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 the new data as

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

Combination of differnet features are also used in dimansion raising, e.g. given scalar x_i and y_i, another kind of new data is computed as

$$(x_i, y_i, x_i y_i).$$

Dimension raising

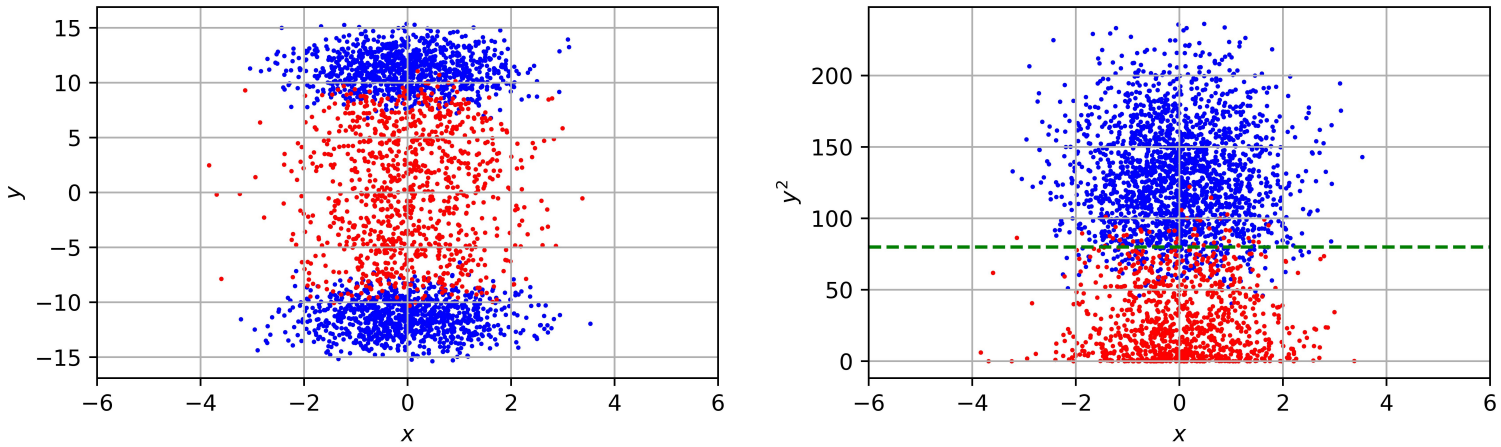

Figure: Example for dimension raising. The left shows that origin data, where the data is not linearly separable, while the right shows the data after dimension raising, where the y_2 is easy to be separated by a line.

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

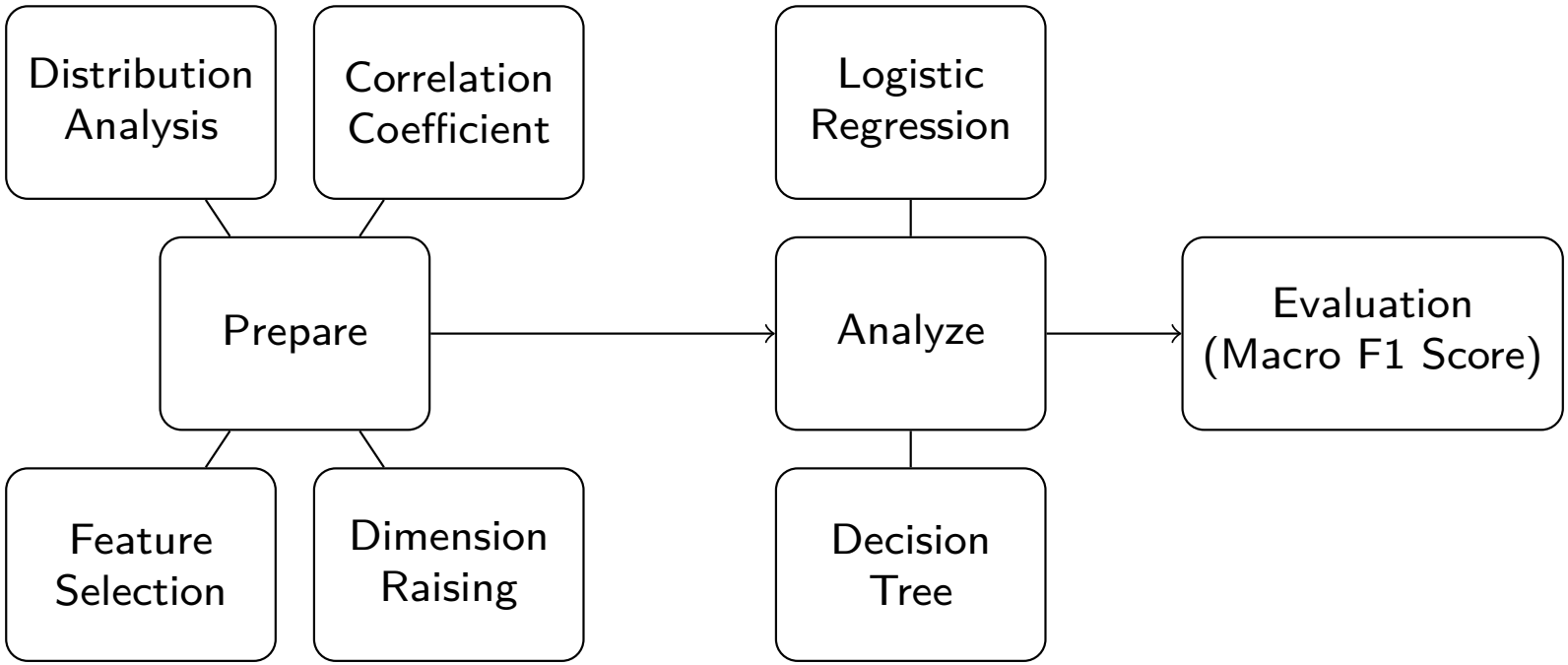

Figure: Overview Flowchart.

Task 1 - Preformance - Logistic Regression

Penalty	Dimension Raising (Degree)	F1 Score	Time/Mem (s/MB)
None	None	0.7909	$1.55/735$
None	2	0.7905	$1.41/730$
None	3	0.7909	$1.28/741$
Lasso	None	0.7912	$1.66/736$
Ridge	None	0.7915	$1.65/735$

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 1 - Preformance - Decision Tree

Max Depth	Number of Threshold	Criterion	F1 Score	Time/Mem (s/MB)
10	64	gini	0.7502	17.5/497
10	128	gini	0.7485	16.4/498
12	64	gini	0.7266	25.2/497
10	64	entropy	0.7428	16.4/501

Table: Preformance for decision tree. 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 2 - Overview

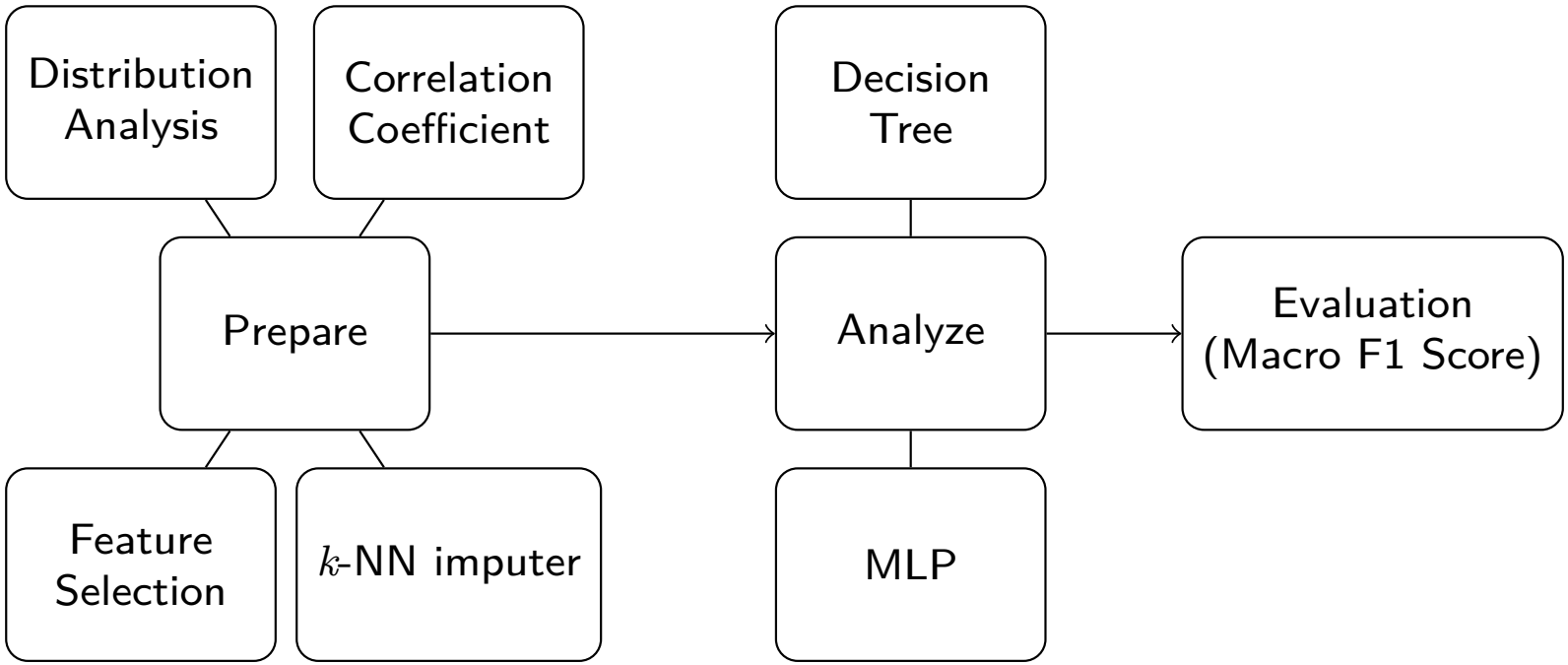

Figure: Overview Flowchart.

Task 2 - Distribution Analysis

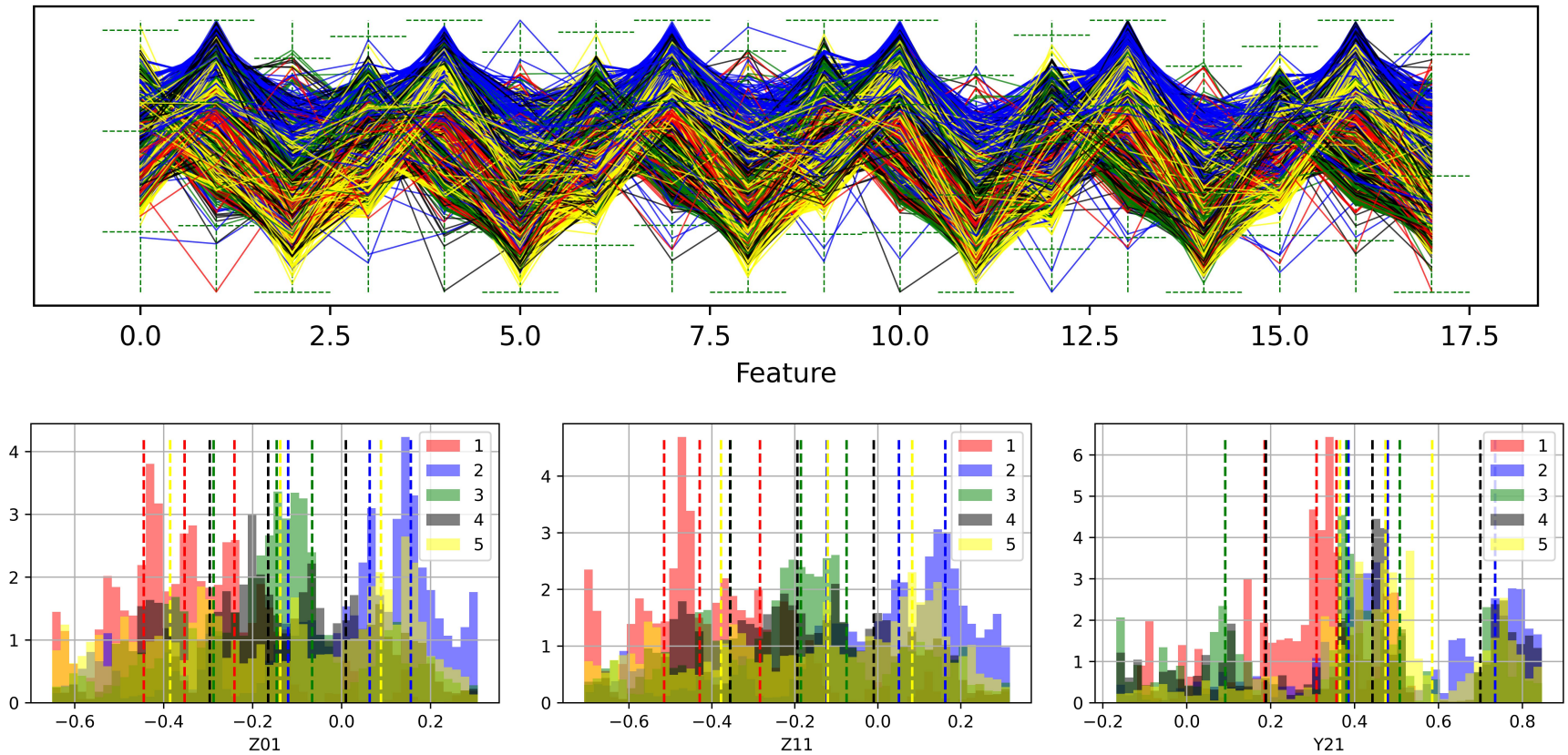

Figure: Data visualization for task 2. The top shows the parallel coordinate plot for each features (we chose 10% of total data), and the below shows even the features can't be easily separated, but the distribution of the labels can be differnet (the top and bottom 5% points are considered as outliers, and will not be shown).

Task 2 - Correlation Coefficient

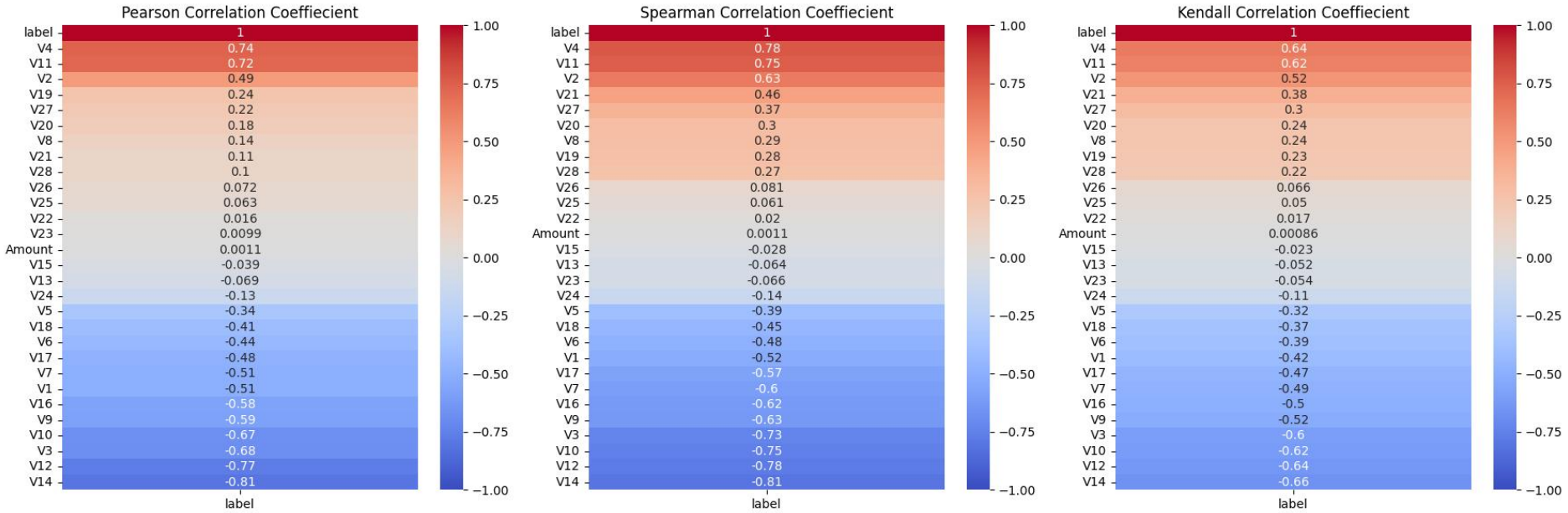

Figure: Correlation coefficient for each features.

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

Task 2 - Preformance - Decision Tree

Max Depth	Num of Threshold	Criterion	Feature Selection	F1 Score	Time/Mem (s/MB)
10	64	gini	False	0.8097	153.6/606
10	128	gini	False	0.8043	168.5/650
12	64	gini	False	0.8511	172.8/601
10	64	entropy	False	0.8184	147.7/608
10	64	gini	True	0.8052	80.7/585

Table: Preformance for decision tree. 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 2 - Preformance - MLP

Dropout	Hidden Size	Feature Selection	Batch Size	F1 Score	Time/Mem (s/MB)
0.3	128	False	512	0.9617	219.4/856
0.3	64	False	512	0.9173	212.9/812
0.3	256	False	512	0.9767	221.7/856
0.3	128	False	1024	0.9629	148.6/857
0.3	128	False	512	0.9644	381.4/859
0.3	128	True	512	0.9474	171.7/848

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.

Task 3 - Overview

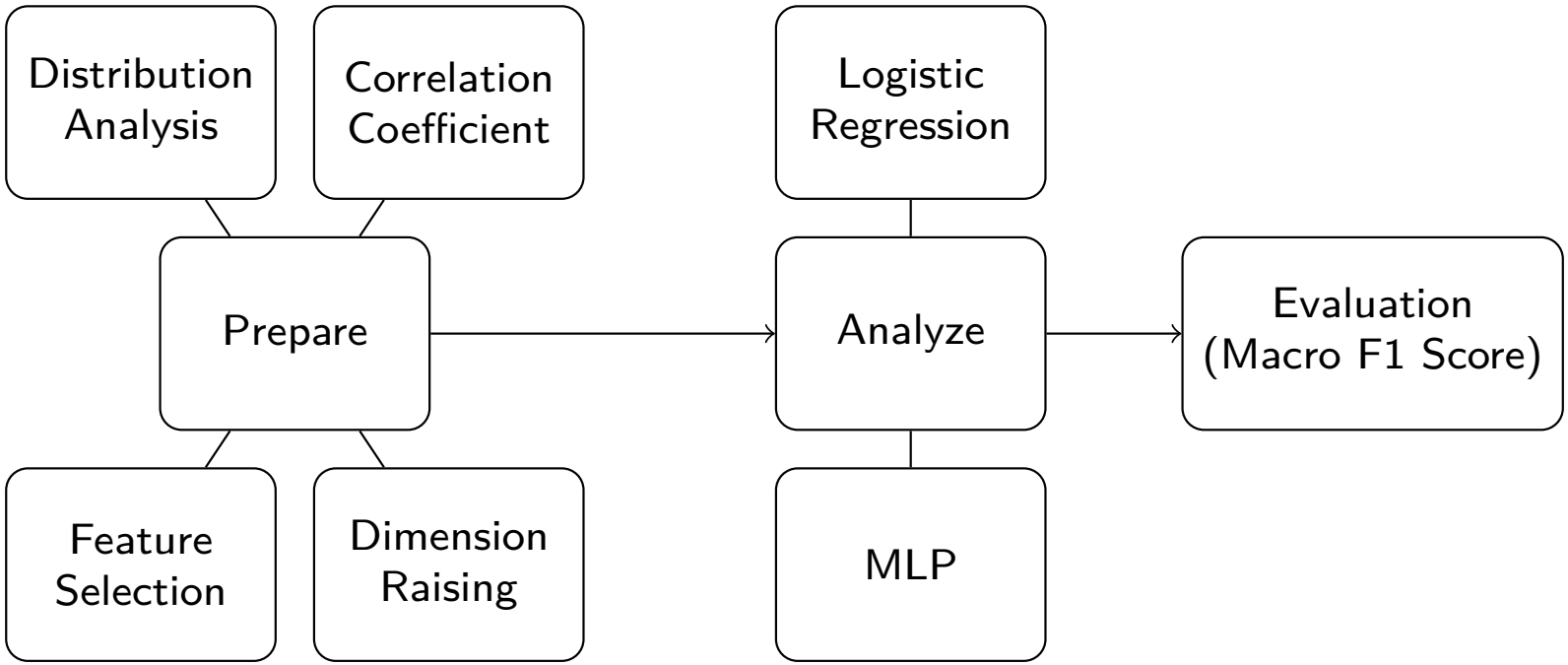

Figure: Overview Flowchart.

Task 3 - Distribution Analysis

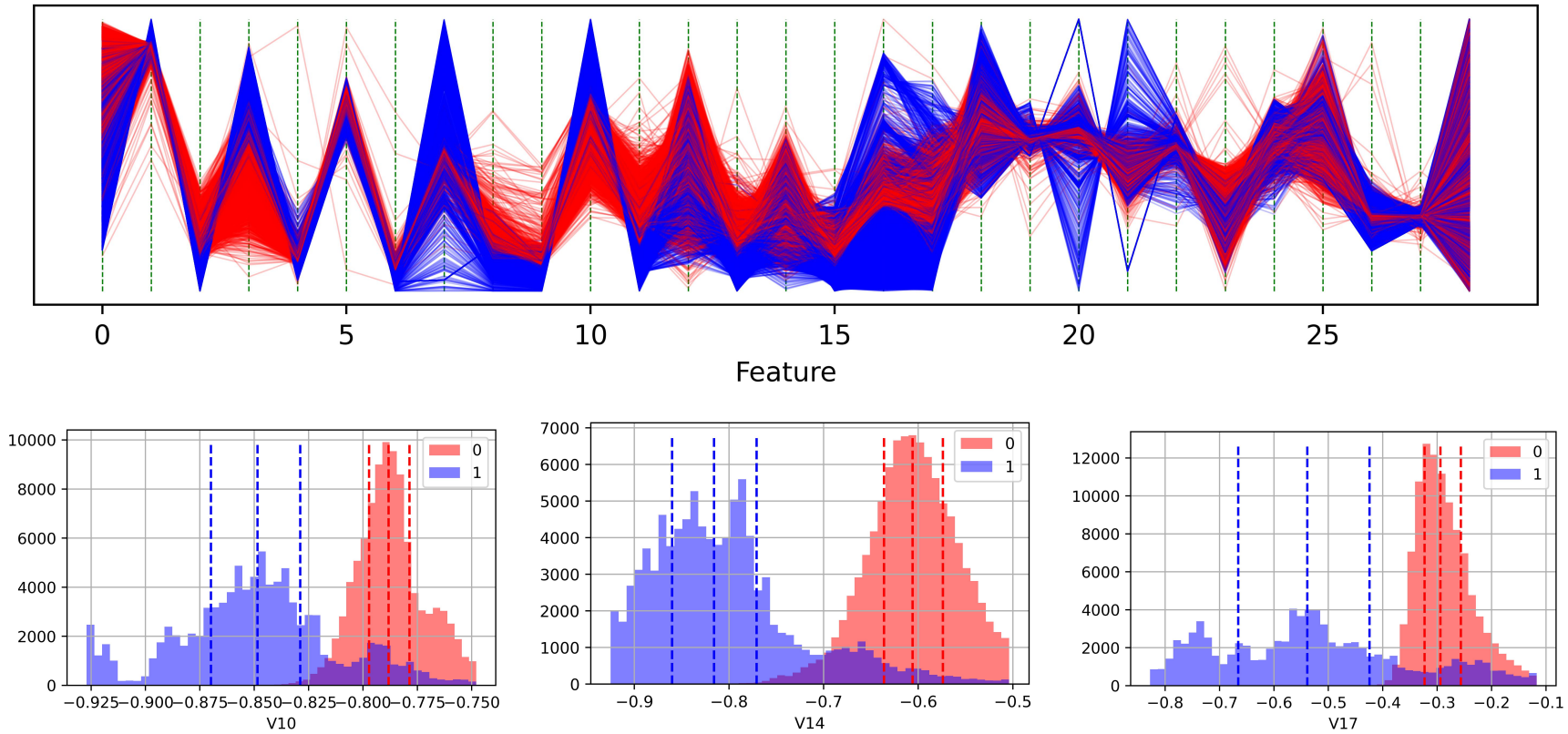

Figure: Data visualization for task 3. The top shows the parallel coordinate plot for each features (we chose 10% of total data), and the below shows some features that can be linearly separable (the top and bottom 5% points are considered as outliers, and will not be shown).

Task 3 - Correlation Coefficient

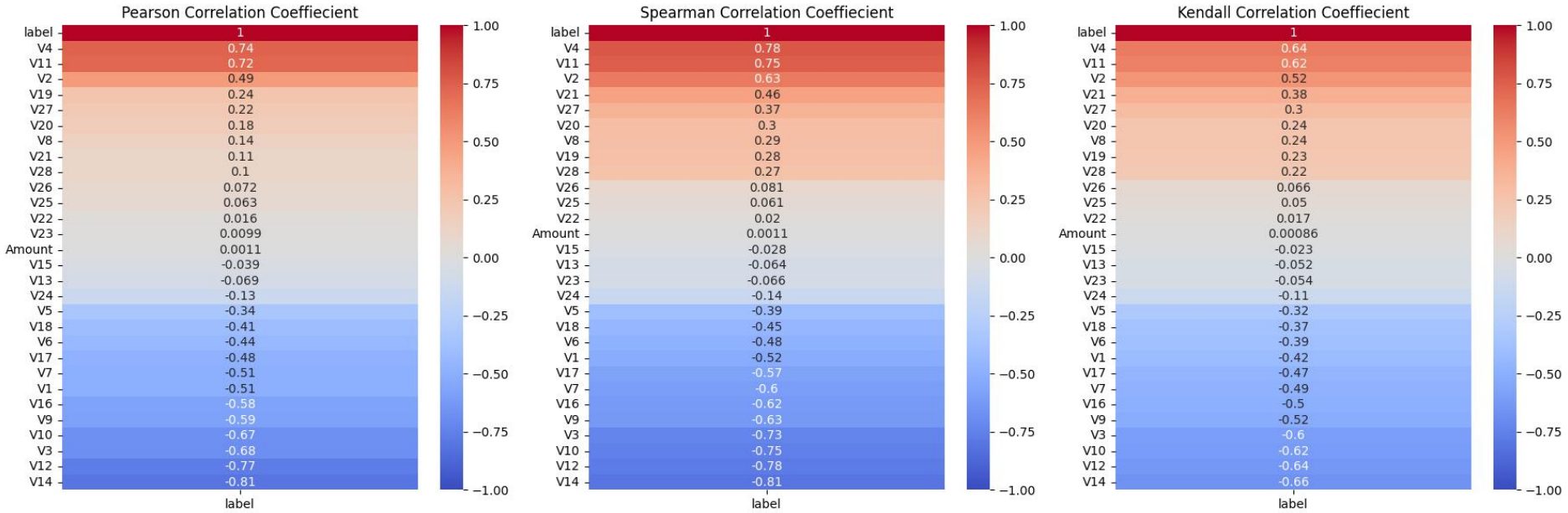

Figure: Correlation coefficient for each features.

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 (s/MB)
None	None	False	0.9586	$19.5/866$
None	2	False	0.9619	$33.5/915$
None	3	False	0.9621	$49.6/973$
Lasso	None	False	0.9579	$21.5/854$
Ridge	None	False	0.9553	$19.8/853$
None	None	True	0.9582	$16.5/825$

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 (s/MB)
0.1	32	False	0.9522	31.9/929
0.2	32	False	0.9000	31.1/916
0.1	32	True	0.9481	31.3/899
0.1	16	False	0.9503	31.7/929
0.1	64	True	0.9500	29.7/902

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

- Well designed feature engineering, including feature selection and dimension raising, can lead to a better model performance;
- The decision tree will be more prone to overfitting the data;
- The MLP can handle both linear and nonlinear data, but it is sensitive to hyperparameters.

Reference I