

**Credit Card**

**Fraud**

**Detection**

Machine Learning Project

-

Amit Parmar

**Objective:**

•

In today's world we are on the express train to a cashless society.

Whenever we say cashless, credit card is one of the best way for

transaction. But as many people are becoming the

victim of fraud so

it is important to notice this kind of fraud transactions through credit

card details.

•

The primary goal of

my machine learning web app is

to determine

whether a customer is

trying to perform a fraudulent transaction or

not.

**Benefits:**

•

Manual inspection if fraud is identified.

•

Detection of upcoming fraud.

•

Gives better insight of customer base.

•

Prevents

customer

from fraud traps.

**Data Sharing Agreement**

❖

Sample file name(CrediCardFraud.csv)

❖

Length of date stamp(8 digits)

❖

Length of time stamp(6 Digits)

❖

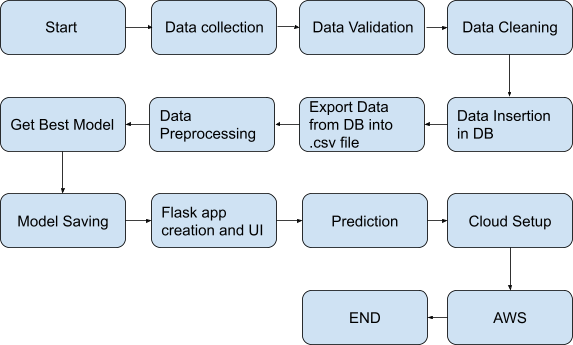
Number of columns

❖

Column names

❖

Column data type

**Architecture**

**Data Validation:**

•

File name Validation: File name validation as per the DSA.

•

Name and No. Of columns: It will check for number

of columns and

Name of the columns.

•

Data types of

columns: The datatype of columns is given

in the

schema file.

•

Null Values

:

If any columns in

the

data contains null values

then

the

respective

detail

of the transaction will be

dropped.

**Database:**

Database creating and connection: Create the database

with

the key space name passed.

Insertion: The dataset in the form of csv is inserted into

the

database

**Model Training:**

**Data Export from DB**

:

The data is

stored in the database which will be

exported for further model training purpose.

**Exploratory Data Analysis and Data Preprocessing:**

•

Missing value count

•

No of

rows and columns

•

Categorical/Numerical Columns

•

Null value handling

•

feature Selection using

selectKbest

**Model Selection**

•

Compute metrics and generate graphs for model evaluation

and important analysis.

•

I viewed AUC Values

for each model

and plot the

ROC curves

•

After testing several classification algorithms and comparing

their performance, Logistic Regression was selected for our

model building

with an accuracy of around 90%.

**Prediction**

•

The testing files are shared and we perform the same

validation operations, data transformation and

data

insertion on them.

•

The accumulated data from database is exported in csv

format for

prediction.

•

We perform

data pre

-

preprocessing techniques in it.

**Q&A:**

1.

Explain about the Project.

Credit Card Fraud Detection Web app is a machine

learning

Web app designed to detect where a transaction performed

using a credit card is fraudulent or not. It will help to identify

some serious fraud performed using credit cards. Based on

the details of the transaction provided by the

user the model

will predict

whether the transaction is fraudulent or not.

2

. What's the source of data?

The dataset is taken from

kaggle

problem statement.

Click to add text

3.

What was the data type?

All the

columns contains numerical values only.

4.

What is the

complete flow you followed in this

Project?

Refer slide 5th for the better Understanding.

5.

How logs are managed?

Here I am using different Logs as per the

steps

that I followed in

validation

and

modeling like file validation log, data insertion log, model training

log,

predicting log

etc.

6.

What is the size of data?

The size of the dataset is 152Mb.

o Perfoming feature selection using selectkbest

5

.How logs are managed?

I am using different logs

as per the steps that I followed in modelling like Data

insertion, model training log, prediction log etc.

6

.What techiniques are you using for data preprocessing?

o

Visualizing relation of independent variables with each other and

with

dependent

variable.

o

Cleaning data and removing the null

rows.

o

Checking

and changing Distribution of continuous values.

o

Removing

outliers

o

Handling the

unbalanced data

7.How training was done or what models were used?

Before training the data, first I worked in balancing the dataset since it was a highly unbalanced dataset.Feature selection was performed over training and test dataset.I have used several classificatin algorithms but logistic regression suited best for my model builing.After working with the training dataset, I have perfomed same steps for my test set and the result was quite good.

8.How prediction was done?

The user need to provide the required details of his/her trasaction done through credit card then the model will take the input and provide prediction output which will be shown to the user through the Ul.

9.Which is more important to you model accuracy or model performance?

Well, you must know that model accuracy is only a subset of model performance. The accuracy of the model and performance of the model are directly proportional and hence better the performance of the model, more accurate are the predictions.

10.How can you optimize your solution?

1. Model optimization depends on various factors
2. Train with better data or do data pre-processing in efficient way.
3. Increase the quantity of training data etc.
4. Try and use multithreaded approaches.

11.Explain the Confusion Matrix with Respect to Machine Learning Algorithms.

A confusion matrix (or error matrix) is a specific table that is used to measure the performance of an algorithm. It is mostly used in supervised learning; in unsupervised learning, it’s called the matching matrix. The confusion matrix has two parameters: • Actual

* Predicted

It also has identical sets of features in both of these dimensions.

12.What is Overfitting, and How Can You Avoid It?

Overfitting is a situation that occurs when a model learns the training set too well, taking up random fluctuations in the training data as concepts. These impact the model’s ability to generalize and don’t apply to new data. When a model is given the training data, it shows 100 percent accuracy technically a slight loss. But, when we use the test data, there may be an error and low efficiency. This condition is known as overfitting.

There are multiple ways of avoiding overfitting, such as:

* Regularization: It involves a cost term for the features involved with the objective function
* Making a simple model with lesser variables and parameters, the variance can be reduced
* Cross-validation methods like k-folds can also be used
* If some model parameters are likely to cause overfitting, techniques for regularization like LASSO can be used that penalize these parameters

13.How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem?

While there is no fixed rule to choose an algorithm for a classification problem, you can follow these guidelines:

* If accuracy is a concern, test different algorithms and cross-validate them
* If the training dataset is small, use models that have low variance and high bias • If the training dataset is large, use models that have high variance and little bias

14.What is AUC Curve ?

AUC stands for "Area under the ROC Curve" .AUC measures the entire 2D area underneath the entire ROC curve.

15.

Which Tool You Are Used For Implementation This Model?

1)

Ide : VS Code

2)

Cloud : Heroku

3)

Data Base : MongoDB/Cassandra

16.

What Kind of challenges have u faced during the project?

Since I was working alone in this

project I faced many challenges. But slowly I came to

know that those were really silly small problems. The

most

challenging part for this

project was on the

backend development

of my web app.

17.

In which technology you are most comfortable?

I am pretty confident in machine learning.

18.How were you maintain the failure cases?

If our model is not predicting correctly for data then that dataset goes to database . There will be a report triggered to the support team at the end of the day with all failure scenarios where they can inspect the failure. Once we have a sufficient number of cases we can label and include those data while retraining the model for better performance.

19.What are your expectations?

I expect to work on different projects to enhance my technical skill and learn new things simultaneously.

20.What is your future objective?

My future objective is to learn new things in AI field because it changes continuously, and my aim is to pursue my career as a data scientist.