4/30/19

**Jane Adams (Seeing the Future of AI)**

* Sentiment of measuring happiness (Twitter); hedonmeter.org
* SciPy : ‘hieratical clustering’
* T-test vs. KS Correlation/Test
* Proejcts: USA Sensus Data (How to determine a place is livable)
* Commonly Data Viz Tools: Plotly; Processing.org (Virtual Reality); CARTO (Open Source Mapping)

**Adam McElhinney (Chief of ML & AI Strategy at Uptake)**

* **Uptake** (Chicago, IL): AI Application to detect/measure/diagnose the health condition of the machines (eg. cars). Failure Prediction
* Data Collection:
  + Place sensor on the machines to collect data
  + Manual Plug in Machines;
  + Satellite;
  + CellPhone;
  + WiFi\* to collect data
* Challenge:
  + Connectivity (Cellular connectivity): most rural areas have worst connectivity, but also more likely to have vehicles broken
  + Data transmitted only when machines are ON (but when machine breaks, it is OFF)
  + Failure/Repair Data are recorded inconsistently (data entry failure because mechanics are entering the data inaccurately)
  + Root cause of failure difficult to ascertain
  + Large quantity of signal data (storage challenge) from machine
  + Highly imbalanced data – less positive samples
  + Count for machine’s normal aging vs. failure (how to distinguish and define the two category?)
  + Seasonality of data (machine act differently under different seasons)
* Solutions:
  + Physics-based models
  + ML Algo
    - Knowledge Based Models
    - Life Expectancy
    - ANN
    - Transfer Learning

**Ben Vigoda (CEO of Gamalon)**

* <https://gamalon.com/>
* Worth request a demo: the company created a NLP application to better understand customer sentiment (Sales/Marketing, Consumer Goods etc.)
* Ben Vigoda (CEO) – giving the talk in ODSC
  + <https://blog.gamalon.com/news/author/ben-vigoda>
  + Unsupervised Deep Learning Models
  + Order of words & order of phrases & Rewording matters (not only counting or statistics of the words)

5/1/2019

**Rajiv Shah – Deciphering the Black Box: Latest Tools and Techniques for Interpretability**

* Permutation feature importance (Use “R”)
  + Same model but with different seeds (to see how feature importance shifts – to add variance in the model to begin with)
  + Run feature importance in different models (which features are constantly on top vs. ones not)
    - Eliminate features which has its importance variance
    - Build random features in dataset (to understand the level of randomness of the model)
  + Papers: Strobl/Lundberg
  + Explain AI Blog: <https://explained.ai/>
* Random Component Analysis
* Partial Dependence
  + **Isolate the features**, and change the values of the age to see how that impact the performance
  + Partial dependence: averaging the ‘coefficient’
  + ICE – Individual Conditional Expectation
  + Papers: Friedman on PDP; Goldstein on ICE Plots
* LIME and its shortcomes (Don’t use LIME)
  + Can apply to all models – modeling agnostic
  + LIME is making an assumption that all the features are normally distributed.
    - Use LIME to create a model
    - Then generate cloud of data points (data neighborhood) around the feature (normally distributed?)
  + **Local interpretation**
  + If a feature is always the #1 in every feature combination, how useful it is in predicting the model?
  + LIME’s – using different hyperparameter, might have different results in explaining the models
  + Slow
* SHAP
  + Shapley Value: When there are different features cooperate, which are doing the magic (eg. winter time pushing car out of snow, which ppl are helping)

**Brian Lucena – Explaining XGBoost Models**

Log-loss: estimate how reliability of probability

Improve log loss -> improve AUC

But improve AUC, not necessarily improve logloss – might include noise

* Partial Dependency (Average of ICE) vs. Individual Conditional Expectation (ICE Plot)
* Shapley Value: Bring in the feature in the model at different order (along with other features), see each step what the change (bump or dip)

**Sourav Dey, Alex Ng - Reproducing Data Science Using Orbyter**

Docker is your friend.

Orbyter is a framework and toolsets for helping ML teams move to a container-first workflow

Cookiecutter in for Data Science (by Peter Bull) Python (make it easy to onboard)

Klematic (after downloading Docker, it comes with the visualization of docker)

Containers are like apartments and VMs are like houses

12factors.net

Flexibility, Observability, Reproducibility

Setup\_logging() is better than print()

‘mlflow’ package (OpenSource by DataBricks)

<https://medium.com/manifold-ai/torus-a-toolkit-for-docker-first-data-science-bddcb4c97b52>

**Stephen Pushkarev - Automate ML Models Using Kubeflow**

[**http://odsc.k8s.hydrosphere.io/**](http://odsc.k8s.hydrosphere.io/)

[**https://github.com/Hydrospheredata/odsc-workshop**](https://github.com/Hydrospheredata/odsc-workshop)

Kubeflow vs. Mlflow (Databricks) vs. Airflow

* Research
* Data Preparation
* Model Training
* Model Cataloguing
* Model Deployment
* Model Integration Testing
* Production Inferencing
* Modeling Performance Monitoring
* Model Maintenance

Pipeline, experiments and runs

Kubeflow: Low level API; Projects in MLflow; Have ML pipelines

MLflow: Spark level API; Projects; Do not care about the containers; hard to scale in future

Airflow: only pipeline

5/2/2019

**Regina Barzilay (PhD): How AI Changes the HealthCare**

* Add classical risk factors
* Hathoway Phenomenon
* DNN Suffers/Limitations:
  + DNN powerful is you train the RIGHT data
  + DNN needs to more data to be more robust
  + Not robust towards distribution shift
  + DNN learn data bias (interpretable neural models; monkey vs. monkey with guitar - human)

**Maureen McElaney – IBM Developer Advocate: Trust and transparency in ML lifecycle**

* Bias in the model today – eg. Prison defendants scoring / tend to be racist
* Data 4 Black Lives – trying to use data for social justice
* **Joy Boulamwini**
* **Anil Dash** on the biases of tech (prodcast)
* Usefull tools and potential projects ideas:
  + <https://github.com/IBM/AIF360>
  + <https://developer.ibm.com/code/open/centers/codait/projects/>

**Hilary Mason – Found and CEO of Cloudera: The Future of AI and ML Products**

<https://blog.fastforwardlabs.com/>

<https://www.ibm.com/blogs/industries/little-known-story-first-iot-device/>

**strategic framework** – sometimes it is not just two dimentional (value vs. cost of working on this projects)

* Sometimes it is worth spending time and effort to try a product/direction even it does not work
* It could become an investment, because your previous trial could save lots of time to eg. get the data ready/learn from the process – same time in a long run
* ‘Agile Sounds Good’ , but does not work
* Product design is crucial – it is not only about the model’s accuracy, but also the usage of the product ultimately
* Do we have the right technology in place?