My project is about Loan application process.

I would like to discuss about my client first. He was from banking sector. He came with an idea that he wanted to identify the genuine customers whom bank could trust and provide them loan.

> With the increase in banking sector many people are applying for loans in bank. All these loans are not approvable.
> The main source of income in banking sectors are loans. The main objective of banks is to invest their assets in safe customers.
> Today many banks approve loan after many process of verification and validation but still there is no surety that selected customer is safe or not.
> Therefore it is important to apply various techniques in banking sector for selecting a customer who pays loan on time.

Purpose of this project is to help financial industry or bank to collect all the information of customers in a place ,So that bank will easily identify who is applicable for loan or not.

I started this project by collecting all the important documents that was required i.e. Aadhar Card, Bank Statement, Pan Card and Salary Slip. That was the most important and difficult part in my project.

Because we did not have continuous flow of data. Data cycle was not good. And the exact data was not available for our use case and it took us around few months to collect all the data or documents.

We made different folder for different documents and store particular documents in particular folders.

Important information like , name , address, date of birth, gender, pan number, Aadhar number, salary slip , etc.

For Name, Date of Birth, Address and Gender I used Aadhar card.

For Pan Card number and fathers name I used Pan Card.

To get Total salary I used salary slip and

For Account number I used Bank Statement.

These are the most essential information that bank need to proceed loan application process. Now I have blue print from where I could get all the information, here I had everything but in image form.

So, the first step was to convert image into text and to do that I used OCR i.e. Optical Character Recognition.

OCR= OCR stands for "Optical Character Recognition."

It is a technology that recognizes text within a digital image.

It is commonly used to recognize text in scanned documents and images.

There are lots of OCR tecnologyies…

pyTesseract , google vision, amazon tesseract

and we used Pyteseract because it is flexible and free to use.

First we installed tesseract and the most important thing was the path that we needed to remember while we were installing the tesseract.

Because when we use wrong path in code, it will not work.

Now we needed to import some python libraries like

Pytesseract,

Pdf2image and

From pdf2image import cover_from_path

To extract all information from text I needed to do regex and for that I need to import some more libraries. Like…

1. **Import re** (it is a sequence of character used to check whether the pattern is in the given string or text , and to perform different regex operation we can use different regex functions such as …. Re.findall(), re.match(), re.sub(), re.research(), re.compile() )

2. **Import OS** (The OS module in Python is a part of the standard library of the programming language. With the help of this library we create a folder , access the path and fetching its contents.)

3. **Import panda as pd** ( pandas is a python library used for working with data sets and it has a function for analysing, cleaning, manipulating and exploring the data.
   Here we use the panda to create a database and with this database we convert this into csv file.

   When we start writing a regex code for a particular information
   We need to pay attention:
   For example : we want to extract name from the particular document
   Here we faced different format of name for example,
   Some customer has two words name , some has three words name.
   For date of birth …we have also different format
   We need to write codes according .
   Once we got all information we made the features accordingly

After that I created a Dataframe that contains all the information in rows and columns. and at the end I stored all data in an excel file or csv file.