Applying Machine Learning to Bank Churn Prediction

Rajbir Bhattacharjee

**Abstract** Regarding *Lorem Ipsum* and according to Wikipedia, in publishing and graphic design, *lorem ipsum* is common placeholder text used to demonstrate the graphic elements of a document or visual presentation such as font, typography, and layout. Although it resembles Latin, it does not have any inherent meaning. Publishers often use *lorem ipsum* when displaying typeface to direct focus to the presentation. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aliquam sapien. Proin euismod metus id elit. Aliquam posuere orci nec lacus. Vivamus consectetuer, turpis non vulputate faucibus, nisl lectus pretium urna, at mollis turpis arcu non quam.

1 Introduction

The data-set considered in this paper is labelled sample of a bank’s customer data. The data is labeled whether the customer left the bank or not. The task is a classification task to predict whether a customer will leave the bank or not.

The dataset has the following fields:

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Notes** |
| RowNumber | Integer | Row Number |
| CustomerId | Integer | Customer Id, no specific pattern noted |
| Surname | String | Surname of the customer |
| CreditScore | Integer | Credit score of the customer |
| Geography | String | Country of the customer, has 3 distinct values, France, Germany and Spain |
| Gender | String | Male or Female |
| Age | Integer | Age of the customer |
| Tenure | Integer | Number of years with the bank |
| Balance | Float | Account balance |
| NumOfProducts | Integer | Number of products customer is subscribed to |
| HasCrCard | Integer | Denotes whether a customer has a credit card or not |
| IsActiveMember | Integer | Denotes whether customer is active or not |
| Estimated Salary | Float | Estimated salary of the customer |
| Exited | Integer | Target class label: 0/1 indicating whether the customer exited the bank or not. |

The data set has 10000 examples with no missing values. However, the data set suffers from moderate class imbalance with about 25% of the samples have the label Exited=1, and the remaining 75% have the label Exited=0.

The motivation to build a machine learning model is to be able to predict which customers will exit the bank after training on the data. A further objective of the study is to compare how different models perform, along with techniques to remedy data imbalance.

2 Research

This work compared the performance various several machine learning models to the dataset. A further goal of this work is to research techniques to remedy data imbalance and compare how they perform on the dataset.

In a classification problem, the ideal situation is where each target label has equal probability of occurrence in the real world, as well as the sample data-set. However, that is seldom the case, and in most real-world problems, the class labels are highly skewed. Consider the case of predicting cancer – only a small percentage of the population have cancer, and the vast majority of the population do not have cancer. Further imbalance maybe introduced by the sampling methodology.

If we are predicting two classes, and the chances of a sample being a +ve is 1%, and the chance of a sample being a -ve is 99%, a machine learning model can achieve an accuracy of 99% by simply predicting -ve all the time. However, such a prediction is of little use. Hence accuracy is not a good judge of the quality of prediction. Better measures of judging accuracy are measures like precision, recall and F-1 score. The F-1 score is particularly interesting because it uses just one number to indicate both false positives and false negatives. Another metric that is frequently used is the confusion matrix.

3 Methodology

4 Evaluation

5 Conclusion

References

1.