PROCESS METHODOLOGY

The first web app will provide data visualizations of certain trends, distributions of attributes with respect to traffic volumes that are particularly insightful in answering the data science questions for the target customer. Through these visualizations of the tendencies of the outcome variable; traffic volume with respect to predictors that impact the volumes will be observed. The second app will deploy a machine learning technique to predict future traffic volumes guided by the significance of predictors determined in the first app.

WATERFALL VS. AGILE

In order to choose the most appropriate process methodology, first the method for organizing the project must be selected.

Agile frameworks manage a project by breaking it up in several smaller phases, they require constant collaboration with stakeholders to make meaningful refinements. The iterative nature of agile methods allow each step of the cycle to be revisited as many times as desired to refine the understanding and results. Through this collaborative cycle, information is shared between different tasks, hence continuous improvements at each stage can be observed. Whereas waterfall is a more traditional approach which follows a linear flow between steps and hence requires finalization of previous steps in order to proceed to the next. With advancement in technology, Big Data analytics, AI… data science is beyond basic analysis and many problems faced by data scientists are non-linear. This linear dependency could potentially cause data analysis and software development processes to become inflexible and less iterative as there is no opportunity to revisit prior steps. This causes the process to be inflexible to change, as analytics won't be able to pivot back and forth through the process. Also it requires detailed requirement specifications up front. As most data scientists don't know what exploring the data yields before actually tackling it, determining data science questions at the beginning with no chance to refine it can be problematic in later stages. (Thurber, 2020)

With agile methodologies information flow with stakeholders and data science teams is maximized in a sense that stakeholders get value sooner and provide meaningful feedback in short timeframes. Data scientists can also evaluate model performance earlier and adjust their project plan in accordance with feedback from stakeholders and their prior findings about the model. (Hotz, 2022)Hence, overall a process methodology which can be aligned well with agile principles must be selected as it holds several advantages over waterfall approaches such as: early time-to market, great collaboration, early risk revealing through iterative development. (Saltz and Shamshurin, 2017)

SELECTION OF PROCESS METHODOLOGY

Two popular project management methodologies that work well with agile principles are Kanban and Scrum. Kanban is about visualizing and improving the flow of work. using kanban boards, teams

focus on reducing project durations. Whereas, through set intervals called sprints, Scrum relies on completing an increment of work. The objective of the team is to create learning loops to quickly gather and integrate customer feedback. (Rehkoph, no date)

However, for the specifications of coursework 1, a data science methodology would be more fit to explore, prepare and visualize the data. Hence another data science methodology stands out with its emphasis on data science workflow, which is CRISP-DM. CRISP-DM differs from Scrum and Kanban where the focus is on team collaboration.(Saltz, 2022) One advantage of CRISP-DM is that, just like Kanban, it is very adaptable in a sense that it can be implemented with minimal training, organization role changes or controversies. (Hotz, 2022)

Another potential fit methodology under these same considerations is Team Data Science Project; TDSP. TDSP is an agile and iterative data science methodology to deliver predictive analytics solutions efficiently, it is more modern than CRISP-DM with updated technology stacks. However, from time to time it can make the data science process unnecessarily complex, in that sense CRISP-DM would be a better fit based on its straightforwardness with an effective and simple data analytics process. (Thurber, 2020)

The chosen process methodology is CRISP-DM, it stands for CRoss-Industry Standard Process for Data Mining. When there is a clear business goal that translates into a data mining goal, CRISP-DM can become very useful in converting data into knowledge. Based on KDnuggets latest Poll, with 43% share, CRISP-DM is suggested as the most popular methodology for data analytics, data mining and data science projects.(Piatetsky, 2014) CRISP-DM has an iterative nature/structure with 6 main stages shown in Figure 1 (Chapman et al., 2000)

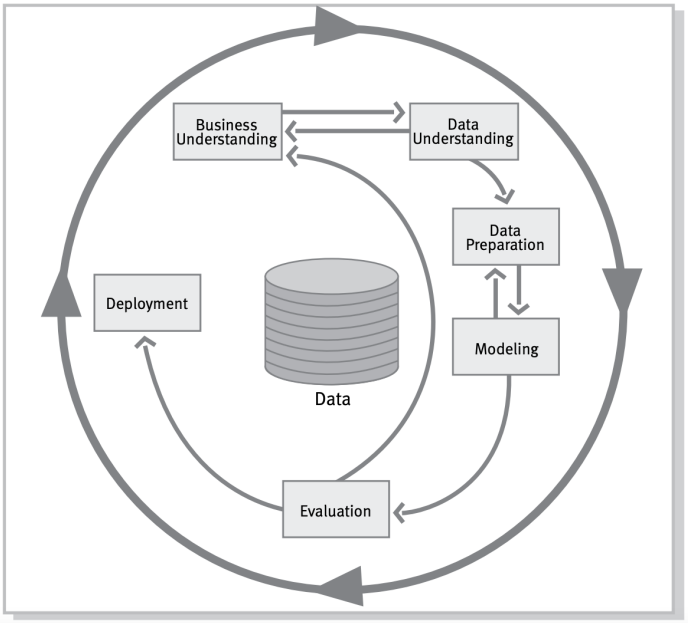

Figure 1: Steps of CRISP-DM Process Model

Advantages of CRISP-DM:

- **Business Understanding**

CRISP-DM emphasizes highly on the business understanding stage ensuring business goals are kept at the center of the project at all times. It encourages data scientists to focus on business objectives and goals to ensure project findings actually provide tangible benefits with high business value to the organizations. Certain relationships which the company may be well aware of may be trivial to discover throughout data analysis. In that sense thoroughly understanding business goals and problems beforehand may enable variables in our analysis to be chosen in a way that prevents such problems.

- **Cross-Industry Standard**

CRISP-DM is a cross-industry standard meaning that it can tackle any data science project notwithstanding its domain or destination, it can be applicable to every data mining project. Although it does not provide a detailed step-by step guide to each data science process, it still provides a simple framework that is easy to understand with very generalizable structures and intuitive steps. “CRISP-DM provides strong guidance for even the most advanced of today’s data science activities”.(Vorhies, 2016)

- **Flexible**

Fundamental reason why CRISP-DM is chosen is attributed to its iterative approach allowing frequent refinement at each stage of the cycle. This allows evaluation of the project progress with respect to business goals and objectives throughout the process. Through this flexibility, risk of getting an end result that does not address business goals gets minimized. (Sridharan, 2018) The flexible nature of CRISP-DM allows the benefits of agile principles. Through each iteration of the cycle, data science gains a deeper understanding of the data and the fundamental problem. Information and knowledge gained from prior iterations feed into following cycles. As the model and data analysis methods can be refined and improved on later iterations, it makes it possible for models to be flawed at the very beginning. If some of the distributions or correlations of certain attributes turn out to be very different from our initial business considerations or if certain outliers were omitted by accident, we may choose to go back to business understanding stage and refine our business our objectives or go back to data preparation and try determining different useful trends and distribution more aligned with the business objectives.

- **Long-term Strategy**

Another advantage that comes from the flexible and iterative nature of CRISP-DM is its ability to create long-term strategy for the organization enforced by its short and frequent iterations especially at the beginning stages. Through the first iteration of the cycle, a simple model can be generated that can be easily reinforced and improved in further iterations. This is done through information sharing between stages that allow preliminary analysis to be improved with respect to the latest insights.

Disadvantages of CRISP-DM:

A disadvantage of CRISP-DM is that it is not considered a project management methodology. This is because it implicitly assumes that its user is either a single person or a small team, and hence ignores the necessary teamwork coordination required for larger projects. (Hotz, 2022) Also, although the 6 stages of CRISP-DM are considered a great framework for analytics processes, it requires the specifications to be updated regularly to adapt to the challenges of Big Data and modern data science. (Piatetsky, 2014)

Improvements for CRISP-DM:

- As CRISP-DM is not well applicable to modern data science problems and challenges of Big Data, integrating aspects of modern technology could improve the robustness and practicability of this methodology. For example, when needed, leveraging cloud technologies and modern software practices like Git version control used in this project, could increase accuracy and performance. (Piatetsky, 2014)

- For the requirements of this coursework, in respect to data mining goals CRISP-DM is a great data science process methodology. Although data science and software development have common tools, to better tackle a software design project in later stages of deployment of machine learning algorithms in coursework 2, supporting CRISP-DM by coordinating with an additional software development process methodology like Kanban or Scrum could be more appropriate

References

Chapman, P. *et al.* (2000) *CRISP-DM 1.0: Step-by-step data mining guide*. The CRISP-DM consortium. Available at: https://web.archive.org/web/20220401041957/https://www.the-modeling-agency.com/crisp-dm.pdf (Accessed: November 1, 2022).

Hotz, N. (2022) *What is CRISP DM?*, *Data Science Process Alliance*. Available at: https://www.datascience-pm.com/crisp-dm-2/ (Accessed: October 29, 2022).

Piatetsky, G. (2014) *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*, *KDnuggets*. Available at: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html (Accessed: October 29, 2022).

Rehkoph, M. (no date) *Kanban vs. scrum: which agile are you?*, *Atlassian*. Available at: https://www.atlassian.com/agile/kanban/kanban-vs-scrum (Accessed: November 1, 2022).

Saltz, J. (2022) *CRISP-DM is still the most popular framework for executing data science projects*, *Data Science Process Alliance*. Available at: https://www.datascience-pm.com/crisp-dm-still-most-popular/#:~:text=CRISP%2DDM%20comes%20out%20as%20the%20most%20popular&text=Note%20that%20CRISP%2DDM%20is,possible%20and%20indeed%2C%20often%20occur. (Accessed: October 29, 2022).

Saltz, J.S. and Shamshurin, I. (2017) *Big Data Team Process Methodologies: A literature review and the identification of key factors for a project's success*, *IEEE Xplore*. Available at: https://ieeexplore.ieee.org/abstract/document/7840936/figures#figures (Accessed: October 28, 2022).

Sridharan, M. (2018) *CRISP-dm - a framework for Data Mining & Analysis*, *Think Insights*. Available at: https://thinkinsights.net/data-literacy/crisp-dm/ (Accessed: November 1, 2022).

Thurber, M. (2020) *A holistic framework for managing data analytics projects*, *Elder Research*. Available at: https://www.elderresearch.com/blog/a-holistic-framework-for-managing-data-analytics-projects/ (Accessed: October 28, 2022).

Vorhies, W. (2016) *CRISP-DM – a standard methodology to ensure a good outcome*, *Data Science Central*. Available at: https://www.datasciencecentral.com/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome/ (Accessed: November 1, 2022).