CS Capstone Problem Statement

October 26, 2017

CDK Data Stream AI

Prepared for

CDK Global

Chris Smith _____ _____
Signature *Date*

Prepared by

Group 65

Jacob Geddings _____ _____
Signature *Date*

Inhyuk Lee _____ _____
Signature *Date*

Juan Mugica _____ _____
Signature *Date*

Abstract

Our team has been assigned to assist in the development of AI for application to CDKs existing Data Streams. The goal in doing this is to gain insight from said data streams, use this data to predict future events, and detect when an anomaly is present. These three goals need to be independent of one another and be capable of functioning as a black box. This entails the ability to write applications and/or functions that are then fed through the system in various ways. Tools that will be necessary for project completion include the AWS platform, Docker, Linux platform, and several open source AI resources.

Contents

1 Main Goal

ROUGH DRAFT PRE-MEETING WITH CLIENT

CDK as a dealership oriented company routinely deals with the management of thousands of documents. Given the diverse nature of how these are gathered, often incorrect, incomplete, or lacking key information. Moreover, many of these documents are not gathered electronically, making it basically impossible to utilize existing computing resources to address these issues. This leaves CDK with only one option, to waste human resources in performing very mundane and automatable tasks. Our job for this project is to create standalone software that can parse pictures in PDF format to gather from these details of the following nature:

- Is the document signed.
- Should the document be signed, if so, in how many places. More so, where are there signatures missing, how many are missing and are these necessary to the integrity of the document.
- Information regarding different legal forms contained within the document which will first be presented in an image format but must be parsed into text. These include, but are not limited to:
  - License information
  - Permit information

Due to a wide ranging number of formats for different types of forms,the program is to utilize modern machine learning techniques, more specifically neural networks and neural network capabilities. To do so it is required that existing machine learning frameworks be used. Some examples include DL4J, TensorFlow, and OpenCV. The program is to be deployed on many different machines, making it essential that it be compatible with cross platform container software such as Docker. Given that the software is to learn from a large number of data points that will be provided by our client, it is also essential that it be able to run without problems for extended amounts of time. The software must be coded with scalability in mind; It should keep track of previously parsed documents and be able to consistently classify new documents into the aforementioned categories (fully signed, lacking signatures, lacking space for a signature / not applicable).

Once the program structure and functionality is up to the client's standard in regards to basic functionality (85% classification success rate) , it must become cloud deploy able utilizing Amazon Web Services or services of similar nature.

2 Stretch Goals

Once the software can perform ALL of the aforementioned tasks with an industry satisfactory success rate (¿95% classification success rate), our client is interested in expanding the software to outline key features pertaining to pictures of vehicles clients send in. These include but are not limited to :

- The model of the vehicle
- The make of the vehicle
- License plate information
- Legal information regarding such licenses
- The condition of the vehicle: dents, cracks, scratches...etc.

Given that this is a stretch goal, our client put extra emphasis on taking it one feature at a time. These features are outlined in a manner where they each provide a level of significance to the overall project .The software must first be able to detect that the picture is of a car. Once it can satisfactorily do this (again, 85% classification success rate), we may move on to the next feature: detecting the license plate, and parsing it into machine readable text. If we are able to achieve a satisfactory rate in this aspect, we are then tasked to move on to detecting the make and model of the vehicle. Lastly, if all other milestones have been successfully completed, we are tasked with detecting key features regarding the vehicle's condition.