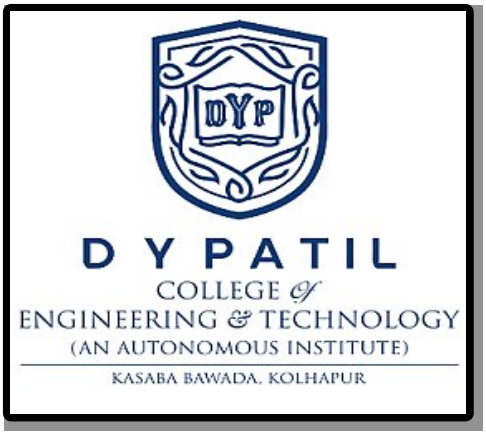
A PROJECT II SYNOPSIS

ON

"SPAM MAIL DETECTION MODEL"

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING



Submitted by :-

1) Viraj Arun Raut - 06
2) Atharv Shivaji Adasul - 22
3) Ayush A. Deolekar - 23

Class :- TY CSE **Division :- C**

Under the Guidance of :-

Prof. S. N. Patil

D.Y. PATIL COLLEGE OF ENGINEERING & TECHNOLOGY, KOLHAPUR

(Academic Year 2022-2023)

INDEX

INTRODUCTION

We've all been the recipient of spam emails before. Spam mail, or junk mail, is a type of email that is sent to a massive number of users at one time, frequently containing cryptic messages, scams, or most dangerously, phishing content.

While spam emails are sometimes sent manually by a human, most often, they are sent using a bot. Most popular email platforms, like Gmail and Microsoft Outlook, automatically filter spam emails by screening for recognizable phrases and patterns. A few common spam emails include fake advertisements, chain emails, and impersonation attempts. While these built-in spam detectors are usually pretty effective, sometimes, a particularly well-disguised spam email may fall through the cracks, landing in your inbox instead of your spam folder.

Clicking on a spam email can be dangerous, exposing your computer and personal information to different types of malware. Therefore, it's important to implement additional safety measures to protect your device, especially when it handles sensitive information like user data.

In this tutorial, we'll use Python to build an email spam detector. Then, we'll use machine learning to train our spam detector(https://blog.logrocket.com/best-javascript-machine-learning-libraries-in-2021/) to recognize and classify emails into spam and non-spam. Let's get started!

Problem Statement

1. Unwanted e-mails irritating internet connection.

2. Critical e-mails message are missed and / or delayed.

3. Millions of compromised computers.

4. Billions of dollars lost worldwide.

5. Identity theft.

6. Spam can crash mail servers and fill up hard drives.

What is Spam e-mail?

Spam email is unsolicited and unwanted junk email sent out in bulk to an indiscriminate recipient list. Typically, spam is sent for commercial purposes. It can be sent in massive volume by botnets, networks of infected computers.

Challenges for Spam Mail Detection:

1. Lowering Case.

2. Special Characters.

3. Removal of stopwards.

4. Removal of Hyperlinkes.

5. Removal of numbers.

6. Removal of whitespaces.

Overview of Spam Mail Detection Model

All these tasks are done through Natural Language Processing (NLP), which processes text into useful insights that can be applied to future data. In the field of artificial intelligence, NLP is one of the most complex areas of research due to the fact that text data is contextual. It needs modification to make it machine-interpretable and requires multiple stages of processing for feature extraction.

Classification problems can be broadly split into two categories: binary classification problems, and multi-class classification problems. Binary classification means there are only two possible label classes, e.g. a patient's condition is cancerous or it isn't, or a financial transaction is fraudulent or it is not. Multi-class classification refers to cases where there are more than two label classes. An example of this is classifying the sentiment of a movie review into positive, negative, or neutral.

There are many types of NLP problems, and one of the most common types is the classification of strings. Examples of this include the classification of movies/news articles into different genres and the automated classification of emails into a spam or not spam. I'll be looking into this last example in more detail for this article.

Solution Offered for Enabling Fraud Detection with Machine Learning

- Fraud Detection Model follows the Random Forest algorithm which takes hundreds of decision trees and aggregates and builds the model to avoid overfitting due to class imbalance.
- Analysis of at least 20-50 last transactions of the customer is done to identify an underlying pattern.
- Python and Machine Learning used to build the model.

Spam Mail Detection Process and Techniques

Real-Time Spam mail Detection requires time between normal mail and spam mail, History of user and email behavior. Spam mail detection involves the identification of fraut mail , finding hidden defects, detection of spam mail .

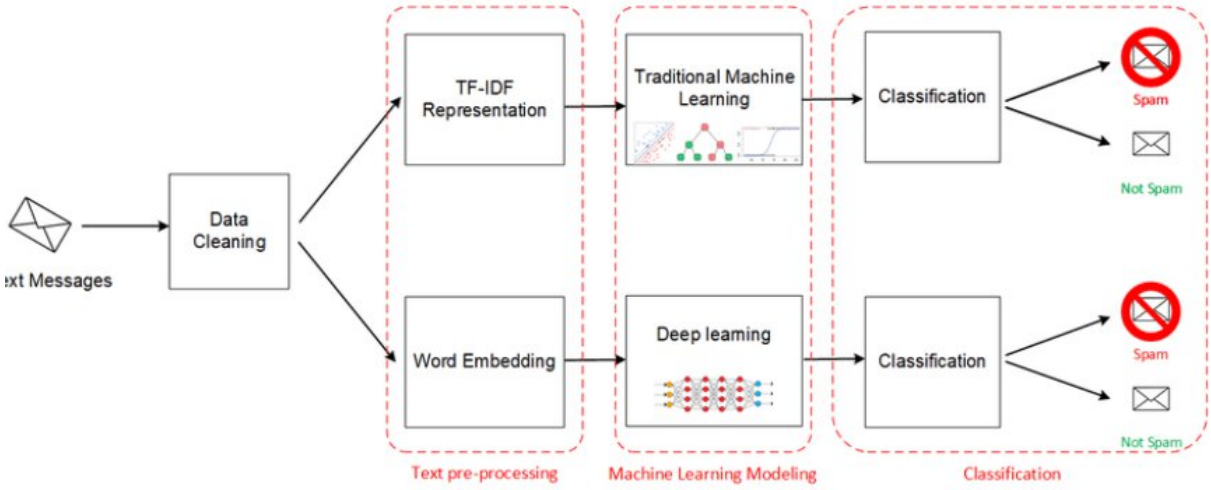
Steps for Credit Card Fraud Detection:

1. **Login module**
2. **Registration module**
3. **Administration module**
4. **User module**
5. **Mailing module**

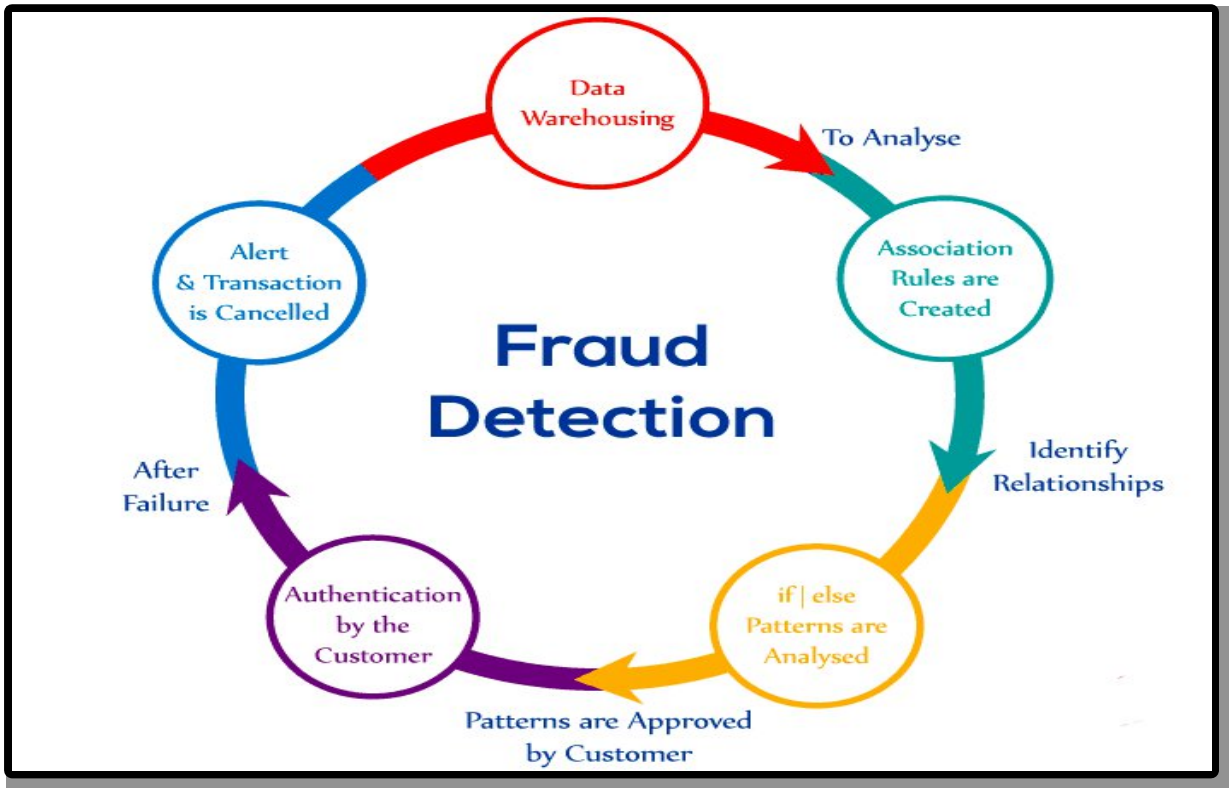
Benefits of Fraud Detection

* Availability of online alerts to detect any suspicious activity on the mails.
* Better Analytics and Predictive Forecasts.
* Stay Safe Online

SYSTEM ARCHITECTURE



rchitecture of the spam detection model.



MODULES OF SYSTEM

* Module 1 :- Fraud Cleaning

Fraud cleaning module contains various steps such as including hyperlinks , checking user address .

*** Module 2:- Data Extraction**

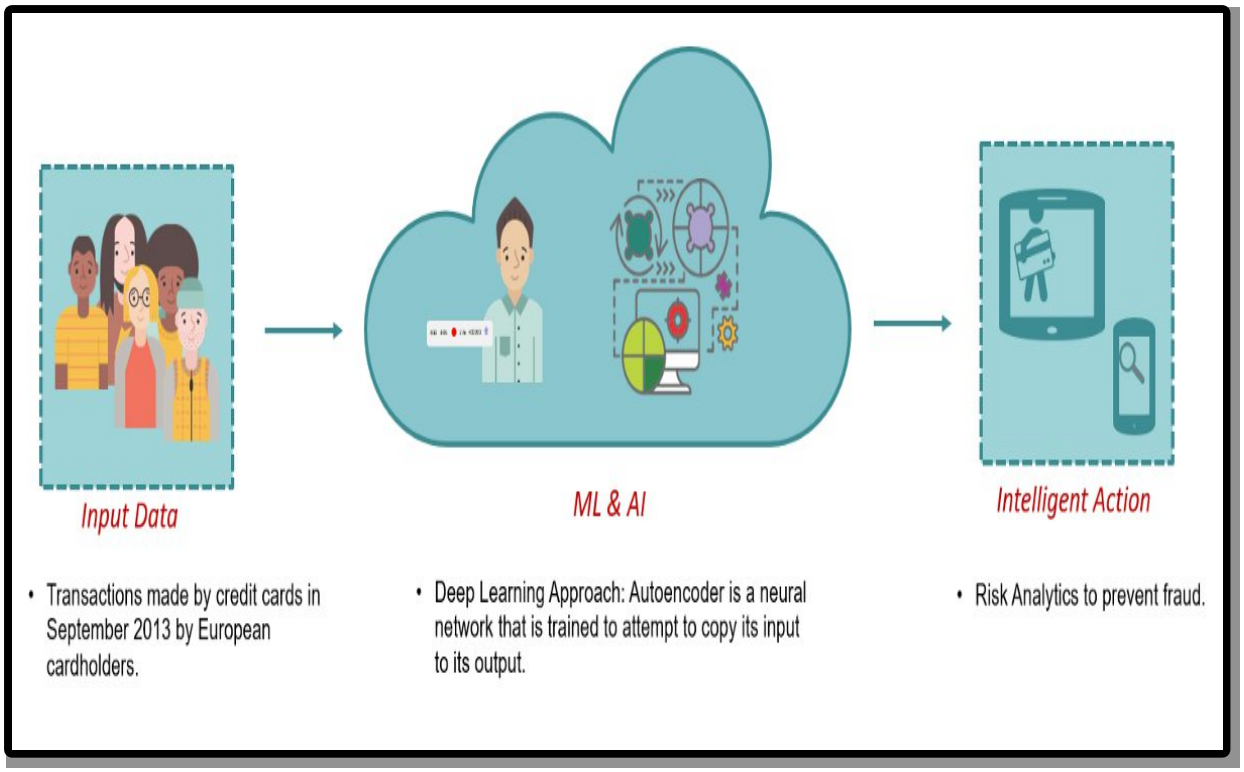
It includes data extraction using the training dataset , testing and cross validation through predictions.

*** Module 3:- Building Models and Feature Engineering**

It includes building modules using Logistic Regression , Decision Tree, Random Forest .

*** Module 4:- Streaming Data Ingestion**

It includes data preprocessing , data visualization , data modle deployment.



SYSTEM REQUIREMENTS

1) Software Requirements :-

*** Language :- Python**

*** Operating System :- Windows 7 and above .**

*** Platform :- Jupyter Notebook & Anconda IDE .**

*** Interpreter :- Python Compiler .**

*** Domain :- classification – Machine Learning model**

2) Hardware Requirements :-

- ❖ Processor :- Intel Core i3 and above .
- ❖ RAM :- 128 MB
- ❖ Hard Disk :- 4 GB
- ❖ Monitor

CONCLUSION

spam filtering is an incredibly powerful statistical technique—with acceptable computational complexity—for identifying spam messages. techniques address many weaknesses of other methodologies:

* The entire message can be examined, not just special parts.

* All words are significant, not just special keywords or addresses.

* Updating is, in practice, infrequent (never more than one or two email messages per week through the training program; often none).

* So far, spam attacks on Bayesian filters have been relatively unsuccessful.

* When combined with other techniques, Bayesian filters can be a very strong component of an institution's global spam system.

REFERENCES

[2] CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2 " A Comprehensive Survey of Data Mining-based Fraud Detection Research" published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia

[8] https://jespublication.com/upload/2020-110465.pdf

[9]https://www.semanticscholar.org/paper/Fraud-Detection-in-Credit-Cards-using-Logistic-Alenzi-Nojood/524b1a8ba37129f7ebdd0fff4528de75b991fec6

[10] https://github.com/prabhatk579/fraud-detection-using-logistic-regression

"SPAM MAIL DETECTION MODEL"

Class :- TY CSE Division :- C Batch – T1

Roll No.	Name	Sign
06	Viraj Arun Raut	
22	Atharv Shivaji Adasul	
23	Ayush A. Deolekar	

Date :-

Place :- Kolhapur

Prof. S.N.Patil	Prof . A. P. Budaragade	Prof. R. J. Dhannal
[Project Guide]	[Project Co-ordinator]	[H.O.D]