BI/AI/MI project

Final Project

ABSTRACT

This project aims to optimize duplex PDF documents by detecting and removing unnecessary blank pages. Using job-provided PDF samples, it follows the CRISP-DM methodology: business understanding, data preparation, analysis, model development, evaluation, and deployment. The goal is to reduce processing times and improve document preparation efficiency.

Cesar Ortega
Bus Intel & Ds Sys CIDM 5310

BI/AI/ML Storyboard for PDF Blank Page Detection and Removal

Problem Statement

In the process of managing and preparing PDF documents for printing, unnecessary blank pages can lead to inefficiencies and unnecessary processing speeds. This project aims to develop a model that can accurately identify and remove unnecessary blank pages from PDF files of all sizes, ensuring documents are prepared optimally for duplex printing.

Objectives

1. Accurate Blank Page Detection: Develop a model that can identify blank pages in PDF documents with high accuracy.
2. Removal of Unnecessary Blank Pages: Ensure that only necessary blank pages are retained for proper duplex formatting.
3. Maintain Document Integrity: Ensure that the removal of blank pages does not affect the logical structure and readability of the document.

Stakeholders

- **Businesses**: Companies that need to prepare large volumes of documents for printing, looking to optimize processing speeds.
- **Print Service Providers**: Businesses offering printing services that need to ensure efficient and quick processes.
- **Document Managers**: Professionals responsible for managing and preparing documents for printing and distribution.
- **Software Developers**: Teams developing document management and printing optimization software.

Data Requirements

- **PDF Documents**: A diverse set of PDF documents in duplex format, with varying content and structure.
- **Page Metadata**: Information about each page, such as text content, images, and graphical elements, ect..
- **User Annotations**: Data from users indicating which pages are considered blank or unnecessary in their context.

Tools and Technologies

- **User Layer**: PowerBI for data visualization and interactive dashboards to monitor model performance and document preparation status.
- **PaaS Layer**: Azure for data handling, storage, and analysis, ensuring scalability and reliability.

- **Python Libraries**: PyMuPDF (fitz) for PDF processing and scikit-learn for model development.

Steps and Processes

1. **Data Collection**: Gather a variety of PDF documents with different structures and formats.
2. **Data Cleaning and Preparation**: Process the collected PDFs to extract text, images, and or metadata, and label pages as blank or non-blank.
3. **Exploratory Data Analysis (EDA)**: Analyze the data to understand patterns and characteristics of blank pages in different document types.
4. **Model Development**: Develop predictive models using machine learning techniques to accurately identify blank pages. Consider both image-based and metadata-based features.
5. **Model Evaluation**: Evaluate the performance of the models using metrics such as accuracy, precision, and recall to ensure reliable blank page detection.
6. **Visualization and Reporting**: Use PowerBI to create interactive dashboards and reports to visualize the model's performance and the status of document preparation.
7. **Deployment**: Deploy the predictive model on Azure, enabling real-time blank page detection and removal in document processing workflows.

Potential Challenges

- **Data Quality**: Ensuring the accuracy and consistency of the PDF data and annotations used for training.
- **Model Accuracy**: Developing a model that can accurately identify blank pages in a wide variety of document formats.
- **Integration**: Integrating the model with existing document management systems and workflows.
- **Scalability**: Ensuring the solution can handle large volumes of PDF documents and provide real-time processing.

Outcome and Benefits

- **Improved Efficiency**: Businesses can reduce processing steps and improve efficiency by removing unnecessary blank pages from PDFs.
- **Optimized Document Preparation**: Documents are prepared optimally for duplex printing, ensuring better resource utilization.
- **Enhanced User Experience**: An interactive PowerBI dashboard provides users with an intuitive way to monitor and manage document preparation.

- **Scalable Solution**: The deployment on Azure ensures the solution can scale to meet the needs of large organizations with high volumes of documents.

Data Sources

Primary Source: Provided Samples from My Job

Description:

- **Provided Samples from My Job**: These samples include a variety of duplex PDF documents used in day-to-day operations. These documents will serve as the primary dataset for training and validating the model.

Dataset Includes:

- **Duplex PDFs**: Documents formatted for double-sided printing, where every content page is followed by a back side, which could be blank or contain content.
- **Complex Documents**: PDFs with varying levels of graphical content, text, and invisible elements such as white text.

Current Implementation

Setup	Actions	Outcomes	Results
Describe the current process of managing PDF documents.	Describe current methods used to identify and handle blank pages.	Explain the current manual or automated methods used.	Describe the current outcomes of these processes.

Future Implementation

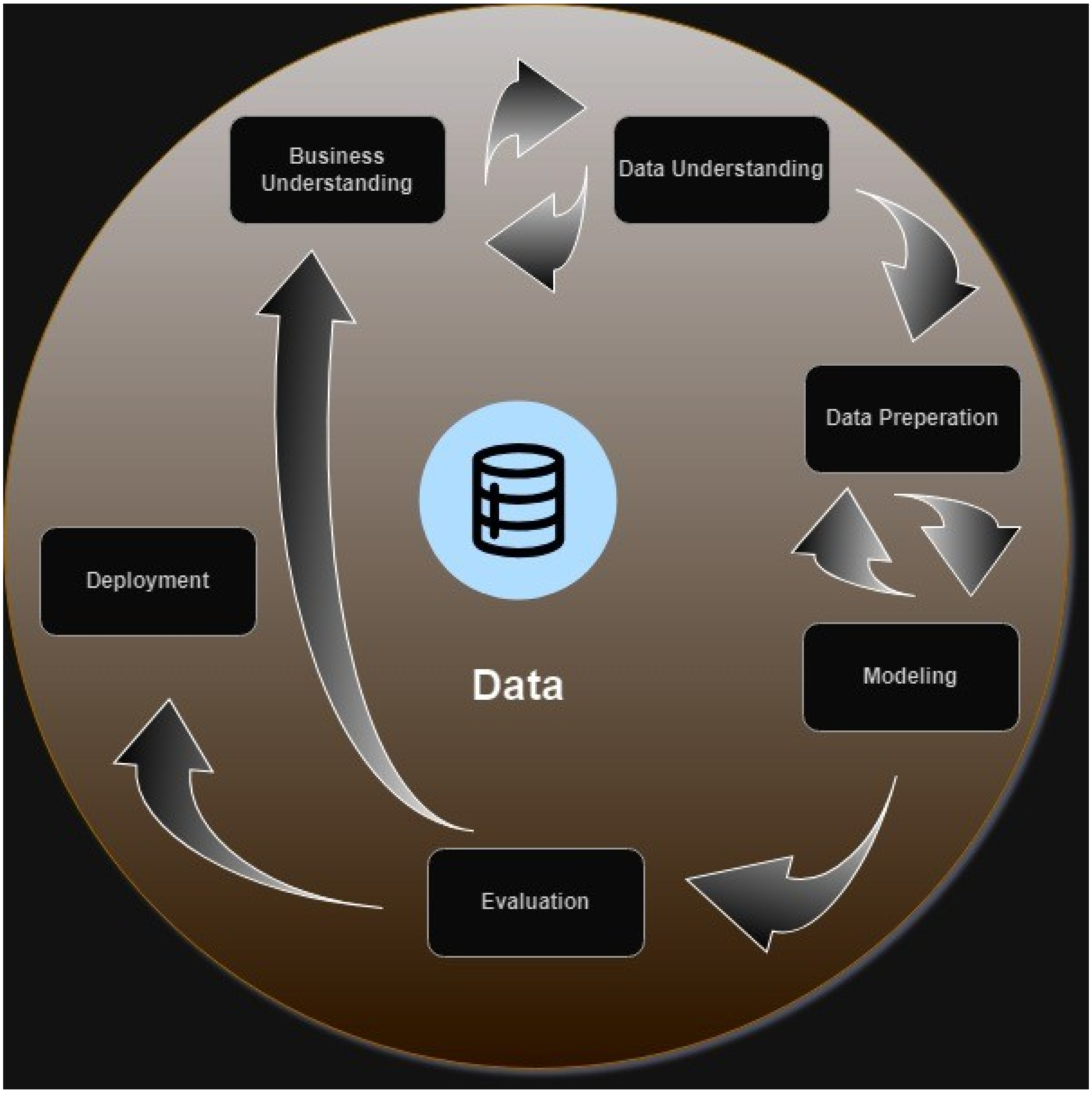
Setup	Actions	Outcomes	Results
Collect historical data on PDF usage, including blank and non-blank pages. Involve experts in document management and data analysis.	Analyze historical data to identify patterns in blank page occurrences. Develop a predictive model to identify unnecessary blank pages. Use explainable AI to understand why a page is classified as blank or non-blank.	Each PDF document will be analyzed, and unnecessary blank pages will be identified and flagged for removal. These insights will be integrated into document management systems.	The predictive model will enable real-time detection and removal of blank pages. This will improve document processing efficiency and reduce unnecessary processing steps.

4V Model Analysis for PDF Blank Page Detection and Removal Project

Data Source: Provided samples from my job.

1. **Volume**:
    - The dataset contains a substantial number of PDF documents, each potentially having multiple pages.
    - The provided samples cover a wide range of document types and formats used in day-to-day operations.
    - The volume of data is sufficient to train and validate the model for various scenarios of blank page detection and removal in duplex PDFs.
2. **Velocity**:
    - The dataset includes documents that are frequently updated and generated as part of regular business operations.
    - New samples are continuously added as more documents are processed, ensuring that the dataset remains current and relevant.
    - The ongoing addition of new data helps in maintaining the model's performance and relevance over time.
3. **Variety**:
    - The dataset comprises a diverse set of duplex PDF documents, including reports, invoices, manuals, and other business-related documents.
    - Each document varies in terms of content structure, graphical elements, text density, and potential invisible elements such as white text.
    - This variety ensures the model is exposed to different types of content and formatting, enhancing its ability to generalize and perform accurately across various document types.
4. **Veracity**:
    - The dataset is sourced from actual business operations, ensuring that the data is representative of real-world scenarios.
    - Initial inspection shows a high level of accuracy, with clear labeling of blank and non-blank pages based on business requirements.
    - While there may be occasional inconsistencies, the overall data quality is high, making it suitable for developing a reliable and robust model for blank page detection and removal.

CRISP-DM:


Business Understanding
Data Understanding
Data Preperation
Modeling
Evaluation
Deployment
Data