Visualising Logistic Regression

Eoin Kenny
16206131

1 Proposed Hypothesis

The question this paper asks is twofold, "What is the best aid to communicate what a logistic regression model has learned, and why it predicted what it did?"

Classification has always been of the utmost importance in business intelligence, with the recent advent of vast datasets, inexpensive computational hardware and advances in artificial intelligence this responsibility has become increasingly shifted from human workers towards machine learning algorithms. Common applications of this which apply to logistic regression are insurance fraud, banking loans, and any decision-making process which is binary in nature.

It is now possible to process limitless amounts of these aforementioned decisions with modern technology, but the person who may be subject in these decisions is still entitled to a detailed breakdown of why they may or may not have been (for example) accepted on their house loan or life insurance policy, which is why it is important to communicated the decision making process to people with perhaps no formal education and certainly no understanding of the underlying statistical processes involved in a clear, and concise manner. Basel II's recommendations on banking laws has now made this issue increasingly mandatory for financial companies in particular [1].

Some solutions in the literature exist to aid in this problem, but as Kosara [2] has stated the problem of model visualisation is largely underexplored in visualisation and he references studies in relation to this by Harrison & Yang [3] and Kay & Heer [4]. Both studies attempt to find out if the correlation of nine commonly used visualisations can be modelled using Weber's Law, but the focus of this paper will be on explaining why such correlations happen, rather than simply showing it.

More relevant research has been done by Royston & Altman [5], they believe that graphical aids are important in understanding the logistic model, and that whilst the receiver-operating characteristic (ROC) curve is familiar, it is not necessarily easy to interpret. Their solution to this problem was to use a simple histogram or dot plot of the risk score in the outcome groups. Whilst I personally agree this solution is helpful, histograms are not easily understood by the population at large and could present problems.

Other attempts include using the "effects" package [6] by Fox et al which models with some typical values to make predictions, plugging in various combinations of independent values and getting predicted probabilities, which allows us to see how they change as we vary our independent variables. Plots like these are called "effect plots" [7].

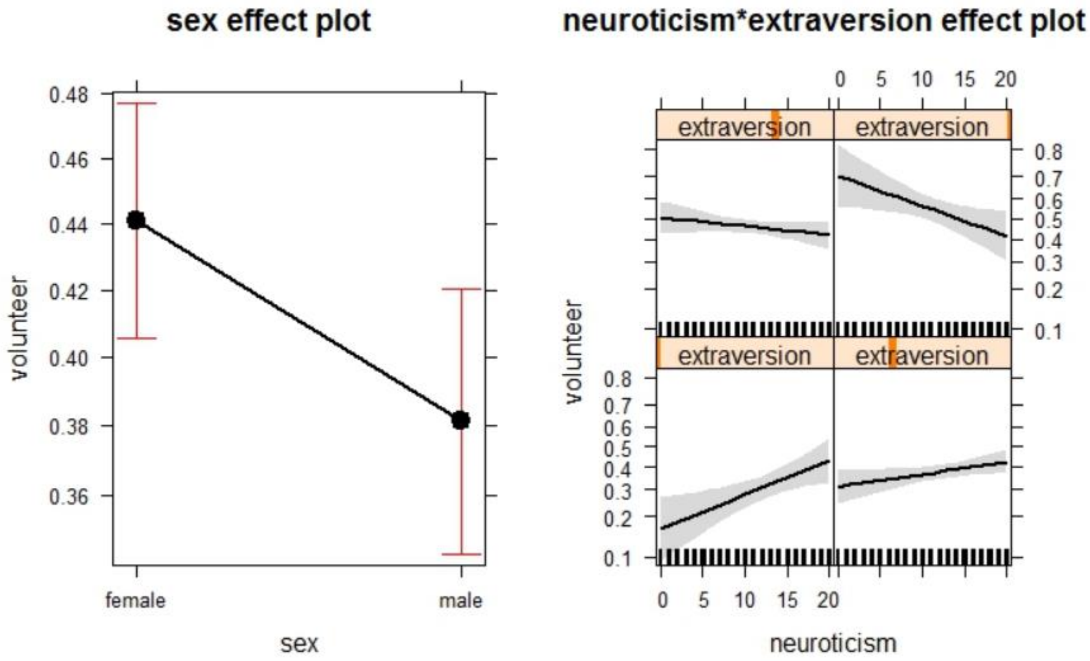


Figure 1: Effect plot of what genders are more likely to volunteer for psychological research

The left of *Figure 1* shows an "effect plot" of how women are more likely to volunteer for psychological research when compared to men. A plot like this is perhaps useful but does little to aid a user in easily determining what features contributed most to the final classification.

Perhaps the most popular method to convey a company/model's decisions is the use of scorecards. Each attribute is given points based on statistical analyses, taking into consideration numerous factors such as the predictive strength of the characteristics and correlation between characteristics. The total score of an applicant is the sum of the scores for each attribute present in the scorecard for that applicant. If the score raises above a certain threshold, then the model's binary decision changes.

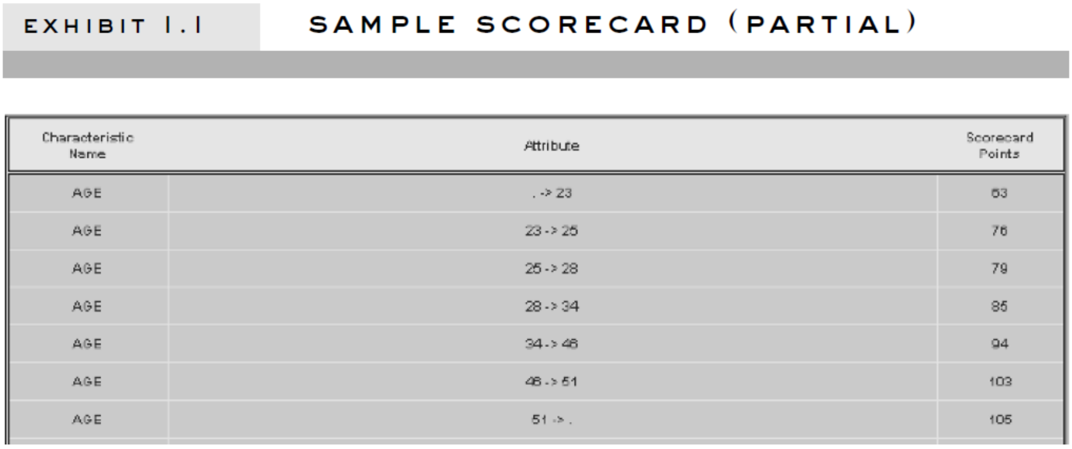


EXHIBIT 1.1 SAMPLE SCORECARD (PARTIAL)

Characteristic Name	Attribute	Scorecard Points
AGE	. -> 23	63
AGE	23 -> 25	76
AGE	25 -> 28	79
AGE	28 -> 34	85
AGE	34 -> 46	94
AGE	46 -> 51	103
AGE	51 -> .	105

Figure 2: Scorecard for Credit Scoring

Figure 2 shows a typical "Scorecard" [8], whilst these can convey well what a model has learned and why it makes decisions they have no visual element and lack potential engagement with human sensory abilities.

As machine learning models begin to take over sensitive decisions in everyday life, it has become increasingly important to understand why they make the decisions they do. Insurance and banking issues aside, some judicial models have even been discovered to be inherently racist at predicting reoffenders [9]. Issues and concerns such as these could perhaps be more easily identified if what the model had learned and why it makes the decisions it does were more easily communicated with effective visuals (the previous examples shown are lacking). The area of artificial intelligence is currently underexplored in information visualisation and represents a gap which warrants research.

2 Experimental Method

2.1 Overview

The experiment shall be performed in two rounds, both happening one after the other in the same conditions with no break. The first round shall attempt to answer the question "How best to visualise what a logistic regression has learned?", and the second "How best to communicate why a logistic regression made a certain classification?"

In round one people will be shown fifteen different visualisations of models using three methods, a bar chart, a bubble chart and a treemap (five each). People will be asked to rank the top three significant feature predictors in order, and provide a guess of the model's decision when presented with a set of feature values (an instance). Since continuous features may have a range of values, the mean of the dataset will be shown to compare its ranking against the binary ones.

In round two they will be given twenty questions which focus again on the previous three visualisations but with scorecards as a forth aid (henceforth, "aid" refers to visuals/scorecards), giving five questions for each again. People will be asked to rank the top three features which contributed most to the model's decision and try guess what that decision is.

Round one's independent variables are the top three feature rankings for each question, the time taken to answer, and the user's guess of classification for a given instance. The dependent variables are the precision/recall for all above, the exception is time because a simple measure of total time taken for each type of visualisation will be sufficient.

For round two the independent variables are the user's ranking of the top three features which caused the model's decision, their guess for the model's actual prediction and the time taken. The dependent variables are again total time (for each type of aid), and precision/recall for rankings and guesses.

As participants are going through each example in both rounds they will have the opportunity to learn a model's characteristics to memory, which will skew later questions. To account for this, a different model from a different dataset must be used for each question which will vary the visualisations and scorecards. To compare results fairly, all questions must have an evenly distributed level of difficulty.

The experimental conditions will simply be each participant answering these questions on a computer, asked to do so as fast and accurately as possible.

The overall design of the experiment focuses on trying to ascertain the fastest, easiest and most accurate way to communicate a model's training and decision-making process. The first round is particularly for developers/academics to help in the model design phase, the second round is for explaining to clients why a model made a certain decision and compares the popular method of scorecards against visually centred approaches (although clients may also wish to see what a model has learned, i.e. round one).

2.2 Data collection

During round one the data recorded for each question will be the ranking of the model's three most dominant features chosen by the user, the guess for the classification of a given sample instance, and the time taken to answer. The ranking data will allow us to see which visualisation allows people to understand the model's weights, features and possible range of values best. The classification measurement will allow us to see how well people can predict the model's predictive behaviour without the visual aids made available in round two. Time will give us a way to gauge accuracy against speed for all types of visualisations.

During round two the data recorded will be the user's prediction of the classification, the top three features which contributed to it, and the total time. The top three features will show us if the aids can effectively communicate the major reasons behind the model's decision, the guess of classification will be a further measure of how clear the classification is (without textual assistance, although this may be given by a financial institution etc. if desired), and again time will give us a way to gauge accuracy and speed for all types of questions.

There will be no subjective measurements in this experiment, all the data is highly objective.

My first question was how to understand visually what a logistic regression has "learned", to do this we can use its weights, bias and features and the first round allows us to do this in three separate ways and compare them objectively. Round one also allows us to see how well that understanding can be applied mentally to new instances presented, which is a further test of the visual's communicative power. My second question was to find the best way to help people understand why the model has made decisions it has. To do this we can examine the total contribution of each feature after it is multiplied by its respective weight alongside the bias, round two allows us to do this with four different aids and compare them objectively again.

2.3 Selected subjects

The subjects used in this experiment will fall into two camps, professionals in a quantitative field and the lay person with an average understanding of statistical concepts. Both demographics will answer all thirty-five questions.

Since round one of the experiment is directed towards helping professionals better refine predictive models, half of the subjects will be from industry and familiar with concepts such as bias and weights. Round two is intended for the lay person to clearly and swiftly understand why a model made a certain classification, so these must represent random samples from the population. For comparison, both demographics will participate in both rounds.

The industry subjects will be sourced from any Irish analytics company willing to participate. To make such a company more interested, promising them first access to the research results for business use would be appropriate. To gain a random sample of the general population Mechanical Turk would be both suitable and fast.

For statistically significant results, thirty subjects from both demographics will be needed.

2.4 Data analysis

To compare the three visuals in round one and gauge the most suitable for said task the precision and recall shall be used to measure how accurately people can rank the model's three most sensitive features. The totals for these will be averaged across each visual's five examples, the time taken will also be averaged. Finally, the total correct answers for guessing each classification of sample instances will be totalled over five for each visual. These measures will allow us to see which visual aid allows people to understand a model most accurately, and which communicates it fastest. Since difficulty is evenly distributed across questions, an average will suffice for meaningful results.

For round two precision and recall shall measure all data here, except for time, which again will be averaged across all five examples for each aid. The aim here is to help a lay person understand why a machine learning algorithm behaved as it did, if the average person from Mechanical Turk can use these aids to accurately and quickly rank the top three influential features in the prediction then they are effective at doing so. To further test the aids, information about the actual prediction shall be withheld and volunteered by the participant, this will be a further test if the person truly understands how to interpret the visuals and scorecards (for example, is it easier to understand area on a treemap or a bubble chart?).

2.5 Practical setup

The experiment can be run from any computer the participants have access to in their own time. The entire assessment will be online, there is no offline element to the experiment.

The participants will be instructed to allow thirty minutes to complete the experiment, it must be at a time when they will be undisturbed and free from any possible distractions. Participants must also complete the experiment alone without any outside aid. They should also be advised that questions are to be completed as fast and as accurately as possible, although there is no time limit for completing the questions.

For continuous features in round one, participants should be advised to take the mean value as the feature's influence when ranking them against categorical ones, which are all binary in nature. These mean values are shown as dividing lines in the visualisations which attempt to make it as clear as possible.

Participants using laptops are advised to ensure their respective machine is fully charged or plugged in with the capability to stay on for at least thirty minutes.

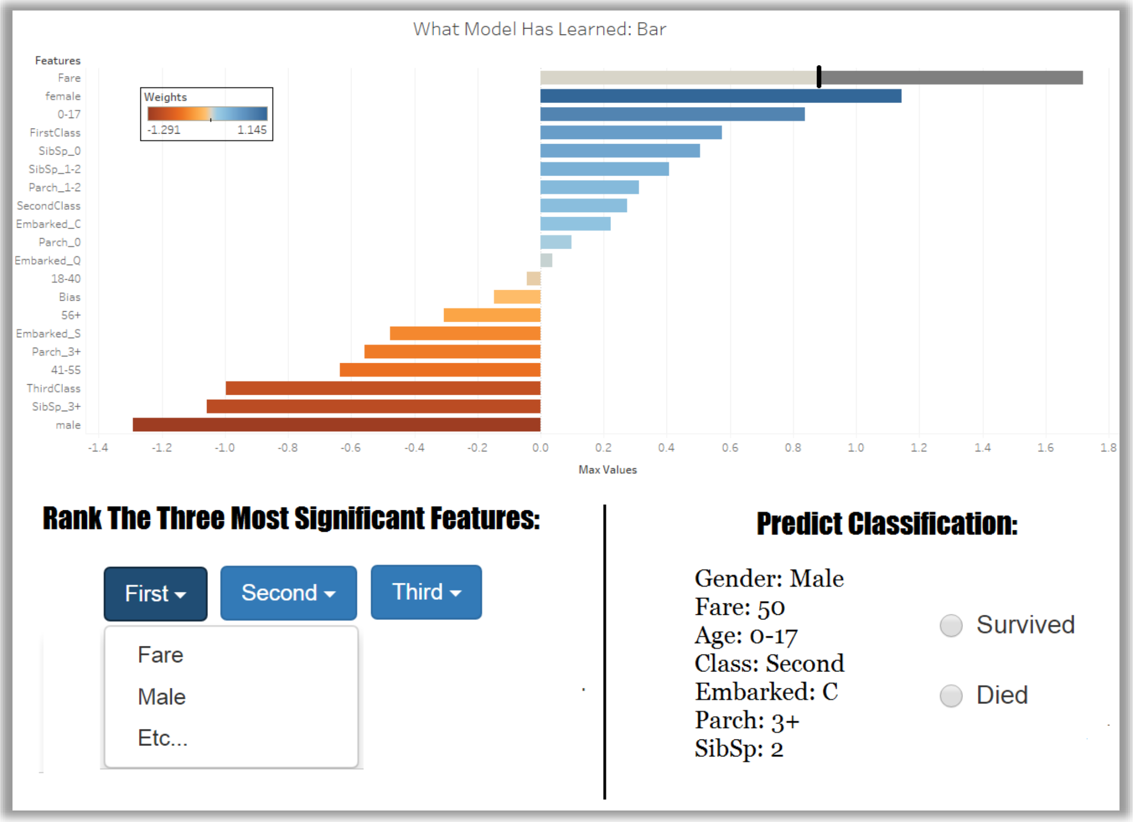
The experiment can take place in any room at any time provided the participant will not be distracted or disturbed and has internet access.

All visualisations will appear on the computer screen, there is no printed paper element to the experiment.

3 Data Visualisations

3.1 Round One Visualisations

3.1.1 Bar Chart



Rank The Three Most Significant Features:

Predict Classification:

Gender: Male
Fare: 50
Age: 0-17
Class: Second
Embarked: C
Parch: 3+
SibSp: 2

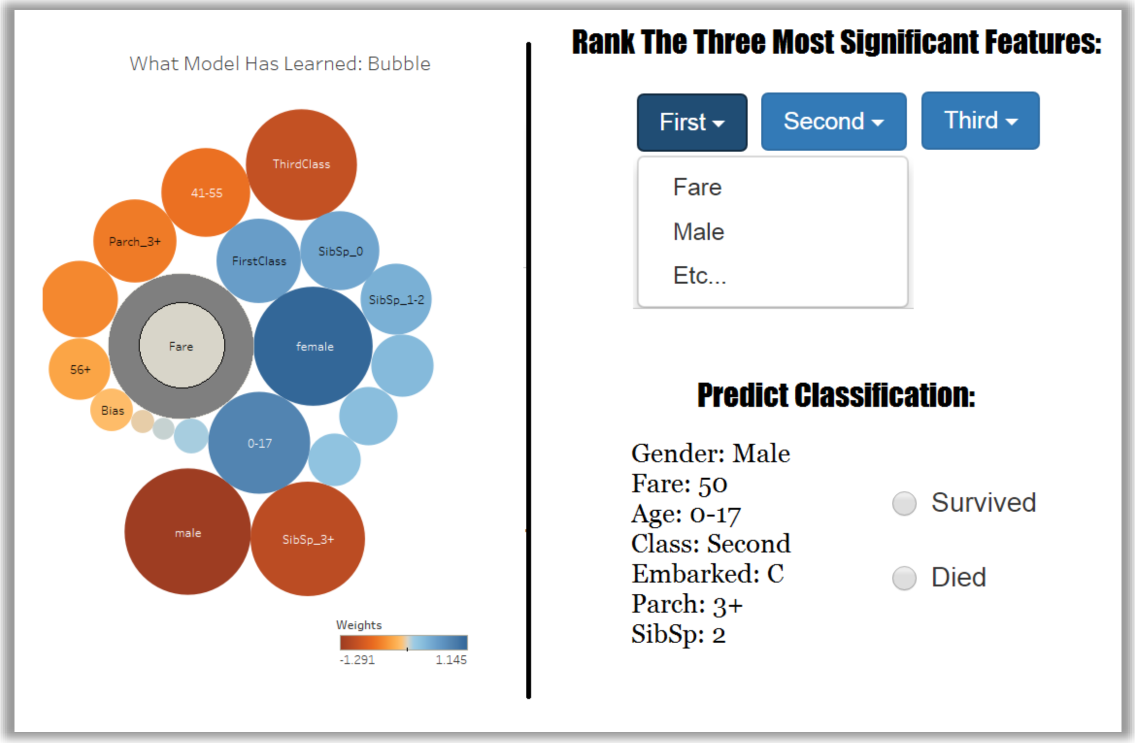
○ Survived

○ Died

Figure 3: Bar Chart for Round One of Experiment

The mean of the feature *Range* can be seen dividing the bar above, this is to be taken as the features significance when ranking them in order.

3.1.2 Bubble Chart



Rank The Three Most Significant Features:

Predict Classification:

Gender: Male
Fare: 50
Age: 0-17
Class: Second
Embarked: C
Parch: 3+
SibSp: 2

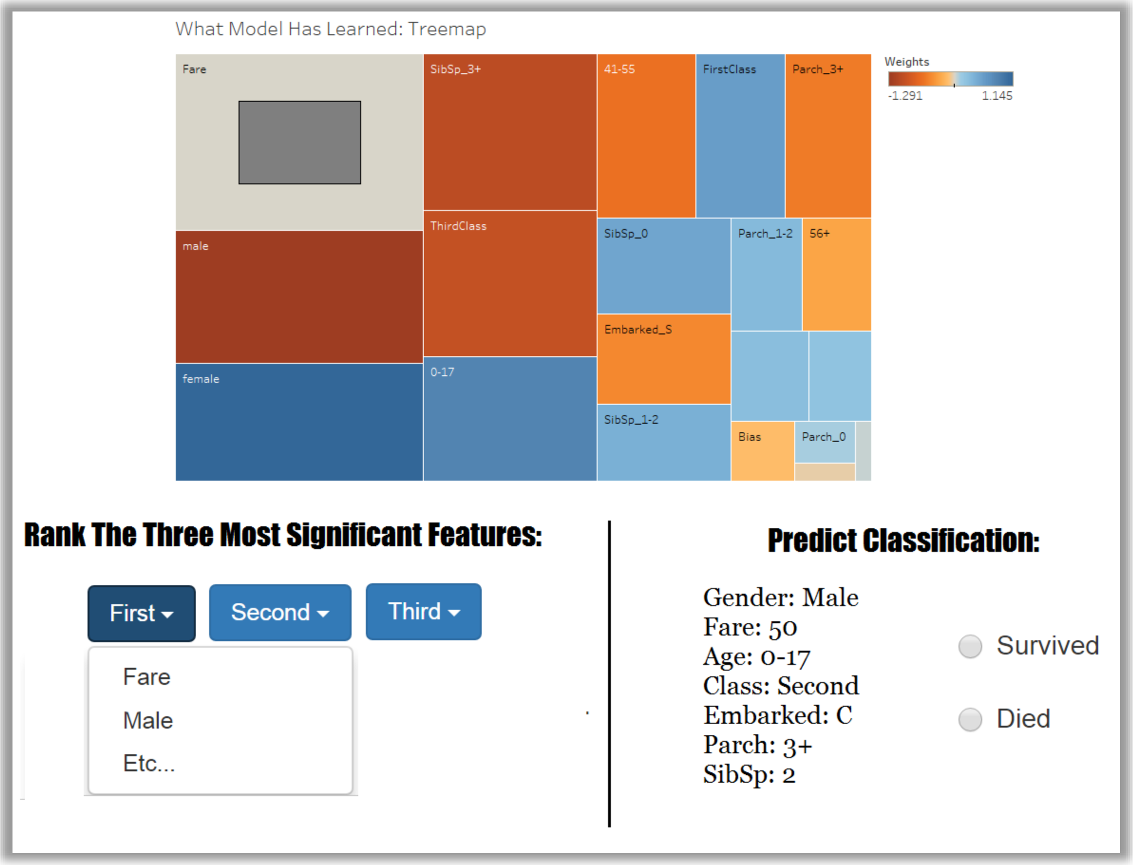
○ Survived

○ Died

Figure 4: Bubble Question for Round One

Similarly, the feature *Fare* has its mean shown as a sub-radius of the bubble.

3.1.3 Treemap


What Model Has Learned: Treemap

Rank The Three Most Significant Features:

Predict Classification:

Gender: Male
Fare: 50
Age: 0-17
Class: Second
Embarked: C
Parch: 3+
SibSp: 2

○ Survived

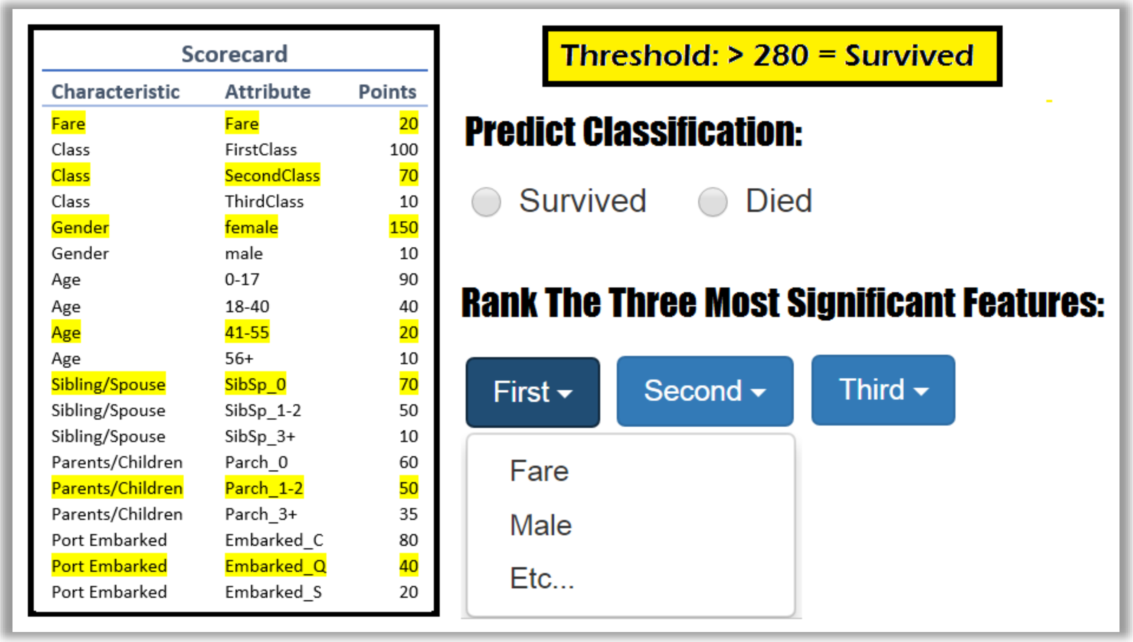
○ Died

Figure 5: The Treemap for Round One

Finally, the feature *Fare* is also divided at its mean in the treemap by a smaller rectangle.

3.2 Round Two Visualisations

3.2.1 Scorecard: This Example Survives



Scorecard

Characteristic	Attribute	Points
Fare	Fare	20
Class	FirstClass	100
Class	SecondClass	70
Class	ThirdClass	10
Gender	female	150
Gender	male	10
Age	0-17	90
Age	18-40	40
Age	41-55	20
Age	56+	10
Sibling/Spouse	SibSp_0	70
Sibling/Spouse	SibSp_1-2	50
Sibling/Spouse	SibSp_3+	10
Parents/Children	Parch_0	60
Parents/Children	Parch_1-2	50
Parents/Children	Parch_3+	35
Port Embarked	Embarked_C	80
Port Embarked	Embarked_Q	40
Port Embarked	Embarked_S	20

Threshold: > 280 = Survived

Predict Classification:

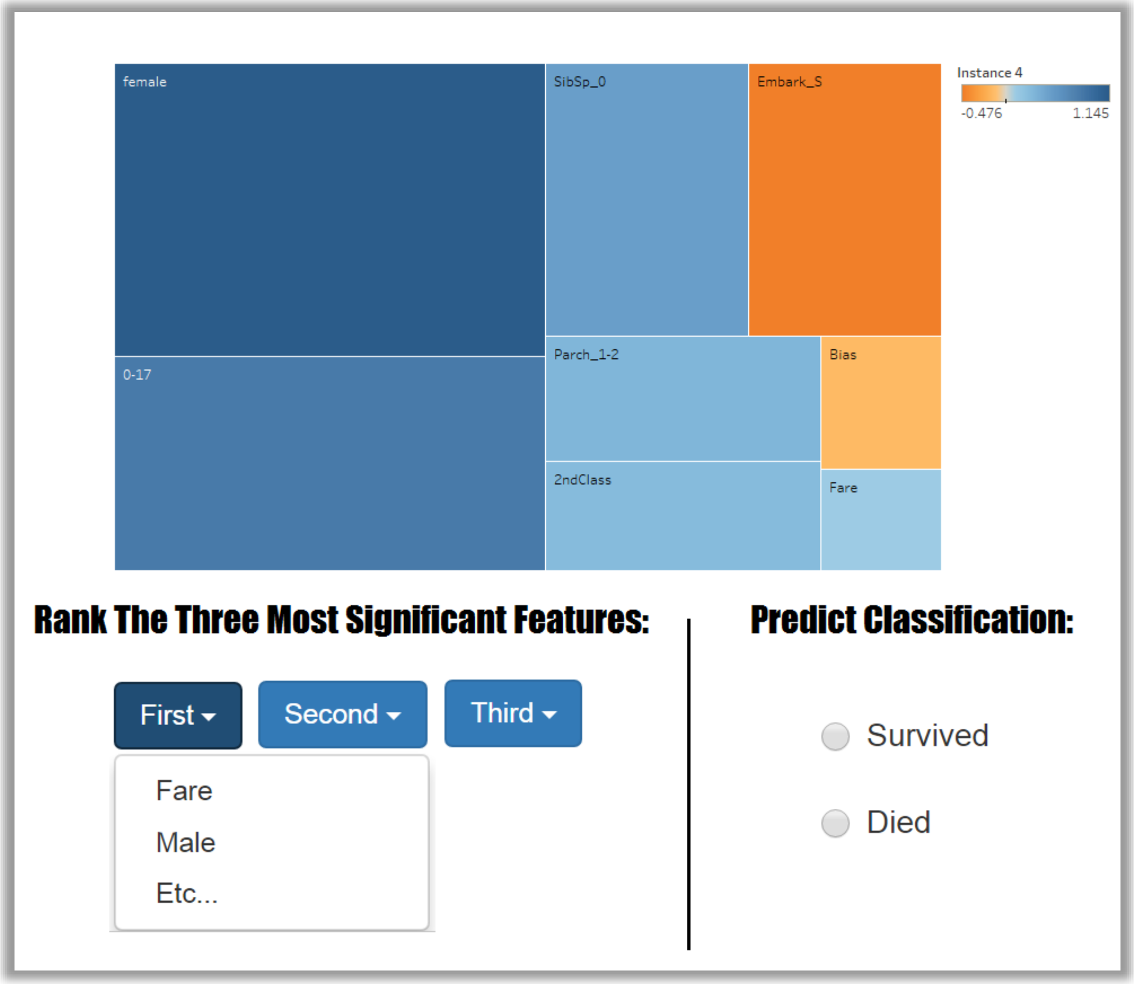
○ Survived ○ Died

Rank The Three Most Significant Features:

Figure 6: Scorecard Aid for Round Two

The scorecard aid for round two has the features to be tallied highlighted to fairly compare it to the other three visuals aids which naturally have them highlighted. The threshold is also shown, which represents the points needed to be classified as “Survived” rather than “Died”.

3.2.2 Treemap: Example 1 (likely to survive)

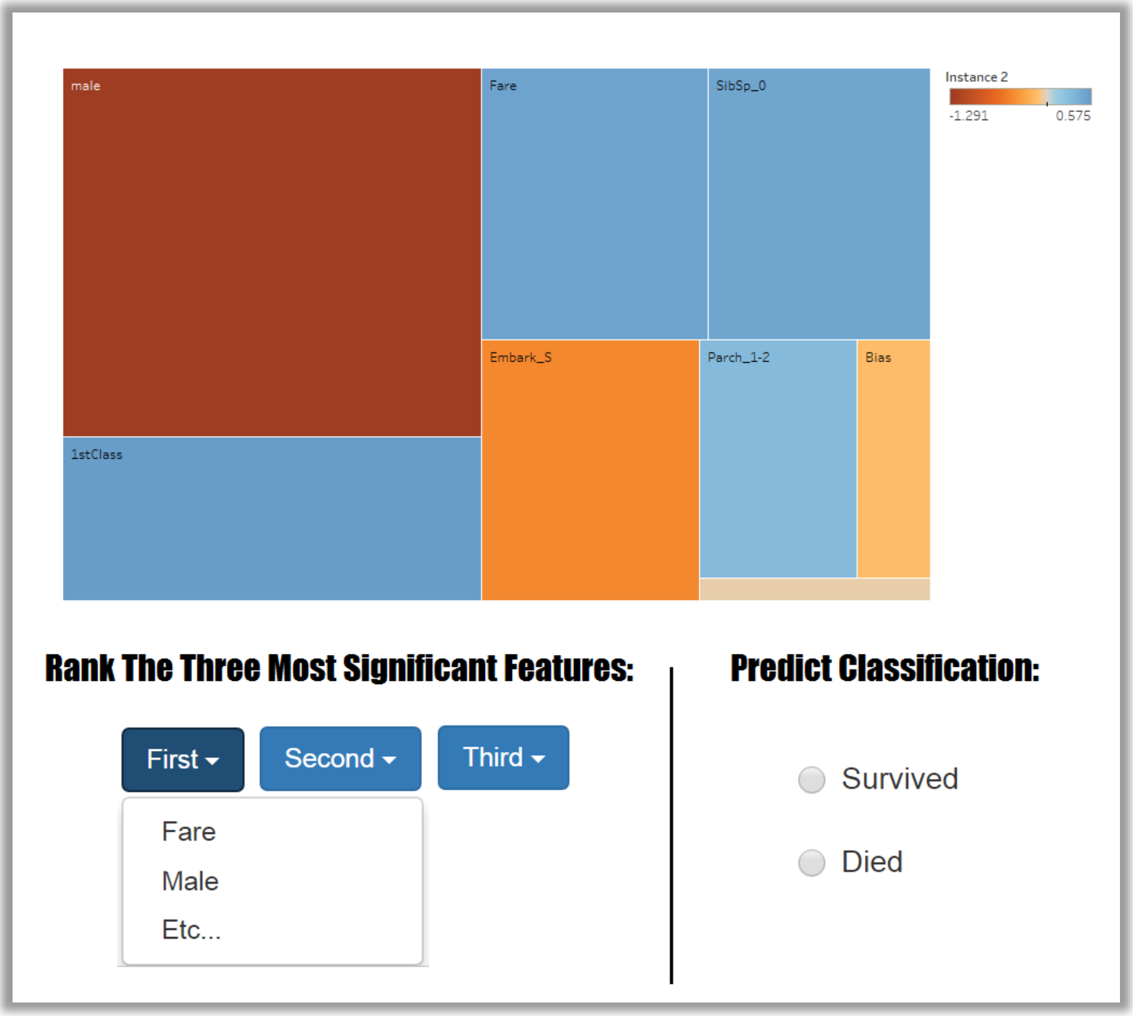


Rank The Three Most Significant Features:

Predict Classification:

Figure 7: An Instance Likely to Survive

3.2.3 Treemap: Example Two (not likely to survive)



Rank The Three Most Significant Features:

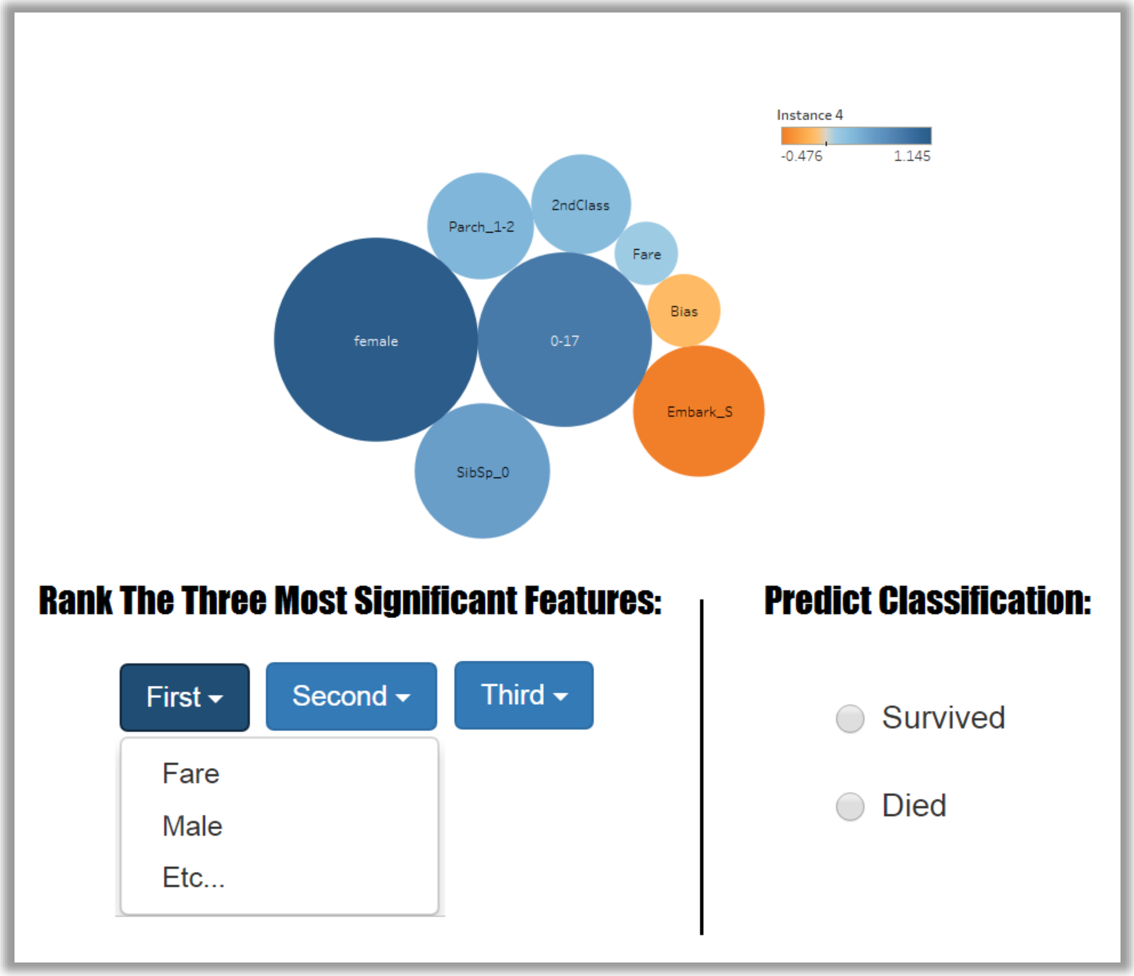
Predict Classification:

○ Survived

○ Died

Figure 8: An Instance Not Likely to Survive

3.2.4 Bubble Chart: Example 1 (likely to survive)



Rank The Three Most Significant Features:

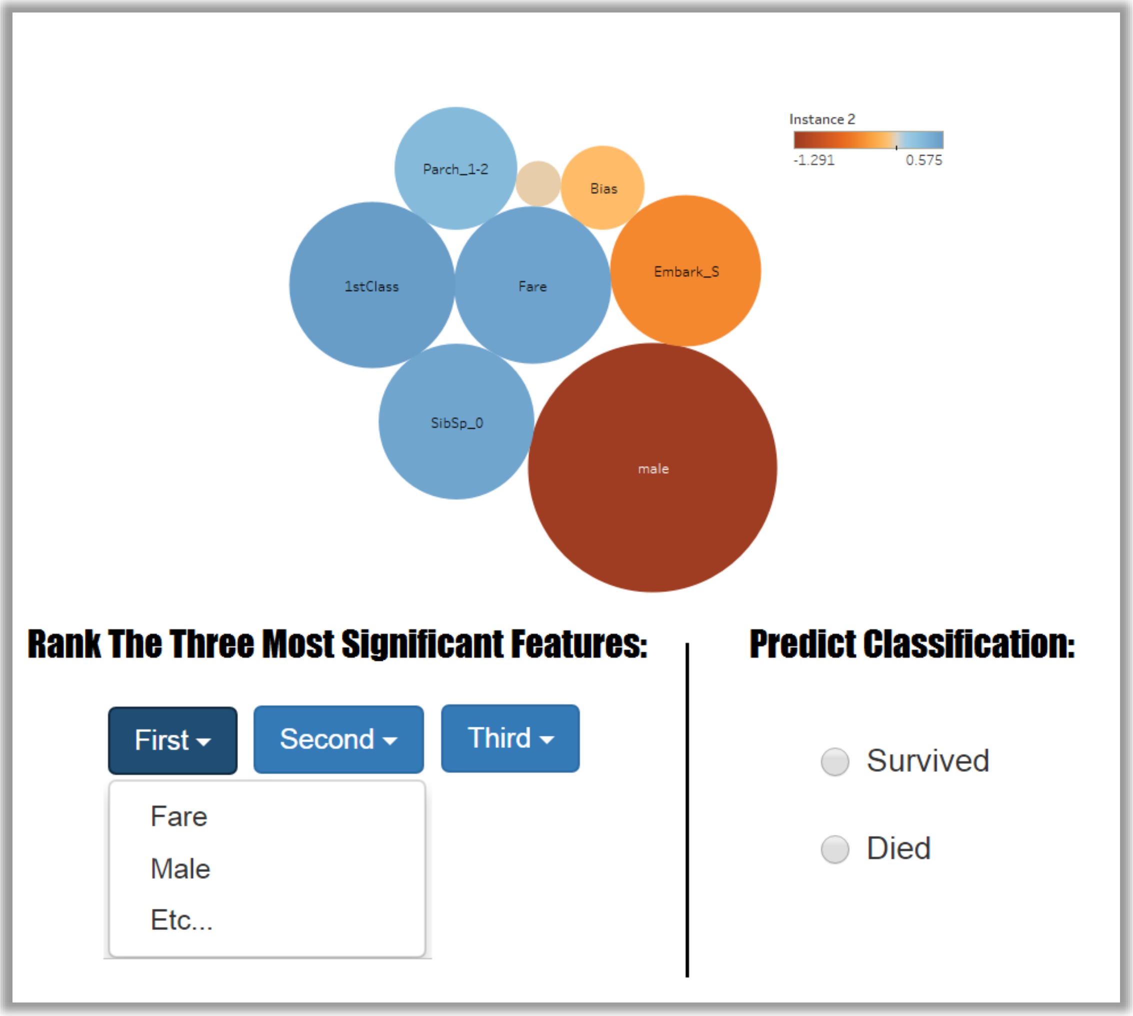
Predict Classification:

○ Survived

○ Died

Figure 9: An Instance Likely to Survive

3.2.5 Bubble Chart: Example 2 (not likely to survive)



Rank The Three Most Significant Features:

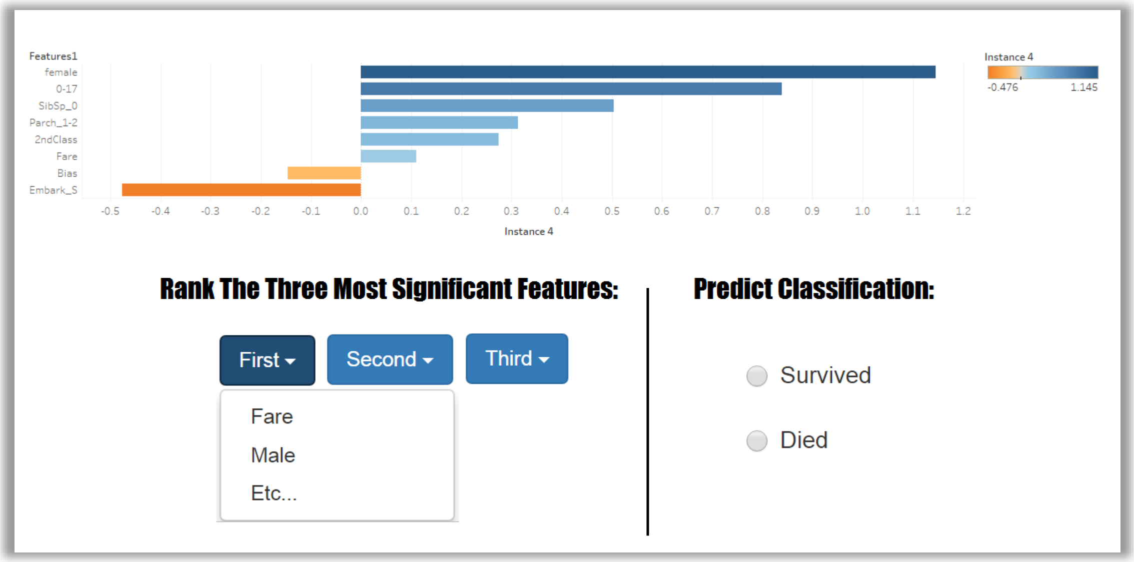
Predict Classification:

○ Survived

○ Died

Figure 10: An Instance Not Likely to Survive

3.2.6 Bar Chart: Example 1 (likely to survive)



Rank The Three Most Significant Features:

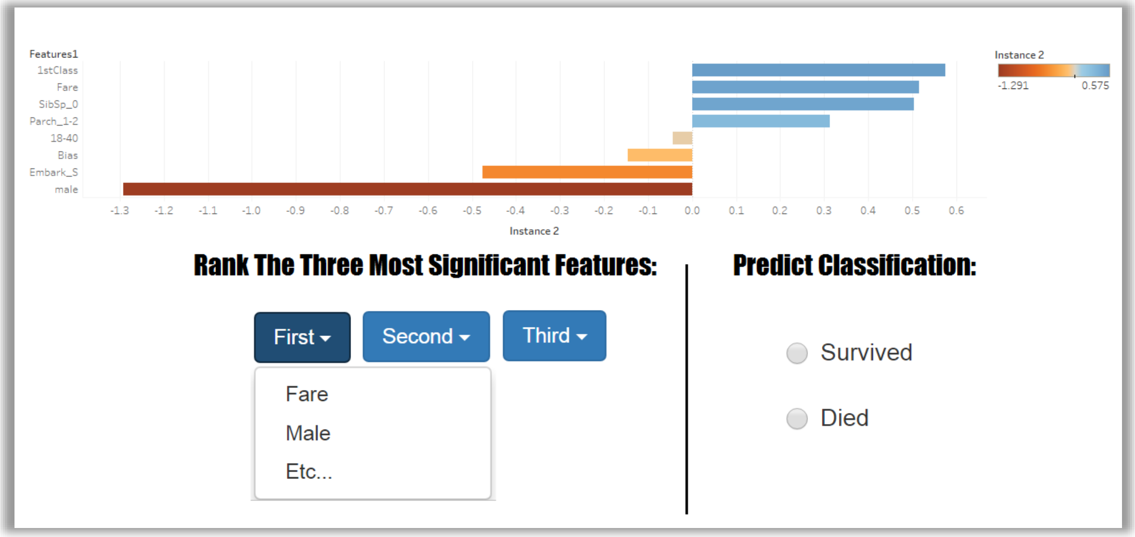
Predict Classification:

○ Survived

○ Died

Figure 11: An Instance Likely to Survive

3.2.7 Bar Chart: Example 2 (not likely to survive)



Rank The Three Most Significant Features:

Predict Classification:

○ Survived

○ Died

Figure 12: An Instance Not Likely to Survive

References

[1] Basel Committee on Banking Supervision. (2006). *Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework - Comprehensive Version*, Bank for International Settlements. Retrieved at September 16, 2008, from the website temoa : Open Educational Resources (OER) Portal at http://www.temoa.info/node/7721

[2] Kosara, R., 2016, October. An Empire Built On Sand: Reexamining What We Think We Know About Visualization. In *BELIV* (pp. 162-168).

[3] Harrison, L., Yang, F., Franconeri, S. and Chang, R., 2014. Ranking visualizations of correlation using weber's law. *IEEE transactions on visualization and computer graphics*, *20*(12), pp.1943-1952.

[4] Kay, M. and Heer, J., 2016. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE transactions on visualization and computer graphics*, *22*(1), pp.469-478.

[5] Royston, P. and Altman, D.G., 2010. Visualizing and assessing discrimination in the logistic regression model. *Statistics in medicine*, *29*(24), pp.2508-2520.

[6] Fox, J., 2003. Effect displays in R for generalised linear models. *Journal of statistical software*, *8*(15), pp.1-27.

[7] Ford, C. (2017). *Visualizing the Effects of Logistic Regression | University of Virginia Library Research Data Services + Sciences*. [online] Data.library.virginia.edu. Available at: http://data.library.virginia.edu/visualizing-the-effects-of-logistic-regression/ [Accessed 10 Nov. 2017].

[8] Siddiqi, N., 2012. *Credit risk scorecards: developing and implementing intelligent credit scoring* (Vol. 3). John Wiley & Sons.

[9] Lim, K. (2017). *The racist, fascist, xenophobic, misogynistic, intelligent machine*. [online] The Business Times. Available at: http://www.businesstimes.com.sg/brunch/the-racist-fascist-xenophobic-misogynistic-intelligent-machine [Accessed 10 Nov. 2017].