User Manual

N.B. This project is primarily research based and, as such, is not exactly geared towards having a 'user'. Hence, rather than an in depth instruction manual for an end user, this document will focus on giving information which would be useful for recreation of the results outlined in the technical manual, assuming later researchers to be the 'user'.

Methods used

Data scraping and storage
The PRAW Reddit API was used in scraping all datasets used in this project. Some custom error handling and further processing of the PRAW object returned was carried out in some cases.

Statistical methods

Standard methods to test association were applied to select variables for inclusion in the model, as outlined in the table below. The standard functions for each test in the Python module scikit-learn were used, with the exception of Cramér's V which was implemented as shown in the technical manual. This was carried out to select the best predictive variables and remove extremely collinear variables.

Variable types to be tested	Pre-processing	Statistical test
Categorical/categorical	-	Chi-squared, extended by Cramér's V
Numeric/numeric	-	Pearson's R
Categorical/numeric	Classification of numerics	Chi-squared, extended by Cramér's V

Classification of the numeric variables was performed through logarithmic binning as all numerics had heavily skewed distributions.

Regression

The standard function for ridge regression in scikit-learn was used for building the model, validation functions in scikit-learn were used to derive R^2 values for the various versions of the model. Any filtering or sampling steps carried out on training data sets are outlined in detail in the technical manual.