MANU Policy Report

Implementer: Dat Tien Nguyen

Part I – Prediction Task

1. Context Analysis

As can be depicted in task description, the data consists of 86 customer features and the objective is to use the first 86 features to predict the final one to examine whether the tested customers will have the high probability to possess the MANU policy so that efficient sale campaigns could be implemented on them.

The given training dataset is completed without any missing data and all features are recorded in integer data type. The target contains 5474 samples of class 0 and 348 samples of class 1 representing customers not taking and taking the policy respectively. For this reason, the training data is suffered from highly unbalanced context for a binary classification problem.

```
df_train['86'].value_counts()
✓ 0.0s

0    5474
1     348
Name: 86, dtype: int64
```

Figure 1: Value Counts for Majority and Minority Class.

2. Training Strategies with Different Models and Data Processing Methods

To find the appropriate model for this context, different classifiers is projected to be applied including normal method (Decision Tree, Bagging), ensemble (Gradient Boosting, Ada Boost, Random Forest) and also balanced ones (Balanced Random Forest, Easy Ensemble and RUS Boost). If using the normal and ensemble methods to train directly on this unbalanced database, the model would still produce high accuracy since it can still keep giving 0 prediction for this majority class. However, this would bias this class and ignore the sensitivity on the true or class 1 which is significantly important in this customer context.

As a result, different data preprocessing methods are also applied to handle this unbalanced data. This includes under sampling, over sampling, combined over & under sampling, PCA dimensional reduction and customized ensemble major voting.

The given 'data_train.txt' will firstly be utilized to test the model's reliability with those models before being fully used for the selected one to predict the target feature in 'data_test.txt'. By doing this, the generalization of the model for the data's context could be enhanced to potentially produce higher predicting results for the test set. To test the model's reliability, f1

score is applied instead of accuracy score since this evaluates strictly the model's performance on the minority targets.

3. Result Analysis

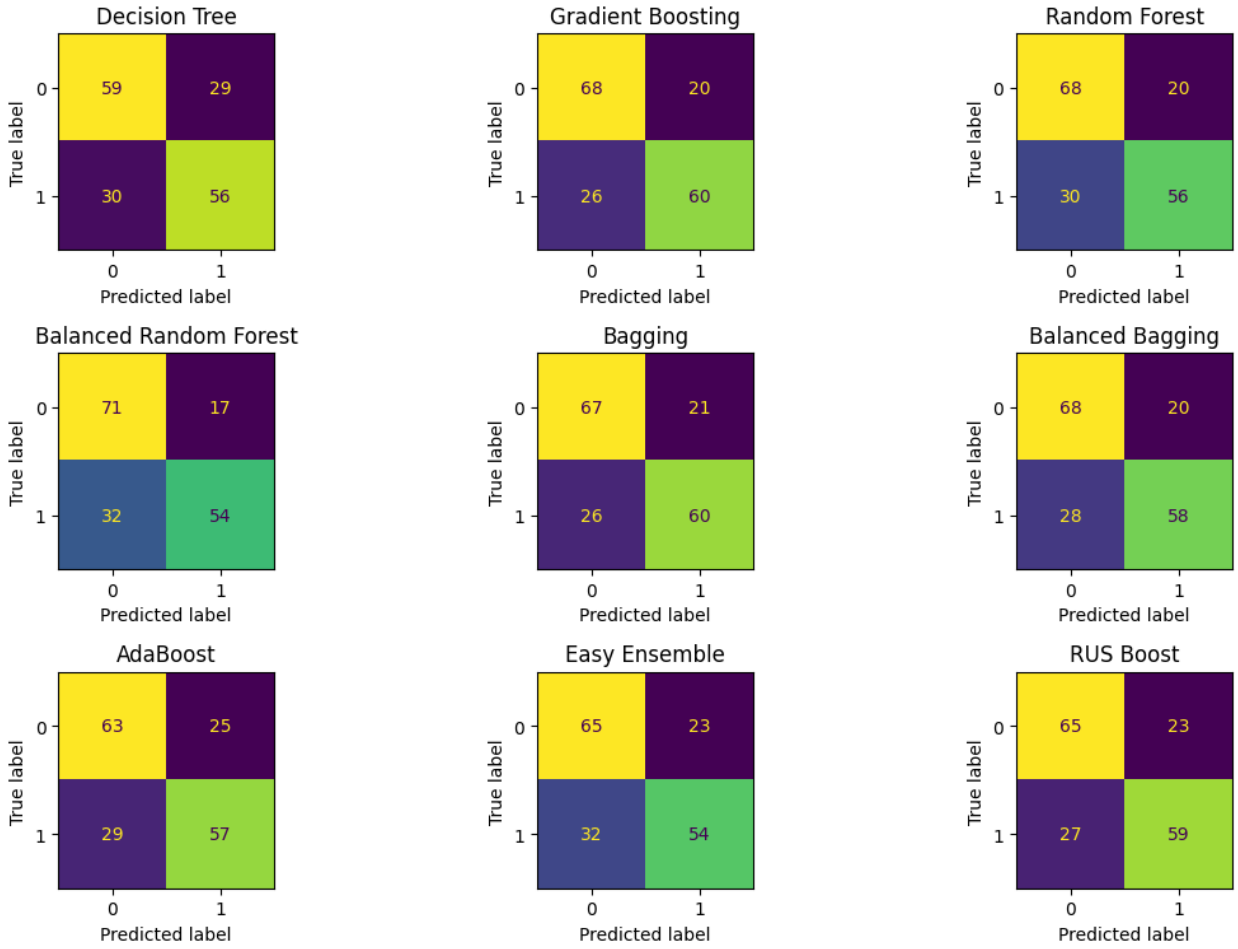
Figure 2: f1 Score of Different Data Processing Methods for Tested Machine Learning Classifiers

	original data	PCA	RandomUnderSample	SMOTE	Combined
Decision Tree	0.150289	0.109756	0.654971	0.140000	0.138169
Gradient Boosting	0.021277	0.021277	0.722892	0.021978	0.200364
Random Forest	0.069565	0.054054	0.691358	0.082645	0.210101
Balanced Random Forest	0.187726	0.201521	0.687898	0.216561	0.194332
Bagging	0.064516	0.018692	0.718563	0.066667	0.190476
Balanced Bagging	0.227273	0.206718	0.707317	0.239748	0.179104
AdaBoost	0.081633	0.040404	0.678571	0.079208	0.188679
Easy Ensemble	0.174326	0.189893	0.662577	0.198444	0.188679
RUS Boost	0.119554	0.187364	0.702381	0.142684	0.188679

As cane be seen in Figure 2 demonstrating the performances of multiple data processing measures, there are only models with Random Under Sample method producing high f1-score for the minority class (customers buying MANU polity) while that of the others demonstrates highly poor performance (lower than 0.3).

Regarding model categories, Balanced Random Forest, Gradient Boosting and Bagging Classifiers are the methods giving high performance with f1 score of 0.688, 0.723 and 0.719 respectively. Particularly as can be witnessed in Figure 2, Balanced Random Forest reduces the number of False Positive (customers buying MANU but predicted not buying) whereas Gradient Boosting and Bagging show a smaller number of False Negative (customers not buying MANU but predicted buying).

Figure 3: Confusion Matrices of Tested Models with Under Sampling



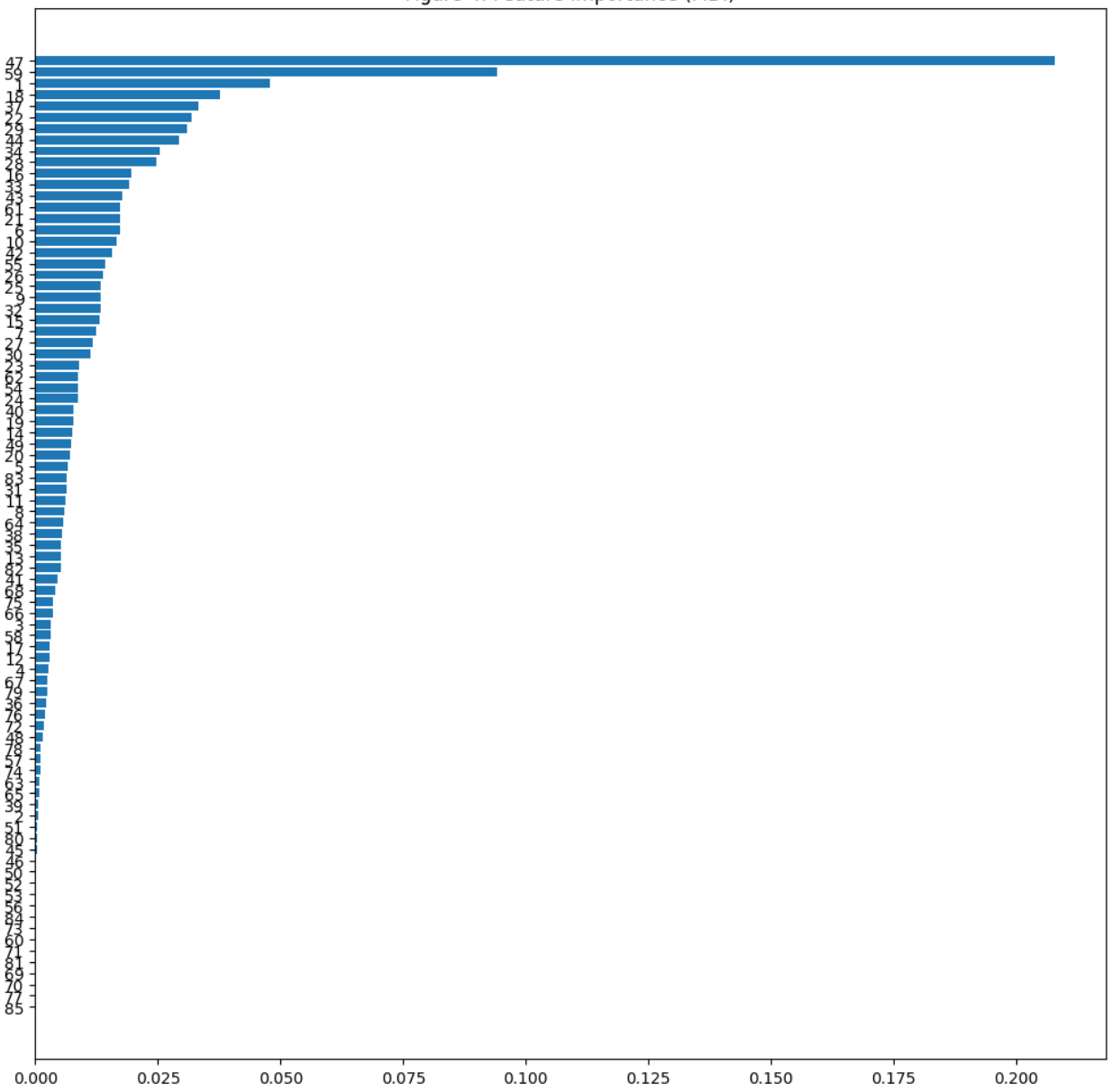
In the case of prediction for marketing strategy, FP could be more important than FN since promoting for wrong people that will not actually possess the polity will probably produce less potential benefit loss than not marketing for ones that will definitely take it.

Therefore, Bagging and Gradient Boosting Classifier should be more prioritized to be selected in this case. As a result, a text file with predicting results of these 2 selected Machine Learning models will be generated for further consideration of difference case scenarios in reality. For generation of top 800 customer's IDs sorted in descending order on their probability buying Banca will be implemented based on the results of Gradient Boosting Classifier.

Since the training dataset is limited with only 348 samples for each class, it is highly recommended to gather more data for the minority class so that the model's performance could be potentially enhanced for the reliability if applied on real practical cases.

Part II – Explanation Task

Figure 4: Feature Importance (MDI)



In this highly dimensional scenario, it is considered to be not convenient and time-consuming to examine the tendence of each dependent feature with respect to the target one. Therefore, feature importance of the selected method (Gradient Boosting Classifier) will be utilized to visualize and evaluate the worth-concerning characteristics.

As can be seen in Figure 4, the importance weights are distributed abundantly in 47^{th} and 59^{th} features before decreasing exponentially towards the other attributes. In this case, the top 5 features with highest importance coefficients will be selected for further exploratory data analysis (EDA).

Figure 5: Pairplot for Selected Important Features

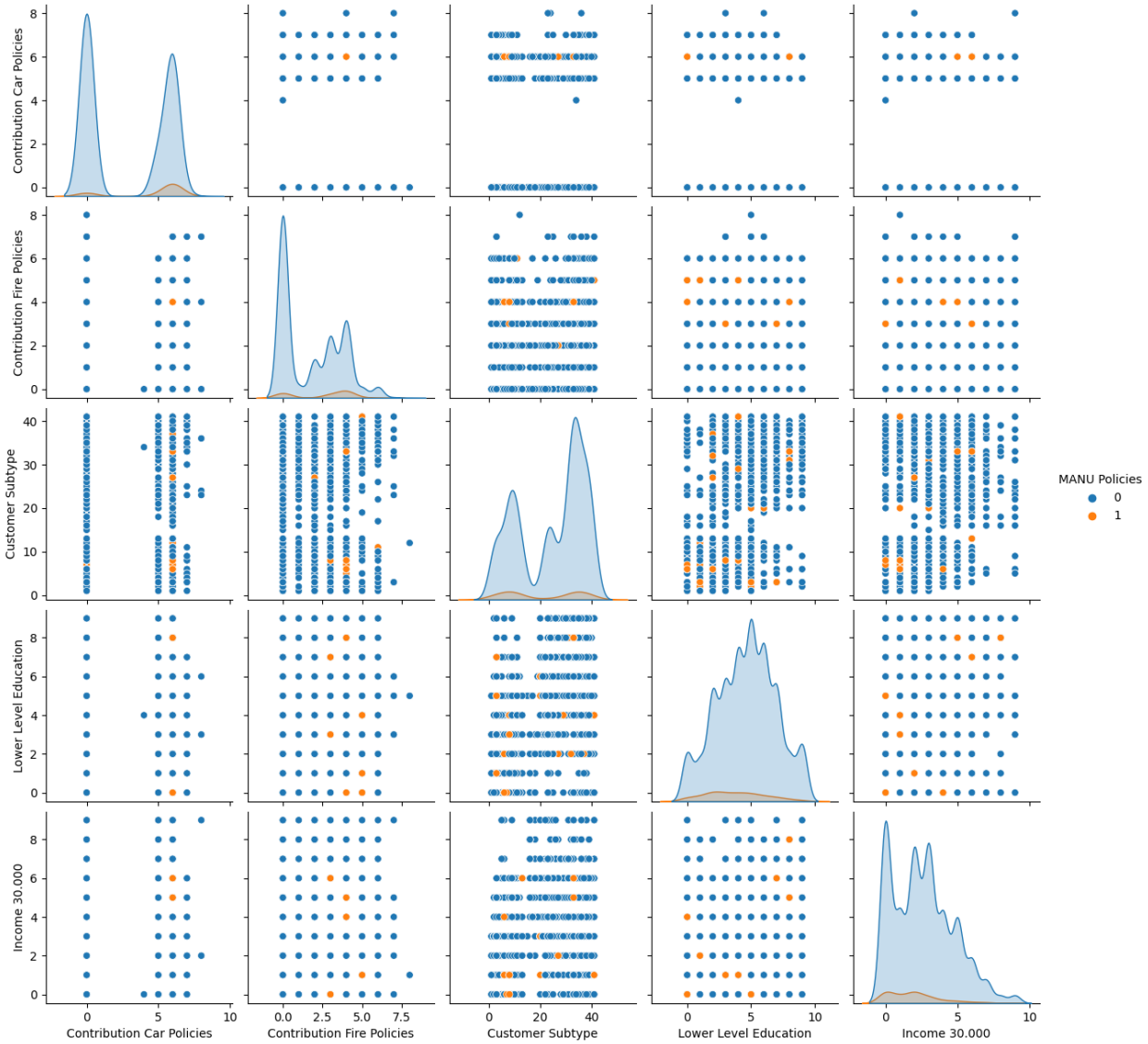


Figure 5 generally shows that the distribution of MANU policies is unclear concerning the 3 last selected features. Notably regarding the 2 first attributes, it is witnessed that the customers that possess the policy is likely to be categorized in the type 6 of Contribution Car Policies and type 3, 4, 5 of Contribution Fire Policies.

As a result, when considering the sale campaign, it is recommended to consider the customers belonging to those indicated types.

NOTE: For more detailed information on the implementation methods and processes, please take a look at the notebook file: Implementation.ipynb included in the submission file.