Identification of Customers with High Likelihood to Purchase

Chien-Chun Liao, Yi-Wen Shih, Jane Song

[chien-chun.liao@utdallas.edu](mailto:chien-chun.liao@utdallas.edu), [Yi-Wen.Shih@UTDallas.edu](mailto:Yi-Wen.Shih@UTDallas.edu), [jie.song@utdallas.edu](mailto:jie.song@utdallas.edu)

Abstract

Super Server Company should identify and pursue clients with the highest likelihood of purchase using data analysis in order to maximize profit. Machine learning classification models, including K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree Classifier, Logistic Regression, and Neural Network, were used to label each customer account as either a potential buyer (1) or not a potential buyer (0). The most accurate models were then used in an ensemble to produce the best predictive model. In conclusion, we predicted the probability of each account belonging to the “buyer” class with 0.91 accuracy.

Introduction

Super Server Company (“The Company”) intends to carry out a marketing campaign targeted towards potential buyers with the highest likelihood of purchase. Incorrect classification of potential buyers as non-buyers results in missed income opportunities. On the other hand, misidentification of a non-buyer as a customer with purchase potential results in wasted resources. As such, precise and accurate data analysis must be performed while taking into account the limitations of the dataset provided.

Methods

The data obtained on customer accounts includes five attributes: The Company sells two types of products, cloud and networking, and has three known competitors.

The target variable, “purchase\_event”, indicates whether a customer has previously purchased from the Company. The goal is to label customer accounts as 0 (not a buyer) or 1 (a buyer). The probability of each account belonging to the “buyer” class is also of interest since the task at hand is to generate a ranked list of accounts from the highest to lowest likelihood.

*Unbalanced Dataset*

The distribution of target classes is severely imbalanced with 2,351(5.07%) instances in the minority class and 43,981(94.83%) instances in the majority class. The imbalanced nature of the dependent variable may hinder machine learning models’ ability to learn and distinguish between classes.

Common methods of handling unbalanced datasets include over-sampling, under-sampling, or both. We chose to utilize a combination method in which the minority class is over-sampled and the majority class is under-sampled incrementally and simultaneously in equal folds until the composition is uniform. In this case, the minority class was enlarged approximately 4.32 times while the majority class was reduced approximately 4.32 times. This resulted in a balanced dataset with 20,338 rows.

*Dealing with Errors*

Descriptive analysis performed on the Firmographic attributes revealed potential errors. For example, many accounts were categorized into multiple categories that should be mutually exclusive. Some firms were recorded as both having less than 1,000 employees and having more than 10,000 employees. Others did not belong to any firm size category. Similar issues were detected with the other firmographics attribute. To mitigate this, we recategorized these types of abnormal instances as “employee number unknown” and “industry unknown”.

*Missing Values*

The dataset provided had a minimal amount of missing data (469), in the “persona\_tech” column. We choose to impute missing values with the mode (‘1’), of the existing observations in the same column.

*Methods and Tools*

We used Python for all data analysis and machine learning tasks performed. Python packages used include *Scikit-learn*, *Pandas, imbalanced-learn.*

We chose to include all attributes, except for the two unknown columns, at the start of our analysis. In order to preserve the interpretability of the final result and provide useful recommendations, we chose not to use dimensionality reduction methods such as Principal Component Analysis (PCA) or Truncated Singular Value Decomposition (SVD) unless absolutely necessary.

The list of models we tested includes K-Nearest Neighbor, Support Vector Machine, Decision Tree Classifier, Logistic Regression, and Neural Network. First, we used the *sklearn.model\_selection.GridSearchCV* function to obtain the best hyperparameters and to evaluate each model with cross validation procedure. Next, we combined the most suitable algorithms to produce a model with better predictive power using the Weighted Average Ensemble method, with weight equal to the proportion of each model’s accuracy rate to the sum of all model’s accuracy rate.

Results

The industry classification (Retail, Healthcare, Finance, or Infrastructure) of 82% of all the customer accounts that have purchased from the Company or a competitor is unknown. Similarly, we do not know the business size of 75% of past buyers. However, if the unknown instances have the same distribution as the known instances, we can speculate that the biggest buyers of the Company’s products are businesses in the Finance industry with 10,000 to 50,000 employees.

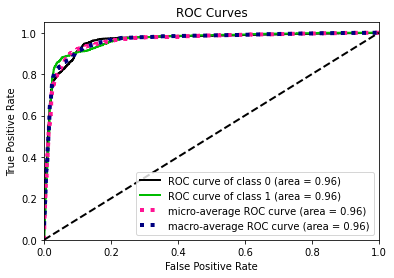
Apart from the descriptive insights the data provided, we also developed a model that is capable of predicting probability of purchase.

The test scores (R-square) of individual models are summarized in the table below.

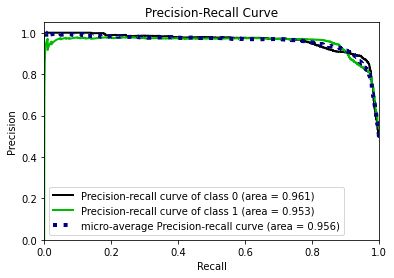
|  |  |
| --- | --- |
| **Models** | **Score** |
| SVC with RBF kernel | 0.97 |
| Decision Tree Classifier | 0.95 |
| SVC with polynomial kernel: | 0.91 |
| KNN | 0.87 |
| Linear SVC | 0.63 |
| Logistic Regression | 0.63 |

The metrics produced by the Weighted Average Ensemble model are shown below.

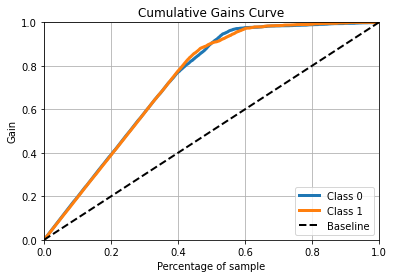
1. The ROC Chart  
   AUC score: 0.91



1. Precision-Recall Chart



1. Gain Chart



Conclusions

In conclusion, we predicted the probability that an account, in the test dataset, belongs in the potential “buyer” class with 0.91 accuracy. The final algorithm detected 595 accounts with a statistical probability of purchasing. We recommend that the Company take the resulting list and its budget into consideration to tune its marketing strategy.