Prediction of Credit Card Fraudusing Machine Learning

T.Manikandan

AssistantProfessor
DepartmentofComputerScienceandEngineeringGalgotiasUniversity

Abstract—
Itisvitalthatcreditcardcompaniesareabletoidentifyfraudulentcreditcardtransactionssothatcustomersarenotcharg edforitemsthattheydidnotpurchase.SuchproblemscanbetackledwithDataScienceanditsimportance,alongwithMac hineLearning,cannotbeoverstated.Thisprojectintendstoillustratethemodellingofadatasetusingmachinelearningwi thCreditCardFraudDetection.TheCreditCardFraudDetectionProblemincludesmodellingpastcreditcardtransacti onswiththedataoftheonesthatturnedouttobefraud.Thismodelisthenusedtorecognizewhetheranewtransactionisfrau dulentornot.Ourobjectivehereistodetect100%ofthefraudulenttransactionswhileminimizingtheincorrectfraudclas sifications.CreditCardFraudDetectionisatypicalsampleofclassification.Inthisprocess,wehavefocusedonanalysing andpre-
processingdatasetsaswellasthedeploymentofmultipleanomalydetectionalgorithmssuchasLocalOutlierFactorandI solationForestalgorithmonthePCAtransformedCreditCardTransactiondata.
Keywords—
Creditcardfraud,applicationsofmachinelearning,datascience,isolationforestalgorithm,localoutlierfactor,automa tedfrauddetection.

I.   INTRODUCTION

'Fraud'increditcardtransactionsisunauthorizedandunwantedusageofanaccountbysomeoneotherthantheo wnerofthataccount. Necessarypreventionmeasurescanbetakentostopthisabuseandthebehaviourofsuchfraudulentpr acticescanbestudiedtominimizeitandprotectagainstsimilaroccurrencesinthefuture. Inotherwords, CreditCardFrau dcanbedefinedasacasewhereapersonusessomeoneelse'screditcardforpersonalreasonswhiletheownerandthecardis suingauthoritiesareunawareofthefactthatthecardisbeingused.

Frauddetectioninvolvesmonitoringtheactivitiesofpopulationsofusersinordertoestimate, perceiveoravoido bjectionablebehaviour, whichconsistoffraud, intrusion, anddefaulting.

Thisisaveryrelevantproblemthatdemandstheattentionofcommunitiessuchasmachinelearninganddatascie ncewherethesolutiontothisproblemcanbeautomated.

Thisproblemisparticularlychallengingfromtheperspectiveoflearning, asitischaracterizedbyvariousfactors suchasclassimbalance. Thenumberofvalidtransactionsfaroutnumberfraudulentones.

LavyaAggrawal,AbhinavSingh,ShivangiJadon.
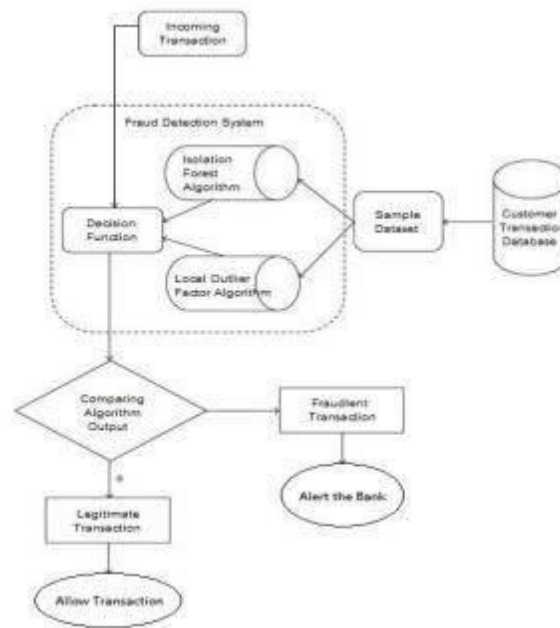DepartmentofComputerScienceandEngineeringGalgotiasUniversity
abhinav. 20scse1290006 @ galotiasuniversity. edu. inlavya. 20scse1180011 @ galgotiasuniversity. edu. inshiv angi. 1050014 @ galgotiasuniversity. edu. in

Thesearenottheonlychallengesintheimplementationofareal-
worldfrauddetectionsystem,however. Inrealworldexamples, themassivestreamofpaymentrequestsisquicklyscann edbyautomatictoolsthatdeterminewhichtransactionstoauthorize.
Machinelearningalgorithmsareemployedtoanalysealltheauthorizedtransactionsandreportthesuspiciousones. Thes ereportsareinvestigatedbyprofessionalswhocontactthecardholderstoconfirmifthetransactionwasgenuineorfraudul ent.
Theinvestigatorsprovideafeedbacktotheautomatedsystemwhichisusedtotrainandupdatethealgorithmtoeventually improvethefraud- detectionperformanceovertime.

Frauddetectionmethodsarecontinuouslydevelopedtodefendcriminalsinadaptingtotheirfraudulentstrategies. Thesefraudsareclassifiedas:
• CreditCardFrauds: OnlineandOffline
• CardTheft
• AccountBankruptcy
• DeviceIntrusion
• ApplicationFraud
• CounterfeitCard
• TelecommunicationFraud

Someofthecurrentlyusedapproachestodetectionofsuchfraudare:
• ArtificialNeuralNetwork
• FuzzyLogic
• GeneticAlgorithm
• LogisticRegression
• Decisiontree
• SupportVectorMachines
• BayesianNetworks
• HiddenMarkovModel
• K- NearestNeighbour

II.   LITERATUREREVIEW

Fraudactastheunlawfulorcriminaldeceptionintendedtoresultinfinancialorpersonalbenefit. Itisadeliberateactthatisagainstthelaw, ruleorpolicywithanaimtoattainunauthorizedfinancialbenefit.
Numerousliteraturespertainingtoanomalyorfrauddetectioninthisdomainhavebeenpublishedalreadyandareavailableforpublicusage. AcomprehensivesurveyconductedbyCliftonPhuaandhisassociateshaverevealedthattechniquesemployedinthisdomainincludeddataminingapplications, automatedfrauddetection, adversarialdetection. Inanotherpaper, Suman, ResearchScholar, GJUS&TatHisarHCEpresentedtechniqueslikeSupervisedandUnsupervisedLearningforcreditcardfrauddetection. Eventhoughthesemethodsandalgorithmsfetchedanunexpectedsuccessinsomeareas, theyfailedtoprovideapermanentandconsistentsolutiontofrauddetection.
Asimilarresearchdomainwaspresentedby Wen-FangYUandNaWangwheretheyusedOutliermining, OutlierdetectionminingandDistancesumalgorithmstoaccuratelypredictfraudulenttransactioninanemulationexperimentofcreditcardtransactiondatasetofonecertaincommercialbank. Outlierminingisafieldofdataminingwhichisbasicallyusedinmonetaryandinternetfields. Itdealswithdetectingobjectsthataredetachedfromthemainsystemi. e. thetransactionsthataren' tgenuine. Theyhavetakenattributesofcustomer' sbehaviourandbasedonthevalueofthoseattributesthey' vecalculatedthatdistancebetweentheobservedvalueofthatattributeanditspredeterminedvalue. Unconventionaltechniquessuchashybriddatamining/ complexnetworkclassifi

---

cationalgorithmisabletoperceiveillegalinstancesinanactualcardtransactiondataset, basedonnetworkreconstruction algorithmthatallowscreatingrepresentationsofthedeviationofoneinstancefromareferencegrouphaveprovedefficien ttypicallyonmediumsizedonlinetransaction.

Therehavealsobeeneffortstoprogressfromacompletelynewaspect. Attemptshavebeenmadetoimprovethealert-feedbackinteractionincaseoffraudulenttransaction.

Incaseoffraudulenttransaction, theauthorisedsystemwouldbealertedandafeedbackwouldbesenttodenytheongoingt ransaction.
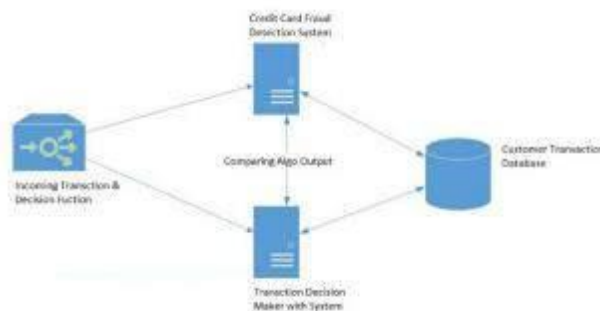
ArtificialGeneticAlgorithm, oneoftheapproachesthatshednewlightinthisdomain, counteredfraudfromadifferentdir ection.

Itprovedaccurateinfindingoutthefraudulenttransactionsandminimizingthenumberoffalsealerts. Eventhough, it wasaccompaniedbyclassificationproblemwithvariablemisclassificationcosts.

III. METHODOLOGY

Theapproachthatthispaperproposes, usesthelatestmachinelearningalgorithmstodetectanomalousactivities, calledo utliers.

Thebasicrougharchitecturediagramcanberepresentedwiththefollowingfigure:

Whenlookedatindetailonalargerscalealongwithreallifeelements, thefullarchitecturediagramcanberepresentedasfol lows:
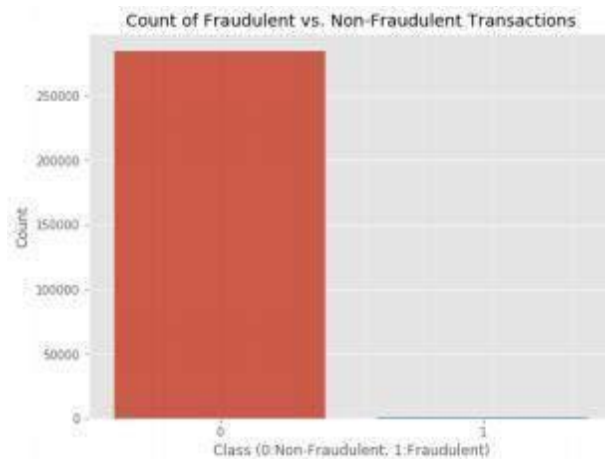

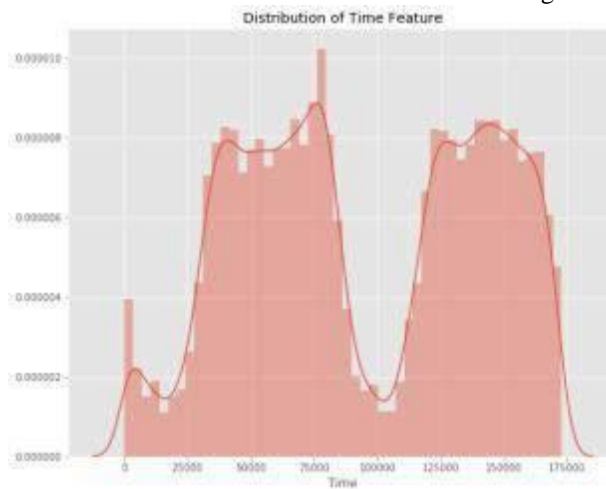Firstofall, weobtainedourdatasetfromKaggle, adataanalysiswebsitewhichprovidesdatasets.

Insidethisdataset, thereare31columnsoutofwhich28arenamedasv1-v28toprotectsensitivedata.

TheothercolumnsrepresentTime, AmountandClass. Timeshowsthetimegapbetweenthefirsttransactionandthefollo wingone. Amountistheamountofmoneytransacted. Class0representsavalidtransactionand1representsafraudulento ne.

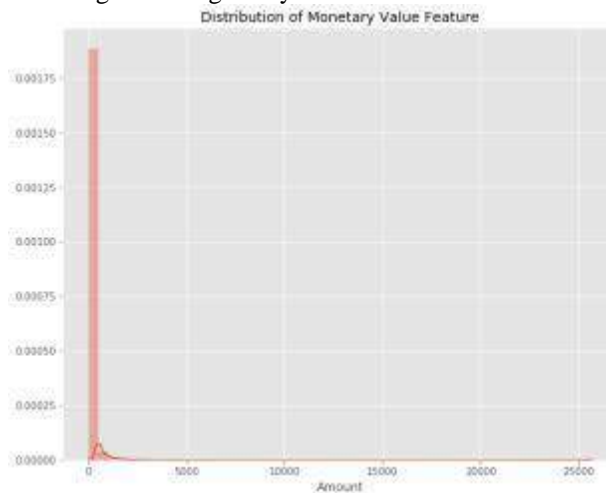Weplotdifferentgraphstocheckforinconsistenciesinthedatasetandtovisuallycomprehendit:

Count of Fraudulent vs. Non-Fraudulent Transactions

Thisgraphshowsthatthenumberoffraudulenttransactionsismuchlowerthanthelegitimateones.

Distribution of Time Feature

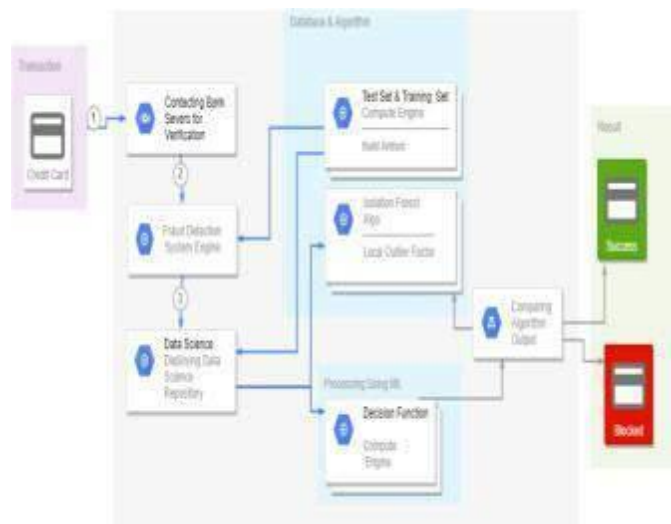Thisgraphshowsthetimesatwhichtransactionsweredonewithintwodays. Itcanbeseenthattheleastnumberoftransacti
onsweremadeduringnighttimeandhighestduringthedays.

Distribution of Monetary Value Feature

Thisgraphrepresentstheamountthatwastransacted. Amajorityoftransactionsarerelativelysmallandonlyahandfuloft
hemcomeclosetothemaximumtransactedamount.
Aftercheckingthisdataset,weplotahistogramforeverycolumn. Thisisdonetogetagraphicalrepresentationofthedatas
etwhichcanbeusedtoverifythattherearenomissing
anyvaluesinthedataset. Thisisdonetoensurethatwedon'trequireanymissingvalueimputationandthemachinelearnin
galgorithmscanprocessthedatasetsmoothly.

Afterthisanalysis, weplotaheatmaptogetacolouredrepresentationofthedataandtostudythecorrelationbetweenoutpredictingvariablesandtheclassvariable. Thisheatmapisshownbelow:

Thedatasetisnowformattedandprocessed. ThetimeandamountcolumnarestandardizedandtheClasscolumnisremovedtoensurefairnessofevaluation. Thedataisprocessedbyasetofalgorithmsfrommodules. Thefollowingmodulediagramexplainshowthesealgorithmsworktogether: Thisdataisfitintoamodelandthefollowingoutlierdetectionmodulesareappliedonit:

• LocalOutlierFactor
• IsolationForestAlgorithm

Thesealgorithmsareapartofsklearn. Theensemblemoduleinthesklearnpackageincludesensemble-basedmethodsandfunctionsfortheclassification,regressionandoutlierdetection.

Thisfreeandopen-sourcePythonlibraryisbuiltusingNumPy, SciPyandmatplotlibmoduleswhichprovidesalotofsimpleandefficienttoolswhichcanbeusedfordataanalysis andmachinelearning. Itfeaturesvariousclassification, clusteringandregressionalgorithmsandisdesignedtointeroperatewiththenumericalandscientificlibraries. We'veusedJupyterNotebookplatformtomakeaprograminPythontodemonstratetheapproachthatthispapersuggests. ThisprogramcanalsobeexecutedonthecloudusingGoogleCollabplatformwhichsupportsallpythonnotebookfiles. Detailedexplanationsaboutthemoduleswithpseudocodesfortheiralgorithmsandoutputgraphsaregivenasfollows:

A. LocalOutlierFactor

ItisanUnsupervisedOutlierDetectionalgorithm. 'LocalOutlierFactor' referstotheanomalyscoreofeachsample. Itme
asuresthelocaldeviationofthesampledatawithrespecttoitsneighbours.

Moreprecisely, localityisgivenbyk-

nearestneighbours, whosedistanceisusedtoestimatethelocaldata. Thepseudocodeforthisalgorithmiswrittenas:

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import IsolationForest

rng = np.random.RandomState(42)

# Generate train data
X = 0.3 * rng.randn(100, 2)
X_train = np.r_[X + 2, X - 2]
# Generate some regular novel observations
X = 0.3 * rng.randn(20, 2)
X_test = np.r_[X + 2, X - 2]
# Generate some abnormal novel observations
X_outliers = rng.uniform(low=-4, high=4, size=(20, 2))

# fit the model
clf = IsolationForest(behaviour='new', max_samples=100,
                      random_state=rng, contamination='auto')
clf.fit(X_train)
y_pred_train = clf.predict(X_train)
y_pred_test = clf.predict(X_test)
y_pred_outliers = clf.predict(X_outliers)

# plot the line, the samples, and the nearest vectors to the plane
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
```
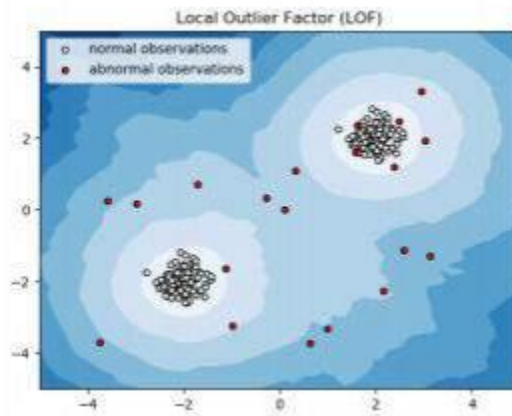
OnplottingtheresultsofLocalOutlierFactoralgorithm, wegetthefollowingfigure:

Bycomparingthelocalvaluesofasampletothatofitsneighbours, onecanidentifysamplesthataresubstantiallylowertha
ntheirneighbours. Thesevaluesarequiteamanousandtheyareconsideredasoutliers.

Asthedatasetisverylarge, weusedonlyafractionofitinoutteststoreduceprocessingtimes.

Thefinalresultwiththecompletedatasetprocessedisalsodeterminedandisgivenintheresultssectionofthispaper.

B. IsolationForestAlgorithm

TheIsolationForest'isolates' observationsbyarbitrarilyselectingafeatureandthenrandomlyselectingasplitvaluebet
weenthemaximumandminimumvaluesofthedesignatedfeature.

Recursivepartitioningcanberepresentedbyatree, thenumberofsplitsrequiredtoisolateasampleisequivalenttothepathl
engthrootnodetoterminatingnode.

Theaverageofthispathlengthgivesameasureofnormalityandthedecisionfunctionwhichweuse.

Thepseudocodeforthisalgorithmcanbewrittenas:

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import LocalOutlierFactor

np.random.seed(42)

# Generate train data
X = 0.3 * np.random.randn(100, 2)
# Generate some abnormal novel observations
X_outliers = np.random.uniform(low=-4, high=4, size=(20, 2))
X = np.r_[X + 2, X - 2, X_outliers]

# fit the model
clf = LocalOutlierFactor(n_neighbors=20)
y_pred = clf.fit_predict(X)
y_pred_outliers = y_pred[200:]

# plot the level sets of the decision function
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = clf._decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
```
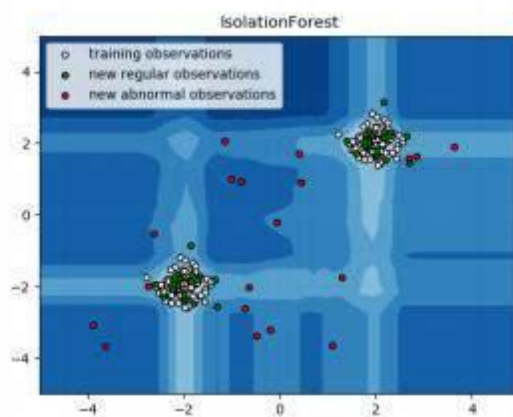
OnplottingtheresultsofIsolationForestalgorithm,wegetthefollowingfigure:

Partitioningthemrandomlyproducesshorterpathsforanomalies. Whenaforestofrandomtreesmutuallyproducesshort erpathlengthsforspecificsamples,theyareextremelylikelytobeanomalies.
Oncetheanomaliesaredetected,thesystemcanbeusedtoreportthemtotheconcernedauthorities. Fortestingpurposes, wearecomparingtheoutputsofthesealgorithmstodeterminetheiraccuracyandprecision.

IV.IMPLEMENTATION

Thisideaisdifficulttoimplementinreallifebecauseitrequiresthecooperationfrombanks, whicharen'twillingtoshareinformationduetotheirmarketcompetition,andalsoduetolegalreasonsandprotectionofdataoftheirusers. Therefore, we lookedupsomereferencepaperswhichfollowedsimilarapproachesandgatheredresults. Asstatedinoneofthesereferencepapers:

"ThistechniquewasappliedtoafullapplicationdatasetsuppliedbyaGermanbankin2006.Forbankingconfidentialityreasons, onlyasummaryoftheresultsobtainedispresentedbelow. Afterapplyingthistechnique,thelevel1listencompassesafewcasesbutwithahighprobabilityofbeingfraudsters.
Allindividualsmentionedinthislisthadtheircardsclosedtoavoidanyriskduetotheirhigh-

riskprofile. Theconditionismorecomplexfortheotherlist. Thelevel2listisstillrestrictedadequatelytobecheckedonacasebycasebasis. Creditandcollectionofficersconsideredthathalfofthecasesinthislistcouldbeconsideredassuspicious fraudulentbehaviour. Forthelastlistandthelargest, theworkisequitablyheavy. Lessthanathirdofthemaresuspicious. Inordertomaximizethetimeefficiencyandtheoverheadcharges, apossibilityistoincludeanewelementinthequery;this

elementcanbethefivefirstdigitsofthephonenumbers, theemailaddress, andthepassword, forinstance, thosenewquer iescanbeappliedtothelevel2listandlevel3list.".

V.RESULTS

Thecodeprintsoutthenumberoffalsepositivesitdetectedandcomparesitwiththeactualvalues. Thisisusedtocalculateth eaccuracyscoreandprecisionofthealgorithms. Thefractionofdataweusedforfastertestingis10% oftheentiredataset. Thecompletedatasetisalsousedattheendandboththeresultsareprinted.

Theseresultsalongwiththeclassificationreportforeachalgorithmisgivenintheoutputasfollows, whereclass0meansth etransactionwasdeterminedtobevalidand1 meansitwasdeterminedasafraudtransaction.

Thisresultmatchedagainsttheclassvaluestocheckforfalsepositives.

Resultswhen10% ofthedatasetisused:

```
Isolation Forest
Number of Errors: 71
Accuracy Score: 0.99750711000316

           precision   recall  f1-score   support

       0       1.00     1.00      1.00     28432
       1       0.28     0.29      0.28        49

    accuracy                       1.00     28481
   macro avg   0.64     0.64      0.64     28481
weighted avg   1.00     1.00      1.00     28481


Local Outlier Factor
Number of Errors: 97
Accuracy Score: 0.9965942207085425

           precision   recall  f1-score   support

       0       1.00     1.00      1.00     28432
       1       0.02     0.02      0.02        49

    accuracy                       1.00     28481
   macro avg   0.51     0.51      0.51     28481
weighted avg   1.00     1.00      1.00     28481
```

Resultswiththecompletedatasetisused:

```
Isolation Forest
Number of Errors: 659
Accuracy Score: 0.9976861523768727

           precision   recall  f1-score   support

       0       1.00     1.00      1.00    284315
       1       0.33     0.33      0.33       492

    accuracy                       1.00    284807
   macro avg   0.66     0.67      0.66    284807
weighted avg   1.00     1.00      1.00    284807


Local Outlier Factor
Number of Errors: 935
Accuracy Score: 0.9967170750718908

           precision   recall  f1-score   support

       0       1.00     1.00      1.00    284315
       1       0.05     0.05      0.05       492

    accuracy                       1.00    284807
   macro avg   0.52     0.52      0.52    284807
weighted avg   1.00     1.00      1.00    284807
```

VI. CONCLUSION

Creditcardfraudiswithoutadoubtanactofcriminaldishonesty. Thisarticlehaslistedoutthemostcommonmethodsoffra udalongwiththeirdetectionmethodsandreviewedrecentfindingsinthisfield. Thispaperhasalsoexplainedindetail, ho wmachinelearningcanbeappliedtogetbetterresultsinfrauddetectionalongwiththealgorithm, itsimplementationand pseudocode, explanationexperi mentationresults.

Whilethealgorithmdoesreachover9 9 . 6 % accuracy, itsprecisionremainsonlyat2 8 % whenatenthofthedatasetistake nintoconsideration. However, whentheentiredatasetisfedintothealgorithm, theprecisionrisesto3 3 % . Thishighperc entageofaccuracyistobeexpectedduetothehugeimbalancebetweenthenumberofvalidandnumberofgenuinetransacti ons.

Sincetheentiredatasetconsistsofonlytwodays' transactionrecords, itsonlyafractionofdatathatcanbemadeavailableif thisprojectweretobeusedonacommercialscale. Beingbasedonmachinelearningalgorithms, theprogramwillonlyinc reaseitsefficiencyovertimeasmoredataisputintoit.

VII. FUTUREENHANCEMENTS

Whilewecouldn' treachoutgoalof1 0 0 % accuracyinfrauddetection, wedidendupcreatingasystemthatcan, withenou ghtimeanddata, getveryclosetothatgoal. Aswithanysuchproject, thereissomeroomforimprovementhere. Theveryn atureofthisprojectallowsformultiplealgorithmstobeintegratedtogetherasmodulesandtheirresultscanbecombinedtoi ncreasetheaccuracyofthefinalresult.

Thismodelcanfurtherbeimprovedwiththeadditionofmorealgorithmsintoit. However, theoutputofthesealgorithmsn eedstobeinthesameformatastheothers. Oncethatconditionissatisfied, themodulesareeasytoaddasdoneinthecode. Th isprovidesagreatdegreeofmodularityandversatilitytotheproject.

Moreroomforimprovementcanbefoundinthedataset. Asdemonstratedbefore, theprecisionofthealgorithmsincrease swhenthesizeofdatasetisincreased. Hence, moredatawillsurelymakethemodelmoreaccurateindetectingfraudsandr educethenumberoffalsepositives. However, thisrequiresofficialsupportfromthebanksthemselves.

REFERENCES

[1]. CreditCardFraudDetectionBasedonTransactionBehaviour-by JohnRichardD. Kho, LarryA. Vea"publishedbyProc. ofthe2017IEEERegion10Conference(TENCON),Malaysia,November5-8,2017

[2]. CLIFTONPHUA1, VINCENTLEE1, KATESMITH1&ROSSGAYLER2"AComprehensiveSurveyofDataMining- basedFraudDetectionResearch"publishedbySchoolofBusinessSystems,FacultyofInformationTechnology,MonashUniversity,Welling tonRoad,Clayton,Victoria3800,Australia

[3]. "SurveyPaperonCreditCardFraudDetectionbySuman",ResearchScholar,GJUS&THisarHCE,SonepatpublishedbyInternationalJourn alofAdvancedResearchinComputerEngineering&Technology(IJARCET)Volume3Issue3,March2014

[4]. "ResearchonCreditCardFraudDetectionModelBasedonDistance Sum–byWen-FangYUandNaWang"publishedby2009InternationalJointConferenceonArtificialIntelligence

[5]. "CreditCardFraudDetectionthroughParencliticNetworkAnalysis- ByMassimilianoZanin,MiguelRomance,ReginoCriado,andSantiagoMoral"publishedbyHindawiComplexityVolume2018,ArticleID 5764370,9pages

[6]. "CreditCardFraudDetection:ARealisticModelingandaNovel LearningStrategy"publishedbyIEEETRANSACTIONSONNEURALNETWORKSANDLEARNINGSYSTEMS,VOL.29,NO.8,AU GUST2018

[7]. "CreditCardFraudDetection- byIshuTrivedi,Monika,Mrigya,Mridushi"publishedbyInternationalJournalofAdvancedResearchinComputerandCommunicationEngi neeringVol.5,Issue1,January2016

[8]. DavidJ.Wetson,DavidJ.Hand,MAdams,WhitrowandPiotrJusczak"PlasticCardFraudDetectionusingPeerGroupAnalysis"Springer,Is sue2008.