CRISP-DM Methodology

J P Morgan classification for legal documents

Introduction:

J.P. Morgan facing a big problem: checking lots of legal documents is slow and can lead to mistakes. This can cost them a lot of money and time.

To fix this, we want to use smart computers to read and understand these documents. We will use a special method called CRISP-DM to help us. Using this method, we will define our problems and objectives and reach our goal.

By using this new technology, we can make the process much faster and more accurate. This will save J.P. Morgan time and money.

My name is Ruturaj T. Saravane and here I shared my approach to given challenge using CRISP-DM framework.

Phase 1: Business Understanding

Overview:

We should involve with management in JP Morgan and explore what the bank expects from the data analysis. Our involvement with management and lead staff will help us in this discussion to derive desired outcomes and to conclude possible results. It will also help us to determine objectives for our study.

1. Defining business objectives

We will try to get possible insights into the bank's scenario and plan our data mining goal accordingly. Also discussing with lead people from JP Morgan will help us decide specific business objectives.

As JP Morgan is primarily involved in working with legal documents and loan servicing, bank wants to stay ahead defining new working system to minimize the work delay and avoid human errors while reviewing legal documents. In short, the business objective is,

- To automate the classification of various legal documents.

- To minimize loan servicing mistakes.

the manual review of contracts is time-consuming and prone to errors. Automating this process will reduce review time by 80% and increase accuracy in loan processing by at least 95%.

2. Assessing the current situation

After understanding business objectives, it is crucial to assess bank's current situation. We will review current working environment and will go through available resources that will help us for our analysis.

In JP Morgan, we would assess the available data, current situation of bank's service and all the resources. Currently, reviewing contracts and legal documents consumes 360,000 hours of work each year by lawyers and loan servicing agents that means it requires a significant amount of time to review the numbers of contracts and there are increasing loan servicing mistakes due to manual work. Also, it is difficult to interpret new regulations about bank.

We have to address,

- How can we handle incomplete or inconsistent documents?

- What existing systems can support this automation?

3. Determining data mining goals

We will focus on deriving data mining goals associated with business objectives. These will be our technical goals.

- To convert manual documents into scanned form using image recognition.

- To convert scanned documents into machine-readable text using OCR tools.

- To develop a machine learning model to classify clauses into every possible attribute and to classify further those legal documents and contracts.

- To build an AI model for complex fillings, such as credit default swaps and custody agreements.

- To build a model which will help to calculate loan payments and fees, apply payments, and give timely notices.

4. Producing a project plan

In this phase, we will discuss our tasks and propose our project plan with our team and with superiors in JP Morgan. Estimating the time required for our project and considering the budget is crucial.

We should be prepared for problems with the data, like errors or inconsistencies. Adjust our plans to handle these issues. For example, if different sections of the data look different, we'll use tools to make them look the same.

We will assign time for each phase of our project and define what resources will each phase requires. We should consider risk factors which may arise while performing our project work and should inform our team and superior staff on priority.

Phases	Time required	Resources/Tasks	Challenges	Cost or Budget
Business Understanding	0 – 1 week	Meetings, Discussions	Issues interacting with staff, etc.	
Data Understanding	0 – 1 week	All documents and data sources	Documents and other data in Irrelevant form, etc.	
Data Preparation	2 – 3 weeks	Handling documents	Incomplete data and inconsistencies, etc.	
Modeling	1 – 2 weeks	High end computer system and tools	Availability and affordability of this system.	
Evaluation	0 – 1 week	Presentation with superiors	Unwanted results and problem while reviewing.	
Deployment	2 – 3 weeks	High end computer system, Management team	Issues in deployment and management.	

(time and other allocation are given approximately, as I do not have much knowledge about planning of this project)

Phase 2: Data understanding

This phase focuses on availability of data and understanding it for our analysis. It holds a significant role because it sets a clearer and effortless path for the next phase of data preparation.

1. Gathering initial data

The JP Morgan uses various data sources which we can consider.

- All the legal documents and contracts.
- Documents related to commercial loans and finance services.
- Agreements such as credit-default swaps and custody agreements.
- We will examine historical data on errors and delays in loan processing.

2. Describing data

We will characterize the quantity and quality of the data and describe the conditions of the data we have.

- Data description includes clause types, positions, and common terms.
- Quality and quantity of legal documents.
- 1000 contracts with an average of 20 clauses each.
- Documents contains numerical data for payments and punctuation marks.
- Attributes like clauses, terms and conditions, deadlines, payment invoices, etc.

3. Exploring Data

We can explore data to summarize its usefulness and how it be utilized in our project work. It will help us to understand how our data is aligned with our data mining goals. Main task is to analyze language patterns, clause structures, etc. and to understand types of data and data distribution.

- Clauses contain dates of payments, due dates.
- Legal documents may have 100+ attributes.

- Custody agreements contains both beneficial and asset owner with terms of document.

- There are 2-3 clauses in medium sized documents and more than 5 clauses in large legal documents.

- Wrong calculation of payments and fees in loan servicing documents.

4. Verifying data quality

It is important to ensure that our data is in complete form and relevant. Also, we should look for any inconsistencies like missing data or duplicate values.

- There are missing payment terms.

- Contextual information from contracts is useful for missing data.

- Custody agreements are in complete form but owner name is missing.

- There are duplicate due dates printed on loan agreements.

Phase 3: Data Preparation

1. Selecting data

We will focus on relevant data types and attributes for our next step.

- Portions of contracts, some specific clauses, terms, and conditions, etc.

- Clauses related to penalties for late payments.

- Loan types and terms, Payment methods, etc.

- We have Chosen owner names and demographic details from custody agreements.

2. Cleaning data

It is a crucial task because data cleaning will reduce problems in our analysis. Handling missing values or estimate them using same type of documents is essential.

- Using interpolation or WMA or such statistical technique we have replaced missing numerical values.

- We excluded irrelevant data from custody agreements.

- We are handling symbols, punctuation marks from contracts and agreements with python libraries and removing them.

3. Integrating data

We will combine those datasets or and data types which are related to each other. Integrating data based on what machine learning model you are going to use is important.

- Text data of contracts and numerical data of payments and loans.

- We Linked demographic data of borrowers with legal clauses.

- Integrating Clauses from legal documents and demographic data of customer.

- Combining Loan agreements with behavioral data of customers.

4. Formatting data

We will format our data based on our tools and techniques of modeling.

- We will require an image recognition tool so further we can tokenize text data.

- Checking formats of clauses.

- Ex. We will format loan data like loan amount – payments – dues.

Phase 4: Modeling

1. Selecting modeling techniques

We must find proper model related to our goals of data mining.

In our scenario, we are selecting following models,

- Model that will convert our scanned data in machine readable format.(OCR)

- Model to recognize signatures and stamps.

- Model for defining various types of clauses and other attributes. (NLP)

- Model for combining related data.

- Techniques of forecasting for payments and due dates.

2. Designing test

We will work on designing test and splitting the data into training and testing sets. We will choose appropriate testing model and adjust it with our testing sets.

- We have defined data sets for designing test.

- Team is deciding which test should be chosen and then work designing it.

- We tested the model's ability to classify clauses by comparing it with a manual dataset.

3. Building the model

We will build various models based on our goals and objectives. Comparing the results from those models and then finalizing the model for our analysis will be a good practice. Then we will train those models using prepared data.

- Building the model for classifying clauses, terms, etc. into attributes.

- Building the model to optimize all attributes in our data.

- Building model to combine related data to extract insights about loan servicing.

- Building the model to calculate loan payments and fees.

- Building the model to derive relations between beneficial owners and asset owners through agreements.

4. Assessing the model

We run our trained model on real data to check performance based on accuracy.

- Our team ran model on data and saw if it classifies clauses as we desired.

- The model is deriving 100+ attributes from legal documents.

- The model is missing some critical clauses from custody agreements.

- Decision tree model is calculating upcoming payments correctly.

Phase 5: Evaluation

1. Evaluating the results

We will evaluate the results by measuring the accuracy and using statistical techniques.

- Comparing results of the model's output with manual documents.

- Suppose we run the model to classify the clauses then we will check if it correctly classifies those clauses and cross check with manual documents.

- Our model gave the calculation of payments and fees which is slightly differs from manual calculation.

- There is overlapped data extraction related to assets owner.

- Model classified type of contracts correctly.

2. Reviewing the process

We will review our model and testing process and verify that it meets our business objectives. We will review If any test code or model need improvement.

- Reviewing the wrong data fetch related to asset owners in previous task.

- Reviewing why the model calculated wrong payments in fees in last step.

- Reviewing clauses in manual with model attributes.

3. Determining the next steps

If needed we may go back to business understanding process until we get satisfactory results. If the output does not fulfill our goals we may go back to the modeling process.

- Suppose the model fail to classify clauses or fetch wrong custody agreements data for correlation then again we will improve our model and try.

- Our model is calculating upcoming payments incorrectly so we will work on data preparation and modeling.

- Classifying model missed some of the main clauses we are updating our code for that model.

- If we get optimal results then we will determine to continue to deployment phase.

Phase 6: Deployment

1. Planning deployment

We will inform our team and the bank's team for deployment planning. We will take help from experts and keep our team in touch. We will ensure the management system is compatible with this change.

- First, we deployed our models on sample documents and planned to implement it.

- Then, we deployed our model via a platform accessible to JP Morgan's team.

- We connected the system to our current document software so it's easy to use.

2. Monitoring and Maintenance

We should monitor if our software is in good work and classifying those documents as per bank's requirements. Collecting feedback in every case will be helpful. We get to define the maintenance for our models and after continuous monitoring we can conclude the success of our project work.

- Our team is tracking how well the model is working overtime using analytical tools.

-We are checking how accurate the classifications are every month and updating the model with new contracts for maintaining the good results.

3. Reviewing the project

We will see the results of all the above processes on JP Morgan and we will conclude overall impact of the solution on business. We will look for any required updates and changes and proceed to finalizing the project.

- We presented our results to JP Morgan and explained them how it will help them.

- Models are classifying the documents, calculating the payments and fees and minimizing mistakes in loan servicing.

- We achieved our data mining goal of time reduction and accurate work and project is ready to be finalized.

- We assigned a name for our modeling software as CoIn, Contracts Intelligence Software.

4. Finalizing the project

The final step of or project will be reviewing results, recording those for future reference, documenting all our planning, work, tools, techniques and all the methodology.

- We documented all our findings, including how we prepared the data, the models we used, and the results we got.

- We created a user guide to make it easy for JP Morgan's team to use CoIn software.

- Our team Handing over our work and well working CoIn software to JP Morgan and their team for implementation.

We used CRISP-DM framework to present documentation on how to teach computers to understand and sort legal documents for J.P. Morgan. We used smart tools to help the computers, like to read text from images and to understand the meaning of words.

This new system named as CoIn software makes the process much faster and more accurate. It saves time and reduces mistakes. It is a great way to handle the many complex legal documents that banks like J.P. Morgan deal with.

Link to Video: https://shorturl.at/Z7oRR