Model Interpretation guidelines for the enterprise

August 25, 2017
PREDICTIVE MODELING: FUN OR MISERY?
PREDICTIONS OFTEN GO WRONG

Is it going to rain this weekend? I have a thing.

Lemme check. *type type*

...Uhh. What?

**YOUR 10-DAY FORECAST:**

<table>
<thead>
<tr>
<th>Today</th>
<th>Tomorrow</th>
<th>Friday</th>
<th>Saturday</th>
<th>Sunday</th>
</tr>
</thead>
<tbody>
<tr>
<td>☀️</td>
<td>☁️ ☀️</td>
<td>⚡️</td>
<td>⚡️ ☁️</td>
<td>☁️</td>
</tr>
</tbody>
</table>

...Oh! You typed a minus sign in the zip code. The negative zip codes are all like that.

Let's never move there.

Monday Tuesday Tuesday Tuesday Tuesday
I am a Lead data scientist at DataScience.com. I enjoy applying and optimizing classical (Machine Learning) and Bayesian design strategy to solve real-world problems. Currently, I am exploring on better ways to evaluate and explain Model learned decision policies. I am also a member of AAAI and organizer of PyData So Cal meet-up group.
AGENDA

• DEFINE MODEL INTERPRETATION

• UNDERSTAND THE NEED FOR MODEL INTERPRETATION

• DISCUSS DICHOTOMY BETWEEN PERFORMANCE AND INTERPRETATION

• INTRODUCE SKATER

• DISCUSS FIT TO ANALYTICAL WORKFLOW WITH datascience.com PLATFORM

• DEMO

• Q&A
DEFINE INTERPRETATION

- Definition is subjective - Data Exploration to understand the data well
DEFINE INTERPRETATION

- Definition is subjective - overlaps with Model Evaluation
WHAT IS MODEL INTERPRETATION?

- High-level definition: Model interpretation is about understanding machine learning/statistical modeling behavior.

- With model interpretation, one should be able to answer the following questions:
  - **Why** did the model behave in a certain way? What are the relevant variables driving the model outcome?
  - **What** other information can a model provide to avoid prediction errors?
  - **How** can we trust the predictions of a “black box” model?
ERROR VS MODEL COMPLEXITY

Error(x) = Bias^2 + Variance + Irreducible Error

**Reference: Scott Fortmann-Roe**
PREDICTIVE OPTIMISM

Under-fitting:

Over-fitting:

Model Prediction Error vs. Model Complexity

Prediction Error for New Data vs. Training Error

Optimism
WHY MODEL INTERPRETATION?

- Helps in exploring and discovering latent or hidden feature interactions (useful for feature engineering/selection)

- Helps in understanding model variability as the environment changes (once the model is operationalized in a non-stationary environment)

- Helps in model comparison

- Helps an analyst or data scientist build domain knowledge about a particular use case by providing an understanding of interactions
WHY MODEL INTERPRETATION?

- Brings transparency to decision making to enable trust
  - Fair Credit Reporting Act (FCRA) U.S. Code § 1681

SUBCHAPTER III—CREDIT REPORTING AGENCIES

§ 1681. Congressional findings and statement of purpose

(a) Accuracy and fairness of credit reporting

The Congress makes the following findings:

1. The banking system is dependent upon fair and accurate credit reporting. Inaccurate credit reports directly impair the efficiency of the banking system, and unfair credit reporting methods undermine the public confidence which is essential to the continued functioning of the banking system.

2. An elaborate mechanism has been developed for investigating and evaluating the credit worthiness, credit standing, credit capacity, character, and general reputation of consumers.

3. Consumer reporting agencies have assumed a vital role in assembling and evaluating consumer credit and other information on consumers.

4. There is a need to insure that consumer reporting agencies exercise their grave responsibilities with fairness, impartiality, and a respect for the consumer’s right to privacy.

Mandate by U.S. government on Fair and Accurate Credit reporting. Predictive models should not be discriminative (biased) toward any group.
Are all predictive models interpretable?

An interpretable model may or may not predict well

For example, we tried to benefit from an extensive set of attributes describing each of the movies in the dataset. Those attributes certainly carry a significant signal and can explain some of the user behavior. However, we concluded that they could not help at all for improving the accuracy of well tuned collaborative filtering models. Beyond selecting which features of the data to model, working with well designed models is also important ...

Solution to the Netflix Prize Bell et al., ‘08
PERFORMANCE VS. INTERPRETABILITY

Variable A

Credit card approved

Variable B

Credit card denied

Simple decision boundary

Complex decision boundary
HOW ABOUT A MORE DIFFICULT RELATIONSHIP?

Data

Learned decision boundaries
HOW DO WE SOLVE THIS PROBLEM?

● Problems:
  ○ Data scientists are choosing easy-to-interpret models like simple linear models or decision trees over high-performing neural networks or ensembles, effectively sacrificing accuracy for interpretability
  ○ Community is struggling to keep pace with new algorithms and frameworks (H20.ai, sklearn, R packages)

● Possible Solution: **What if** there was an interpretation library that…
  ○ Is model agnostic
  ○ Provides human-interpretable explanation
  ○ Is framework agnostic (scikit-learn, H20.ai, Vowpal Wabbit)
  ○ Is language agnostic (R, Python)
  ○ Allows one to interpret third-party models (Algorithmia, indico)
  ○ Supports analytical workflows during modeling process and post deployment
INTRODUCING SKATER
WHAT IS SKATER?

- Lack of good intuitive libraries to understand analytical results
- Help Human in the loop: A python library designed to demystify the inner workings of black-box models
- Uses a number of techniques for model interpretation to explain the relationships between input data and desired output, both globally and locally
- One can interpret models both before and after they are operationalized
### SKATER USES

- Model-agnostic variable importance for global interpretation
  - Helps in evaluating the importance of each independent input variable

<table>
<thead>
<tr>
<th>Feature</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>worst area</td>
<td></td>
</tr>
<tr>
<td>worst perimeter</td>
<td></td>
</tr>
<tr>
<td>area error</td>
<td></td>
</tr>
<tr>
<td>worst radius</td>
<td></td>
</tr>
<tr>
<td>worst texture</td>
<td></td>
</tr>
<tr>
<td>mean radius</td>
<td></td>
</tr>
<tr>
<td>mean perimeter</td>
<td></td>
</tr>
<tr>
<td>mean area</td>
<td></td>
</tr>
<tr>
<td>mean concavity</td>
<td></td>
</tr>
<tr>
<td>worst concave points</td>
<td></td>
</tr>
<tr>
<td>mean concave points</td>
<td></td>
</tr>
<tr>
<td>perimeter error</td>
<td></td>
</tr>
<tr>
<td>worst compactness</td>
<td></td>
</tr>
<tr>
<td>radius error</td>
<td></td>
</tr>
<tr>
<td>worst concavity</td>
<td></td>
</tr>
<tr>
<td>mean compactness</td>
<td></td>
</tr>
<tr>
<td>worst symmetry</td>
<td></td>
</tr>
<tr>
<td>texture error</td>
<td></td>
</tr>
<tr>
<td>worst smoothness</td>
<td></td>
</tr>
<tr>
<td>mean texture</td>
<td></td>
</tr>
<tr>
<td>worst fractal dimension</td>
<td></td>
</tr>
<tr>
<td>concavity error</td>
<td></td>
</tr>
<tr>
<td>compactness error</td>
<td></td>
</tr>
<tr>
<td>mean fractal dimension</td>
<td></td>
</tr>
<tr>
<td>mean smoothness</td>
<td></td>
</tr>
<tr>
<td>fractal dimension error</td>
<td></td>
</tr>
<tr>
<td>symmetry error</td>
<td></td>
</tr>
<tr>
<td>smoothness error</td>
<td></td>
</tr>
<tr>
<td>mean symmetry</td>
<td></td>
</tr>
<tr>
<td>concave points error</td>
<td></td>
</tr>
</tbody>
</table>

![Variable Importance Chart](chart.png)
SKATER USES

- Partial dependence plots for global interpretation
  - A visualization technique that can be used to understand and estimate the dependence of the joint interaction of the subset of input variables to the model’s response function

One-way interaction with variance

Two-way interaction
SKATER USES

- Improved local interpretable model-agnostic explanations (LIME) for local interpretation
  - A novel technique developed by Marco, Sameer and Carlos to explain the behavior of any classifier or regressor in an human interpretable and faithful manner using surrogate models

Text with highlighted words

I firmly believe that one of the major aspects of what makes House of Cards so good is the ability to watch all the episodes back-to-back with no commercials or programming schedules to get in the way
SKATER USES

- LIME for image interpretability (experimental)

highlight the feature boundaries

highlight the feature boundaries
**WITHOUT SKATER ...**

- Data → $f(Y|X)$ → Black box model → Evaluate

**WITH SKATER ...**

- Data → Unboxed model → Evaluate

Unboxed model:
- R or Python model (linear, nonlinear, ensemble, neural networks)
- Scikit-learn, caret and rpart packages for CRAN
- H20.ai, Algorithmia, etc.

How do I understand my models?
- Partial dependence plot
- Relative variable importance
- Local Interpretable Model Explanation (LIME)
- More coming ...
HOW DOES IT FIT INTO AN ANALYTICAL WORKFLOW?

Define Hypothesis
Use relevant key performance indicators

Handle Data
Handle Missing Data
Data Partitioning

Engineer and Select Features
Transform data
Select relevant features

Build Model
Build a predictive model

Deploy Model
Operationalize analytics as scalable REST APIs

Test and Monitor Model
1. Log and track behavior
2. Evaluate
3. Conduct A/B or multi-armed bandit testing

Model Interpretation: In-Memory Models
- Model assessment
- Explain model at a global and local level
- Publish insights, make collaborative and informed decisions

Model Interpretation: Deployed Models
- Explore and explain model behavior
- Debug and discover errors to improve performance

Retrain
Evaluate

Improve existing hypothesis/Generate a new one
HOW DO WE DO IT - datascience.com platform

Convenient resource selection

Choice of version control repository

Choice of kernel
INTERACTIVE DEPLOY

reproducibility - native git integration

convenient resource scaling

interactive user-friendly deploy
SKATER DEMO
A QUICK GLIMPSE INTO THE FUTURE

- Bayesian Rule Lists: An interpretable model, with series of decision statement

Learned interpretable model
Trained RuleListClassifier for detecting diabetes

================================================================================
IF Glucose concentration test : 159.5_to_inf THEN probability of diabetes: 16.7% (9.3%-25.6%)
ELSE IF Body mass index : -inf_to_27.3499995 THEN probability of diabetes: 93.2% (88.7%-96.7%)
ELSE IF Glucose concentration test : -inf_to_99.5 THEN probability of diabetes: 85.7% (78.2%-91.9%)
ELSE IF Age (years) : 30.5_to_inf THEN probability of diabetes: 40.1% (32.4%-48.1%)
ELSE IF Glucose concentration test : 99.5_to_130.5 THEN probability of diabetes: 80.6% (72.6%-87.4%)
ELSE probability of diabetes: 53.2% (39.0%-67.1%)
================================================================================

**Reference: Benjamin Letham**
A QUICK GLIMPSE INTO THE FUTURE

**Dining room**

- **Frequent object:**
  - wall: 0.99
  - chair: 0.98
  - floor: 0.98
  - table: 0.98
  - ceiling: 0.75
  - window: 0.73

- **Informative object:**
  - table: 0.96
  - chair: 0.85
  - chandelier: 0.80
  - plate: 0.73
  - vase: 0.69
  - flowers: 0.63

**Bathroom**

- **Frequent object:**
  - wall: 1
  - floor: 0.85
  - sink: 0.77
  - faucet: 0.74
  - mirror: 0.62
  - bathtub: 0.56

- **Informative object:**
  - sink: 0.84
  - faucet: 0.80
  - countertop: 0.80
  - toilet: 0.72
  - bathtub: 0.70
  - towel: 0.54

**Reference:** B. Zho et al.
INTERPRETATION ROADMAP

- More improvements to Deployed Model - H20, VW, Spark-MLLib

- Possible Future Work
  - Individual Conditional Expectation
  - Local Interpretation: e.g. Anchors for Local Interpretation
  - Global Interpretation: e.g. Probabilistic Rule based Models
  - Better support for Image Interpretability
    - extension of LIME
    - Class Activation Maps
  - Better ways to detect Interaction effect among model variables

- Help wanted we can’t do it alone: https://goo.gl/W17q4i
SUMMARIZE VALUE OF datascience.com platform, jupyter, Skater

- Reproducible experiments with native integration with version control
- Convenient resource selection
- Scalable Jupyter Notebook with support for multiple kernels - python2, python3, R, and spark kernels (pyspark, spark) to name a few
- datascience.com provides the ability to work with the Machine Learning library of choice
- Interactive Jupyter environment for data exploration and building model
- Model Interpretation with Skater for In-Memory model
- datascience.com interactive deploy to operationalize the model
- Model Interpretation with Skater for Deployed Model
- Interactive Jupyter.widgets to enable human in the loop interaction for both In-Memory and Deployed model
- datascience.com deploy to maintain multiple model versions for further model testing and improvement
Thanks

- Special thanks to Aaron Kramer(*one of the original authors of Skater*), Ben Van Dyke and rest of the datascience.com teammates for helping out with Skater

- Thank you to [jupytercon](https://www.jupytercon.com) for providing us the opportunity to share our thoughts with a wider community