**A comparative study of Classification Models in Machine Learning**

|  |  |  |  |
| --- | --- | --- | --- |
| Pravesh Gupta | Manjot Kaur Dherdi | Konstantin Chemodanov | Manish Yadav |

**Abstract**

This report aims to answer the following questions. Which classification model is the best choice, given a dataset? How different models behave when trained on different datasets? For the project, we have trained all the models on few datasets selected from UCI repository and the choice of ‘best model’ for dataset is made by evaluating their performances in making predictions. Comparisons among models are made based on ROC AUC scores, training and testing accuracy while also keeping into account the time taken to train the model. Observations are made about the behavior of a model with different types of datasets. Need for data pre-processing is assessed and required feature transformations are implemented.

**Introduction**

The classification task is one of the classical machine learning problems and we decided to investigate binary and multi-class classification to acquire practical knowledge of different machine learning algorithms. Next step on our ML journey was a search of datasets suitable for our study. We sifted through hundreds of datasets available on the website of University of California, Irvine. Finally, we selected 7 datasets from UCI [1] and 1 dataset from city of Montreal open data portal [6]. Montreal crime dataset is an example of the data from our immediate environment collected since 2015 and representing criminal situation while protecting private information, criminal events are localized up to the nearest street intersection.

List of datasets:

|  |  |
| --- | --- |
| **#** | **Name** |
| 1 | Occupancy Detection [2] |
| 2 | Activities of Daily Living Recognition Using Binary Sensors [4] |
| 3 | BitcoinHeist Ransomware Address [7] |
| 4 | Bank Marketing [3] |
| 5 | Montreal Crime [6] |
| 6 | Default of Credit Card Clients [5] |
| 7 | Census Income |
| 8 | Yeast |

Table 1. List of datasets

Datasets #2, #5 and #8 are multiclass datasets, the rest are binary. After datasets inspection we decided that some preprocessing is required, namely categorical data encoding and parsing date values. Additional dataset investigation was done to determine if the model contains outliers.

We agreed on using grid for hyperparameters search and cross validation strategy. We decided to calculate ROC AUC score and plot ROC AUC curve for visual interpretation. Besides the accuracy and confusion matrix we also opted to calculate training and execution time for different classification models. We considered to investigate the need of feature transformations for successful implementation of classification algorithms. Our ultimate goal of the project is comparison of ML models based on their performance and determining if some models work better with specific datasets.

Finally, after careful review of the goals and techniques we started the implementation of machine learning classification algorithms. We randomly divided 8 models between 4 team members, and everybody had a chance to implement their 2 models in a best possible way. Everyone used all 8 datasets for model training and testing. After many days and nights of coding, collaborations, discussions, and troubleshooting we came up with some observations, calculations, and results.

**Methodology and Experimental Results**

Classification Experiments

Regression Experiments

Interpretability Experiments

Conclusions

**References**

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Luis M. Candanedo, Véronique Feldheim. Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. Energy and Buildings. Volume 112, 15 January 2016, Pages 28-39.

[3] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[4] Ordóñez, F.J.; de Toledo, P.; Sanchis, A. Activity Recognition Using Hybrid Generative/Discriminative Models on Home Environments Using Binary Sensors. Sensors 2013, 13, 5460-5477.

[5] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.

[6] Ville de Montréal (2020). Criminal acts. [https://donnees.montreal.ca/ville-de-montreal/actes-criminels].

[7] Akcora, Cuneyt & Li, Yitao & Gel, Yulia & Kantarcioglu, Murat. (2019). BitcoinHeist: Topological Data Analysis for Ransomware Detection on the Bitcoin Blockchain.