Predicting Paper Retractions

William Davis and Jack Kendrick Github

Overview

Peer reviewed articles are a foundational pillar of academia. In theory, the process of peer review ensures that high quality, credible, and trustworthy research that advances our current knowledge can be disseminated to the community at large. However, it is not a foolproof process and some papers do fall through the cracks and end up being retracted. Retractions indicate seriously flawed and unreliable research, errors, fraud, ethical issues, or other serious concerns. Our aim was to build a classifier that identifies papers that have a high risk of retraction. We hope that our model could be helpful in the peer-review and publication process but want to emphasize that it cannot and should not replace rigorous scrutiny from an expert.

Stakeholders**:** Our main stakeholders are academic journals and those involved in the peer review process.

KPIs: Our key performance indicators are precision, recall, and F_1-score.

Dataset

We used the Retraction Watch database to identify PLOS One as a reputable journal that has a large number of retractions. Our dataset consists of 424,223 papers published in the journal PLOS One from 2010-2020. Data was collected from OpenAlex using the PyAlex API. We used the raw data available from OpenAlex to create features of interest such as the proportion of a paper's authors that have been retracted previously, and various measures of how many retractions institutions associated with a paper have received.

Approach

We built a baseline logistic regression model and used forward stepwise subset selection to choose a small subset of features to focus on. We then compared this baseline model to nearest neighbor, random forest, and support vector classification methods. Throughout this process, we used stratified 10-fold cross validation to choose hyperparameters. The stratification here is important since our two data classes are hugely imbalanced.

Results

Our final model is a random forest classifier consisting of 500 estimators with a maximum depth of 20. In testing, this method has a F_1-score of 0.289, precision of 0.691, and recall of 0.182. In particular, the high precision and low recall suggests that much of our inaccuracy stems from false negatives, rather than false positives. The overall accuracy of our classifier in training is 99.8%.