FraudVision project

Saima Binth Zubair

Abstract

FraudVision is an AI-powered solution designed to detect fraudulent credit card transactions in real-time, leveraging machine learning techniques. Utilizing 31 features, including anonymized PCA-transformed data, the solution aims to provide robust fraud detection through a user-friendly interface built using Streamlit. The core of the project is a Logistic Regression model that has demonstrated high accuracy in identifying fraudulent transactions. This project proposes a roadmap for product development, financial modeling, and integration into financial systems to safeguard transactions and reduce fraud-related losses.

Introduction

With the rapid growth of e-commerce and digital payments, fraudulent activities in the credit card sector have surged, causing billions in losses for businesses and consumers. FraudVision offers a machine learning-driven approach to addressing this problem, providing an interactive platform that detects fraudulent transactions in real-time. The web application, built using Streamlit, aims to empower users with an intuitive tool that not only flags suspicious transactions but also helps in real-time analysis. The project initially focuses on training a high-accuracy model using Logistic Regression, with plans to further enhance and commercialize the solution.

Problem statement

Credit card fraud is a major threat to the global financial system, resulting in significant economic losses for both institutions and consumers. Existing fraud detection systems struggle to keep pace with increasingly sophisticated fraud techniques and often fail to provide real-time detection. This results in delays in identifying and preventing fraudulent transactions. FraudVision aims to address these issues by offering a scalable, accurate, and real-time fraud detection solution that integrates seamlessly with financial systems, enhancing proactive fraud management and reducing financial losses.

Prototype Selection

a. **Feasibility**

FraudVision is a highly feasible project for development within the short term (2-3 years). Several factors contribute to its feasibility:

- Technology Availability: FraudVision leverages existing machine learning techniques, including Logistic Regression and PCA, which are mature and well-understood. The use of platforms like Streamlit makes the development of a user-friendly web application quick and cost-effective.
- Data Availability: Credit card transaction datasets are widely available, with many financial institutions already maintaining detailed transaction logs.
- Development Timeline: The development of a minimum viable product (MVP) can be accomplished within 12-18 months. A real-time fraud detection engine, based on the current project framework, can be launched within this timeframe, with potential for further enhancements through iterative development cycles.

b. Viability

FraudVision is positioned to remain relevant over the long term (20-30 years) due to several key factors:

- Growing Digital Payments: As digital transactions continue to grow globally, the need for real-time fraud detection will only increase. The product can adapt to new transaction methods (such as cryptocurrencies and mobile payments) and evolving fraud tactics.
- Scalability and Flexibility: The solution can be scaled and enhanced as more sophisticated machine learning models and technologies become available, ensuring the product's longevity.
- Regulatory Trends: Financial regulations are likely to increase over the coming decades, driving demand for automated fraud detection systems. FraudVision can evolve in compliance with these regulations, offering regulatory adherence as a service

c. Monetization

FraudVision can be directly monetized, making it suitable for this project:

- Subscription Model: Financial institutions, such as banks and credit card companies, can pay for FraudVision as a SaaS product. Subscriptions can be tiered based on transaction volume or features (e.g., basic fraud detection vs. advanced analytics).
- Transaction-Based Fees: A per-transaction pricing model can be implemented, where businesses pay for each transaction analyzed for fraud, especially for large enterprises dealing with high transaction volumes.
- Cost Savings: By preventing fraudulent transactions, FraudVision reduces the costs of chargebacks and fraud-related losses for businesses, which adds value that clients are willing to pay for.

Prototype Development

Data collection and preprocessing

In [1]:

```python
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

In [2]:

```python
credit_card_data = pd.read_csv('creditcard.csv')
credit_card_data.head()
```

Out[2]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.22577
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.24799
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.10830
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431

5 rows × 31 columns

```python
# checking the number of missing values in each column
credit_card_data.isnull().sum()
```

[5]:

```
Time      0
V1        0
V2        0
V3        0
V4        0
V5        0
V6        0
V7        0
V8        0
V9        0
V10       0
V11       0
V12       0
V13       0
V14       0
```

Evaluation

In [13]:

```
# statistical measures of the data
legit.Amount.describe()
```

Out[13]:

```
count    284315.000000
mean         88.291022
std         250.105092
min           0.000000
25%           5.650000
50%          22.000000
75%          77.050000
max       25691.160000
Name: Amount, dtype: float64
```

"count" indicates the total number of transactions in the dataset that have a valid "Amount" value (i.e., non-missing values). "mean" is the average transaction amount for all the transactions in the dataset. In this case, the mean is 88.291022, indicating that the average transaction amount is around $88USD."std"isthestandarddeviationofthetransactionamounts.Itisameasureofhowspreadoutthetransactiona$ 5.65 USD. "50%" is the median transaction amount, which is the value that separates the lower half of the transactions from the upper half. In this case, the median is 22.0, indicating that half of the transactions have a value less than or equal to $22USD."$7577.05 USD. "max" is the largest transaction amount in the dataset. In this case, the largest transaction amount is 25,691.16, indicating that there may be some transactions with very large values.

Training

Model Training

Logistic Regression

In [28]:

```
model=LogisticRegression()
```

In [29]:

```
# training the Logistic Regression Model with Training Data
model.fit(X_train, Y_train)
# accuracy on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
print('Accuracy on Training data : ', training_data_accuracy)
```

```
Accuracy on Training data :  0.9377382465057179
```

In [30]:

```
# accuracy on test data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
print('Accuracy score on Test Data : ', test_data_accuracy)
```

```
Accuracy score on Test Data :  0.8984771573604061
```

Business Model Development for FraudVision

1. **Value Proposition**
   - Real-Time Fraud Detection: Provide financial institutions with an advanced AI-powered tool that detects fraudulent transactions instantly, minimizing financial losses and protecting consumer trust.
   - High Accuracy: Leverage a Logistic Regression model with proven accuracy (93% training, 89% testing) to deliver reliable fraud detection.
   - User-Friendly Interface: Offer an intuitive, interactive web application built with Streamlit, enabling easy integration and management of fraud detection processes.
2. **Revenue Model**
   - Subscription-Based Pricing:
     - Basic Plan: For small to mid-sized businesses, offering essential fraud detection capabilities.
     - Premium Plan: For large enterprises, including advanced features such as detailed analytics and customized reporting.
     - Enterprise Plan: Tailored solutions with full integration support, dedicated account management, and advanced customization options.
   - **Transaction-Based Fees:**
     - Per Transaction Fee: Charge a fee for each transaction analyzed, suitable for businesses with high transaction volumes.
     - Tiered Pricing: Offer different pricing tiers based on transaction volume or risk level.
   - **Licensing:**
     - Partnerships with Payment Processors: License the technology to payment processors and fintech firms, creating a new revenue stream and expanding market reach.
   - **Consulting and Integration Services:**
     - Customization Services: Provide consulting to tailor the solution to specific business needs and integrate it into existing systems.

Financial Modeling

Financial Equation

Profit (y) = Revenue from Sales (mx) - Costs (c)

Where:

- Revenue from Sales (mx) = Price per Subscription × Number of Subscriptions Sold
- Costs (c) = Fixed Costs + Variable Costs

Assumptions for FraudVision:

- **Price per Subscription** = $100 per month
- **Number of Subscriptions Sold** = 100 (e.g., a mix of Basic, Premium, and Enterprise plans)
- **Fixed Costs** = $30,000 (includes development, infrastructure, and initial marketing expenses)
- **Variable Costs** = $10,000 (includes ongoing operational costs, customer support, and additional marketing)

Calculations:

1. **Revenue from Sales:**

$$\text{Revenue from Sales (mx)} = \text{Price per Subscription} \times \text{Number of Subscriptions Sold}$$

$$\text{Revenue from Sales} = 100 \times 100 = 10,000$$

2. **Costs:**

$$\text{Costs (c)} = \text{Fixed Costs} + \text{Variable Costs}$$

$$\text{Costs} = 30,000 + 10,000 = 40,000$$

3. **Profit:**

$$\text{Profit} = \text{Revenue from Sales} - \text{Costs}$$

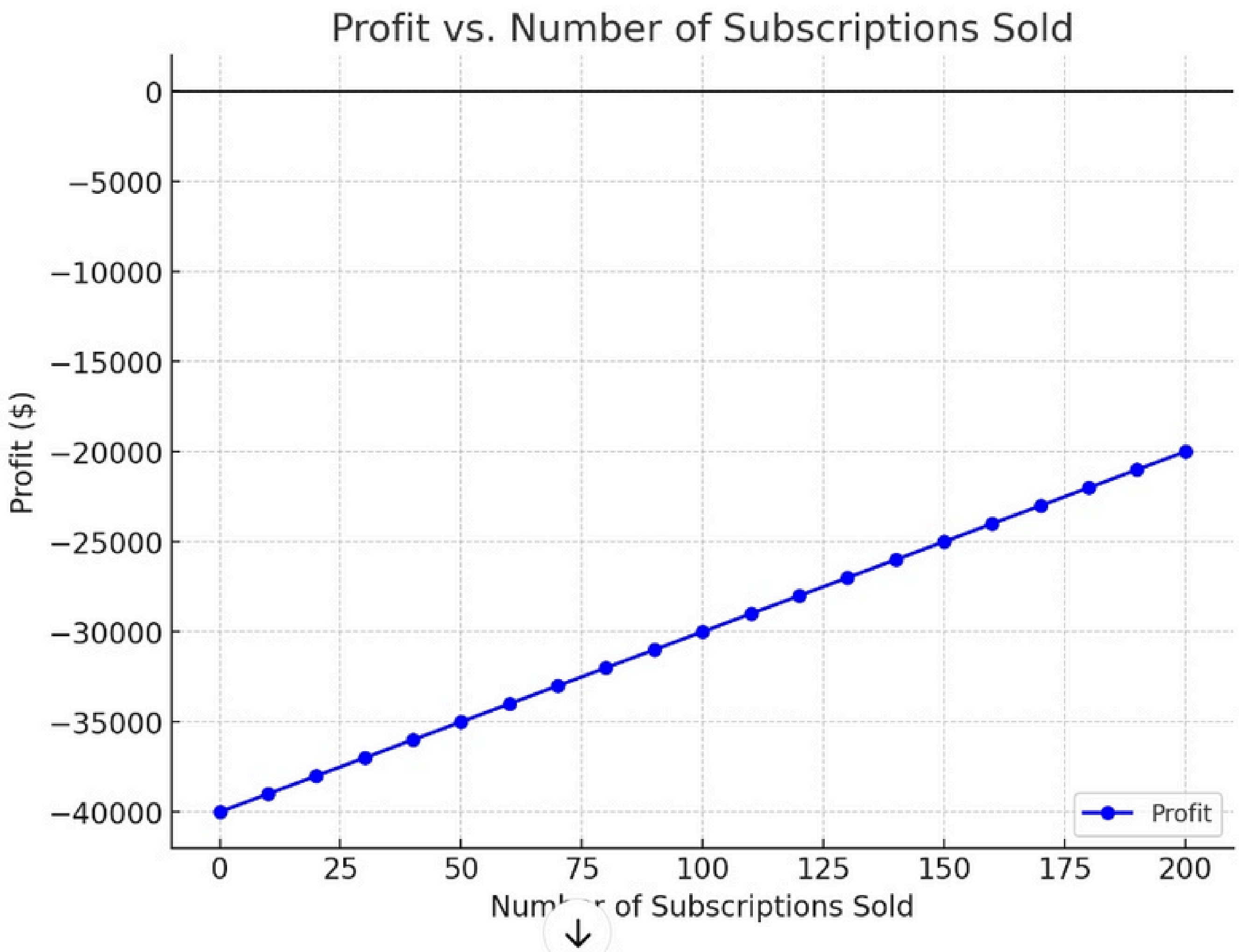
$$\text{Profit} = 10,000 - 40,000 = -30,000$$

Financial Equation for FraudVision:

$$\text{Profit} = (100 \times 100) - (30,000 + 10,000)$$

$$\text{Profit} = 10,000 - 40,000$$

$$\text{Profit} = -30,000$$



Here is the graph showing the relationship between the number of subscriptions sold and the profit for FraudVision. As you can see, the profit starts negative and increases as more subscriptions are sold.

Github link:

https://github.com/SaimaBZ/Feynn_labs-final_project