[**EX 11**][**REPORT**] Aran Coll

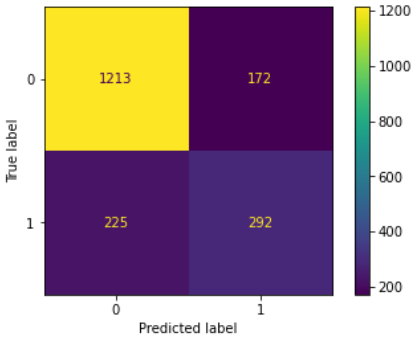
Calculate the best model and define the process you have followed to achieve it. Some questions to answer are:

- Which variables you used and why to train the model.
- Hyperparameters used along the different trainings.
- Define how the cutoff has influenced your decision. Insert the different density charts you have used

As seen in class, the best model is the one that better fits  by understanding our use case. In this seminar, we are a Telecommunication company that has noticed an increasing volume of churners. To prevent that we study the data in order to let the Marketing team know the potential customers that want to give up on our company.

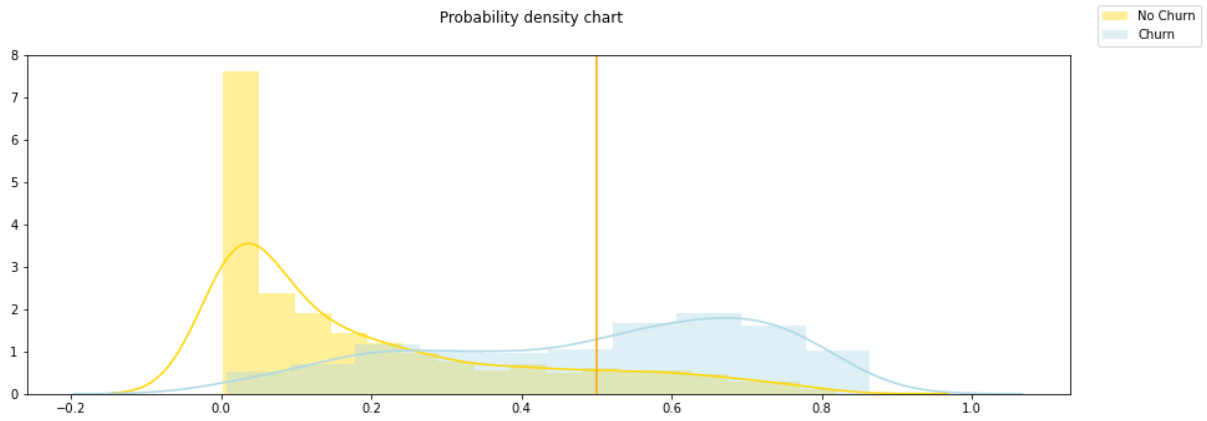
The values that we will use to determine the quality of our model are the precision and recall, extracted from the Confusion matrix, and modified according to the decisions taken from visualizing the Probability Density Chart and the Feature Importance bars.

Our initial conditions, using LogisticRegression with the liblinear solver, all the columns, and a 0.5 value as cutoff, are these:




Classification report:

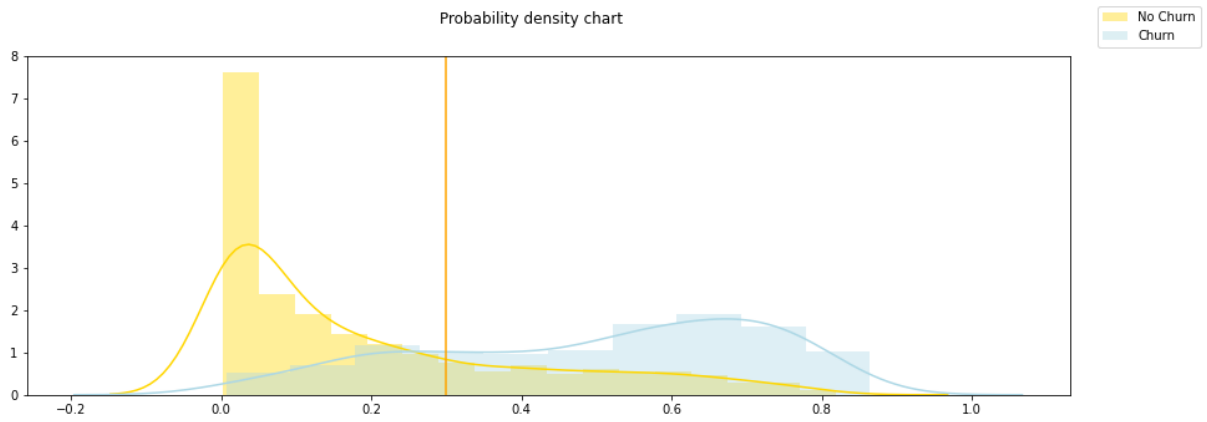
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.84 | 0.88 | 0.86 | 1385 |
| 1 | 0.63 | 0.56 | 0.60 | 517 |
| accuracy | | | 0.79 | 1902 |




Let's first assume that the Marketing team comes up with an amazing idea which is not expensive at all, and soon enough we would have benefits from those companies that stayed. In other words, the prediction of churners does not have to be very precise since we want to reach as many as possible.
For these reasons, we decrease the cutoff value down to 0.3 and we get these values:

Classification report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.75 | 0.82 | 1385 |
| 1 | 0.54 | 0.76 | 0.63 | 517 |
| accuracy | | | 0.76 | 1902 |



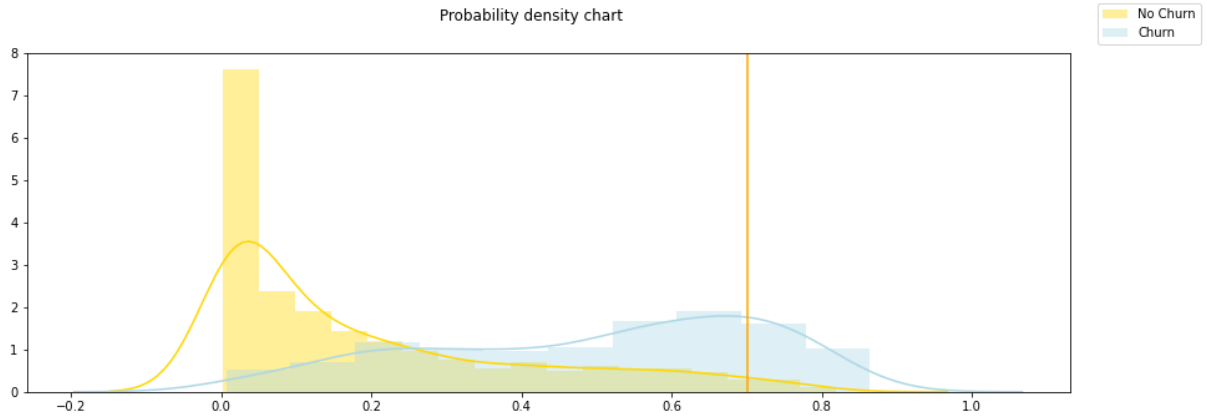

The recall, the relevant instances from the total that are retrieved, increases considerably from 56% to 76%, so more companies that were churners are identified and thus, marketing campaigns can change their decisions. The cost of this is the precision, since almost only

half of the instances retrieved are relevant churners. The accuracy of the model stands high as before.

In this other case scenario where we have to better filter the potential churners to focus the marketing logistics, we will have to adjust our model to get better results.

Classification report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.81 | 0.93 | 0.86 | 1385 |
| 1 | 0.69 | 0.40 | 0.51 | 517 |
| accuracy | | | 0.79 | 1902 |




We put the cutoff to 0.6 to increase the precision without lowering the recall too much (since with a cutoff of 0.7 the recall goes down to 0.22). By doing this, the model has a better precision identifying the churners and thus, the marketing campaigns can be more personalized and effective on those companies that want to leave us. However, the price to pay is that only 4 every 10 churners are detected.

Other aspects can also be taken into consideration.
The gender and Partner variables are oftentimes the less important features (absolute value) and can be dropped off the table. Also those variables with an importance value larger than |0.25|, explain most part of the model, and therefore, the others can be dropped off too.
For the hyperparameters of the model, the number of iterations can be reduced to more than the half since it converges quite quickly; and the solver does not provide relevant changes in the output results.

To prove my points, with only 11 features the precision and recall values stay really similar to a cutoff of value 0.3. The probability of well predicting a churn gets narrower to less than 0.8, as seen in the Probability density chart .

Classification report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.74 | 0.81 | 1385 |
| 1 | 0.52 | 0.77 | 0.62 | 517 |
| accuracy | | | 0.74 | 1902 |