BSDS November Hackathon

Buongiorno a tutti!

Bocconi Students for Data Science is hosting its first Machine Learning focused hackathon of the academic year! We are really looking forward to it and we hope that this challenge suits your interests and skills. Anyway, less of the chit-chat and let me explain to you what this challenge is all about.

If my colleagues were diligent, you should have received an email on Monday 16/11/2020 in which, along with some important info, a file called *BSDS_November.csv*. If you try to open that file, probably with excel, you can see a dataset, more specifically a training dataset (if you don't know what that is don't panic, i'll explain that to you in a sec). So, what do you have to do with this dataset? What the heck is that about?

This database reports data related to a "**Churn analysis**" for a *Telecommunication* company. You may already know about it, but churn analysis studies the churn rate of a company, where the churn rate represents the portion of customers that leave that specific company. Therefore, it's fairly easy to understand why firms need and strive to have relevant information about why, how and when a customer leaves the company, the most immediate one being trying to retain that customer and prevent that to happen in the future with another customer.

Let's try to dive into this dataset and let me explain the meaning of the following columns.

1) "gender", Whether the customer is a male or a female
2) "Partner", Whether the customer has a partner or not (Yes, No)
3) "PhoneService", Whether the customer has a phone service or not (Yes, No)
4) "InternetService", Customer's internet service provider (DSL, Fiber optic, No)
5) "OnlineSecurity", Whether the customer has online security or not (Yes, No, No internet service)
6) "Contract", The contract term of the customer (Month-to-month, One year, Two year)
7) "PaymentMethod", The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)
8) "MonthlyCharges", The amount charged to the customer monthly
9) "TotalCharges", The total amount charged to the customer
10) "Churn",Whether the customer churned or not (Yes or No). This represents the label that your model should try to predict, so you'll find yourself facing with a **supervised machine learning problem** (supervised means that you know what you're trying to predict, i.e. the "Churn" label).

What you're asked to do is to try to deploy a machine learning algorithm in order to train a model on the set we will give you trying to predict as precisely as possible if the customer will or won't leave the company in the next future. The "training set" that you have been given has all the 10 columns described above, while the "test set" that I, the keys-keeper, greedly enshrine in my computer, will miss the label column, the one called "Churn", and on this dataset your ML model will be tested.

As a strategy I advise you to split the dataset into two different dataset, one being the test set (where the churn label should be omitted) and the other the training set where the actual models are trained. Having done that you can start to implement different models and find the best hyperparameters to put in that model. As you probably saw in last week's Oracle meeting there are many algorithms, such as Support Vector Machines, Regression Trees and many more, that can be used as techniques in supervised machine learning.

Who's the winner then? The group that will send us the most accurate model. To grade that, we will perform a simple loop on the label column of the test dataset with your predictions and the actual labels (i.e. the values of the "Churn" column) that checks what portion of the labels you got right. Moreover, the winners, along with envy and admiration from their fellow associates, will receive some space on our social media platforms, especially Linkedin, and, if they want (but I'm sure they will) they will receive personal congratulations from the best ever associate in the world Vittorio Costa (he is so dreamy my god).

Lastly, I have to tell you the logistics of this project. We will require you to load the dataset we have given you on Jupyter and then send us your Jupyter notebook where your model is explained. If you don't already have Jupiter on your computer, I recommend you to download it from Anaconda Navigator. Before the deadline, which is **midnight of Sunday 30/11/2020**, you must send your work to this email: "vittoriocosta281099@gmail.com". The email object has to be "*BSDS November - name of your group* " (or your surname if you're working alone). Moreover, the jupyter notebook should be called "*BSDS_November_nameofyourgroup*" (and the same as before for those working alone).

So, let me say good luck to you, in bocca al lupo, buena suerte, daje e tutt e cos. Should any concern arise, don't mind contacting the organizers through their mails, which are specified below.

BOCCONI STUDENTS FOR DATA SCIENCE

Organized by Vittorio Costa (vittoriocosta281099@gmail.com) and Federico Fiorani (federico.fiorani@studbocconi.it) with the supervision of Mert Tekdemir (mert.tekdemir@studbocconi.it)