SUICIDE RATE AROUND THE WORLD FROM 1985-2016

Angelina Kiman | ISM 6356 – Data Mining Analytics & Visualization | Winter 2019

My data is about suicide rates from 1985 and 2016. I obtained from www.kaggle.com and it is publicly available (https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016). The goal is to investigate the suicide rates that is greater than year 1999 and country that has GDP per capita below $20,000. Simply, are people more prone to commit suicide during a recession?

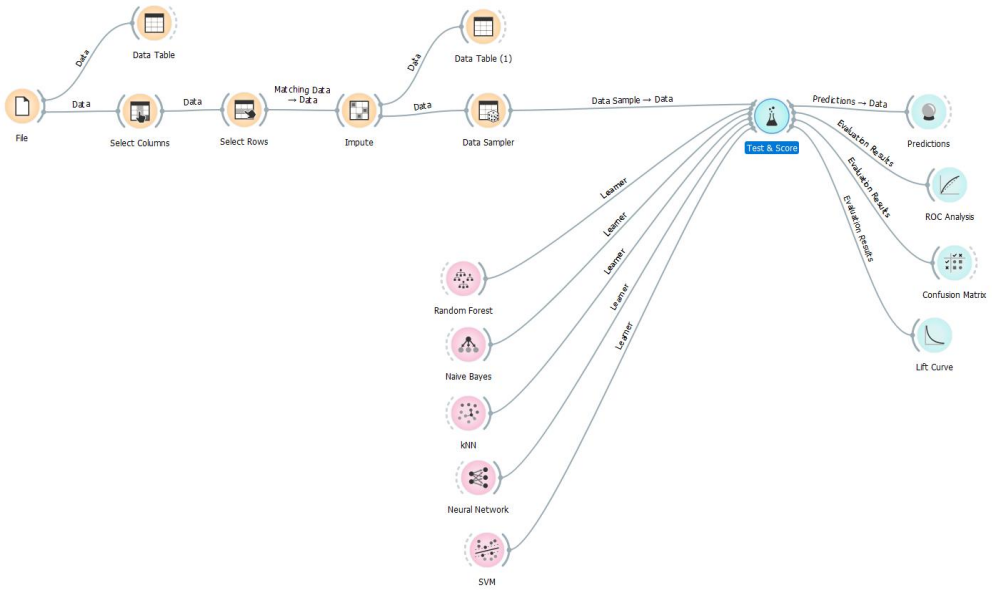


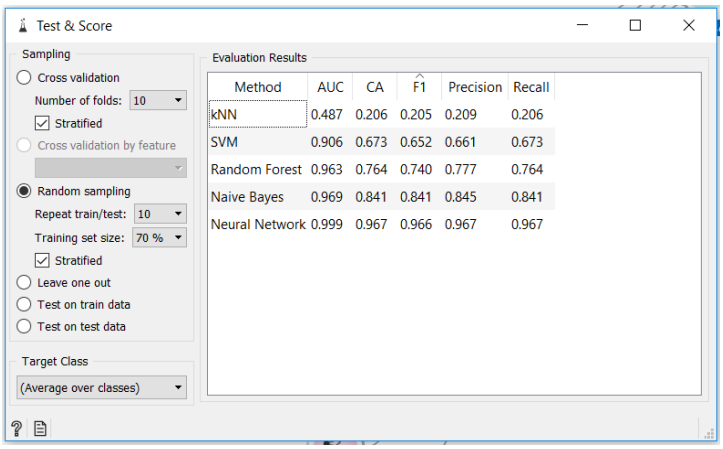
Figure 1.1. Suicide rates data mining application in Orange tools

There are two algorithms conditions that I choose: year and GPD per capita. As the world's economy progressively grows, I would like to know specifically year after 1999 and country that has GPD per capita less than $20,000, might have or have not higher suicide rates.

Since my dataset falls under logistic regression, I used five types of algorithms: random forest, naïve bayes, kNN, and neural network, and support vector machine (SVM). Random forest and naïve bayes are insensitive with specific hyper-parameters, do not require tweaking any parameters, and versatile. kNN suitable for large data as my dataset is more than 10,000. SVM and Neural Network are suitable for

large dataset as well and both are machine learning algorithms that may boosts the test accuracy results amongst other algorithms.

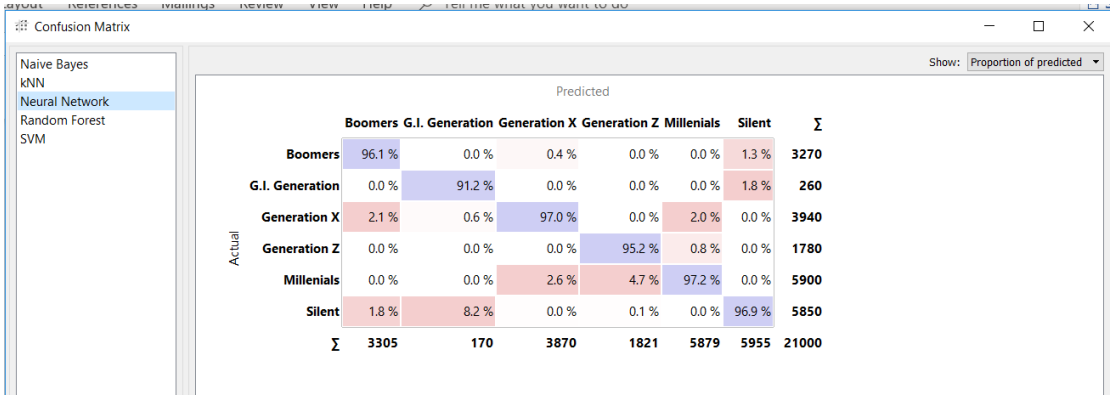
Additionally, I use predictions to gain insight which generation is likely to get suicide during recession era. I use ROC and Lift Curve to test the graph of goodness fit as well as to give visual aids for model performance.



Method	AUC	CA	F1	Precision	Recall
kNN	0.487	0.206	0.205	0.209	0.206
SVM	0.906	0.673	0.652	0.661	0.673
Random Forest	0.963	0.764	0.740	0.777	0.764
Naive Bayes	0.969	0.841	0.841	0.845	0.841
Neural Network	0.999	0.967	0.966	0.967	0.967

Figure 1.2. Test Score Result

Based on *Figure 1.2.*, the data shows that my thesis are supported. Neural network has the highest classification accuracy of 96.7%. It shows many people are decided to take out their lives during the recession. On the contrary, kNN's accuracy is the lowest one in comparison to others algorithm methods.



	Boomers	G.I. Generation	Generation X	Generation Z	Millenials	Silent	Σ
Boomers	96.1 %	0.0 %	0.4 %	0.0 %	0.0 %	1.3 %	3270
G.I. Generation	0.0 %	91.2 %	0.0 %	0.0 %	0.0 %	1.8 %	260
Generation X	2.1 %	0.6 %	97.0 %	0.0 %	2.0 %	0.0 %	3940
Generation Z	0.0 %	0.0 %	0.0 %	95.2 %	0.8 %	0.0 %	1780
Millenials	0.0 %	0.0 %	2.6 %	4.7 %	97.2 %	0.0 %	5900
Silent	1.8 %	8.2 %	0.0 %	0.1 %	0.0 %	96.9 %	5850
Σ	3305	170	3870	1821	5879	5955	21000

Figure 1.3. Confusion Matrix

Based on the *Figure 1.3.*, Millennials are prone to commit suicide and followed by the Generation X. G.I. Generation is the lowest one to commit suicide. Nevertheless, the number of percentages of suicide rates are sitting close with each other.

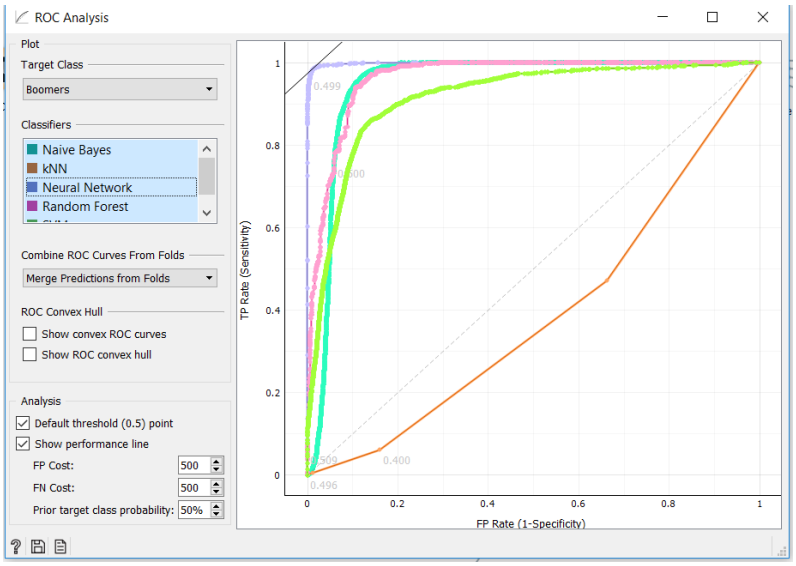


Figure 1.4. ROC Analysis

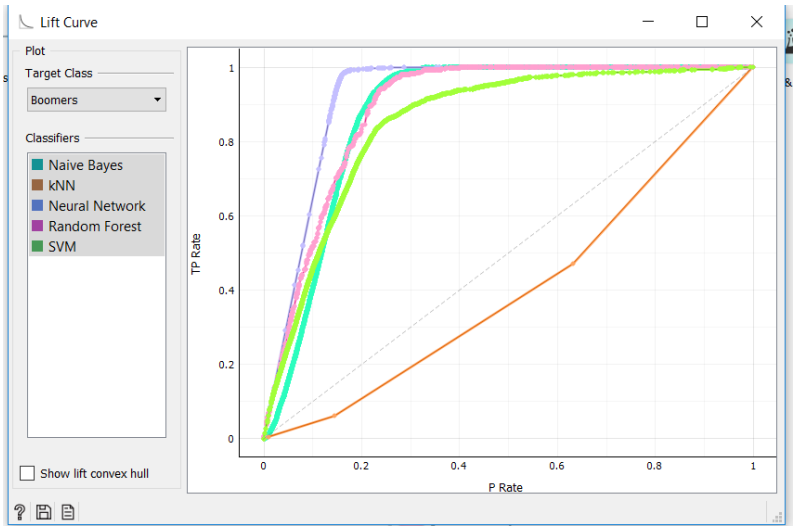


Figure 1.5. Lift Curve

Based on Figure 1.4. and Figure 1.5., the results are very similar. As we can see, the outsider is kNN (orange line). This result gives me confidence that my theory is supported. After running this model, I am curious whether the GDP per capita is the main factor that many people are suicide. If yes, what is the "safe number" for the GDP per capita that will minimize number of people are going for suicide.

What went well? I did a lot of trial-and-error when executing the dataset. I did not understand the concept in Orange tools at first; so, I decided to perform under RapidMiner to better grasp the ideas (thanks to Dr. LaBrie's advice!). I watched many tutorials on YouTube and did Google some articles to understand the evaluation results.

What did not go well? Things did not go as expected. Originally, I would like to use Petfinder.com data; but, for some reasons, I was not confident with the results and wondered if I executed properly. This data taught me a lot as it was more complicated; so, when I changed my dataset into suicide, I did not encounter many problems like I did previously. In addition, I was quite indecisive with tools that I wanted to use. I was in dilemma between using Orange and Alteryx. I watched many tutorials of both. I tried using both as well and at the very end, I noticed Alteryx required some "R" knowledge in which I did not know how to code. I go with Orange because it is similar with RapidMiner. I realized Orange have some different widgets connotations in comparison to RapidMiner, but still manageable to figure it out!

What would I do differently next time?

I would better plan about the project. I pushed my limit –getting out from the comfort zones. I knew the Petfinder.com dataset is more challenging than the suicide ones. I tried my best even though I failed. Besides, I should have been more decisive which data tools that I wanted to use; hence, I could deep dive better on that tools instead of both, Alteryx and Orange. I should have watched Week 5 lecture on time! To be honest, I watched it quite last minute and power outage happened. The Week 5 lecture specifically features selection, helped me to understand the evaluation results better.