Data 622 Machine Learning and Big Data_HW1

By: Jiaxin Zheng

Exploratory Data Analysis and Algorithm Selection

The Bank Marketing dataset collect information from a marketing campaign of a Portuguese bank. The goal is to predict whether a client will subscribe to a term deposit. This is a binary classification analysis with both categorical variables (e.g., job, education, marital, housing loan) and numeric variables (e.g., age, balance, campaign, pdays, previous).

The analysis revealed several important patterns. The clients with tertiary education and students tend to have higher subscription rates compared to those with primary and secondary education. Marital status also matters; single clients are slightly more likely to subscribe than married ones. Employment type plays a strong role, with students and retired clients responding more positively than blue-collar workers. Financial factors such as balance show a skewed distribution, with a small number of clients having extremely high balances, while most clients have relatively low balances. Age is another differentiator, where younger age (18-25) and seniors (76+) show higher subscribing compared to middle-aged groups. Overall, the dataset shows clear imbalance, most clients did not subscribe to a term deposit.

Algorithm Selection

For this dataset, I recommend Random Forest, Logistic Regression, and XGBoost.

- Random Forest:
    - Pro: Compared to single decision tree, random forest is an ensemble method that can reduce overfitting. It can naturally handle mix dataset, and robust to outliers and noise.
    - Cons: More computationally intensive.
- Logistic Regression:
    - Pro: Simple, efficient, and interpretable. Coefficient can be converted into odds ratios, making it easier to understand which variables are most strongly influence subscription.
    - Cons: May not capture nonlinear relationships.
- XGBoost:
    - Pro: Can handle class imbalances using build-in function.
    - Cons: More complex, and training can be slow.

Preprocessing Requirement:

- **Data Cleaning:** The data does not have missing values but several “unknown” categorical variables. Rather than treating these “unknown” categorical variables as missing values, I interpret them as valid categories that may carry predictive meaning. I would like to apply transformation to reduce the influence of extreme values without discarding data.
- **Dimensionality Reduction:** Check correlations among numeric features (e.g., campaign, pdays, previous). I would consider combining highly related features.
- **Feature Engineering:** Create a new feature make it easier to interpret, such as group age into bins (18–25, 26–35, etc.)
- **Sampling:** Subsampling may be unnecessary.
- **Data Transformation:** For logistic regression, numeric features may be normalized or standardized. Tree-based models do not require scaling.