CS3111 - Introduction to Machine Learning
Lab 01 -Feature Engineering

Dr.R.T.Uthayasanker

February 16, 2024

This is an individual assignment!
Due Date: 29 February 2024 by 11.59 PM

Background

Feature engineering refers to the process of using domain knowledge to select and transform the most relevant variables from raw data when creating a predictive model using machine learning or statistical modeling. The goal of feature engineering and selection is to improve the performance of machine learning (ML) algorithms. Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data pre-processing is to improve the quality of the data. The objective of this task is to explore various data processing and feature engineering techniques to develop the best prediction machine learning model.

About Data

You have been provided with a loan default prediction data set from a finance company in the United States. The data provided contain information on previous loan applicants and whether they 'defaulted' or not. The aim is to identify patterns that indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. Features, also known as variables, include credit scores, the quantity of financing queries, address details such as zip codes and states, and collections, among other factors. Details about the features are given in the data dictionary. The data set is a matrix of around 860,000 observations and 150 variables. Note that there are missing values, outliers, and features of multiple data types.

Data files format

The dataset can be downloaded from the provided Google Drive link. Link to Data Folder. The details of the files in the drive folder;

- **DataDictionary.xlsx :** The collection of names, definitions, attributes, and descriptions of the data objects or items in the data set.
- **train.csv :** The training data with all variables (features) and target column ('loan_status'), used to train the machine learning model (e.g. XGBoost Classifier)
- **valid.csv :** Validation data with all variables (features) and target column ('loan_status'), used to validate and check the performance of the machine learning model trained using training data.
- **X_test.csv :** Test data with all available variables (features) but **without the target column**, used to predict unknown observations using your developed model.

Assignment tasks

- Your task is to apply all the knowledge you have learned about data preprocessing, feature selection & engineering for the given data set in order to develop the best-performing machine learning classification model.

    1. Feature Selection / Removal: Eg. using data cleaning/feature scoring techniques
    2. Feature encoding
    3. Feature crossing
    4. Any other advanced feature engineering techniques
    5. Dimensionality Reduction
    6. Etc...

- Develop an XGBoost Classifer model using the training data and with your feature engineering techniques. Then, use the validation data to check the accuracy of your model and iterate among various feature engineering techniques. Guide to develop XGBoost classifer model

- Finally, give the reduced set of features (**minimum number of features)** enough to predict the target label more accurately.

- Moreover, use Shapely values as an explainable AI technique to interpret your developed model with your features.

Evaluation

- You should transform the features given in the X_test.csv using your developed feature engineering and data preprocessing techniques while training and validation **Don't shuffle test data**

- You should be able to upload a CSV file with your final set of features for the test data (minimum number of feature for higher accuracy).

- The csv file that is submitted should be in the following format and named the files with your index number (e.g. **210001X.csv**) (a submission link is provided).

- The expected csv file should have the following columns in the right order. (We will evaluate your submission using coderunner; hence, strictly follow the formatting guide lines)

    1. Predicted label (0 or 1)
    2. Feature 1
    3. Feature 2
    4. Feature 3
    5. etc.

- In addition, you should submit a **report (5-10 pages)** stating the feature engineering and other processing techniques that you used, your **Python notebook** or link to the notebook, comprising the code for data pre-processing and feature engineering and **SHAP analysis for explainable AI**. (A submission link is provided for the submission of Lab 1 report, and the file should be named **210001X_report.pdf**)

References

1. Feature selection techniques in machine learning
2. Feature engineering techniques
3. Machine Learning Explainability - A kaggle short course