MachineLearningrNanodegree Shanmukha Mudigonda CapstoneProposal Date:August 02,2019

DomainBackground

AsreportedbyBusinessInsider[1],lastyear,CreditCards.comfoundthatcreditcardfraud wasontherise.Bothnumberoffraudsandtypesofcreditcardscamsaremoreandmore. Thesenumbersseemunfortunatelydestinedtogrowevenmoreinthefuture.Forthisreason, itisimportanttofindawaytoautomaticallyrecogniseanomalies.Alotofresearchhasbeen doneinordertofindasolutiontothisproblem.References[6]and[7]proposesolutionsusing ArtificialNeuralNetworks.

**ProblemStatement**

Theaimofthefrauddetectionsystemistodetectfraudaccuratelyandbeforefraudiscom- mitted.Thegoalistodetectleastandaccuratefalsefrauddetection.Themostcommonly techniquesusedfrauddetectionmethodsareRandom Forest and XGB Classifier machine learningalgorithms.

Eachtransactionhasasetofuniquefeatures,suchasthevalueofthetransaction,thetimeatwhichitoccured,issuerandrecipient,anidandothersensitivedata.Theproblemwillbestructuredasaclassificationproblem,whereeachtransactioncouldbeclassifiedas”Normal”or”Fraud”.

**DatasetsandInputs**

Inordertoreproducethiskindofproblem,IfoundausefuldatasetavailableonKaggle[4].The datasetscontainstransactionsmadebycreditcardsinSeptember2013byeuropeancardholders. Thisdatasetpresentstransactionsthatoccurredintwodays,wherewehave492fraudsoutof 284,807transactions.Thedatasetishighlyunbalancedandthepositiveclass(frauds)account for0.172%ofalltransactions.

ItcontainsonlynumericalinputvariableswhicharetheresultofaPCAtransformation.Due toconfidentialityissues,theauthorscannotprovidetheoriginalfeaturesandmorebackground informationaboutthedata.FeaturesV1,V2,...V28aretheprincipalcomponentsobtained withPCA,theonlyfeatureswhichhavenotbeentransformedwithPCAare”Time”and ”Amount”.Feature”Time”containsthesecondselapsedbetweeneachtransactionandthe firsttransactioninthedataset.Thefeature”Amount”isthetransactionAmount,thisfeature canbeusedforexample-dependantcost-senstivelearning.Feature”Class”istheresponse variableandittakesvalue1incaseoffraudand0otherwise.

**SolutionStatement**

Autoencodersareaparticularclassofneuralnetworksthattakenaninput,”compress”that inputdowntocorefeaturesandtriestoreconstructtheoriginalinputfromthissqueezed representation.Thiskindofapproach,isusedforapplicationsinwhichwewanttorecognise anomaliesor”distortions”ondata.

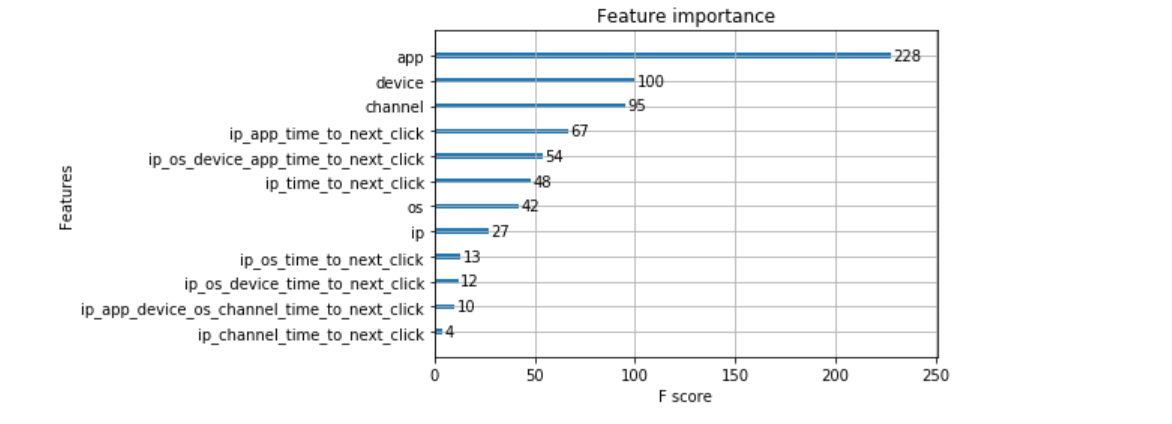
WiththisprojectIwanttoexperimentwithautoencodersinordertodetectfraudsand measureitsresults.Theaimofthisprojectistomeasurethequalityofthistechniquetomake acomparisonwithothers.ThereasonwhyIchoosethismodelisthatIwanttoplayalittle bitwithan”alternative”and(maybe)wrongmodel,tomeasurehowmuchdoesitdifferfrom

1

mostusedimplementations.Thehugeclassimbalanceislikelytocausehugeproblemsfor anautoencoderifIdon’tusesomeverycleverdataaugmentation.Perhapsbyintegratingan adversarialfactorintoautoencoders(ontheencoderend),resultscouldbegoodenoughtobe comparedwith”standars”solutions.

BenchmarkModel

Inordertomeasurethequalityofthisexperimentandtoevaluatetheresults,weneedsome referencemodel.Itseemsreallydifficulttofindothersthatconductedandsharedtheresult ofthiskindofanalysis.Aftersomeresearch,Ifoundtwoothersolutionsontheexactsame datasetthatIselected.Reference[5]recognisefraudsusingmachine learning algorithm XGB with93.5% ofaccuracywhilereference[6]



EvaluationMetrics

IwilltrytooptimisetheparametersoftheAutoencoderinsuchwaythatthereconstruction errorisminimised.ROCandPrecision-Recallcurveswillbeusedinordertoevaluatethe results

.

**Data Preprocessing**

The first step taken in pre-processing the dataset was **Normalization.** The Normalization   
procedure was applied only on the **Amount** Feature since it wasn’t on the same scale as the other features.

**ProjectDesign**

Theprojectwillbedevelopedfollowingthosesteps:

1.Iwillconductatfirstanexploratoryanalysistohaveabetterunderstandingofthedata;

2.theneuralnetworkwillbetrainedandfixedwiththerightparameters;

3.attheendIwillevaluatetheresultsandIwillcomparethemtothebenchmarkreferences.

2

NoticethatinthedatasetthatIhavechosen,”Time”isthesecondselapsedbetweeneach transactionandthefirsttransactioninthedataset.Itisunknownduringwhattimeofthe daythetransactionsactuallybegan.Then,thecolumncanatmostinformushowclosethe transactionsweremadeinbetween2fraudulentones.SinceIwanttopredictfraudsregardless oftransactiontime,thisfeaturewillbedroppedearlier.Noticethattimeandamounthave verydifferentmagnitudesinthedataset,whichwilllikelyresultinthelargemagnitudevalue ”washingout”thesmallmagnitudevalue.Itwouldbebettertoscalethedatatosimilarmag- nitudes.MostofthedataresultfromtheproductofaPCAanalysis.Iwilldothesametothe ”amount”column(rememberthattimewillbedropped!).

Ineedtofaceanotherproblemalso.Fraudulenttransactionsaresignificantlylowerthan normalhealthytransactionsi.e.accountingittoaround1or2%ofthetotalnumberofobser- vations.WiththeautoencoderIwillgeneratedataforaugmentingthedataset.

.

**Implementation:**

The implementation stage involved creating a training and predicting pipeline .This stage involved testing four algorithms to see which best suits the problem. The following   
algorithms were used;

I. XGB Classifier

II. Random Forest

The following metrics were used to measure performance as the data set scaled;

I. Time taken to train the model.

II. Time taken to make predictions on the train and cross validation set.

References

[1]CreditCardFraudDetection-Kaggle.Availableat:<https://www.kaggle.com/mlg->   
ulb/creditcardfraud.

[5]CaseStudy:HowtoImplementCreditCardFraudDetectionUsingMachine Learning.   
Availableat: https://www.altexsoft.com/whitepapers/fraud-detection-how-machine-learning-systems-help-reveal-scams-in-fintech-healthcare-and-ecommerce/

[[6]](https://www.romexsoft.com/blog/implement-credit-card-fraud-detection/%5b6%5d)CreditCardFraudDetectionviaXGB Classifer and Random Forest.Availableat:   
<https://www.kaggle.com/yuridias/credit-card-fraud-detection-RandomForest>

3