Case Study: An AI-driven Model to Classify Legal Documents

Introduction

J P Morgan announced it had developed and deployed new software called COIN—shorthand for Contract Intelligence—that automates document review for a certain class of contracts. COIN does the mind-numbing job of interpreting commercial loan agreements that consumed 360,000 hours of work each year by lawyers and loan officers within seconds.

Learning Outcome

You will learn to convert a business problem into an analytical problem and the ability to break the process down using CRISP-DM.

Background Information & Scenario

Background

The software employs image recognition to identify patterns in these agreements. While JP Morgan has been tight-lipped about the details of the proprietary technology, the bank has stated that the algorithm digests data on the bank's numerous contracts and it can identify and categorize repeated clauses. The bank reports that the algorithm classifies clauses into one of about one hundred and fifty different "attributes" of credit contracts. For example, it may note certain patterns based on clause wording or location in the agreement.

Scenario

The software reviews in seconds the number of contracts that previously took lawyers over 360,000 person-hours. Apart from shortening the time it takes to review documents, COIN has also managed to help JP Morgan decrease its number of loan-servicing mistakes. J P Morgan intends to deploy COIN for more complex filings, such as credit-default swaps and custody agreements. In the medium and long term, the bank also hopes to use machine learning to interpret altogether new regulations.

Problem Statement / Business objectives

How would you break the problem/scenario down with the help of CRISP-DM methodology: "Automate the classification of various legal documents"

Data, Information for Case Analysis

Use the steps in the CRISP-DM methodology.

Question, Deliverables, and Rubric

What's your solution?

Distinct the process based on all the steps in CRISP-DM.

Deliverables and Rubric: Graded Assessment Requirements

1. Required deliverables – a word document detailing the breakdown of all 6 steps as shown above in the solution.
2. Submission templates – N/A
3. Student facing and faculty rubrics – Total of 12 points where:
   - ❖ The ability to break down the overall process – 6 points.

- ❖ Grouping relevant procedures/information based on domain knowledge and the business problem provided in each step – 1 point per step.

4. Sample solution - The solution for this exercise is not constrained to a single answer. A general solution is provided above.

Solution

1. Step 1 :
Business understanding / Objective : To Analyze the Software **"COIN"** developed which automates **classification of various legal documents** for a **certain class of contracts** and examine the **relationship between the attributes** to adopt a model so that the consumed 360,000 hours of work each year by lawyers and loan officers is **reduced to seconds**.

Step 2 :Data Understanding :
2.1 Data Collection / Data Attributes : based on clause wording or location mentioned in the agreement. Uses **credit-default swaps** and **custody agreements.**
Attributes : Based on the main features like Capacity, capital, collateral, Conditions, Character of the contract.
2.2 Data Description :
File type : Image

Step 3 : Data Preparation
3.1 Data Cleaning & Pre-processing
This process take 80% performance of the case study.
Remove the dirt in the attributes. To fill the missing and incorrect values such as zeros, multiple zeros, 'none,' 'no,' spaces, various spaces, NULLs, etc.
Converting the data types to a correct format.
Step 4 : Modeling & Evaluation
The bank reports that the algorithm classifies clauses : based on clause wording or location and relate the attributes to one another finding a required pattern and classify it and categorize repeated clauses to give a better accuracy and speed . And adopt machine learning model to interpret altogether new regulations.

Conclusion:

A software "COIN" is designed and developed to automates classification of various legal documents for certain contracts using an algorithm which classifies
the repeated attributes found in the document images, which as a result gives a better accuracy and to review the documents within seconds and reduces the loan-servicing mistakes.