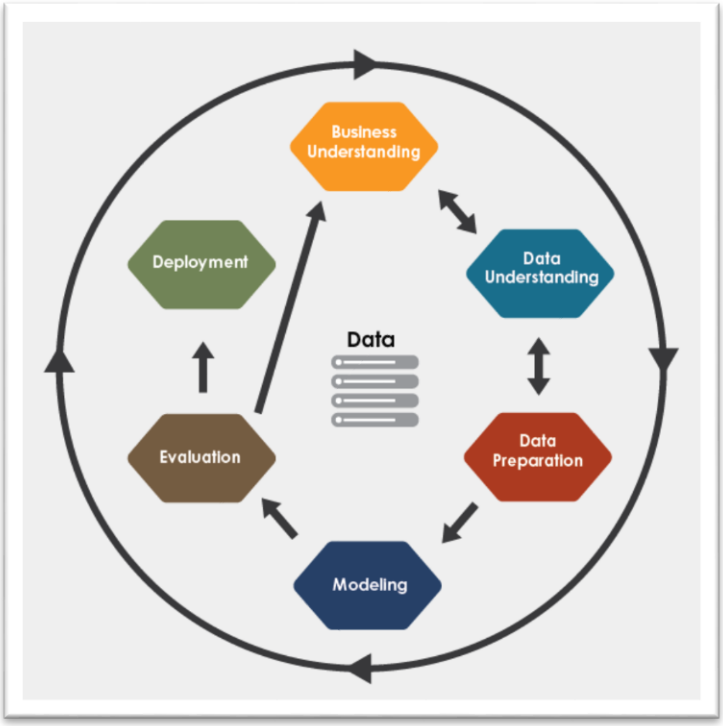
Project: Automation of Legal Document Classification Using CRISP-DM Framework for JP Morgan's COIN System

Problem Statement: "Automate the classification of various legal documents"

Learning Outcome: Convert a business problem into an analytical problem and the ability to break the process down using CRISP-DM.

CRISP DM: CRISP-DM stands for cross-industry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. This project can be break down using CRISP-DM methodology to provide industrial level approach to create a software COIN to classify legal documents.



CRISP-DM Methodology Breakdown:

Phase 1: Business Understanding

This phase focuses on Objective and Requirement of the Project from a Business perspective. This phase consists:

1. **Defining Business Objective**: In this project we want to create and deploy a software called **COIN** (Contract intelligence) to Automate the classification of the legal documents.

2. **Accessing the current situation**: As of now it takes a lot of time (360,000 Hour/Year) To interpret commercial loan agreements by various lawyers and loan officers and that too with many errors.
3. **Determining data mining goal:** Goal is to build a system that can accurately classify the clauses withing legal documents into predefined attributes and reduce the errors.

Phase 2: Data Understanding

This phase focuses on initial collection of data and familiarizing with the data. This phase consists:

1. **Gathering Initial Data:** We start with collection of most of the data we can gather which in our case would be**:**
   1. Historical Legal Documents: These would be all the existing contract and document.
   2. **Annotations or labels:** These would be those documents that are already reviewed by Experts and sorted in the attributes.
   3. **Scanned Documents:** These documents would that were scanned and converted into machine readable text using OCR tools.
2. **Describing Data:** The collected legal document contains various clauses written in various formats and patterns. Patterns such as keywords, positioning and formatting of clauses are necessary for **COIN** to learn and classify these clauses in accurately into attributes.
3. **Exploring Data:** The collected data is thoroughly analysed to reveal various meaningful patterns, variation in clause structure, common keywords. This analysis results in how different legal document with different legal clauses are framed and which will further ensure that the system will accurately distinguish and classify these clauses into 150 predefined attributes.
4. **Verifying Data Quality:** After collecting data and going through the data a thorough check needs to be initiated where all the non-legal and irrelevant documents need to filtered. Non-Legal and irrelevant documents can result in incorrect data to enter into training model which will make the performance of model poor. Misclassification can also happen as non-legal and irrelevant data can hinder in the process of finding the correct keyword, patterns.

Phase 3: Data Preparation

This phase is dedicated to cleaning and transforming raw legal clauses from the document to suitable data for modelling. This phase contains:

1. **Selecting Data:** Selecting data from those documents which were already labelled by experts. Selecting attributed document from 150 predefined attributed documents that can be used in classification task. Focusing of such documents that clearly shows or represent the clauses and attributes.

2. **Cleaning Data:** Data need to be cleared of all the irrelevant information that can potentially make the performance of the model poor. Different tool will be used such as using OCR to extract text from image based and scanned documents and fix any OCR error, removing unnecessary elements such as special character or irrelevant metadata and handling the missing data from all the gathered data by ensuring that all the required clauses and attributes are available for training the model.
3. **Integrating Data:** Combining all the various sources of data that have been gathered and cleaned and make one single data that is easier for model to test and train sets.
4. **Formatting Data:** Structing the data into a format that is suitable for model to perform correctly. Tokenizing text in smaller units to create structed datasets. Converting labels into machine-readable formats.

Phase 4: Modelling

With clean data this phase focuses on utilizing that clean data and building model from that data. in this phase various modelling techniques are used to find the correct model that can produce highly accurate result and that model will be used at the time of deployment phase. The main tool that will be utilize in this project will be **PYTHON** for creating model, evaluating the models and finalizing the correct model**.** This phase contains:

1. **Selecting Modelling Techniques:** Objective of this step is to choose a suitable model for classifying legal clauses into predefined attributes. Traditional machine learning models such as Logistic Regression, Random Forest, SVM can be used for simpler patters and smaller datasets. Deep learning model such as LSTM or CNN or transformer-based models such as BERT or RoBERTa can be used for handling complex, context rich text data.
2. **Designing Tests:** Certain criteria need to be established to evaluate the performance of the model. Splitting the data into training, validation, and testing sets (70-20-10 split) so that we can use 70% of data to train the model, use 20% of data to test the model based on data its trained and finally validating the model with 10% of the data. Metrics tools in python library such as **accuracy, precision, recall** and **F1-score** will be used to evaluate the performance of the model. **Cross-validation** will be performed to ensure the model is generalized well to unseen data.
3. **Building the Model:** Training the selected models on the prepared dataset. Preprocessing the data like tokenization, embedding generation to make the data suitable for the selected models. Training the traditional models using features like keyword counts or TF-IDF scores. Fine tuning pre-trained transformer models like BERT on the legal dataset for better contextual understanding. Optimizing the hyperparameters (e.g. learning rate, batch size) to the performance.

4. **Assessing the Model:** Comparing all the model used (Traditional models vs deep learning models) and finding the best model is objective of this step. Models are to be tested on validation set to avoid overfitting and tuning parameters for even better performance from the selected model. Evaluating the model on the test set using predefined metrics to assess real world performance. Compare results from all the models (Traditional vs deep learning) to select the most effective model.

Phase 5: Evaluation

Before proceeding to deployment phase, the model's performance is need to thoroughly evaluated. This ensures that model meets the business objective that were set in Business understanding phase. This phase contains:

1. **Evaluating Results**: Asses the models performance based on predefined success criteria. Analysing the model's classification performance using evaluation metrics such as **accuracy, precision, recall** and **F1-score.** Use the unseen legal documents to test the data to verify model's ability to generalize to new data. Measure improvements against the baseline.
2. **Reviewing the Process**: Evaluate the overall workflow to identify area of improvement before deploying the model. Analysing each phase thoroughly to ensure that all the phase matches the requirement as per the project. Identifying the challenges faced such in each phase such as gathering the data, cleaning the data to meet the required criteria as per the project, selecting the correct model based on the model that performed best as per the required criteria as per the project and based on the evaluation metrics results for all the models.
3. **Determining the Next Steps**: Selecting the next course of action based on evaluation results. If the model meets the objective set at the business understanding phase proceeding to the deployment phase for automating the classification of legal document in the real-world scenario. Start next phase of the project to make model able to handle even more complex filings such as credit-default swaps and custody agreements. But if the model does not meet the objectives that were set at the start of phase 1 start with identifying area for improvement such as gathering even more data, refining the data cleaning process to fix irrelevant data, training the model with different parameters again to improve performance of the model, refining the feature engineering process, iterating through **CRISP-DM** process with adjustment based on feedback.

Phase 6: Deployment

The final phase involves deploying the model into a real-world environment. This phase focuses on making the developed model operational in real-world and ensuring that the model delivers consistent results in real-world use. This phase contains:

1. **Planning the Deployment:** Planning how the **COIN** system will be integrated into JP Morgan's existing workflow. Defining the deployment environment like keeping the model on on-premises servers or cloud platforms. Ensuring necessary tools such as OCR and clause classification models are integrated. Developing a user-friendly interface for lawyers and loan officer to further smooth the process classifying legal documents and reducing the human error.
2. **Monitoring and Maintenance:** Ensuring the system is performing reliably and adapts to changes on regular basis by setting up performance monitoring to track the accuracy, process speed, and error rates of the model. Implementing a logging and error trackers to identify and address issues regarding the model swiftly and promptly. Updating the system regularly to readily interact correctly with new types of legal document or attributes.
3. **Reviewing the Project:** Evaluate the project's success in meeting its business objectives. Comparing the initial goal with the final outcomes of the project on the bases of key points such time saving, error reduction, and classification accuracy. Collecting feedbacks on regular basis and implementing them on the model to ensure smooth operation of the model these feedbacks can be taken from end users such as lawyer and loan officers.
4. **Finalizing the Project:** Officially close the project and transition it to operational status. Confirming that all the deliverables such as **COIN** system and supporting documentation are completed. Transfer the system ownership to the maintenance and operation team. Preparing a final report summarizing the project's objective, outcomes, challenges, and future recommendations.

Video Link:

https://drive.google.com/file/d/1PL1CSxPgeKXBnsKlsrlhWIVXbjgBw8jg/view?usp=sharing

Summary:

This project is about automation of classification of legal documents. JP Morgan is global financial services company and they have announced that they will be developing a software know as **COIN** a software that will attain the powerful functioning of classifying the legal documents into various categories of cases. We are tasked to create a CRISP-DM for this project. The main objective of this project is to reduce the time it took previously by the lawyers and loan officers which was more than 360,000 hours per year and reducing it to few seconds with the power of powerful machine learning algorithms. And also reducing the error which were made by lawyers and loan officers. We used previous legal documents to train our model and cleaned it various tools provided in **PYTHON** and selected a best performing model out of various Traditional models, NLP models and Deep Learning model**.** We phase by phase completed the required steps needs for **CRISP-DM** and finalize our project to deploy where we setup few task teams for task such as error handling, model evaluations and its regular performance check. After finalizing the project, we ran final tests to ensure proper functioning of the **COIN** as a software in reducing the error and time management. After this we finally deployed the Project.