Customer churn analysis

Report by:
Mauro Gioè

Chapter 1

Introduction

The main goals of this analysis are to develop a model which is able predict clients which are going to drop their subscription in the future and which features help the most predicting such behavior. Data come from a sample provided by Telco which was uploaded on Kaggle [3]. Additionally to the information about clients who left the company within the last month, some details regarding clients subscription plan and profile are available.
Given the limited amount of explanatory variables available, the model used for prediction show a good ability to identify clients who are going to leave the company.Aside from variables which are kind of obvious predictors of the client's behavior, such as the length of contract or the tenure, the most interesting variable is the internet service, with clients who signed up for the optical fiber showing a tendency to leave the company. It is probably no coincidence that clients whose monthly charges are around the subscription cost for the optical fiber are also more likely to drop, this probably shows a less competitive offer than competitors, at least in terms of value perceived by the consumers. The body of this report will follow a question-oriented structure, starting from the setup of the analysis, the model predictive power usefulness will be established , then the features that provided the most information about clients actions will be highlighted, showing whether they increase or decrease the probability of the client leaving the company. Lastly, the main results will be discussed.

Chapter 2

Analysis

2.1 Predicting customer churn

2.1.1 Methods

Churn analysis is a typical unbalanced classification problem, in this dataset the customer attrition rate is approximately equal to 0.27 . Therefore in order correctly analyze this dataset, both the data splitting (70/30 split) and the tuning of the model parameters through cross validation were carried out through stratification of the response variable.
The method chosen for predicting the outcome was a random forest but in a real scenario it would have been ideal to fit also a gradient boosting, since depending on the true-data-generating mechanism one method may perform better than the other (the former focuses on reducing the variance of the ensemble model whereas the latter focuses on reducing the bias).
The parameters involved in the tuning are:

- Number of features to consider at each split (mtry);
- Maximum depth of the tress ($\max_{\text{depth}}$);

these are usually the most useful parameters to tune (the number of trees is usually set to be as the highest number which allows reasonable computations times). The metric used to evaluate the model performance is the F_2 score, without any real stakeholder feedback, or cost benefit evaluation of the future intervention to reduce customer churn, I simply assumed that to give double the weight to recall rather than precision was reasonable, i.e. it is twice as important to identify true leaving clients than wrongly predict as leaving a client which won't churn.

2.1.2 Analysis

The best model identified though cross validation held the following parameters:

- mtry=4;
- $\max_{\text{depth}}$ =4.

The number of trees was set equal to 200. Afterwards a crossvalidation was carried out to identify the classification threshold that maximizes the F_2 score on the train set. With a threshold approximately equal to 0.38, the model holds a score of 0.74.

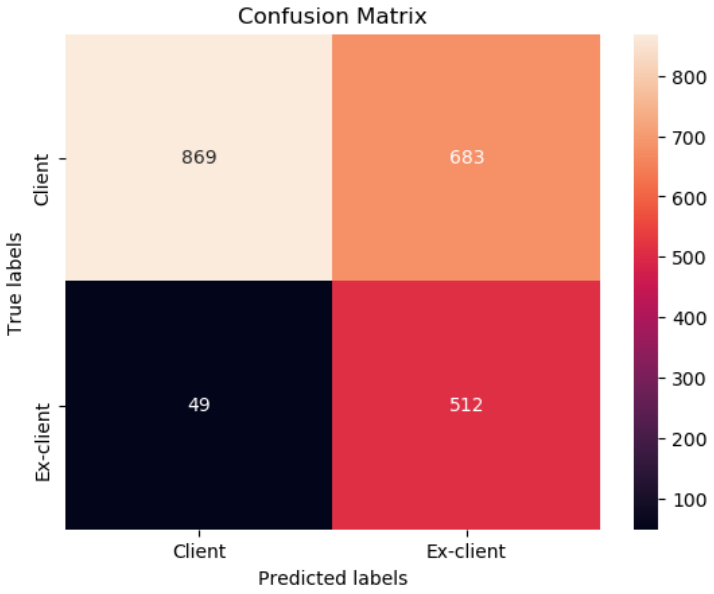

Figure 2.1: Confusion matrix for the optimal threshold

2.1.3 Conclusions

Despite the lack of strong predictors (i.e. information related to the satisfaction of the clients or to the competitors offer) the model seems to perform well, predicting less then 10% of false negatives while managing to correctly identify a client as leaving 43% of the time.

2.2 Evaluation of model predictive ability

2.2.1 Methods

In order to better evaluate the performance of the model, in the foresight of the implementation of a future customer retention intervention, it is useful to look at the cumulative gain chart and the relative lift chart. These graph allow to identify on which percentiles of customers focus on in order to maximize the benefit cost ratio (obliviously prioritizing those clients which are subscribed to the services which ensure a larger profit).

2.2.2 Analysis

The charts below show again a good performance of the model. In a real scenario we could opt to aim for the maximum point of the gain chart, Fig. 2.3 in order to be efficient or stick to the constraint related to budget we could spend on the future intervention. The fourth decile represent the point of diminishing returns, therefore after that point we would obtain smaller marginal gains. That is also the point after which we are able to capture less than twice the amount of clients

prone to leaving than if we were to select clients for the intervention randomly, fig 2.2.

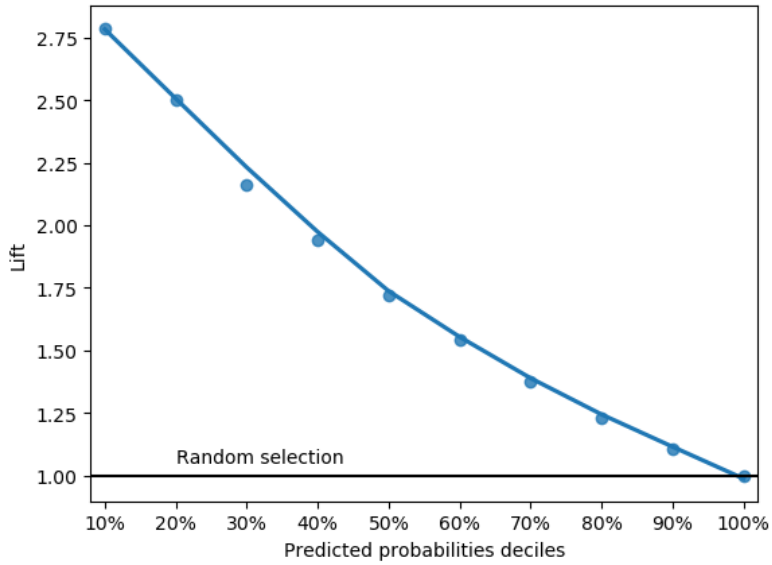

Figure 2.2: Lift chart

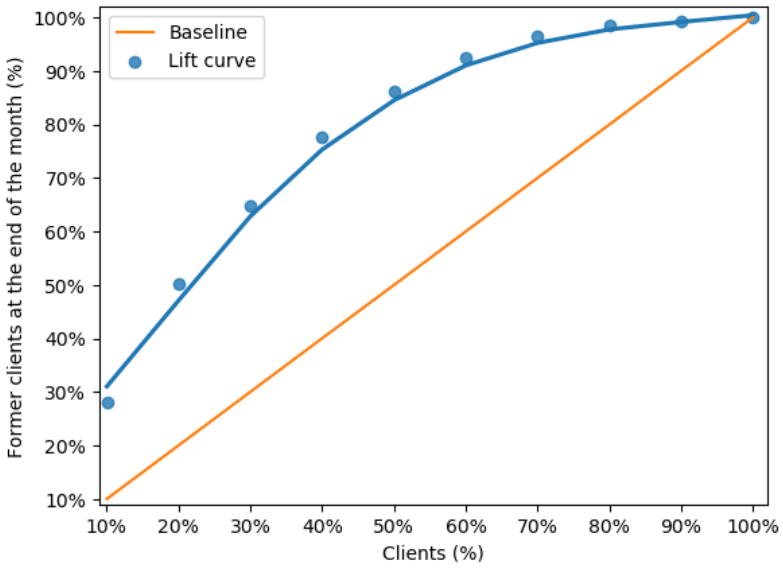

Figure 2.3: Cumulative gain chart

2.2.3 Conclusions

The model shows a good performance up until the fourth decile. On the basis of the strategy followed by the company for client retention, the percentage of

clients to contact may vary drastically, but without any economical information at disposal (without having any available information), selecting the forty percent of clients with the highest predicted probability of leaving the company seems the most reasonable choice.

2.3 Feature importance

2.3.1 Methods

In order to appropriately evaluate in a correct way the features importance in the prediction it is important to apply a (group-wise) permutation to features [1]. The grouping is necessary in order to avoid splitting importance across correlated features, while the permutation introduced in the features allows to measure the drop in the scoring metric and thus their importance in the prediction.

2.3.2 Analysis

To start with a cluster analysis was carried out to create the features cluster (with features having been already encoded). The clusters where created on the basis of dissimilarity, computed starting from a correlation matrix created using the following metrics:

- Pearson's correlation coeffient: for quantitative varible pairs;
- Correlation ratio: for quantitative/nominal variable pairs;
- Cramer's V: for nominal variable pairs;

Afterwards a complete linkage clustering was carried out and by setting the threshold to 0.4, in order to capture most of the encoded categorical features, the groups in Fig. 2.4, were identified.

The results of the subsequent group-wise permutation feature importance can be observed in Fig. 2.5.

2.3.3 Conclusions

The most relevant feature seems to be the length of contract by far. As shown in the next section, among the other features, the internet service is probably the variable which provide the greatest insight on possible changes in the company policy.

2.4 Features effect on customer churn

2.4.1 Methods

The graphs in this section show the distribution of the probabilities predicted from the random forest.

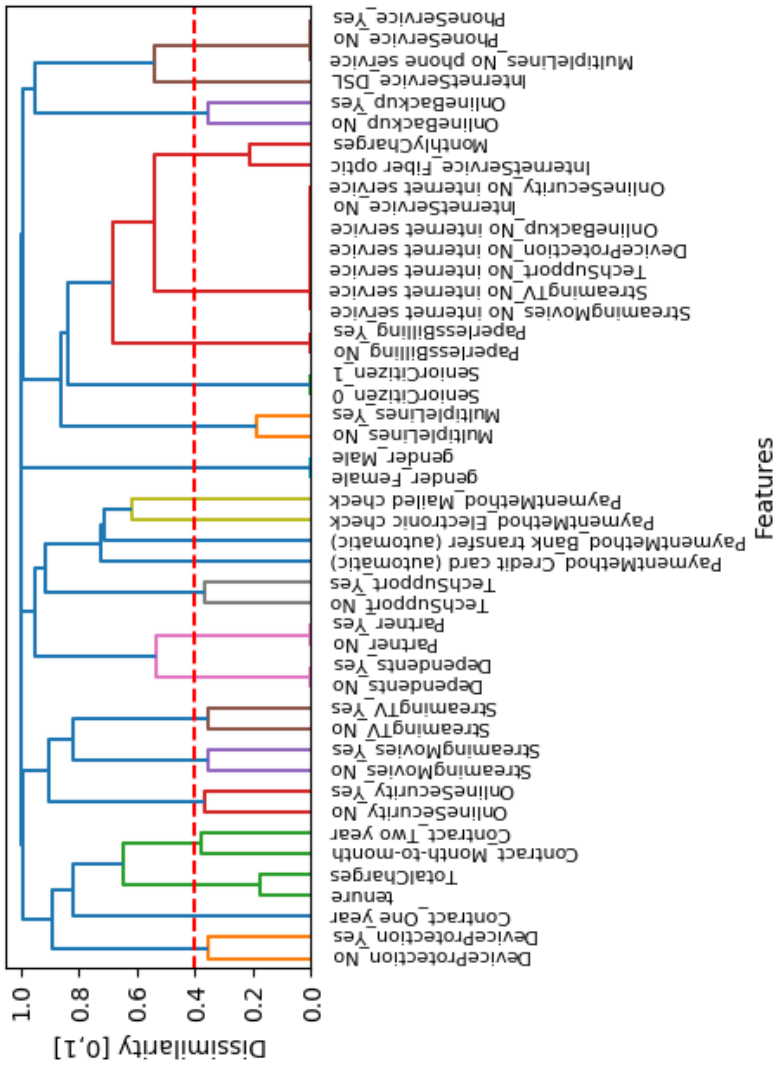

Figure 2.4: Cluster analysis

2.4.2 Analysis

The distribution of predicted probabilities for the kind of contract predictably shows that the shorter the contract the more likely will be for him to drop, Fig 2.6. It is instead more interesting to look at the distribution of internet service and monthly charges (the two are highly correlated). Client who happen to be subscribed to fiber optic service are more like to drop than others, Fig. 2.7 (the one with no internet are probably old people who are usually really unlikely to churn). These clients have an average price per month of 91.27 ($\pm$12.70), which happens to fall in the range of prices in which clients are also more likely to drop, Fig. 2.8. For space reasons the other graphs are available at https://churn-project.herokuapp.com/

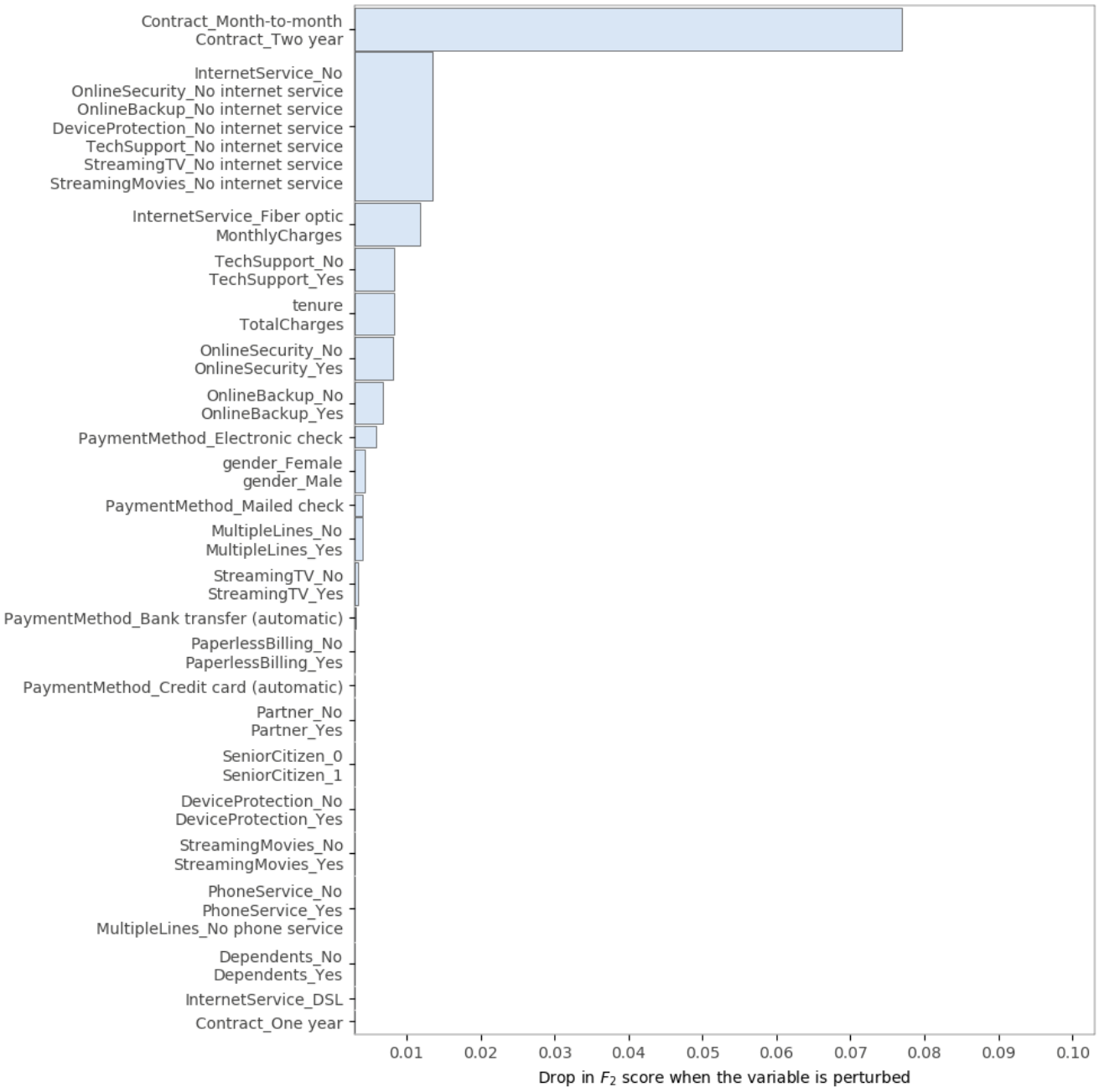

Figure 2.5: Cluster analysis

2.4.3 Conclusions

The results show a tendency of client subscribed to the fiber optic service to leave the company, this may be due to a not competitive price/offer with respect to other internet service providers. It would be advisable to look deeply into the matter, one way to do so would be to recover information about others company offer, and try to evaluate where the production efficiency of the company stands with respect to others, using methods such as DEA (data envelopment analysis) [2].

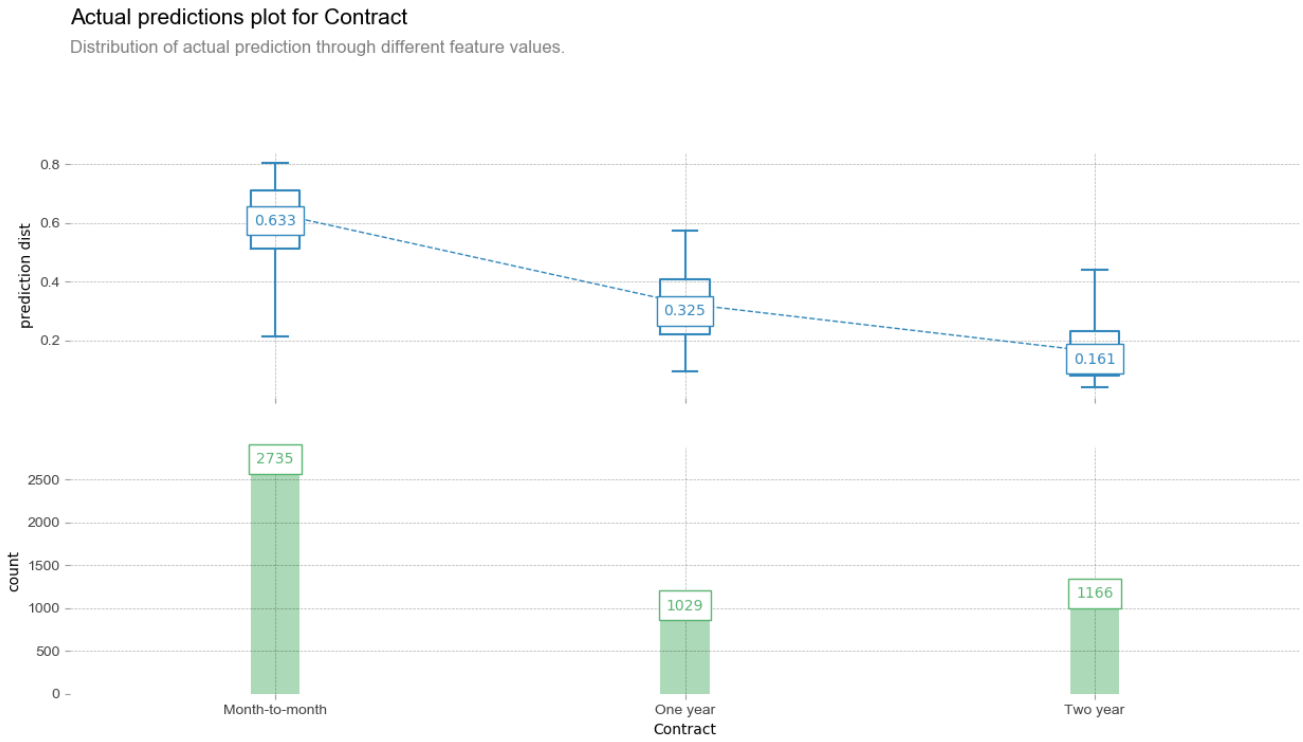

Figure 2.6: Boxplots of the predicted probabilities for different kinds of contract

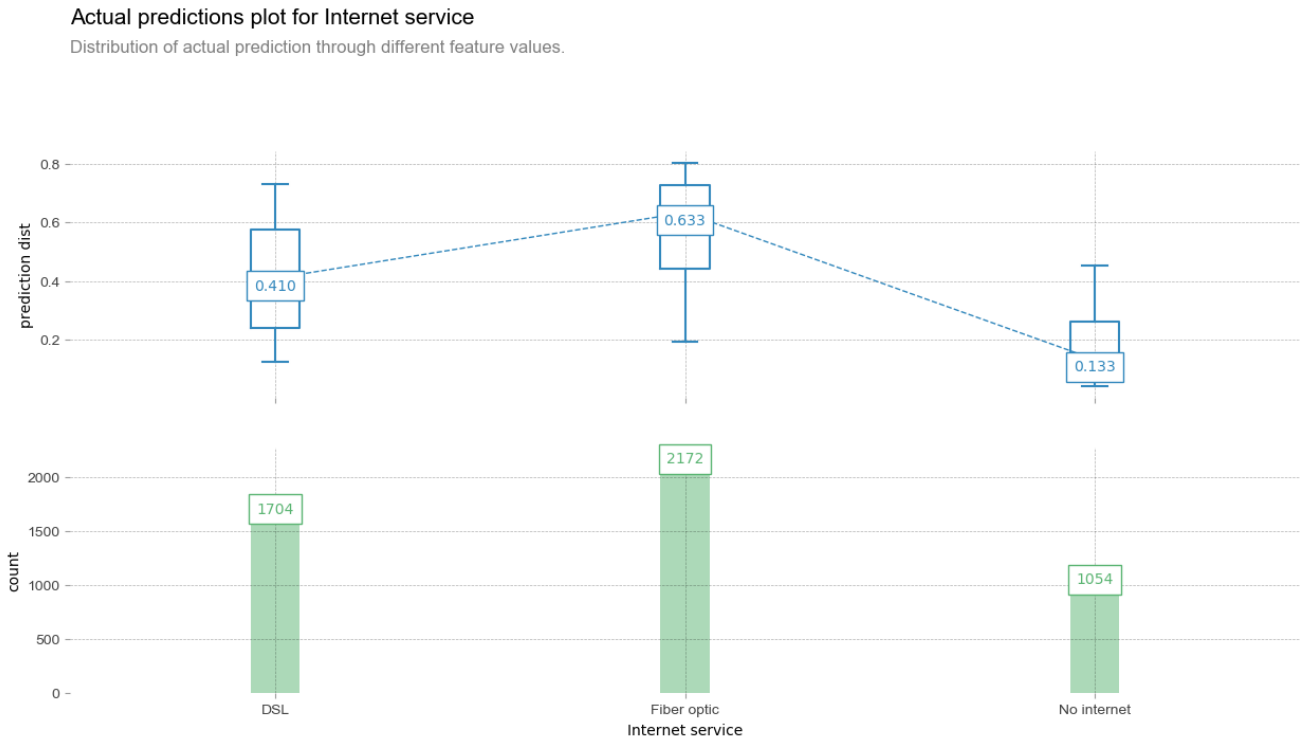

Figure 2.7: Boxplots of the predicted probabilities for the internet service

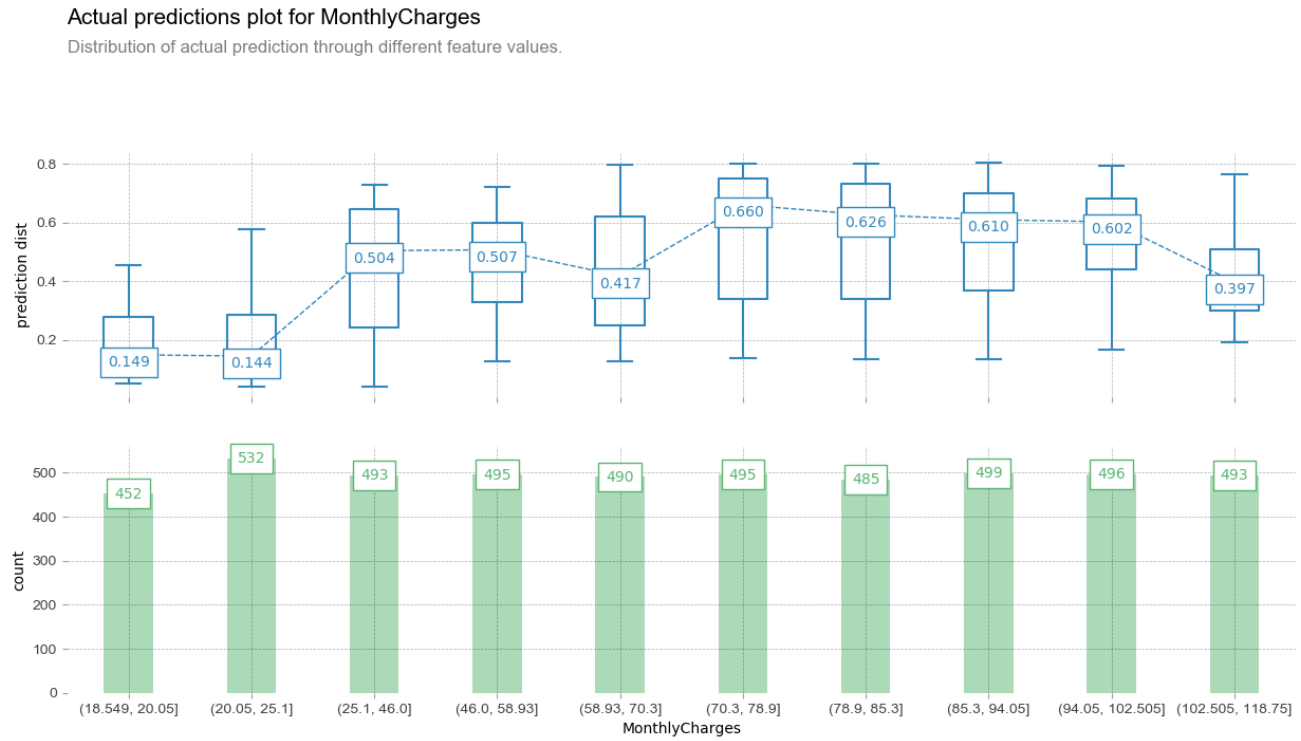

Figure 2.8: Boxplots of the predicted probabilities for the monthly charges by deciles

Chapter 3

Conclusions

Despite the lack of tangible information about clients reasons for leaving the company, the model was able to capture most of the clients who left the company at the end of the month (91%) while avoiding to classify too many clients as leaving when they actually did not (57%). These results would probably already allow the implementation of several interventions aimed to offer some bonus content to this clients while agreeing to extend the contract length, barring any too expensive cost/effiency intervantion, this may be a good solution to tackle the problem. Another idea would be to look at other companies offer in order to underhand where ours company offer stands in terms of efficiency, in order to either adjust the costs sustained by the company or improve the quality/price of the service.

References

[1] André Altmann et al. “Permutation importance: a corrected feature importance measure”. In: *Bioinformatics* 26.10 (2010), pp. 1340–1347.

[2] Abraham Charnes, William W Cooper, and Edwardo Rhodes. “Measuring the efficiency of decision making units”. In: *European journal of operational research* 2.6 (1978), pp. 429–444.

[3] Telco. *Customer churn sample*. `https://www.kaggle.com/blastchar/telco-customer-churn`, visited 07/07/2020.