Statement of Independence

Please refer to statement.txt in the same encompassing folder of this document.

Citations

Dataset

- https://zindi.africa/competitions/ai-hack-tunisia-4-predictive-analytics-challenge-1/data

Resampling techniques

- https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html

Evaluation Metrics

- https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.plot_confusion_matrix.html
- https://machinelearningmastery.com/fbeta-measure-for-machine-learning

Multi-Layer Perceptron

- https://www.researchgate.net/figure/Multilayer-Perceptron-Advantages-and-Disadvantages_tbl4_338950098
- https://carpentries-incubator.github.io/machine-learning-novice-sklearn/06-neural-networks/index.html
- https://scikit-learn.org/stable/modules/neural_networks_supervised.html
- https://benjaminobi.medium.com/5-minutes-tutorial-on-how-to-compute-and-visualize-the-covariance-matrix-2597ab98d9ee
- https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

Support Vector Machines

- https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier
- https://towardsdatascience.com/svm-and-kernel-svm-fed02bef1200

- https://towardsdatascience.com/support-vector-machine-simply-explained-fee28eba5496
- https://www.baeldung.com/cs/svm-hard-margin-vs-soft-margin
- https://towardsdatascience.com/the-support-team-svm-555d2c30b1b3
- https://stackoverflow.com/questions/12355434/svm-with-hard-margin-and-c-value
- https://stats.stackexchange.com/questions/74499/what-is-the-loss-function-of-hard-margin-svm
- https://www.section.io/engineering-education/regularization-to-prevent-overfitting/
- https://neptune.ai/blog/fighting-overfitting-with-l1-or-l2-regularization
- https://towardsdatascience.com/solving-svm-stochastic-gradient-descent-and-hinge-loss-8e8b4dd91f5b
- https://medium.com/@divakar_239/stochastic-vs-batch-gradient-descent-8820568eada1

Decision Tree

- https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html
- https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html?highlight=rfe#sklearn.feature_selection.RFE
- https://scikit-learn.org/stable/modules/generated/sklearn.tree.plot_tree.html?highlight=tree
- https://ai.plainenglish.io/hyperparameter-tuning-of-decision-tree-classifier-using-gridsearchcv-2a6ebcaffeda
- https://www.section.io/engineering-education/hyperparmeter-tuning/
- https://machinelearningmastery.com/cost-sensitive-decision-trees-for-imbalanced-classification/how-to-penalize-decision-tree-for-misclassifying-a-certain-class

XGBoost

- https://xgboost.readthedocs.io/en/stable/

k-Nearest Neighbours

- https://towardsdatascience.com/curse-of-dimensionality-a-curse-to-machine-learning-c122ee33bfeb
- https://towardsdatascience.com/the-surprising-behaviour-of-distance-metrics-in-high-dimensions-c2cb72779ea6

Presentation Slides

Background/Motivation

- https://www.sciencedirect.com/science/article/abs/pii/S0045790621003013
- https://www.sciencedirect.com/science/article/pii/S0973082618311177#:~:text=Electricity%20theft%20is%20considered%20as,%2C%20%26%20Rao%2C%202016).
- https://www.prnewswire.com/news-releases/world-loses-893-billion-to-electricity-theft-annually-587-billion-in-emerging-markets-300006515.html
- https://www.oecd.org/derec/unitedkingdom/40700982.pdf

Problem statement

- https://medium.com/razorthink-ai/4-major-challenges-facing-fraud-detection-ways-to-resolve-them-using-machine-learning-cf6ed1b176dd

Discussion

- https://machinelearningmastery.com/fbeta-measure-for-machine-learning
- https://towardsdatascience.com/oversampling-and-undersampling-5e2bbaf56dcf

Images

- LeNail, (2019). NN-SVG: Publication-Ready Neural Network Architecture Schematics. Journal of Open Source Software, 4(33), 747, https://doi.org/10.21105/joss.00747

Proposal

- https://doi.org/10.1038/s41598-022-07337-7
- https://doi.org/10.1016/j.compeleceng.2021.107329
- https://www.kaggle.com/code/khsamaha/lightgbm-fraud-detection-in-elec-and-gaz

Documents

statement.txt - Declaration of Statement of Independence
README.pdf - This document describes this document.
CS3244 Team 05 Presentation Slides.pdf - Slides used in the presentation video
01 - Datasets

- Test
    - invoice_test.csv - raw test invoice data given as part of the competition that was not used in this project. Included for completeness
    - client_test.csv - raw test client data given as part of the competition that was not used in this project. Included for completeness
- Train
    - invoice_with_target.csv - raw invoice data with target information added.
    - invoice_train.csv - raw invoice data
    - client_train.csv - raw client data

- fixed_data_train.csv - our group curated dataset after feature engineering

02 - Feature Engineering
- FinalGenerator.ipynb - The Python notebook that was used to generate our curated dataset
- CovMatrixPlot.png - Plot of covariance matrix
- Header and Meaning - Documentation explaining the various features.
- ML-Dataset-Analysis.ipynb - Scratch notebook used for data analysis
- ShawnAnalysis.ipynb - Scratch notebook used for data analysis
- ZX-CovMatrixPlot.py - Script containing the code that plots the covariance matrix but does not include the data preprocessing.
- ZX-ML-Dataset-Analysis.ipynb - Scratch notebook used for data analysis

03 - Models
- Eth-KNN.ipynb - k-Nearest Neighbours
- NG-DecisionTree.ipynb - Decision Tree
- JY-SVM.ipynb - Support Vector Machine
- ST-XGBClassifier.ipynb - XGBoost Classifier
- ZX-MLP - Folder containing the source code for multi-layer perceptron classifier.
    - ZX-MLP-DataAnalysis-v2.ipynb - notebook that was used to analyse the results.
    - ZX-MLPClassifier-VaryingHidden-HPC-v2.py - script that tries out various combinations of hidden layers after improvements are added.
    - ZX-MLPClassifier-VaryingHidden-HPC.py - script that tries out various combinations of hidden layers.
    - ZX-MLPClassifier.ipynb - notebook used for testing of various code

Extra Notes

The above-mentioned models are trained on the same dataset, fixed_data_train.csv, which was curated from the original datasets given, included in the folders, Test and Train. The additional features are obtained using FinalGenerator.ipynb following the document, Header and Meaning. The raw dataset provided was analysed using the following Python scratch notebooks, ML-Dataset-Analysis, ShawnAnalysis, and ZX-ML-Dataset-Analysis. The covariance matrix of the curated dataset was plotted using the ZX-CovMatrixPlot.py.

Acknowledgements

We would like to thank the NUS IT Research Computing team for providing us with the NUS High Performance Computing to compute our various Machine Learning models.