Detection of Credit Card Fraud using ML techniques.

Ing. Alexandru-Gabriel Enache

Agenda

- Introduction
- Dataset and QA
- Data Exploration and Visualization
- Methodology employed
- Model I: Logistic Regression
- Model II: Autoencoder
- Model Evaluation
- Results and conclusion

Introduction

- Credit Card Fraud is the most common form of identity theft right now.
- According to Experian, a UK based credit bureau, Credit Card Fraud has soared to a 10-year high as of 2022.
- Over 150 million Americans have been victims of fraud in 2022, up from 127 million in 2021.

Introduction

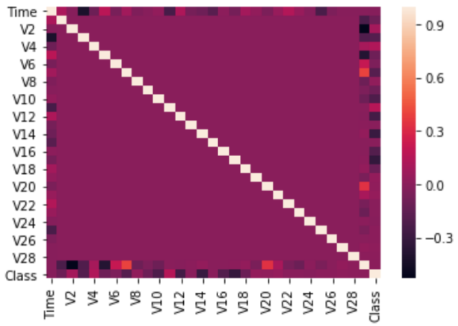
- To counter and minimize exposure of Credit Card holders/consumers, we can use data from Credit Card consumers as well as transaction information in order to trace fraudulent transactions.
- The proposed solution for today’s presentation is two Machine Learning based algorithms to identify a fraudulent credit card transaction.

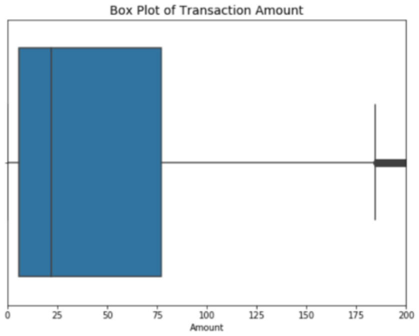
Dataset and QA

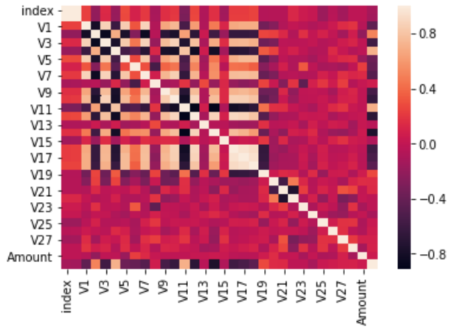
- The Dataset is available on Kaggle in a .csv format.
- The dataset contains transactions made by credit cards in September 2013 by European cardholders.
- This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.
- It contains only numerical input variables which are the result of a PCA transformation.
- Unfortunately, due to confidentiality issues, the dataset doesn't provide the original features and more background information about the data.
- The Dataset is very clean, as it has no NULL values, no empty rows and the values are all unique.

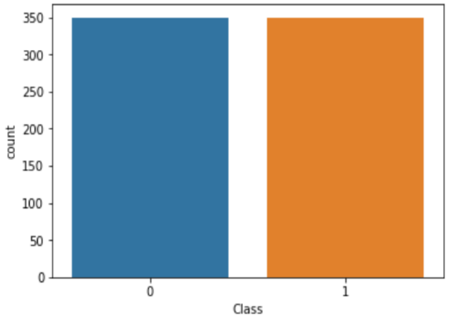
Data Exploration and Visualization

- Using a correlation heatmap there is very little correlation when utilizing the entire Dataset.
- Undersampling the non-fraudulent data, then having a 1:1 ratio of fraudulent data to non-fraudulent data, reveals correlations between features and target variable.









Methodology

Three separate model definitions

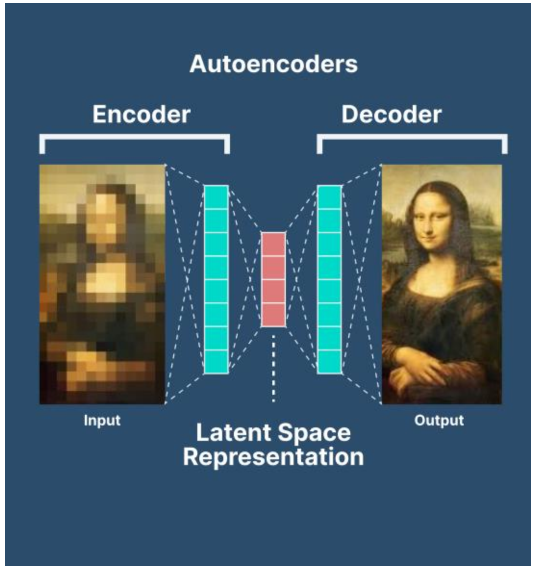
- One local scikit-learn based logistic regression model.
- Two Spark ML based models: A Logistic Regression model and a Autoencoder model.

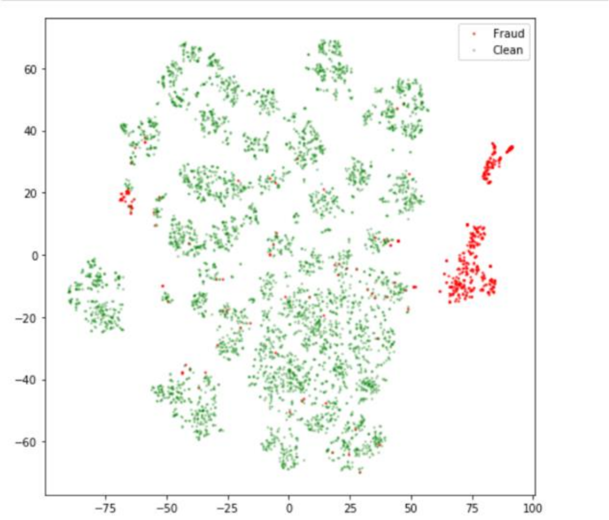
Feature Engineering

- Created Pipeline consisting of
  - Normalization/Scaling
  - Vector Assembler
  - PCA Transformation

Model I: Logistic Regression

- Method I: Local Machine Implementation
  - To have a base target on the dataset.
  - Trained model on subsample of the dataset.
  - Took a 1:1 ratio.
  - Only predicted and evaluated on 2,000 samples to save time.
- Logistic Using Apache Spark
  - Spark ML implementation.
  - Utilized a subset equal to local machine implementation to train the model.
  - Fit model to entire dataset.



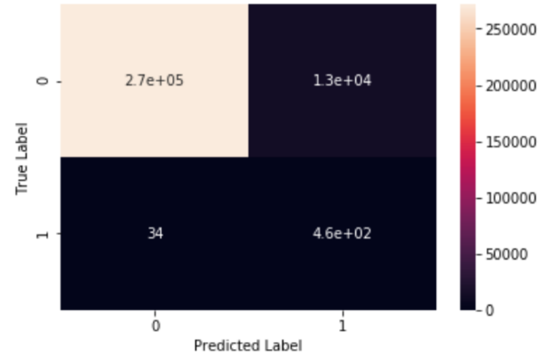


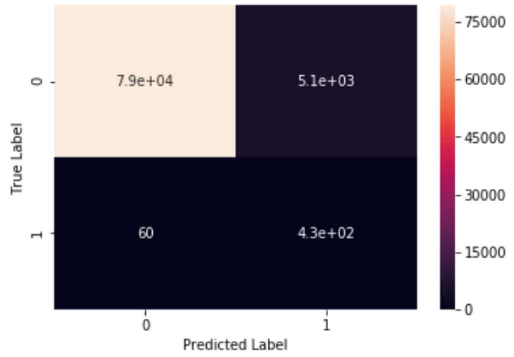
- Utilized a 3 Dense Layer Sequential model.
- Trained only on a subsample of the dataset, 5,000 samples of only non-fraudulent data.
- The Autoencoder was trained to be able to reconstruct non-fraud data.
- Take weights from the encoder part then feed the decoder and see the error deviation of the learned representation.

Model II: Autoencorders

Model Evaluation

- A Confusion Matrix was utilized to assess the Models’ accuracy. The results are as follows:
- Logistic Regression
  - Accuracy on Fraudulent transactions 93.1% detection rate
  - F1-Score: .93
- Autoencoder
  - Accuracy on Fraudulent Transactions 87.8% detection rate
  - F1-Score: .91





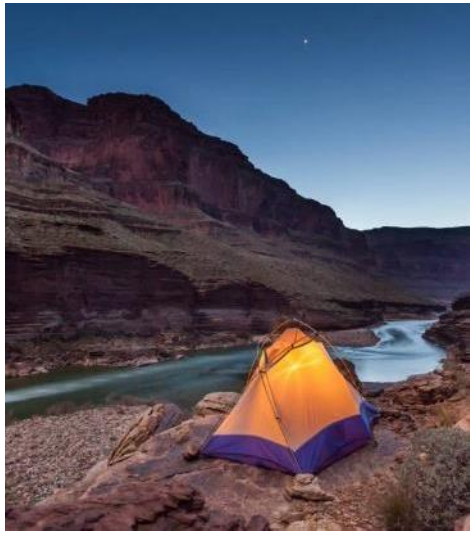
Summary

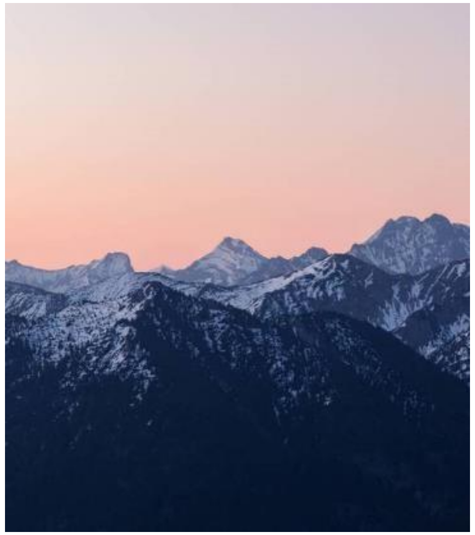
The Logistic Regression was ~5% better in detecting fraudulent transactions than the Autoencorder.

Better results with the Autoencorder can be achieved with more layers and to limit the compression layer of the network.

We eliminated the features of "time" and "amount" as they seemed to have little to no correlation with the "label" class and only got a tenth of a percentage increase in accuracy in both models.

A CNN model on the Local Machine is considered in the works.





Appendix

- The resources are publicly available and can be found on https://github.com/aenache99/ibm-advanced-data-science-capstone
- The contents of the repository are under the Apache-2.0 License.
- More about Credit Card Fraud information:
- https://www.experianplc.com/media/latest-news/2023/credit-card-fraud-soars-to-10-year-high/
- https://www.security.org/digital-safety/credit-card-fraud-report/
- The dataset: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Ing. Alexandru Gabriel Enache

Thank you!