Project Report

Team #12

WANG Zeyu	YANG Xirui	Wu Tianxiao
24056788G	24135668G	24084591G

April 7, 2025

Contents

Problem definition

- **Task 1**: Train models based on 10 features, including academic metrics, aptitude scores, and soft skill ratings, to predicting whether a student will be successfully placed in a job.
- **Task 2**: Train models based on 18 anonymity features and labeled with 5 classes, then predict the labels based on the training set.
- **Task 3**: Train models based on 28 anonymity numerical with the transaction amount, to indicate whether it is a fraudulent transaction or a legitimate transaction.

Correlation coefficient[1][2]

- **Pearson correlation coefficient**: Measures linear correlation between two sets of data;
- **Kendall correlation coefficient**: Measures the rank correlation by counting the concordant pairs;
- **Spearman correlation coefficient**: Measures the rank correlation based on the Pearson correlation coefficient;

[1] Hervé Abdi (2007). "The Kendall rank correlation coefficient". In: *Encyclopedia of measurement and statistics* 2, pp. 508–510.

[2] Jan Hauke and Tomasz Kossowski (2011). "Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data". In: *Quaestiones geographicae* 30.2, pp. 87–93.

k-nearest neighbors imputer[3][4]

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;
- Is robust in many applications.

[3] Olga Troyanskaya et al. (2001). "Missing value estimation methods for DNA microarrays". In: *Bioinformatics* 17.6, pp. 520–525.

[4] Afaq Juna et al. (2022). "Water quality prediction using KNN imputer and multilayer perceptron". In: *Water* 14.17, p. 2592.

Dimension raising

Given the scalar data x_i and a given degree d, we compute the new data as

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

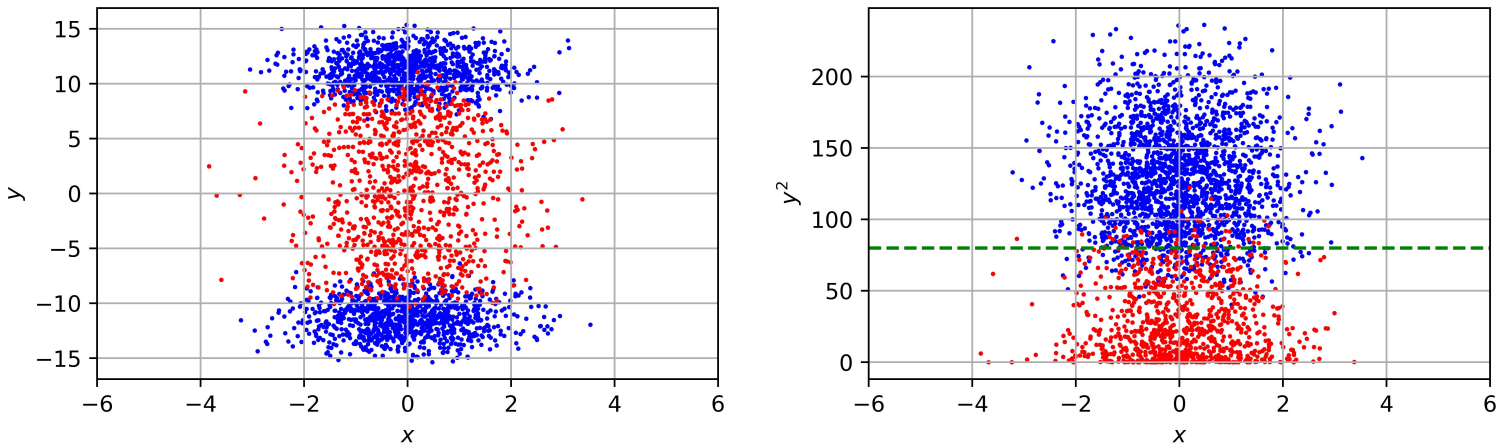

Figure: Example for dimension raising. The left shows that origin data is not linearly separable, while the right shows the data after dimension raising, where the y_2 is easy to be linearly separated.

Dimension raising

For the meaningful features, we can do the dimension raising manually, e.g., in task 1,

- Internships, Projects, Workshops/Certifications $\Longrightarrow$ Practice;
- AptitudeTestScore, SoftSkillsRating $\Longrightarrow$ Potential Ability;
- SSC_Marks, HSC_Marks $\Longrightarrow$ Progress.

Data Analysis

- **The logistic regression**[5][6] is a widely used linear classification model, which gives a probability value ranging between 0 and 1;
- **The decision tree**[7][8] is a supervised learning method used for classification which predict the label with piecewise constant approximation;
- **The multilayer perceptron (MLP)**[9][10] is a basic kind of neural network which learns a function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ to approximate the input and output, with the ability of hierarchical feature extraction.

---

[5] Strother H Walker and David B Duncan (1967). "Estimation of the probability of an event as a function of several independent variables". In: *Biometrika* 54.1-2, pp. 167–179.

[6] Maja Pohar, Mateja Blas, and Sandra Turk (2004). "Comparison of logistic regression and linear discriminant analysis: a simulation study". In: *Metodoloski zvezki* 1.1, p. 143.

[7] Paul E Utgoff (1989). "Incremental induction of decision trees". In: *Machine learning* 4, pp. 161–186.

[8] Sotiris B Kotsiantis (2013). "Decision trees: a recent overview". In: *Artificial Intelligence Review* 39, pp. 261–283.

[9] Frank Rosenblatt (1958). "The perceptron: a probabilistic model for information storage and organization in the brain.". In: *Psychological review* 65.6, p. 386.

[10] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams (1986). "Learning representations by back-propagating errors". In: *nature* 323.6088, pp. 533–536.

Solution Overview

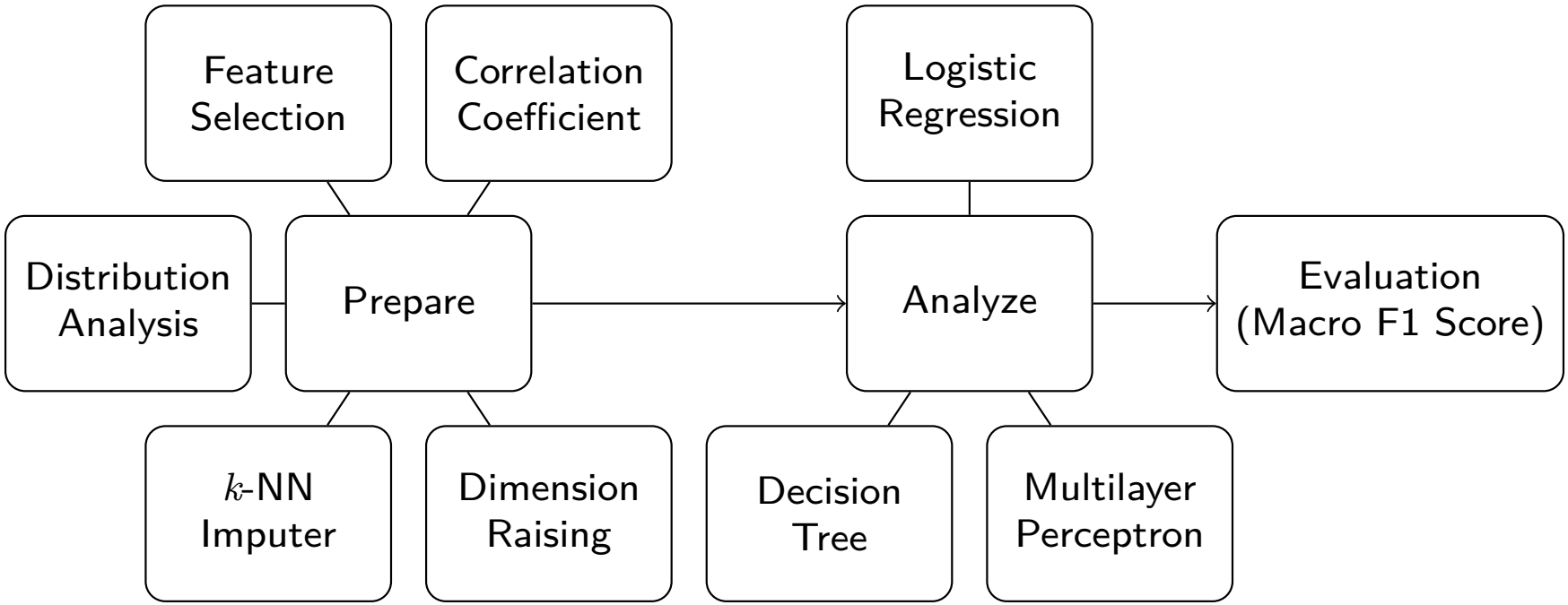

Figure: Overview flowchart.

Discovery & insights

- The features engineering can help imrpove the model preformance, with the a higher time and memory cost;
- the effectiveness of feature engineering varies by task, which needs adjustments based on the data;
- The logistic regression preforms a better reliability and efficiency especially on linearly separable data, while MLP will be better in the complex multi-class problem;
- Regularization will reduce the influence of noise and outliers, which prevent the model from overfitting and ensuring the generalization;
- The parallel computing will significantly accelerate the training, but must be balanced against memory cost and convergence stability.

Future Work

- Employ automated feature engineering tools insteal of manual design, which can improve feature selection and generation;
- Ensemble methods (e.g., random forests or gradient boosting trees) may help deal with the overfitting of decision tree;
- Adapt the models to the dynamic environment and input, especially can help detected the fraud in time;
- Use the distributed computing frameworks to speed up the model and also train with the large-scale datasets.

Reference I

- Abdi, Hervé (2007). "The Kendall rank correlation coefficient". In: *Encyclopedia of measurement and statistics* 2, pp. 508–510.
- Hauke, Jan and Tomasz Kossowski (2011). "Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data". In: *Quaestiones geographicae* 30.2, pp. 87–93.
- Juna, Afaq et al. (2022). "Water quality prediction using KNN imputer and multilayer perceptron". In: *Water* 14.17, p. 2592.
- Kotsiantis, Sotiris B (2013). "Decision trees: a recent overview". In: *Artificial Intelligence Review* 39, pp. 261–283.
- Pohar, Maja, Mateja Blas, and Sandra Turk (2004). "Comparison of logistic regression and linear discriminant analysis: a simulation study". In: *Metodoloski zvezki* 1.1, p. 143.
- Rosenblatt, Frank (1958). "The perceptron: a probabilistic model for information storage and organization in the brain.". In: *Psychological review* 65.6, p. 386.

Reference II

- Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). "Learning representations by back-propagating errors". In: *nature* 323.6088, pp. 533–536.
- Troyanskaya, Olga et al. (2001). "Missing value estimation methods for DNA microarrays". In: *Bioinformatics* 17.6, pp. 520–525.
- Utgoff, Paul E (1989). "Incremental induction of decision trees". In: *Machine learning* 4, pp. 161–186.
- Walker, Strother H and David B Duncan (1967). "Estimation of the probability of an event as a function of several independent variables". In: *Biometrika* 54.1-2, pp. 167–179.