Improving the explainability of Random Forest classifier – user centered approach

Dragutin Petkovic† 1, 3, Russ Altman2, Mike Wong3, Arthur Vigil4

1Computer Science Department, San Francisco State University (SFSU)
2Department of Bioengineering, Stanford University
3SFSU Center for Computing for Life Sciences, 1600 Holloway Ave., San Francisco, CA 94132
4Twist Bioscience, 455 Mission Bay Boulevard South, San Francisco, CA 94158
Mary is deciding whether to adopt a ML-based diagnostic method
Mary has to make a decision based on *Current* state-of-the-art of presenting ML data

- *ML algorithm used*
- *Quality and details of the Training DB*
- *Information about specific SW used*
- *Accuracy and methods used to estimate it*

Mary’s decision is critical for patients’ well being and for the company.
TO TRUST OR NOT TO TRUST?
What could have happened? *(Besides SW bugs and errors)*

• MLDA could have performed *correctly* based on blindly following training data

**BUT**

• decision might be *fundamentally wrong* - some examples can be found in:
  
  
  – “Can AI be Taught to Explain Itself”, NY Times Magazine Nov 2017
Seminar Outline

• Brief summary of Workshop on “Machine learning and deep analytics for biocomputing: call for better explainability” held at PSB 2018 January 2018 (joint work with Prof. L. Kobzik and Prof. C. Re)

• Improving RF Explainability (RFEX) and case study using Stanford Feature data (joint work with Prof. R. Altman, M. Wong and A. Vigil)
What is ML Explainability?

• Easy to use information explaining why and how the ML approach made its decisions
  – Model Explainability: helps explain the ML model as a whole
  – Sample explainability: helps explain decision on specific sample (often user confidence is guided by ML accuracy on specific samples they know about)

• Targeted to both ML experts and non-experts
Demand for better MLDA explainability is growing including mainstream media and public.

NY Times Magazine
Nov 21 2017

Economist Feb 2018
And will likely be regulated

• New EU General Data Protection laws
  – https://www.eugdpr.org/

• New IEEE Standard 70001 on Transparency of Autonomous Systems
  – https://standards.ieee.org/develop/project/7001.html
BBC Y factor – excellent podcast on machines and morals

Rethinking about Safety of autonomous cars

Uber Won’t Seek Calif. Permit Renewal To Test Self-Driving Vehicles After Fatal Crash
March 27, 2018 2:49 PM ET

Updated at 9:00 a.m. ET Wednesday

The future of Uber’s self-driving car program is likely non-existent in California, at least for now.

The company announced it will not renew Uber’s permit through the state’s Department of Motor Vehicles to continue testing a fleet of autonomous-driving cars on California roads, following last week’s deadly crash in Tempe, Ariz.

“We decided to not reapply for a California DMV permit with the understanding that our self-driving vehicles would not operate on public roads in the immediate future,”
And of course...what about autonomous military machines....

Pathways to Banning Fully Autonomous Weapons

October 23rd, 2017

On 16 October 2017, the Permanent Mission to the United Nations of Mexico partnered with the International Committee for Robot Arms Control, Human Rights Watch, Seguridad Humana en Latinoamérica y el Caribe and the Campaign to Stop Killer Robots to host a panel discussion entitled "Pathways to Banning Fully Autonomous Weapons" as part of the First Committee side event series for the 72nd Session General Assembly. Ambassador Juan Sandoval Mendóza, Deputy Permanent Representative of Mexico to the United Nations, introduced the discussion by noting that Mexico has joined other states in calling for a ban on lethal autonomous weapons systems, also known as fully autonomous weapons. Referencing possible implications of the 2030 Agenda for Sustainable Development, Ambassador Mendóza stressed the importance of continued research and education on these weapons systems and their possible dangers.

Ms. Marly Wareham, global coordinator of the Campaign to Stop Killer Robots and moderator of the panel, described autonomous weapons as systems that select targets and use force without meaningful human control. She said the side event would consider the role for prohibiting lethal autonomous weapons systems, as well as pathways to concluding a new international treaty that would ban them. Professor Noel Sharkey, from the International Committee for Robot Arms Control, continued the discussion by focusing on the technologies in such systems. He mentioned the difficulty of formulating a precise definition of an autonomous weapon, and he noted that the United States Department of Defense references a need for "appropriate levels of human judgment" over these systems. Professor Sharkey urged the international community to further discuss controls on fully autonomous weapons. Countries view these weapons...
Benefits of better ML Explainability

• Increased **confidence and trust** of application and domain experts as well as public in adopting ML;
• Better **validation, audit** and prevention of cases where ML approach produces results based on fundamentally wrong reasons or can behave in **unsafe** manner
• Simplification and **reduction of the cost** of application of ML in practice (e.g. by knowing which smaller feature subsets produce adequate accuracy)
• **Improved “maintenance”** where ML method has to be changed or tuned to new data or decision needs;
• Possible **discovery of new knowledge and ideas** (e.g. by discovering new patterns and factors that contribute to ML decisions)
Goals

- Discuss *challenges in explainability* of current Machine Learning and Deep Analytics (MLDA) used in biocomputing
- Start the discussion on *concrete ways to improve it*

Workshop involved both MLDA researchers and users/adopters


*Support of Stanford Mobilize Project is gratefully acknowledged*
Interest was high
Workshop organization

Panel 1: What are the needs and problems
View of “Users “
Moderator: Prof. Les Kobzik
• Dr. R. Ghanadan - Google (since September 2017, previously at DARPA Explainable AI)
• Dr. W. Kibbe - Chief for Translational Biomedical Informatics in the Department of Biostatistics and Bioinformatics and chief data officer for the Duke Cancer Institute, Professor, Duke University since August 2017
• Dr. B. Percha - Assistant Professor, Icahn School of Medicine at Mount Sinai; Head of R&D, Health Data and Design Innovation Center (HD2i) Institute for Next-Generation Healthcare

Panel 2: Some current examples and possible solutions – View of “Developers”
Moderator: Prof. Christopher Re
• Dr. R. Roettger - Assistant Professor, University of Southern Denmark, Odense
• Dr. R. Scheuermann - Dir. Of Bioinformatics, J. Craig Venter Institute
• A. Esteva - Ph. D. Candidate, Stanford University

Panel 3: discussion with panelists and audience
Highlights of some panelists’ talks
From Dr. R. Ganahan: Explainable Models will accelerate the development and impact of ML/AI systems
Data and startups: some observations

- Nonlinear relationship between accuracy and usefulness
- Big/medium-sized/small/tiny data
- Often consistency is more important than correctness
- Choose the most explainable model (sales, internal knowledge, etc.)
- Ongoing technical advice/education to teams are vital

From: B. Percha
How to improve ML explainability?

Data

1. Understand your data

From A. Esteva

1. Understand the algorithm’s response to the data
Any model may pick out *unintended signal*. Deep models may pick out *more* unintended signal.

**Upshot:**
Picked up on *mascara*

From: C. Re

Kuehlkamp et al. *Gender-from-Iris or Gender*
Toward Explainable Machine Learning - RFEX: Improving Random Forest Explainability

Prof. D. Petkovic
SFSU

Random Forest ML

- Widely used
- Excellent performance
- Abundance of SW tools available
- Based on ensemble of trees
- Amendable to explanation and offers feature importance ranking
Current approaches in RF Explainability

• **Feature ranking** uses RF-provided variable importance measures like e.g. RF-provided Gini, MDA (mean decrease in accuracy) or others, to present them in *tables or horizontal bar charts* sorted by chosen variable importance measure.
  – Too simplistic, not done for + vs. – class separately. Lack of tradeoffs between features used and accuracy.

• **Rule extraction from trained RF.** This method consists of:
  a) performing standard RF training;
  b) defining rules by analyzing trained RF trees (resulting in very large set of rules, order of 100 K); and
  c) reducing the number and complexity of extracted rules by optimization to reduce to 10s – 100s of rules, each with 1-10 or so conditions.
  – Hard to interpret by humans; rules often complex; lack of tradeoffs between accuracy and number of rules used.

• **No “user design and evaluation”** with key adopters who are often non-RF expert users – the key constituency
Traditional RF classification

Training Data

Random Forest Training

Accuracy:
F1, OOB, confusion Matrices, ntree, mtry...

Trained RF tree Ensemble

RFEX: RF Explainability enhancement

RFEX: Creation of Explainable RF Model

RFEX summary Report (one page)

RFEX Explainability Model data

User needs
RFEX explainable model – driven by User/Adopter Needs/Questions

• What is the loss/tradeoffs of accuracy if I use only certain subset of most important features?

• What are most important features contributing to ML prediction and how do they rank in importance?

• Also, tell me more about features:
  – What is the relationship of most important features for + vs. – class, is there any overlap?
  – What is “direction” of features? Abundance (“more of it” or “presence”) or deficiency (“less of it” or “absence”)? What thresholds I can use to determine this? What are basic class specific feature stats?
  – Which features interact together?

• Can explainable ML model be presented in an easy to understand and simple summary for ML/domain experts and non-experts?

• Finally: evaluate if the RFEX explainable model is helpful and intuitive to domain experts?
The \texttt{featurize} program takes points in a biomolecule to build microenvironments (Figure A) and produce computational representations of said microenvironments (Figure B).

The \texttt{featurize} program analyzes mutually exclusive concentric spherical volumes (called shells) around a given point in a biomolecule. These shells collectively describe a microenvironment. The \texttt{featurize} program tallies physicochemical properties for each atom contained in each shell. These tallies form the computational representation describing the microenvironment. Groups of microenvironment computational representations are called Feature Vectors and are stored in Feature Vector Files.

Machine Learning can be trained on Feature Vectors to produce biomolecule functional class models. Biomolecules of unknown function can be characterized as Feature Vectors and scored against functional class models to predict functionality.

From AWS case study of Stanford-SFSU collaboration
Note *unbalanced* training data (many fewer positive samples)

Our previous work achieved good RF prediction but *we were not sure why!*

RF feature ranking and accuracy

• **Feature Ranking:** We use **MDA** – Mean Decrease in Accuracy (part of RF alg.) - provided by all RF implementations
  – for each feature in dataset:
    - randomly permute feature;
    - make predictions on this permuted data;
    - record average decrease in accuracy vs. using unpermuted data;
  – Permuting more important features result in larger decrease in accuracy

  (Permutation base ranking is more robust and less biased (in R tool 4 and later) )

• RF consists of ensemble of disjoint trees, so best features used for + class may not be the same as those for – class → MDA can be computed for + and – class *separately* (**MDA+; MDA-**) – important in case of *highly unbalanced data* (FEATURE data is unbalanced)

• **RF Accuracy:** we use **F1 score** for + class
  – F1 = 2* (precision*recall)/(precision + recall)
  – Precision/recall optimized by varying *cutoff* for ensemble tree voting
New RFEX measures to explain how features are used by RF: *Feature Direction* and *Mutual Feature Interaction*

- **Feature Direction** - $\text{DIR}(I) + (n)$ or $– (n)$: denoting fraction of times ($n$) when feature $I$ was above (+) *(abundance)* or below (-) *(deficiency)* the threshold when making correct prediction, for all trees in the forest making a correct prediction, and for all test samples.

- **Mutual Feature Interaction** $\text{MFI}(I,J)$ for features $I$ and $J$ - count of times features $I$ and $J$ appear on the same tree path making a correct prediction, for all trees in RF ensemble, and for all test samples.
  - Note that MFI only measures statistical pair-wise feature co-occurrences and not necessarily causality.
### Ranking of top 20 features with MDA+ and MDA for ASP_PROTEASE.4.ASP.OD1

<table>
<thead>
<tr>
<th>Top Features by +MDA</th>
<th>+/-</th>
<th>Top Features by -MDA</th>
<th>+/-</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEG_CHARGE_s2</td>
<td>+ (0.91)</td>
<td>RESIDUE_NAME_IS_GLY_s2</td>
<td>- (0.99)</td>
</tr>
<tr>
<td>RESIDUE_CLASS1_ISUNKNOWN_s2</td>
<td>+ (0.84)</td>
<td>RESIDUE_CLASS1_ISUNKNOWN_s2</td>
<td>- (0.99)</td>
</tr>
<tr>
<td>RESIDUE_NAME_IS_GLY_s2</td>
<td>+ (0.82)</td>
<td>RESIDUE_CLASS2_IS_POLAR_s2</td>
<td>- (0.93)</td>
</tr>
<tr>
<td>SECONDARY_STRUCTURE1_IS_STRAND_s5</td>
<td>+ (0.96)</td>
<td>RESIDUE_NAME_IS_LEU_s5</td>
<td>- (0.96)</td>
</tr>
<tr>
<td>RESIDUE_NAME_IS_GLY_s3</td>
<td>+ (0.88)</td>
<td>SECONDARY_STRUCTURE1_IS_STRAND_s5</td>
<td>- (0.88)</td>
</tr>
<tr>
<td>RESIDUE_CLASS1_ISUNKNOWN_s3</td>
<td>+ (0.89)</td>
<td>PEPTIDE_s2</td>
<td>- (0.85)</td>
</tr>
<tr>
<td>SOLVENT_ACCESSIBILITY_s5</td>
<td>- (0.93)</td>
<td>SOLVENT_ACCESSIBILITY_s1</td>
<td>+ (0.83)</td>
</tr>
<tr>
<td>SOLVENT_ACCESSIBILITY_s4</td>
<td>- (0.82)</td>
<td>RESIDUE_NAME_IS_GLY_s3</td>
<td>- (0.96)</td>
</tr>
<tr>
<td>RESIDUE_NAME_IS_THR_s4</td>
<td>+ (0.86)</td>
<td>ATOM_TYPE_IS_O2_s2</td>
<td>- (0.95)</td>
</tr>
<tr>
<td>ATOM_TYPE_IS_O2_s2</td>
<td>+ (0.86)</td>
<td>NEG_CHARGE_s2</td>
<td>- (0.95)</td>
</tr>
<tr>
<td>SECONDARY_STRUCTURE1_IS_TURN_s3</td>
<td>+ (0.90)</td>
<td>RESIDUE_CLASS1_ISUNKNOWN_s3</td>
<td>- (0.96)</td>
</tr>
<tr>
<td>RESIDUE_CLASS2_IS_BASIC_s4</td>
<td>- (0.99)</td>
<td>MOBILITY_s5</td>
<td>+ (0.92)</td>
</tr>
<tr>
<td>CHARGE_WITH_HIS_s2</td>
<td>+ (0.95)</td>
<td>SOLVENT_ACCESSIBILITY_s4</td>
<td>+ (0.92)</td>
</tr>
<tr>
<td>CHARGE_s2</td>
<td>+ (0.93)</td>
<td>SECONDARY_STRUCTURE1_IS_TURN_s3</td>
<td>- (0.89)</td>
</tr>
<tr>
<td>NEG_CHARGE_s3</td>
<td>+ (0.88)</td>
<td>RESIDUE_NAME_IS_THR_s4</td>
<td>- (0.94)</td>
</tr>
<tr>
<td>RESIDUE_NAME_IS_THR_s3</td>
<td>+ (0.77)</td>
<td>RESIDUE_CLASS2_IS_POLAR_s3</td>
<td>- (0.95)</td>
</tr>
<tr>
<td>SECONDARY_STRUCTURE1_IS_TURN_s2</td>
<td>+ (0.84)</td>
<td>SOLVENT_ACCESSIBILITY_s5</td>
<td>+ (0.92)</td>
</tr>
<tr>
<td>RESIDUE_CLASS2_IS_POLAR_s3</td>
<td>+ (0.83)</td>
<td>RESIDUE_CLASS1_IS_HYDROPHOBIC_s5</td>
<td>- (0.82)</td>
</tr>
<tr>
<td>SECONDARY_STRUCTURE1_IS_STRAND_s4</td>
<td>+ (0.94)</td>
<td>NEG_CHARGE_s3</td>
<td>- (0.86)</td>
</tr>
<tr>
<td>RESIDUE_NAME_IS_ASP_s3</td>
<td>+ (0.93)</td>
<td>ELEMENT_IS_ANY_s4</td>
<td>+ (0.50)</td>
</tr>
</tbody>
</table>

- Some top features overlap in + vs. – classification
- All directions very consistent (high %)
- If features overlap their direction is opposite
Trade-offs in using subset of top ranked features vs. accuracy

Model Performance using a subset of features

We varied the number of features used to train our RF model from 2 to 20 and plotted the f-score for each trained model to show how model performance varied as we increased the number of features used.

Observations:
- RF classifiers trained on just a few (between 10 and 20) features performed very close to RF using all 480 features
- Some of our models perform well even with just 2 or 3 features. Others showed a steeper drop off
- Standard deviation over CV trials is small

These charts reveal that one needs very few (from 2-6 depending on a model) top ranked features to achieve 90% or better of the accuracy when all 480 features are used.
RFEX One Page Summary report

**Trade-off in accuracy using only top N (e.g. 4) best features**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature</th>
<th>Direction</th>
<th>f-score</th>
<th>Importance</th>
<th>MFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SECONDARY_STRUCTURE1 IS STRAND</td>
<td>+</td>
<td></td>
<td></td>
<td>(5; 3; 7)</td>
</tr>
<tr>
<td>2</td>
<td>NEG CHARGE</td>
<td>+</td>
<td>0.8916</td>
<td></td>
<td>(3; 7; 5)</td>
</tr>
<tr>
<td>3</td>
<td>RESIDUE NAME IS GLY</td>
<td>+</td>
<td>0.8919</td>
<td></td>
<td>(5; 7; 6)</td>
</tr>
<tr>
<td>4</td>
<td>RESIDUE_CLASS1 IS UNKNOWN</td>
<td>+</td>
<td>0.9133</td>
<td></td>
<td>(7; 5; 1)</td>
</tr>
<tr>
<td>5</td>
<td>RESIDUE NAME IS THR</td>
<td>-</td>
<td>0.9525</td>
<td></td>
<td>(3; 1; 7)</td>
</tr>
<tr>
<td>6</td>
<td>SOLVENT ACCESSIBILITY</td>
<td>+</td>
<td>0.9531</td>
<td></td>
<td>(3; 1; 4)</td>
</tr>
<tr>
<td>7</td>
<td>SOLVENT_ACCESSIBILITY</td>
<td>+</td>
<td>0.9781</td>
<td></td>
<td>(3; 4; 5)</td>
</tr>
<tr>
<td>8</td>
<td>RESIDUE NAME IS GLY</td>
<td>+</td>
<td>0.9806</td>
<td></td>
<td>(5; 1; 6)</td>
</tr>
<tr>
<td>9</td>
<td>RESIDUE_CLASS2 IS BASIC</td>
<td>+</td>
<td>0.9844</td>
<td></td>
<td>(5; 3; 7)</td>
</tr>
<tr>
<td>10</td>
<td>RESIDUE_CLASS1 IS UNKNOWN</td>
<td>-</td>
<td>0.9841</td>
<td></td>
<td>(5; 1; 4)</td>
</tr>
<tr>
<td>11</td>
<td>CHARGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>SECONDARY_STRUCTURE1 IS TURN</td>
<td>+</td>
<td>0.9901</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>RESIDUE NAME IS ASP</td>
<td>+</td>
<td>0.9905</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>SECONDARY_STRUCTURE1 IS STRAND</td>
<td>+</td>
<td>0.9905</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>CHARGE WITH HIS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>RESIDUE_CLASS2 IS POLAR</td>
<td>+</td>
<td>0.9914</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>RESIDUE_NAME IS THR</td>
<td>+</td>
<td>0.9914</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>RESIDUE_NAME IS THR</td>
<td>+</td>
<td>0.9914</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>ATOM TYPE IS O2</td>
<td>+</td>
<td>0.9917</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>NEG CHARGE</td>
<td>+</td>
<td>0.9933</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SECONDARY_STRUCTURE1 IS TURN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Base RF accuracy using ALL features**

RF trained on 48577 negatives, 1585 positives;

Performance f-score with all 480 features=0.9952

*Classic way Of presenting RF accuracy*

**Bold underlined Feature also in top 20 for - class**

**Ranking of most important features with MDA+**
<table>
<thead>
<tr>
<th>rank</th>
<th>feature</th>
<th>direction</th>
<th>f-score</th>
<th>importance</th>
<th>MFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RESIDUE_NAME IS ASP_s3</td>
<td>+</td>
<td></td>
<td></td>
<td>(6,7,8)</td>
</tr>
<tr>
<td>2</td>
<td>SECONDARY_STRUCTURE1 IS_4-HELIX_s4</td>
<td>+</td>
<td>0.6288</td>
<td></td>
<td>(6,7,8)</td>
</tr>
<tr>
<td>3</td>
<td>SECONDARY_STRUCTURE1 IS_4-HELIX_s5</td>
<td>+</td>
<td>0.6933</td>
<td></td>
<td>(6,7,1)</td>
</tr>
<tr>
<td>4</td>
<td>SECONDARY_STRUCTURE1 IS_BEND_s3</td>
<td>+</td>
<td>0.8330</td>
<td></td>
<td>(6,7,8)</td>
</tr>
<tr>
<td>5</td>
<td>SECONDARY_STRUCTURE2 IS_BETA_s3</td>
<td>+</td>
<td>0.8804</td>
<td></td>
<td>(8,7,6)</td>
</tr>
<tr>
<td>6</td>
<td>SOLVENT_ACCESSIBILITY_s4</td>
<td>+</td>
<td>0.9002</td>
<td></td>
<td>(7,1,4)</td>
</tr>
<tr>
<td>7</td>
<td>PEPTIDE_s3</td>
<td>+</td>
<td>0.9310</td>
<td></td>
<td>(6,1,8)</td>
</tr>
<tr>
<td>8</td>
<td>SOLVENT_ACCESSIBILITY_s3</td>
<td>+</td>
<td>0.9305</td>
<td></td>
<td>(7,6,1)</td>
</tr>
<tr>
<td>9</td>
<td>RESIDUE_CLASS2 IS_ACIDIC_s3</td>
<td>+</td>
<td>0.9325</td>
<td></td>
<td>(6,7,4)</td>
</tr>
<tr>
<td>10</td>
<td>RESIDUE_NAME IS ILE_s4</td>
<td>+</td>
<td>0.9382</td>
<td></td>
<td>(6,1,7)</td>
</tr>
<tr>
<td>11</td>
<td>CARBONYL_s2</td>
<td>+</td>
<td>0.9439</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>SECONDARY_STRUCTURE2 IS_HELIX_s5</td>
<td>+</td>
<td>0.9478</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>ELEMENT_IS_ANY_s3</td>
<td>+</td>
<td>0.9486</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>SECONDARY_STRUCTURE2 IS_HELIX_s4</td>
<td>+</td>
<td>0.9465</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>CARBONYL_s4</td>
<td>-</td>
<td>0.9475</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>RESIDUE_NAME IS GLY_s3</td>
<td>-</td>
<td>0.9453</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>RESIDUE_NAME IS GLY_s2</td>
<td>+</td>
<td>0.9461</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>NEG_CHARGE_s2</td>
<td>-</td>
<td>0.9507</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>ATOM_TYPE_IS_C_s3</td>
<td>+</td>
<td>0.9512</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>RESIDUE_CLASS1 IS UNKNOWN_s2</td>
<td>+</td>
<td>0.9521</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RF trained on 48145 negatives, 1811 positives
Performance f-score with all 480 features=0.95439
RFEX pipeline summary – *general* steps in providing more explainable RF

1. Establish **Base RF Accuracy** using all features (use F1 score)
2. **Rank features/variables** (e.g. use MDA and do it *separately* for + and – class if data is unbalanced)
3. Provide **tradeoffs** between features used and accuracy (e.g. what accuracy we can get using only top K ranked features)

Then work only with Top N features **(N usually 2-5% of total number of features for 90% of original accuracy!!!)**

4. Explain how **features are used** by RF
   - Determine class-specific feature stats: e.g. **feature direction** namely its **abundance** (more of it) or **deficiency** (less of it) or some other feature statistics (AV/SD/RANGE)
   - Determine which features **interact** with each other (MFI, correlation)
5. Create **easy to use RFEX** data and report (**one page**)
RFEX Usability review – anonymous survey of 13 expert and non-expert users
Measure how RFEX increases *user confidence* vs. using only traditional RF results

<table>
<thead>
<tr>
<th>Question</th>
<th>ALL users (13)</th>
<th>FEATURE and RF NON-experts (4)</th>
<th>FEATURE experts and RF NON-experts (3)</th>
<th>FEATURE NON-experts and RF experts (2)</th>
<th>FEATURE and RF experts (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate your <em>increase in confidence</em> of FEATURE data after using RFEX summaries</td>
<td>2.7 (SD 2.2)</td>
<td>2.5</td>
<td>2</td>
<td>0.5</td>
<td>4.5</td>
</tr>
<tr>
<td>Estimate your <em>increase in understanding why and how RF works on FEATURE data</em> (e.g. RF Explainability) using RFEX one page summaries</td>
<td>3.3 (SD 1.7)</td>
<td>3.25</td>
<td>3.7</td>
<td>1</td>
<td>4.25</td>
</tr>
<tr>
<td>I believe RFEX approach will be useful for other applications of RF</td>
<td>4.4 (SD 0.5)</td>
<td>4.5</td>
<td>4.3</td>
<td>4</td>
<td>4.5</td>
</tr>
<tr>
<td>I believe RFEX approach (or its modifications) will be useful for other machine learning methods</td>
<td>4.0 (SD 0.8)</td>
<td>4.5</td>
<td>3.3</td>
<td>3.5</td>
<td>4.25</td>
</tr>
</tbody>
</table>

(1 low...5 high)
Future work

• Try RFEX on other RF applications
• Work on RFEX sample explainability
• Develop RFEX toolkit

• From repeatability to explainability to ethical, moral and safe AI

Acknowledgements

We thank Prof. L. Kobzik, Dr. L. Buturovic and Prof. K. Okada for valuable feedback; T. Murray and J. Schwartz for help in organizing RFEX usability study; We also thank Prof. R. Altman and Dr. J. Ku for encouragement to pursue this work.

The work has partially been supported by NIH grant R01 LM005652 and by Stanford Mobilize Project and SFSU Center for Computing for Life Sciences.

We also thank 13 anonymous usability reviewers.
Thank You