Customer brand preference

1. Modelo utilizado y justificación.

The best model that adjusts to the predictions is the Random Forest, however, to optimize the predictions and improve the score, a Bagging is combined, which is a boostrap model that randomly replicates the random forest X number of times to optimize the predictions. It is important to clarify that this is achieved because the data is not massive and does not require so much computational load. This exception is important because for more extensive data the model would have such a high computational load that it would exceed the computational hours, which is inefficient. For the designed case, it manages to make 5000 predictions in 1.3 min.

¿Why is the best model?

Scores	Bagging Classifier	Random Forest	Gradient Boost
Accuracy	0.926	0.924	0.911
Classification Matrix	[[2534 274] [278 4338]]	[[2534 274] [287 4329]]	[[2543 265] [390 4226]]

It can be seen in the table above that in accuracy the Bagging Classifier is the highest value and has a balance in the false negatives and positives that were predicted. This makes the Bagging Classifier model the best choice.

It is important to clarify that other methods such as clustering and Logistic Regression were modeled but did not obtain satisfactory results.

On the other hand, the ROC AUC scores, the ROC curve and the log of this model are calculated since it is not enough to be able to determine a good model only by its precision.

To understand these terms a little, they are scores used in binary classification problems such as the problem in execution.

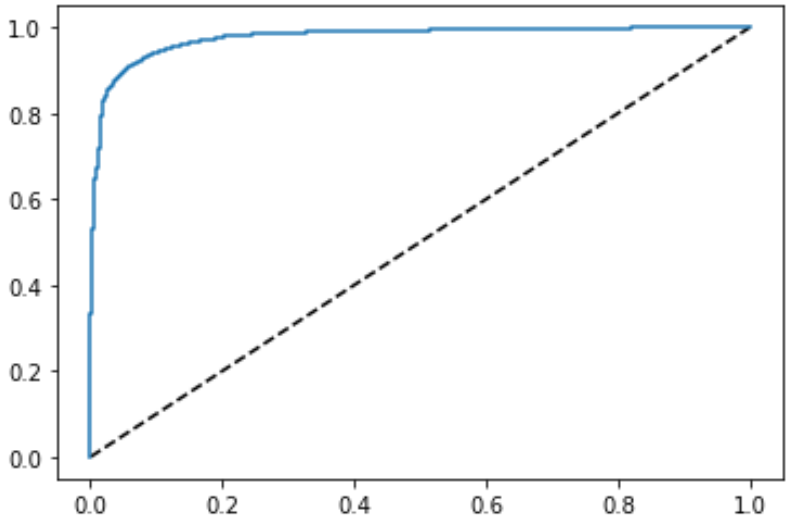
The log los is defined as the distance between the probability of the prediction with respect to the real value. This means that the closer the probability of the predictions, the lower the log los and as a consequence we have a model with a very good predictive capacity.

The ROC AUC is an indication of the separability or distinction that can be seen between the two classifications. The more this distinction can be appreciated and the smaller the overlap between the classifications, the higher the ROC-AUC will be and as a consequence we will have a more efficient model.

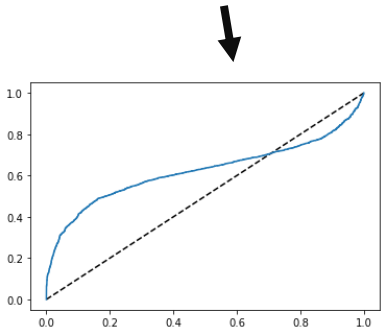
Finally, the ROC curve is nothing more than all the possible ROC values according to the true positive values with the false positives. This means that the larger the area under the curve, the more optimal the model.

The results of the Bagging model are shown below.

Area under the high curve of our model



This shows an inefficient model with a low and negative area under the curve: attempt Logistic regression of the problem at hand.



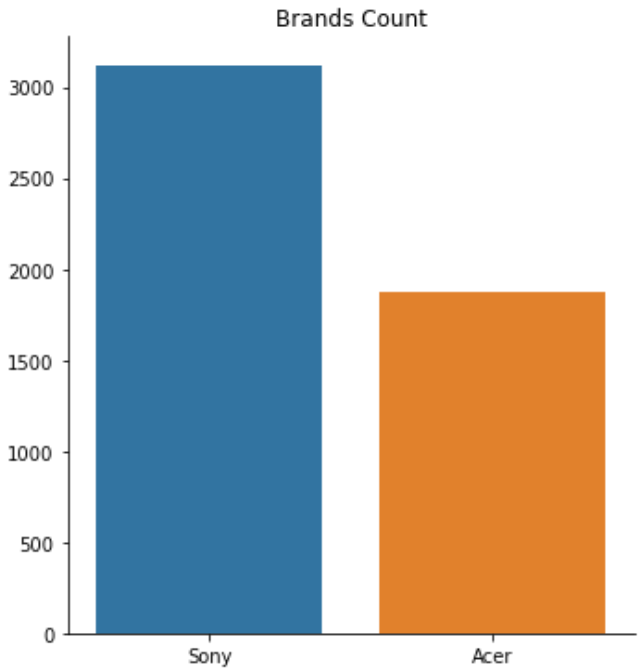
Finally, the log Los and ROC AUC values were 0.20 and 0.976 respectively.

With this we can conclude that the model has quite high standards and can be used as a reliable predictor to complete incomplete data.

Below is an excerpt of the data completed with the model, as well as the probabilities of each value to detail the choice of that value.

	salary	age	elevel	car	zipcode	credit	brands	proba_1	proba_0
0	150000.00000	76	1	3	3	377980.10160	1	0.000107	0.999893
1	82523.83897	51	1	8	3	141657.60660	0	0.821385	0.178615
2	115646.63620	34	0	10	2	360980.35850	1	0.035645	0.964355
3	141443.39330	22	3	18	2	282736.31910	1	0.000367	0.999633
4	149211.27030	56	0	5	3	215667.28960	1	0.000536	0.999464
...	...	...	...	...	...	...	...	...	...
4995	83891.55966	52	2	14	5	28685.22963	0	0.810730	0.189270
4996	125979.28910	71	0	12	7	276614.82930	1	0.117844	0.882156
4997	74064.71053	24	2	2	2	202279.57880	0	0.980576	0.019424
4998	106485.56710	46	3	16	0	381242.08810	0	0.938850	0.061150
4999	50333.57979	70	1	5	5	224871.17020	0	0.910549	0.089451

Finally, we can observe the count for each brand and we can see that out of the 5000 customers, more than 3000 prefer the sony brand.



The data is stored in a CSV named Survey.

In case you want to review the code used and the process using the Python tool check Github