Phase :5

Final project submission

Ai -driven exploration and prediction of company registration trends with registrar of companies(ROC)

|  |  |
| --- | --- |
| Name | J.JEEVITHA |
| Date | 01.11.2023 |
| Team ID | Proj\_2121\_Team\_4 |
| Project Name | Registrar of Companies(ROC) |
| Maximum marks |  |

Index:

1. [List of figures and tables: 3](#_Toc149739308)
2. [Abstract 3](#_Toc149739309)
3. [Introduction: 4](#_Toc149739310)
4. [Literature and survey: 4](#_Toc149739311)
5. [Problem definition 5](#_Toc149739312)
6. [Design thinking 7](#_Toc149739313)
7. [Importing the dataset and performing data cleaning and analysis 8](#_Toc149739314)
8. [Data visualization 9](#_Toc149739315)
9. [Module development and evaluation: 11](#_Toc149739316)
10. [code sample: 13](#_Toc149739317)
11. [output Screenshots: 16](#_Toc149739318)
12. [Conclusion and enhancement: 18](#_Toc149739319)

# 

# List of figures and tables:

1. [Problem definition 5](#_Toc149756559)
2. [Design thinking 7](#_Toc149756560)
3. [Idea Prioritization: 8](#_Toc149756561)
4. [Importing the dataset and performing data cleaning and analysis 8](#_Toc149756562)
5. [Data visualization 9](#_Toc149756563)
6. [For plotting the dataset: 9](#_Toc149756564)
7. [Scatterplot using x-label and y-label: 10](#_Toc149756565)
8. [Module development and evaluation: 11](#_Toc149756566)
9. [code sample: 13](#_Toc149756567)
10. [output Screenshots: 15](#_Toc149756568)

# 

# Abstract

In this naanmudhalvan project I have learned about data exploration and prediction using (Registrar of companies ) .

The Registrar of Companies ( ROC ) is an office under the Ministry of Corporate Affairs (MCA), which is the body that deals with the administration of companies and Limited Liability Partnerships (LLPs) in India. At present, Registrar of Companies (ROCs) are operating in all the major states/UT’s.

However, states like Tamil Nadu and Maharashtra, have more than one ROC. Some ROCs have jurisdiction of two or more states/UT like Chennai ROC has jurisdiction of Tamil Nadu state and UT of Andaman and Nicobar Islands.

As per section 609 of the Companies Act, 1956, the ROCs are tasked with the principal duty of registering both the companies and LLPs across the states and the union territories. Currently, after the introduction of Companies Act, 2013, the same powers conferred under section 609 is provided under section 396 of the Companies Act, 2013 to the ROCs.

The Registrar of Companies also certifies that LLPs (Limited Liability Partnerships) comply with the legal requirements contained in the Limited Liability Partnership Act, 2008.

# Introduction:

Registrar of Companies maintains a registry of records concerning companies which are registered with them and allows the general public to access this information on payment of a stipulated fee. The Central Government preserves administrative control over the Registrar of Companies with the help of Regional Directors. As of today, there are seven Regional Directors, supervising the operations of ROCs within their relevant regions.

# Literature and survey:

##### **Definitions: Classification and Regression**

Below, we jump into explaining the different types of supervised learning algorithms. All of them can either be used for classification or regression, or both.

**Classification** means an either/or result using binary (0 = no, 1 = yes). So the algorithm will classify something as either one or another, but never both. There is also **multi-class classification**, which deals with organizing data into defined categories or types relevant to a specific need.

**Regression** means the result will end with a real number (either round or a decimal point). You usually have a dependent variable and an independent variable, and the algorithm will use both points to estimate a possible other result (either forecast or generalized estimate).

#### **Decision Tree**

One of the most common supervised learning algorithms, decision trees get their name because of their tree-like structure (even though the tree is inverted). The “roots” of the tree are the training datasets and they lead to specific nodes which denote a test attribute. Nodes often lead to other nodes, and a node that doesn’t lead onward is called a “leaf”.

Decision trees classify all the data into decision nodes. It uses a selection criteria called Attribute Selection Measures (ASM) which takes into account various measures (some examples would be entropy, gain ratio, information gain, etc). Using the root data and following the ASM, the decision tree can classify the data it is given by following the training data into sub-nodes until it reaches the conclusion.

#### **Naive Bayes**

The reason this algorithm is called “Naive Bayes” is that it’s based on Bayes’ Theorem, and also relies heavily on a large assumption: that the presence of one particular feature is unrelated to the presence of other features in the same class. That major assumption is the “naive” aspect of the name.

Naive Bayes is useful for large datasets with many different classes. It, like many other supervised learning algorithms, is a classification algorithm.

# Problem definition

# The problem is to perform an AI-driven exploration and predictive analysis on the master details of companies registered with the Registrar of Companies (RoC). The objective is to uncover hidden patterns, gain insights into the company landscape, and forecast future registration trends. This project aims to develop predictive models using advanced Artificial Intelligence techniques to anticipate future company registrations and support informed decision-making for businesses, investors, and policymakers.

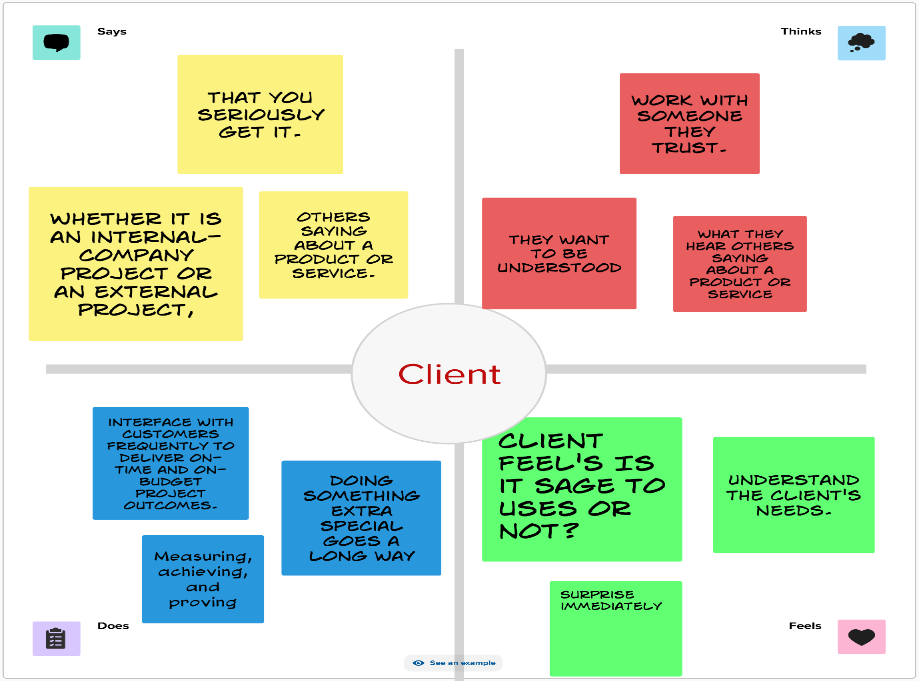
Problem Statement:

|  |  |
| --- | --- |
| Biggest challenges | One of the biggest challenges of machine learning prediction models is their inability to generalize beyond the data they were trained on. |
| Errors of prediction | Errors of prediction are defined as the differences between the observed values of the dependent variable and the predicted values for that variable obtained using a given regression equation and the observed values of the independent variable. |
| Predictive models are positive or negative | ROC curves are a nice way to see how any predictive model can distinguish between the true positives and negatives. In order to do this, a model needs to not only correctly predict a positive as a positive, but also a negative as a negative |
| Graphs | An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate. |
| Training datas | In other words, they may struggle to make accurate predictions when presented with data that is significantly different from the training data. |

## 

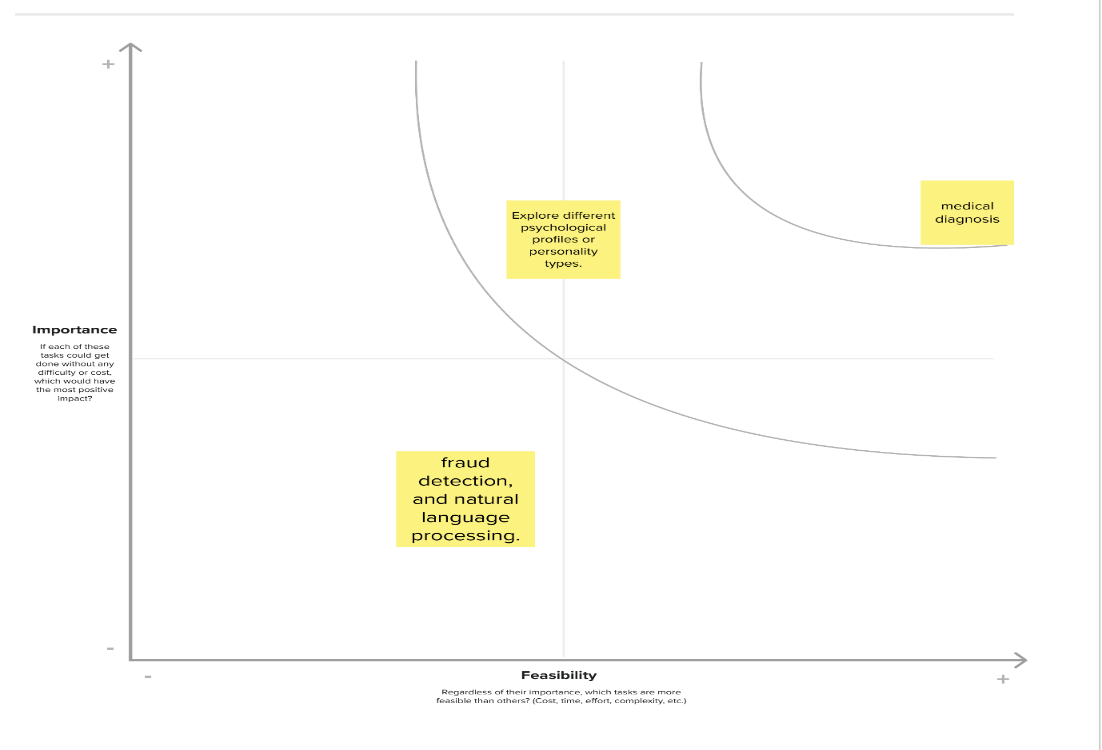
## Design thinking

The first step is to empathize with the stakeholders involved in the RoC's services. This includes entrepreneurs, business owners, investors, government officials, and the public. Conduct interviews, surveys, and observations to gain insights into their needs, pain points, and expectations when interacting with the RoC.



## Idea Prioritization:

Prioritization is a critical process for individuals and organizations to allocate their time, resources, and efforts effectively. It involves determining which tasks, projects, or goals are most important and should be tackled first.



# Importing the dataset and performing data cleaning and analysis

***#Importing the packages needed for the above given problem***

import numpy as np

import pandas as pd

➢Here, you are importing the pandas library with the alias "pd," which is a common practice. However, you also attempted to import pandas with the alias "np," which is usually used for NumPy, another popular Python library. It's better to use "pd" consistently for pandas

***#reading the dataset as csv file***

df=pd.read\_csv(r"C:\Users\91908\Downloads\Data\_Gov\_Tamil\_Nadu.csv")

***#cleaning the data which is unwanted for the dataset***

Code:

df.COMPANY\_SUB\_CATEGORY =df.EMAIL\_ADDR.fillna("unknown") print(df.isnull().sum())

output:

CORPORATE\_IDENTIFICATION\_NUMBER 0

COMPANY\_NAME 0

COMPANY\_STATUS 0

COMPANY\_CLASS 334

COMPANY\_CATEGORY 334

COMPANY\_SUB\_CATEGORY 0

DATE\_OF\_REGISTRATION 39

REGISTERED\_STATE 0

AUTHORIZED\_CAP 0

PAIDUP\_CAPITAL 0

INDUSTRIAL\_CLASS 310

PRINCIPAL\_BUSINESS\_ACTIVITY\_AS\_PER\_CIN 0

REGISTERED\_OFFICE\_ADDRESS 90

REGISTRAR\_OF\_COMPANIES 174

EMAIL\_ADDR 38129

LATEST\_YEAR\_ANNUAL\_RETURN 75889

LATEST\_YEAR\_FINANCIAL\_STATEMENT 75782

dtype: int64

# Data visualization

## For plotting the dataset:

Code:

sns.FacetGrid(df,hue="AUTHORIZED\_CAP",height=4)\

.map(sns.histplot,"LATEST\_YEAR\_ANNUAL\_RETURN")\

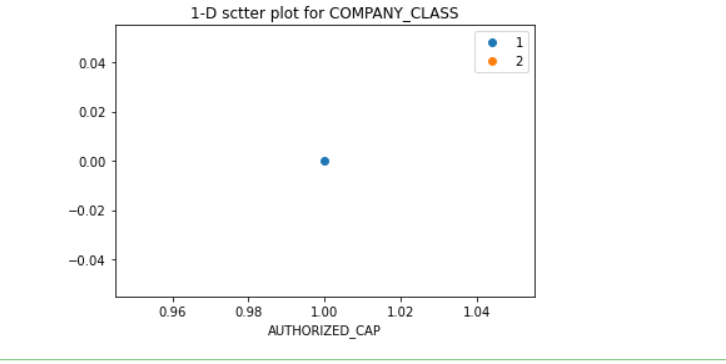
.add\_legend();

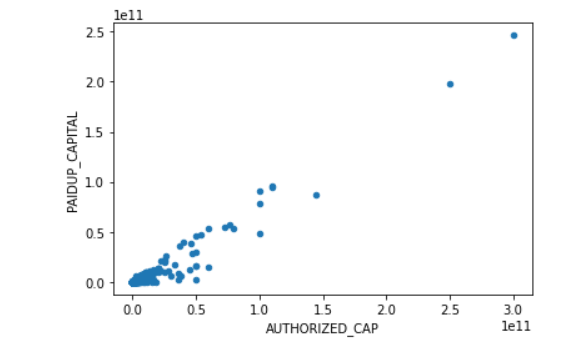
plt.show();

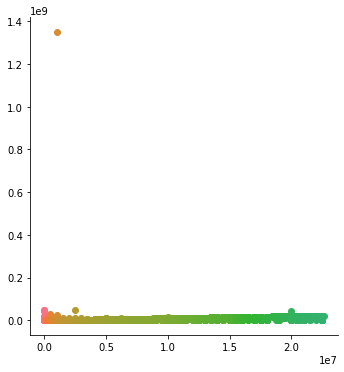
output:



## Scatterplot using x-label and y-label:







Using some types of plots , I have executed data visualization of my dataset

All the code and the plots(output) are executed successfully.

# Module development and evaluation:

Algorithm :

Step1:

Start the program .

Step2:

Now import all the packages in the program

Like pandas ,numpy , matplotlib , seaborn and etc

that you want .

Step3:

Next step is to execute the csv file for our given dataset and then go to step 4.

Step4:

In this step train the dataset and test it for more than one time to get some clear idea.

Step5:

After completing these steps just display you dataset.

Step6:

Stop the program.

*Source code:*

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn import metrics

from sklearn import datasets

from sklearn.metrics import r2\_score

*csv file execution:*

df=pd.read\_csv(r"C:\Users\91908\Downloads\Data\_Gov\_Tamil\_Nadu.csv")

*here the data training and testing is performed:*

train\_df=pd.read\_csv(r"C:\Users\91908\Downloads\Data\_Gov\_Tamil\_Nadu.csv")

test\_df=pd.read\_csv(r"C:\Users\91908\Downloads\Data\_Gov\_Tamil\_Nadu.csv")

train\_df.columns

*output:*

Index(['CORPORATE\_IDENTIFICATION\_NUMBER', 'COMPANY\_NAME', 'COMPANY\_STATUS',

'COMPANY\_CLASS', 'COMPANY\_CATEGORY', 'COMPANY\_SUB\_CATEGORY',

'DATE\_OF\_REGISTRATION', 'REGISTERED\_STATE', 'AUTHORIZED\_CAP',

'PAIDUP\_CAPITAL', 'INDUSTRIAL\_CLASS',

'PRINCIPAL\_BUSINESS\_ACTIVITY\_AS\_PER\_CIN', 'REGISTERED\_OFFICE\_ADDRESS',

'REGISTRAR\_OF\_COMPANIES', 'EMAIL\_ADDR', 'LATEST\_YEAR\_ANNUAL\_RETURN',

'LATEST\_YEAR\_FINANCIAL\_STATEMENT'],

dtype='object')

# code sample:

Code:

classes=pd.value\_counts(df["EMAIL\_ADDR"])

classes.plot(kind='bar')

plt.title("Class distribution Histogram")

plt.xlabel("COMPANY\_STATUS")

plt.ylabel("COMPANY\_CLASS")

plt.show();

sns.FacetGrid(df,hue="AUTHORIZED\_CAP",height=4)\

.map(sns.histplot,"LATEST\_YEAR\_ANNUAL\_RETURN")\

.add\_legend();

plt.show();

one=df[df['AUTHORIZED\_CAP']==1]

two=df[df['PAIDUP\_CAPITAL']==2]

label=['1','2']

plt.plot(one["AUTHORIZED\_CAP"],np.zeros\_like(one["AUTHORIZED\_CAP"]),'o')

plt.plot(two["AUTHORIZED\_CAP"],np.zeros\_like(two["AUTHORIZED\_CAP"]),'o')

plt.title("1-D sctter plot for COMPANY\_CLASS")

plt.xlabel("AUTHORIZED\_CAP")

plt.legend(label)

plt.show()

df.plot(kind='scatter',x="AUTHORIZED\_CAP",y="PAIDUP\_CAPITAL")

plt.show()

sns.FacetGrid(df,hue="AUTHORIZED\_CAP",height=5).map(plt.scatter,"AUTHORIZED\_CAP","PAIDUP\_CAPITAL").add\_legend()

plt.show()

train\_df.describe()

print(df.shape)

print("\n")

print(df.dtypes)

# get the locations

X = df.iloc[:, :-1]

y = df.iloc[:, -1]

# split the dataset

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

X, y, test\_size=0.05, random\_state=0)

print(X,y)

df.info()

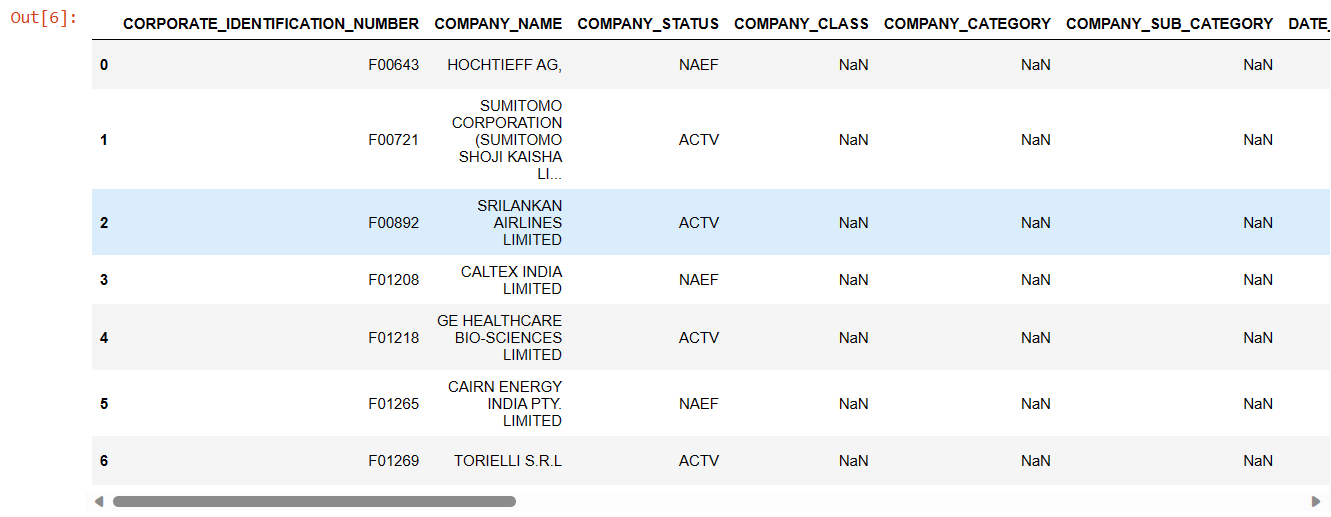
print('\_'\*40)

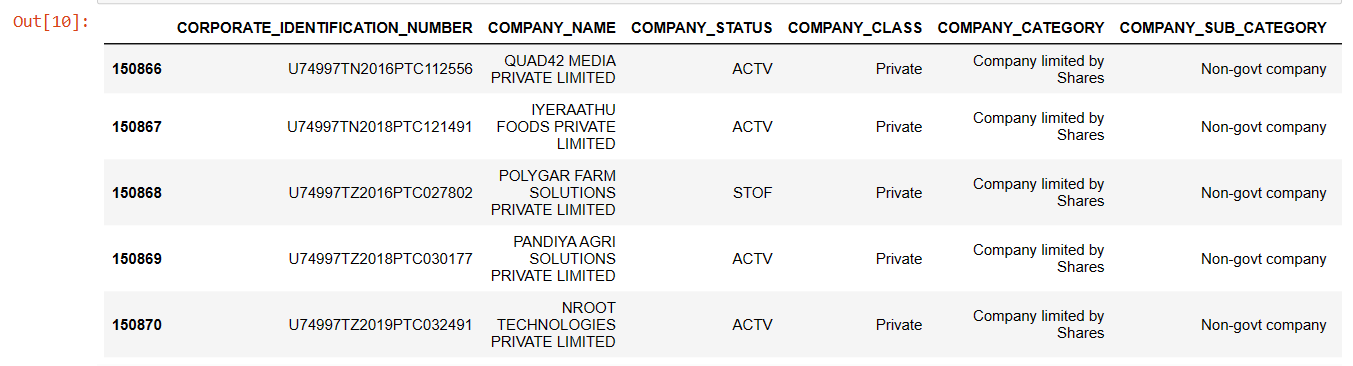
df.info()

df.COMPANY\_SUB\_CATEGORY =df.EMAIL\_ADDR.fillna("unknown")

print(df.isnull().sum())

## output Screenshots:





(150871, 17)

CORPORATE\_IDENTIFICATION\_NUMBER object

COMPANY\_NAME object

COMPANY\_STATUS object

COMPANY\_CLASS object

COMPANY\_CATEGORY object

COMPANY\_SUB\_CATEGORY object

DATE\_OF\_REGISTRATION object

REGISTERED\_STATE object

AUTHORIZED\_CAP float64

PAIDUP\_CAPITAL float64

INDUSTRIAL\_CLASS object

PRINCIPAL\_BUSINESS\_ACTIVITY\_AS\_PER\_CIN object

REGISTERED\_OFFICE\_ADDRESS object

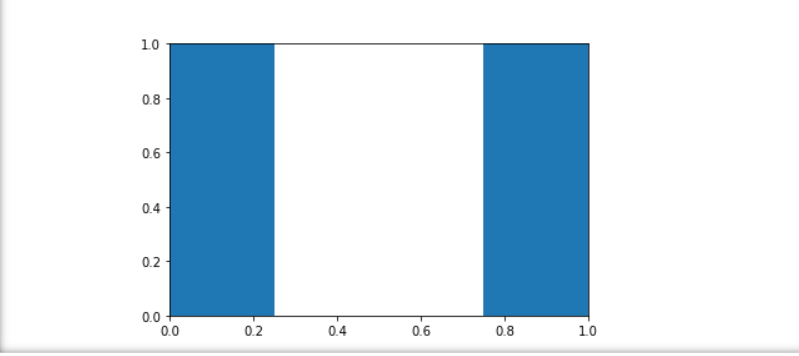
REGISTRAR\_OF\_COMPANIES object

EMAIL\_ADDR object

LATEST\_YEAR\_ANNUAL\_RETURN object

LATEST\_YEAR\_FINANCIAL\_STATEMENT object

dtype: object



output:

ganravi@gmail.com 182

compliance@kanakkupillai.com 176

secretarial@stjohntrack.com 161

smrajunaidu@gmail.com 144

pcschn1@gmail.com 133

...

info@skymaxlogistics.com 1

vishnu2444@yahoo.com 1

rashahuljob@gmail.com 1

baskar.mrl@gmail.com 1

nroottechnologies@gmail.com 1

Name: EMAIL\_ADDR, Length: 79940, dtype: int64

# 

# Conclusion and enhancement:

In this phase 5 final project submission I have learned the following topics are problem definition, design thinking , problem solving ,importing data and data cleaning and data visualization ,model development and evaluation.

And also I get some more knowledge and ai exploration and prediction of the registrar of the companies (ROC).