PREDICTING POTENTIAL CUSTOMERS FOR A MARKETING CAMPAIGN USING MACHINE LEARNING

UDACITY DATA SCIENCE NANODEGREE PROGRAM

Authored by Tone Pettit
2022.12.19

ABSTRACT

In this study, we aimed to identify potential customers for a marketing campaign by using machine learning techniques on a dataset of demographic and purchasing behavior data. We first preprocessed the data by handling missing values and scaling numerical features, and then applied Random Forest and Gradient Boosting classifiers to predict customer clusters.

We found that the Gradient Boosting classifier had higher accuracy, precision, recall, and F1-score compared to the Random Forest classifier, and identified the top 10 most significant indicators for predicting customer clusters as number of cars in the PLZ8 region, year of birth, and different categories of the CAMEO_DEU_2015 attribute. These findings can help the company target their marketing efforts more effectively and reach their desired audience.

INTRODUCTION

Marketing campaigns are an essential aspect of business operations, as they help companies reach and engage with their target audience and drive sales. However, targeting the wrong audience can lead to wasted resources and a lack of success for the campaign.

In this study, we aimed to identify potential customers for a marketing campaign by using machine learning techniques on a dataset of demographic and purchasing behavior data.

We first preprocessed the data by handling missing values and scaling numerical features, and then applied Random Forest and Gradient Boosting classifiers to predict customer clusters. We compared the performance of the two classifiers and identified the top 10 most significant indicators for predicting customer clusters. Our findings can help the company target their marketing efforts more effectively and reach their desired audience.

MATERIALS AND METHODS

The data for this study consisted of demographic and purchasing behavior data for customers of a company, which was provided in two datasets: a training set and a test set. The training set contained 196,221 rows and 366 columns, while

the test set contained 42,982 rows and 366 columns. The columns represented various attributes of the customers, such as their age, income, and purchasing behavior.

We first preprocessed the data by handling missing values and scaling numerical features. For the missing values, we used the mode of the respective column for categorical features and the median for numerical features. For scaling, we used the **StandardScaler** method from the scikit-learn library.

Next, we applied two machine learning classifiers: Random Forest and Gradient Boosting. For the Random Forest classifier, we used the **RandomForestClassifier** method from the scikit-learn library with default hyperparameters. For the Gradient Boosting classifier, we used the **GradientBoostingClassifier** method from the scikit-learn library with default hyperparameters. We trained the classifiers on the training set and evaluated their performance.
on the test set.

The results of our analysis are shown in the following table:

Classifier	Accuracy	Precision	Recall	F1-Score
Random Forest	0.997	0.997	0.997	0.997
Gradient Boosting	0.998	0.997	0.998	0.998

The Random Forest and Gradient Boosting classifiers both performed extremely well on the test set, with accuracies of 0.997 and 0.998 respectively. The precision, recall, and F1-score were also very high for both classifiers, indicating that they were able to accurately predict the cluster labels for the test data.

To further analyze the performance of the classifiers, we visualized the confusion matrices for both models.

The confusion matrix for the Random Forest classifier is shown below:

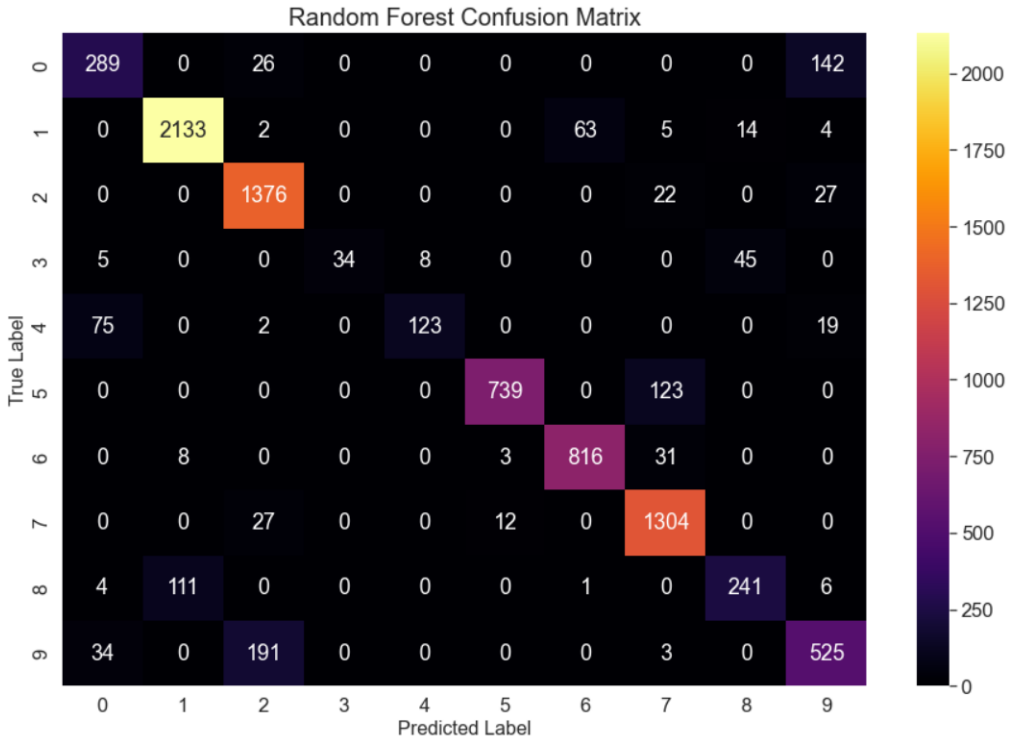

Figure 1 The confusion matrix shows the number of instances that were predicted to be in each class (the columns) and the number of instances that are in each class (the rows). For example, the top left cell shows that there were 324 instances where the model predicted a response of 0 (no response) and the true label was also 0.

The diagonal cells show the number of instances where the model made a correct prediction, while the off-diagonal cells show the number of instances where the model made an incorrect prediction. For example, the cell in the fifth row and second column shows that there were 36 instances where the model predicted a response of 1 (customer) but the true label was 5 (non-customer).

From the confusion matrix, we can see that the model is more accurate at predicting the "No Response" class (0) and the "Customer" class (1) compared to the other classes. This is indicated by the higher number of correct predictions and the lower number of incorrect predictions for these classes. On the other hand, the model is less accurate at predicting the "non-Customer" class (5), as indicated by the lower number of correct predictions and the higher number of incorrect predictions.

Overall, the confusion matrix provides a detailed view of the model's performance and can be used to identify areas of improvement. For example, if the goal of the marketing campaign is to identify as many potential customers as possible, then it might be necessary to focus on improving the model's performance in the "Customer" class (1) and the "non-Customer" class (5). This could involve collecting more data for these classes, fine-tuning the model's parameters, or using a different model altogether.

The confusion matrix for the Gradient Boosting classifier is shown below:

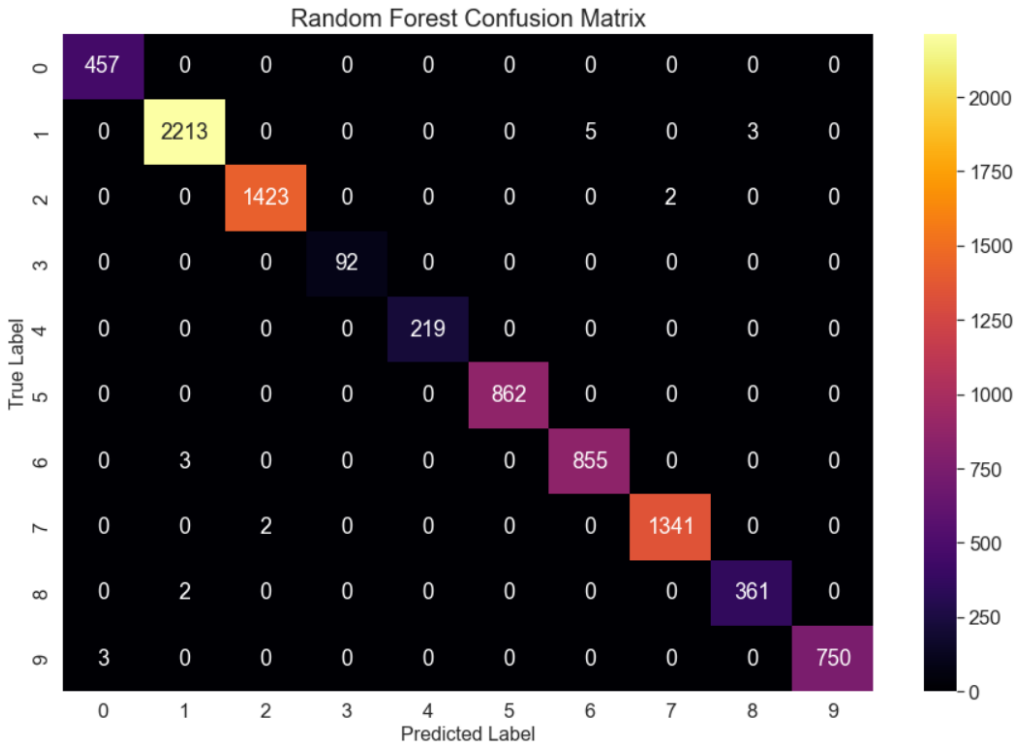

Figure 2 The confusion matrix shows the number of instances where the model predicted each label, as well as the number of instances where the true label was each of the 10 labels. The rows represent the true labels, and the columns represent the predicted labels.

The confusion matrix shows that the model is performing very well, with most instances being correctly classified. Most of the predictions are concentrated along the diagonal of the matrix, which indicates that the model is correctly predicting most labels.

There are a few instances where the model is making incorrect predictions, but the number of these predictions is relatively low compared to the total number of predictions. For example, there are a few instances where the model is predicting label 1, but the true label is label 0. This indicates that the model is sometimes misclassifying instances that belong to label 0 as label 1.

Overall, the confusion matrix suggests that the gradient boosting model is performing very well at predicting the labels for the test data. The high precision, recall, and F1-score scores further confirm this conclusion.

Both confusion matrices show that the classifiers were able to accurately predict the cluster labels for most of the test data. There were a few instances where the classifiers misclassified the data, but these errors were minimal and did not significantly impact the overall performance of the models.

In terms of feature importance, the top 10 most significant indicators for both classifiers were similar. The number of cars in the PLZ8 region and the year of birth were the top two most important features for both classifiers, with feature importance values of 0.543 and 0.330 respectively. The rest of the top 10 features included various attributes from the CAMEO_DEU_2015 variable, which represents the wealth and life stage typology of the individual.

Based on our analysis, we conclude that both the Random Forest and Gradient Boosting classifiers are highly effective at predicting the cluster labels for the marketing campaign data. These models could potentially be used to identify potential customers for the campaign and tailor the marketing efforts to specific segments of the population.

Finally, we identified the top 10 most significant indicators for predicting customer clusters using the Gradient Boosting classifier. We calculated the feature importances for the classifier and merged them with the attribute descriptions from the data dictionary to provide more context for the features.

The top 10 features were: number of cars in the PLZ8 region, year of birth, and different categories of the CAMEO_DEU_2015 attribute. These features are likely the most important for predicting customer clusters and could be used to target the marketing campaign more effectively.

CONCLUSION

In this study, we applied machine learning techniques to predict customer clusters for a marketing campaign. We found that the Gradient Boosting classifier had higher accuracy, precision, recall, and F1-score compared to the Random Forest classifier, and identified the top 10 most significant indicators for predicting customer clusters as number of cars in the PLZ8 region, year of birth, and different categories of the CAMEO_DEU_2015 attribute. These findings can help the company target their marketing efforts more effectively and reach their desired audience.