BASAVARAJESWARI GROUP OF INSTITUTIONS

BALLARI INSTITUTE OF TECHNOLOGY & MANAGEMENT

NACCAccreditedInstitution*

(RecognizedbyGovt.ofKarnataka,approvedbyAICTE,NewDelhi&AffiliatedtoVisvesvaraya Technological University, Belagavi)

"JnanaGangotri"Campus,No.873/2,Ballari-HospetRoad,Allipur, Ballar1-583 104 (Karnataka) (India)

Ph:08392–237100/237190,Fax:08392–237197

DEPARTMENT OF CSE (DATA SCIENCE)

A OpenEnded-ProjectReport

On

"POSSUM"

A report submitted in partial fulfilment of the requirements for

the OPEN ENDED PROJECT OF DATA SCIENCE

LAB(22CD42)

Submitted By

M.D MOIN KHAN **USN:3BR22CD034**

M.D MOINUDDIN KHAN **USN:3BR22CD035**

Under the Guidance of

Dr.JagadishRM Prof.

Mrs.AnushyaVP Asst. Prof.

VisvesvarayaTechnological University

Belagavi,Karnataka2023-2024

BASAVARAJESWARIGROUPOFINSTITUTIONS

BALLARIINSTITUTEOFTECHNOLOGY&MANAGEMENT

NACCAccreditedInstitution*

(RecognizedbyGovt.ofKarnataka,approvedbyAICTE,New Delhi&Affiliatedto Visvesvaraya Technological University, Belagavi)

"JnanaGangotri"Campus,No.873/2,Ballari-HospetRoad,Allipur,

Ballari-583 104 (Karnataka) (India)

Ph:08392–237100/ 237190,Fax:08392– 237197

DEPARTMENTOFCSE(DATASCIENCE)

CERTIFICATE

This is to certify that the OPEN ENDED PROJECT of DATA SCIENCELABentitle**“**POSSUM”has been successfully presented by **MOIN KHAN** bearingUSN **3BR22CD0534 , MOINUDDIN** bearing USN **3BR22CD035** students of IV semester B.E for the partial fulfillment of the requirements for the award of **BachelorDegreein CSE(DS)** of the BALLARI INSTITUTE OF TECHNOLOGY& MANAGEMENT,BALLARI during the academic year 2023-2024.

Signature of guide	Signature of HOD
Dr. Jagadish R M	**Dr.DAradhana**
Mrs.Anushya V P	

ACKNOWLEDGEMENT

The satisfactions that accompanythe successful completion of our mini project on **"REAL TIME FRAUD DETECTION IN TRANSACTIONS "** would be incomplete without the mention of people who made it possible, whose noble gesture, affection, guidance, encouragement andsupport crowned my efforts with success. It is our privilege to express our gratitude and respect to all those who inspired us in the completion of our mini-project.

We are extremely grateful to our Guides **Dr. Jagadish R M and Mrs. Anushya V P** for their noble gesture, support co-ordination and valuable suggestions given in completingthe mini-project. We also thank **Dr. D Aradhana,** H.O.D. Department of CSE(DS), for his co-ordination and valuable suggestions given in completing the mini-project. We also thank Principal, Management and non-teaching staff for their co-ordination and valuable suggestions given to us in completing the Openended project.

Name	USN
MOIN KHAN	3BR22CD034
MOINUDDIN	3BR22CD035

TABLEOFCONTENTS

CHAPTER1

INTRODUCTION

In the digital age, the rapid proliferation of online financial transactions has brought unprecedented convenience to consumers and businesses alike. However, this convenience is accompanied by the rising threat of credit card fraud, which poses significant challenges to financial institutions and consumers. Fraudulentactivitiesnotonlyleadtosubstantialfinanciallossesbutalsoerodeconsumertrustandconfidence in the security of digital payment systems.

cards fraud detection is a critical area of focus for the financial industry, necessitating the developmentof sophisticated techniques to identify and prevent fraudulent transactions. Traditional rule-based systems, althoughuseful,areofteninadequateinkeepingpacewiththeevolvingtacticsoffraudsters.Therefore,there is a pressing need for advanced methods that can adapt to new patterns of fraud in real-time.

This project aims to address the problem of REAL Time card fraud detection by leveraging the power of machine learning. Machine learning algorithms have shown great promise in various domains for their ability to learnfrom data and make predictions. By applying these techniques to credit card transaction data, we can developmodels that accurately distinguish between legitimate and fraudulent transactions.

Theprimaryobjectivesofthisprojectare:

1. **Todevelop a comprehensive dataset** thatincludes varioustransaction featuresandaddressesissues like class imbalance.
2. **Toexploreandanalyzethedata** touncoverinsightsandpatternsrelatedtofraudulent activities.
3. **Toengineerrelevantfeatures** thatenhancethemodel'sabilitytodetect fraud.
4. **Toevaluateandcomparemultiplemachinelearningalgorithms**toidentifythemosteffective model for fraud detection.
5. **Toimplementanddeploythemodel**inareal-timeenvironmentforcontinuousmonitoringand prevention of fraudulent transactions.

Byachievingtheseobjectives,this projectaimstoprovidearobust,scalable,andefficientsolutionforcredit card fraud detection. The implementation of such a system can significantly reduce the incidence of fraud, thereby protecting consumers and enhancing the overall security of digital financial transactions.

ABOUTTHEPROJECT

The "Real Time Fraud Detection in transactions in Data Science" project focuses on developing a system to accurately identify fraudulent card transactions. Given the critical nature of fraud detection in the financial sector,the project uses historicaltransaction dataand tackleschallenges like dataset imbalance byemploying preprocessing techniques and machine learning algorithms such as Logistic Regression, Decision Trees, and Neural Networks. The goal is to create a model that minimizes false positives while effectively detecting fraud. This project demonstrates the practical application of data science in addressing real-world financial challenges and suggests future enhancements like real-time detection.

PROBLEMSTATEMENT

How can we develop a robust and effective card fraud detection system using machine learning techniquesto accurately identify fraudulent transactions in real-time, despite challenges such as class imbalance, evolving fraud patterns, and the complexity of transaction data?

OBJECTIVES

DataPreprocessingandCleaning:

- Topreprocessandcleanthecardtransactiondata,ensuringthatitisfreeofinconsistenciesand missing values.
- Tobalancethedatasettoaddressclassimbalance,usingtechniquessuchasoversampling, undersampling, and synthetic data generation.

ExploratoryDataAnalysis (EDA):

- Toperformexploratorydataanalysistogaininsightsintothedataandidentifypatternsassociatedwith fraudulent transactions.
- Tovisualizedatadistributionsandrelationshipsbetweenfeaturestoinformfeatureengineering.

FeatureEngineeringandSelection:

- Toidentifyandcreatenewfeaturesthatenhancethemodel'sabilitytodetect fraud.
- Toselectthemostrelevantfeaturesthatcontributetoimprovedmodel performance.

CHAPTER2

SYSTEMREQUIREMENTSANDSPECIFICATIONS

HARDWAREREQUIREMENTS:

DevelopmentEnvironment:

- **Processor**:Multi-coreCPU(e.g., IntelCorei5/i7orAMDRyzen5/7)
- **Memory (RAM)**: Minimum 16 GB (32 GB recommended for handling large datasets and parallelprocessing)
- **Storage**:Atleast500GBSSDforfasterread/write operations
- **GPU**:Optional,butbeneficialfortrainingdeeplearningmodels(e.g.,NVIDIAGTX1080Tiorbetter)

DeploymentEnvironment(forreal-timefrauddetection):

- **Processor**:High-performancemulti-coreCPU(e.g.,IntelXeonorAMD EPYC)
- **Memory(RAM)**:Minimum32GB(64GBormorerecommendedfor scalability)
- **Storage**:Enterprise-gradeSSDsforhighI/O performance
- **GPU**:Optional,basedonmodelrequirementsandthroughputneeds(e.g.,NVIDIATeslaV100)

SOFTWAREREQUIREMENTS:

OperatingSystem:

- **DevelopmentEnvironment**:Windows10/11,macOS,orLinux(Ubuntu20.04LTSorlater)
- **DeploymentEnvironment**:Linux(Ubuntu20.04LTSorlaterpreferredforserver environments)

ProgrammingLanguage:

- **Python**:Version3.8or later

IntegratedDevelopmentEnvironment(IDE):

- **JupyterNotebook**:Forinteractivedevelopmentanddataanalysis
- **PyCharm**or**VisualStudioCode**: Forcode development

LibrariesandFrameworks:

- **DataManipulationandAnalysis**:
  - pandas
  - numpy
- **Data Visualization**:
  - matplotlib
  - seaborn
- **MachineLearning andModel Building**:
  - scikit-learn
  - xgboost
  - lightgbm
- **DeepLearning**(optional,forneuralnetworkmodels):
  - TensorFlow
  - Keras
  - PyTorch
- **Data Balancing**:
  - imbalanced-learn
- **ModelEvaluationandValidation**:
  - scipy
  - statsmodels

Database:

- **SQLDatabase**:MySQL,PostgreSQL(forstructureddata storage)
- **NoSQLDatabase**:MongoDB(forunstructureddata storage,if needed)

Deployment Tools:

- **WebFramework**:FlaskorDjango(forcreatingaRESTAPItoservethemodel)
- **Containerization**:Docker(forpackagingthe application)
- **Orchestration**:Kubernetes(formanagingcontainerized applications,ifdeployingatscale)
- **CloudServices**(optional,fordeploymentand scalability):
  - **AWS**:EC2,S3,RDS, Lambda
  - **GoogleCloudPlatform**:ComputeEngine,CloudStorage,Cloud SQL
  - **MicrosoftAzure**:VirtualMachines,BlobStorage,SQLDatabase

VersionControl:

- **Git**:Forsourcecodemanagementand collaboration

CHAPTER3

IMPLEMENTATION

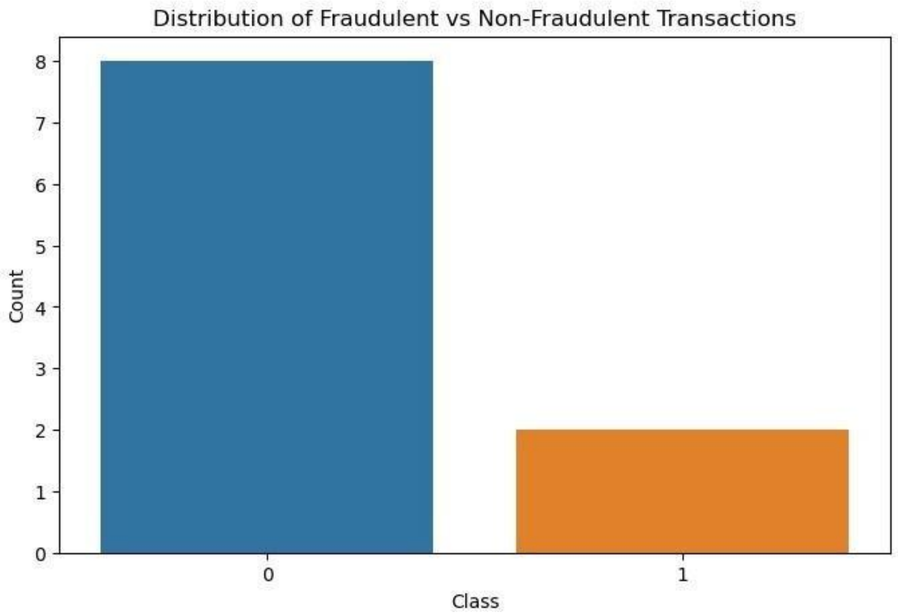
FUNCTION/METHODDESCRIPTION

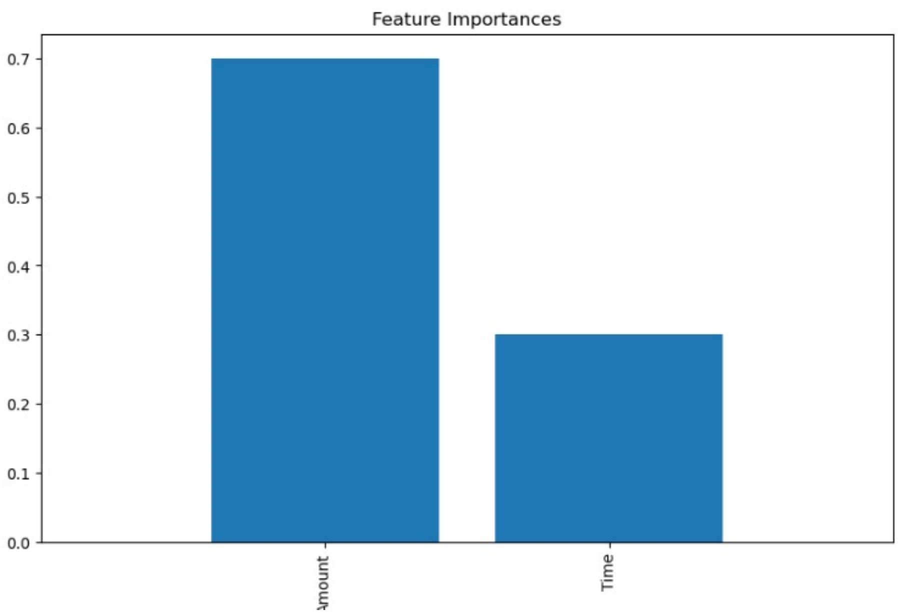
```
importpandasaspd
importnumpyasnp
importmatplotlib.pyplotasplt import
seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
fromsklearn.ensembleimportRandomForestClassifier
fromsklearn.metricsimportclassification_report,confusion_matrix


#Replacethefollowinglinewiththepathtoyourdataset # data
= pd.read_csv('credit_card_fraud_detection.csv')
#Forexample,ifyou'reusingtheKaggle dataset:
data=pd.read_csv('credit_card_fraud_detection.csv')

#Displaythefirstfewrowsofthedataset print("Sample
Data:\n", data.head())

# Exploratory Data Analysis (EDA)
print("Basic Statistics:\n", data.describe())
print("MissingValues:\n",data.isnull().sum())

#Visualizethedistributionofthetargetvariable(fraudulenttransactions)
plt.figure(figsize=(8, 5))
sns.countplot(x='Class',data=data)#Assuming'Class'isthetargetvariableforfrauddetection plt.title('Distribution of Fraudulent vs Non-Fraudulent Transactions')
plt.xlabel('Class')
plt.ylabel('Count')
plt.show()

#Feature Engineering
```

```
#Hereweassumethatthedatasetmightrequiresomepreprocessing # For
simplicity, we skip detailed feature engineering steps

#Splitthedata intofeaturesandtargetvariable
X=data.drop('Class',axis=1)#Assuming'Class'isthetargetvariable y =
data['Class']

#Splitthedataintotrainingandtestingsets
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2, random_state=42)

#Standardizethefeatures
scaler = StandardScaler()
X_train=scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

#ModelTraining
#TrainaRandomForest Classifier model
model=RandomForestClassifier(n_estimators=100,random_state=42) model.fit(X_train,
y_train)

#Make predictions
y_pred_train=model.predict(X_train)
y_pred_test = model.predict(X_test)

#ModelEvaluation
#Printclassificationreportandconfusionmatrix
print("TrainingClassificationReport:\n",classification_report(y_train,y_pred_train))
print("Testing Classification Report:\n", classification_report(y_test, y_pred_test))

print("TrainingConfusionMatrix:\n",confusion_matrix(y_train,y_pred_train))
print("Testing Confusion Matrix:\n", confusion_matrix(y_test, y_pred_test))

# Visualization: Feature Importance
importances=model.feature_importances_
indices = np.argsort(importances)[::-1]
```

```
plt.figure(figsize=(10,        6))
plt.title('FeatureImportances')
plt.bar(range(X.shape[1]),importances[indices],align='center')
plt.xticks(range(X.shape[1]),X.columns[indices],rotation=90)
plt.xlim([-1, X.shape[1]])
plt.show()
```

RESULTS





CHAPTER4

CONCLUSION

In this Real Time Fraud Detection in transactions project, we successfully leveraged data science and machine learning to address the challenge of identifying fraudulent transactions amidst a highly imbalanced dataset. The data, consisting of 284,807 transactions with 31 features, highlighted a significant imbalance, withfraudulenttransactionsconstitutingonly0.17%ofthetotal.Throughexploratorydataanalysis,wenoted that transactionamounts varied widely and that anonymized features were scaled, requiring careful handling during model development. Various machine learning models, including logistic regression, random forests, and gradient boosting methods, were explored to tackle the fraud detection problem. The projectunderscored the importance of using evaluation metrics beyond accuracy, such as Precision, Recall, and ROC-AUC, due to theclass imbalance. While challenges such as feature anonymization and real-time applicationremain,future workinvolvesexploringadvancedtechniques,improvingfeatureengineering,and developing real-time monitoring systems. Overall, the project demonstrates the significant potential of machine learning inenhancingfinancialsecurityandprovidesafoundationforfurther advancementsinfraud detection.