Classification Projects on Machine Learning for Beginners - 1
Project Overview

Overview

The Classification algorithm is a type of supervised machine learning technique used to categorize a set of data into classes. For a given example of input data, a classification algorithm assigns the most probable class label. An easy-to-understand example is classifying email as spam or non-spam. There are several use cases of classification in real-world scenarios. This project aims to give you the basic idea related to different algorithms used for classification.

Aim

To predict license status for the given business.

Data Description

The dataset used is a licensed dataset. It contains information about 86K different businesses over various features. The target variable is the status of license which has five different categories.

Tech Stack

- ➔ Language: Python
- ➔ Libraries: pandas, scikit_learn, category_encoders, numpy, os, seaborn, matplotlib

Approach

1. Data Description
2. Exploratory Data Analysis
3. Data Cleaning
    a. Missing Value imputation
    b. Outlier Detection
4. Data Imbalance

5. Data Encoding
6. Model Building
    a. KNN classifier
    b. Naive Bayes algorithm
    c. Logistic Regression
    d. Decision Tree classifier
7. Classification Metrics
    a. Precision
    b. Recall
    c. F1 score
    d. Accuracy
    e. Macro average
    f. Weighted average
8. Feature importance

Modular code overview

```
input
  |_License_Data.csv

lib
  |_EDA.ipynb
  |_Licence Status Multi Label Classification.ipynb

output
  |_model_report.xlsx

src
  |_Engine.py
  |_ ML_pipeline
            |_model_selection.py
            |_preprocessing.py
            |_utils.py

requirements.txt
```

Once you unzip the modular_code.zip file, you can find the following folders within it.

1. input
2. src
3. output
4. lib

5. requirements.txt

1. The input folder contains the data that we have for analysis. In our case, it contains Licence_Data.csv.

2. The src folder is the heart of the project. This folder contains all the modularized code for all the above steps in a modularized manner. It further includes the following.
    a. ML_pipeline
    b. engine.py

The ML_pipeline is a folder that contains all the functions put into different python files, which are appropriately named. These python functions are then called inside the Engine.py file.

3. The output folder contains an excel file for classification metrics scores of each model.

4. The lib folder is a reference folder. It contains the original ipython notebook that we saw in the videos.

5. The requirements.txt file has all the required libraries with respective versions. Kindly install the file by using the command pip **install** -r **requirements**.**txt**

Project Takeaways

1. What is classification?
2. Types of classification
3. Understanding the Business context and objective
4. Data Cleaning
5. What is Data Imbalance?
6. How to deal with imbalanced data?
7. Feature Encoding
8. Importance of splitting data
9. K Nearest Neighbours(KNN) algorithm
10. Naive Bayes algorithm
11. Logistic Regression

12. Decision Tree classifier
13. Confusion matrix
14. Accuracy measurement
15. Precision, Recall, F1 Score
16. Feature Importance
17. Model Predictions
18. Model Evaluation