ST4240: ASSIGNMENT 1

Semester 2: 2017-2018

1. Please, write your group number as well as the matriculation number and name of each member of the group. Note that your group number may have been updated: you can double check your group number **here**.
2. The assignment is to be uploaded on the IVLE by the **1st of April**, 11:59pm.
3. Late assignment will not be accepted. Only typed **pdf** files can be submitted.
4. if your group number is XX, please name your file:

   `assignment_1_XX.pdf`

   For example, if you are submitting a report for the group 13, your file should be named
   `assignment_1_13.pdf`

5. Only one report per group is submitted.
6. You can use any programming language you deem appropriate.

The assignment is composed of three independent exercises consisting in building predictive models for classification and regression problems.

- Describe your approach for building your predictive models. Your report should be **brief** but should also contain all the steps used in your approach. The quality of your report will be an important part of your mark. It should be clear and to the point.
- You are indeed allowed to use any statistical model/approach you deem appropriate – but you should describe why you think your approach is correct. Using a very fancy / complicated model when a more simple / robust one works as well is generally not a good idea. Only use models that you do understand. You will be penalized for using a model wrongly.
- Do describe all the approaches that you have tried, even the ones that were not succesfull / useful.
- If you do feature engineering, and you are encouraged to do so, please describe the approach in details.
- Data cleaning and exploration are usually important steps. Give details of your approach.
- For all three exercises, you will give an estimate of the performances of your algorithm. (eg. use cross validation)
- Be independent, read widely, and remember that google is your friend ...

1. Rossmann Store Sales

You are provided with historical sales data for 1115 Rossmann stores. The task is to build a statistical model for predicting the "Sales" volume. Training data available at at **here**.

2. Give Me Some Credit

Banks play a crucial role in market economies. They decide who can get finance and on what terms and can make or break investment decisions. For markets and society to function, individuals and companies need access to credit. Credit scoring algorithms, which make a guess at the probability of default, are the method banks use to determine whether or not a loan should be granted. Your task is to design an algorithm for predicting the probability that somebody will experience financial distress in the next two years.

Historical data are provided on 250000 borrowers and is available **here**.

3. SMS Spam Classification

The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged acording being ham (legitimate) or spam. Your task is to bild a classifier for detrmining whether a SMS is a span or a ham. Data is available **here.**

Remarks:
- There are plenty of R/Python packages for text manipulation.
- You are allowed to use any vocabulary / stop-word / etc.. dataset you deem appropriate.

4. House Prices

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, your task is to build a statistical model for predicting the final price of houses. Data is available **here**.