MGT415 - Data Science in Practice - Problem Set 2

De Becker, Sophie; Di, Yao; Gallese, Mattia; Martiriggiano, Giacomo
College of Management of Technology, EPFL, Switzerland

I. Introduction

This report will present the results of a Customer Churn analysis carried out on a customer dataset of Tesco. After performing an Exploratory Data Analysis (EDA), it was highlighted the importance of analysing Tesco's negative aspects, such as Customer churn, in order to mitigate Tesco's strategy weakness. The main goal is to predict which customers are more likely to leave the company.

II. Logistic Regression Analysis

The first analysis is performed using logistic regression in order to predict Churn and NoChurn customers. Two logistic regressions are built, and they take respectively as input the monthly charge and the tenure variables. However, this solution can not be implemented since it predicts never churn. Indeed, in Figure 1 the probability of belonging to a class never reaches above 0.5 for churn.

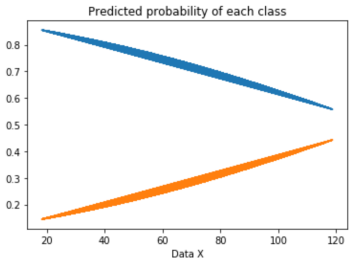

Figure 1. Logistic Regression

III. Classification Analysis

In order to obtain a more accurate and reliable result, two different classification algorithms are applied, random forest and support vector machine. For comparing and evaluating these two models, the recall metric is used. It was chosen, among the other metrics, because Tesco needs the most possible reliable model for a Customer retention strategy against churn.

A. Random Forest

Random Forest model is performed using multiple object inputs. The results obtained by implementing Random forest model are shown in the confusion matrix (Figure2).

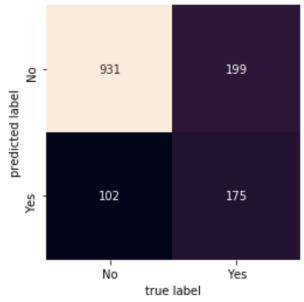

Figure 2. Random Forest-Confusion Matrix

In the confusion matrix it is represented the true churn, given by the test set, and the predicted churn, given by the model. The recall score obtained with Random Forest model is $46,8\%$.

B. Support Vector Machine

Support vector machine is performed using both numerical and object variables. It was noted that the dataset was not balanced based on the churn parameter. Therefore, in order to get a reliable result, the No Churn set of customer was downsampled. The results obtained implementing the Support Vector Machine model are shown in the Figure 3.

	Churn_pred	No Churn_pred
Churn_true	476	85
No Churn_true	468	1081

Figure 3. Support Vector Machine

The recall score obtained with the support vector machine is $84,8\%$, almost two times the one obtained with random forest model.

In order to improve the performance of this classifier, cross-validation and grid search models were applied to support vector machine. The cross-validation model is implemented in order to be sure that the model is not overfitting. While, the grid search model is implemented to find the optimal hyperparameters. The results obtained after applying the grid search model are shown in Figure 4.

	Churn_pred	No Churn_pred
Churn_true	469	92
No Churn_true	490	1059

Figure 4. After Implementing Grid Search Model

The results obtained implementing this optimization are worse than the previous one, since the recall score is $83,6\%$.

IV. Conclusion

To conclude, it is advised to adopt the support vector machine model to predict the customer churn since it has the highest recall score. In this way, the customer retention strategy can be adapted with the lowest amount of losses.