WGU C951

Task 3

MACHINE LEARNING PROJECT PROPOSAL

Kevin Joseph

Student ID #

Date

**A. Project Overview**

This proposal describes a machine learning solution that can identify documents that are missing critical information to flag for human review

**A.1. Organizational Need**

The organization needs to implement a machine learning solution that can identify documents that are missing critical information to flag for human review. They need it to be accurate and get better over time on identifying these cases to save them money on extensive human review.

**A.2. Context and Background**

The project is based on a intelligent document processing whitepaper from AWS. The solution is designed to be able to read from a document and identify information that needs to be present but is missing so that it can be flagged for human review. The company has been losing money and trust from their customers because of delivering documents that are missing critical information. We are tasked with developing the solution for this problem.

**A.3. Outside Works Review**

The AWS solution relates to my solution because it provided an end to end solution in the cloud space that can extract text from docs, process those docs, and then classify them based on if they are missing information.

The Kili solution relates to my solution because it provides information on how the process of training a dataset to the preperation of data for the model is conducted. It also provided info on how to label data for the model to easily classify.

The altexsoft solution relates to my solution because it gives various examples of real world use cases of document classification in action and how those solutions were built. I can use the examples to architect my own solution for the company.

**A.4. Solution Summary**

We need to develop a machine learning classifer that can detect when a document is missing critical information. The classifer would be continously fed data to improve its accuracy over time and be maintained to prevent model drift. The classifer would have end to end logic that will extract text from the document, process that text, and then classify that text through a ETL pipeline.

**A.5. Machine Learning Benefits**

Machine Learning would make the manual human review that is currently in place automated and much faster. It will also be able to learn on the fly and adapt to changes to the structure of the document.

**B. Machine Learning Project Design**

**B.1. Scope**

* Use bullet points to describe in scope and out of scope items. A minimum of 1 out of scope item must be discussed, and 3 in scope items.
* In Scope: The project will provide a solution that will extract text from the document.
* In Scope: The project will provide a solution that will be able to use ETL for the documents until it is flagged for further review.
* In Scope: The project will provide a solution that will have a ready stream of training data that will periodically train the model over time, which will increase the accuracy.
* Out of Scope: The project will not have the capability to conduct final reviews for the information personally without human assistance.

**B.2. Goals, Objectives, and Deliverables**

Goals

• The solution must build upon the existing ETL pipeline to get all the documents neccessary for training and processing.

* The cost should cost less than the total cost of hiring humans to manually review the documents.

Objectives

• The model must be atleast 90% accurate at identifying missing information

* The model should be up 24/7 to allow for continous data processing.

Deliverables

• The solution will provide a automated document procesisng system that will flag documents that have missing information for additonal human review.

* The solution will provide an enhanced ETL pipeline that will provide the claims team all the neccessary data/ documents as well as metadata to improve the exisiting claims validation system.

**B.3. Standard Methodology**

Development will follow the SEMMA methodology.

• Sample: We will gather documents for training the model. The model needs documents that have instances of complete information and instances of incomplete information

• Explore: We will analyze the gathered documents and predict patterns that occur often.

• Modify: We will modify the documents to better fit the training parameteres of the model so that it would be able to easily flag the documents that are missing information.

• Model: We will use the machine learning classifier to train on the labeled data and flag the missing information.

• Assess: We will use the F1 metric as our accuracy metric. This is to make sure we have a standard and industry grade metric to base our accuracy on.

**B.4. Projected Timeline**

**Sprint Schedule**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sprint** | **Start** | **End** | **Tasks** |
| 1 | nov 1 | nov 30 | create AI framework |
| 2 | dec 1 | dec 15 | build and train data pipeline |
| 3 | dec 15 | dec 31 | do validation testing |
| 4 | jan 1 | jan 18 | final review and deployment |

**B.5. Resources and Costs**

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| GPU | this is the hardware to train our model | $20,000 |
| AWS storage | this is the host the data for processing | $1000 |
| Software Engineer | this is the person that builds our model | $50000 |
|  | **Total** | $71,000 |

**B.6. Evaluation Criteria**

|  |  |
| --- | --- |
| **Objective** | **Success Criteria** |
| model accuracy | We need the model to recognize missing info atleast 90% of the time |
|  |  |
| continous data processing | we need to process each document through the model within 30 seconds continously throughout the day |

**C. Machine Learning Solution Design**

**C.1. Hypothesis**

The machine learning classifer can identify missing information documents reducing the need for manual review and increase the efficiency of the exisiting process by 50%.

**C.2. Selected Algorithm**

We will implement the supervised learning algorithm. The random forest algothim is the best algorithim to use in this case.

**C.2.a Algorithm Justification**

I chose this algothim because it is the best algorithim for identifying patterns in documents.

**C.2.a.i. Algorithm Advantage**

This algorithim is the best because it makes final predicitions based on smaller decisions which will make our model more accurate.

**C.2.a.ii. Algorithm Limitation**

The algorithim might take a long time processing the documents because of the amount of decision it has to make for each prediction.

**C.3. Tools and Environment**

We are going to use NVIDIA gpus to train our model.

We will run on the AWS cloud enviroment for our storage needs.

We will use tensorflow for building our model as a programmer.

We will use linux as our operating system for the best compatibility.

**C.4. Performance Measurement**

The training phase involves using the F1 score metric as our evalution benchmark so that we can validate that our model is accurate. The validation phase will involve continous testing of our model to ensure the model drift is within an appropiate range, and if not then we would do retraining. The production phase would involve we are running the model continously 24/7. In terms of ROI and profitability we will evaluate the total cost of the model and compare it to the respective cost savings against employing humans to manually review.

**D. Description of Data Sets**

**D.1. Data Source**

The data for the model would come from the company's existing document database. We will use the avaliable dataset for in house training and make sure our data is secure within the confines of the company.

**D.2. Data Collection Method**

We will run an ETL pipeline from the company database to use in our model in a way that is efficent and in an usable format for the model to easily understand.

**D.2.a.i. Data Collection Method Advantage**

The advantage of using an ETL pipeline to load our datasets is that we are able to scale our pipeline to meet the needs of the model. If the model needed more documents to increase the accuracy, we are able to increase the throughput of the ETL pipeline to supply the model with the neccessary amount.

**D.2.a.ii. Data Collection Method Limitation**

The internal document database can have documents in a format that may not be compatible with the model, so this means we would have to transform the data in a way that is digestable by the model which would increase the time required for processing and the need for compute resources.

**D.3. Quality and Completeness of Data**

The data would be prepared for the algorithim by using th ETL pipeline to load the data into the model. We will make sure to have quaility assurance test cases in place that check to see if the documents are blank or in a corruputed format. The labeling process would be through with explict labels for each type of document to ensure accuracy.

**D.4. Precautions for Sensitive Data**

We will make sure to encrypt the files while at rest and also at transit so that outside entities cannot access it. For access to any cloud microsystems we will use the AWS guideline of least privilege to ensure that employees are only able to access data which they need to and nothing more.

**References**

amazon web services. (n.d.). machine learning for document processing. AWS. retrieved october 28, 2024, from <https://aws.amazon.com/machine-learning/ml-use-cases/document-processing/>

kili technology. (n.d.). machine learning for unstructured document analysis: a guide. retrieved october 28, 2024, from <https://kili-technology.com/data-labeling/machine-learning/machine-learning-for-unstructured-document-analysis>

altexsoft. (n.d.). document classification with machine learning. retrieved october 28, 2024, from <https://www.altexsoft.com/blog/document-classification/>