Schaffhausen Institute of Technology | Assignment 1
2021/22 Term 2 | Dynamic Analysis and Machine Learning
Assigned Date: 14/03/2022
Due Date: 27/03/2022

Due Date

Late assignments will not be accepted and will receive ZERO mark.

Objectives

The objectives of this assignment are as follows.

- **You will practice Linear regression by applying it to solve the problem of Software Defects Prediction and Logistic regression on the problem of Malware Detection in Scikit-Learn library.**
- **You will perform the necessary preprocessing steps such as: imputation, rescaling and encoding of categorical features.**
- **You will also evaluate the learned models using appropriate metrics from the Scikit-Learn package.**
- **Finally, you will solve a case that needs some research on your behalf which would help us to assess your understanding of how to handle a learning problem when the dataset might be imbalanced.**

1 Linear Regression [3.5 Points]

In this task, you will implement a Linear Regresison model to predict the of number of defects in a software. You will work with the dataset of KC1-class-level from the PROMISE repository. You can visit the dataset link to read more about it, such as the description of different features.

More specifically, you are required to perform the following tasks on the 'defects.csv' dataset:

1. **Read data in Python. Then explore and clean it; display datatypes, and check for missing values.**
2. **Split your data into train and test sets (80% and 20% respectively).**
3. **Apply imputations for the missing values.**
4. **(optional) Some features can be filtered out. Remove them and provide justification(s) for their removal.**
5. **Rescale the dataset with a scaler of your choice.**
6. **Train a linear regression model, validate it on the test set, and report its performance.**
7. **Train a polynomial regression model (try different degrees). Plot the relationship between the degree vs the test loss and the degree vs the training loss on the same plot.**
8. **Answer these questions: Which degree would you choose for your model? Would it be better to choose a higher degree?**
9. **Write your own code to split the the dataset into train and test sets, and use it to split the dataset three times, such that each split has different train and test sets. Use these splits to train three Linear Regression models. Did the models turn out to be the same, or were there any differences? Why?**

10. **List the MSE for the test set in each case and mention the most and the least important feature according to the three built models.**

2 Logistic Regression [1.5 Points]

Android is one of the most famous mobile OS worldwide thanks to its open-source code and its technological impact. However, due to the possibility of installing applications from third parties without any intensive central monitoring, Android has recently become a malware target.

To prevent malware attacks, researchers and developers have proposed different security solutions including static analysis, dynamic analysis, and artificial intelligence. Indeed, data science has become a promising area in cybersecurity, since analytical models based on data allow for the discovery of patterns that can help to predict malicious activities.

In this task, you will use some network layer features to build a machine learning classification model that will detect malware applications, using public benchmarks (datasets).

More specifically, you are required to perform the following tasks on the 'android_traffic.csv' dataset:

1. **Read data in Python. Then explore it; display datatypes, and check for missing values.**
2. **Rescale the dataset with a scaler of your choice.**
3. **Split your data into train and test sets (80% and 20% respectively).**
4. **Train a logistic regression model and evaluate it on test set. Report accuracy, precision, recall, and f1-score.**
5. **Answer the following question: Which metric from the previous step is more appropriate for this task and why?**

3 Imbalanced Dataset [1 Point]

On a different datasets of defects, that shows if a class has a defect or no 'defects2.csv', You will do the following:

- **Print the proportion of each category in the dataset.**
- **Split your data into train and test sets (80% and 20% respectively), while keeping the same proportion of categories in each.**
- **Check the arguments of the LogisticRegression model in Sklearn. Which of them could be used to overcome the problem of imbalance? Why?**
- **train a model with and without using the argument from previous step.**
- **For both models: obtain the best prediction threshold that maximize the precision on the testset.**
- **Report both the precision and threshold from previous step for both models.**

Deliverables

You are required to submit your solutions as a single ipynb file. Please, put your name and university email as the first line in the notebook.

Notes

- **Cheating is a serious academic offense and will be strictly treated for all parties involved. So delivering nothing is always better than delivering a copy.**
- **Organize your notebook appropriately. Divide it into sections and cells with clear titles for each task and subtask.**
- **Make your code clean with the appropriate naming conventions. So maybe you want to take a look at some references: Link 1 and Link 2.**
- **Make sure your plots are descriptive and self-contained.**