Using *R* and *RapidMiner Auto Model* to rapidly and reliably choose a great red from 40,000 *Kaggle* wine review texts.

Dr John Gwinyai (Gwin) NYAKUENGAMA
DatAnalytics
Email: DatAnalytics@iinet.com.au
Webpage: https://dat-analytics.net/

KEY WORDS
Kaggle wine reviews; *R*; *word2vec, h2o* routine; *RapidMiner Auto Model*; Automatic Feature Engineering; Supervised Machine Learning Models; Naive Bayes; Generalized Linear Model; Logistic Regression; Deep Learning; Random Forest; Gradient Boosted Trees; Support Vector Machine; Model performance; Receiver Operator Curve; Confusion Matrix

ACKNOWLEDGEMENT
We thank for materials used in this blog:

- *Kaggle (2018)* for their wine reviews; and
- *FreeGreatPicture (2019)* for the image in title.

INTRODUCTION
Choosing a good red for that all-important dinner using wine review texts has never been easier, faster or more reliable! Not only is choosing a dodgy red a money-waster, but a meal spoiler as well.

This blog show-cases a very fast machine learning (*ML*) tool for avoiding the choice of a bad red wine, without even tasting a single drop! This has to be better than relying on one's knowledge or a mate's "recommendation" of a great red wine – both of which are as unreliable as they are unscientific.

METHOD

We accessed some 40,000-odd, red wines on *Kaggle* wine reviews from *Kaggle (2018)* and processed their text comments using the *Word2Vec, h2o* routine in the *R* programming language. We saved 100 features from the process which represented the vectorised text characteristics of the *Kaggle* wines.

The features were subsequently used to choose a good wine using *RapidMiner Auto Model*, a high-end, modelling tool which employs automatic feature engineering. We evaluated seven *ML* models, following Nyakuengama (2018).

At the end of the *ML* process, we had the complete *Kaggle* wine reviews tagged as good or bad. We ignored the average wines. Not only were we able to rapidly choose the top 10 good wines, but we confidently avoided the dodgy ones! *RapidMiner Auto Model* allowed us to see the confidence level upon which the decision to pick those great red wines, and avoid the dodgy ones, was made!

The added bonus is that we were able to save our *RapidMiner Auto Model* models for repeated use until next year when we will update the models with the new wine-list and generate another wine ranking / classification, literally in seconds.

Please feel free to email the author (DatAnalytics@iinet.net.au) for technical details regarding the *word2vec* routine in *R Studio* or the *RapidMiner Auto Model* experiments.

RESULTS

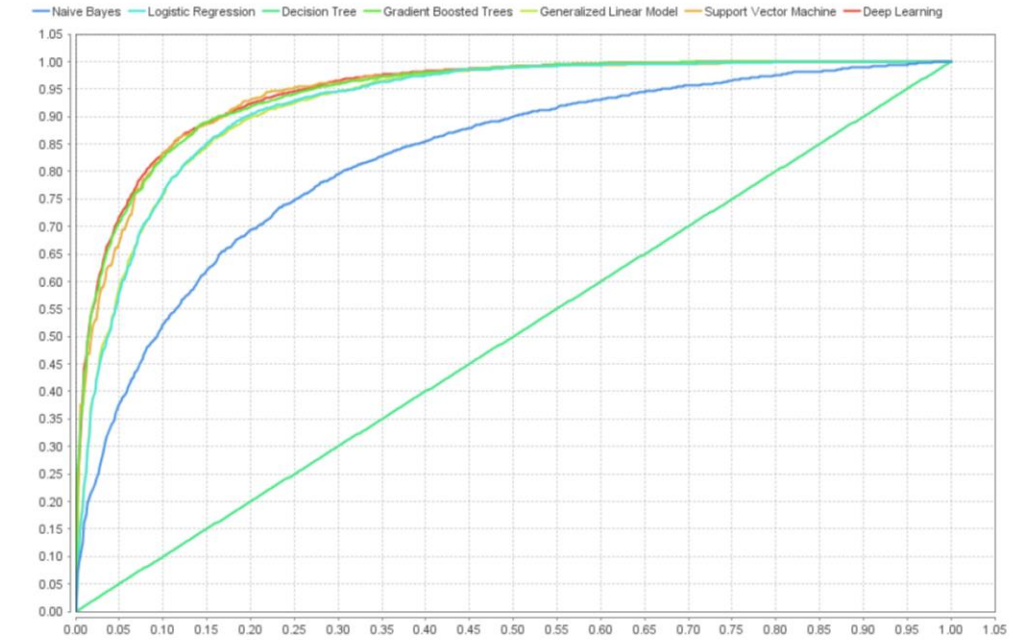
Table 1 suggests that in our *RapidMiner Auto Model* experiments, the best models were *Gradient Boosted Trees*, *Support Vector Machine* and *Deep Learning* while the *Decision Tree* model was the worst – based on accuracy, precision and area under the curve (auc).

Table 1: *RapidMiner Auto Model* performance parameters (in percentages).

Model	parameter		
	accuracy	precision	auc
Naïve Bayes	71.6	56.5	82.2
Generalised Linear Model	85.3	78.8	92.5
Logistic Regression	85.3	79.9	92.5
Deep Learning	87.5	80.4	94.5
Decision Tree	65.1		50.0
Gradient Boosted Trees	87.4	82.8	94.3
Support Vector Machine	87.4	82.2	94.0

The *Receiver Operator Curve* below (Figure 1), also suggests that the best *RapidMiner Auto Model*s were *Gradient Boosted Trees*, *Support Vector Machine* and *Deep Learning* while the *Decision Tree* model was the worst.

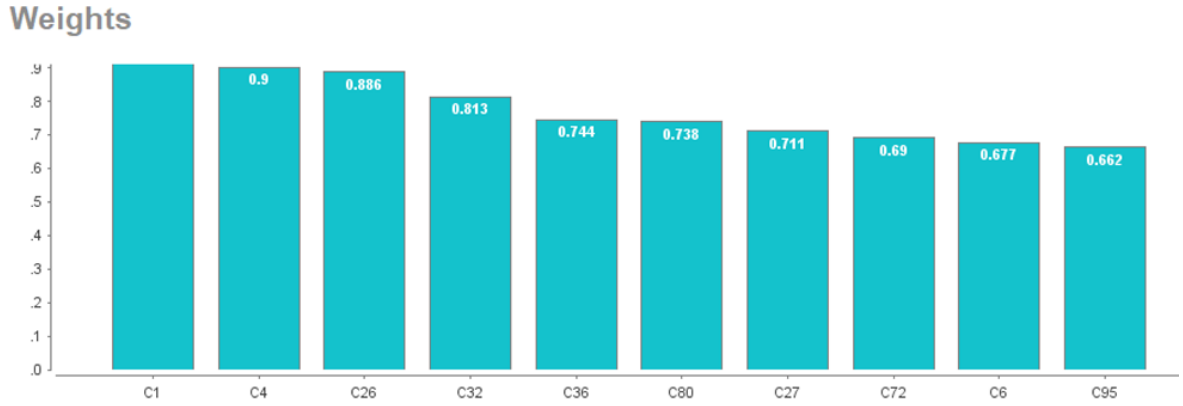
Figure 1: Receiver Operator Curve (ROC) from *RapidMiner Auto Model.*



INFLUENCIAL FEATURES

The most influential features from the *Kaggle* wine review text are shown below in Figure 2. Features importance (weights) ranged from 0.662 to 1.000.

Figure 2: Feature weights from *RapidMiner Auto Model*.



The red wines we care about!

A list of the top 10 good red wines predicted by the *Gradient Boosted Trees* model in *RapidMiner* is shown below (Table 3). Also, shown below are the prediction probabilities of a wine label.

Table 3: *Crème de la crème*, 10 red wines and prediction confidence levels.

RowNb	label	prediction(label)	confidence(bad)	confidence(good)
16030	great	great	0.002	0.998
15697	great	great	0.003	0.997
15960	great	great	0.003	0.997
15915	great	great	0.003	0.997
15961	great	great	0.004	0.996
14075	great	great	0.004	0.996
15722	great	great	0.004	0.996
14644	great	great	0.004	0.996
15695	great	great	0.004	0.996
14806	great	great	0.004	0.996

The real identities of the wines have been suppressed for commercial reasons.

Dodgy red wines we should avoid

A list of 10 dodgy, bottom of the barrel, red wines predicted by the *Gradient Boosted Trees* model in *RapidMiner Auto Model* is shown below (Table 4). Also, shown below are the prediction probabilities of a wine label.

Table 4: Bottom of the barrel, 10 red wines and prediction confidence levels.

RowNb	label	prediction(label)	confidence(bad)	confidence(good)
3212	bad	bad	1.000	0.000
3882	bad	bad	1.000	0.000
378	bad	bad	1.000	0.000
4453	bad	bad	1.000	0.000
476	bad	bad	1.000	0.000
296	bad	bad	0.999	0.001
80	bad	bad	0.999	0.001
614	bad	bad	0.999	0.001
10219	bad	bad	0.999	0.001
5873	bad	bad	0.999	0.001

The real identities of the wines have been suppressed for commercial reasons.

NEXT BLOG

Next time we will investigate if there is any relationship between the listed prices and wine quality using the *Kaggle* wine reviews, so please stay tuned!

BIBLIOGRAPHY

FreeGreatPicture (2018) red wine image: https://www.freegreatpicture.com/drinks/red-wine-52861

Kaggle (2018): Wine Reviews. https://www.kaggle.com/zynicide/wine-reviews/home

Nyakuengama (2018): Use of *RapidMiner* - Auto Model To Predict Customer Churn. https://dat-analytics.net/2018/07/28/use-of-rapidminer-auto-model-to-predict-customer-churn/

RapidMiner Studio: https://rapidminer.com/products/studio/