Banking Dataset - Marketing Targets

Ece KOBANÇ
Computer Engineering Department
Dokuz Eylul University
İzmir, Turkey

Efe Kaan KABAKAŞ
Computer Engineering Department
Dokuz Eylul University
Izmir, Turkey

Keywords—machine learning, class imbalanced, smote, binary classification, decision tree, logistic regression, k-nearest neighbour, support vector machine, random forest

I. INTRODUCTION

A time deposit is a type of bank account that is deposited in cash with a financial institution and yields interest to the depositor over a specified period of time. Time deposits are the most important source of income, especially for commercial banks. For this reason, banks look for a customer profile that can sell time deposit accounts and develop marketing strategies. Today, the most common form of marketing is to reach the appropriate customer profile by phone, but huge call centres are established for this method and the cost of this method is quite high. For this reason, banks target the customer profile that they will not waste time on. This study was carried out with a dataset examining the customer profile of a Portuguese bank in seventeen different categories. The aim of the study is to find a relationship between these categories and subscriptions and to determine which categories directly affect the sale of time deposit subscriptions. The classification objective of the study is to predict whether the sought-after customer will subscribe to the time deposit.

II. RELATED WORKS

1) In the real world, the number of data is quite large and the distribution of these data is uneven. In this study, different oversampling techniques proposed for the solution of the class imbalanced problem were examined and these solutions were compared by trying different data sets.

   Four different oversampling methods, namely SMOTE (Synthetic minority oversampling approach), ADASYN, Borderline-SMOTE, and Safe-Level SMOTE, were compared. Since it is an unbalanced data set, sensitivity, specificity, precision, F-mean and g-mean metrics were used instead of accuracy and error rate.

   When these models are examined on different data sets, it is seen that the Safe Level SMOTE technique achieves more accuracy than other techniques (Gosain., 2017).

2) In this study, the same data set data mining methods as ours were applied. MLPNN, LR, TAN and C5.0 were used as models. As a result of the study, they determined that the C5.0 model was ahead of the other models, even if it was a small amount (Elsalamony.,2017).

A. Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. MATERIAL AND METHODS

The data set used for the study was obtained from kaggle.com. The title of the dataset is “Banking Dataset – Marketing Target”. First, the data were analyzed to make the model ready for training. When the data set was examined, it was seen that there was no empty space or duplicated data. Then, data quality reports were made as separate categorical and numerical data. Then, using appropriate visualization techniques, the relationships between the data were examined and an inference was tried to be made.

First, categorical data were created in a pie chart and examined. Then the data were analyzed in a bar graph. Especially when the bar graph is applied to the target feature, a very serious difference has emerged in the distribution between the data in the target feature. This means that we are facing a class imbalance problem. While examining the relationship between target feature and data, categorical data were obtained by using bar plot graphs; Continuous data were analyzed using histograms and distribution graphs.

The data were analyzed by the IQR method to find outlier values. However, since the data was very scattered and there were too many outlier values, interfering with all the outlier values changed the dataset drastically. So we visualized the data and eliminated values that were very outliers by putting a threshold value. The number of data objects, which was 45211 in its original state, decreased to 45147 after this process.

Although there is no missing or duplication, the biggest problem in this data set is that the number of "unknown" values is quite high. The “job”, “education”, “contact” and “poutcome” columns contain a lot of unknown data. In order to cope with this situation, a relationship has been established between the "job" and "education" features. The lines were combined so that the "job" and "education" features remained as labels and the ratio was found. From the sub-features of the "job" feature, which occupational

groups and which education level they have been shown in this new table, and the rates are determined.

In this way, the educational status of the "unknown" values in the "job" feature was examined. The occupational group is assigned to the "unknown" value in whichever occupational group that education level is higher. Similarly, the professions of the lines with "unknown" value in the "education" property were examined. Unknown values of a vocational group are assigned to the most common educational level in that vocational group.

In addition, the mode value of the "contact" property is assigned to the unknown values in the "contact" property.

In order to use this categorical data in mathematical algorithms, it had to be converted into numerical data, and some encoding techniques were applied to do this. When the categorical data in the data set are examined, it is seen that there are 10 of them. And since the value range of 5 of these 10 categorical variables has only 2 values, these data were mapped as 0 and 1 and encoded. The values of these 2 variables (housing, default, loan, target) have been replaced with 1 if 'yes' and '0' if 'no'.

Creating dummy attributes is used for variables with more than 2 values (job, marital, education, poutcome). After this process, while the original columns were removed, a new column with a value of 0 and 1 was added for each value of that column. Finally, when the encoding process was finished, a 31-column data set was formed.

When the numeric data is considered, it is seen that each of them has very different value ranges. These differences are clearly visible in the data quality report, and for these numeric features to be of equal importance in mathematical algorithms, they all had to have the same range of values. To do this, min-max normalization was used on the relevant features. In this way, all the features have new values between 0 and 1.

When the distribution of the target feature is examined, it is seen that the majority of the customers (11.6%) are in the 'No' class. While 39922 customers are in the 'No' class, 5289 customers are in the 'Yes' class.

This leads to the imbalanced class problem. In order for the model to be able to generalize the test data correctly, it should be trained with a balanced training dataset from all features. In this case, to solve this problem for this dataset with 2 features, it was necessary to either apply the undersampling technique to the class with the majority of the instances or to apply the oversampling technique to the class with the minority of the instances.

The method of applying the oversampling technique to the class with the minority of the samples was chosen in order not to experience data loss and decrease the quality of the model.

SMOTE (Synthetic Minority Oversampling Technique), one of the oversampling techniques, was used to perform this operation. This method generates synthetic data from minority class samples using the k-nearest algorithm and eliminates the imbalance by equating the sample numbers of the features. As a result, 'No' - 'Yes' values equalled 39879 to 39879, which were 39879 to 1869 before SMOTE.

IV. DATASET INFORMATION

In this study, there are 45211 data objects in the dataset that predicts whether the customer will subscribe or not with supervised binary classification techniques. Also, each row has 17 columns. 10 of these columns are categorical and 7 of them are numerical. While some of the categorical variables have 2 values, some others have more than 2 values. The value ranges of numeric data are also different from each other. The description and type information for all variables are detailed below as shown by Elsalamony [1], as previously described.

TABLE I. ATTRIBUTE DESCRIPTION

Attributes	Type	Description	Range
Age	Numeric	How old the customer is	18:95
Job	Categorical	Job type	-
Marital	Categorical	Marital status	-
Education	Categorical	Education level	-
Default	Categorical	Has credit in default?	-
Balance	Categorical	Average yearly balance, in euros	-8019: 102127
Housing	Categorical	Has housing loan?	-
Loan	Categorical	Has customer loan?	-
Contact	Numeric	Communication type	-
Day	Numeric	Last contact day of the month	1:31
Month	Numeric	Last contact month of the year	-
Duration	Numeric	Last contact duration, in seconds	0:4918
Campaign	Numeric	Number of contacts performed for this client (includes the last contact)	1:63
Pdays	Numeric	Number of days that passed by after the client was last contacted from a previous campaign	-1:871
Previous	Numeric	Outcome of the previous marketing campaign	0:275

Poutcome	Categorical	Has the client subscribed a term deposit?	-
Target	Categorical	Has the client subscribed?	-

V. RESULTS

TABLE II. ACCURACY COMPARISON

Models	Acc 25/75, Before SMOTE	Acc 25/75, After SMOTE	Acc 10-Fold CV, Before SMOTE	Acc 10-Fold CV, After SMOTE
LR	0.89	0.82	0.90	0.81
KNN	0.88	0.87	0.89	0.87
NB	0.89	0.82	0.89	0.82
DT	0.85	0.87	0.85	087
RF	0.89	0.92	0.90	0.90
SVM	0.89	0.84	0.89	0.83

TABLE III METRIC COMPARISON

Models	Accuracy	Recall	Precision	F1 Score
LR	0.81	0.80	0.80	0.82
KNN	0.87	0.94	0.84	0.82
NB	0.82	0.83	0.83	0.89
DT	087	0.88	0.87	084
RF	0.90	0.95	0.91	0.93
SVM	0.83	0.84	0.84	0.84

VI. CONCLUSION

In the study, five different model algorithms, namely Logistic Regression, K-nearest Neighbor, Support Vector Machine, Naive Bayes and Random Forest, were used in order to make an accurate binary classification and compare the results.

When the models are examined, Logistic Regression has an accuracy value of 0.89 and a cross-validation value of 0.90 before the smote process. However, these values are not healthy due to the class imbalanced problem. After the Smote process, the accuracy in the Logistic Regression model was 0.82 and the cross-validation value was 0.81.

In the K-nearest Neighbor model, the accuracy value was 0.88 and the cross-validation value was 0.89 before the smote process. These values were calculated as 0.87, both the accuracy value and the cross-validation value, after the oversampling process.

In the Naive Bayes model, the accuracy and cross-validation values were calculated as 0.89 before the smote process, and both values were calculated as 0.82 after the smote process.

In the Decision Tree model, the accuracy and cross-validation values were calculated as 0.85 before the smote process, and both values were calculated as 0.87 after the smote process.

In the Support Vector Machine model, accuracy was calculated as 0.89 before smoothing. The cross-validation value was also found to be 0.89. After the smoothing process, the accuracy value was calculated as 0.84 and the cross-validation value as 0.83.

Finally, in the Random Forest model, the accuracy was 0.89 before smoothing and the cross-validation value was 0.90. After oversampling, these values accuracy and cross-validation values were found to be 0.92.

When the results are examined, it is seen that the accuracy value is decreased after the oversampling process in all models except the Decision Tree. However, recall and precision values are in a healthy balance.

According to the results, the Random Forest method gave the highest values after the highest oversampling process. Random Forest model should be used for this data set.

REFERENCES

[1] Hany A. Elsalamony, Bank Direct Marketing Analysis of Data Mining Techniques, International Journal of Computer Applications (0975 – 8887), vol. 85, Jan 2014.

[2] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 79-85, doi: 10.1109/ICACCI.2017.8125820.