Programming Assignment 2 final report

Xuanzhen Lao

Keheng Cai

Washington University in St. Louis

CSE 514A Data Mining

Professor Cynthia Ma

April 25, 2024

GitHub: https://github.com/kingfroglao/CSE514_PA2_.git

1. Introduction

Banks play an important role in economic growth, providing a range of financial services to customers. Among many service strategies, telemarketing is often used as an important means for banks to promote new products and services. However, this strategy is often time-intensive and requires employees to develop strong connections with customers to achieve the company's business goals. With the development of marketing technology, the large amount of customer data accumulated throughout history has become an asset for formulating precise marketing strategies. This project aims to use this accumulated customer information to reveal customer behavior patterns through appropriate data analysis and predict which customers are more likely to respond positively to the bank's products or services. We will use machine learning technology to build models to predict customer responses to telemarketing activities, thereby improving the efficiency and effectiveness of marketing activities.

2. Materials

In this project, we use data mining methods to predict the success of telemarketing calls selling bank fixed deposits. Our dataset involves the marketing activities of Portuguese banking institutions between 2008 and 2010, involving 41,188 instances and 17 categories of variables (16 features and 1 target), which contains data such as customer personal information, contact information, education level, etc. The dataset also contains data on when and how many times salespeople contacted customers, as well as the results of previous marketing campaigns.

Table 1. Variables table of dataset [1]

Variable Name	Role	Type	Demographic	Missing Values
age	Feature	Integer	Age	no
job	Feature	Categorical	Occupation	no
marital	Feature	Categorical	Marital Status	no
education	Feature	Categorical	Education Level	no
default	Feature	Binary		no
balance	Feature	Integer		no
housing	Feature	Binary		no
loan	Feature	Binary		no
contact	Feature	Categorical		yes
day_of_week	Feature	Date		no
month	Feature	Date		no
duration	Feature	Integer		no
campaign	Feature	Integer		no
pdays	Feature	Integer		yes
previous	Feature	Integer		no
poutcome	Feature	Categorical		yes
y	Target	Binary		No

3. Method

To study the impact of marital status on classification, we divided the dataset into three independent binary classification tasks and ensured a balanced distribution of samples from each task: single, married, and divorced. In each classification task, we set aside 10% of relevant samples for final validation of the model.

In this project, we selected four classification models: Naive Bayes classifier, decision tree, artificial neural network, and random forest. For each model, we selected five hyperparameters and performed 5-fold cross-validation on the training data set to tune these hyperparameters to achieve better performance improvements in accuracy. We selected the combination of hyperparameters that performed best in cross-validation and trained the model on this set of parameters on the entire training dataset. Then, we use these trained models to make predictions on the final validation set. Furthermore, we test the performance of the existing model by performing PCA dimensionality reduction to evaluate the results with and without dimensionality reduction. This step is to understand the specific impact of dimensionality reduction on model performance. We compare the accuracy with and without dimensionality reduction and consider other performance metrics when necessary.

4. Models Fitting

For each model, we selected 5 hyperparameters and ran 5-fold cross validation to test the hyperparameter values.

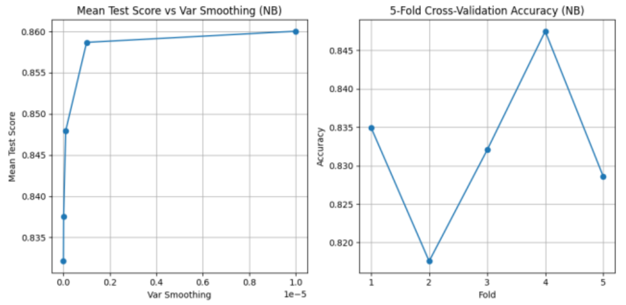
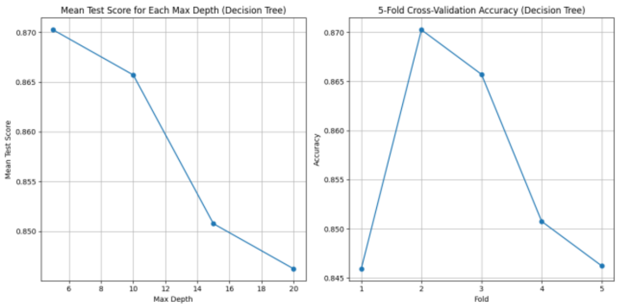
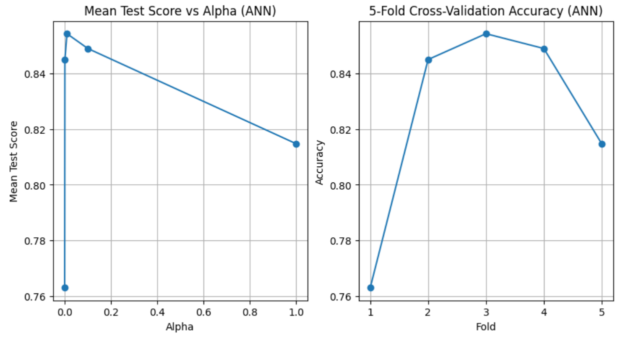
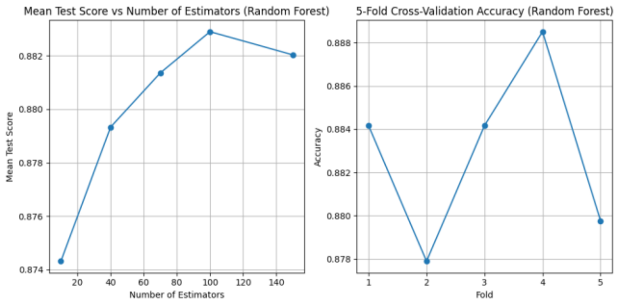
- Naïve Bayes Classifier: 'var_smoothing': [1e-9, 1e-8, 1e-7, 1e-6, 1e-5]
- Decision Tree: 'max_depth': [None, 5, 10, 15, 20]
- Artificial Neural Network: alpha': [0.0001, 0.001, 0.01, 0.1, 1]
- Random Forest: 'n_estimators': [10, 40, 70, 100, 150]

We will run the function calls of the model using selected hyperparameter values and find the highest accuracy. Then, we will perform 5-fold cross-validation on the training data set and visualize the cross-validation results.

4.1 Single Classification

4.1.1 Visualization

Naïve Bayes Classifier best hyperparameter: 'var_smoothing': 1e-05	Decision Tree best hyperparameter: 'max_depth': 5

<table>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>ANN best hyperparameter: 'alpha': 0.01</td>
<td>Random Forest best hyperparameter: 'n_estimators': 100</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
</tr>
</table>

4.1.2 Prediction results on the final validation set

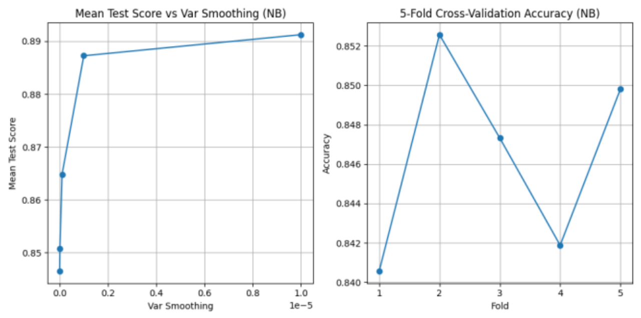
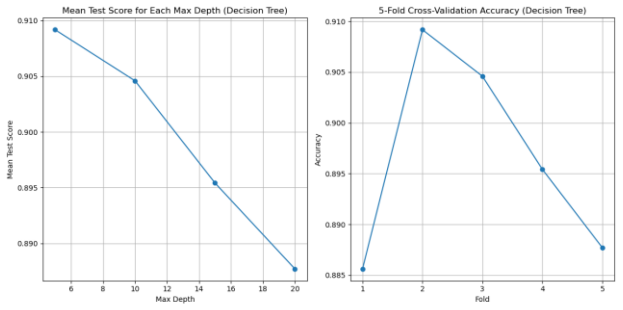
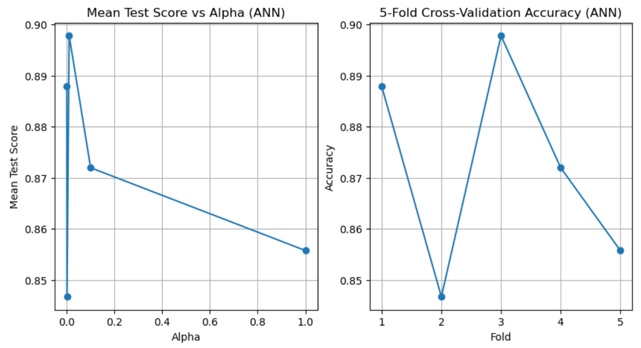
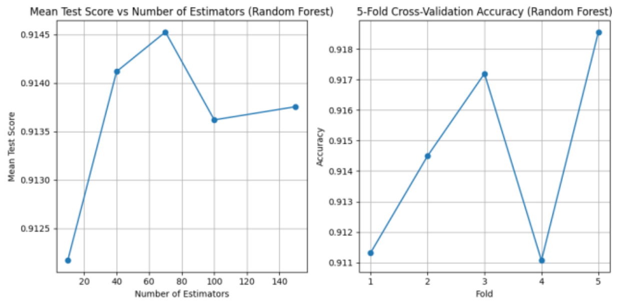
Table 2: Prediction results for Single Classification

Model	**Model performance**	**Training runtime (sec)**	**Prediction performance (accuracy)**	**Testing runtime (sec)**
Naïve Bayes	0.85069	0.00265	0.76562	0.01107
Random Forest	0.86371	0.96967	0.83593	0.30111
ANN	0.84982	1.45259	0.7969	0.42175
Decision Tree	0.86545	0.019995	0.8125	0.00701

4.2 Married Classification

4.2.1 Visualization

Naïve Bayes Classifier best hyperparameter: 'var_smoothing': 1e-05	Decision Tree best hyperparameter: 'max_depth': 5

	
ANN best hyperparameter: 'alpha': 0.01	Random Forest best hyperparameter: 'n_estimators': 70
	

4.2.2 Prediction results on the final validation set

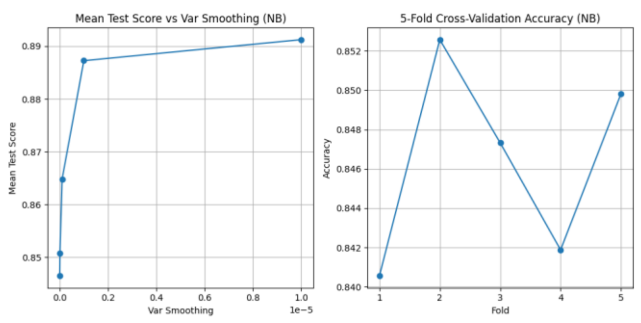
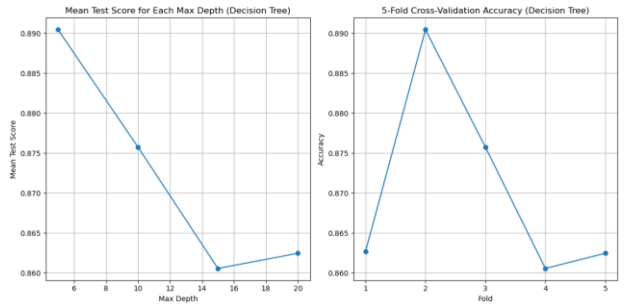
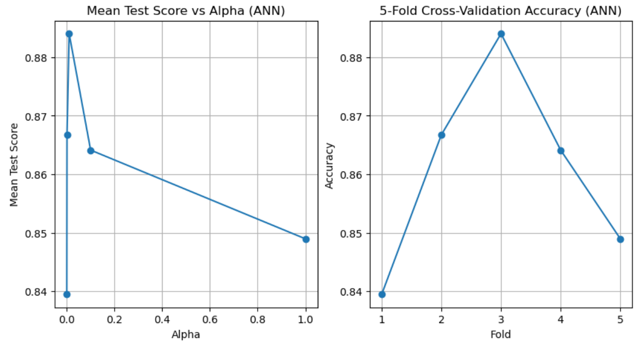
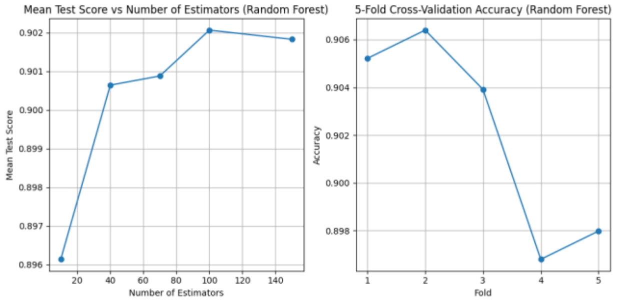
Table 3: Prediction results for Married Classification

Model	Model performance	Training runtime (sec)	Prediction performance (accuracy)	Testing runtime (sec)
Naïve Bayes	0.8902	0.005195	0.92308	0.002124
Random Forest	0.9184	1.4572	0.9414	0.170841
ANN	0.8906	1.412	0.90842	0.274999
Decision Tree	0.9171	0.03799	0.92308	0.007998

4.3 Divorced Classification

4.3.1 Visualization:

Naïve Bayes Classifier best hyperparameter: 'var_smoothing': 1e-05	Decision Tree best hyperparameter: 'max_depth': 5

	
ANN best hyperparameter: 'alpha': 0.0001	Random Forest best hyperparameter: 'n_estimators': 100
	

4.3.2 Prediction results on the final validation set

Table 4: Prediction results for Divorced Classification

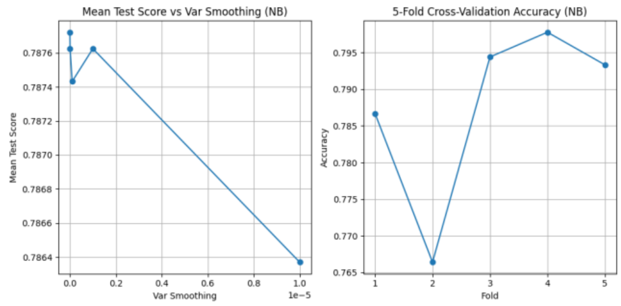
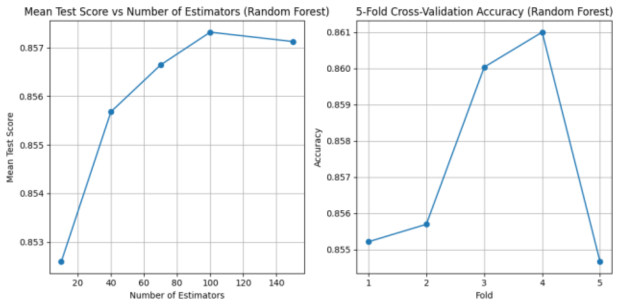
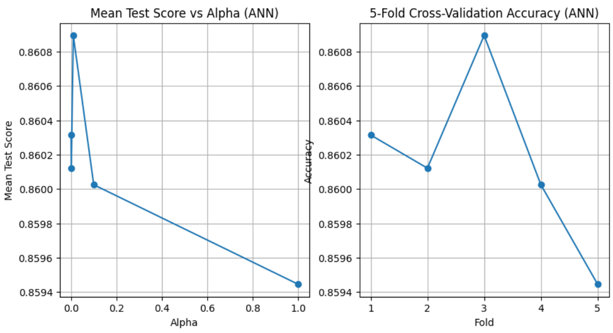
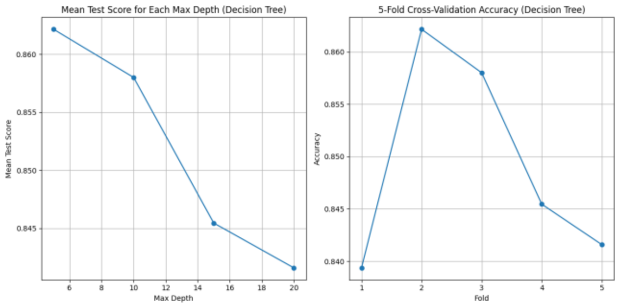
Model	Model performance	Training runtime (sec)	Prediction performance (accuracy)	Testing runtime (sec)
Naïve Bayes	0.8763	0.00202	0.86792	0.002124
Random Forest	0.898	0.41688	0.86792	0.10547
ANN	0.85928	0.390	0.90566	0.1039
Decision Tree	0.8849	0.007998	0.81132	0.003993

5. Dimension Reduction

For this project, we chose Principal Component Analysis (PCA) as the dimension reduction method. PCA is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space while preserving the most important information. It begins by standardizing the data to have zero mean and unit variance, followed by computing the covariance matrix of the standardized data. The eigenvectors and eigenvalues of this covariance matrix are then calculated, where the eigenvectors represent the principal components, and the eigenvalues indicate the amount of variance explained by each component. The principal components are ordered by their corresponding eigenvalues, and a subset of these components is selected to retain the desired amount of variance (half of the original features). Finally, the data is transformed into the new lower-dimensional space using the selected principal components.

5.1 Single Classification

5.1.1 Visualization

<table>
<tr><td>Naïve Bayes Classifier best hyperparameter: 'var_smoothing': 1e-09</td><td>Random Forest best hyperparameter: 'n_estimators': 100</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>ANN best hyperparameter: 'alpha': 0.0001</td><td>Decision Tree best hyperparameter: 'max_depth': 5</td></tr>
<tr><td>
</td><td>
</td></tr>
</table>

5.1.2 Prediction results on the final validation set

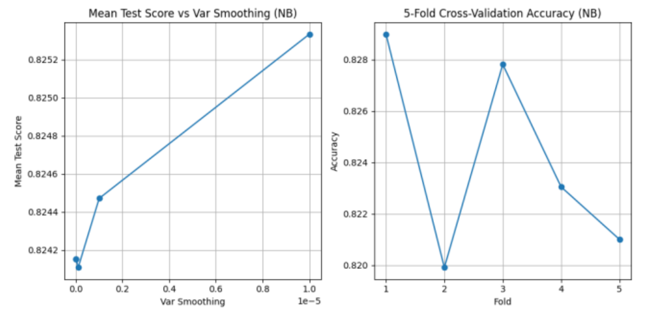
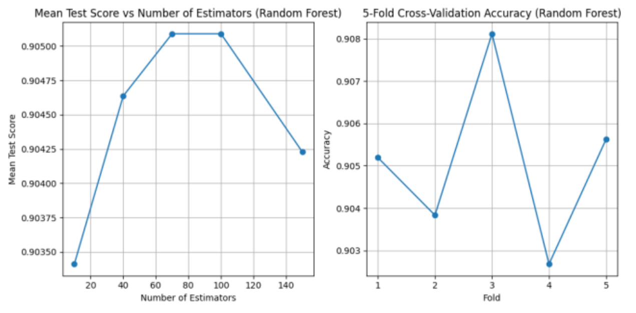
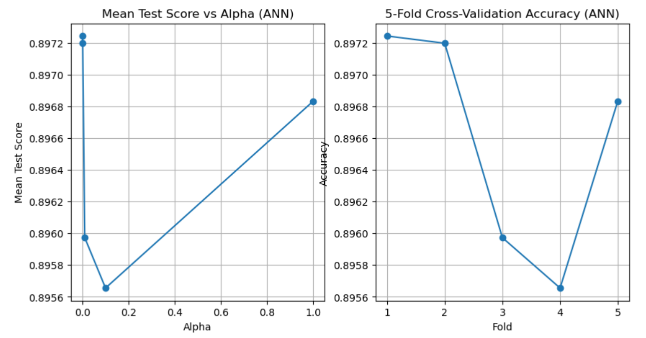
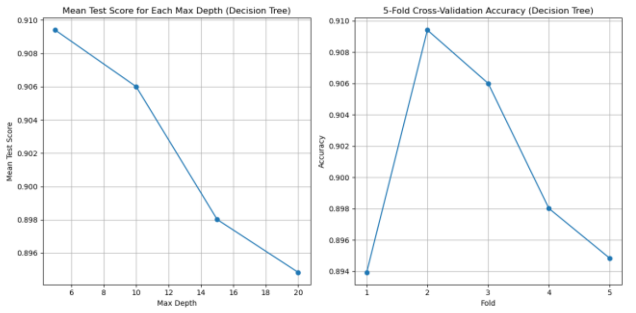
Table 5: Prediction results for Single Classification (dimension reduction)

Model	**Model performance**	**Training runtime (sec)**	**Prediction performance (accuracy)**	**Testing runtime (sec)**
Naïve Bayes	0.7986	0.003628	0.765625	0.002320
Random Forest	0.86111	0.4638	0.8125	0.0984
ANN	0.858506	15.8406	0.8125	3.56093
Decision Tree	0.852430	0.00618	0.8125	0.0050

5.2 Married Classification

5.2.1 Visualization

Naïve Bayes Classifier best hyperparameter: 'var_smoothing': 1e-05	Random Forest best hyperparameter: 'n_estimators': 100

	
ANN best hyperparameter: 'alpha': 1	Decision Tree best hyperparameter: 'max_depth': 5
	

5.2.2 Prediction results on the final validation set

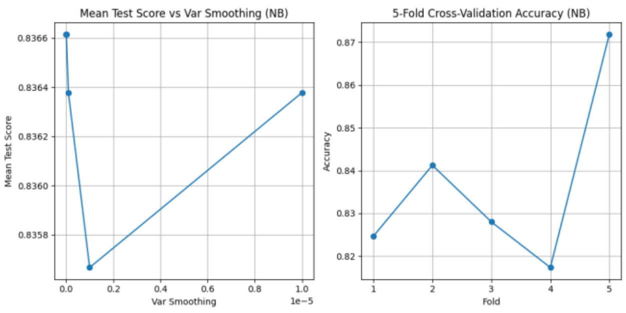
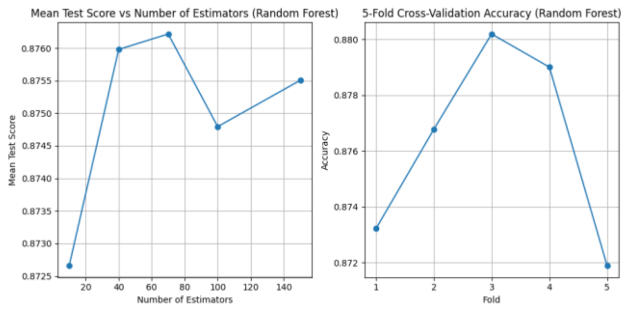
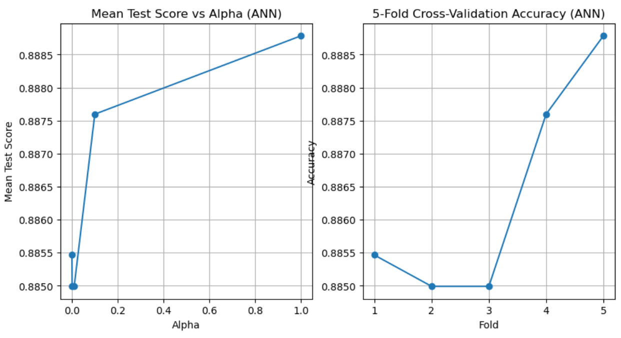
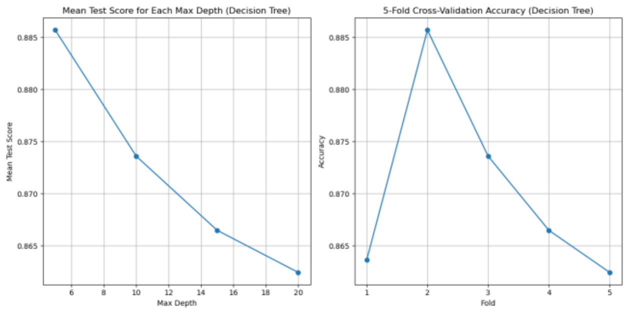
Table 6: Prediction results for Married Classification (dimension reduction)

Model	Model performance	Training runtime (sec)	Prediction performance (accuracy)	Testing runtime (sec)
Naïve Bayes	0.82489	0.0035	0.80952	0.00202
Random Forest	0.90326	1.0718	0.91941	0.1752
ANN	0.90489	1.41716	0.91941	0.6792
Decision Tree	0.90734	0.01099	0.93040	0.0040

5.3 Divorced Classification

5.3.1 Visualization

Naïve Bayes Classifier best hyperparameter: 'var_smoothing': 1e-05	Random Forest best hyperparameter: 'n_estimators': 70

	
ANN best hyperparameter: 'alpha': 1	Decision Tree best hyperparameter: 'max_depth': 5
	

5.3.2 Prediction results on the final validation set

Table 7: Prediction results for Divorced Classification (dimension reduction)

Model	**Model performance**	**Training runtime (sec)**	**Prediction performance (accuracy)**	**Testing runtime (sec)**
Naïve Bayes	0.82089	0.002852	0.8868	0.00202
Random Forest	0.86354	0.15861	0.81132	0.0609
ANN	0.87420	0.533322	0.8868	0.02073
Decision Tree	0.87420	0.00204	0.868	0.0009996

6. Results

There were significant differences in predictability across the three binary classification tasks (single, married, divorced). In a single classification task, the accuracy of different models in predicting customer responses to telemarketing ranged from approximately 76.56% to 81.25% (average 80.27%). In contrast, the married classification task consistently showed higher prediction performance, with accuracy ranging from 90.82% to 94.14% (average 92.39%). The accuracy for the divorced classification task ranged from approximately 81.13% to 88.68% (average 86.23%). From this information, we know that the married classification shows more predictable results. The single classification task had a moderate degree of predictability, while the divorced classification was intermediate between the single and married tasks.

Classification models have different performances in analyzing different tasks. In the single classification task, the model performance of the four models is very close, the range is from 84.98% to 86.54%, and the decision tree has the highest model performance of 86.54%. It shows that the prediction capabilities of different models are relatively uniform in this task. However, random Forest (83.59%) showed the highest accuracy. In the married classification task, the Random Forest has the highest model performance of 91.84% and an accuracy of 94.14%, becoming the best-performing model. In the model performance of the Divorced classification task, the random forest model performed best (89.98%), and the range of other models was 85.92%-88.49%. However, the ANN (90.56%) shows higher accuracy, and the accuracy of other models ranges from 81.13% to 86.97%. Overall, Random Forests generally performs well in all classification tasks. Therefore, we cannot simply say that all models perform equally "good" on all tasks. What is clear, however, is that the data from all the models met the expected metrics.

After comparing before and after dimension reduction, we can see that in a single classification task, after dimension reduction, the model performance of the Naïve Bayes dropped significantly, but the prediction performance did not change. The model performance of the other three models did not change much and the accuracy only fluctuated by about ±2%. In married classification, the performance and accuracy of the Naïve Bayes model also dropped significantly after dimension reduction. However, the performance of other models did not change much, with a fluctuation of about ±1%. Also, the prediction performance of Decision Trees increased by 1% while the prediction performance of Random Forests decreased by approximately 2%. In divorced classification, the model performance and accuracy of both the Naïve Bayes and Random Forest dropped significantly. The ANN model's performance increases but the accuracy decreases. Similarly, the model performance of Random Forest did not change much, while the accuracy became higher.

Overall, the Naïve Bayes seems to be more sensitive to dimension reduction, especially in model performance, which has a negative impact. Random Forests and Decision Trees have varying effects on different tasks. ANN shows potential for model performance improvements that dimension reduction may have a positive impact.

Given that our motivation is to optimize telemarketing strategies for banking products, we recommend focusing on married classification and using random forest models for in-depth analysis. By fine-tuning model parameters, feature selection, and cross-validation, we can improve the model's prediction accuracy and stability. At the same time, the prediction results of the model are used to identify the customer groups most likely to be interested in the product, thereby achieving precise positioning and personalized marketing of target customers. Through data analysis and planning implementation, we expect to significantly improve marketing efficiency, reduce costs, and increase customer satisfaction.

Reference

[1] Moro,S., Rita,P., and Cortez,P.. (2012). Bank Marketing. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.