Churn Prediction Model

Introduction:
In this task, we set out to find a set of features that have a great impact on the cancellation of a reservation. After choosing a model that we were satisfied with, we calculated the correlation coefficient of each feature to the response. The top 15 features appear in *Figure 1*. 15 is too large a number therefore we decided to focus on the 5 most important features as seen in *Figure 2*.

Features:
You may observe that there are three features that represent almost identical categories. Moreover, they were derived from the same original feature, as such we will explain the relevance of their mutual category.

These features, *cancellation_policy_<period>*, represent the penalty given to a customer for canceling their reservation during the specified period. It is no surprise that the original feature, *cancellation_policy*, has such an impact on the result, considering it represents the amount of money one would pay for canceling their reservation. After observing the data, we realized that there are uncountably many different policies. As such we decided to break them down into the four most relevant categories. Doing so, three of these four features remained highly impactful on the response, as can be seen in *Figure 2*.

The next feature we will discuss is *no_of_adults*. Once again, it is easy to see that the number of adult guests is significant enough to affect the result. Indeed, the more people involved in a reservation gives it a measure of stability, especially since single cancellations may be replaced.

Finally, we will explore the feature *time_ahead*. This feature describes the number of days between the booking and the check-in dates. During our discussions we concluded that the-

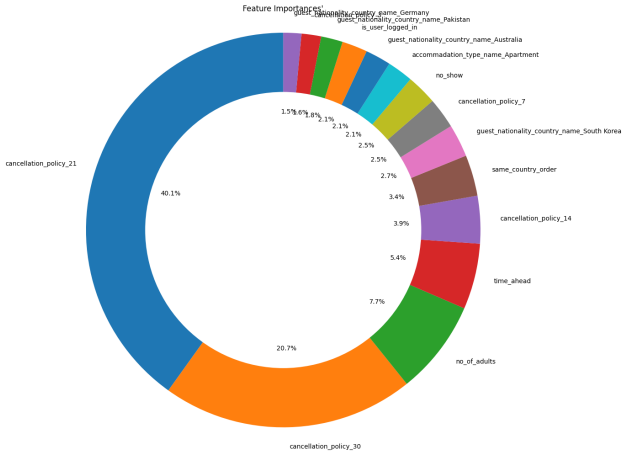

Figure 1. Top 15 features

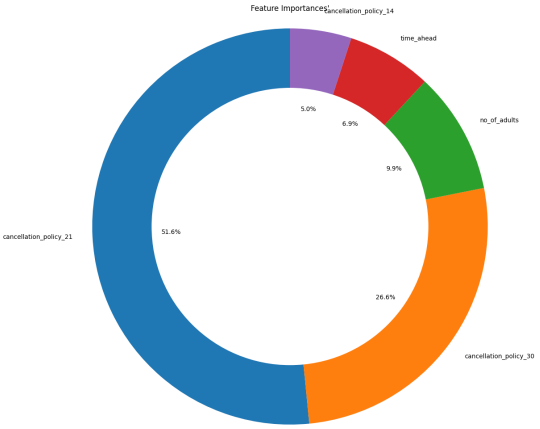

Figure 2. Top 5 features

-longer the time, the higher the chance for cancellations. Thus, we conclude that this feature has a direct impact on the response of our model.

Performance

Previously, we claimed that there were five features that were the “most important”. In order to prove said claim, and to show their enormous contribution to our conclusion; we will explore the performance of our chosen model as seen in *Figure 3*. We will do so with different sets of features over different training-set sizes, in order to emphasize the magnitude of their importance.

It is clear that when fitting the model with all of its features, the best results are achieved. This is no surprise since our model uses much more relevant data to learn from. Even so, using only the top five features, we get a model that is minimally worse than our best result, with only a fraction of the data. This goes to show that our five best features have the greatest impact over our prediction. Emphasizing their importance even further, the last model shows the results we receive when excluding the top five features, resulting in an extremely bad result.

Lastly, *Figure 4* once more proves that using only the top 5 features for training, we get sufficiently good F1 results when predicting the test set and the train set. Clearly, the score we receive when predicting over the training set is much higher than that of the test set, which is expected, considering we abuse already seen data to fit the model.

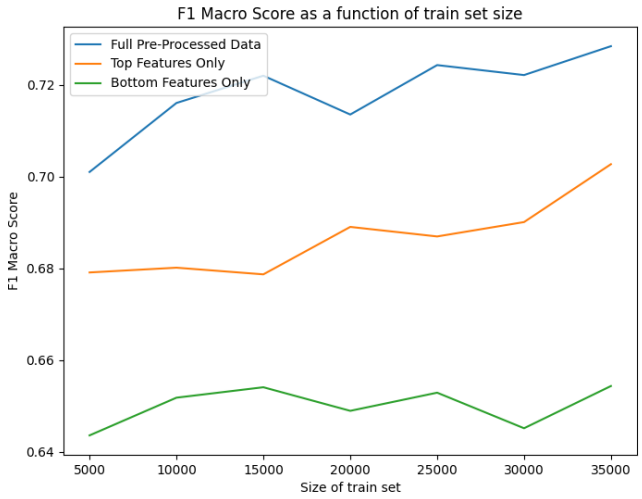

Figure 3. Performances over different sets of features

Figure 4. F1 Score of the Test & Train sets over an increasing percentage of training data