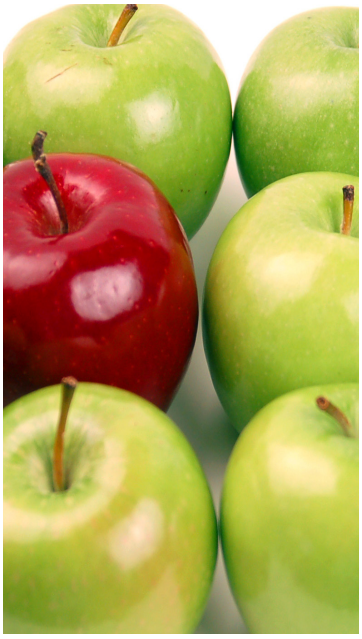
Automating Web Page Classifcation through AI

Rittman Mead was contacted by a client to determine whether a manual web page classification process could be automated using Artificial Intelligence (AI).

Background



The client is involved in the fight against online piracy. As part of this they trawl the internet to determine whether pages have links to pirated material, this is done through a manual classification process.

They were looking to automate this; currently pages are manually checked by a third party and then rechecked by the client. This process is highly labour intensive, and therefore costly. As such, they would like to create a predictive model that would perform scoring within the operational data pipeline and therefore automate checks.

The client placed more stringent requirements on the ability of the model to correctly predict the “negative class” over the “positive class”. This means that incorrectly predicting the positive class can be tolerated to a greater degree.

Solution

Rittman Mead’s Insight Lab was used to deliver the solution. Following the Insight Lab methodology the data is first explored and cleaned for use.

The data composition was then investigated, and feature engineering was used to identify useful predictors within the data set. The data was class imbalanced, that is to say that there were more instances classified as positive than negative. Therefore to produce models with a high accuracy in identifying the “negative class” sites, many sampling and training methods were tested.

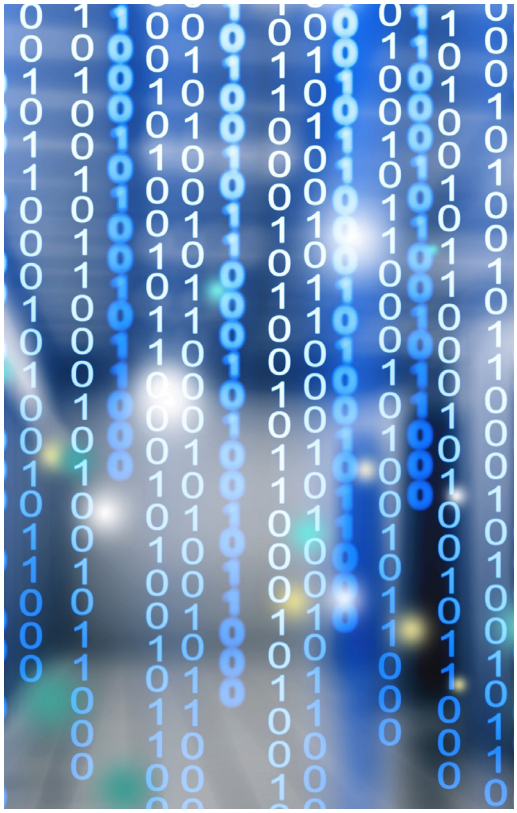
rittmanmead
A DATA AND ANALYTICS COMPANY

A baseline model was produced using only the supplied URL and site information, built by investigating and implementing a number of different models and sampling techniques.

Models were trained and tested on separate data sets to ensure that the true performance of the model against unseen data is evaluated. If a model is tested on the same data set as it is trained, then the reported accuracy of the model is likely to be an overestimate.

During training, the models are evaluated in accordance with an appropriate metric, the improvement of which is the “goal” of the development process. The predictions produced from the trained models, when run on the test set, will determine the accuracy of the model, how closely its predictions align with the unseen real data.

The best fitting model was identified as that which has the largest sensitivity (true positive rate). This was found to be a GLMNET (Generalized Linear Model using Elastic-Net regularization) model trained on accuracy using a down sampled training data set.



Benefits

The produced model, with 98% accuracy, allows the automation of a time consuming task, enabling resources to be focussed on more demanding tasks. Although for the moment some manual checks will still be required to monitor the accuracy of the model, and identify discrepancies or deviations from the model’s fit, there will be a significant reduction in the amount of manual checks that will need to be completed both within the client and by the third party, saving time and money.

Products and Services used

The Rittman Mead Insight Lab offers on-demand access to an experienced data science team, using a clearly defined and proven process to deliver one-off analyses and production-ready predictive models.

With the expert knowledge at Rittman Mead we can ensure the right models are used and no biases are introduced into the data. A clear and concise description of the investigation from start to finish will be provided to ensure that justification and context is given for all decisions and actions.