INDIAN INSTITUTE OF TECHNOLOGY, BOMBAY

CS725: Project Report

On

Analysis of performance of various classification models for a binary classification task on Bank marketing dataset

Under the guidance of

Prof. Preethi Jyothi

By:
Anmol Namdev (213050044)
Prishat Bachhar (213050078)
Shivam Gautam (213050003)
Subodh Latkar (213050047)
Kalpankur Pandey (204277002)

About the Dataset

The dataset we chose to work on is taken from the following link: UCI Bank Marketing Dataset
The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. The output will be if the product (bank term deposit) would be ('yes') or not ('no') subscribed.
The CSV file of the data contains 17 columns, containing 16 attributes and 1 output(whether subscribed or not for the term deposit).
The dataset has in total 45211 samples out of which 39922 are negative samples and 5289 are positive samples.

The Attributes it have and their detailed description is as follows:
1 - age (numeric)
2 - job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
3 - marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown')
4 - education (categorical: 'tertiary', 'secondary', 'primary')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - balance: How much account balance a person have if he have an account with the bank
7 - housing: has a housing loan? (categorical: 'no','yes','unknown')
8 - loan: has a personal loan? (categorical: 'no','yes','unknown')
9 - contact: contact communication type (categorical: 'cellular','telephone')
10 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
11 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
12 - duration: last contact duration, in seconds (numeric).
13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
15 - previous: number of contacts performed before this campaign and for this client (numeric)
16 - poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'unknown', 'success')

Data Cleaning

The steps followed for cleaning the data are as follows:

- In the balance column , We have dropped the outlier values which are more than 2.5 standard deviations away from the mean.
- We dropped the Contact field and changed the duration of call from seconds to minutes and changed the month from words to integers.
- We dropped calls that are less than 10s and we dropped rows with education, job, poutcome stated as other.
- We have also dropped Senior Citizens acting as outliers (age more than 65).

Feature Importance

The most important features are**:**

1. Account Balance
2. Age
3. Duration
4. Number of Contacts

So the main outcomes of the modelling are:

- The bank should not call a customer more than 7 times as it increases dissatisfaction.
- The bank should target elderly and young people. They have the most acceptance rate.
- The bank should prioritize people with high balances.

Support Vector Machines

Results:

Techniques	Precision		Recall		F1 Score		Accuracy
	0	1	0	1	0	1	
No balancing	0.92	0.67	0.98	0.32	0.95	0.43	0.91
Oversampling Minority	0.98	0.39	0.84	0.83	0.90	0.53	0.84
Undersampling Majority	0.98	0.36	0.82	0.85	0.89	0.51	0.82
Using SMOTE	0.96	0.41	0.88	0.70	0.92	0.52	0.86

Observations:

1. The RBF kernel provides the best results among linear, polynomial, sigmoid and rbf.
2. Increasing and decreasing the value of regularization parameter C in the SVC class provided by Sklearn showed no significant improvement in accuracy and other metrics.

3. For all the balancing techniques the Area Under the Curve (AUC) of the SVM model was 0.91 which is more than that of without balancing.
4. The F1 score is also better for the positive class when balancing is applied.
5. Although the accuracy suffers the balanced models provide better True Positive rates at lower False Positive rates (since, higher AUC) and better F1 score for the positive class.

Neural Networks

We tried to solve the selected Bank Marketing Binary classification problem using 2 strategies and 3 balancing techniques.

1) The first strategy involved label encoding the fixed valued features, i.e., do not create a separate column for each class in the feature and just by simply label encoding different classes in the same column.
   The result of this strategy using the 3 balancing techniques are:

Techniques	Precision		Recall		F1 Score		Accuracy
	0	1	0	1	0	1	
No balancing	0.94	**0.53**	**0.95**	0.47	**0.94**	0.50	**0.90**
Oversampling Minority	0.94	0.48	0.93	0.51	0.94	0.50	0.89
Downsampling Minority	**0.98**	0.35	0.81	**0.84**	0.88	0.50	0.81
Using SMOTE	0.94	0.48	0.92	0.56	0.93	**0.52**	0.88

2) The second strategy involved One-Hot encoding the fixed valued feature and a separate binary column was created for each class in the column.
   The Neural Network and the result of this strategy using the 3 balancing techniques are:

Techniques	Precision		Recall		F1 Score		Accuracy
	0	1	0	1	0	1	
No balancing	0.93	0.56	**0.95**	0.45	**0.94**	0.49	0.89
Oversampling Minority	**0.96**	**0.91**	0.90	**0.97**	0.93	**0.94**	**0.93**
Downsampling Minority	0.81	0.84	0.85	0.80	0.83	0.82	0.82
Using SMOTE	0.94	0.45	0.91	0.57	0.92	0.51	0.87

Interpretations:

It can be seen that the balancing techniques used under the first strategy doesn't seem to help much. But as we move to the second strategy, Oversampling Minority classes technique works quite effectively and impressive results are obtained.

Reason to why SMOTE didn't worked

To understand the reason as to why SMOTE does not work here in dealing with imbalance in our case, we must first look into How SMOTE actually generates the Synthetic Samples of the minority class. Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbors for that example are found (typically k=5). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space. But the caveat here is that it does not take care about the classes to which the nearest neighbors belong to. So this can in turn introduce noise in the data and the accuracy may fall. Also, SMOTE is less feasible and not practical for very high dimensionality data.

Decision Tree Classifier

Decision Tree

Easy to extract feature importance and help explain the model.

Parameters :-

a. max_depth: how deep is the decision tree allowed to split?

b. Criterion: information gain is criterion or gini impurity.?

While performing training, the Decision Tree algorithm was imported from scikit-learn.

Created a Decision Tree classifier. 'max_depth' (best = 7), 'criterion (entropy) were progressively tuned while comparing parameters by examining best_parameter output. Accuracy And AUC for ROC increased for max_depth of 7 although Gini Impurity did not show such Improvement in accuracy and AUC for ROC.

	Recall	Precision	Accuracy	Support
0(No cases)	0.92	0.97	0.95	7077
1(Yes cases)	0.56	0.33	0.42	839
Macro avg.	0.74	0.65	0.68	7916
Weighted avg.	0.89	0.90	0.89	7916

Model Performance and Evaluation

While evaluating model performance using ROC curve and Evaluation metrics, the original imbalanced training set was used with the test set to reflect the low subscribe rate in the natural setting.

1) Imported all the trained models from the local environment.
2) Plotted ROC curves for training set and test set .

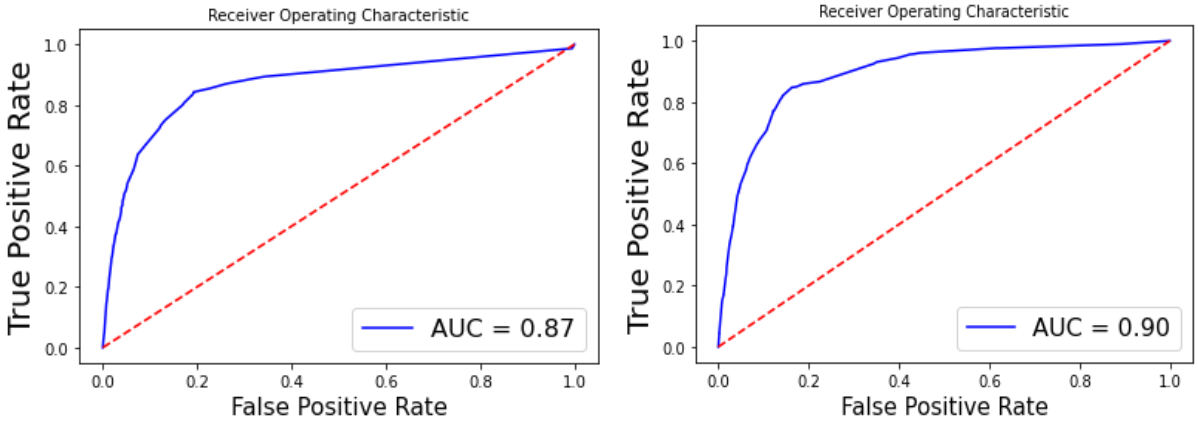


Figure:- Left Hand Diagram Shows AUC for Gini impurity and right hand curve shows AUC for Information gain.

When evaluating performance in imbalanced training set and test set the final model not only predicted higher probability value for true subscribers but also achieved the highest area under curve score (AUC).

Confusion matrix Analysis :

Analysis of Confusion Matrix indicates that with only 20% clients
contacted, it can identify 332 of subscribers in the final model. This is clearly intuitive: use minimal effort to maximize return.

[[6847 230]
[507 332]]

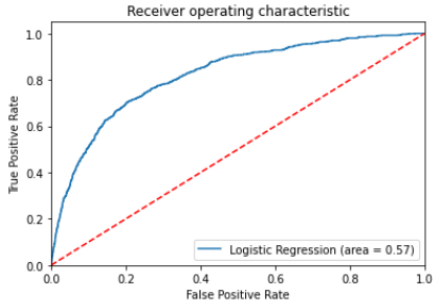
Results:

Techniques	Precision		Recall		F1 Score		Accuracy
	0	1	0	1	0	1	
No balancing	0.93	**0.60**	**0.97**	0.40	**0.95**	0.50	**0.90**
Oversampling Minority	0.97	0.39	0.84	0.83	0.90	**0.53**	0.84
Undersampling Majority	**0.98**	0.34	0.78	**0.89**	0.87	0.49	0.79
Using SMOTE	0.96	0.40	0.86	0.73	0.91	0.52	0.85

Logistic Regression

One of the regression techniques used in the case study was logistic regression.The library used is SKLearn, we started with finding the test and the train XG Boost accuracy and found that
XGB accuracy score for train: 0.937: test: 0.906
Model was build using statsmodel library and optimization terminated successfully after 8 Iterations, The P value of all the independent variable was very less barring 0.571 for pdays.Education and duration had positive coefficients and others had negative however pdays and previous had very insignificant positive coefficient.

Accuracy of the logistic regression classifier on the test set was found 0.90. 10-fold cross validation average accuracy was found 0.896.Confusion matrix and other parameters on unbalanced data was as follows:



We had identified that the data is highly skewed so we tried all 3 balancing techniques with following results:. Overall comparison in a table is as follows

Techniques	Precision		Recall		F1 Score		Accuracy
	0	1	0	1	0	1	
No Balancing	0,91	0.6	0.99	0.16	0.94	0,25	0.9
Oversampling Minority	0.91	0.53	0.98	0.18	0.95	0.27	0.9
Undersampling Majority	0.91	0.58	0.98	0.22	0.95	0.31	0.9
Using SMOTE	0.91	0.56	0.99	0.15	0.94	0.23	0.9

Ablation Study

We divided the set of 16 features into 4 groups:
Group 1: Client Information related features---**age, job, marital, education, default**
Group 2: Banking related attributes---**housing, loan, balance, month_int**
Group 3: Related to the last contact made---**day, month, duration, campaign, pdays, previous, poutcome**
Group 4: Most important 5 features as can be seen from Feature Importance---**duration, balance, day, age**

Groups	SVM			Neural Network			Logistic Reg.			Decision Tree		
	F1 Score		Acc	F1 Score		Acc	F1 Score		Acc	F1 Score		Acc
	0	1		0	1		0	1		0	1	
1	0.94	0	0.89	0.94	0	0.89	0.94	0	0.89	0.94	0.03	0.89
2	0.94	0	0.89	0.94	0.15	0.89	0.94	0	0.89	0.94	0	0.89
3	**0.95**	**0.44**	**0.91**	**0.96**	**0.53**	**0.92**	0.95	0.30	0.90	**0.95**	**0.48**	**0.91**
4	0.95	0.28	0.90	0.94	0.34	0.90	0.95	0.29	0.90	0.95	0.33	0.90
1,2	0.94	0	0.89	0.94	0.17	0.88	0.94	0	0.89	0.94	0	0.89
1,3	0.95	0.41	0.91	0.94	0.47	0.89	0.95	0.30	0.90	0.95	0.45	0.91
1,4	0.95	0.22	0.90	0.95	0.33	0.89	0.95	0.27	0.90	0.95	0.28	0.90
2,3	0.95	0.40	0.91	0.95	0.49	0.91	**0.95**	**0.31**	**0.90**	0.95	0.44	0.90
2,4	0.94	0.29	0.89	0.94	0.45	0.90	0.94	0.27	0.89	0.94	0.31	0.90
3,4	0.95	0.44	0.91	0.95	0.51	0.90	0.95	0.28	0.90	0.95	0.44	0.90
1,2,3	0.95	0.40	0.91	0.94	0.49	0.89	0.95	0.30	0.90	0.95	0.42	0.90
1,2,4	0.95	0.23	0.90	0.94	0.44	0.89	0.95	0.27	0.90	0.95	0.30	0.90
1,3,4	0.95	0.42	0.91	0.95	0.43	0.89	0.95	0.31	0.90	0.95	0.46	0.91
2,3,4	0.95	0.40	0.90	0.95	0.53	0.90	0.94	0.29	0.89	0.94	0.42	0.90
1,2,3,4	0.95	0.40	0.90	0.94	0.52	0.90	0.94	0.30	0.89	0.94	0.45	0.90

Github Link

https://github.com/anmolnamdev/CS725_Final_Project

References

1. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
2. https://www.datacamp.com/community/tutorials/diving-deep-imbalanced-data
3. https://towardsdatascience.com/balancing-is-unbalancing-5f517936f626
4. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/
5. https://www.datacamp.com/community/tutorials/decision-tree-classification-python
6. https://towardsdatascience.com/get-your-decision-tree-model-moving-by-cart-82765d59ae09
7. https://www.einfochips.com/blog/data-cleaning-in-machine-learning-best-practices-and-methods/
8. https://machinelearningmastery.com/logistic-regression-for-machine-learning/
9. https://towardsdatascience.com/machine-learning-for-beginners-an-introduction-to-neural-networks-d49f22d238f9
10. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/