Lab 1 | Data Science

Example Data Science Process

Painting of an Orange Sunset

Main Goal & Objectives

The main goal of this lab is to gain familiarity with the process of doing data science by studying an example. The goal is to read the paper "Predicting seasonal movements and distribution of the sperm whale using machine learning algorithms."

Q1 | What were the authors trying to accomplish?

The abstract of the paper gives an overview into the paper. As per the abstract, the problem the authors were planning was to implement a conservation plan in the Mascarene Archipelago in the south-west Indian Ocean for Sperm Whales. The study aimed at the following:

- Investigating the seasonal movements
- Predicting the potential distribution
- Assessing the diel vertical behavior of the sperm whale

As a result, this had meant that the authors were trying to **predict the potential distribution of sperm whales during the wet and dry season separately** as mentioned in Section 2.5.

Q2 | Describe their data. How many whales did they track? On average, how long was each whale tracked? During that time, how many locations did they visit on average and how far did they travel on average?

The number of whales being tracked can be found in Section 2.1, where as quoted "Sperm whales (n=21) were instrumented..." meant that **21 sperm whales** were tracked.

The numbers of locations the sperm whale visited on average can be found in Section 3.1, where as quoted "the number of locations recorded per sperm whale ranged from 7 to 176. The tracking duration was on average 34 ± 20 days." Using this information, we can deduce that the average was **0.2 ± 8.8 location / day**.

The distance travelled by the sperm whales on average can be found in Section 3.1, where as quoted "the total distance travelled varied between 154 and 3,112 km, and the average horizontal speed was 2.7 ± 0.3 km/h." Using this data, we determine the average distance covered as **154 to 3,112 km**.

Q3 | In Section 2.5, the authors describe their approach to applying machine learning models to perform species distribution modeling. In this problem, the area of interest is divided into a grid, and the model is applied to each grid cell to make a prediction. What were the eight input values for each grid cell.

When it came to dividing the area of interest into a grid, where the model is applied to each grid, the authors were looking at strong relationships between cetaceans distribution and dynamic environmental variables. This can be found in Section 2.4, as these are the following **eight variables**:

1. Sea Surface Temperature (SST)
2. Sea Surface Height (SSH)
3. Ocean Currents (U and V components)
4. Ocean Current Velocity
5. Mixed Layer Depth (MLD)
6. Bottom Temperature
7. Bathymetry
8. Seafloor Roughness

Q4 | Is this framed as a classification or regression problem? Describe what the dependent (output) variable represents and how it is interpreted.

This problem is faced as a **classification** problem. According to Section 2.5, it is stated that "The 14 algorithms were ran for each simulation run using the presence of sperm whales (1: presence or 0: pseudo-absence) as a response variable".

A response variable means the variable is dependent. Since the goal was to check for the presence of sperm whale, this had meant that the authors framed the problem as a classification problem.

Using the dependent variables, the results were interpreted to created a "tuned model" which was then used to generate ten prediction maps of the sperm whale's distribution, which were then averaged to provide a **final map of the distribution of sperm whales during the wet and dry season separately**.

Q5 | In Section 3.4, the authors describe the results of testing 14 different machine learning models plus a combined "stacking" model. Which models performed best?

Between the 14 different machine learning models, there were six performance metrics used to determine the best performing models, such as accuracy, kappa, sensitivity, specificity, F1 score, and TSS.

After finding the mean values of the different performance metrics, it was found that the **best model was the Random Forest (RF) for both seasons**, with values ranging from 0.93 to 0.99.

Q6 | What differences do you observe between the maps in Figure 6 for the wet and dry seasons? What do these differences mean in terms of the movement of the whales?

According to Figure 6, it can be seen that there is a large swath of ocean and two islands which are depicted as grey (Mauritius & Rodrigues). The wet season graphs are depicted as purple and the dry season graphs are depicted as orange.

When comparing the wet and dry season graphs, it can be seen that when comparing the same latitudes and longitudes on the two different season graphs, there is a higher probability of a sperm whale at those exact locations comparatively. Another observation is that there is a high probability of the sperm whales near the islands during the dry season.

In Section 4.2, a few things can be deduced in terms of the movement of the whales:

1. During the wet season, a significant proportion of the individuals left the island's coastal waters to perform a short migration towards Rodrigues

2. In Mid-December, it is shown that 70% of tracked whales showed a synchronized departure from Mauritius.

We can conclude from such observations that the whales begin to migrate away from the island's coastal waters during the wet season.

Q7 | Do you think it was worthwhile to evaluate 14 different types of machine learning models?

Yes. Having different types of machine learning models allows for capturing the different patterns in data, as well as complex relationships. In Section 4.3, it is mentioned that "the machine learning based approach used in this study has the ability to model complex polynomial relationships without relying on unrealistic assumptions".

Q8 | In Section 4.3, the authors discuss the advantages of their approach versus more traditional approaches. Summarize their argument.

The authors argue that because of their approach, they have been able to produce new insights of the movements of sperm whales. They strongly believe that instead of using traditional regression methods, their machine learning based approach allowed them to model complex relationships without relying on assumptions. Furthermore, they argue that their approach provides a new way to combine predictions from several algorithms.

Q9 | What might you do differently, if anything, if you were given this data set?

Given that the number of whales that were observed was 21 as mentioned in Section 2.1, it seems that the dataset is relatively small. Hence, it would not be ideal to use deep learning or convolutional neural networks, as such models require larger datasets.

While other models were used in the paper, such as KNN's and Bayesian Additive Regression Trees (BART), both of which were models in my knowledge domain, I don't have any further suggestions as such approaches used in the paper were approaches already known to me.

Q10 | Explain the six phases of the CRoss Industry Standard Process for Data Mining (CRISP-DM) model. For each phase, describe its purpose and what activities or decisions are typically involved. Illustrate how these phases interconnect to guide the entire data science process.

Phase 1 | Business Understanding

The purpose of this phase is understanding the objectives and requirements of the project. Such examples of activities are determining business objectives, assessing the situation, determine the data mining goals and producing the project plan.

Phase 2 | Data Understanding

This phase is driven by the focus to identify, collect and analyze the data sets that can help with accomplishing the project goals. There are four tasks in this phase, which are collecting initial data, describing the data, exploring data, and verify the data quality.

Phase 3 | Data Preparation

Data Preparation is referred to as "data munging", which prepares the final dataset for modeling. There are five tasks, which are selecting data, cleaning the data, constructing the data, integrating and formatting the data.

Phase 4 | Modeling

When it comes to the modeling phase, we work on building and assessing various models based on several different modeling techniques. There are four tasks in this phase, which are selecting modeling techniques, generating test designs, building the model and assessing the model.

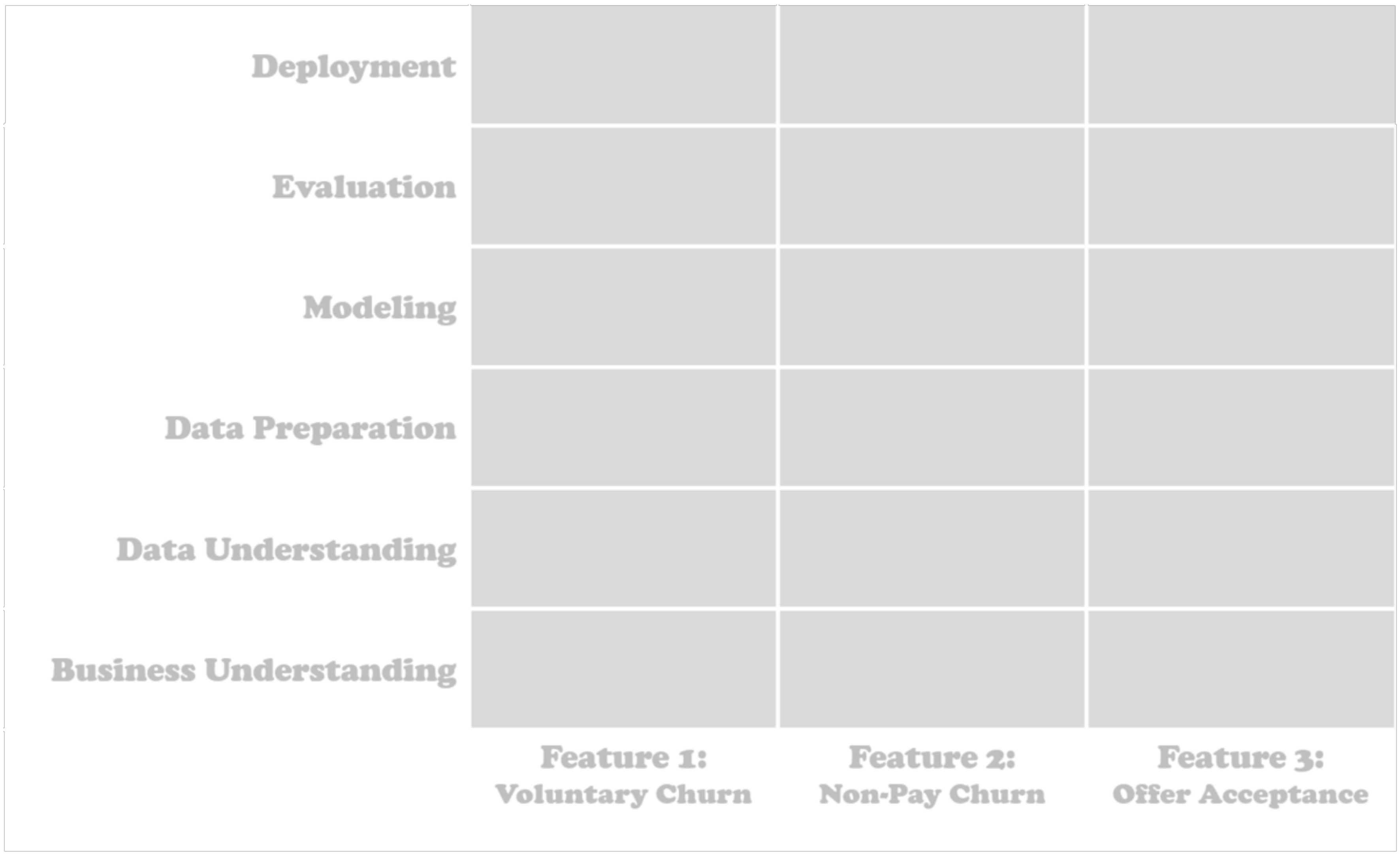
Phase 5 | Evaluation

The evaluation phase looks more broadly at which model best meets the business and what to do next. The phase has three tasks, which are evaluating the results, reviewing the process, and determining next steps.

Phase 6 | Deployment

The deployment phase is designed to allow customers to access the results. This is achieved through four tasks, which are planning the deployment, planning the monitoring and maintenance, produce final report and review the project.

Illustration | Vertical Slicing



Vertical Workflow Diagram

As per the illustration above, vertical slicing is a agile implementation, where the team would narrowly focus on quickly delivering one vertical slice up the value chain from a feature at a time.

As a result, this would allow for multiple smaller vertical releases and allows space for feed back from feature to feature.

Sources

Hotz, Nick. “What Is CRISP DM? - Data Science Process Alliance.” Data Science Process Alliance, 10 Sept. 2018, www.datascience-pm.com/crisp-dm-2/#What_are_the_6_CRISP-DM_Phases.

Chambault, Philippine, et al. “Predicting Seasonal Movements and Distribution of the Sperm Whale Using Machine Learning Algorithms.” Ecology and Evolution, vol. 11, no. 3, 12 Jan. 2021, pp. 1432–1445, https://doi.org/10.1002/ece3.7154.