MONASH University

ETX2250/ETF5922: Data Visualization and Analytics

Analytics

Lecturer: *Lauren Kennedy*

Department of Econometrics and Business Statistics

lauren.kennedy1@monash.edu

Week 6

Prediction

- Prediction arises in many business contexts.
- There is some unknown variable that is the target of the prediction.
  - This is usually denoted y and may be called the *dependent variable*, or *response* or *target variable*.
- There are some known variables that are used to make the prediction.
  - These are usually denoted $\mathbf{x}$ and may be called the *independent variables*, or *predictors* or *features*.

Supervised Learning

- For some observations data will be available for both y AND $\mathbf{x}$.
- We can use these observations to *learn* some rule that gives predictions of y as a function of $\mathbf{x}$.
- This prediction is denoted $\hat{y} = \hat{f}(\mathbf{x})$
- This general setup is often called *supervised learning*.

Unsupervised Learning

- There is only information available for $\mathbf{x}$.
- We group together similar observations to create clusters, or learn relationships between the observations
- In machine learning, we would call this training with `unlabelled data`.
- As a general rule, supervised learning is more likely to have greater accuracy.

Summary : Supervised learning

Variable	Training	Evaluation
Predictor 1 (X1)	Data available	Data available
Predictor 2 (X2)	Data available	Data available
Dependent Variable (Y)	Data available	Data NOT available

Example : Supervised learning

Variable	Old Customers	New Customer
Age (X1)	Data available	Data available
Limit (X2)	Data available	Data available
Default (Y)	Data available	Data NOT available

Summary : Unsupervised learning

Variable	Training	Evaluation
Variable 1 (X1)	Data available	Data available
variable 2 (X2)	Data available	Data available
Class	Data NOT available	Data NOT available

Example : Unsupervised learning

Variable	Observed penguin	New penguin
Bill length (X1)	Data available	Data available
Bill depth (X2)	Data available	Data available
Species	Data NOT available	Data NOT available

Supervised learning

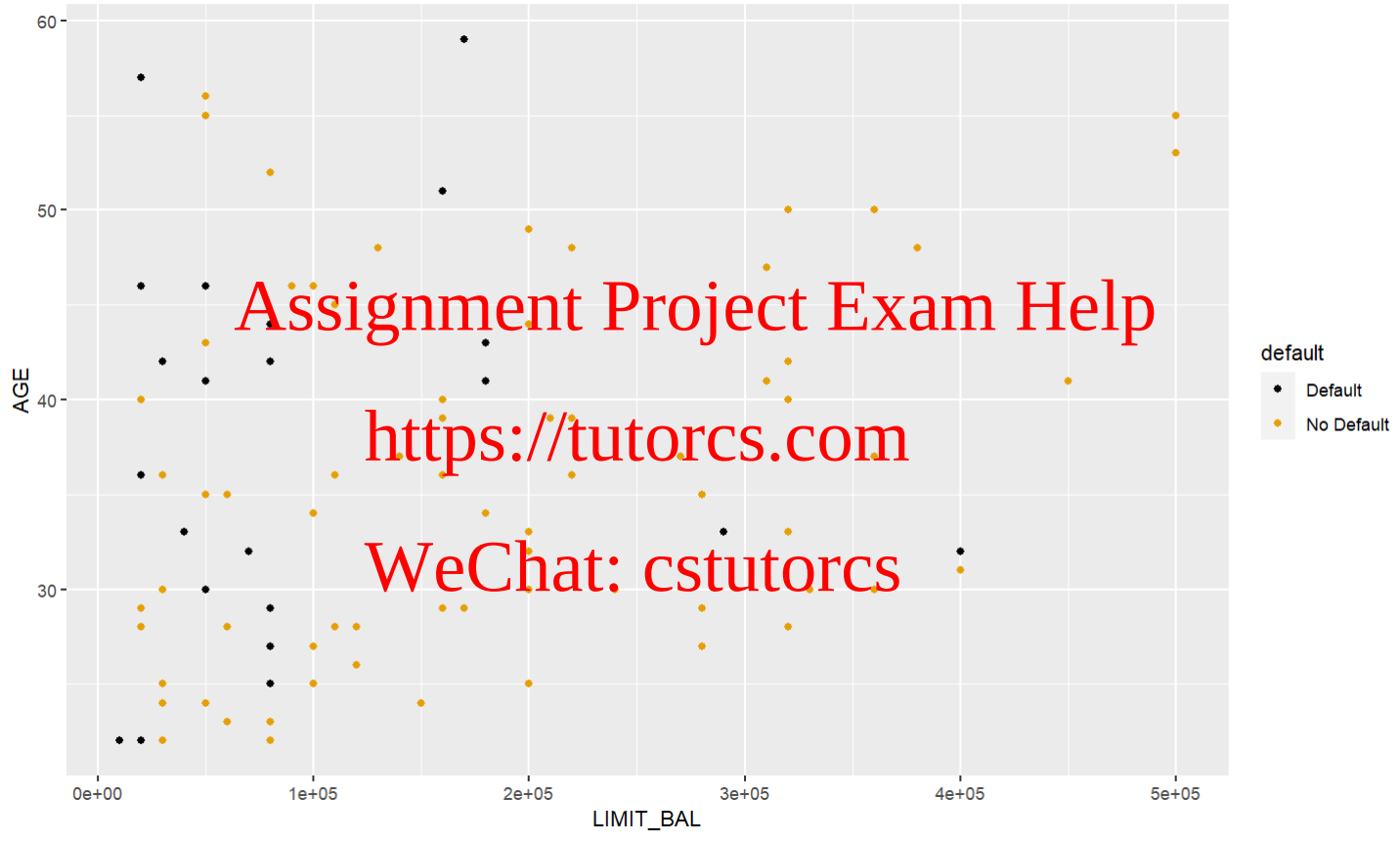
Regression

- Sometimes y is a numeric (metric) variable. For example
  - Company profit next month.
  - Amount spent by a customer.
  - Demand for a new product.
- In this case we are doing *regression*.
- This can be more general than the linear regression that you may be familiar with.

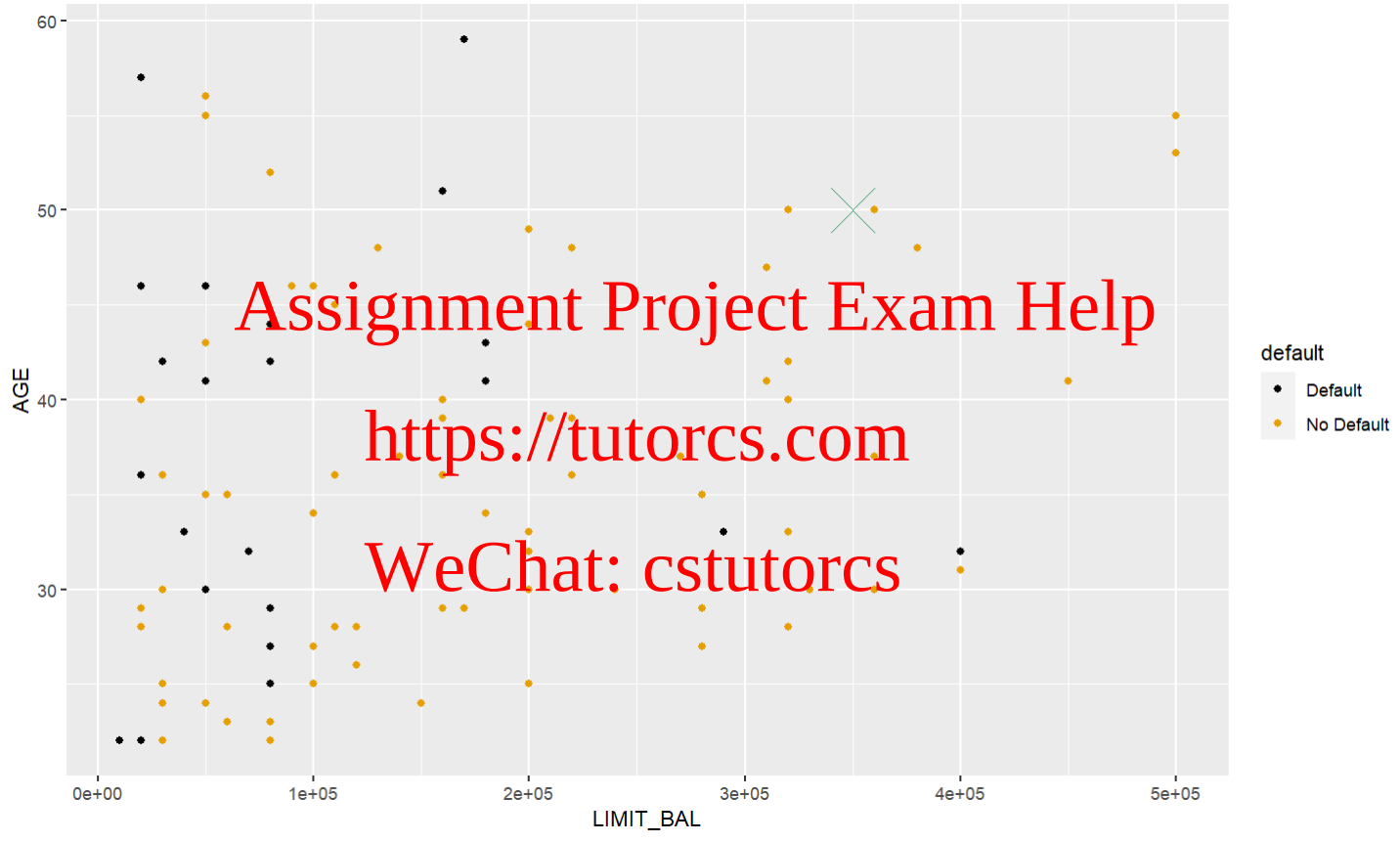
Classification

- Sometimes y is a categorical (nominal, non-metric) variable. For example
  - Will a borrower default on a loan?
  - Can we detect which tax returns are fraudulent?
  - Can we predict which brand customers will choose?
- In this case we are doing *classification*.

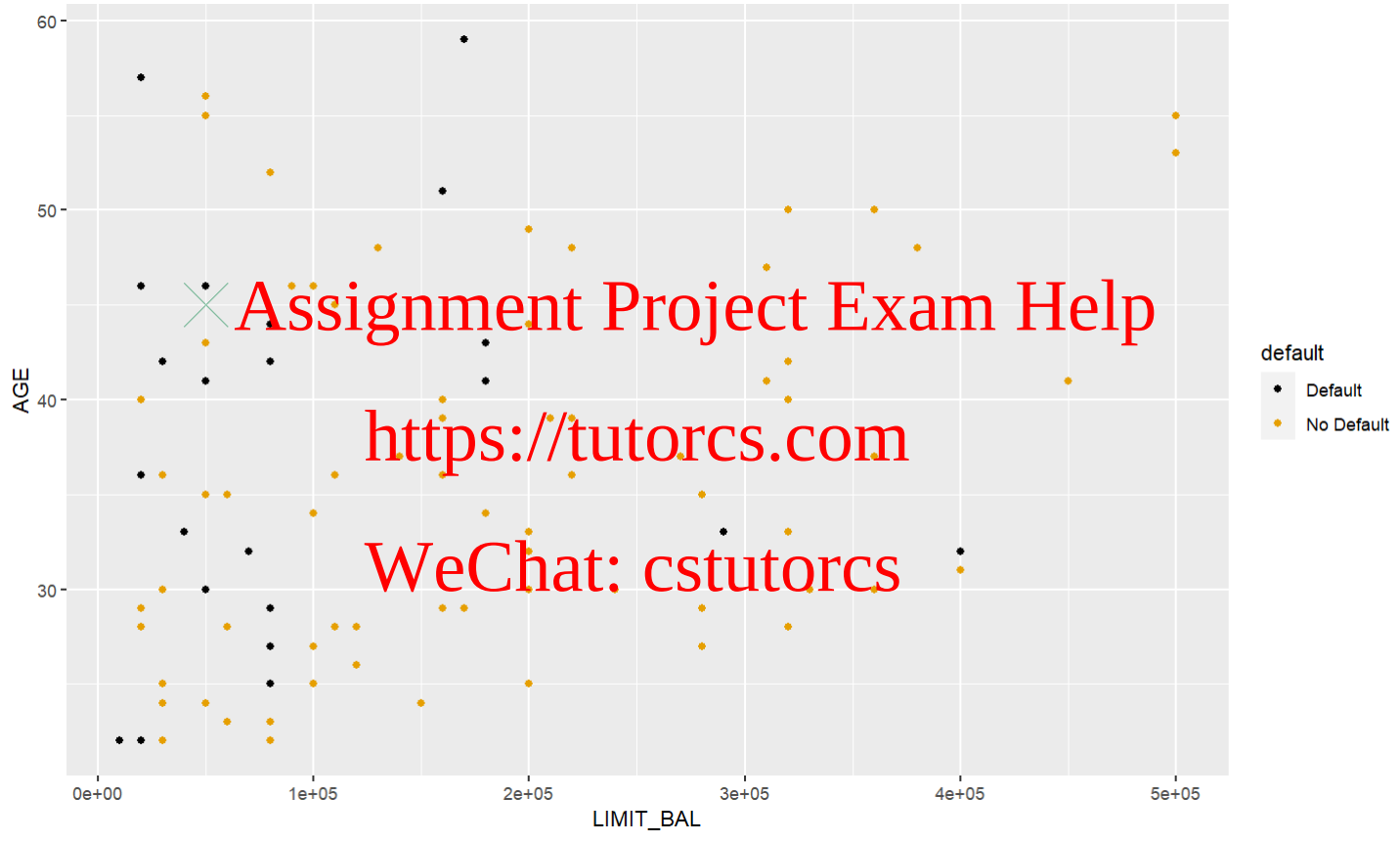
Classification example: Credit Data



Default or not?



Default or not?



Assessing Classification

Some math

- Generally data y_i and $\mathbf{x}_i$ available for $i = 1, 2, 3, \ldots, n$.
- An algorithm is trained on this data. Some function of $\mathbf{x}_i$ is derived where $\hat{y} = \hat{f}(\mathbf{x})$.
- How to decide if $\hat{f}$ is a good classifier or bad classifier?

Misclassification

- The **misclassification error** is given by

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

- Here $I(.)$ equals 1 of the statement in parentheses is true and 0 otherwise.
- Large numbers imply a worse performance.
- Since all n points are used for training and evaluation this measures *in-sample* performance.

Training v Test

- In practice we want predictions for values of y that are not yet observed.
- To artificially create this scenario the data we have available can be split into two
  - *Training* sample used to determing $\hat{f}$
  - *Test* sample used to evaluate $\hat{f}$
- The y values of the test sample will be treated as unknown during training.

Notation

- $\mathcal{N}_1$ is the set of indices for training data.
- $|\mathcal{N}_1|$ is the number of observations in training data.
- $\mathcal{N}_0$ is the set of indices for test data
- $|\mathcal{N}_0|$ is the number of observations in test data.

Example

- Suppose there are five observations, $(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_5, \mathbf{x}_5)$
- Suppose observations 1,2 and 4 are used as training data.
- Suppose observations 3 and 5 are used as test data.
- Then $\mathcal{N}_1 = \{1, 2, 4\}$ and $|\mathcal{N}_1| = 3$
- And $\mathcal{N}_0 = \{3, 5\}$ and $|\mathcal{N}_0| = 2$
- Only the data in $\mathcal{N}_1$ is used to determine $\hat{f}$.

Training v Test

Training error rate

$$\frac{1}{|\mathcal{N}_1|}\sum_{i\in\mathcal{N}_1} I(y_i \neq \hat{y}_i)$$

Test error rate

$$\frac{1}{|\mathcal{N}_0|}\sum_{i\in\mathcal{N}_0} I(y_i \neq \hat{y}_i)$$

Overfitting

- Some methods perform very well (even perfectly) on training error rate.
- Usually these same methods will perform poorly on test error rate.
- This phenomenon is called *overfitting*.
- Generally achieving a low **test** error rate (also called out-of-sample or generalisation error) is more important.

A simple example

- Consider a test set of a single observation $\mathcal{N}_0 = \{j\}$.
- The classifier is trained using all data apart from j.
- This classifier is then used to predict the value of y_j.
- The choice of j may seem arbitrary.

Extending the idea of test and training

- The process can be repeated so that each observation is left out exactly once.
- Each time all remaining observations are used as the training set.
- This process is called **Leave-one-out cross validation(LOOCV)**

k-fold CV

- A faster alternative to LOOCV is **k-fold cross validation**
- The data are randomly split into k partitions.
- Each observation appears in exactly one partition, i.e. the partitions are *non-overlapping*.
- Each partition is used as the test set exactly once.

Confusion matrix

- Consider predicting one of two categories (default or no default)
- We have the observed data (the truth) and our predicted outcome. The set of all possible outcomes is in the confusion matrix.

Actual vs Predicted	Default (predicted)	No default (predicted)
Default (actual)	True Positive	False Negative
No default (actual)	False Positive	True Negative

- The total error rate is the sum of misclassification over total number of classications

$$ErrorRate = \frac{FalsePostive + FalseNegative}{TruePositive + TrueNegative + FalsePositive + FalseNegative}$$

Sensitivity v Specificity

- In a 2-class problem think of one class as the presence of a condition and the other class as the absence of a condition.
- In an auditing example the condition can be that the person is guilty.
- *Sensitivity* refers to the true positive rate (also called recall). The proportion of guilty classified as guilty.

$$TruePositiveRate = \frac{TruePositive}{TruePositive + FalseNegative}$$

- *Specificity* refers to the true negative rate(also called selectivity. The proportion of innocent classified as innocent.

$$TrueNegativeRate = \frac{TrueNegative}{FalsePositive + TrueNegative}$$

- False positive rate

$$FalsePositiveRate = \frac{FalsePositive}{FalsePositive + TrueNegative}$$

Probabilistic Classification

- In many cases an algorithm will predict a single "best" class.
  - Predict a customer will purchase Gucci.
- In other instances an algorithm will provide probabilities.
  - The customer has a 40% chance of purchasing Gucci, a 35% of chance of purchasing Givenchy and a 25% chance of purchasing YSL.

Probabilistic Classification

- A probabilistic prediction can be converted to a point prediction.
- Simply choose the class with the highest probability.
- In the example on the previous slide the choice would be Gucci.

Two class case

- In the two class case, choosing the class with highest probability is simple.
- Assign to a class if the probability is greater than 0.5
- In some applications a different threshold may be used.
- This is particularly the case if there are asymmetric costs involved with different types of misclassification.

An example

- Suppose you work for the tax office.
- You need to decide who should be audited and who should not be audited.
- When doing classification you can make two mistakes
  - Audit an innocent person
  - Fail to audit a guilty person
- Are these mistakes equally costly?

Tax example

- Auditing an innocent person is costly since resources are used for no gain.
  - Suppose it costs $100 to audit a person.
- Failing to audit a guilty person is costly since there is a failure to recover tax revenue.
  - Let $500 be recovered from the guilty.
- In this example, it is more costly to fail to audit the guilty.
- However, misclassification rate treats both errors the same.

Sensitivity v Specificity

- Consider that we audit when the probability of being guilty is greater than 50%.
- Changing this threshold can change the sensitivity and specificity.
- Reducing the threshold to 0 means everyone is audited. The sensitivity will be perfect but specificity will be zero.
- Raising the threshold to 1 means no one is audited. The specificity will be perfect but sensitivity will be zero.

Example

Person	Pred. Pr. Guilty	Truth
A	0.3	Not Guilty
B	0.4	Guilty
C	0.6	Guilty
D	0.7	Guilty

Questions

- For a threshold of 0.5
  - What is your prediction for each individual?
  - What is the misclassification error?
  - What is the sensitivity?
  - What is the specificity?
  - What is the cost?

Answer

Person	Pred. Pr. Guilty	Prediction	Truth
A	0.3	Not Guilty	Not Guilty
B	0.4	Not Guilty	Guilty
C	0.6	Guilty	Guilty
D	0.7	Guilty	Guilty

Answers

- Misclassification error is 0.25.
- Sensitivity is 0.6667
- Specificity is 1
- Cost is $500

Your turn

- How do the answers change when the threshold is 0.2?
- How do the answers change when the threshold is 0.65?

Answer (Threshold 0.2)

Person	Pred. Pr. Guilty	Prediction	Truth
A	0.3	Guilty	Not Guilty
B	0.4	Guilty	Guilty
C	0.6	Guilty	Guilty
D	0.7	Guilty	Guilty

Answers

- Misclassification error is 0.25.
- Sensitivity is 1
- Specificity is 0
- Cost is $100

Answer (Threshold 0.65)

Person	Pred. Pr. Guilty	Prediction	Truth
A	0.3	Not Guilty	Not Guilty
B	0.4	Not Guilty	Guilty
C	0.6	Not Guilty	Guilty
D	0.7	Guilty	Guilty

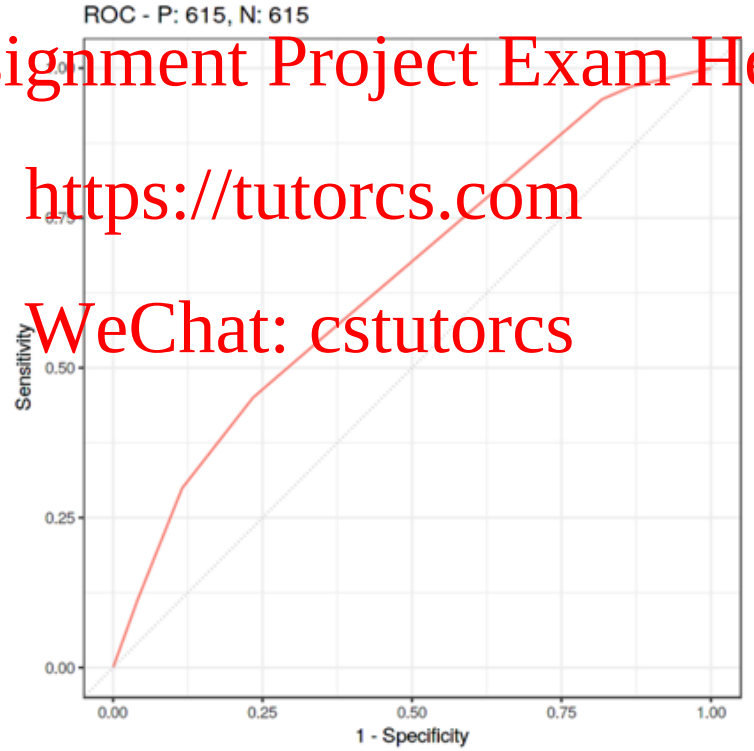
Answers

- Misclassification error is 0.5.
- Sensitivity is 0.333
- Specificity is 1
- Cost is $1000

Receiver operating characteristic curve

-Vary the threshold and and plot the TPR against the FPR

- When AUC (area under curve) is close to 1, close to perfect!
- When AUC (area under curve) is .5, no separation between groups



Unsupervised learning

Cross validation

- Split the data into two sub samples (randomly splitting)
- Run your unsupervised method on each seperately
- Compare the two cluster solutions for consistency
  - number of clusters
  - cluster profiles

External validation

- Examine the difference on variables not included in the cluster analysis but for which you'd expect theoretical and relevant reason to expect variation across the clusters

Additional issues with prediction

Predict v Explain

- In this unit our emphasis will be on *prediction*.
- This is very different to *explanation* or *causality*.
- Consider the example of predicting sales of a Toyota by looking at number of internet searches for "Toyota".
- If there is a large number of people searching for "Toyota" it is more likely for sales of Toyota in the following period to be higher.

Causality

- This relationship is not easy to manipulate.
- For instance, if Toyota instructs its employees to spend the afternoon searching the word "Toyota" on Google, sales will not go up.
- In this case there is a *common cause* for browsing for cars and buying cars; namely the intent to buy cars.
- Unlike intent to buy a car, browsing behaviour is observable and can be used for prediction.

Causal language vs correlation language

Causal language

- A causes B
- A explains B

Correlation language

- A is associated with B

- An increase in A is related to an increase in B

Ethics and analytics

Data driven decision making

- Increasingly in society we see data used to drive decision making
- This is often to the strength and benefit of society
  - recent use of data driven decisions to govern decisions surrounding lockdowns and safety precautions during Covid-19
- However, data driven decisions can also have negative consequences, unintended at initial outset
  - Robodebt saga
- We do not have time for a full exploration of data ethics, but we will touch on some ideas today
- For a checklist of ethics in data and more resources on each particular topic see https://deon.drivendata.org/

Dataset bias

- The data used to train the algorithm is not representative of the people it will be used to make decisions for, or contains undesirable instances of societal bias.
- Accordingly, the algorithm might learn false or undesirable relationships
- If the test data is random sample of the dataset used for training, this bias may not be apparent until rollout
- An example: A word network is trained using Google news archives. It learns relationships between words - like female and queen. As the dataset contains biases, the algorithm also learns gendered associations, like male and engineer.

Feedback loops

- In deployed algorithms, the data is sometimes fed back in to continuously update predictions.
- This can lead to feedback loops - where the decisions made from the algorithm impact the data, which then impacts the decisions made by the algorithm
- If the data is initially biased, this can exaggerate the bias considerably.
- An example: We use an algorithm to predict where crimes are likely to occur, and allocate police officers accordingly. The areas predicted to have high crime record more crime because there are more police officers. This is fed into the algorithm, which results in even higher predictions of crime in those areas, which leads to a higher police presence etc.

Proxy discrimination

- It is against the law to discriminate against certain classes of people (e.g., gender, race/ethnicity, religion), and so these variables are not always used in predictive analytics
- However, proxy variables which are strongly related to these protected classes can have the same effect.
- An example: Amazon explores (but doesn't use) an algorithm to assist with hiring decisions. As they are hiring in tech, there is already a substantial gender bias in the existing workforce. The algorithm rates candidates from all women universities lower than those from other universities.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Fairness across groups

- In this lecture we've discussed the error rate over the dataset (testing and/or training)
- However, we might also look at the error rate for particular groups of individuals, particularly those in protected classes (gender, race/ethnicity, age etc)
- The algorithm error should not be significantly higher for these classes when compared to the broader dataset.
- An example: Recidivism (when criminals re-offend) prediction error is higher when the candidate is black when compared to white candidates.

Conclusion

- For the remainder of the unit the focus is on different analytic techniques. The next three weeks will be unsupervised methods, followed by three weeks on supervised methods.
- In a business (and any other setting) be aware that
  - Correlation does not imply causation
  - Prediction should be thought about probabilistically.
  - Cost should be taken into account when classification is used in decision making.

MONASH University

Lecturer: *Lauren Kennedy*

Department of Econometrics and Business Statistics

lauren.kennedy1@monash.edu

Week 6