Report for Hwk 2

Huanou Liu

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature Date: 10/16/2017

Preprocessing Stage

After loaded the data set of *directMarketing.csv,* I studied patterns of the whole data set. Findings are listed as following:

* There is no NA values in the data set by checking the info attribute of the data frame.
* "saleSizeCode" attribute is category and its alphabetical order has positive correlation with sale value. So this attribute should be encoded into numeric with original ordinal as {D:1, E:2, F:3, G:4}.
* Instead of the attributes as “firstDate” and “lastDate”, we may be more interested in the whole time span. So I transformed these two columns into a new one as “timeSpan” by calculating the difference of the dates.
* The confirmation of star customer is labeled as “X” which cannot be processed by sklearn module. I thus transform this attribute into {positive : 1, negative : 2}.
* By examining the correlation matrix of all attributes, I found lastSale and avrSale have a strong correlation of 0.8. Then I decide to drop one of them to satisfy the assumption of naïve Bayes model.
* The whole data set is randomly splitted into training and testing sets by sklearn.model\_selection.train\_test\_split with test\_size=0.3

Results Analysis

|  |  |  |
| --- | --- | --- |
| Models | accuracy score | area under the ROC curve |
| decision tree | 0.577 | 0.595 |
| SVM | 0.575 | 0.572 |
| naive Bayes | 0.578 | 0.613 |
| logistic regression | 0.578 | 0.616 |

Table 1. The summary of all models studied

All the 4 models have close accuracy score. For the area under the ROC curve, naive Bayes and logistic regression models exhibit better performance.

A picture containing text

Description generated with high confidence

Figure Confusion Matrix of all models

All the models have similar confusion matrix. While naïve Bayes model has the highest true positive rate and SVM has the highest true negative rate.

A close up of a map

Description generated with high confidence

Figure ROC Curve and AUC of All Models

When compared with logistic regression model, all three models we studied have slightly lower AUC. While all models exhibit similar curve shape.

A picture containing screenshot

Description generated with high confidence

Figure Precision-Recall Curve for all Models

The AUC curves of naïve Bayes and logistic regression models are very smooth generally. While the ones of decision tree and SVM models consist of ridged lines.

Conclusion

The three machine learning models I implemented does not provide better predictive ability in general when compared with the logistic regression model. I may assume the target variable of this project is difficult to model in the machine learning or regression method (with approximate 60% accuracy). I may further tune the models or acquired more data to improve the predictive ability.