My Understanding of Case Study: Video Link

https://drive.google.com/file/d/1dvxCIhbRBCkL1a8bdVAS0ZI3HEba2e00/view?usp=sharing

JP MORGAN: AUTOMATION OF CLASSIFICATION OF LEGAL DOCUMENTS

Data Science PGC, Internshala, Course I Project

Abstract

Converting JP Morgan's business problem into analytical Problem using CRISP-DM

Amit Hiremath

a.hiremath94@gmail.com

J P Morgan: Automation of the classification of various legal documents

Amit Hiremath
Data Science PGC, Course I, Internshala – 04/07/2024

Abstract:

To create an CRISP-DM framework for the JP Morgan COIN project via break down the process into the six phases of CRISP-DM (Cross-Industry Standard Process for Data Mining). This approach provides a structured or systematic way to solve business problems.

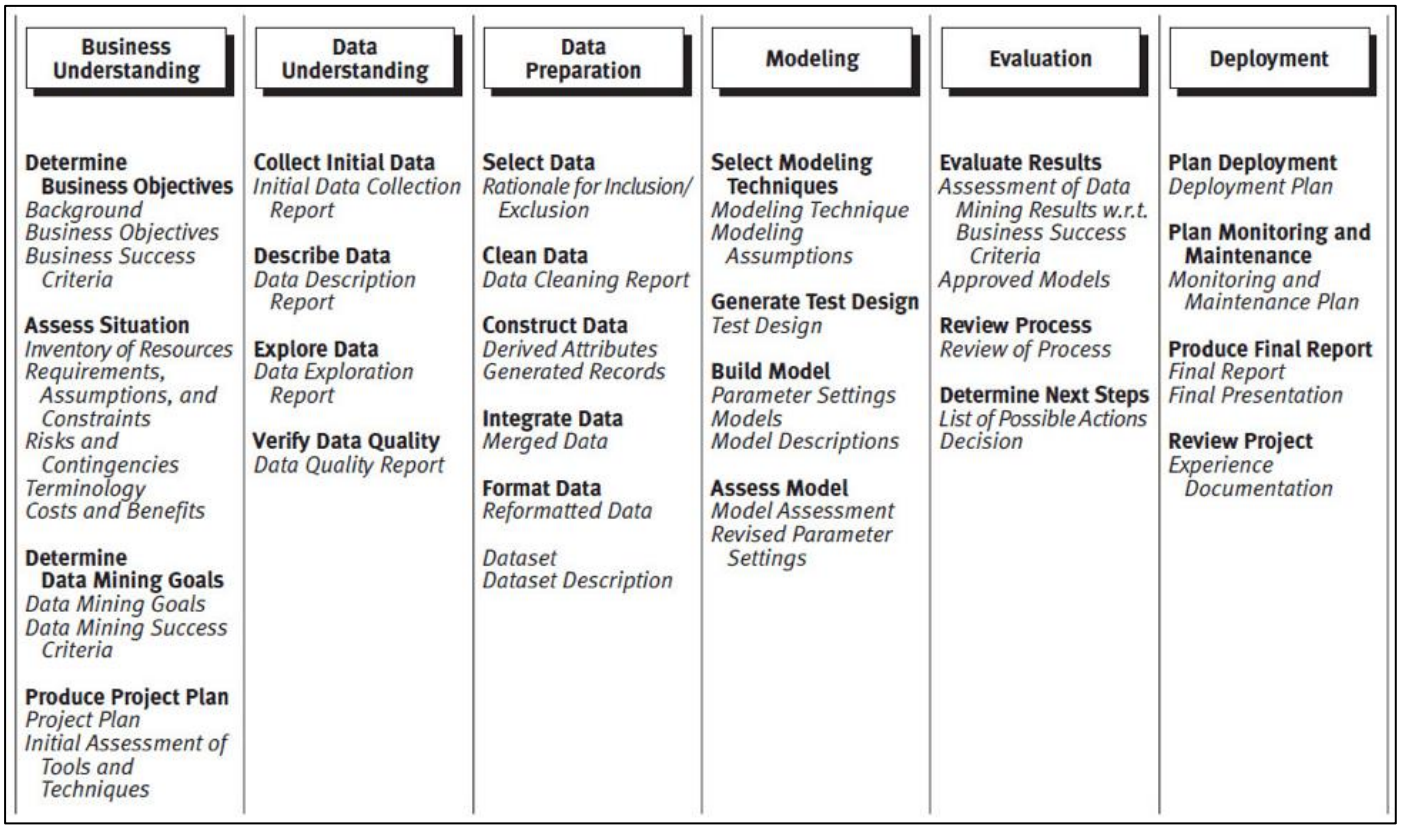
Introduction

About JP Morgan:

JP Morgan is a global financial services firm with over 240,000 employees and operations in more than 100 countries. They are a leader in investment banking, commercial banking, financial transaction processing, and asset management. JP Morgan serves millions of customers, including many of the world's most prominent corporate, institutional, and government clients. They have been operating in Europe for nearly 200 years and have a strong local market presence across Europe, the Middle East, and Africa (EMEA). Their regional headquarters for their EMEA business is in London, and they have offices in all major financial centers.

About CRISP-DM

CRISP-DM is a comprehensive and flexible framework for data mining projects, ensuring that they are systematically and efficiently executed. There are 6 phases in the CRISP DM framework, namely: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment. Following chart shows sub-phases of each of the phases.



Case Study: The COIN software of JP Morgan employs image recognition to identify patterns in these agreements. 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.

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.

What needs to be done? For the above-mentioned business problem, COIN software project needs to be implemented using CRISP DM for classification of legal documents including for credit swap, custody agreements and ML for interpretation of new regulations.

We can straight away jump to the CRISP DM framework for Planning & Deploying COIN software for various purposes mentioned above. Key deliverables are as follows:

- Business Understanding Document: Detailing objectives, benefits, and key questions.
- Data Understanding Report: Summarizing data characteristics and initial findings.
- Data Preparation Plan: Outlining preprocessing steps and techniques used.
- Modelling Report: Describing the chosen algorithms, training process, and performance metrics.
- Evaluation Report: Assessing the model's effectiveness and potential impact.
- Deployment Plan: Detailing steps for implementation, monitoring, and user training.

Phase I: Business Understanding (Document)

Business Objective

The primary objective of implementing the COIN software at JP Morgan is to automate the review of legal documents and classification of credit attributes, significantly reducing the time and resources spent on these tasks. This will enhance efficiency, accuracy, and compliance within the organization.

Current Situation

JP Morgan, a leading global financial services firm, deals with a vast amount of legal documentation daily. The traditional manual process of reviewing and classifying these documents is time-consuming, prone to errors, and requires substantial human resources. As the volume of transactions and legal documents grows, so does the need for a more efficient solution.

Problem Statement

The manual review process for legal documents at JP Morgan is inefficient, error-prone, and resource-intensive. There is a need for an automated solution that can quickly and accurately classify legal documents into one of 150 different credit attributes, ensuring compliance and reducing operational costs.

Stakeholders

Senior Management: Interested in reducing costs and improving efficiency.

IT Department: Responsible for the implementation and maintenance of the COIN software.

Data Science Team: Tasked with developing and refining the machine learning models for document classification.
Operational Staff: End users who will interact with the software and rely on its outputs for decision-making.

Expected Outcomes
Efficiency: Significant reduction in the time required to review and classify legal documents. Accuracy: Improved accuracy in the classification of credit attributes, reducing the risk of errors. Cost Reduction: Lower operational costs due to reduced manpower requirements for document review. Compliance: Enhanced compliance with regulatory requirements through consistent and accurate document classification.

Business Constraints
Data Security: Ensuring the security and confidentiality of legal documents.
Integration: Seamless integration of COIN with existing IT infrastructure and workflows.
User Training: Training staff to effectively use the new software. And also Adhering to all relevant regulations and standards.

Risks and Mitigations
Data Privacy: Risk of sensitive data being exposed. Mitigation: Implementing robust data encryption and access controls.
Model Accuracy: Risk of the machine learning model misclassifying documents. Mitigation: Regularly updating and validating the model with new data.
User Resistance: Risk of staff being resistant to change. Mitigation: Providing comprehensive training and support.

Key Performance Indicators (KPIs)
Time Reduction: Measuring the time taken to review and classify documents before and after implementation.
Accuracy Rate: Tracking the accuracy of document classification by the COIN software.
Cost Saving: Calculating the reduction in operational costs due to automation.

Phase II: Data Understanding (Report)

Data Collection
The data collection for the JP Morgan COIN software project will include historical legal documents that have to be reviewed and classified manually. These documents are used to train the machine learning models to automate the classification process.

Data Sources
Legal Documents - These documents contain information about various credit attributes that need to be classified. The dataset will be of thousands of legal documents.

Data Formats
Text Documents - The primary data format is text, including PDF, Word, and plain text files. The dataset comprises several key attributes necessary for the classification task. These attributes include both text data from the documents and structured metadata i.e., Document Content, Labels indicating the credit-related categories each document falls into e.g., loan agreements, credit default swaps or custody agreement other attributes like Document ID, Creation Date, Author, The type of legal document e.g., contract, agreement, memorandum etc.

Data Quality
Ensuring high-quality data is critical for the success of the machine learning models. The following aspects of data quality are to be considered: Completeness - All necessary fields and attributes should be present in each document. Missing data should be identified and addressed, either through imputation or by excluding incomplete records. Consistency - Ensuring that the data is consistently formatted across all records. Resolving any discrepancies

in the labeling or categorization of credit attributes. Accuracy - Verify that the text data accurately reflects the content of the original documents. Timeliness - Ensuring that the data is up-to-date and relevant to the current legal and business context.

Data Exploration

Data exploration involves analyzing the dataset to understand its structure, distribution, and

Phase III: Data Preparation

The objective of the data preparation plan is to outline the steps necessary to clean, transform, and prepare the dataset for building a machine learning model to classify legal documents into one of the 150 different credit attributes. Steps in Data Preparation:

Data Collection and Aggregation

Gathering all relevant legal documents and metadata from the identified sources. Aggregate the documents and metadata into a unified dataset for processing.

Data Cleaning

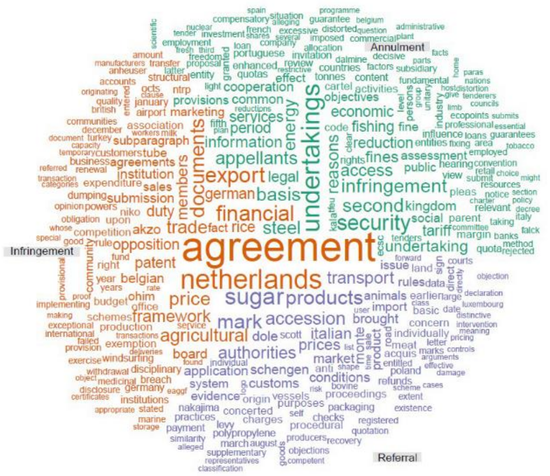
Removing Duplicates, Handle Missing Values - For text data: If a document has missing sections or paragraphs, decide whether to exclude the document or impute the missing sections using techniques like interpolation or external references. For metadata - Impute missing values where possible, using techniques such as mean imputation for numerical data or mode imputation for categorical data. Alternatively, remove records with missing critical information. Correct Inconsistencies - Standardize inconsistent entries in the metadata, such as date formats, author names, and document types.

Text Preprocessing

Tokenization - Spliting the text of each document into individual tokens (words or

any underlying patterns. Key steps in this process include: Summary Statistics

Calculate basic statistics (e.g., mean, median, standard deviation) for numerical attributes. Generate frequency distributions for categorical attributes. Text Analysis - Performing word frequency analysis to identify common terms and phrases in the documents, Using natural language processing (NLP).



phrases). Lowercasing - Converting all text to lowercase to ensure uniformity. Removing Stopwords: Removing common stop-words (e.g., "and," "the," "is") that do not contribute significant meaning to the text. Removing Punctuation and Special Characters: Eliminating punctuation marks and special characters to clean the text. Stemming/Lemmatization - Reducing words to their root form (e.g., "running" to "run") to consolidate similar words.

Feature Engineering

Term Frequency-Inverse Document Frequency (TF-IDF): Calculating the TF-IDF scores for words in the documents to identify important terms. N-grams - Creating n-grams (bigrams, trigrams) to capture context and phrases. Metadata Features - Encoding categorical metadata features using techniques like one-hot encoding. Text Embeddings - Using pre-trained word embeddings (e.g., Word2Vec, GloVe) or transformer-based embeddings (e.g., BERT) to convert text into numerical vectors.

Data_Transformation:

Normalization/Standardization: Normalizing or standardizing numerical features to ensure they have a similar scale. Feature Selection - Selecting the most relevant features based on correlation analysis, feature importance, or other selection techniques. Dimensionality Reduction - Applying techniques like Principal Component Analysis (PCA) to reduce the dimensionality of the feature space if necessary. Data Splitting - Splitting the dataset into training, validation, and test sets. Data Augmentation - If the dataset is imbalanced (i.e., some credit attributes are underrepresented), use data augmentation techniques such as oversampling, undersampling, or synthetic data generation (e.g., SMOTE) to balance the classes.

Phase IV: Modelling

The objective of this phase is to document the modeling phase of the project, detailing the steps taken to build and evaluate the machine learning model for classifying legal documents into one of 150 different credit attributes.

Model Selection

Given the nature of the classification task, we can consider several machine learning algorithms: Logistic Regression, Random Forest, Support Vector Machine (SVM), Gradient Boosting, and Multinomial Naive Bayes etc. Considering the multi-class nature of the problem, we can select the Random Forest classifier as our primary model due to its ability to handle a large number of features and classes, and its robustness against overfitting.

Data Preparation

As described in the Data Preparation Plan, we can preprocess the data.

Model Building

Random Forest Classifier Hyperparameters - Number of trees in the forest: 100, Maximum depth of the tree: None (nodes are expanded until all leaves are pure or contain less than the minimum samples required to split); Minimum samples split: 2 ; Minimum samples leaf: 1; Bootstrap: True (samples are drawn with replacement)

Training

The model can be trained on the preprocessed training dataset using the Random Forest Classifier from the sklearn library.

Evaluation

The model can be evaluated on the validation set using accuracy, precision, recall, and F1-score metrics.

Phase V: Evaluation (Report)

The evaluation document aims to assess the performance and effectiveness of the machine learning model developed for classifying legal contract documents into one of 150 different credit attributes. This evaluation covers metrics, results, and insights drawn from model performance on the test dataset.

Model Overview

The primary model used for this task will be a Random Forest classifier, selecting for its robustness, ability to handle multi-class classification, and effectiveness with the given dataset.

Data Overview

The dataset comprised legal contract documents, each associated with one of 150

credit attributes. The data was pre-processed to clean and tokenize text, handle missing values, and perform feature engineering. Key features included TF-IDF scores, text embeddings, and metadata such as document type and author.

Evaluation Metrics

The following metrics can be used to evaluate the model's performance: Accuracy: The proportion of correctly classified instances among all instances.

Precision: The proportion of true positive predictions among all positive predictions.

Recall: The proportion of true positive predictions among all actual positives.

F1-Score: The harmonic mean of precision and recall.

Confusion Matrix: A table used to describe the performance of the classification model by showing the actual versus predicted classifications.

Phase VI: Deployment (Plan)

The deployment plan outlines the necessary steps to integrate the COIN software into JP Morgan's existing IT infrastructure, ensuring a smooth transition from development to production. The goal is to automate the classification of legal contract documents accurately and efficiently.

Pre-Deployment Preparation

Infrastructure Assessment: Evaluate the current IT infrastructure. Identify necessary hardware and software upgrades. Ensure compatibility with existing systems.

Data Preparation: Ensure all historical and real-time legal documents are preprocessed. Store data in a secure, scalable database system. Validate data quality and consistency.

Security and Compliance: Implementing data encryption protocols. Ensure compliance with legal and regulatory standards.

Deployment Steps:

Environment Setup Provision of cloud resources or on-premise servers for deployment. Set up development, testing, and production environments. Installing necessary software and libraries, including the machine learning framework. Model Integration - Integrate the trained Random Forest classifier into the production environment. Developing APIs to allow other applications to interact with the model. Ensuring seamless data flow between data sources, the model, and downstream applications. Testing and Validation - Conducting end-to-end testing in the staging environment. Validating model predictions against known outcomes. Performing load testing to ensure system scalability.

Monitoring and Logging - Setting up real-time monitoring tools to track system performance and model accuracy. Establish alert systems for potential issues like system downtime or prediction errors. User Training and Documentation - Trainning end-users and stakeholders on the new system. Providing comprehensive documentation covering system use, troubleshooting, and maintenance.

Deployment Rollout - Beginning with a phased rollout to minimize disruption. Deploying the system to a small user group for initial feedback. Gradually expanding deployment to the entire organization based on feedback and performance.

References:

https://web.actuaries.ie/sites/default/files/2021-10/211005%20Final%20Slide%20Deck_CRISP%20DM%20Talk%20SAI%20webinar.pdf

https://keithmccormick.com/wp-content/uploads/CRISP-DM%20No%20Brand.pdf

https://www.ibm.com/docs/it/SS3RA7_18.3.0/pdf/ModelerCRISPDM.pdf

https://www.datascience-pm.com/wp-content/uploads/2021/08/CRISP-DM-for-Data-Science.pdf

https://research-information.bris.ac.uk/ws/portalfiles/portal/220614618/TKDE_Data_Science_Trajectories_PF.pdf

https://www.datascience-pm.com/crisp-dm-2/

https://hpi.de/fileadmin/user_upload/fachgebiete/rabl/Lectures/PDE_Poster/PDE_Patricia_Sowa.pdf

https://www.elevenjournals.com/tijdschrift/ELR/2021/1/ELR-D-20-00035