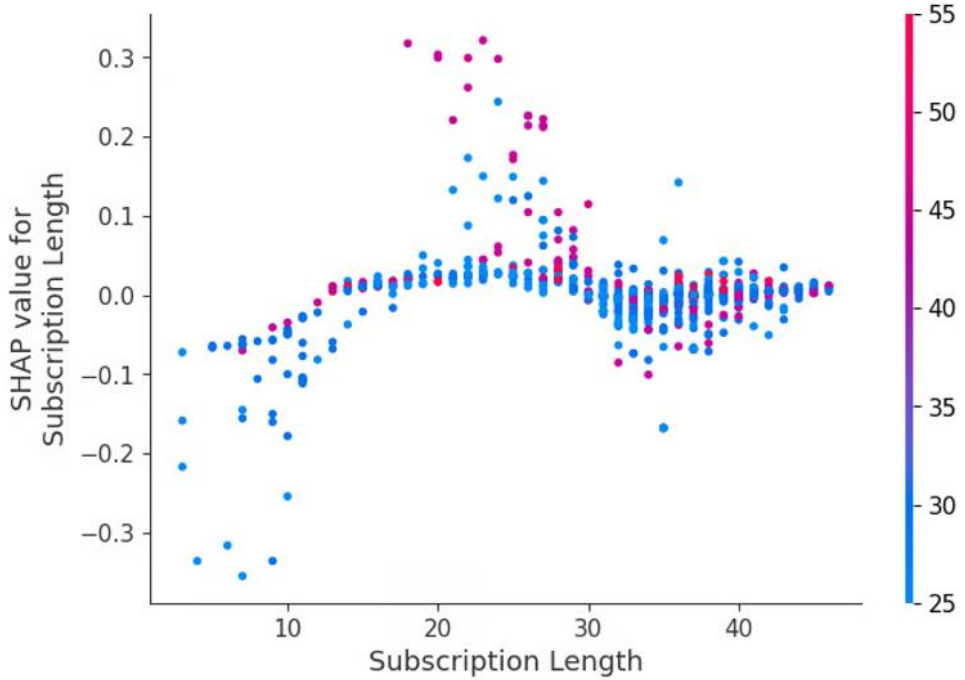
```
shap.summary_plot(shap_values, X_test)
```



Dependence Plot

- A dependence plot is a type of scatter plot that displays how a model's predictions are affected by a specific feature (Subscription Length). On average, subscription lengths have a mostly positive effect on the model.

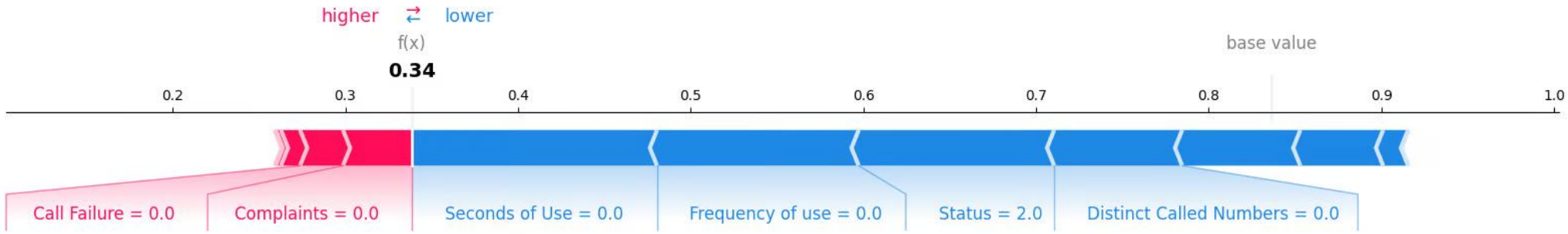
- **X-axis:** Actual values of a selected input feature.
- **Y-axis:** SHAP values corresponding to that feature.
- **Each dot:** One observation (sample) from the dataset.
- **Color:** Values of another feature, often representing a potential interaction with the selected feature.



Force Plot

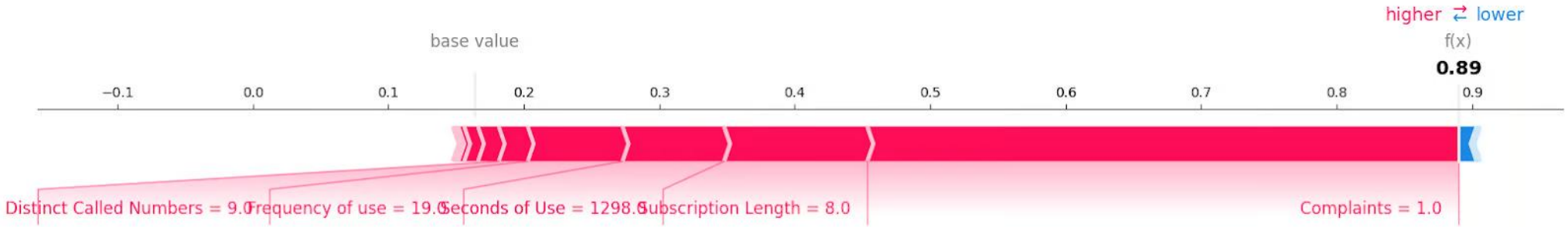
- We will examine the first sample in the testing set to determine which features contributed to the "0" result. To do this, we will utilize a force plot and provide the expected value, SHAP value, and testing sample.

- We can clearly see that zero complaints and zero call failures have contributed to negative to loss of customers.



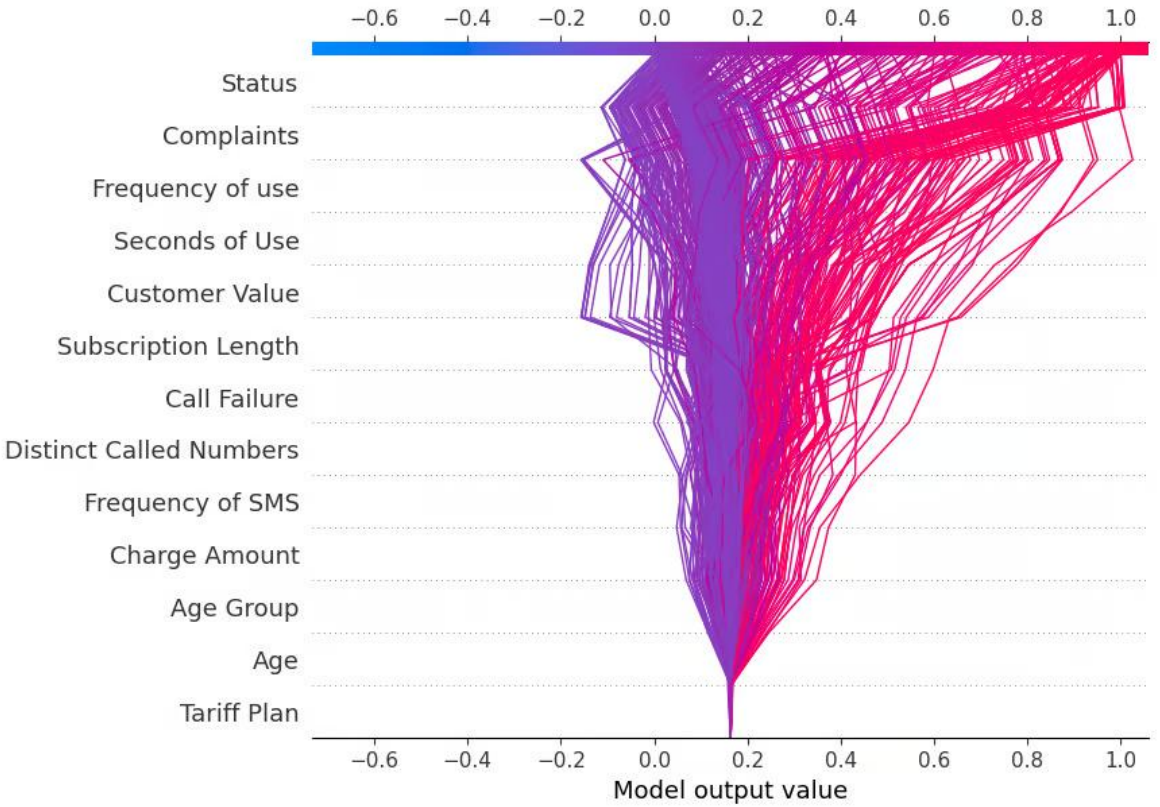
Force Plot

- Let's look at customer churn samples with label "1".
- You can see all of the features with the value and magnitude that have contributed to a loss of customers.
- It seems that even one unresolved complaint can cost a telecommunications company.



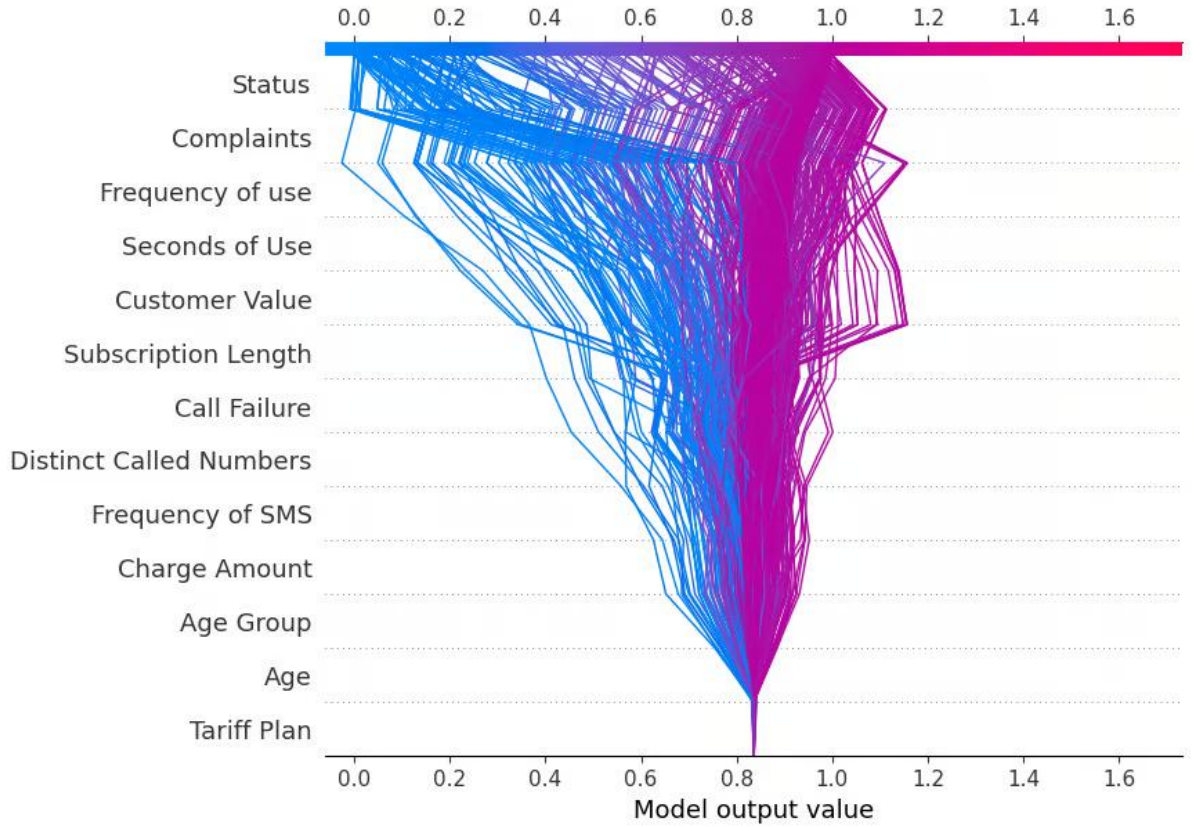
Decision Plot

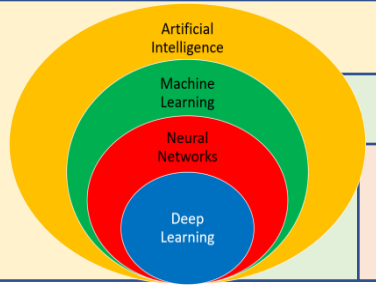
- We will now display the decision plot.
- It visually depicts the model decisions by mapping the cumulative SHAP values for each prediction.
- Each plotted line on the decision plot shows how strongly the individual features contributed to a single model prediction, thus explaining what feature values pushed the prediction.
- The target label "1" decision plot is tilted towards "1".



Decision Plot

- For the decision plot is tilted towards “0”.





Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

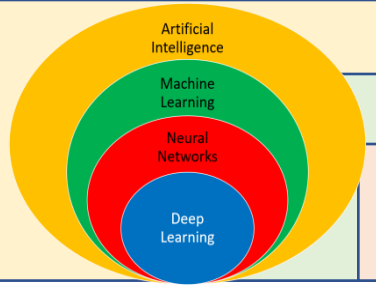
Application of SHAP Values

- Apart from machine learning interpretability and explainability, SHAP value can be used for:

- **Model debugging**: By examining the SHAP values, we can identify any biases or outliers in the data that may be causing the model to make mistakes.

- **Feature importance**: Identifying and removing low-impact features can create a more optimized model.

- **Anchoring explanations**: We can use SHAP values to explain individual predictions by highlighting the essential features that caused that prediction. It can help users understand and trust a model's decisions.



Artificial Intelligence (AI)

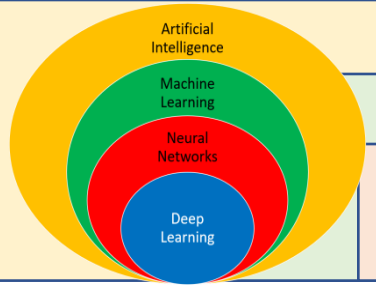
Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

Application of SHAP Values

- **Model summaries**: It can provide a global summary of a model in the form of a SHAP value summary plot. It gives an overview of the most important features across the entire dataset.
- **Detecting biases**: The SHAP value analysis helps identify if certain features disproportionately affect particular groups. It enables the detection and reduction of discrimination in the model.
- **Fairness auditing**: It can be used to assess a model's fairness and ethical implications.
- **Regulatory approval**: SHAP values can help gain regulatory approval by explaining the model's decisions.



Artificial Intelligence (AI)

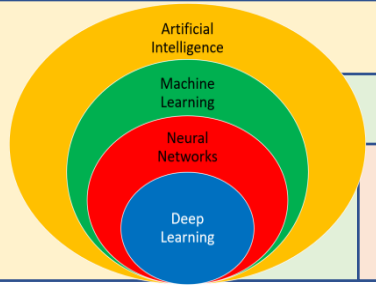
Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

Conclusion

- We have explored SHAP values and how we can use them to provide interpretability for machine learning models.
- While having an accurate model is essential, companies need to go beyond accuracy and focus on interpretability and transparency to gain the trust of users and regulators.
- Being able to explain why a model made a particular prediction helps debug potential biases, identify data issues, and justify the model's decisions.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

