

## Model name

Potential Customer Classification

## Model date and version

Built February 2022

No numbered versioning. The model was trained in February 2022 and has not been updated since.

All parties registered as a user of this model will be informed by email if we release an update.

## Overview Model type

Pre-trained categorisation model

This model categorises potential customers into two categories: likely to convert to customers and unlikely to convert to customers.

Potential Customer Classification uses DBScan to classify customer engagement on our website and social media, to facilitate identifying engagement patterns most likely to convert to a purchase. It was trained on a data set of 20,000 customer records from our web analytics data, dating from January 2015 to January 2022. The data set was labelled automatically from those who converted to customers and appear in our orders database.

## Questions or comments

Please send any questions to:

[team@datascience\\_ioa.org](mailto:team@datascience_ioa.org)

## Primary intended users

The marketing team to identify priorities in their outreach strategy and social media campaigns.

The sales team may prioritise calls to potential customers based on their classification.

## Primary intended uses

To identify early which initial engagements on our social media site are most likely to result in conversion to a purchase and to use that information to inform decision making around marketing approaches.

## Out of scope uses

Potential Customer Classification is intended for low-stakes uses such as informing marketing decisions. It should not be used for decisions that have a real-world effect on the customer or potential customer such as pricing decisions.

## Limitations

The Potential Customer Classification algorithm only supports users of our English-based content and was not trained on international content.

The Potential Customer Classification is not updated in real time and therefore changes in consumer behaviour may not be reflected in the algorithm.

### The following factors can affect classifications:

Level of engagement on social media (accuracy of prediction for those with low engagement may be reduced)

Age (older customers feature less in our data set and their behaviour is less reflected in the results)

## Metrics

The model is optimised for recall over precision, meaning that it is optimised to ensure that the algorithm does not miss anyone likely to convert to a customer. Though this also means that a larger-than-usual number of people will be labelled as potential customers who will not eventually convert to customers.

## Training and evaluation data

Please speak to the data team if you would like to inspect the training data:

[team@datascience\\_ioa.org](mailto:team@datascience_ioa.org)

## Quantitative analysis

### Accuracy and reliability

The model has an accuracy level of 72%, meaning that it was able to place subjects into the correct category 72 times out of a hundred.

The data set was imbalanced, with only 9% of enquiries converting in the training data set resulting in a sale, therefore accuracy rates can be deceptive.

The training set of 20,000 customers should be adequate in size and given the number of parameters to give reliable results.

### Precision

Precision measures the accuracy of the positive (will convert to a customer) label. A low score, near to 50%, on precision means that there are a lot of false positives in the end results. This was rather low at 76%, meaning that around a quarter of those labelled by the machine as likely to convert did not, in fact, convert.

## Recall

Recall measures the ability of the algorithm to identify all potential positives (that no customer that converted was missed). A low score, near 50%, on recall means that there were a lot of false negatives in the end results. This was 93%.

These are within our benchmark criteria, but the limitations should be considered when making decisions.

## Ethical considerations

The intended use of this data set is for low-stakes sales decisions. No personally identifiable information was included in the train and test data sets. The role that any sensitive information (ethnicity, gender, religious affiliation etc.) may have played in the classification process was not investigated.

The model was accurate up until February 2022. Further work is needed to identify whether the algorithm still generalises to potential customers today. Generalisation is the concept of whether an algorithm that was trained on one specific data set works equally efficiently when used on a new data set.

The model does not capture information on potential customers with limited engagement with social media and the findings should not be generalised to that group.

## Feedback

All users are encouraged to share concerns or comments about the performance of the algorithm by completing the form below:

[ioaglobal.org/modelfeedback](https://ioaglobal.org/modelfeedback)

You should receive an email from the team within 48 hours with information on how your comments will be processed and any teams that it will be escalated to. We thank you for your support in developing our data processes.

## Additional notes and any other relevant factors

None