Case Study

TEXT MINING AUDIT REPORTS TO DETERMINE GRANT RECIPIENT RISK

Elder Research partnered with Excella Consulting to build an end-to-end grant risk estimation solution in the client's AWS cloud. It used text mining and document classification to extract CPA Findings from audit reports and assign risk scores to federal grant recipients.

INDUSTRY

» Government

BUSINESS NEED

» Text mine audit report documents to extract CPA findings and assign risk scores to grant recipients

SOLUTION

» Built an end-to-end solution in the client's AWS cloud

» Delivered an interactive investigative tool using Looker to provide end users with appropriate context to make informed decisions

» Identified audit findings with 81% precision and 95% recall rates

BENEFIT

» Centralized data, with search/filter and drill-down capabilities to enable efficient access to data

» Improved data traceability and transparency to inform investigations

The Challenge

The client needed to optimize strategies to fight fraud, waste, and abuse for federal grant applications. Grant recipients must undergo a Single Audit performed by an independent certified public accountant (CPA) as defined in Circular A-133 by the U.S. Office of Management and Budget. The audit is to ensure a recipient complies with the federal program's requirements for how the money can be used. One of its key elements is the findings section where independent auditors list where the auditee is not following best financial or government grant program practices and requirements. The project goal was to use text mining and machine learning to extract the independent CPA findings from the reports and to use them to evaluate grant recipient risk. The team set a target of 80% precision and 95% recall for the project.

The Solution

Elder Research partnered with Excella Consulting to build an end-to-end solution in the client's AWS cloud. The solution involved data ingestion, unsupervised and supervised machine learning, and a powerful dashboard visualization and drill down tool based on Looker. The dashboard included search/filter and drill-down capabilities to enable users to locate, access, and work with data efficiently and independently.

The client receives approximately 50 thousand audits per year. Audit reports are multi-document PDFs ranging in size from dozens to hundreds of pages and comprised of a mix of machine-readable text and scanned images. We extracted approximately 12 million PDF pages (for about five years of audits), performed text mining, and incorporated other structured data sources to assign risk scores to recipients. A model ensemble that included a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) was used to classify pages during text mining. We found that the structured data only documented about half of the actual findings.

As a next step we are currently working on extracting and analyzing the text of each individual finding using a hybrid CNN/RNN model working at the granularity of characters.

Headquarters
300 W. Main Street, Suite 301
Charlottesville, VA 22903
(434) 973-7673
www.elderresearch.com

ELDER RESEARCH
DATA SCIENCE ◆ MACHINE LEARNING ◆ AI

Office Locations
Arlington, VA
Linthicum, MD
Raleigh, NC
London, UK

Results

More than 260 auditors, investigators, evaluators, and lawyers now use the tool and it has helped launch or support eight audits in four different regions, three evaluations in three regions, and one major investigations project. Our client has named this project one of its five most important initiatives.

Elder Research helped the client throughout the entire agile development process, from road mapping machine learning goals, to the selection of infrastructure/tools and data sources. The project has been extended to include more data sources, text mining, graph analysis, and other leading-edge technologies and goals.

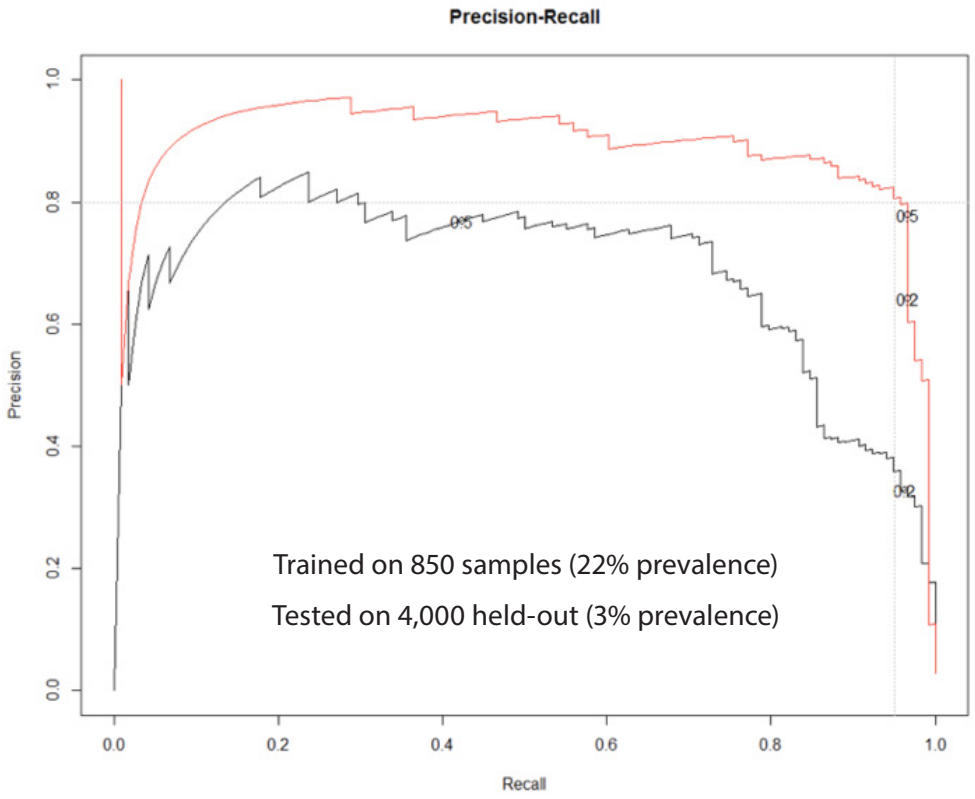

Figure 1. The Precision-Recall curve of our page classification algorithm. The black line is a baseline Naive Bayes model, and the Red line is our CNN/RNN hybrid algorithm, which dominates the baseline model, and exceeded the project goal of 80% precision and 95% recall.

About Elder Research

Elder Research is a recognized leader in the science, practice, and technology of advanced analytics. We have helped government agencies and Fortune Global 500® companies solve real-world problems across diverse industries. Our areas of expertise include data science, text mining, data visualization, scientific software engineering, and technical teaching. With experience in diverse projects and algorithms, advanced validation techniques, and innovative model combination methods (ensembles), Elder Research can maximize project success to ensure a continued return on analytics investment.

Headquarters
300 W. Main Street, Suite 301
Charlottesville, VA 22903
(434) 973-7673
www.elderresearch.com

Elder Research
Data Science • Machine Learning • AI

Office Locations
Arlington, VA
Linthicum, MD
Raleigh, NC
London, UK