Target

The main aim of this report is to predict if a user of the Citizen Science web project will disengage from the current session within the next 5 minutes.

The Dataset

The data source for this report is a stream of HIT data from the Citizen Science web project. A user HIT consists of a user ID and a timestamp to indicate when a user made a contribution to a project. A user session is defined as a sequence of HITs for the same user where consecutive HITs are not separated by more than 30 minutes.

Firstly we organised the data into sessions as defined above. This would give us a good view of how a users HITs are spaced over a particular session. We then created additional columns for each HIT that would provide adequate information about a HIT and it's place in a user's session. We grouped the HITs by session and created three normalised fields namely:

Session_duration
The amount of time that has elapsed since the start of the current session

Idle_time
The amount of time that has elapsed since the last HIT in the current session

Total_hits
The number of HITs that have occured in the current session

These three fields then meant that each HIT could be evaluated individually and gives a comprehensive explanation of the session leading up to the HIT.

The final field we added was **will_churn** that marks a HIT if it is within 5 minutes of the end of the session.

After creating these fields we found that there was an imbalance in the classes where roughly 30% of the HITs were within 5 minutes of the end of a session. We would therefore have to ensure our models took this imbalance into account. To train we split the model 75/25 into train and test sets and first trained a logistic regression model and a random forest classifier. Finally, we balanced our training dataset and then modelled a neural network.

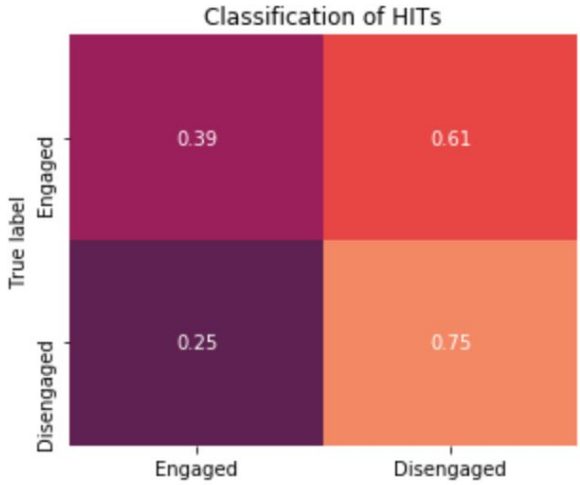
Training basic models

The problem we are trying to solve is a binary classification problem and the most important metric we want to evaluate is the False Negative Rate. This is because it is much more

damaging to miss a user who is about to disengage than it is to falsely mark a user as someone who is about to disengage.

Logistic Regression

Therefore the first model we tried was a simple logistic regression model. We fit the standard

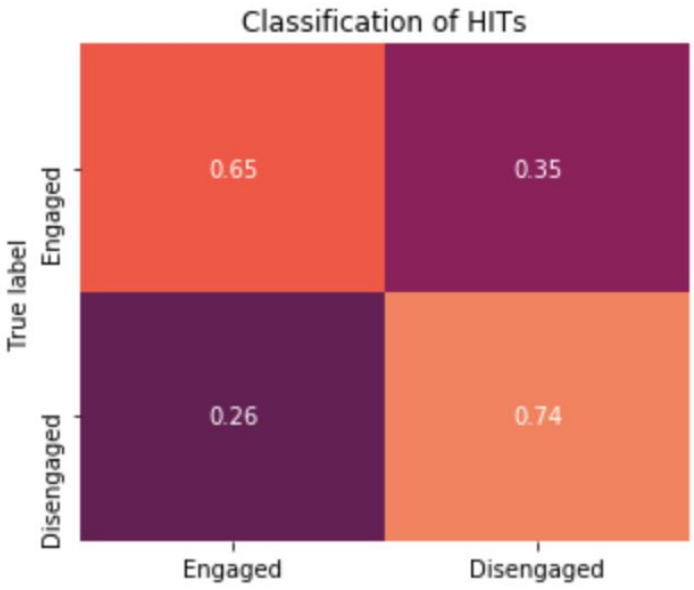


Logistic Regression model from SciKit to the data, being sure to use the “balanced” mode, which automatically adjusts the weights inversely proportional to class frequencies in the input data.

This was trained on the train data and produced the above confusion matrix.

Random Forest

The next model we tried was a random forest model. We fit the scikit random forest model, being sure to again use the class weight balanced mode to adjust for the imbalanced classes.

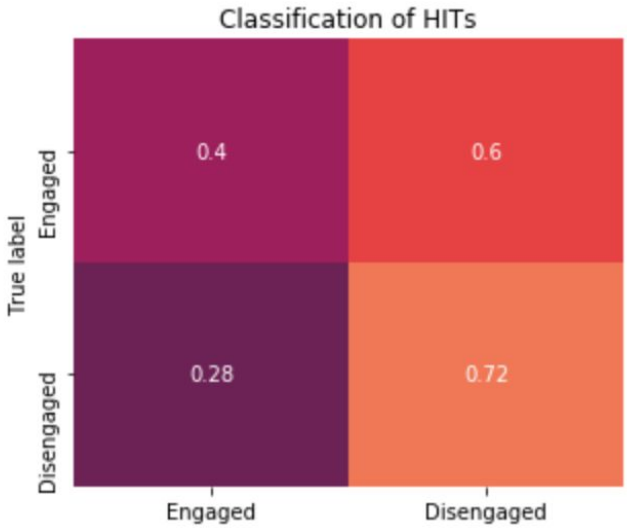


Training a neural network

Finally we trained a neural network to perform classification. To handle the class imbalance we under represented the majority class (will_churn = 0) in the training set.

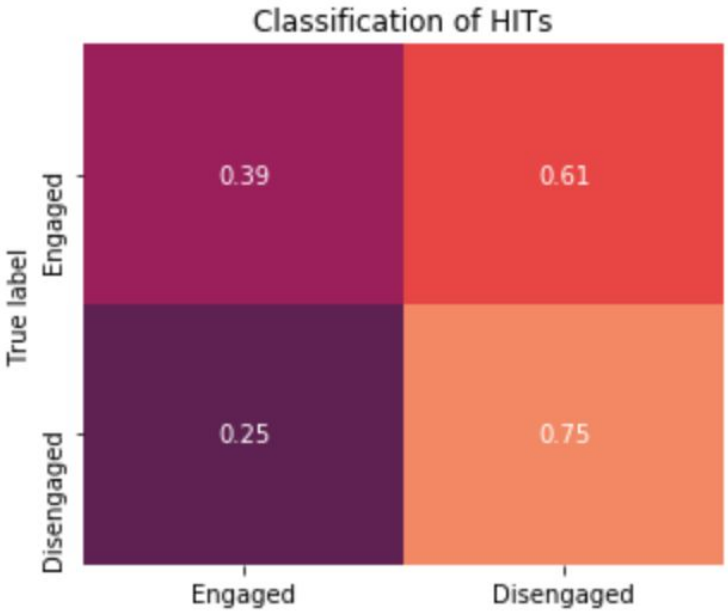
The neural net consisted of 2 layers, and the logistic function activations Y, trained with stochastic gradient descent.

This resulted in the following confusion matrices.



Results

The best performing model of all that we tried was the random forest. This model was very accurate in predicting both the engaged and disengaged class and therefore was the model we decided to evaluate on the test set. This produced the following confusion matrix,



We can see from this that the model didn't perform as well on the test set since it incorrectly predicted the disengaged class 61% of the time. This is probably due to the weights being adjusted during training to handle the imbalanced classes.

Future Work

With more time and resources, future work on this problem could include:

- Use a user's previous behaviour as a predictor on when there session will end. Rather than a model that averages over all the users.
- Spend more time examining different architectures of the neural network and it's activations
- Examine the effect a false positive (i.e. nudging a user when they were not going to disengage) has on a users behaviour and if this should be strongly avoided