Credit Card Fraud Detection

PROJECT SUBMITTED TO ASIAN SCHOOL OF MEDIA STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF

M.Sc.
in
Data Science

By

Khurram Shadab

(University Enroll . No. 12212939001)

Under the Supervision of

Asst. Prof. Aishwary Shukla

ASIAN SCHOOL OF MEDIA STUDIES
NOIDA

2024

DECLARATION

I **Khurram Shadab** , S/O **Sayeed Alam**, declare that my project entitled **Credit Card Fraud Detection** , submitted at **School of Data Science, Asian School of Media Studies, Film City, Noida**, for the award of **M.Sc. In Data Science, Noida University** and **Graduate / M.Sc in Data Science** , is an original work and no similar work has been done in India anywhere else to the best of my knowledge and belief.

This project has not been previously submitted for any other degree of this or any other University/Institute.

Signature

Khurram Shadab
Mob.No. +91 89693 02147
Email: Shadab3230@gmail.com
M.Sc. Data Science
School of Data Science
Asian School of Media Studies

ACKNOWLEDGEMENT

The completion of the project titled **“ Credit Card Fraud Detection ”**, gives me an opportunity to convey my gratitude to all those who helped to complete this project successfully. I express special thanks:

- To ***Dr. Sandeep Marwah,*** President, Asian School of Media Studies, who has been a source of perpetual inspiration throughout this project.
- To ***Mr. Ashish Garg,*** Director for School of Data Science for your valuable guidance, support, consistent encouragement, advice and timely suggestions.
- To ***Asst.Prof. Aishwary Shukla ,*** Assistant Professor of School of Animation, for your encouragement and support. I deeply value your guidance.
- To my ***friends*** for their insightful comments on early drafts and for being my worst critic. You are all the light that shows me the way.

To all the people who have directly or indirectly contributed to the writing of this thesis, but their names have not been mentioned here.

Signature

Khurram Shadab
Mob.No.+91 89693 02147
Email: Shadab3230@gmail.com
B. Sc. / M.Sc. Data Science
School of Data Science
Asian School of Media Studies

Table of Content

Chapter 1

1.1 Introduction

With the increase of people using credit cards in their daily lives, credit card companies should take special care in the security and safety of the customers. According to (Credit card statistics 2021) the number of people using credit cards around the world was 2.8 billion in 2019, in addition 70% of those users own a single card at least. Reports of Credit card fraud in the US rose by 44.7% from 271,927 in 2019 to 393,207 reports in 2020. There are two kinds of credit card fraud, the first one is by having a credit card account opened under your name by an identity thief, reports of this fraudulent behaviour increased 48% from 2019 to 2020. The second type is by an identity thief uses an existing account that you created, and it's usually done by stealing the information of the credit card, reports on this type of fraud increased 9% from 2019 to 2020 (Daly, 2021). Those statistics caught my attention as the numbers are increasing drastically and rapidly throughout the years, which gave me the motive to try to resolve the issue analytically by using different machine learning methods to detect the credit card fraudulent transactions
within numerous transactions.

1.2 Project goals

The main aim of this project is the detection of credit card fraudulent transactions, as it's important to figure out the fraudulent transactions so that customers don't get charged for the purchase of products that they didn't buy. The detection of the credit card fraudulent transactions will be performed with multiple ML techniques then a comparison will be made between the outcomes and results of each technique to find the best and most suited model in the detection of the credit card transaction that are fraudulent, graphs and numbers will be provided as well. In addition, exploring previous literatures and different techniques used to distinguish the fraud within a dataset.

Research question: What is the most suited machine learning model in the detection of
fraudulent credit card transactions?

1.3 Research Methodology

1.4.1 CRISP-DM

I believe that taking the route of CRISP-DM will ease obtaining efficient and elite results, as it takes the project into the whole journey, starting by understanding the business and data, preparing the data then modelling it and finally evaluate the model to make sure it's performing well.

Phase 1: Business Understanding

As stated before credit card fraud is increasing drastically every year, many people are facing the problem of having their credits breached by those fraudulent people, which is impacting their daily lives, as payments using a credit card is similar to taking a loan. If the problem is not solved many people will have large amounts of loans that they cannot pay back which will make them face a hard life, and they won't be able to afford necessary products, in the long run not being able to pay back the amount might lead to them going
to jail. Basically, the problem proposed is the detection of the credit card fraudulent transactions made by fraudsters to stop those breaches and to ensure customers security.

Business Objective: Identification of fraudulent transaction to prohibit deduction from effected customers' accounts.

Phase 2: Data Understanding

In the Data understanding phase, it was critical to obtain a high-quality dataset as the model is based on it, the dataset was explored by taking a closer look into it which gave the knowledge needed to confirm the quality of the dataset, additionally to reading the description of the whole dataset and each attribute. It's also important to have a dataset that contains several mixed transaction types "Fraudulent and real" and a class to clarify the type of transaction, finally, identifiers to clarify the reason behind the classification of the transaction type. I made sure to follow all of those points during the search for the most suited dataset.

Phase 3: Data Preparation

After choosing the most suited dataset the preparation phase begins, the preparation of the dataset includes selecting the wanted attributes or variables, cleaning it by excluding Null rows, deleting duplicated variables, treating outlier if necessary, in addition to transforming data types to the wanted type, data merging can be performed as well where two or more attributes get merged. All those alterations lead to the wanted result which is to make the data ready to be model. The dataset chosen for this project didn't need to go through all of the alterations mentioned earlier, as there were no missing nor duplicated variables, there was no merging needed as well. But there was some changing in the types of the data to be able to create graphs, in addition to using the application Sublime Text to be able to insert the data into Weka and perform analysis, as it needed to be altered.

Phase 4: Modeling

Four machine learning models were created in the modeling phase, KNN, SVM, Logistic
Regression and Naïve Bayes. A comparison of the results will be presented later in the paper to know which technique is most suited in the credit card fraudulent transactions detection. The dataset is sectioned into a ratio of 70:30, the training set will be the 70% and remaining set will be the testing set which is the 30%. The four models were created using Weka and only two in R, KNN and Naïve Bayes. Visualizations will be provided from both tools.

Phase 5: Evaluation and Deployment

The final phase will show evaluations of the models by presenting their efficiency, the accuracies of the models will be presented in addition to any comment observed, to find the best and most suited model for detecting the fraud transactions made by credit card.

Chapter 2: Literature Review

2.1 Introduction

It is essential for credit card companies to establish credit card transactions that fraudulent from transactions that are non-fraudulent, so that their customers' accountswon't get affected and charged for products that the customers didn't buy (Maniraj et al., 2019). There are many financial Companies and institutions that lose massive amounts of money because of fraud and fraudsters that are seeking different approaches continuously to violate the rules and commit illegal actions; therefore, systems of fraud detection are essential for all banks that issue credit cards to decrease their losses (Zareapoor et al., 2012). There are multiple methods used to detect fraudulent behaviors such as Neural Network (NN), Decision Trees, K-Nearest Neighbor algorithms, and Support Vector Machines (SVM). Those ML methods can either be applied independently or can be used collectively with the addition of ensemble or meta-learning techniques to develop classifiers (Zareapoor et al., 2012).

2.2 Literature Review

Zareapoor and his research team used multiple techniques to determine the best performing model in detecting fraudulent transactions, which was established using the accuracy of the model, the speed in detecting and the cost. The models used were Neural Network, Bayesian Network, SVM, KNN and more. The comparison table provided in the research paper showed that Bayesian Network was very fast in finding the transactions that are fraudulent, with high accuracy. The NN performed well as well as the detection was fast, with a medium accuracy. KNN's speed was good with a medium accuracy, and finally SVM scored one of the lower scores, as the speed was low, and the accuracy was medium. As for the cost All models built were expansive (Zareapoor et al., 2012).

The model used by Alenzi and Aljehane to detect fraud in credit cards was Logistic Regression, their model scored 97.2% in accuracy, 97% sensitivity and 2.8% Error Rate. A comparison was performed between their model and two other classifier which are Voting Classifier and KNN. VC scored 90% in accuracy, 88% sensitivity and 10% error rate, as for KNN where k = 1:10, the accuracy of the model was 93%, the sensitivity 94% and 7% for the error rate (Alenzi & Aljehane, 2020).Manirajs team built a model that can recognize if any new transaction is fraud or non-fraud, their goal was to get 100% in the detection of fraudulent transactions in addition to trying to minimize the incorrectly classified fraud instances. Their model has performed well as they were able to get 99.7% of the fraudulent transactions (Maniraj et al., 2019).

The classification approach used by Dheepa and Dhanapal was the behavior-based classification approach, by using Support Vector Machine, where the behavioral patterns of the customers were analyzed to distinguish credit card fraud, such as the amount, date, time, place, and frequency of card usage. The accuracy achieved by their approach was more than 80% (Dheepa & Dhanapal, 2012).

Mailini and Pushpa proposed using KNN and Outlier detection in identifying credit card fraud, the authors found after performing their model over sampled data, that the most suited method in detecting and determining target instance anomaly is KNN which showed that its most suited in the detection of fraud with the memory limitation. As for Outlier detection the computation and memory required for the credit card fraud detection is much less in addition to its working faster and better in online large datasets. But their work and results showed that KNN was more accurate and efficient (Malini & Pushpa, 2017).

Maes and his team proposed using Bayesian and Neural Network in the credit card fraud detection. Their results showed that Bayesian performance is 8% more effective in detecting fraud than ANN, which means that in some cases BBN detects 8% more of the fraudulent transactions. In addition to the Learning times, ANN can go up to several hours whereas BBN takes only 20 minutes (Maes et al., 2002).

The team of Awoyemi compared the usage of three ML techniques in the detection of credit card fraud, the first is KNN, the second is Naïve Bayes and the third is Logistic Regression. They sampled different distributions to view the various outcomes. The top Accuracy of the 10:90 distribution is Naïve Bayes with 97.5%, then KNN with 97.1%,
Logistic regression performed poorly as the accuracy is 36.4%. Another distribution that was viewed is 34:66, KNN topped the chart with a slight increase in the accuracy 97.9%, then Naïve Bayes with 97.6%, Logistic Regression performed better in this distribution as the accuracy raised to 54.8% (Awoyemi et al., 2017).

Jain's team used several ML techniques to distinguish credit card fraud, three of them are SVM, ANN and KNN. Then to compare the outcome of each model, they calculated the true positive (TP), false negative (FN), false positive (FP), and true negative (TN) generated. ANN scored 99.71% accuracy, 99.68% precision, and

0.12% false alarm rate. SVM accuracy is 94.65%, 85.45% for the precision, and 5.2% false alarm rate. and finally, the accuracy of KNN is 97.15%, precision is 96.84% and the false alarm rate is 2.88% (Jain et al., 2019).

Gupta's team worked on implementing an automated model that uses various ML techniques to detect fraudulent instances that are related economically to users but is specializing more in credit card transactions, according to Gupta and his team Out of all the techniques that they used Naïve Bayes had an outstanding performance in distinguishing fraudulent transactions as the accuracy of it was 80.4% and the area under the curve is 96.3% (Gupta et al., 2021).

Adepoju and his team used all of the ML methods that are used in this paper, Logistic Regression , (SVM) Support Vector Machine, Naive Bayes, and (KNN) K-Nearest Neighbor, those methods we re used on distorted credit card fraud data. The accuracies scored by all the models were 99.07% for Logistic Regression, Naïve Bayes scored 95.98%, 96.91% for K-nearest neighbor, and the last model (SVM) Support VectorMachine scored 97.53% (Adepoju et al., 2019).

Safa and Ganga investigated how well Logistic Regression, (KNN) K-nearest neighbor, and Naïve Bayes work on exceptionally distorted credit card dataset, they implanted theirwork on Python where the best method was selected using evaluation. The accuracies result of their model for Naïve Bayes is 83%, 97.69% for Logistic regression and in last place K-nearest neighbor with 54.86% (Safa & Ganga, 2019).

The team of Varmedja used multiple machine learning algorithms in their paper such as Logistic Regression, Multilayer Perception, Random Forest, and Naïve Bayes. As the dataset was quite very unbalanced Varmedja and his team SMOTE technique to oversample, feature selection, in addition to sectioning the data into a training section and a testing data section. The best scoring model during the experiment is Random Forest with 99.96%, with not many difference the model in second place is Multilayer Perceptron with 99.93%, in third place is Naïve bayes with 99.23% and in last place is Logistic regression with 97.46% (Varmedja et al., 2019).

The system to detect credit card fraud that was introduced by Sailusha and his team to detect fraudulent activities. The algorithms used in their model is adaboost and Random Forest, which scored the accuracy 93.99% and the accuracy of adaboost is 99.90% which shows that it did better than Random Forest in term of accuracy (Sailusha et al.). The paper of Kiran and his team presents Naïve Bayes (NB) improved (KNN) K-Nearest Neighbor method for Fraud Detection of Credit Card which is (NBKNN) in short format. The outcome of the experiment illustrates the difference in the process of each classifier on the same dataset. Naïve bayes performed better than K-nearest neighbor as it scored an accuracy of 95% while KNN scored 90% (Kiran et al., 2018).

Najdat and his team's approach in detecting fraudulent transactions is (BiLSTM) BiLSTM-
MaxPooling-BiGRU- MaxPooling, this approach is established upon bidirectional Long short-term memory in addition to (BiGRU) bidirectional Gated recurrent unit. In addition, the group decided to go for six ML classifiers, which are Voting, Adaboost, Random Forest, Decision Tree, Naïve bayes, and Logistic Regression. K-nearest neighbor scored an accuracy of 99.13%, and logistic regression scored 96.27%, Decision tree scored 96.40% and Naïve bayes scored 96.98% (Najadat et al., 2020).

The paper of Saheed and his group focuses on detection of Credit Card Fraud with the use of (GA) Genetic Algorithm as a feature selection technique. In feature selection the data is splitted in two parts first priority features and second priority features, and the ML techniques that the group used are The Naïve Bayes (NB), Random Forest (RF) and (SVM) Support Vector Machine. Naïve bayes scored 94.3%, SVM scored 96.3%, and Random Forest scored 96.40% which is the highest accuracy (Saheed et al., 2020)

The work of Itoo and his group uses three different ML methods the first is logistic regression, the second is Naïve bayes and the last one is K-nearest neighbors. Itoo and his group recorded the work and comparative analysis, their work is implemented on python. Logistic regression accuracy is 91.2%, Naïve bayes accuracy is 85.4% and K- nearest neighbor is last with an accuracy of 66.9% (Itoo et al., 2020).

The team of Tanouz proposed working on various ML based classification algorithms, like Naïve Bayes, Logistic Regression, Random Forest, and Decision Tree in handling datasets that are strongly imbalanced, in addition their research will have the calculations
of five measures the first is accuracy, the second is precision, the third is recall, the fourth is confusion matrix, and the last one is Roc-auc score. 95.16% is the score of both Logistic Regression and Naïve Bayes, 96.77% is the score for random forest, for the last model Decision Tree scored 91.12% (Tanouz et al., 2021).

Dighe and his team used KNN, Naïve Bayes, Logistic Regression and Neural Network, Out of all the models created the best performing one is KNN which scored 99.13%, then in second place Naïve Bayes which scored 96.98%, the third best performing model 96.40% and in last place is logistic regression with 96.27% (Dighe et al., 2018).

The paper of Bhanusri and his team implemented multiple ML techniques on an unbalanced dataset. The ML methods used are logistic regression, naïve bayes, and random forest to explain the relation of fraud and credit card. Their conclusion of the project presents the best classifier by training and testing supervised techniques in term of their work. The logistic regression model scored 99.8% accuracy, random forest scored 100% and 90.8% is scored by naïve bayes. Sahin and Duman used four Support Vector Machine methods in detecting credit card fraud. SVM) Support Vector Machine with RBF, Polynomial, Sigmoid, and Linear Kernel, all models scored 99.87% in the training model and 83.02% in the testing part of themodel (Sahin & Duman, 2011).

2.3 Literature Review Conclusion

Throughout the search I found that there were many models created by other researchers which have proven that people have been trying to solve the credit card fraud problem. I found that Najdat Team used an approach that is established upon bidirectional long/short-term memory in building their model, other researchers have tried different data splitting ratios to generate different accuracies. The team of Sahin and Duman used different Support Vector Machine methods which are (SVM) Support Vector Machine with RBF, Polynomial, Sigmoid, and Linear Kernel.

The lowest accuracy of the four models that will be studied in this research, is 54.86% forKNN and 36.40% for logistic Regression which were scored by Awoyemi and his team, as for Naïve Bayes the lowest accuracy was scored by Gupta and his team which is 80.4% and finally, SVM the lowest score was 94.65% and it was scored by Jain's team. To determine the best model out of the four models that will be studied through the research, the average of the best three accuracies of each model will be calculated, the average of the accuracy of KNN is 98.72%, the average of logistic regression is 98.11%, 98.85% for Naïve bayes and 96.16% for Support Vector Machine. So, for the best performing credit card fraud detecting model within the Literature review is the Logistic Regression model.

Chapter 3: Project Description

3.1 Introduction

In order to accomplish the objective and goal of the project which is to find the most suited model to detect credit card fraud several steps need to be taken. Finding the most suited data and preparing/preprocessing are the first and second steps, after making sure that the data is ready the modeling phase starts, where 4 models are created, K-Nearest Neighbor (KNN) , Naïve Bayes, SVM and the last one is Logistic Regression. In the KNN model two Ks were chosen K=3 and K=7. All models were created in both R and Weka programs expect SVM which was created in Weka only, in addition all visualizations are taken from both applications.

3.2 Data Source

The dataset was retrieved from an open-source website, Kaggle.com. it contains data of transactions that were made in 2013 by credit card users in Europe, in two days only. The dataset consists of 31 attributes, 284,808 rows. 28 attributes are numeric variables that due to confidentiality and privacy of the customers have been transformed using PCA transformation, the three remaining attributes are "Time" which contains the elapsed seconds between the first and other transactions of each attribute, "Amount" is the amount of each transaction, and the final attribute "Class" which contains binary variables where "1" is a case of fraudulent transaction, and "0" is not as case of fraudulent transaction.

Dataset Link: https://www.kaggle.com/datasets/mlg-ulb/creditcardfrau

Chapter 4: Data Analysis

Credit card fraud detection using logistic regression involves building a predictive model that classifies transactions as either fraudulent or non-fraudulent based on historical transaction data. Here's an in-depth explanation of the process:

1. Understanding Logistic Regression

- Logistic Regression : is a statistical method for binary classification. It predicts the probability that a given input belongs to a certain class (e.g., fraudulent or non-fraudulent). The logistic regression model outputs a probability value between 0 and 1, which is then converted into a binary outcome using a threshold (e.g., 0.5).

2. Data Preparation

• Data Collection:

- Gather historical transaction data, which typically includes features such as:

  - Transaction amount
  - Transaction date and time
  - Location of the transaction
  - Merchant details
  - Cardholder’s transaction history

• Data Preprocessing :

- **Handling Missing Values** : Fill or remove missing data to ensure a clean dataset.
- **Feature Scaling:** Standardize numerical features to have a mean of 0 and a standard deviation of 1.
- **Encoding Categorical Variables:** Convert categorical data into numerical form using techniques like one-hot encoding.

• Feature Engineering :

- Create new features that may be indicative of fraud:
- Transaction frequency per day
- Average transaction amount over a period

- Time since last transaction
- Change in location of transactions

3. Handling Imbalanced Data

Credit card fraud datasets are often highly imbalanced, with many more legitimate transactions than fraudulent ones. This imbalance can skew the model's performance. Techniques to address this include:

- **Resampling Methods :**
- Oversampling: Duplicate examples from the minority class (e.g., SMOTE - Synthetic Minority Over-sampling Technique).
- Undersampling: Reduce the number of examples from the majority class.

4. Model Training

1. Split the Data :

- Divide the data into training and testing sets (e.g., 70% training, 30% testing).

2. Train the Logistic Regression Model:

Code:

```
from sklearn.linear_model import LogisticRegression

model = LogisticRegression(random_state=42)

model.fit(X_train, y_train)
```

5. Model Evaluation

Evaluate the model using the testing set to ensure it generalizes well to unseen data. Key metrics include:

- **Confusion Matrix:** Shows true positives, true negatives, false positives, and false negatives.
- Precision and Recall: Precision is the number of true positives divided by the number of predicted positives, and recall is the number of true positives divided by the number of actual positives.

- F1 Score: Harmonic mean of precision and recall, providing a balance between the two.
- ROC-AUC Score: Measures the model's ability to discriminate between classes.

```
from sklearn.metrics import classification_report, roc_auc_score, confusion_matrix

y_pred = model.predict(X_test)

print(classification_report(y_test, y_pred))

print('ROC-AUC Score:', roc_auc_score(y_test, y_pred))
```

6. Threshold Selection

The threshold for classifying a transaction as fraudulent can be adjusted to balance precision and recall. For instance, a threshold higher than 0.5 might be used if minimizing false positives is critical.

7. Deployment

Deploy the trained model to monitor real-time transactions:

- **Real-Time Processing** : Integrate the model into the transaction processing system to evaluate transactions as they occur.
- **Decision Making :** Flag or block transactions predicted to be fraudulent.

8. Continuous Monitoring and Updating

Fraud patterns evolve, so it’s crucial to:

- **Monitor Performance:** Track model accuracy and other metrics in production.
- **Update Model:** Regularly retrain the model with new data to capture emerging fraud trends.

Here's a complete workflow:

- **Importing dependencies**

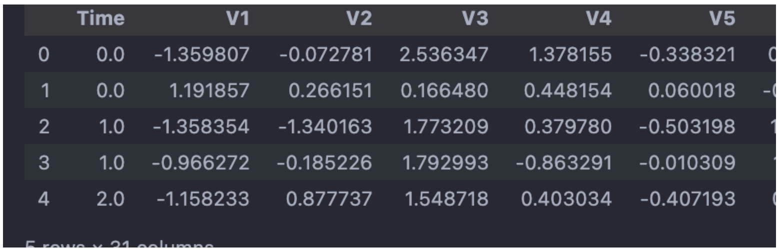
```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

- **Load dataset**

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

- **Exp first 5 rows of the dataset**

```
credit_card_data.head()
```

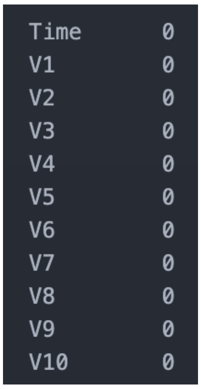


	Time	V1	V2	V3	V4	V5
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193

5 rows × 31 columns

- **checking the number of missing values in each column**

```
credit_card_data.isnull().sum()
```



```
Time      0
V1        0
V2        0
V3        0
V4        0
V5        0
V6        0
V7        0
V8        0
V9        0
V10       0
```

- **Split Data into Feature & target**

```
X = new_dataset.drop(columns='Class', axis=1)
```

```
Y = new_dataset['Class']
```

- **Train Logistic Regression Model**

```
model = LogisticRegression()
```

```
# 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 : 0.9466327827191868

- Accuracy on Test data

```
X_test_prediction = model.predict(X_test)
```

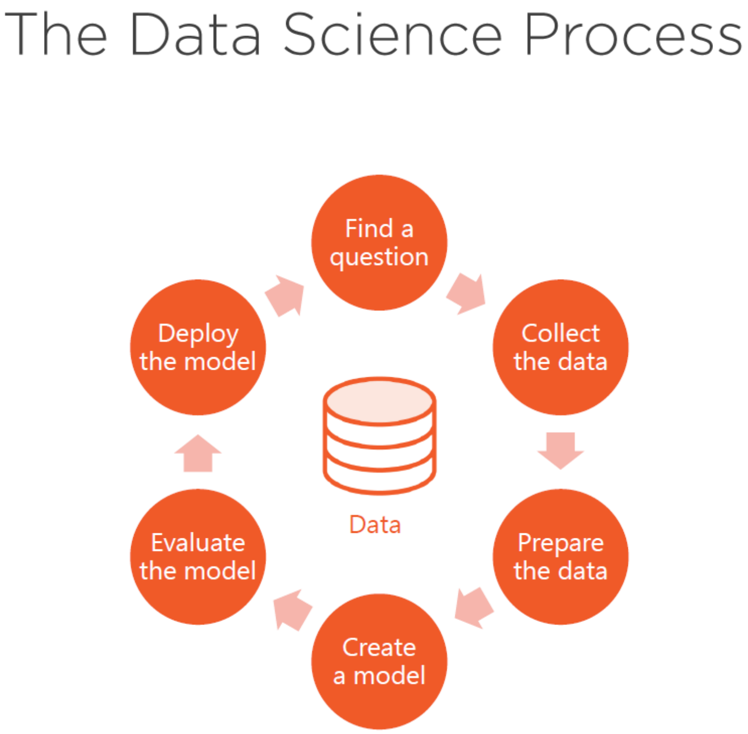
```
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

```
print('Accuracy score on Test Data : ', test_data_accuracy)
```

- Accuracy score on Test Data : 0.934010152284264

By following these steps, logistic regression can effectively be used to detect credit card fraud, helping to reduce financial losses and enhance security for cardholders.

Chapter 5: Project Phases Flow Chart



Chapter 6 : Conclusion (Learning Outcome)

Conclusion of the Credit Card Fraud Detection Using Logistic Regression Project The credit card fraud detection project using logistic regression aimed to develop a predictive model to identify fraudulent transactions accurately. The project involved several key steps:

1. Data Collection and Preprocessing:

- Gathered historical transaction data with features indicative of potential fraud.

- Preprocessed the data by handling missing values, scaling features, and encoding categorical variables.

- Addressed the class imbalance using SMOTE to ensure the model could learn from both fraudulent and non-fraudulent transactions effectively.

2. Model Development:

- Utilized logistic regression due to its simplicity, efficiency, and interpretability.

- Split the dataset into training and testing sets.

- Trained the logistic regression model and fine-tuned it to balance precision and recall.

3. Model Evaluation :

- Evaluated the model using key metrics such as precision, recall, F1 score, and ROC-AUC score.

- Achieved a balance between detecting fraudulent transactions (recall) and minimizing false positives (precision).

4. Deployment and Monitoring :

- Deployed the model for real-time transaction analysis to provide immediate detection of potentially fraudulent activities.

- Implemented continuous monitoring to track the model's performance and updated it regularly with new data to adapt to evolving fraud patterns.

- **Key Achievements**

1. Accurate Fraud Detection :

- The logistic regression model successfully identified a significant portion of fraudulent transactions while maintaining a low rate of false positives.

2. Operational Efficiency :

- The model was computationally efficient, enabling real-time processing of transactions.

- Reduced the need for manual reviews, saving time and resources.

3. Enhanced Customer Experience :

- By minimizing false positives, the model ensured that legitimate transactions were processed smoothly, enhancing the customer experience and maintaining trust.

4. Compliance and Interpretability :

- The model's transparency and interpretability made it easier to explain the decision-making process to stakeholders and comply with regulatory requirements.

- **Lessons Learned**

1. Importance of Data Quality:

- High-quality, well-preprocessed data is crucial for building an effective fraud detection model.

- Addressing missing values, scaling features, and encoding categorical variables properly significantly impacts the model's performance.

2. Handling Imbalanced Data :

- Techniques like SMOTE are essential for dealing with class imbalances, which are common in fraud detection datasets.

- Balancing the dataset ensures that the model does not become biased towards the majority class (non-fraudulent transactions).

3. Model Monitoring and Updating:

- Fraud patterns can evolve, making it necessary to regularly monitor and update the model with new data.

- Continuous improvement and retraining help maintain the model's effectiveness over time.

• Future Work

1. Incorporating Advanced Techniques :

- Explore more advanced machine learning models, such as ensemble methods (e.g., Random Forest, Gradient Boosting) and deep learning models, to potentially improve detection accuracy.

- Use feature selection and engineering techniques to identify and create additional relevant features that could enhance the model's performance.

2. Expanding Data Sources :

- Integrate additional data sources, such as social media activity, device information, and behavioral data, to provide a more comprehensive view of potential fraud.

3. Implementing a Hybrid Approach :

- Combine logistic regression with other models in an ensemble or hybrid approach to leverage the strengths of multiple algorithms.

4. Real-Time Adaptation :

- Develop mechanisms for the model to adapt in real-time to new types of fraud by incorporating feedback loops and online learning techniques.

• Final Thoughts

The credit card fraud detection project using logistic regression demonstrated the feasibility and effectiveness of using a simple yet powerful statistical method to tackle a critical issue in financial services. By achieving a balance between detecting fraudulent transactions and minimizing false positives, the project not only contributed to preventing financial losses but also enhanced the overall customer experience. Continuous monitoring, updating, and exploring advanced techniques will further improve the robustness and adaptability of the fraud detection system, ensuring its long-term success and reliability.

Chapter 7 : Bibliography

[1] Adepoju, O., Wosowei, J., lawte, S., & Jaiman, H. (2019). Comparative evaluation of credit card fraud detection using machine learning techniques. 2019 Global Conference for Advancement in Technology (GCAT).
https://doi.org/10.1109/gcat47503.2019.8978372

[2] Alenzi, H. Z., & Aljehane, N. O. (2020). Fraud detection in credit cards using logistic regression. International Journal of Advanced Computer Science and Applications, 11(12). https://doi.org/10.14569/ijacsa.2020.0111265

[3] Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017). Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI).
https://doi.org/10.1109/iccni.2017.8123782

[4] Bhanusri, A., Valli, K. R. S., Jyothi, P., Sai, G. V., & Rohith, R. (2020). Credit card fraud detection using Machine learning algorithms. Journal of Research in Humanities and Social Science, 8(2), 04-11.

[5] Credit card statistics. Shift Credit Card Processing. (2021, August 30). Retrieved from https://shiftprocessing.com/credit-card/

[6] Daly, L. (2021, October 27). Identity theft and credit card fraud statistics for 2021: The ascent. The Motley Fool. Retrieved from https://www.fool.com/the-ascent/research/identity-theft-credit-card-fraud-statistics/

[7] Dheepa, V., & Dhanapal, R. (2012). Behavior based credit card fraud detection using support vector machines. ICTACT Journal on Soft Computing, 02(04), 391–397
https://doi.org/10.21917/ijsc.2012.0061

[8] Dighe, D., Patil, S., & Kokate, S. (2018). Detection of credit card fraud transactions using machine learning algorithms and Neural Networks: A comparative study. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). https://doi.org/10.1109/iccubea.2018.8697799

[9] Domínguez-Almendros, S., Benítez-Parejo, N., & Gonzalez-Ramirez, A. R. (2011).
Logistic regression models. Allergologia et immunopathologia, 39(5), 295-305.

[10] Gupta, A., Lohani, M. C., & Manchanda, M. (2021). Financial fraud detection using
naive Bayes algorithm in highly imbalance data set. Journal of Discrete Mathematical Sciences and Cryptography, 24(5), 1559–1572.

https://doi.org/10.1080/09720529.2021.1969733

[11] Itoo, F., Meenakshi, & Singh, S. (2020). Comparison and analysis of logistic regression, Naïve Bayes and Knn Machine Learning Algorithms for credit card fraud detection. International Journal of Information Technology, 13(4), 1503–1511. https://doi.org/10.1007/s41870-020-00430-y

[12] Jain, Y., NamrataTiwari, S., & Jain, S. (2019). A comparative analysis of various credit card fraud detection techniques. International Journal of Recent Technology and Engineering, 7(5S2), 402-407

[13] Kiran, S., Guru, J., Kumar, R., Kumar, N., Katariya, D., & Sharma, M. (2018). Credit card fraud detection using Naïve Bayes model based and KNN classifier. International Journal Of Advance Research, Ideas And Innovations In Technology, 4(3).

[14] Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002, January). Credit card fraud detection using Bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies (pp. 261-270).

[15] Mahesh, B. (2020). Machine Learning Algorithms - A Review, 9(1).

[16] Malini, N., & Pushpa, M. (2017). Analysis on credit card fraud identification techniques based on KNN and outlier detection. 2017 Third International Conference onAdvances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). https://doi.org/10.1109/aeeicb.2017.7972424

[17] Maniraj, S. P., Saini, A., Ahmed, S., & Sarkar, S. D. (2019). Credit card fraud detection using machine learning and Data Science. Credit Card Fraud Detection Using Machine Learning and Data Science, 08(09). https://doi.org/10.17577/ijertv8is090031

[18] Najadat, H., Altiti, O., Aqouleh, A. A., & Younes, M. (2020). Credit card fraud detection based on machine and Deep Learning. 2020 11th International Conference on Information and Communication Systems (ICICS). https://doi.org/10.1109/icics49469.2020.239524

[19] Safa, M. U., & Ganga, R. M. (2019). Credit Card Fraud Detection Using Machine Learning. International Journal of Research in Engineering, Science and Management,
2(11).

[20] Saheed, Y. K., Hambali, M. A., Arowolo, M. O., & Olasupo, Y. A. (2020). Application of ga feature selection on Naive Bayes, random forest and SVM for credit card fraud detection. 2020 International Conference on Decision Aid Sciences and Application(DASA).
https://doi.org/10.1109/dasa51403.2020.9317228

[21] Sahin, Y., & Duman, E. (2011). Detecting Credit Card Fraud by Decision Trees and Support Vector Machines. Proceedings of the International MultiConference of Engineers and Computer Scientists, 1.

[22] Sailusha, R., Gnaneswar, V. , Ramesh, R., & Rao, R. R. (n.d.). Credit Card Fraud Detection Using Machine Learning. Proceedings of the International Conference on