Customer churn analysis

Report by:
Mauro Gioè

Chapter 1

Introduction

This report main goal is to answer the questions that naturally arise from a customer attrition analysis, such as "Which clients are going to leave the company in the next future?" and "Which features help me the most in predicting such behavior?". The dataset, provided by IBM, contains information about the clients of a telephone company and has been uploaded on Kaggle [3]. Additionally to the information about clients who left the company within the last month, some details regarding the clients' subscription plan and profile are available.
Given the limited amount of explanatory variables at disposal, the model used for prediction show a good ability to identify the clients who are going to leave the company. The most interesting result regards the internet service provided by the company, with clients who signed up for the optical fiber showing a tendency to leave the company. This might be a signal that the service provided by the company is less competitive than ones that are available on the market, at least in terms of value perceived by the consumers. The body of this report will start by introducing the model used for the prediction, then after evaluating its predictive power, the most important features in the prediction of customers' behavior will be highlighted. Lastly, the main results will be briefly discussed.

Chapter 2

Analysis

2.1 Predicting customer churn

2.1.1 Methods

Churn analyses are 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 and the tuning of the model parameters, which was carried out through cross-validation, were carried out with stratification of the response variable.
The method chosen for predicting the outcome is a random forest but in a real scenario, it would have been better also to fit 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 more important to identify true leavers than to wrongly predict a client as leaving when he had not such intention.

2.1.2 Analysis

The best model, identified though cross-validation, holds the following parameters:

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

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

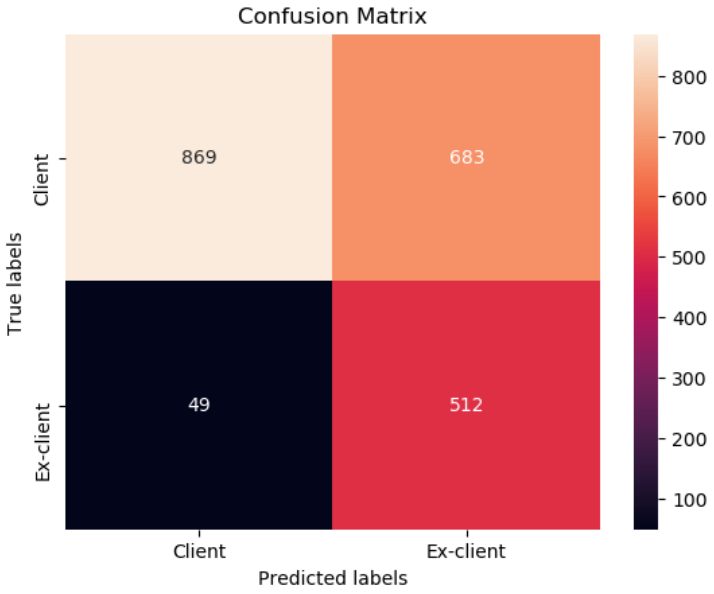

Figure 2.1: Confusion matrix for the optimal threshold

2.1.3 Conclusions

Despite the lack of strong predictors (i.e. information related to the satisfaction of the clients or to the offer of the competitors), the model seems to perform well, predicting less than 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 evaluate the performance of the model, in foresight of the implementation of a customer retention intervention, it is useful to look at the cumulative gain chart and the relative lift chart. These graphs allow the identification of the percentiles of customers on which is better to 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 a good performance of the model. Depending on the intervention to run, we could opt to aim for the maximum point of the gain chart, Fig. 2.3, in order to be as efficient as possible or reach as many clients as the budget allows. In this case, the fourth decile represents 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 intending to leave than if we were to select clients for the intervention randomly, fig 2.2.

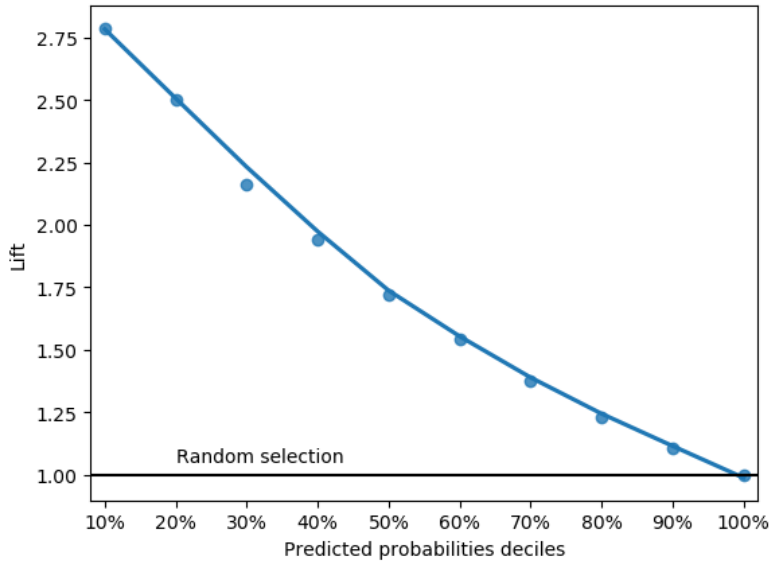

Figure 2.2: Lift chart

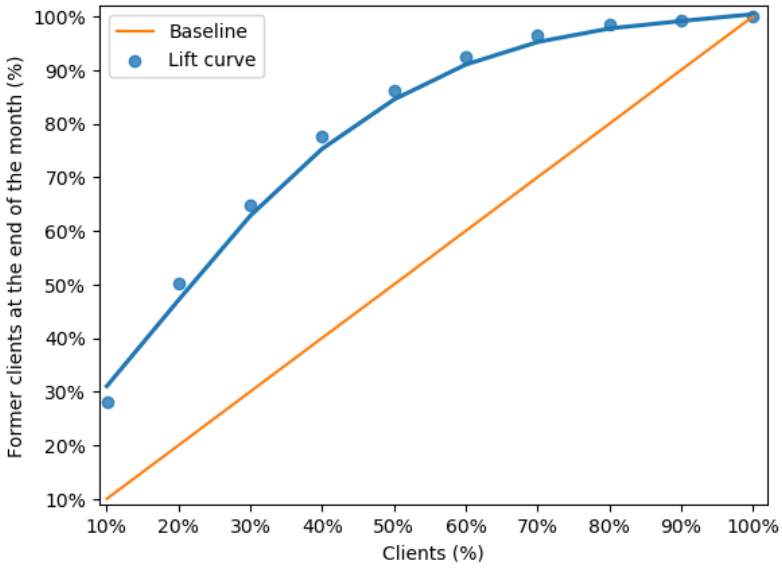

Figure 2.3: Cumulative gain chart

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, selecting the forty percent of clients with the highest predicted probability of leaving the company seems to be the most reasonable choice.

2.3 Feature importance

2.3.1 Methods

In order to appropriately evaluate the importance of the features 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 feature clusters (with features having been already encoded). The clusters where created on the basis of dissimilarity, computed starting from a correlation matrix made up by the following metrics:

- Pearson's correlation coefficient: for quantitative variable 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 provides the greatest insight on possible changes in the company business strategy.

2.4 Effect of the features on customer churn

2.4.1 Methods

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

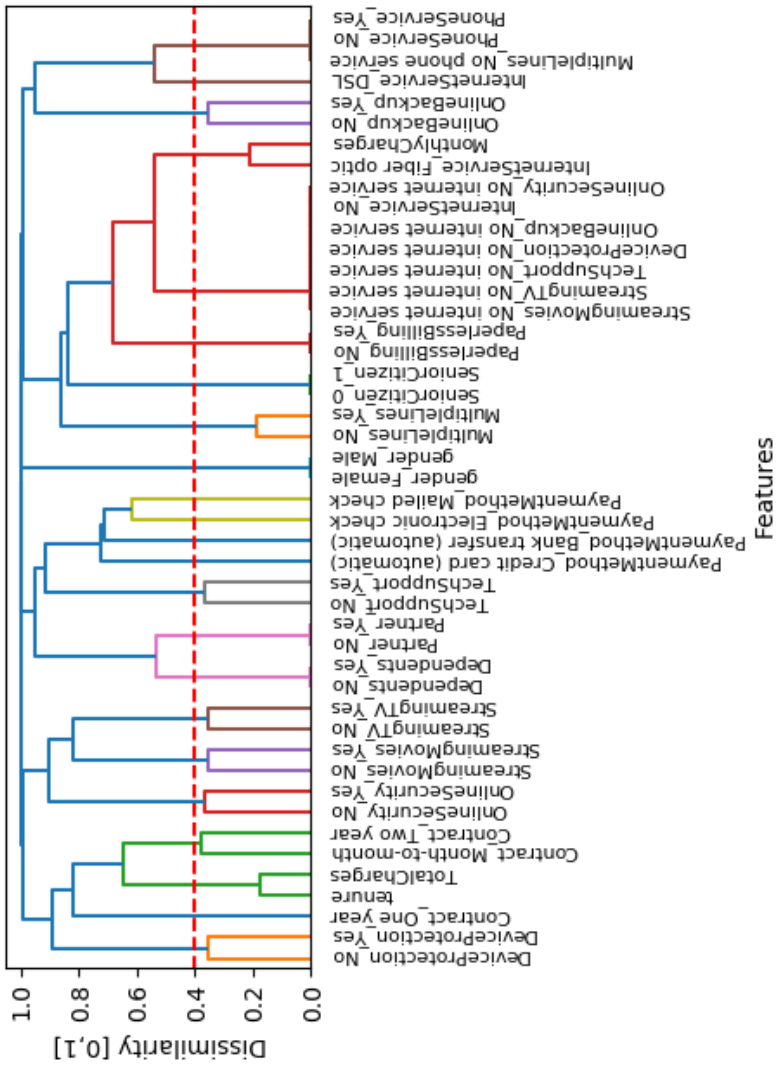

Figure 2.4: Cluster analysis

2.4.2 Analysis

The distribution of predicted probabilities for the kind of contract shows that the shorter the contract the more likely will be for a client to churn, Fig 2.6. This results is quite predictable, it is instead more interesting to look at the distribution of internet service and monthly charges (the two are highly correlated). Clients who are subscribed to the optical fiber service are more like to churn than others, Fig. 2.7, (the ones without 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 within the range of prices in which clients are also more likely to churn, Fig. 2.8. For space reasons the other graphs are available at https://churn-project.herokuapp.com/

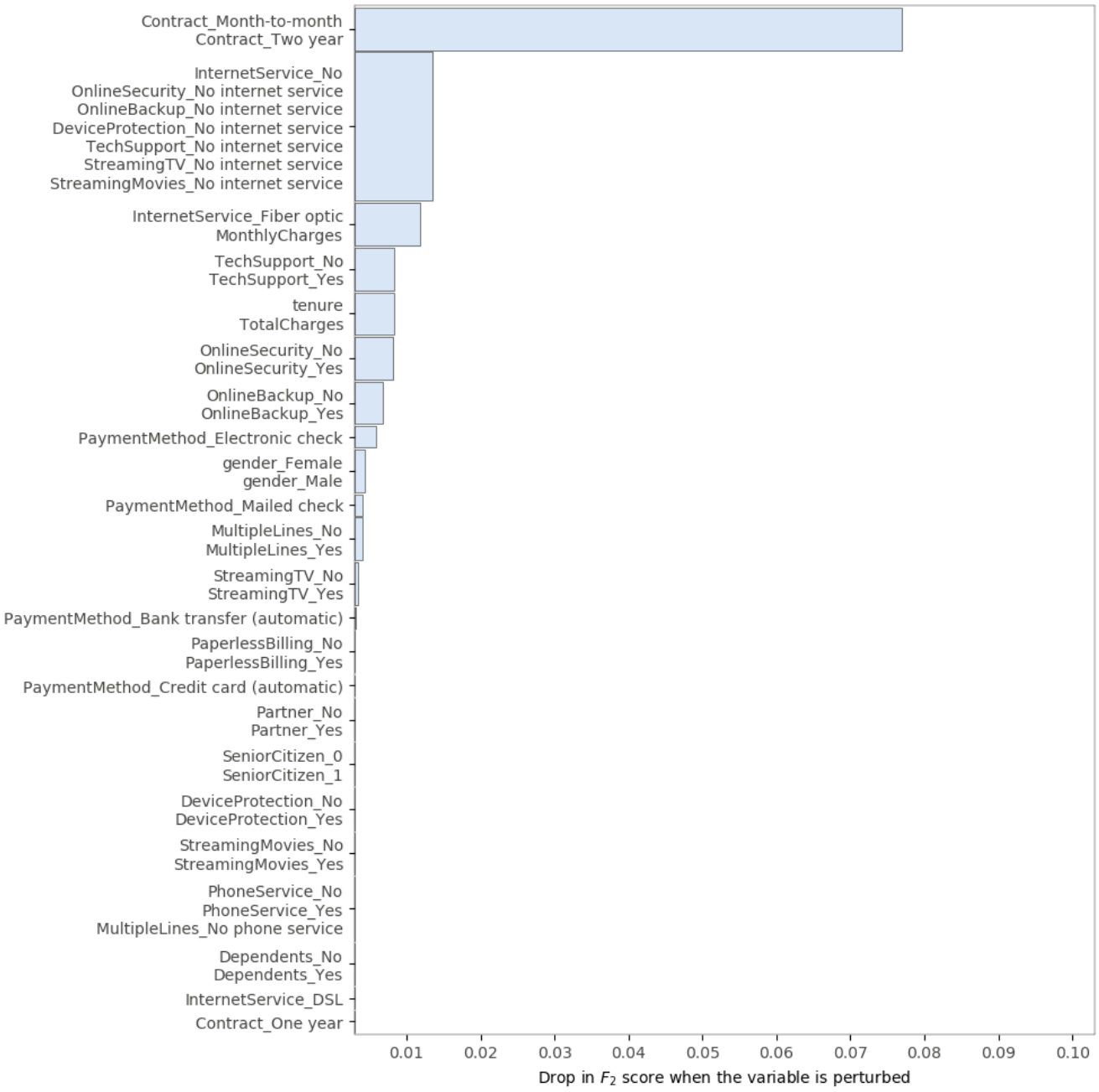

Figure 2.5: Group-wise permutation feature importance

2.4.3 Conclusions

The results show a tendency of clients subscribed to the optical fiber 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 the offers of other companies, and try to evaluate where the production efficiency of the company stands with respect to them, using methods such as DEA (Data Envelopment Analysis) [2].

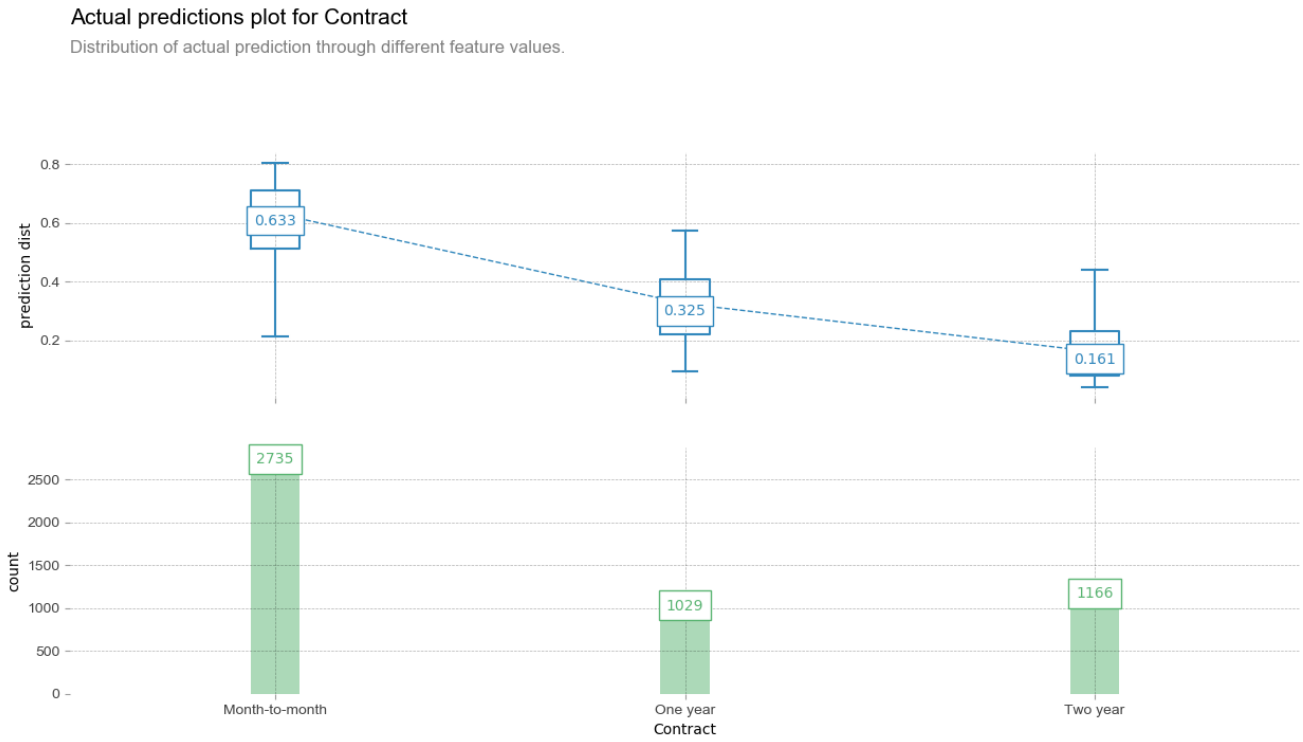

Figure 2.6: Boxplots of the predicted probabilities for different kinds of contract

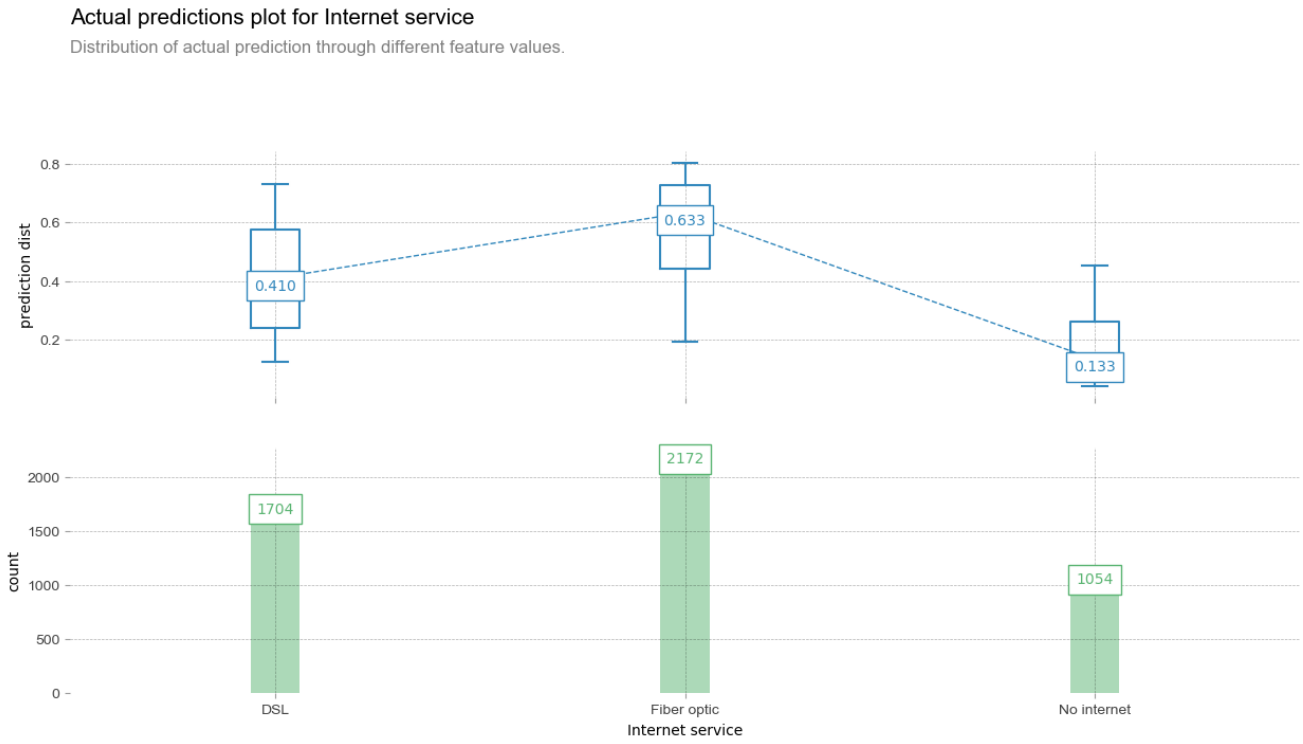

Figure 2.7: Boxplots of the predicted probabilities for the internet service

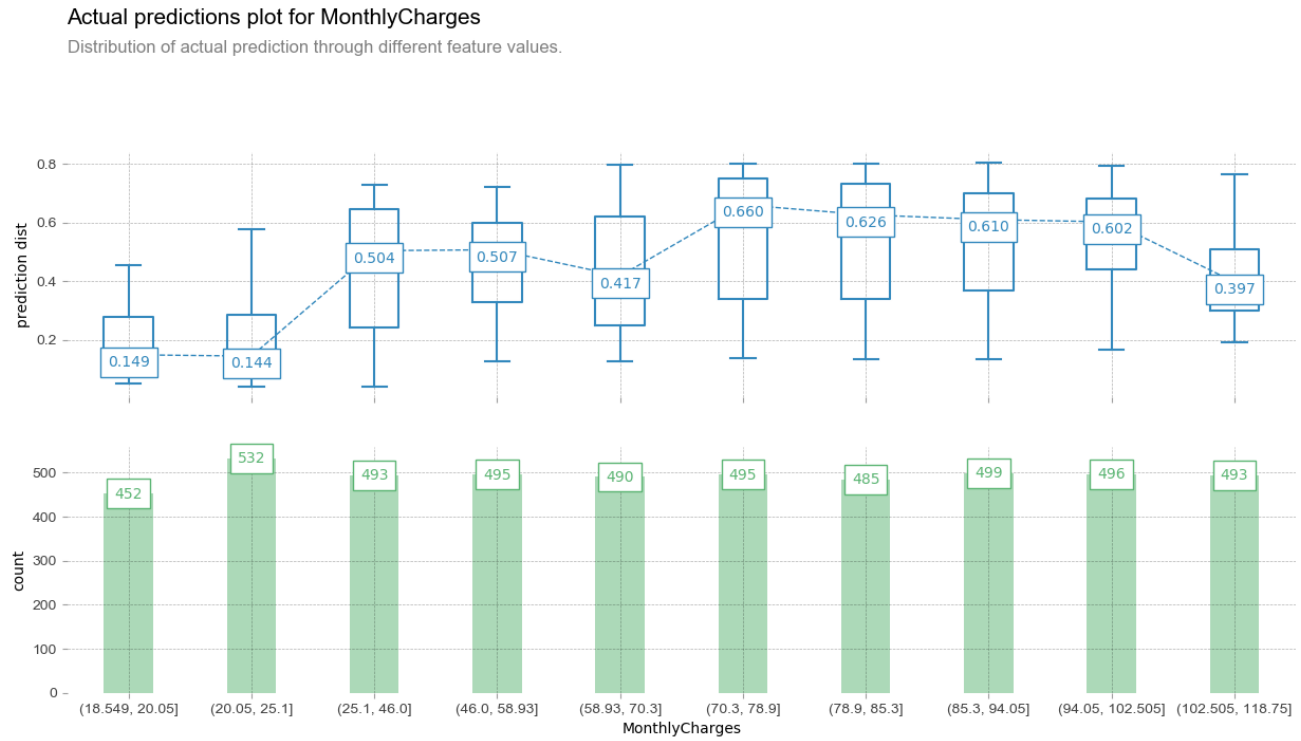

Figure 2.8: Boxplots of the predicted probabilities for the monthly charges by deciles

Chapter 3

Conclusions

Despite the lack of tangible information on the reasons for which clients leave the company, the model is able to capture most of the clients who left the company at the end of the month (91%), while avoiding to commit too many errors when classifying clients as leavers (57%). These results could allow the implementation of several interventions that aim to offer some bonus content to the clients identified as leavers in exchange to an extension of the contract length. Focusing the attention on the optical fiber service, clients tendency to leave, show some dissatisfaction with the service either in terms of quality or price. It could be useful to look at the offers of other companies, in order to understand where the offer of our company stands in terms of efficiency, so to improve the quality (or reduce the price) of the service, or reduce the costs sustained by the company for its implementation.

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] IBM. *Telco customer churn dataset*. `https://www.kaggle.com/blastchar/telco-customer-churn`.