Project Report

Team #12

WANG Zeyu YANG Xirui Wu Tianxiao

April 6, 2025

Contents

1. Problem definition
2. Data Prepare
   - Pearson, Kendall and Spearman correlation coefficient
   - k-nearest neighbors imputer
   - Dimension raising
3. Data Analysis
4. Solution
   - Solution - Task 1
   - Solution - Task 2
   - Solution - Task 3
5. Summary

Problem definition

- **Task 1**: Given the dataset including features (e.g. academic metrics, aptitude scores, and soft skill ratings) of each student, the goal is to predict whether a student will be successfully placed in a job.
- **Task 2**:
- **Task 3**: Given the anonymized numerical features with the transaction amount, the goal is to detect fraudulent transactions.

Pearson, Kendall and Spearman correlation coefficient

- **Pearson correlation coefficient**: Measures linear correlation between two sets of data; Sensitive to the data.
- **Kendall correlation coefficient**: Measures the rank correlation; Robust to outliers; Only effective for monotonic relationships; More robust to ties and suitable for smaller datasets.
- **Spearman correlation coefficient**: Measures the rank correlation; Robust to outliers; Only effective for monotonic relationships; Can be influenced more by large tied ranks and suitable for larger datasets.

k-nearest neighbors imputer

The k-nearest neighbors imputer estimating the missing values using the k nearest neighbors which:

- Works with both numerical and categorical data;
- Don't need any assumption about the data distribution;
- Can be slow for large datasets;
- Sensitive to the neighbors and the distance metric.

Dimension raising

In order to deal with the nonlinearity with the logistic regression, we use the dimension raising where for scalar data x_i and a given degree d, we compute a vector

$$(x_i, x_i^2, \ldots, x_i^d)$$

as the new data.

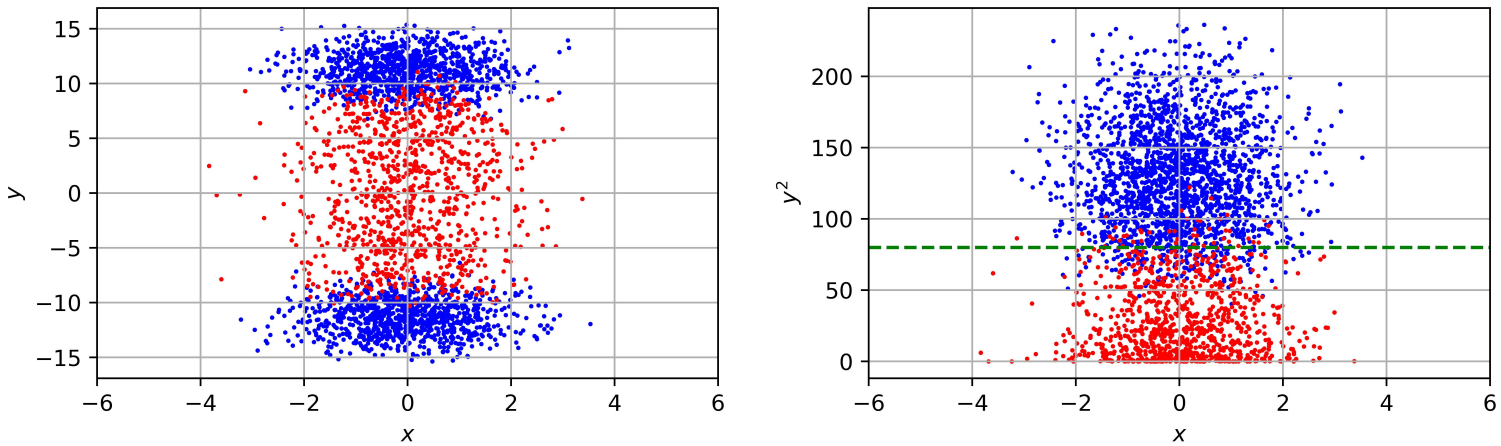

Figure: Example for dimension raising. The left shows that origin data, where the data is not linearly separable, while the right shows the data after dimension raising, where the y_2 is easy to be separated by a line.

Logistic regression

The logistic regression is a widely used linear classification model which:

- Will be less prone to overfitting;
- Can't catch the nonlinearity;
- Might be sensitive to the outliers;
- Can be easily expanded to categorical variables.

Decision tree

The decision tree is a decision support recursive partitioning structure which:

- Don't require normalization or standardization of data;
- Can handle both categorical and numerical data;
- Will be prone to overfitting and sensitive to the noise;
- Not suitable for large dataset.

Multilayer perceptron (MLP)

The multilayer perceptron (MLP) is a basic kind of neural network which:

- Can handle nonlinearity (or approximate any function);
- Can automatically extract features during training;
- Can handle large datasets;
- Will be sensitive to hyperparameters and noise.

Task 1 - Overview

Task 2 - Overview

Task 3 - Overview

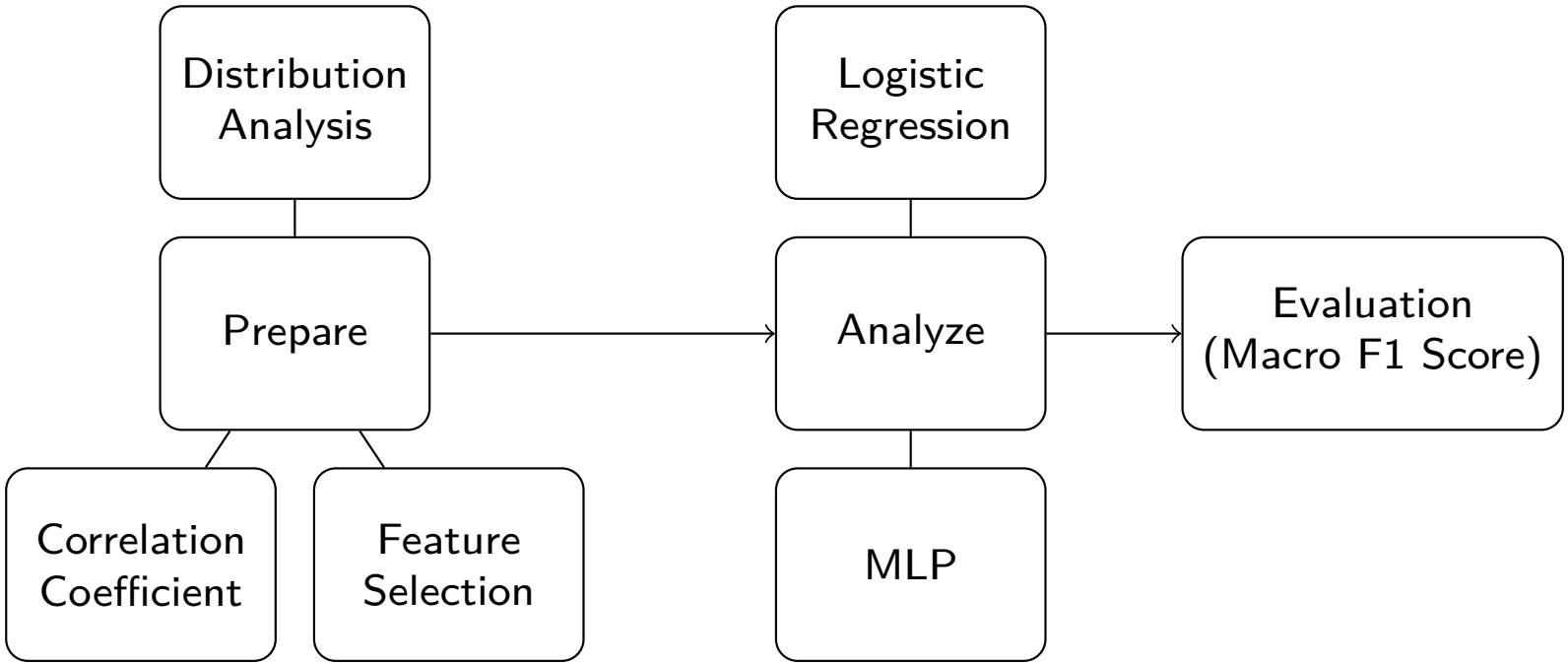

Figure: Overview Flowchart.

Task 3 - Distribution Analysis

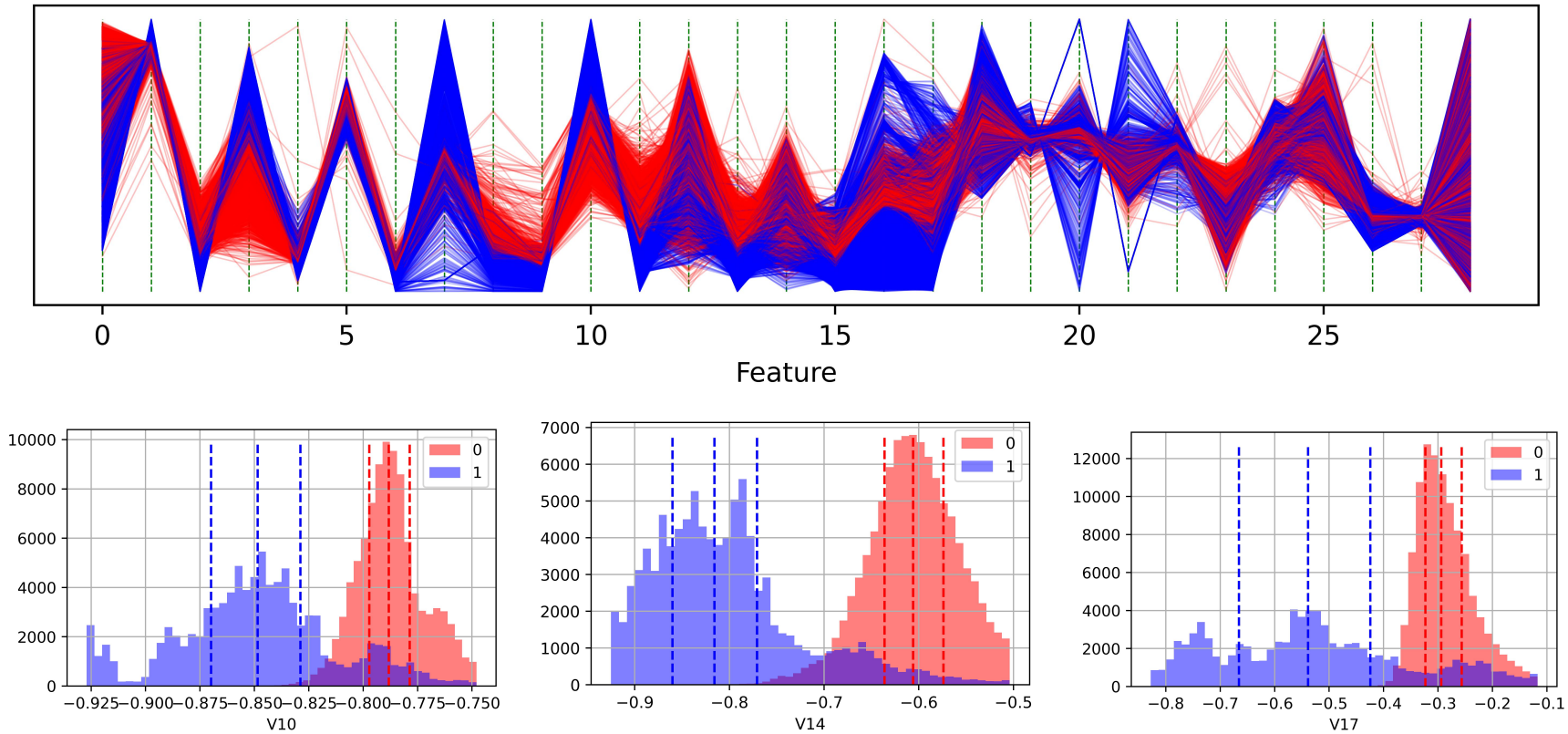

Figure: Data visualization for task 3. The top shows the parallel coordinate plot for each features, and the below shows some features that can be linearly separable.

Task 3 - Correlation Coefficient

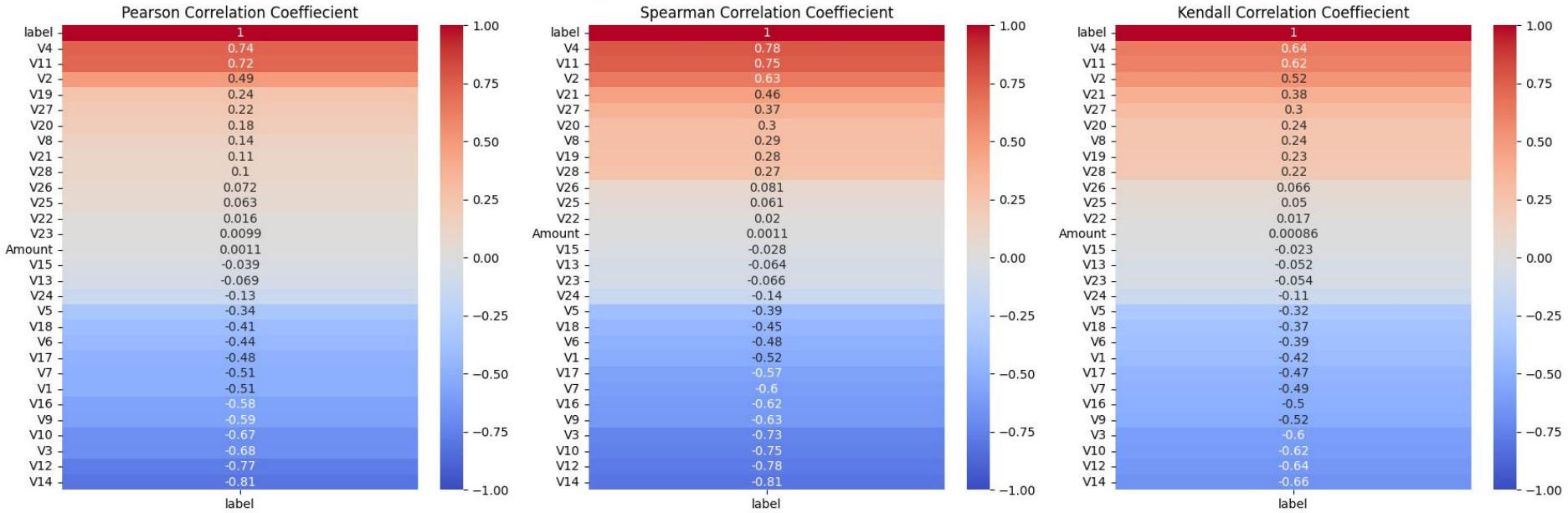

Figure: Correlation coefficient for each features.

We droped the features with all the correlation coefficient less than a threshold (e.g. 0.1).

Task 3 - Preformance - Logistic Regression

Penalty	Dimension Raising (Degree)	Feature Selection	F1 Score	Time/Mem
None	None	False	0.958580	$19.5(s)/866(\text{MB})$
None	2	False	0.961884	$33.5(s)/915(\text{MB})$
None	3	False	0.962137	$49.6(s)/973(\text{MB})$
Lasso	None	False	0.957949	$21.5(s)/854(\text{MB})$
Ridge	None	False	0.955308	$19.8(s)/853(\text{MB})$
None	None	True	0.958186	$16.5(s)/825(\text{MB})$

Table: Preformance for logistic regression. We tested each option with 5 times cross validation (80% for training set, 20% for training set) and 1 times that using all data as training set. The f1 scores showed in the table is the average f1 score for cross validation, the time is the total time for 6 run, and the memory usage is tested using all the data as training set.

Task 3 - Preformance - MLP

Dropout	Hidden Size	Feature Selection	F1 Score	Time/Mem
0.1	32	False	0.952208	$31.9(s)/929(\text{MB})$
0.2	32	False	0.900007	$31.1(s)/916(\text{MB})$
0.1	32	True	0.948050	$31.3(s)/899(\text{MB})$
0.1	16	False	0.950315	$31.7(s)/929(\text{MB})$
0.1	64	True	0.950000	$29.7(s)/902(\text{MB})$

Table: Preformance for MLP. We tested each option with 5 times cross validation (80% for training set, 20% for training set) and 1 times that using all data as training set. The f1 scores showed in the table is the average f1 score for cross validation, the time is the total time for 6 run, and the memory usage is tested using all the data as training set.

Summary

Reference I