ACCT 420: Ensembling and Ethics

Dr. Richard M. Crowley

rcrowley@smu.edu.sg

https://rmc.link/

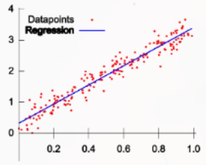
Front Matter

Learning objectives

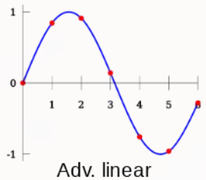
Foundations

Review

Forecasting

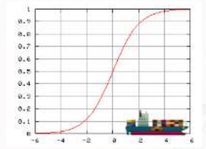


Linear regression



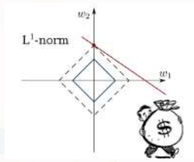
Adv. linear regression

Binary classification



Logistic regression for contracting

Leveraging research for bankruptcy



Lasso regression for fraud

Advanced methods

Natural Language

Anomaly detection

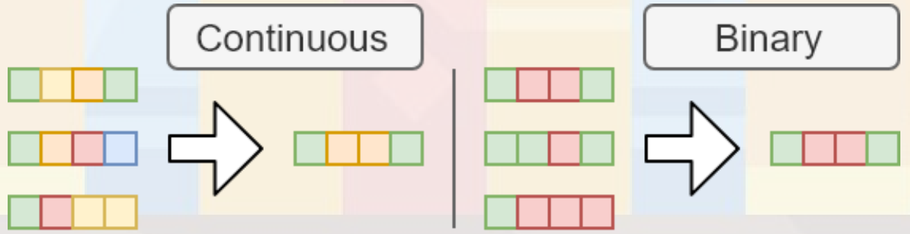
AI/ML

- **Theory:**
  - Ensembling
  - Ethics
- **Application:**
  - Fraud detection
  - Fairness
  - Data security
- **Methodology:**
  - Ensembling
  - SHAP
  - Critical thinking


Ensembles

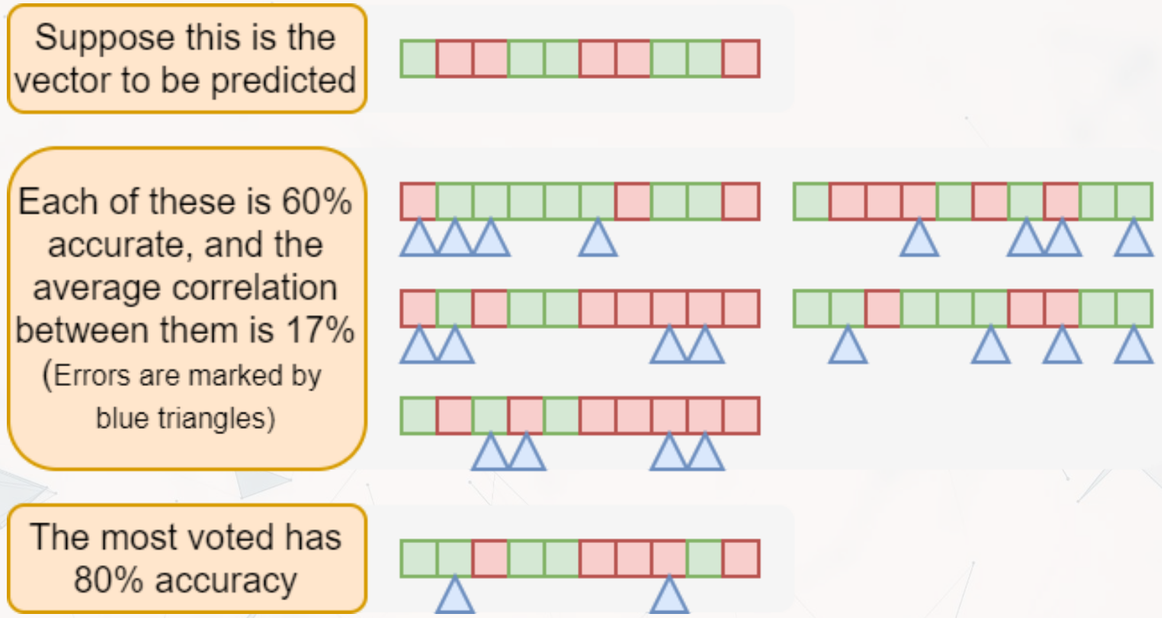
What are ensembles?

- Ensembles are *models made out of models*
- Ex.: You train 3 models using different techniques, and each seems to work well in certain cases and poorly in others
  - If you use the models in isolation, any of them would do an OK (but not great) job
  - If you make a model using all three, you can get better performance if their strengths all shine through
- Ensembles range from simple to complex
  - Simple: a (weighted) average of a few model's predictions



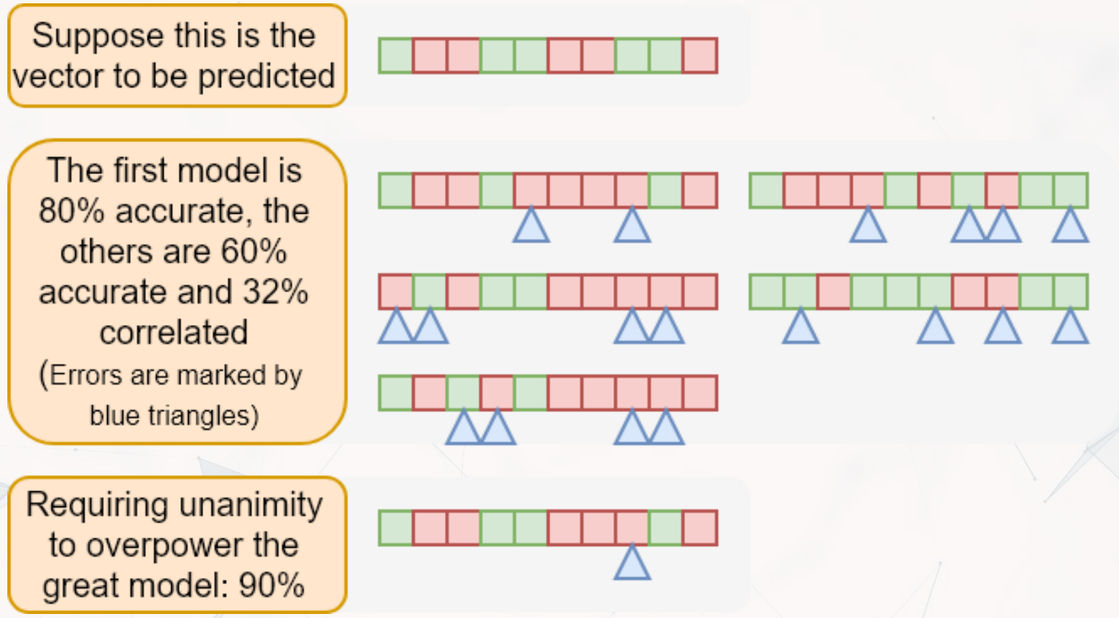
When are ensembles useful?

1. You have multiple models that are all decent, but none are great
   - And, ideally, the models' predictions are not highly correlated



When are ensembles useful?

2. You have a really good model and a bunch of mediocre models
    - And, ideally the mediocre models are not highly correlated

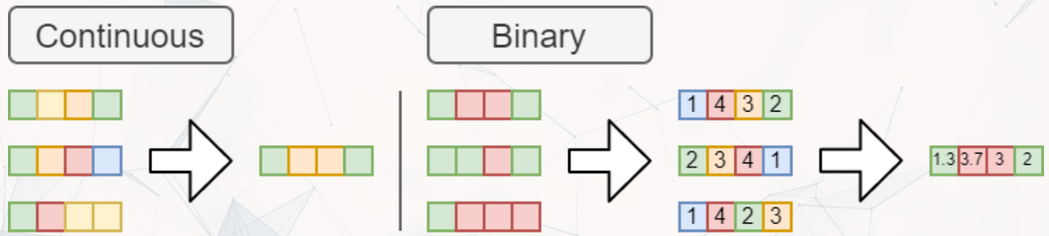


When are ensembles useful?

3. You really need to get just a bit more accuracy/less error out of the model, and you have some other models lying around
4. You want a more stable model
    - It helps to stabilize predictions by limiting the effect of errors or outliers produced by any one model on your prediction
    - Think: Diversification (like in finance)

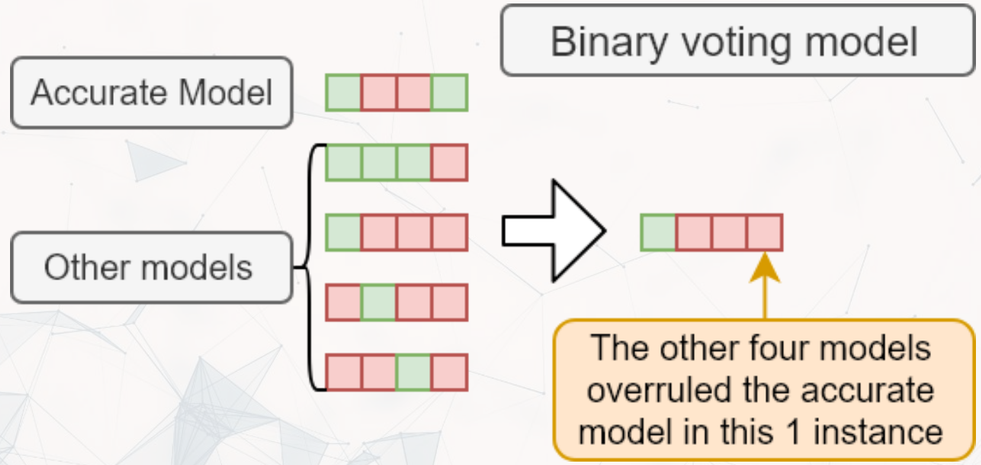
A simple ensemble (averaging)

- For continuous predictions, simple averaging is viable
  - Often you may want to weight the best model a bit higher
- For binary or categorical predictions, consider averaging *ranks*
  - i.e., instead of using a probability from a logit, use ranks 1, 2, 3, etc.
  - Ranks average a bit better, as scores on binary models (particularly when evaluated with measures like AUC) can have extremely different variances across models
    - In which case the ensemble is really just the most volatile model's prediction…
      - Not much of an ensemble



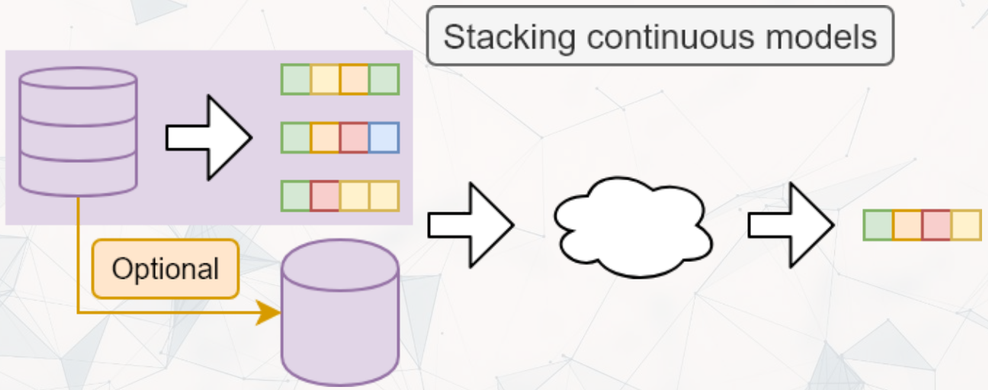
A more complex ensemble (voting model)

- If you have a model the is very good at predicting a binary outcome, ensembling can still help
  - This is particularly true when you have other models that capture different aspects of the problem
- Let the other models vote against the best model, and use their prediction if they are above some threshold of agreement



A lot more complex ensemble

- Stacking models (2 layers)
  1. Train models on subsets (folds) of the training data
  2. Make predictions for each model on the folds it wasn't applied to
  3. Train a new model that takes those predictions as inputs (and optionally the dataset as well)
- Blending (similar to stacking)
  - Like stacking, but using predictions on a hold out sample instead of folds (and thus all models are using the same data for predictions)



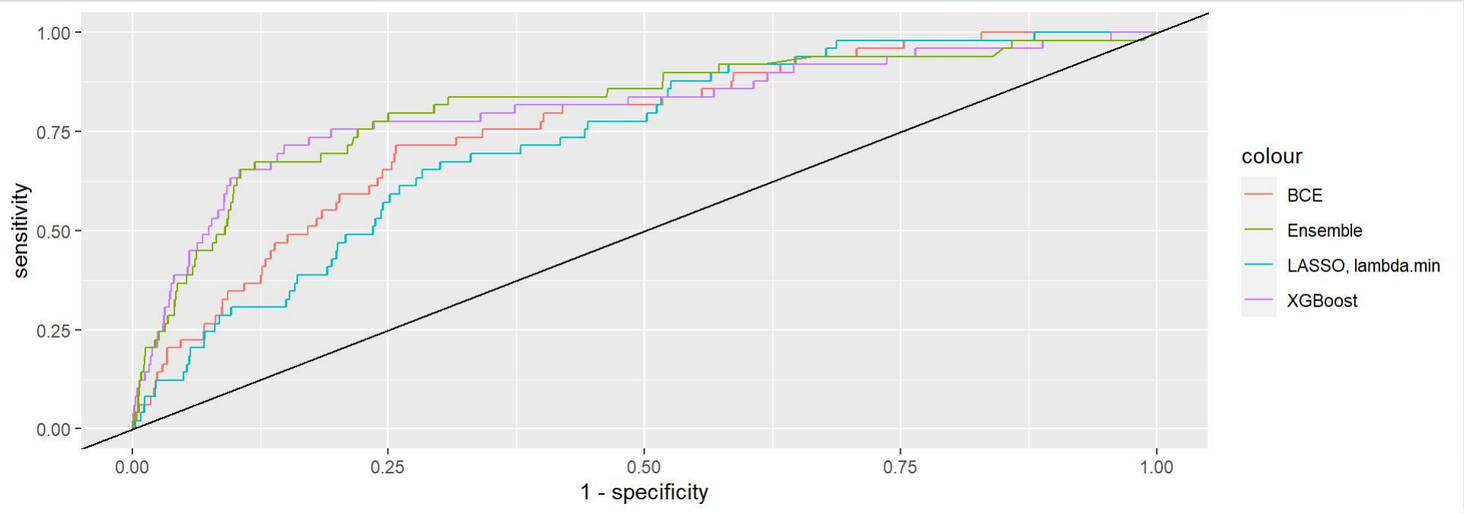
Simple ensemble example

- Ensembling all our fraud detection models from Session 6

```r
df <- readRDS('../../Data/Session_6_models.rds')
df2 <- df %>% mutate(AAER = ifelse(AAER==0,0,1))
library(xgboost)

# Prep data
train_x <- model.matrix(AAER ~ ., data=df2[df2$Test==0,-1])[,-1]
train_y <- model.frame(AAER ~ ., data=df2[df2$Test==0,])[,"AAER"]
test_x <- model.matrix(AAER ~ ., data=df2[df2$Test==1,-1])[,-1]
test_y <- model.frame(AAER ~ ., data=df2[df2$Test==1,])[,"AAER"]

set.seed(468435)  #for reproducibility
xgbCV <- xgb.cv(max_depth=5, eta=0.10, gamma=5, min_child_weight = 4,
                subsample = 0.5, objective = "binary:logistic", data=train_x,
                label=train_y, nrounds=100, eval_metric="auc", nfold=10,
                stratified=TRUE, verbose=0)
fit_ens <- xgboost(params=xgbCV$params, data = train_x, label = train_y,
                   nrounds = which.max(xgbCV$evaluation_log$test_auc_mean),
                   verbose = 0)
```

> We trained an XGBoost model on all our prior model outputs

Simple ensemble results



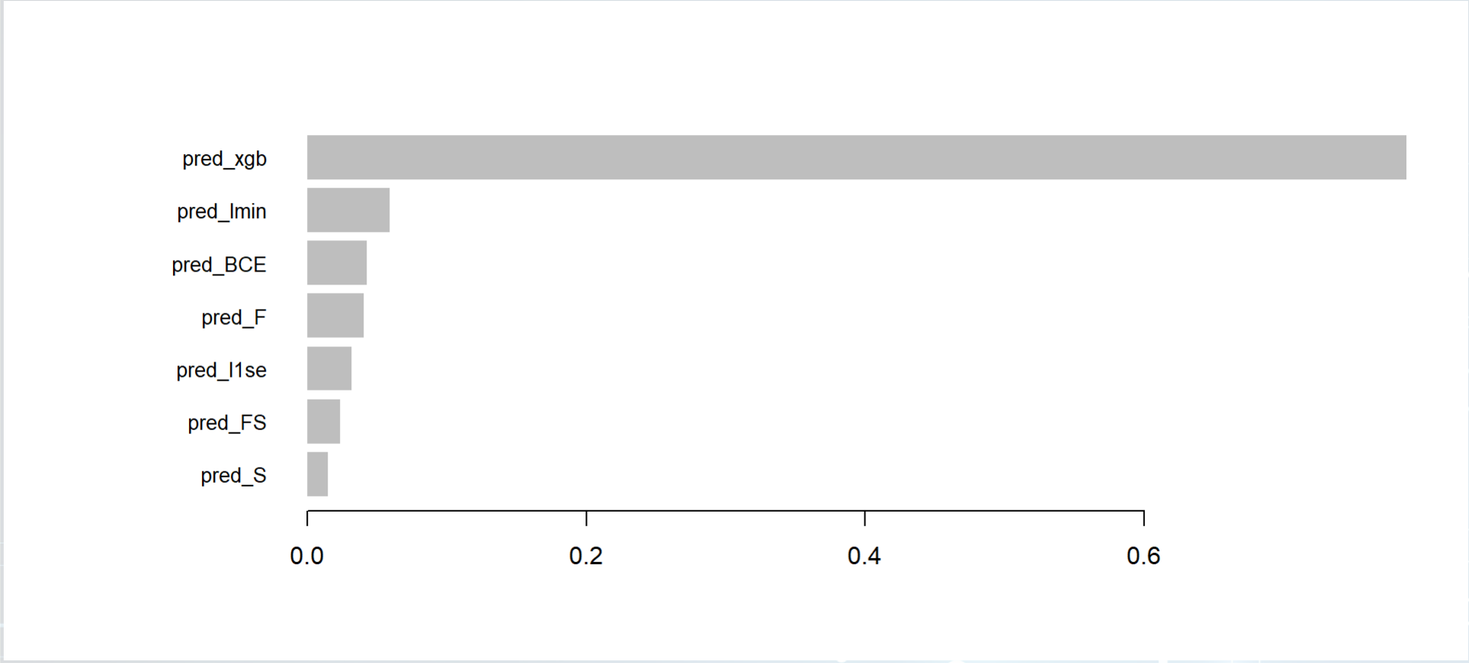

```
aucs  # Out of sample
```

Ensemble	Logit (BCE)	Lasso (lambda.min)	XGBoost
0.8169036	0.7599594	0.7290185	0.8083503

> The ensemble has the best performance

What drives the ensemble?

```r
xgb.train.data = xgb.DMatrix(train_x, label = train_y, missing = NA)
col_names = attr(xgb.train.data, ".Dimnames")[[2]]
imp = xgb.importance(col_names, fit_ens)
# Variable importance
xgb.plot.importance(imp)
```



Practicalities

- Methods like stacking or blending are much more complex than a simple averaging or voting based ensemble
  - But in practice they perform slightly better

> Recall the tradeoff between complexity and accuracy!

- As such, we may not prefer the complex ensemble in practice, unless we only care about accuracy

> Example: In 2009, Netflix awarded a $1M prize to the BellKor's Pragmatic Chaos team for beating Netflix's own user preference algorithm by >10%. The alogorithm was so complex that Netflix never used it. It instead used a simpler algorithm with an 8% improvement.

[Geoff Hinton's] Dark knowledge

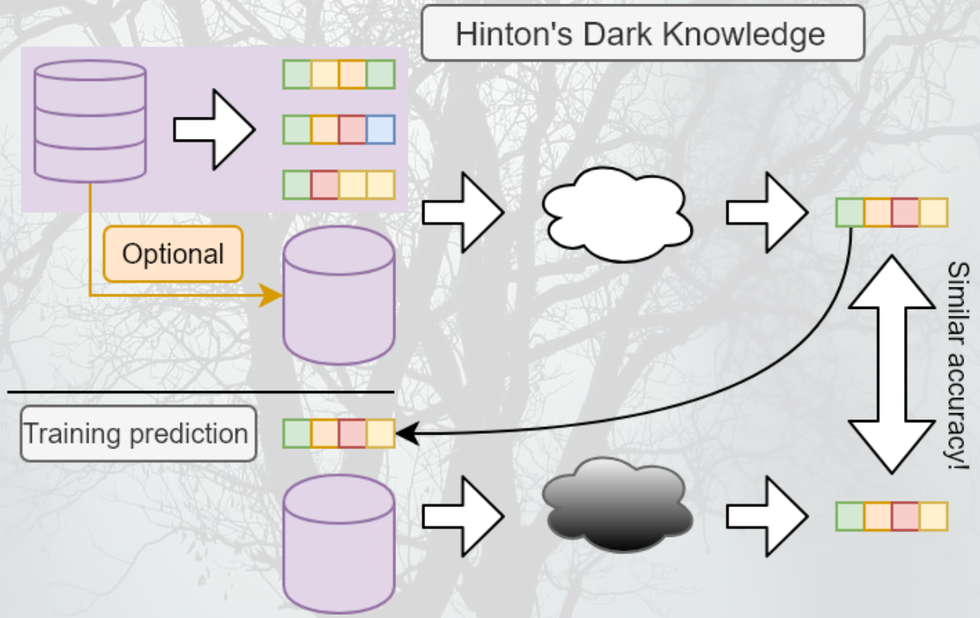
- Complex ensembles work well
- Complex ensembles are exceedingly computationally intensive
  - This is bad for running on small or constrained devices (like phones)

> Dark knowledge

- We can (almost) always create a simple model that approximates the complex model
  - Interpret the above literally – we can train a model to fit the model

Dark knowledge

- Train the simple model not on the actual DV from the training data, but on the best algorithm's (softened) prediction for the training data
- Somewhat surprisingly, this new, simple algorithm can work almost as well as the full thing!



An example of this dark knowledge

- Google's full model for interpretting human speech is >100GB
  - As of October 2019
- In Google's Pixel 4 phone, they have human speech interpretation running locally *on the phone*
  - Not in the cloud like it works on any other Android phone

> How did they do this?

- They can approximate the output of the complex speech model using a 0.5GB model
- 0.5GB isn't small, but it's small enough to run on a phone

Learning more about Ensembling

- Scikit-learn's documentation on ensemble methods it supports
- Geoff Hinton's Dark Knowledge slides
  - For more details on *dark knowledge*, applications, and the softening transform
  - His interesting (though highly technical) Reddit AMA
- Kaggle Ensembling Guide
  - A comprehensive list of ensembling methods with some code samples and applications discussed
- Ensemble Learning to Improve Machine Learning Results
  - Nicely covers bagging and boosting (two other techniques)

> There are many ways to ensemble, and there is no specific guide as to what is best. It may prove useful in the group project, however.


Ethics: Fairness

In class reading with case

- From Datarobot's Colin Preist:
  - Four Keys to Avoiding Bias in AI
    - Short link: rmc.link/420class9
  - *You only need to read the Case portion*
- The four points:
  1. Don't trust black boxes
  2. Check for direct discrimination by the model via inputs
  3. Check for indirect discrimination
  4. Use training data representative of the outcome you want

> What was the issue in the case? Where might similar issues crop up in business? Any examples?

Examples of Reputational damage

- Microsoft's Tay and their response
- Coca-Cola: Go make it happy
- Google: Google Photos mistakenly labels black people 'gorillas'
- Machine Bias
  - ProPublica's in depth look at racial bias in US courts' risk assessment algorithms (as of May 2016)
  - Note that the number of true positives divided by the number of all positives is more or less equal across ethnicities

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

But what about...

Commonsense knowledge

Logical reasoning

Linguistic phenomena

Intuitive physics

...

MACHINE LEARNING DOESN'T CARE

Fairness is complex!

- There are many different (and disparate) definitions of fairness
  - Arvind Narayanan's Tutorial: 21 fairness definitions and their politics
- For instance, in the court system example:
  - If an algorithm has the same accuracy across groups, but rates are different across groups, then true positive and false positive rates must be different!

> Fairness requires considering different perspectives and identifying which perspectives are most important from an ethical perspective

How could the previous examples be avoided?

- Filtering data used for learning the algorithms
  - Microsoft Tay should have been more careful about the language used in retraining the algorithm over time
    - Particularly given that the AI was trained on public information on Twitter, where coordination against it would be simple
- Filtering output of the algorithms
  - Coca Cola could check the text for content that is likely racist, classist, sexist, etc.
- Google may have been able to avoid this using training dataset that was sensitive to potential problems
  - For instance, using a *balanced* data set across races
  - As an intermediary measure, they removed searching for gorillas and its associated label from the app

Examining ethics for algorithms

1. Understanding the problem and its impact
    - Think about the effects the algorithm will have!
        - Will it drastically affect lives? If yes, exercise more care!
    - Think about what you might expect to go wrong
        - What biases might you expect or might find in the data?
        - What biases do people doing the same task exhibit?
2. Manual inspection
    - Check model outputs against problematic indicators
    - Test the algorithm before putting it into production
3. Use methods like SHAP to explain models
4. Some thoughts on the matter from leading companies:
    - Google: Responsible AI practices
    - Meta: Meta's Civil Rights Progress
    - OpenAI: How should AI systems behave, and who should decide?

Examining Fairness with SHAP

What is SHAP?

> SHAP aims to provide an explanation of the importance of model inputs in explaining model output

- SHAP: **SH**apley **A**dditive ex**P**lanations
- SHAP is…
  - game theoretic
  - theory driven
  - a model itself: a model that explains models
- SHAP provides a simple to understand output

SHAP is based on Shapley (1953), “A value for n-person games.” SHAP is derived in Lundberg and Lee (2017)

Key points of SHAP

1. SHAP maintains *local accuracy*
   - It can accurately model outputs on a small subset of data
2. SHAP maintains *missingness*
   - It only uses data used by the model
3. SHAP maintains *consistency*
   - The final output of SHAP follows the same rules as Utility in economics

> SHAP will give us something like a marginal effect, but for *any* model, including complex ML models

Data and approach

- City of Chicago salaries
  - 33,586 employees
- Trained using a simple XGBoost model
- Features:
  - Job title
  - Full time or part time
  - Salaried or hourly
  - Department
  - Gender (based on name)
- Code to replicate is on eLearn

SHAP values

```r
library('SHAPforxgboost')
#https://liuyanguu.github.io/post/2019/07/18/visualization-of-shap-for-xgboost/



vals <- shap.values(xgb_model = fit_xgb, X_train = train_x)
vals$mean_shap_score
```

```
                                  Salaried
                               30789.37053
                 Job.TitlesPOLICE OFFICER
                                3391.77843
                          DepartmentOther
                                2263.00574
                                  Female
                                2223.51487
                           DepartmentOEMC
                                 672.56147
                          Job.TitlesOther
                                 668.91992
                       Job.TitlesSERGEANT
                                 581.33778
                                Full.Time
                                440.57020
```

SHAP values by gender

Female

```
vals$shap_score[train_x[, 'Female'] == 1] %>%
  colMeans()
```

```
                        Job.TitlesMOTOR TRUCK
DRIVER

-46.47701

Job.TitlesOther

693.06793
                                Job.TitlesPOLICE
OFFICER

-669.98481
Job.TitlesPOLICE OFFICER (ASSIGNED AS
DETECTIVE)

-35.20260
```

Male

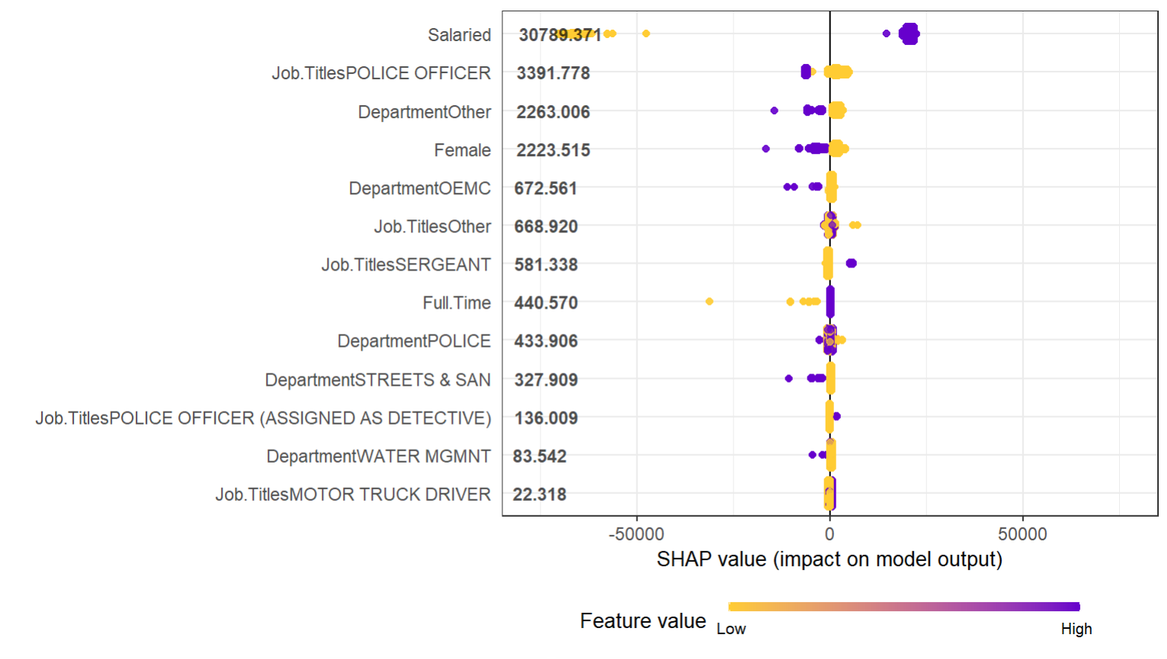
```
vals$shap_score[train_x[, 'Female'] == 0] %>%
  colMeans()
```

```
                        Job.TitlesMOTOR TRUCK
DRIVER

-2.730796

Job.TitlesOther

232.434377
                                Job.TitlesPOLICE
OFFICER

-158.892689
Job.TitlesPOLICE OFFICER (ASSIGNED AS
DETECTIVE)

-9.971078
```

> What do these differences mean?

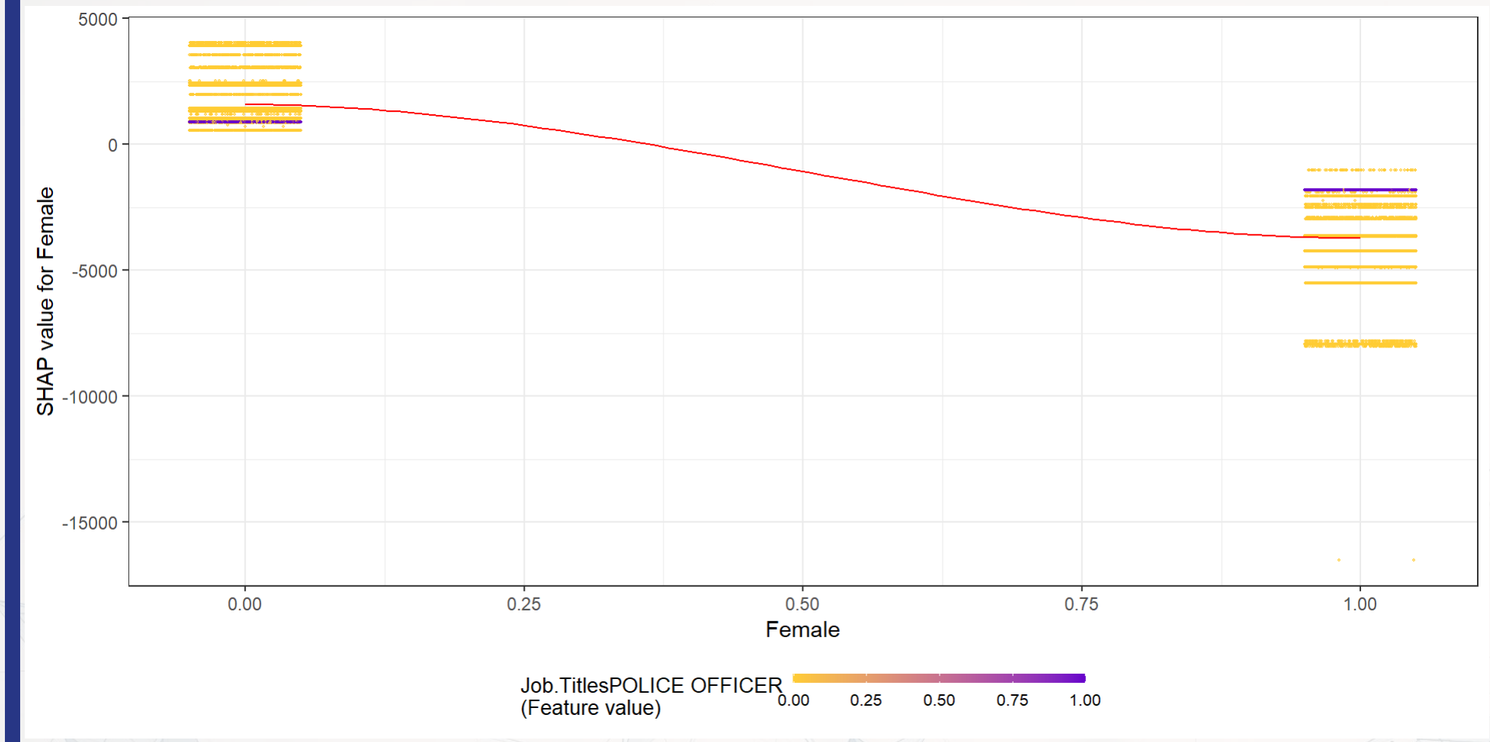
Visualizing SHAP across the population

```r
shap_long <- shap.prep(xgb_model = fit_xgb, X_train = train_x)
shap.plot.summary(shap_long)
```



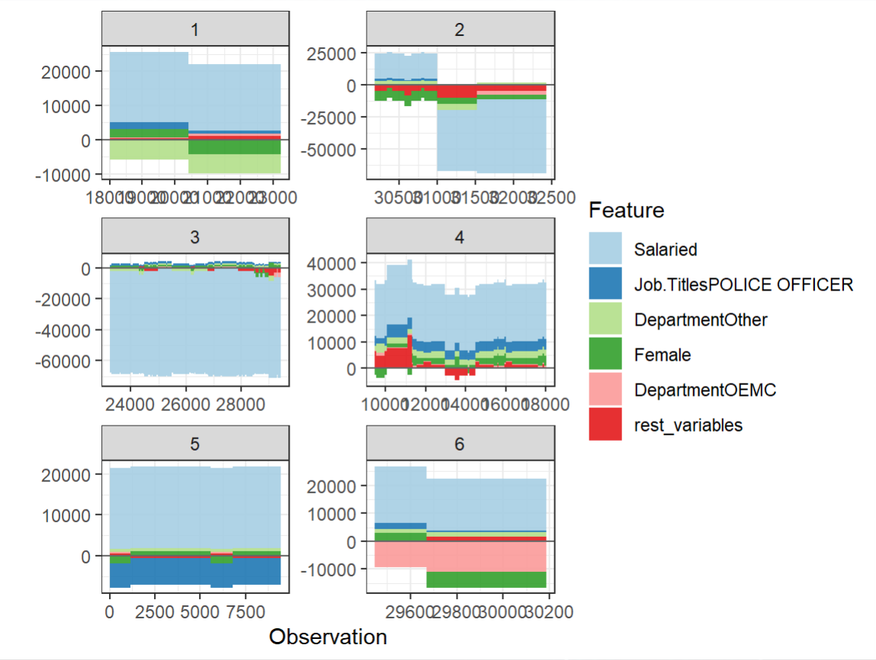
Visualizing differences in gender within a job

```r
shap.plot.dependence(data_long = shap_long, x = 'Female', y = 'Female',
                     color_feature = 'Job.TitlesPOLICE OFFICER', jitter_width = 0.05)
```



SHAP force plot: natural groupings in the data

```r
plot_data <- shap.prep.stack.data(shap_contrib = vals$shap_score,
                                  top_n = 5, n_groups = 6)
shap.plot.force_plot_bygroup(plot_data)
```



Areas where ethics is particularly important

- Anything that impacts people's livelihoods
  - Legal systems
  - Healthcare and insurance systems
  - Hiring and HR systems
  - Finance systems like credit scoring
  - Education
- Anything where failure is catastrophic
  - Voting systems
  - Engineering or transportation systems
    - Such as the Joo Koon MRT Collision in 2017
    - Self driving cars (Results summary)

> Article: Algorithms are great and all, but they can also ruin lives

Research on fairness

- Excerpt below from Universal Sentence Encoder

Target words	Attrib. words	Ref	GloVe		Uni. Enc. (DAN)	
			d	p	d	p
Eur.-American vs Afr.-American names	Pleasant vs. Unpleasant 1	a	1.41	10^{-8}	0.361	0.035
Eur.-American vs. Afr.-American names	Pleasant vs. Unpleasant from (a)	b	1.50	10^{-4}	-0.372	0.87
Eur.-American vs. Afr.-American names	Pleasant vs. Unpleasant from (c)	b	1.28	10^{-3}	0.721	0.015
Male vs. female names	Career vs family	c	1.81	10^{-3}	0.0248	0.48
Math vs. arts	Male vs. female terms	c	1.06	0.018	0.588	0.12
Science vs. arts	Male vs female terms	d	1.24	10^{-2}	0.236	0.32
Mental vs. physical disease	Temporary vs permanent	e	1.38	10^{-2}	1.60	0.0027
Young vs old peoples names	Pleasant vs unpleasant	c	1.21	10^{-2}	1.01	0.022
Flowers vs. insects	Pleasant vs. Unpleasant	a	1.50	10^{-7}	1.38	10^{-7}
Instruments vs. Weapons	Pleasant vs Unpleasant	a	1.53	10^{-7}	1.44	10^{-7}

Table 4: Word Embedding Association Tests (WEAT) for GloVe and the Universal Encoder. Effect size is reported as Cohen's d over the mean cosine similarity scores across grouped attribute words. Statistical significance is reported for 1 tailed p-scores. The letters in the *Ref* column indicates the source of the IAT word lists: (a) Greenwald et al. (1998) (b) Bertrand and Mullainathan (2004) (c) Nosek et al. (2002a) (d) Nosek et al. (2002b) (e) Monteith and Pettit (2011).

> Compares a variety of unintended associations (top) and intended associations (bottom) across Global Vectors (GloVe) and USE

Ethical implications: Case

- In Chicago, IL, USA, they are using a system to rank arrested individuals, and they use that rank for *proactive* policing
- Read about the system here: rmc.link/420class9-2

> What risks does such a system pose?

> How would you feel if a similar system was implemented in Singapore?

Other references

> Kate Crawford’s NIPS 2017 Keynote: “The Trouble with Bias” (video)

Ethics: Data security

Anomymized data

- Generally we anonymize data because, while the data itself is broadly useful, providing full information could harm others or oneself
- Ex.: Studying drug usage can create a list of people with potentially uncaught criminal offenses
  - If one retains a list of identities, then there is an ethical dilemma:
    - Protect study participants by withholding the list
    - Provide the list to the government
      - This harms future knowledge generation by sowing distrust
  - Solution: Anonymous by design
- Website or app user behavior data
  - E.g.: FiveThirtyEight's Uber rides dataset

> What could go wrong if the Uber data wasn't anonymized?

Anonymization is tricky

> Both Allman & Paxson, and Partridge warn against relying on the anonymisation of data since *deanonymisation techniques are often surprisingly powerful*. Robust anonymisation of data is difficult, particularly when it has high dimensionality, as the anonymisation is likely to lead to an unacceptable level of data loss [3]. – TPHCB 2017

- There are natural limits to anonymization, particularly when there is a limited amount of potential participants in the data
  - Example: Web browser tracking at Panopticlick

Responsibilities generating data

- Keep users as unidentifiable as feasible
- If you need to record people's private information, **make sure they know**
  - This is called *informed consent*
- If you are recording sensitive information, consider not keeping identities at all
  - Create a new, unique identifier (if needed)
  - Maintain as little identifying information as necessary
  - Consider using encryption if sensitive data is retained
  - Can unintentionally lead to infringements of *human rights* if the data is used in unintended ways

Informed consent

- When working with data about *people*, they should be informed of this and consent to the research, unless the data is publicly available
- From SMU's IRB Handbook: (2017 SEP 18 version)

- "*Informed consent*: Respect for persons requires that participants, to the degree that they are capable, be given the opportunity to make their own judgments and choices. When researchers seek participants' participation in research studies, they provide them the opportunity to make their own decisions to participate or not by ensuring that the following adequate standards for informed consent are satisfied:
  - *Information*: Participants are given sufficient information about the research study, e.g., research purpose, study procedures, risks, benefits, confidentiality of participants' data.
  - *Comprehension*: The manner and context in which information is conveyed allows sufficient comprehension. The information is organized for easy reading and the language is easily comprehended by the participants.
  - *Voluntariness*: The manner in which researchers seek informed consent from the participants to participate in the research study must be free from any undue influence or coercion. Under such circumstances, participants are aware that they are not obliged to participate in the research study and their participation is on a voluntary basis."

> Also, note the existence of the PDPA law in Singapore

Human rights

- Recall the drug users example
  - If data was collected without their consent, and if it was not anonymized perfectly, then this could lead to leaking of drug user's information to others

> What risks does this pose? Consider contexts outside Singapore as well.

Responsibilities using data

> “The collection, or use, of a dataset of illicit origin to support research can be advantageous. For example, legitimate access to data may not be possible, or the reuse of data of illicit origin is likely to require fewer resources than collecting data again from scratch. In addition, the sharing and reuse of existing datasets aids reproducibility, an important scientific goal. The disadvantage is that ethical and legal questions may arise as a result of the use of such data” (source)

Responsibilities using data

- Respect for persons
  - Individuals should be treated as *autonomous agents*
    - People **are** people
  - Those without autonomy should be protected
- Beneficence
  1. Do not harm (ideally)
  2. Maximize possible benefits and minimize possible harms
     - This can be a natural source of conflict
- Justice
  - Benefits and risks should flow to the same groups – don't use unwilling or disadvantaged groups who won't receive any benefit
    - [Extreme] example: Tuskegee Syphilis study

> Experiments: The Belmont Report; Electronic data: The Menlo Report

End Matter

Wrap up

Today, we:

- Learned about combining models to create an even better model
  - And the limits to this as pointed out by Geoff Hinton
- Discussed the potential ethical issues surrounding:
  - AI algorithms
  - Data creation
  - Data usage
- Learned how to implement a tool for analyzing fairness: SHAP
- Survey on the class session at this QR code:

For next week

- For the next 2 weeks:
  - We will talk about neural networks and vector methods (which are generally neural network based)
    - These are important tools underpinning a lot of recent advancements
    - We will take a look at some of the advancements, and the tools that underpin them
  - If you would like to be well prepared, there is a nice introductory article here (8 parts though)
    - Part 1 is good enough for next week, but part 2 is also useful
    - For those very interested in machine learning, parts 3 through 8 are also great, but more technical and targeted at specific applications like facial recognition and machine translation
  - Keep working on the group project

Fun machine learning examples

- Interactive:
  - Semantris
    - A game based on the Universal Sentence Encoder
  - Draw together with a neural network
    - click the images to try it out yourself!
  - Google's Quickdraw
  - Google's Teachable Machine
  - Four experiments in handwriting with a neural network
- Non-interactive
  - Predicting e-sports winners with Machine Learning

Packages used for these slides

- DT
- downlit
- gender
- kableExtra
- knitr
- leaflet
- plotly
- quarto
- revealjs
- SHAPforxgboost
- tidyr
- tidyverse
- yardstick