CASE STUDY

UNLOCKING VALUABLE DATA FROM SCANNED DOCUMENTS

FOR A LEADING LIFE INSURANCE COMPANY

Elder Research largely automated the extraction of valuable underwriting information from scanned Attending Physician Statement documents using text mining tools and techniques. Extracted text features were transformed into electronic formats suitable for predictive modeling.

INDUSTRY

- Insurance

BUSINESS NEED

- Create a text mining process to unlock valuable applicant information from scanned Attending Physician Statement documents to improve an established underwriting risk model

SOLUTION

- Developed a text extraction pipeline to analyze scanned documents using advanced text mining tools and techniques
- Transformed extracted text features into a format suitable for predictive modeling

BENEFIT

- Saved valuable time by focusing underwriters on the highest risk cases
- Provided additional structured inputs to the client's model for predicting underwriting risk

THE CHALLENGE

Each application for life insurance includes many supporting documents that must be analyzed by underwriters to determine whether a case will be accepted or denied. The analysis process is time consuming for underwriters, as the relevant information is often buried within hundreds of pages of scanned documents. A leading insurance provider identified an opportunity to use text mining to unlock valuable applicant information from scanned images (PDF files) to improve their underwriting risk model. The goal of this project was to create a data capturing process to extract information from Attending Physician Statement (APS) documents and convert the information into electronic formats suitable for predictive modeling. The new data would be used to enhance an existing predictive risk model used by underwriters to determine whether to accept or decline new cases.

THE SOLUTION

When applying for life insurance, the insurance company looks at many different factors to determine the level of risk associated with insuring an applicant. To evaluate the medical risk to the patient a detailed medical history is necessary to make a suitable underwriting decision.

An Attending Physician Statement is a medical history summary from a physician, hospital, or medical facility that has treated the patient and is one of the most sound and proven forms of additional background information. These files often include many pages and contain sizeable amounts of uninteresting information.

Automating this task is attractive, yet mining text from digitized files comprised of multiple formats, and of varying document quality, can be extremely challenging.

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

www.elderresearch.com

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

Office Locations
Arlington, VA
Linthicum, MD
Raleigh, NC

Based on the specific challenges with the APS documents a pipeline architecture was developed to maximize the yield of the important text features, as shown in Figure 1.

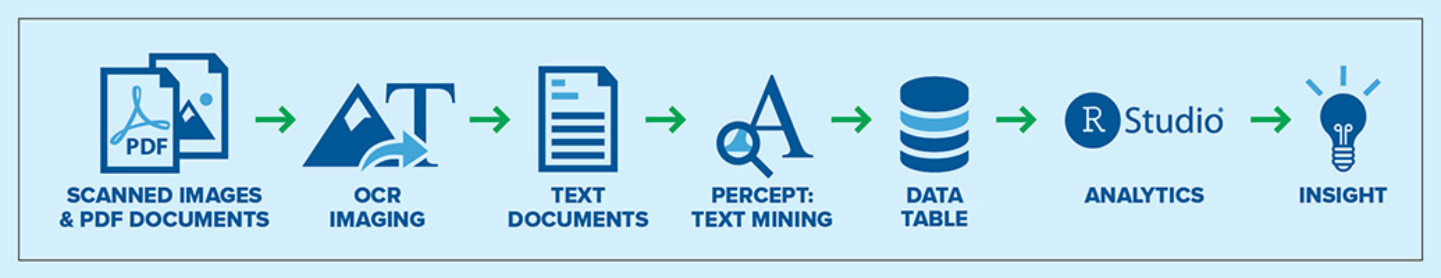

Figure 1. Data Extraction Pipeline

The key challenges encountered when setting up the data pipeline process were:

- Lack of consistent formatting within the APS documents
- Accurate identification of text when using optical character recognition (OCR) on documents with low-quality scans, handwritten information, or both
- Missing data (e.g., no recorded response to a question on a form).

Based on the specific challenges with the APS documents a pipeline architecture was developed to maximize the yield of the important text features (see Figure 1).

The solution employed by Elder Research relied on our text analytics expertise as well as proprietary and open source software tools. All documents required OCR before the data was extracted though documents with lower resolution first needed to be scaled to a larger size. The documents were then processed by Percept, a tool (or set of software routines) developed by Elder Research to extract text from documents in a variety of formats including PDF, Word, and HTML and convert that text into a format suitable for use by predictive models. R Studio was used for post processing, data cleaning, and to transform the text features into an Analytics Base Table—a data table formatted to support analytical modeling.

Results

Creating a structured framework for extracting the critical underwriting information from the APS reports enhanced underwriting risk analysis results. The text extraction process provided valuable inputs for the underwriter in an easy-to-scan structured data format that could be combined with the other structured formats already considered to better prioritize which applications are examined in depth. Prioritizing case decisions based on improved risk prediction will increase efficiency and reduce cost by minimizing the time subject matter experts spend reviewing new applications.

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 & PREDICTIVE ANALYTICS

Office Locations
Arlington, VA
Linthicum, MD
Raleigh, NC