Data Driven Scrum with CRISP - DM

Mark Paradis, Janani, Courtney Hrdy, and Emilio Ramos Monzalvo

I Introduction

Big North Airlines CEO has directly tasked the data science team to come up with a way to reduce customer churn and the effort is of top priority. The team has been working hard on several projects and not been able to produce results that have been valued by the business thus far.

In order to ensure that the Big North Data Science team has the highest probability of achieving success, or conversely quickly “fail” without promising success, we will be using a Data Driven Scum workflow to ensure a high level of communication, collaboration, and accountability across all of the roles and disciplines.

Establishing a collaboration workflow makes certain our team will know how to work together. Additionally, to complete the project as quickly and efficiently as possible we will also be choosing a process lifecycle framework to provide a guiding structure as we assess the business problem and work through the delivering value from the project team. Our framework focuses mostly on the process rather than orienting it by deadline.

II Process Framework

A CRISP-DM

CRISP-DM provides an overall set of guidelines of what should be done during a data science project. CRISP-DM provides a uniform framework for :

1. Standard Phases of a Data Science Project
2. How the Phases are related to each other.
3. Documentation

This methodology is cost-effective as it includes a flow structure for the team to stay on track to meet customer demands. CRISP-DM encourages best practices and allows projects to replicate. Being a cross-industry standard, CRISP-DM can be implemented in any Data Science project irrespective of its domain. The broad framework allows the team to adjust to new challenges as it allows for a team to take steps back in the project.

CRISP-DM has been the de-facto industry standard process model for data mining, with an expanding number of applications across a wide array of industries. It is extremely important that every data scientist and data miner must understand the different steps of this model. Since it is used across multiple industries, it is also easier to communicate to the customers.

The mixture of DDS workflow and CRISP-DM framework will allow the team to not only have a clear path on our project goals, but to also allow the flexibility needed for the dynamic iterations of a DDS workflow. The DDS workflow is meant to adjust to the project's communication based on the status of the project. This means that the iterations will be smaller at the discovery and understanding phase. Then, for bigger phases like Data Preparation, the iterations can be extended for the team to have more time to deliver bigger tasks that need less communication among the team.

We will first use a horizontal slicing approach to address the business understanding and data understanding DM-CRISP Phases for all the possible features that could lower customer churn. Once we have established a "good enough" understanding of the business and data we will switch to a vertical slicing from modeling to deployment to deliver one complete feature per iteration.

B Defining the First Iteration

The analysis will be used to reduce Customer Churn in Big North Airlines. We have four departments to work on the project. The Marketing and Sales Department and Group in charge of partner airlines will work on the survey and are responsible for the questionnaires. The Data Science Team is given access to the database where the survey data is stored. The Data Science team prepares, cleans and analyzes the data and creates some insights for a few columns. As the Data Science team develops new insights, the Marketing team will collaborate with DS to create new strategies.

The Marketing Department will use the insights given by the Data science department to provide feedback that will be used to better design and develop the models required in the following vertical sliced iterations.

Backlog and Priority

The items in the backlog are defined by which phase from the Project Framework the task belongs to. The phases include Business Understanding, Data Understanding,

Data Preparation, Modeling, Evaluation, and Deployment. These phases are meant to categorize the type of task by relating it back to the Data Scientist Workflow in order to understand the priority estimation factors

Table 1: List High level items in the product backlog for the start of the project

Name	**Quick Description**
Review project with project stakeholders	Setup the meetings and request stakeholders and document all questions, answers, configurations, etc
Complete analysis of all aspects of known data	Connect, pull, sample etc the data that the stakeholders are aware of and are possible to access
Investigate possibilities of other data sources	Brainstorm with the team and any possible subject matter experts. Perform research for similar data projects
Investigate possible modeling approaches	Once the data has going through EDA research possible models and techniques that would apply to the data and business problem

The priority is decided through the use of a few factors

- Estimated **Time** To Complete
- Probable/Perceived **Value**
- **Skills** required to complete task
- Amount **Coordination** and dependence on other Teams
- **Dependency** on other items in the backlog

Then they are graded by color, color scale is dependent on factor type:

- Green Examples:
    - **Small** amount Effort, **Low** monetary Cost to business, **Short** Timeframe, **No** Coordination with other teams, **Have All** skills needed on staff, **No** Dependencies to other items, **High** probable Value to business
- Yellow Examples:
    - **Medium** amount of Effort, Dependent **on 1 other task**, **Some** Skills Require training, Coordination with **another** team.
- Red Examples:
    - **Large** amount of effort, **Expensive** monetary cost, **Unknown** Time Frame, **Multiple** Team Coordination, **None - Low** Value to business

During the process of iteration review we will re-examine the grades for the factors on the PBI items. For example if there is a change of dependencies for an item

that was dependent on another item completed in a previous iteration, the color may change from yellow to green if there are no longer any dependencies to other items.

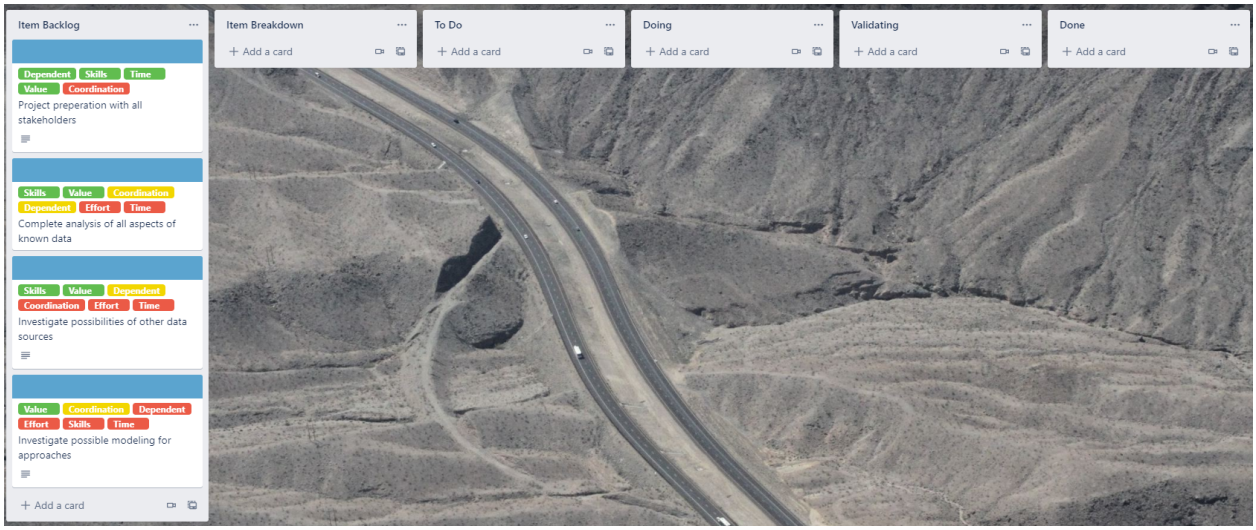
Since we are in the first iteration of the project, we will need time to understand the problem statement at hand. Therefore, we set the deliverable and the iteration to focus on tasks which can alter the subsequent steps that we are planning to take based on the CRISP-DM framework.

For example, understanding whether the survey is going to be a long term provider of data for our model or if the format and questions will change over time will directly impact the Deployment phase. Therefore, we place it at the top of our backlog. Furthermore, our team consists of different skills, and the Business Understanding tasks will not be something up to their skill set. Looking through the Backlog we can identify other important tasks that will be more fitting for a Data Engineer. The combination of the priority estimators and the phases will facilitate the task assignment and make sure we keep on track to our goal.

Kanban Board

The Kanban board will be our tool of choice to visualize our progress in the project. It will consist of six columns to keep track of our high level tasks and their corresponding details breakdown. The columns are Item Backlog, Item Breakdown, To Do, Doing, Validating and Done. The Item Backlog contains the high level items that are needed for our team to complete the project which is sorted by priority. Once a backlog item is selected, it will be broken down into smaller tasks prior to assignment.

Figure 1: Start of Project Kanban Board



In the Sprint Planning meeting, the detailed tasks will be assigned to the members of the team once the iteration length is agreed upon. The tasks will be assigned by the amount of bandwidth and skill the member has and how much he believes he can get done. Once he starts a task, it will then be moved to the Doing column. Once the member believes that their task is done, it will be moved to the Validating column where the team can determine if the task was done appropriately or if it needs any more work done. If there is a consensus on the status of the task, then the task will be passed down to the Done column where it can't be taken out anymore. Once a task is Done, it cannot be opened again.

III Process Framework Application

A First Iteration

For the first set of iteration, we are determining what are specific outcomes of this project. This includes finding actionable items in different departments. For example, if we find churn is related to the Regional Partners, we need to know that a business unit will start a discussion about letting go of some partners. This commitment from the departments is needed prior to a project kickoff as it can be a driving factor of what data we use and what analysis we want to perform. This is why it is at the top of our backlog in terms of priority.

The “complete analysis of all aspects known data” will be placed in the item breakdown and decomposed into tasks by using the rule of thumb to break down an item by type of data science Task:

Create

- Visualized on card via **Pen and Pencil Icon**
- Create a meeting with the CEO and the marketing manager to discuss priorities

Observe:

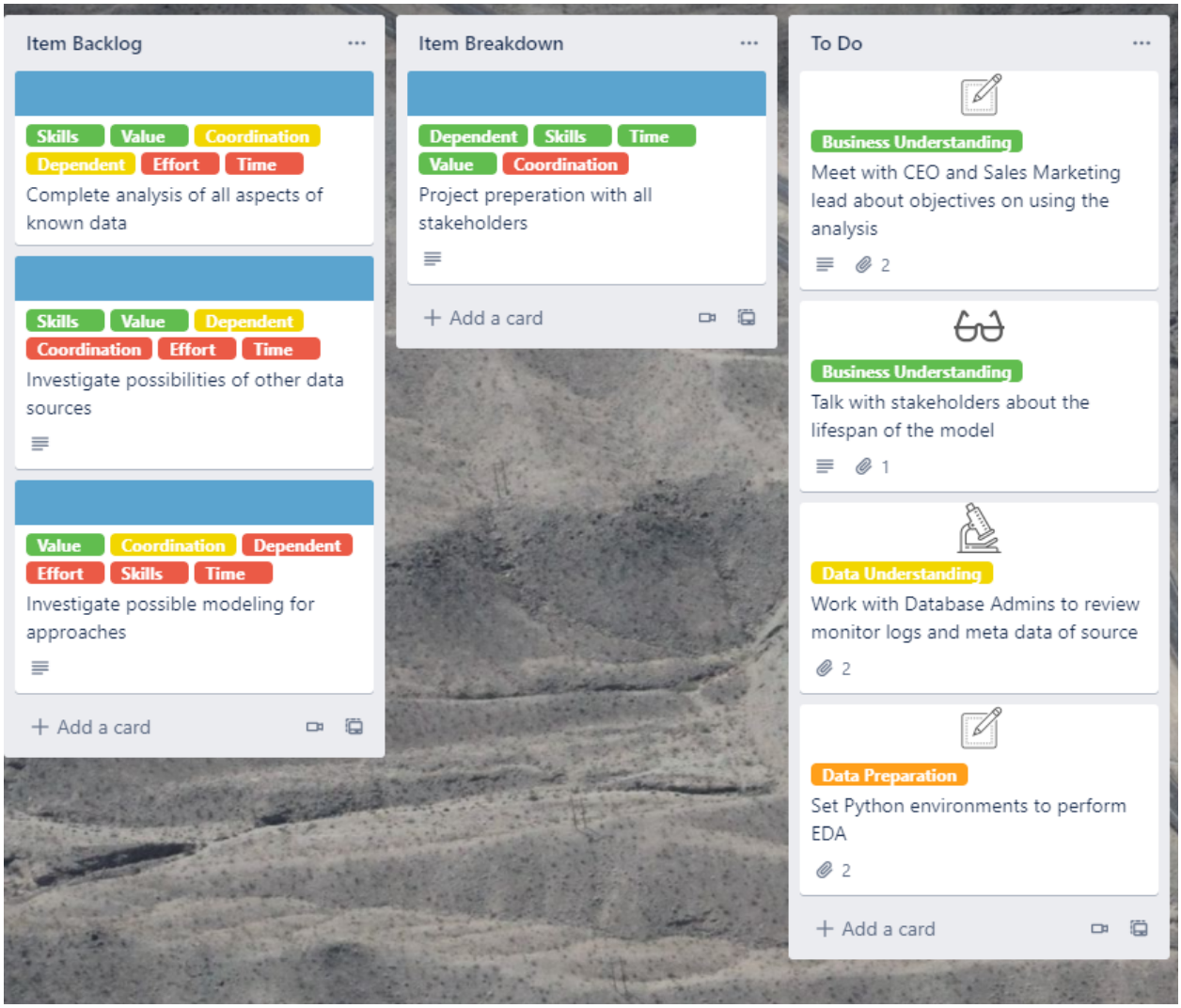
- Visualized on card via **Glasses Icon**
- Document meeting notes

Analyze:

- Visualized on card via **Microscope Icon**
- Assess the availability of data to complete their objectives
- Determine which modeling method to use to meet their priorities

The tasks would be communicated to the team through the Kanban board and through an iteration planning meeting. In this scenario, the team will meet and use common knowledge, the tasks will be assigned based on skill level and availability. Once the assigned task is completed, the team members will look into the board to be able to track which other tasks need to be done and where the rest of the team is. If a team member finishes all their assigned tasks for an iteration, they can pick the next priority task from the Item backlog or help another team member with tasks that are still in the To Do column.

Figure 2: Kanban after First Sprint Planning



-

B Second Iterations

The second iteration will be based on the outcome of the first iteration. For example if the meeting with the CEO was not achieved in the previous iteration, then we will communicate it with the team and replan our approach to the project. On the other hand, if the meeting with the CEO went well, then we will pass the first iterations' tasks into the Done column and start modulating the newest item and most important item from the backlog.

Once we have completed all the project preparation steps we should be able to start diving deeper in the data for the project. During this iteration we will complete an analysis of just the data that was identified during the project preparation iteration. This requires more types of tasks as well as more tasks from the data preparation phase of the CRISP-DM process lifecycle.