**Data Analytics**

**Team Data Challenge**

Team 2:

Alex Pchelintsev

Don-Moses Chiemela

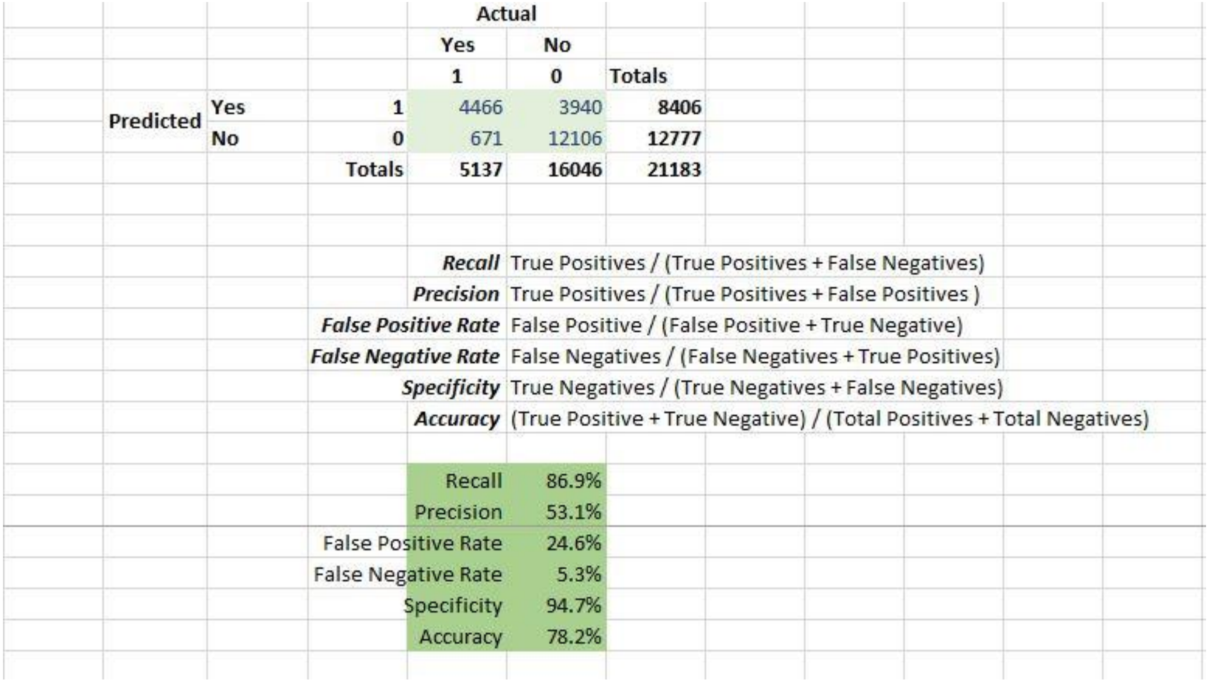
Mohamed Tibourtine

## Summary

We developed a model with an accuracy of 78%, recall of 87% and specificity of 95% and had a >2% change when the model was tested. Overall, it's a good model. Naïve Bayes classification system is a good system for entry data analytics for businesses. Preparing the data and adapting it to the excel template was a challenging task. We were able to make some groupings and dummies with the python script and in the future would use python for all data transformations because it is faster. The team was great, and everyone contributed.

## Model

Based off our first attempt we got a great model and decided not to try and fit it more to preserve the degrees of freedom. We had an accuracy of 78%, recall of 87% and specificity of 95% which is impressive. See figure below.



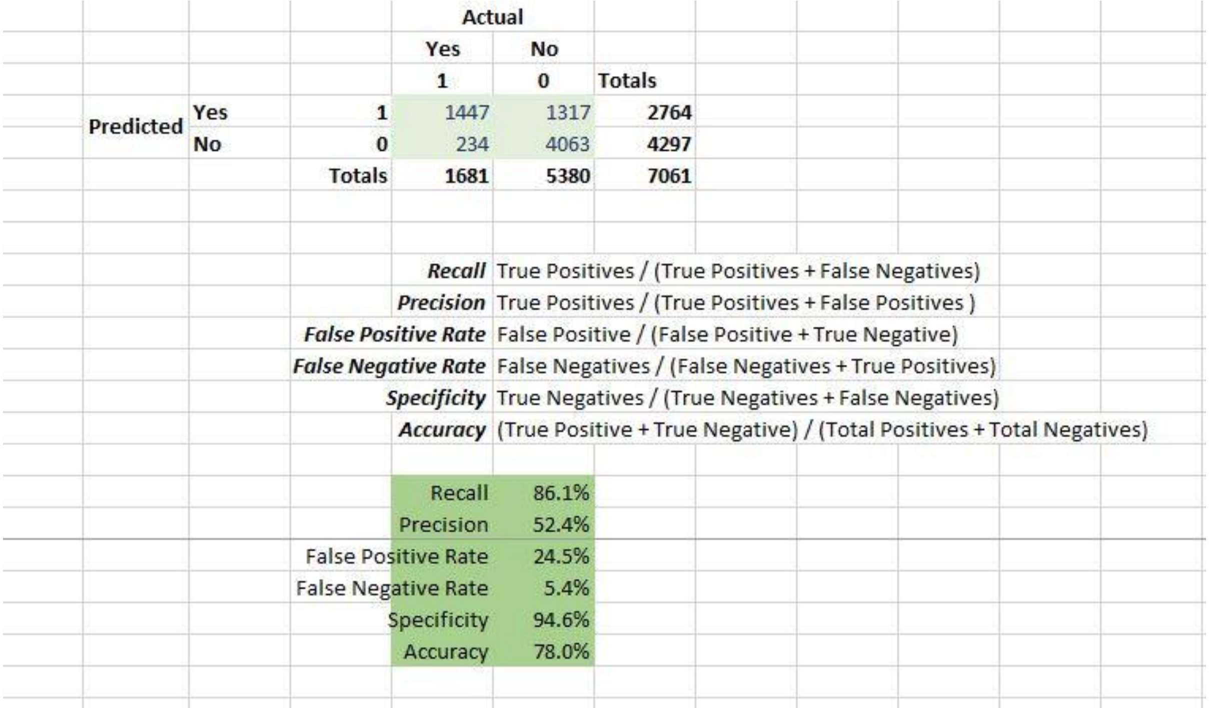
| | | | Actual | | |
|---|---|---|---|---|---|
| | | | Yes | No | |
| | | | 1 | 0 | Totals |
| Predicted | Yes | 1 | 4466 | 3940 | 8406 |
| | No | 0 | 671 | 12106 | 12777 |
| | | Totals | 5137 | 16046 | 21183 |

| | |
|---|---|
| ***Recall*** | True Positives / (True Positives + False Negatives) |
| ***Precision*** | True Positives / (True Positives + False Positives ) |
| ***False Positive Rate*** | False Positive / (False Positive + True Negative) |
| ***False Negative Rate*** | False Negatives / (False Negatives + True Positives) |
| ***Specificity*** | True Negatives / (True Negatives + False Negatives) |
| ***Accuracy*** | (True Positive + True Negative) / (Total Positives + Total Negatives) |

| | |
|---|---|
| Recall | 86.9% |
| Precision | 53.1% |
| False Positive Rate | 24.6% |
| False Negative Rate | 5.3% |
| Specificity | 94.7% |
| Accuracy | 78.2% |

Fig 1: Trained model stats

We ran a test dataset produced from the python script through the model and obtained the results below.



| | | | Actual | | |
|---|---|---|---|---|---|
| | | | Yes | No | |
| | | | 1 | 0 | Totals |
| Predicted | Yes | 1 | 1447 | 1317 | 2764 |
| | No | 0 | 234 | 4063 | 4297 |
| | | Totals | 1681 | 5380 | 7061 |

| | |
|---|---|
| *Recall* | True Positives / (True Positives + False Negatives) |
| *Precision* | True Positives / (True Positives + False Positives ) |
| *False Positive Rate* | False Positive / (False Positive + True Negative) |
| *False Negative Rate* | False Negatives / (False Negatives + True Positives) |
| *Specificity* | True Negatives / (True Negatives + False Negatives) |
| *Accuracy* | (True Positive + True Negative) / (Total Positives + Total Negatives) |

| | |
|---|---|
| Recall | 86.1% |
| Precision | 52.4% |
| False Positive Rate | 24.5% |
| False Negative Rate | 5.4% |
| Specificity | 94.6% |
| Accuracy | 78.0% |

We observed just a little change in the recall, accuracy and specificity. Overall, there was less than 2% change and we would say the model performed excellently. But our model is not free from some inherent limitations of using a naïve bayes classification system. one limitation is that it assumes that the occurrence of a feature is completely independent of the occurrence of another feature. It doesn’t account for multicollinearity. It does not account for variables that were not present in the training dataset or had no occurrence.

A business would use this model to estimate the effectiveness of their different advertisement campaigns to determine when to focus their efforts. Also, it can be used to sort through leads to determine which should be followed up if they are predicted to have a chance to convert.

**Data prep**

Preparing the data and adapting it to the excel template was a challenging task. As young python programmers, we got stuck a lot of times on little bugs, did a lot of research to me able to move forward from each stage. We were able to make some groupings and dummies with the python script and in the future would prefer to complete all categorical data transformation in python as it is faster. We ran into issues understanding the excel template. But once it was understood, developing the model was easy.

The most challenging part, besides the deadlines, was having to merge the different data frames. The team has spent two hours to find out that to do a horizontal concatenation, you would have to add the axis=1 argument to the formula. It was interesting to work across different time zones. However, the real enjoyment comes when the whole code clicks together after finding what caused the glitch. Coding was majorly about 15 minutes of writing codes that work, making up for 95% of the code. And a few hours correcting the 5% that doesn't work.

**Team**

Out of five participants at the beginning of the course, almost immediately, two dropped out and only three are now presenting this work. As a result, the load on each remaining fighter nearly doubled. Despite this circumstance, the participants showed outstanding endurance and mutual assistance in learning and performing every tasks. In constant stress from the extreme schedule of combining two courses and a massive amount of information, the team members, who were close to hysteria, did not lose their enthusiasm and sense of humor, supporting each other again and again in difficult moments. Alex oversaw the write up and excel, Don worked on the python script and excel, Mohammed worked on the python script and write-ups. Then the team member who didn’t work on a deliverable did the final proofing with fresh eyes. In the end strong people come out of the most challenging situations with dignity especially in the middle of the night with a time difference from Boston up to 9 hours.

**Course**

Despite the high quality of educational materials and excellent teachers, the course presented real challenges in terms of self-study and information retrieval and work with its processing and analysis. Yes, we have googled and a lot of resources to help, but, as it was said in the textbook, the language of programmers is quite different from the language of ordinary people, so increasing the level of computer literacy is critical. Of course, we are all very far from the professional level, but some understanding of work principles and the search for solutions has developed.

The team is strong in diversity and mutual reinforcement - synergy! Alex - coordination, humor, moral and psychological support, and low-level technical work. Don and Mohamed - programming wizards. It was an awesome experience.