PDPilot User Study Protocol

Introduction

Duration: ~15 minutes

To begin the study, I will share my screen and use this deck of slides to introduce the participant to the study and review how PDPs and ICE plots are computed and visualized.

Once I start the recording, I will have the participant affirm that they agree to being recorded and that they have received and reviewed the Participant Information Sheet.

Tutorial

Duration: ~15 minutes

Next, I will give the participant a tutorial on how to use PDPilot. The tutorial will be based on the following script. In the tutorial script, text in purple italics represents the actions that I will be performing in PDPilot and showing the participant.

Tutorial script:

PDPilot is an interactive tool that helps machine learning practitioners analyze the behavior of a model through partial dependence plots (PDPs) and individual conditional expectation (ICE) plots.

Here we are analyzing a model trained on the telephone churn dataset. This model predicts how likely a telephone service customer is to switch providers.

The user interface of PDPilot is organized into three tabs: One-way Plots, Two-way Plots, and Detailed Plot. The One-way Plots tab shows a grid of PDP + ICE plots, each containing a single feature. Each plot shows the ICE lines in gray and the partial dependence line in black. Above each plot is a histogram showing the distribution of the feature's values in the dataset.

In the row of controls at the top, the arrow buttons allow you to change pages.

In PDPilot: go to the second page and then back to the first.

The "Plot" dropdown menu controls the type of visualization shown. The default is the standard ICE plot. Alternatively, you can show centered ICE plots, where all of the lines start at y = 0. Lastly, you can cluster the ICE lines and show the mean of each cluster.

Switch between the three visualization types.

The “Scale locally” checkbox determines whether each plot has the same y-axis or each plot has its y-axis scaled to fit its own data. For standard ICE plots, scaling locally may not have any effect if the ICE lines already take up the full range of y-values. In this case, local scaling may have a bigger effect when looking at centered or clustered ICE plots.

Toggle scale locally. Switch to normal ICE plots. Toggle scale locally again.

Brushing the lines on an ICE plot highlights the lines for those instances across all of the plots.

Brush some of the lines on one of the total day minutes plot.

When you brush lines, the histograms update to show both the distribution of the highlighted instances and the distribution of the entire dataset. The green bars show the distribution of the highlighted instances. The transparent bars with black outline show the distribution of the entire dataset. The two plots are overlaid so that you can compare the highlighted distribution to the overall distribution.

Clear the highlighted lines.

The “Sort” dropdown menu controls how the plots are sorted. The default is sorting by importance. This metric ranks the plots in descending order of the average amount of variance in the y-values of their ICE lines. The plots that have more variation are shown first. The first ranked plot is in the top-left corner. The rankings then descend left to right, row by row.

The cluster difference metric ranks the plots in descending order of the total distance from the centers of the clusters to the partial dependence line. The plots that have the most different clusters are shown first. This metric is paired best with the centered or clusters visualizations.

Switch to showing cluster centers and change the sorting metric to cluster difference.

The remaining two sorting metrics are used in coordination with brushing ICE lines.

The highlighted line similarity metric can be useful to identify if a cluster of instances in one plot is also a cluster in any others. Plots where the highlighted lines are closer together and farther from the partial dependence line are ranked first. This metric is best paired with the centered ICE lines visualization.

Switch to showing centered ICE lines.

For example, we see a cluster of lines in the plot for “total international charge” that has a different trend than the rest of the lines. If we highlight these lines and then sort by highlighted similarity, we can see that this subset of instances also shows outlying trends for total international calls and total international minutes.

Select the lines in the plot for total international charge and then sort by highlighted similarity.

Now that we've identified that there is a cluster of instances that show different trends across multiple features, we want to know what is similar about these instances. To help with this, we can sort by the highlighted histogram difference metric. This measures the distance between the feature's distribution for the highlighted instances to the feature's distribution for all instances. The plots are sorted so that the features whose highlighted distributions are most different from the overall distribution are shown first. When this metric is selected, you want to be looking at the histograms and bar charts above the ICE plots.

In this case, sorting by highlighted histogram difference metric ranks the "international plan" feature first. We can see that nearly all of the highlighted instances are for international plans, whereas in the whole dataset, nearly all of the instances are for non-international plans. Now we have a better idea of what defines this subset of instances.

If you change the highlighted instances after selecting one of the highlighted metrics, then clicking the refresh button next to the dropdown menu will update the rankings.

Highlight another set of instances and then refresh the rankings.

If we need a reminder of what the different sorting metrics mean, we can hover over the question mark icon next to the dropdown menu, or reference the documentation.

The left sidebar contains controls for filtering the plots. If no filters are selected, then all of the plots are shown.

We can filter the plots by feature type. Ordered features have values with defined orders, such as quantities. Nominal features have values without defined orders, such as categories.

Toggle the ordered and nominal filters.

For ordered features, we can filter by whether the feature's PDP is generally increasing, decreasing, or sometimes increasing and sometimes decreasing. For example, we can only show PDPs that are generally increasing.

Clear the filters and select "Increasing." Switch to showing cluster centers and scale locally.

It can be easier to see the shape of the PDPs by switching the plot to show clusters and scaling locally.

Switch the plot to show clusters and scale locally.

We could also show PDPs that show both increasing and decreasing behavior.

Deselect “Increasing” and select “Mixed”.

We can also filter the plots by feature name. For example, I can only show plots for features about international calls and plans.

Clear filters. Go back to the centered ICE plots and uncheck scale locally. Search for “int” and select all of the features. Clear the filters again.

If we hover over a plot, there is an expand button in the bottom-left corner. Clicking on this button shows this feature in the Detailed Plot tab. In this tab, we can view a larger version of the plot, compare the feature values to the ground truth labels, and go into more depth on its clusters.

Sort the plots by cluster distance and select the plot for “number of customer service calls.” Expand this plot.

In this tab, the “Clusters” visualization shows each cluster in its own plot. The ICE lines in each cluster are shown in lighter-colored lines. The mean of each cluster is shown in a darker-colored line. The partial dependence line is shown in black in each plot.

For the “number of customer service calls” feature, we see two clusters.

Select the cluster visualization.

However, it looks like the orange cluster should be split into two separate clusters. To improve the clustering, we can try to increase the number of clusters.

Increase the number of clusters to 3.

This looks better. The green cluster stands out in that the model has learned that for some customers, a higher number of customer service calls can decrease their likelihood of churning.

In order to understand what kind of instances make up each cluster, we can select “Cluster Descriptions” on the right. This visualizes the distributions of the instances in each cluster for a handful of features. PDPilot automatically chooses features that are helpful in describing the clusters. The features are ranked top to bottom by how helpful they are in separating the clusters.

With this, we can see that the green cluster consists of customers who have higher total day minutes and total evening minutes than the other two clusters.

Select Cluster Descriptions.

To further investigate the clusters, we can brush the axes in these feature distribution visualizations to filter the ICE lines. For example, we might be interested in seeing what instances with low total day minutes are getting put in the green cluster.

Brush the lower half of total day minutes.

To make the bar charts show percentages instead of counts, we can check “Normalize bar charts”.

Normalize bar charts.

If you are unsatisfied with the default clustering, you can adjust the number of clusters between 2 and 5. Note that adjusting the number of clusters may change the ranking and visualization for this feature in the one-way plots tab.

Increase the number of clusters to 5.

You may need to scroll over the cluster descriptions to see all of the plots.

Scroll over the clusters. Then decrease the number of clusters back to 3.

In addition to adjusting the number of clusters, you can also manually edit the clusters. To do this, you can brush the lines in a cluster and choose to send those lines to a different cluster or to a new cluster.

Brush flat lines in the green cluster and move them to a new cluster.

You can also brush instances in the Detailed Plot tab. This can be useful if you want to select instances based on the value of a given feature. For instance, we can select all instances where the customer has a voicemail plan.

Return to the One-way Plots tab. Select the voicemail plan feature. Brush the instances that have a voicemail plan.

The highlighted histogram difference metric can then tell us what other features make the selected instances distinct from the whole dataset. In this case, we people with voicemail plans have more voicemail messages, as expected.

Sort by the highlighted histogram difference metric.

In addition, we can also sort by highlighted line similarity to see if the lines for people who have voicemails are clustered together in any plot. Here we see they are clustered together for the total day minutes feature.

Sort by the highlighted line similarity.

Next we'll turn our attention to the Two-way Plots tab, which shows a grid of PDPs, each containing two features. The default color scale visualizes the average predictions.

Switch to the Two-way Plots tab.

Using the "Color" dropdown menu, we can change the color scale to visualize the interaction between the pairs of features. A positive value indicates that the features are interacting in a way that makes the average prediction higher than expected if there was no interaction. A negative value indicates that the features are interacting in a way that makes the average prediction lower than expected.

If we need a reminder about what the color scale shows, we can hover over the question mark icon next to the color scale or reference the documentation.

Switch to show interactions. Hover over the question mark icon. Switch to show predictions. Hover over the question mark icon.

The paging and local scaling behave similarly as in the One-way Plots tab.

For filtering two-way plots, you can specify whether both features in the plots must be selected or just one feature in the plots need to be selected. For example, to show all calculated two-way PDPs for a specific feature, you can select that plots must contain "1+ selected feature" and then select a feature.

Select "1+ selected feature" and then select "total eve minutes"

Alternatively, we can only show plots where both features are selected.

Select "account length", "international plan", "total day minutes", and "total eve minutes". Then clear the filters.

Note PDPilot does not compute all possible two-way plots. It only pre-computes a plot when it expects the pair of features to interact with each other.

Opening a two-way plot in the Detailed Plot tab will show the two-way plot in both color scales. The one-way PDPs for the features are shown in the margins. In addition, we also visualize the relationship between the features and the ground truth labels.

Expand one of the two-way plots.

For two-way plots, you can click the “Flip” button to swap the x and y axes.

Click the Flip button.

In the Detailed Plot tab, you can also switch to looking at a different feature or pair of features. If the two-way PDP for a pair of features was not pre-computed, then you can click “Compute Now” to calculate it.

Switch to a pair of features that was not pre-computed. Click “Compute Now”.

After Tutorial:

After going through the tutorial, I will ask the participant if they have any questions about how to use PDPilot.

Verification

Duration: ~20 minutes

Next, I will have them share their screen and run PDPilot on their computer.

This GitHub repository contains the code to get them set up. Once the tool is working, I will verify that the participant knows how to use it. I will have them analyze a model trained on the Bike Sharing dataset.

I will ask them to answer the following questions or perform the following tasks using PDPilot:

- What one-way plot shows the most important feature?
  - What trends do you see in this plot?
- What features have generally increasing shapes?
  - Is there a way that you can answer this question using the filters?
- Visualize the clusters of ICE lines for all of the plots.
  - What feature has the most different clusters of ICE lines?
    - Is there a way that you can answer this question using the rankings?
    - Visualize the clusters in more detail.
      - What is the general shape of each cluster?
      - Describe the instances in each cluster.
- Go back to the “One-way Plots” tab. What’s the distribution of the values for the “humidity” feature in this dataset?
- Switch the visualizations back to show the centered ICE lines. Highlight the ICE lines at the top of the first peak in the plot for the “hour” feature.

    - What instances are represented by these lines? In other words, what are the feature values that distinguish these instances from the other instances?
        - *Make sure they know how to use sorting to answer this question.*
- In the “weather situation” plot, highlight the lines that decrease when it’s raining.
    - What instances are represented by these lines?
- In the “days since 2011” plot, highlight the lines in the top right.
    - Are these instances outliers or clustered together in any other plots?
        - *Make sure they know how to use sorting to answer this question.*
    - What instances are represented by these lines?
- What pair of features has the most interaction?
    - Describe how they interact.
- Visualize interactions instead of predictions. Scale each plot locally.
- For the top pair of features, visualize the interactions and predictions side by side.

Model Exploration

Duration: ~40 minutes

I will provide the participant with the Ames, Iowa Housing dataset, which contains homes and their sale prices, and a pre-trained machine learning model on this dataset. Once they load the model into PDPilot, I will provide them with the following prompt:

We have trained a model that predicts the sale price of homes in Ames, Iowa. You have 30 minutes to analyze the behavior of this model using PDPilot. Please think aloud while you work so that we can understand your process, questions, and insights. You may reference the tool’s documentation and the dataset’s data dictionary. You may ask questions about how to use PDPilot.

Interview

Duration: ~15 minutes

This will be a semi-structured interview that is guided by the questions in the same deck of slides as used in the introduction.