Machine Learning From Imbalanced Data Sets

Kwanda Ngwenduna, Data Scientist & Wits Masters Student, Vodacom

Rendani Mbuvha, Actuary and Machine Learning Researcher, Wits University

3 April 2019, presented at the IAA Colloquium 2019

Hosted by ACTUARIAL SOCIETY OF SOUTH AFRICA

SECTION COLLOQUIUM 2019

AIMS AND OBJECTIVES

- Understand the class imbalance problem and its issues;
- Review existing techniques to address class imbalance;
- Use example mortality data and compare these techniques; and
- Propose a consideration of generative models such as generative adversarial networks (GAN) as an alternative to popular resampling techniques

EVALUATION METRICS

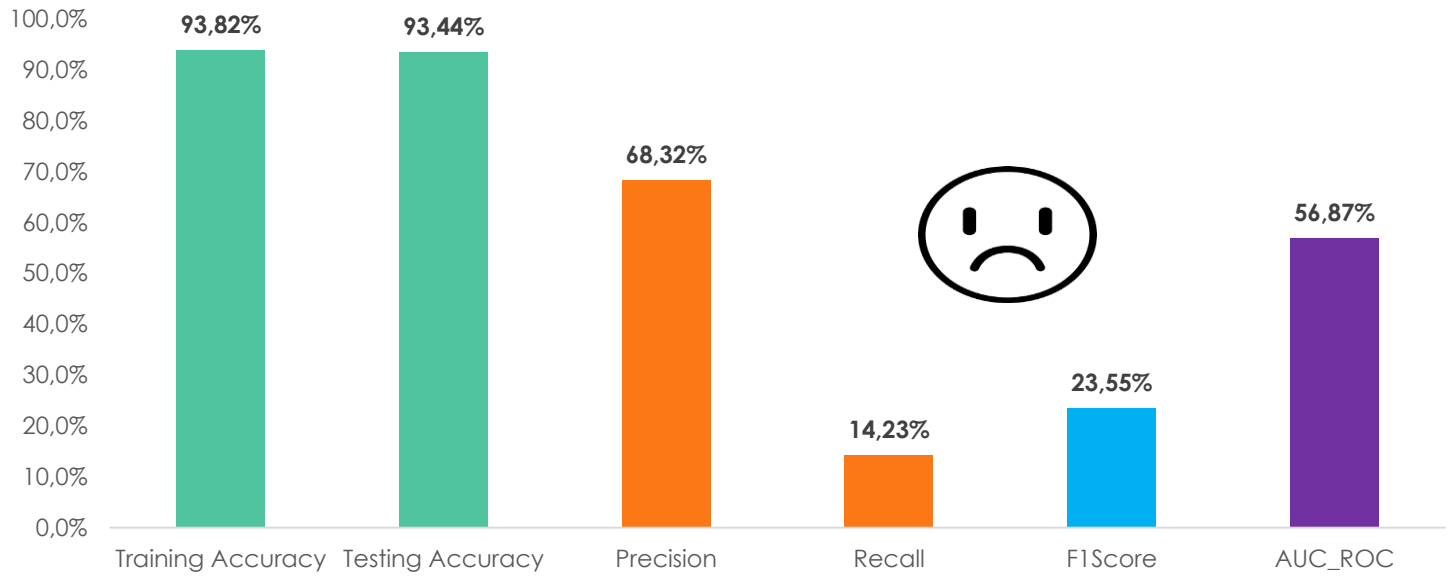
- Accuracy = (TN + TP)/(TN+FN+FP+TP)
- Precision = TP/(TP+FP) i.e. completeness
- Recall = TP/(TP+FN) i.e. exactness
- F-measure: harmonic mean between precision and recall
- Receiver Operating Curve (ROC)
  - Trade-off between TP and FP rate
- **Area under ROC (AUC)**

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

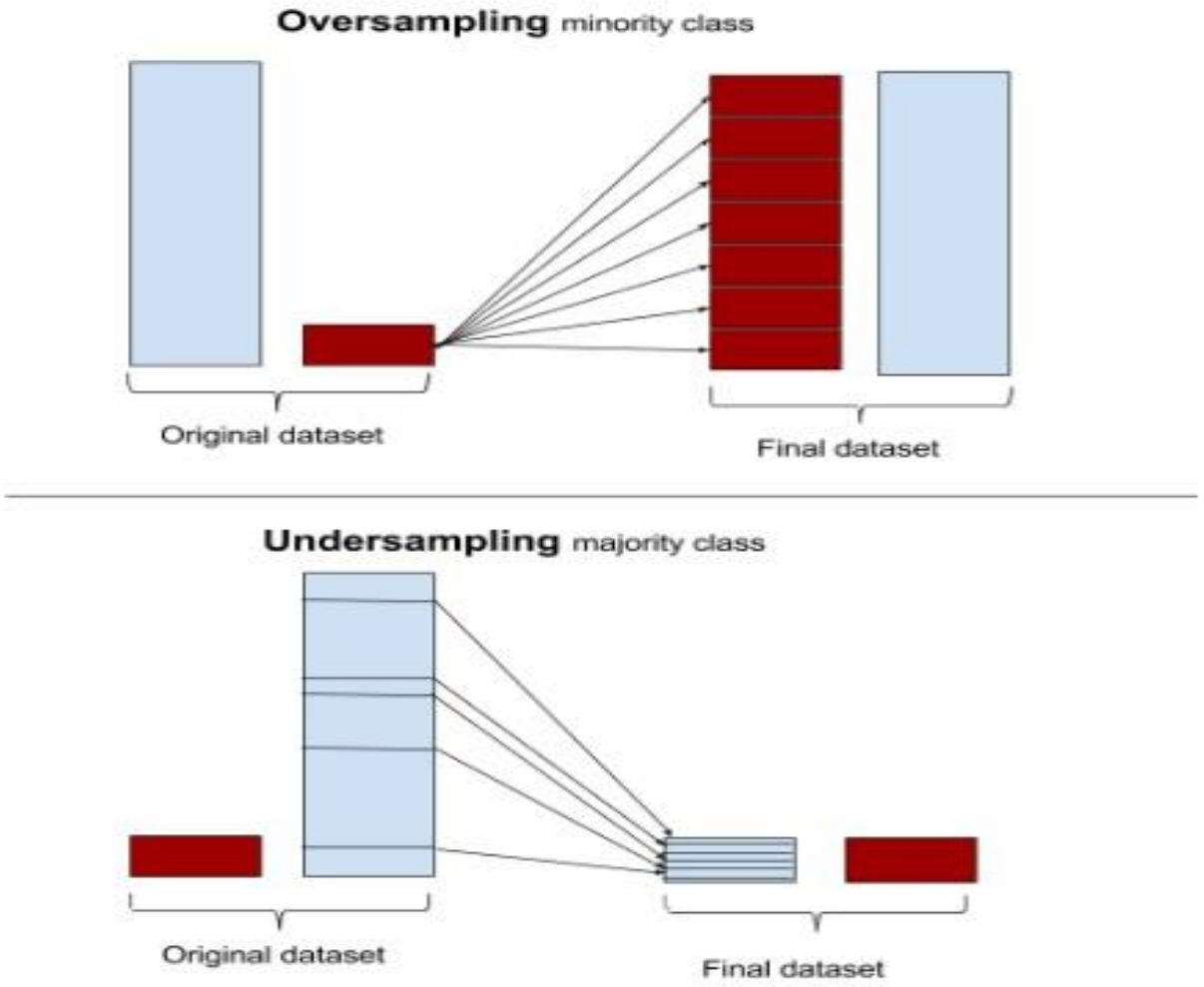
TN **True Negative**
FP **False Positive**
FN **False Negative**
TP **True Positive**

CLASS IMBALANCE

- Overwhelmingly many more instances of a class than others
  - Biasing classification measures
  - Cost of misclassification can be very high
- Accidental mortality: **Survived** (93.2%) and **Death** (6.8%)
- Logistic regression trained in Python for mortality prediction

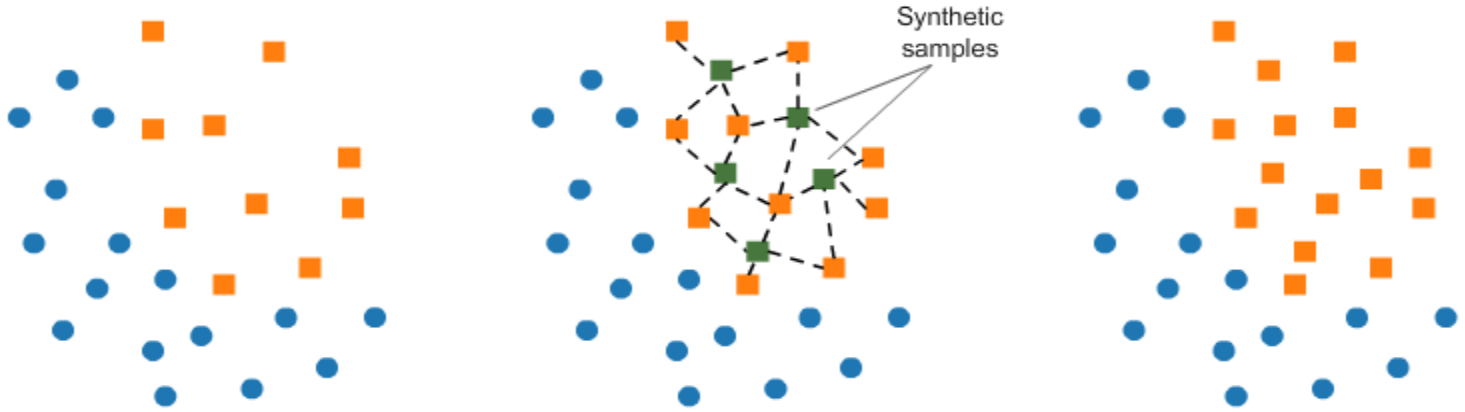


DATA-LEVEL SOLUTIONS



SYNTHETIC CASES

- **Synthetic Over-saMpling TEchnique (SMOTE)**
  - Chawla et e. (2002)
  - State-of-the-art solution and widely popular
- Pre-specified criterion to generate synthetic examples from the minority class such that all classes are balanced
  - Interpolation approach using nearest neighbourhood approach



SMOTE VARIANTS

SMOTE ISSUES

- Over-lapping cases and small disjuncts
- Noisy examples
- True underlying minority case distribution
- Over-generalization; linear distance measure etc.

SMOTE VARIANTS

- Initial selection of instances to over-sample e.g. borderline SMOTE
- Type of interpolation e.g. clustering-based, density-based, etc.
- Dimensionality reduction e.g. manifold learning, kernels etc.
- Integration with informed under-sampling
  - Filtering noisy examples
  - Data cleaning techniques

RESULTS

- Trained a logistic regression to predict mortality using Python
- AUC is the most common metric i.e. closer to 1, the better
- Class imbalance issue is problematic - biased & worst performance
- SMOTE generally improves the evaluation metric significantly
- Integrating SMOTE with an informed under-sampling often improves the results

DATA	AUC
Original Data	56,87%
SMOTE	75,72%
Borderline SMOTE	77,54%
SMOTE with under-sampling	78,96%

ALGORITHMIC LEVEL

PENALIZED MODELS/ENSEMBLES

- **Adjust the algorithm**
  - E.g. decision threshold, adjust cost function, penalised scores
- **Ensemble approaches**
  - E.g. bagging and boosting
- **Cost-sensitive learning**
  - Adjust evaluation metric for mis-classification cost

ONE-CLASS CLASSIFIERS

- Train on majority cases and anything above a threshold is flagged as a minority case
- Phrase as an anomaly/novelty detection
  - Autoencoders are gaining traction
  - Support Vector Machine (SVM)

GENERATIVE MODELS

- SMOTE is based on the feature space rather than the data space of the minority class as a whole.
- Generative models can learn implicitly joint pdf of input and labels simultaneously unlike discriminative models.
  - Understand underlying input structure even without labels
  - New samples, image-to-image translation, text-to-image synthesis, celeb faces, face aging, photo editing, missing data, video generation, animation creation etc.
- Generative Adversarial Nets (GAN) (Goodfellow et al. 2014)

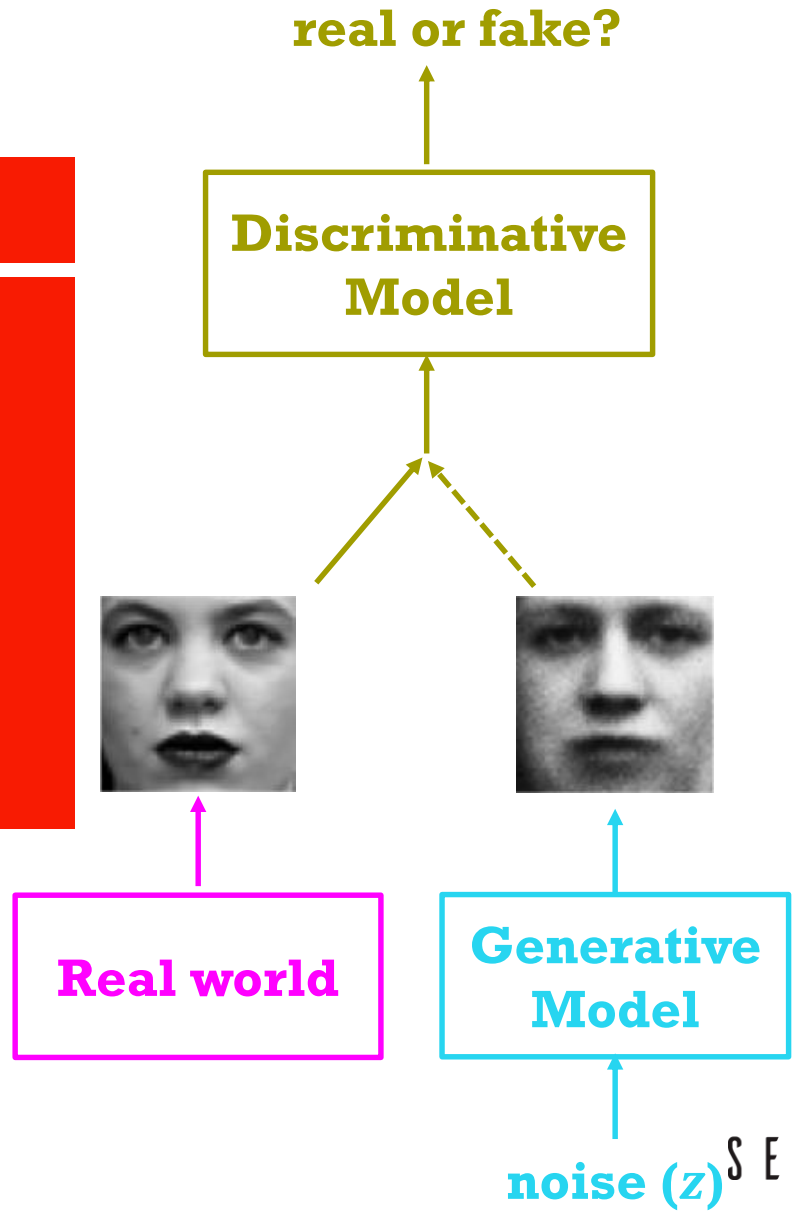
"…GANS and the variations…the most interesting idea in the last 10 years in ML."

- Director of AI Research for Facebook , Yann LeCun

- ***WHAT IF WE USE THIS IDEA TO GENERATE SYNTHETIC AND NEW SAMPLES OF THE MINORITY CLASS?***

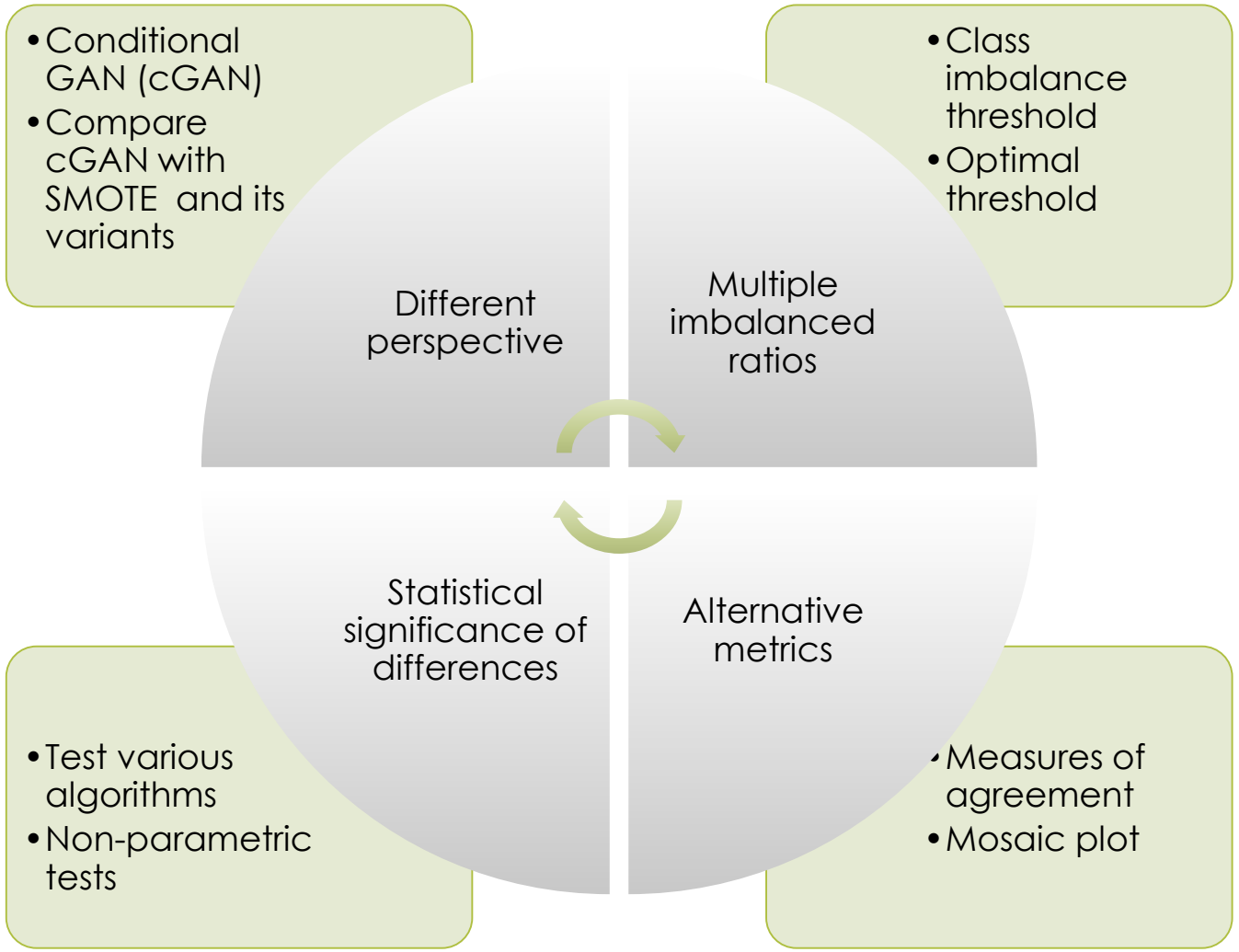
"What I do not understand, I cannot create." Richard Feynman, Nobel Prize Physics in 1965

BASIC IDEA OF GAN



- Generator(G) tries to fool Discriminator(D)
  - G generates new instances from a latent space
- D tries not to be fooled
  - D evaluates instances for autheniticity
- Models are trained simultaneously
  - As G gets better, D has a more challenging task
  - As D gets better, G has a more challenging task
- Ultimately, we don't care about the D
  - Ultimately, we don't care about the D
- Two deep neural networks at play in a zero-sum game

NEXT STEPS


• Conditional GAN (cGAN)
• Compare cGAN with SMOTE and its variants
Different perspective
• Class imbalance threshold
• Optimal threshold
Multiple imbalanced ratios
Statistical significance of differences
• Test various algorithms
• Non-parametric tests
Alternative metrics
• Measures of agreement
• Mosaic plot

CONCLUSION

- Actuarial practise still errs on uncertainty margins which are subjective
  - Class imbalance approaches may offer a different perspective to relook current practises and get creative
- Techniques exist but no consensus on what's best
- Combination of SMOTE + informed under-sampling often works well
- GAN offers a different perspective to sample from the true underlying minority class distribution
- A new algorithmic approach using one-class classifiers such as autoencoders or anomaly detection approach is gaining traction
- Techniques can be extended to regression and time series, e.g. extremes

REFERENCES

- **SMOTE**: *Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P.. 2002. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.*
- **SMOTE VARIANTS:** *Fernandez, A., Garcia, S., Herrera, F., & Chawla, N.V.. 2018. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research, 61, 863-905.*
- **INFORMED UNDER-SAMPLING:** *Batista, G.E., Prati, R.C., & Monard, M.C.. 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20-29.*
- **ALGORITHMS**: *Ganganwar, V.. 2012. An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering, 2(4), 42–47.*
- **SURVEY OF TECHNIQUES:** *He, H., & Garcia, E.A. 2008. Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering, 1263-1284.*
- **GAN:** *Goodfellow, I., Pouget-Abadie, J., Mirza, Mehdi, X., Bing, W., David, O., Sherjil, C., Aaron, & Bengio, Y.. 2014. Generative adversarial nets. Pages 2672–2680 of: Advances in neural information processing systems.*
- **GAN TUTORIAL:** Ian *Goodfellow's NIPS 2016 Tutorial, https://sites.google.com/view/cvpr2018tutorialongans/*
- **ENSEMBLES**: *Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F.. 2012. A review on ensembles for the class imbalance problem: bagging, boosting, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4), 463-484.*
- **MORTALITY:** *Sakr, S., Elshawi, R., Ahmed, A.M., Qureshi, W.T., Brawner, C.A., Keteyian, S.J., Blaha,M.J., & Al-Mallah, M.H.. 2017. Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project. BMC medical informatics and decision making, 17(1), 174.*
- **PACKAGES:** *Python (imblearn & smote_variants); R (smotefamily and unbalanced).*

kwanda.ngwenduna@vodacom.co.za

rendani.mbuvha@wits.ac.za

"…in the process of training generative models, we will endow the computer with the understanding of the world and what is made up of."

OpenAI