Interpreting a machine learning model

Mark Romanowsky, Data Scientist - Explainable AI
Rajiv Shah, Data Scientist - Customer Success
Agenda

- Interpretability: what, why, when, how?
- Understand which features mattered
- Understand what patterns were found
- Understand a particular prediction
Interpretability: what, why, when, how?

Model interpretability tools are important through the whole model lifecycle
When and how can interpretability help?

- Verify the model against domain knowledge
- Curate a dataset
- Improve the model

- Track data drift in important features over time

- Understand what drives an individual model decision and take action
Interpretability, not just transparency

- "Transparent" / "white-box" models aren't the full solution: can still be too complex to understand by inspection (e.g. 100s of terms).
- Some algorithms have good accuracy but don't have coefficients or concise visual descriptions. We still want to understand them too.

We need tools to provide human-level interpretation of what a model learned and how it predicts.

Ideally, they are model-agnostic and provide the required level of detail.
Understand which features mattered

Which data is your model looking at the most?
Feature impact: loan application example

![Feature Impact Chart](chart.png)
Why measure feature impact?

- Verify the model against domain knowledge
  - Validate if important features align with domain knowledge
- Improve the model
  - Detect target leakage
- Curate a dataset and simplify a model
  - Features aren’t free. Data collection, processing, verification all cost time and money. Keep features which add signal, drop uninformative ones.
- Track data drift in important features over time
  - Know when it’s time to retrain
Target leakage: loan application example

Accuracy too good to be true?

Only one feature matters?
Target leakage: loan application example

Whoops! We gave the model a feature that lets it cheat. This is target leakage. Covered in detail in another webinar. (See Additional Resources slide at end)
Feature impact for feature selection

- Features can be detected as redundant with other features. (They have similar effects on the target when permuted.)
- You don't get any extra information from the redundant features, so you should think about dropping them.
- This is a good time to drop other low-importance features to simplify data collection and speed up prediction.
Feature impact for feature selection

- Features can be detected as redundant with other features. (They have similar effects on the target when permuted.)
- You don't get any extra information from the redundant features, so you should think about dropping them.
- This is a good time to drop other low-importance features to simplify data collection and speed up prediction.
Feature impact for model monitoring

Once you deploy a model, how do you know it's still working well as time passes?

In MLOps, look at Feature Drift vs Feature Importance. We want to alert early for drift in the most important features.

MLOps is much more than this. More links at the end...
Now that you know your feature impact...

- Use it as a criterion for model selection
- Use it to refine your feature list
- Don't be afraid to start a new project with an updated and improved dataset!
- Keep an eye on important features over time to monitor model health and know when it's time to retrain
Understand what patterns were found

In your important features, how does the data drive the target?
What is the effect of feature X on the target?

- How does a prediction on a typical row change, as you change the values of a feature?
- This is a "Partial Dependence Plot".
Verify Feature Effects against domain knowledge

- Does the plot tell a reasonable story about the feature?
- Monotonic?
- Major jumps or special values?
- Noisy or smooth?

![Graph showing partial dependence](image)
Verify Feature Effects against domain knowledge

- Does the plot tell a reasonable story about the feature?
- Monotonic?
- Major jumps or special values?
- Noisy or smooth?

In this plot: all else equal, probability of loan default decreases as annual income increases. Does this pattern seem correct?

- But what about that big jump at $42k?
- Is data collection different below and above this breakpoint? E.g. filtering for loan app to be listed at all?
- If this is a real effect, it's very interesting...
Verify Feature Effects against domain knowledge

- Does the plot tell a reasonable story about the feature?
- Monotonic?
- Major jumps or special values?
- Noisy or smooth?

In this plot: all else equal, probability of hospital readmission is highest in the 50-90 age range, especially 70-80. Does this pattern seem correct?
- Why would younger patients (0-40) have a lower probability of readmission?
- Why would age 90-100 patients have a lower probability of readmission?
Effects for text features: Word Cloud

- **Color**: direction of the effect
  - Red is positive, blue is negative
- **Shading**: strength of the effect
  - Darker is stronger
- **Size**: word frequency in the data
  - Bigger is more frequent

*Example: job title in a loan application*
Now that you know your feature effects...

- Compare PDP with domain knowledge and business rules to select a model that learned the right patterns.

- Consider retraining "MONO" models with monotonicity constraints, if domain knowledge suggests the effects should be monotonic.

- Consider training models with an updated feature list. Don't be afraid to start a new project!
Understand a particular prediction

Why did this datapoint lead to that prediction?
Prediction explanations: loan applications

<table>
<thead>
<tr>
<th>PREDICTION</th>
<th>EXPLANATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.484</td>
<td>+++ mths_since_last_record = 114</td>
</tr>
</tbody>
</table>

We interpret the model's behavior row-by-row by answering two related questions:
1. What feature values drive a particular prediction the most?
2. How are they driving that prediction?
# Prediction explanations: loan applications

When the model is confident, do its reasons make sense?

<table>
<thead>
<tr>
<th>ID</th>
<th>PREDICTION</th>
<th>EXPLANATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2775</td>
<td>0.484</td>
<td>+++ mths_since_last_record = 114</td>
</tr>
<tr>
<td>6535</td>
<td>0.455</td>
<td>+++ annual_inc = 16080</td>
</tr>
<tr>
<td>6563</td>
<td>0.372</td>
<td>+++ annual_inc = 25200</td>
</tr>
</tbody>
</table>

For the loan applications with high predicted prob. of default, what are the major factors?

<table>
<thead>
<tr>
<th>ID</th>
<th>PREDICTION</th>
<th>EXPLANATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>4465</td>
<td>0.032</td>
<td>--- mths_since_last_record = 0</td>
</tr>
<tr>
<td>8727</td>
<td>0.030</td>
<td>--- mths_since_last_record = 0</td>
</tr>
<tr>
<td>9337</td>
<td>0.002</td>
<td>--- inq_last_6mths = 25</td>
</tr>
</tbody>
</table>

For the loan applications with low predicted prob. of default, what are the major factors?
# Prediction explanations: hospital readmission

When the model is confident, do its reasons make sense?

<table>
<thead>
<tr>
<th>PREDICTION</th>
<th>EXPLANATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.984</td>
<td>+++ number_emergency = 13</td>
</tr>
<tr>
<td>0.963</td>
<td>+++ number_inpatient = 6</td>
</tr>
<tr>
<td>0.959</td>
<td>+++ number_inpatient = 8</td>
</tr>
</tbody>
</table>

For the patients with high predicted prob. of readmission, what are the major factors?

<table>
<thead>
<tr>
<th></th>
<th>EXPLANATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.061</td>
<td>--- discharge_disposition... = &quot;Expired&quot;</td>
</tr>
<tr>
<td>0.052</td>
<td>--- discharge_disposition... = &quot;Expired&quot;</td>
</tr>
<tr>
<td>0.041</td>
<td>--- discharge_disposition... = &quot;Expired&quot;</td>
</tr>
</tbody>
</table>

For the patients with low predicted prob. of readmission, what are the major factors?
Prediction explanations: hospital readmission

When the model is confident, do its reasons make sense?

<table>
<thead>
<tr>
<th>PREDICTION</th>
<th>EXPLANATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.984</td>
<td>+++ number_emergency = 13</td>
</tr>
<tr>
<td></td>
<td>+++ number_inpatient = 4</td>
</tr>
<tr>
<td></td>
<td>+++ discharge_disposition... = &quot;Discharged/transfer...&quot;</td>
</tr>
<tr>
<td>0.963</td>
<td>+++ number_inpatient = 6</td>
</tr>
<tr>
<td></td>
<td>+++ admission_type_id = &quot;MISSING&quot;</td>
</tr>
<tr>
<td></td>
<td>+++ number_emergency = 2</td>
</tr>
<tr>
<td>0.959</td>
<td>+++ number_inpatient = 8</td>
</tr>
<tr>
<td></td>
<td>+++ number_outpatient = 1</td>
</tr>
<tr>
<td></td>
<td>+++ discharge_disposition... = &quot;Discharged/transfer...&quot;</td>
</tr>
</tbody>
</table>

For the patients with high predicted prob. of readmission, what are the major factors?

For the patients with low predicted prob. of readmission, what are the major factors?

A subtle case of target leakage! You don't need a ML model for these rows...
Prediction explanations for human-in-the-loop

Employee attrition: understand who might leave and why. Explanations can suggest next steps to retain them.

<table>
<thead>
<tr>
<th>ID</th>
<th>PREDICTION</th>
<th>EXPLANATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>3151</td>
<td>0.996</td>
<td>+++ JobRole = &quot;Sales Representative&quot; +++ JobInvolvement = 1 +++ MonthlyIncome = 1359</td>
</tr>
<tr>
<td>2522</td>
<td>0.995</td>
<td>+++ JobRole = &quot;Sales Representative&quot; +++ TotalWorkingYears = 1.1 +++ HourlyRate = 37.4</td>
</tr>
<tr>
<td>4819</td>
<td>0.995</td>
<td>+++ OverTime = &quot;Yes&quot; +++ TotalWorkingYears = 1 +++ HourlyRate = 37</td>
</tr>
</tbody>
</table>

*For the employees with high predicted prob. of quitting, what are the major factors?*

<table>
<thead>
<tr>
<th>ID</th>
<th>PREDICTION</th>
<th>EXPLANATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>4860</td>
<td>0.001</td>
<td>-- -- DailyRate = 1246 -- -- NumCompaniesWorked = 1 -- -- JobRole = &quot;Sales Executive&quot;</td>
</tr>
<tr>
<td>5047</td>
<td>0.001</td>
<td>-- -- MonthlyIncome = 23589.6 -- -- TotalWorkingYears = 29.7 -- -- DailyRate = 1176</td>
</tr>
<tr>
<td>2515</td>
<td>0.001</td>
<td>-- -- MonthlyIncome = 18242.4 -- -- YearsWithCurrManager = 2 -- -- TotalWorkingYears = 25.3</td>
</tr>
</tbody>
</table>

*For the employees with low predicted prob. of quitting, what are the major factors?*
Now that you know your prediction explanations...

- **Model development:**
  - Make sure that your model has good reasons for very confident predictions
  - Identify more subtle cases of target leakage or other data issues
  - Advanced usage: *cluster* prediction explanations to create useful new features (see Additional Resources slide at end)

- **Deployed model:**
  - Provide actionable insights for humans involved in the decision
Conclusion: use interpretability tools!

Every model stakeholder can get something out of feature impact, feature effects, or prediction explanations.

- Verify the model against domain knowledge
- Curate a dataset
- Improve the model

- Track data drift in important features over time

- Understand what drives an individual model decision and take action

- Feature Impact
- Feature Effects
- Prediction Explanations
- Word Cloud

- Feature Impact

- Prediction Explanations
Additional Resources:

- **Other DataRobot Community learning sessions on related topics**
  - Prediction explanation clustering: [https://community.datarobot.com/t5/learning-sessions/explanation-clustering/ba-p/5432](https://community.datarobot.com/t5/learning-sessions/explanation-clustering/ba-p/5432)

- **MLOps walkthrough, including data drift and much more:**

- **Handling multicollinearity, using Feature Associations Matrix and Feature Impact**
  - [https://community.datarobot.com/t5/resources/multicollinearity/ta-p/5305](https://community.datarobot.com/t5/resources/multicollinearity/ta-p/5305)

- **Prediction explanations deeper dive**
  - XEMP white paper: [https://www.datarobot.com/resources/xemp-prediction-explanations/](https://www.datarobot.com/resources/xemp-prediction-explanations/)

- **E-book: how to understand a DataRobot model**
  - [https://www.datarobot.com/blog/how-to-understand-a-datarobot-model-ebook-1/](https://www.datarobot.com/blog/how-to-understand-a-datarobot-model-ebook-1/)
Questions & Answers
DataRobot Community

- **Engage, learn, and accelerate** your AI/ML journey
- **Connect** with peers to find solutions to AI challenges
- **Explore** helpful content to take your AI to the next level
- **Build** your brand as an AI expert & thought leader
- **Join** your peers today at **community.datarobot.com**
- **Questions:** learning_sessions@datarobot.com