Data Science for Business

Situation:

After your first assignment has finished, LAE's staffing department has sent you on to your next project.

This project has already been running for some time. The colleagues are busy crunching numbers and analyzing data. You are asked to help with a specific task on the project.

To know your new project, look at your team's student numbers (r-numbers). Take the lowest number. Of this number, the last digit determines your project:

Last digit	Project
0, 1, 2, 3, 4	A
5, 6, 7, 8, 9	B

Task:

See which project you were assigned to.

For each of the projects, your colleagues have left you detailed information on what you are supposed to do.

There are some similarities between both project A and project B: In both cases, you are presented a classification dataset on several historical observations of some kind. The client asks you to build a model to predict the outcome of new observations.

This is a classical task in the life of a data analyst: You are presented historical data and you want to find a model that can predict the outcome of future data. Use cases include "How does a patient react to a certain treatment?", "Is a certain financial transaction fraudulent?" or "Will a tool break under a certain load?". You will get to know more during your projects.

Due to your training, you know many techniques to build such a model.

As before, LAE has a standardized way of approaching such problems. For a business application of data analytics, we must think about a few more things than just model performance. The reporting template will guide you to answering all questions you client might have.

By the deadline, your project manager asks for your report.

Filling this new template requires more of both creativity and modelling knowledge than the first one. You will be working with RapidMiner again. As before, load the data into RapidMiner to proceed:

1) *Data exploration:*
   First, as always, you are asked to explore the data. Datasets come in various shapes and sizes, hence, there is no straight-forward way of doing this. Two things are needed:
   One, add a visual to the template, e.g., the scatter plot of an interesting relationship you have found between two variables.
   Two, comment on the dataset in at most four bullet points. There are several directions to go, such as "What kind of variables are in the dataset?", "Are there missing values?", "What's the dimensionality of the data?", or "Can you make first guesses on the type of relationships between the variables?". This list in non-exhaustive.
2) Modelling
   Now we will get our hands dirty modelling the data.
   Apply three predictive modelling techniques and compare them in terms of their performance.
   a. In the reporting template, name the three techniques you have applied and provide a brief motivation for why you believe it's a good idea to apply them to the data (think about contents discussed in your 'Data Science for Business' course, such as interpretability or complexity).
   b. Explain briefly how you have evaluated the models, e.g., "Which performance metrics have you used and why?", "How have you set up your training and evaluation?", or "Have you split up your data?".
      Use at most 4 sentences.
   c. State your results. You can decide to write them down in MS Word or paste a screenshot from RapidMiner (or software of choice). Make sure screenshots are readable.
      Next to the raw results, discuss them in at most four bullet points. Directions to go include "What is the seemingly best performing model?", "Are your results conclusive?", "How do the models compare with each other?", "Are these good models, or why not?". This list is non-exhaustive. Keep in mind your audience and be insightful.
   d. Paste a screenshot of your RapidMiner pipeline. If you use a different software, either take a screenshot, or describe the pipeline in words

   Note that for part 2, you do not have to do any pre-processing of the data!
3) Pre-processing
   Note that we have not applied any pre-processing in our modelling. You might know that pre-processing is typically a crucial step in handling data.
   Name three potential reasons why you might need to pre-process data for a data analytics task (not necessarily related to the project you are working on).

Good luck!

Project A: Succession management at WhitesT

WhitesT (spoken "Whites' Tee") is a service provider for dentists is England.

Across departments, WhitesT has problems with succession management, i.e., finding replacements for people leaving the company and employees leaving prematurely.

Your colleagues have already done a good job in gathering all data they were able to find since beginning of the company of people leaving. The csv-file was sent to you.

Each row in the data corresponds to an employee that has left the company.

Your goal is to build a model that can predict whether an existing employee is about to leave the company or not.

Attached is a summary of the data:

Variable	Description
satisfaction_level	The overall workplace satisfaction of the employee
last_evaluation	Last evaluation of the employee performance
number_project	Number of internal projects the employee was assigned to by today
average_montly_hours	Average hours worked per month
time_spend_company	Years spend with WhitesT
work_accident	1: Employee suffered a work accident, 0: otherwise
promotion_last_5years	1: Employee was promoted in last 5 years, 0: otherwise
department	The department the employee is working in
salary	Salary of the employee (low, medium, or high)
left	1: Employee left the company, 0: otherwise

Project B: Churn at Xoom

You are sent to a long-time client of LAE, Xoom. Xoom is a B2C provider of video conferencing solutions.

Next to equipment, like headphones, Xoom also offers their own pre-paid mobile contract, that customers can use to do work-related phone and video calls. The product is not an overall success.

Your colleagues are investigating a potential redesign of the product. You are here to build a model that can predict existing customers from churning.

Each row in the data corresponds to a customer that has left the company.

Attached is a summary of the data:

Variable	Description
Area_Code2	Where the customer is located
Vmail_Message	Number of voice messages in last month
Day_Mins	Average length of phone calls during daytime
Day_Calls	Number of daytime phone calls in last month
Day_Charge	Amount of money charged for daytime calls
Eve_Mins	Average length of phone calls during evenings
Eve_Calls	Number of evening phone calls in last month
Eve_Charge	Amount of money charged for evening calls
Night_Mins	Average length of phone calls during nighttime
Night_Calls	Number of nighttime phone calls in last month
Night_Charge	Amount of money charged for nighttime calls
Intl_Mins	Average length of international phone calls
Intl_Calls	Number of international phone calls in last month
Intl_Charge	Amount of money charged for international calls
CustServ_Calls	Number of phone calls made to customer service
Churn	1: Customer has churned, 0: otherwise