Course 1: Introduction to Data Analytics

Name: Anish Suresh Butle

Date: 08/07/2024

Project Title: J P Morgan classification for legal documents

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

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

First of all before breaking the process down using CRISP-DM (Cross-Industry Standard Process for Data Mining) we should know the six stages of CRISP-DM i.e., Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation & Deployment.

Now we will start breaking the problem statement into these six stages of CRISP-DM.

1. Business Understanding
   As we know that there is focus on objectives, goals and requirements at this stage. In this case with JP Morgan that requires to automate the process of classification of various legal documents to have more time review with fewer errors compare to manual errors as manually it takes 360000 person hours which can increase the chances of errors.

   Project Objectives: To reduce the amount of time spent for analysis and classification of legal documents, to minimize the error in loan agreement service, To improve the efficiency, To deal with new algorithm using machine learning to interpret new regulations.

   Project Goals: The System should maintain high Accuracy and efficiency in order to reduce the possibility of errors and Design a system that can handle these various legal documents for their classification.

2. Data Understanding
   So, at this stage it begins the data collection and familiarizes themselves with the data by exploring data, verifying data quality, and by describing data.

   Data Collection: According to the goals and objectives identified collect all the legal documents that is required for classification and state classification of legal documents and collect samples of each type.

   Data Description: Familiarize yourself with the data characteristics to have understanding for analysis and understand the metadata for each document for more adequate results.

   Data Exploration: Organize the collected data to interpret it and identify the missing values and perform the first level of data analysis for identifying the general distribution of various types of documents.

   Data Quality: Verify the data quality and evaluate the inconsistent data as well as the data duplication.

Course 1: Introduction to Data Analytics

3. Data Preparation
   In this state the model is dedicated to cleaning and transforming raw data into a suitable format for modelling.

   Data Cleaning and Transformation: In this the system must eliminate copies and preprocess the data, handle the missing values properly, Cleaning of the text data in a pre-processing step involves transforming the raw data to a standard format.

   Data Integration and Data Selecting: Transform this huge amount of data into a single set which can be used for further modelling and select the transformed data of suitable format for modelling.

4. Modelling
   In this we have to create a model that would be able to categorize different legal documents. The model name is COIN (Contract Intelligence).

   Model Selection: Select appropriate machine learning algorithm for text classification that can be applied and we can apply Natural Language Processing (NLP) and Deep learning model.

   Model Training: After selecting the appropriate model, we need to train that model after training, we need to build the model, perform certain tests on that model so that the model's accuracy will be increased and also the performance of the model will be enhanced.

5. Evaluation
   In this stage the model's performance is thoroughly evaluated and all the process which is carried till now is reviewed and determined what will be the next steps.

   Model and Business Evaluation: Test the final model also to check whether the model is working properly or not you can take different set of data and evaluate the performance of the system and check whether this model meets the business goals.

   Reviewing Process: All the stages is reviewed and checked whether it is functioning properly or not and also doing error analysis if error is found at the time of reviewing. The next step is also concerned with the identification of potential improvement based on the identified patterns of errors.

6. Deployment
   In this stage the model is deployed into a real-world environment and other process such as planning deployment, monitoring and maintenance of the model and finalizing the project is carried out.

   Deployment Planning: Make sure that the model easily support large number of documents and guide on what will be the best way to integrate the model in real-world.

   Monitoring: for this you can make alert systems to alert when the systems performance is degrading. Before deploying we need to finalize the project.

Course 1: Introduction to Data Analytics

After performing these six stages we can successfully deploy the model COIN (Contract Intelligence) into the real-world. And now we need to train the users how to use this model so that it can be implemented for the betterment of the company. By this way we can create a model which will automate the classification of various legal documents.

Link for the recorded video explanation: project1_link