Tribhuvan University

Godawari College

Itahari, Sunsari, Nepal

An Intern Proposal

On

Data Science project of

“Startup Acquisition Status Predictions”

At

Aspire Tech Pvt. Ltd

Submitted by:

Doleshor khadka (16458/074)

An Intern proposal submitted in partial fulfilment of the requirement for **Bachelor of Science in Computer Science and Information Technology (BSc. CSIT) 8th Semester** of Tribhuvan University, Nepal

6th of June 2022

Abstract

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains. Data science is related to data mining, machine learning and big data. Data science is an interdisciplinary field focused on extracting knowledge from typically large data sets and applying the knowledge and insights from that data to solve problems in a wide range of application domains. The field encompasses preparing data for analysis, formulating data science problems, analysing data, developing data-driven solutions, and presenting findings to inform high-level decisions in a broad range of application domains. As such, it incorporates skills from computer science, statistics, information science, mathematics, data visualisation, information visualisation, data sonification, data integration, graphic design, complex systems, communication and business. Statistician Nathan Yau, drawing on Ben Fry, also links data science to human–computer interaction: users should be able to intuitively control and explore data.

Table of Content

1. About Organisation

Technocolabs was founded in 2019 by a Team of Non-Profit Organisation in Indore, who brought years of experience in Machine learning, Data science and AI Product Development to a new venture to make it a success. In 2019 an experienced leader Yasin Shah joined the team as a CEO and brought the company to a whole new level. Technocolabs is one of the leading Startup Companies with 1+years of experience in Web Development, Machine Learning and Artificial Intelligence, Mobile App Development. Features are :

Global Experience
Technocolabs have a solid track of Experience and have successful Data Science and AI projects completed for a variety of companies and multiple industries.

Value for Results
Technocolabs highly qualified team will take care of all your data needs ensuring high accuracy and quick turnaround.

Favourable/Convenient Terms of Cooperation
Technocolabs team is always ready to offer such terms of cooperation that will be the most suitable for your project needs and goals. T&M and Fixed Price models are offered.

High-Quality Results
Technocolabs focus is on compelling results. It builds its solutions to address the unique requirements and business-specific challenges.

Organisation: Technocolabs Pvt. Ltd

Address: JP Tower, First Floor P1, Dhar Naka Indore

Email: contact@technocolabs.tech

Phone: +91 8319291391

Website: www.technocolabs.com

Department: AI based Solution

2. Internship Supervisor

2.1. Internship Supervisor in Organization

Name of supervisor: Yashin shah

E-mail: yashiinshah@gmail.com

Linked-in: https://www.linkedin.com/in/yasinshah9598/

Facebook: https://www.Facebook.com/yasinshah

Address: Indore, Madhya Pradesh, India

2.2. Internship Supervisor in BSc. CSIT Program

Name of Supervisor: Pratik Gautam

Facebook: https://www.Facebook.com/thatsmepk

E-mail Address: thatsmepk@live.com

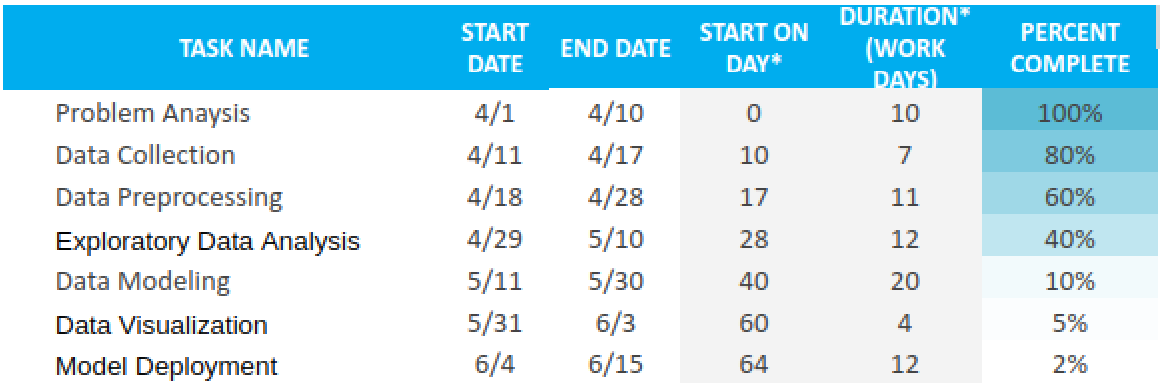
Address: Itahari 9, Sunsari, Nepal

3. Introduction of project

Startups are founded by one or more entrepreneurs who want to develop a product or service for which they believe there is demand. These companies generally start with high costs and limited revenue, which is why they look for capital from a variety of sources such as venture capitalists.Startups come with high risk as failure is very possible but they can also be very unique places to work with great benefits, a focus on innovation, and great opportunities to learn. A startup is a company that's in the initial stages of business.Founders normally finance their startups and may attempt to attract outside investment before they get off the ground.Funding sources include family and friends, venture capitalists, crowdfunding, and loans.Startups must also consider where they'll do business and their legal structure.Startups come with high risk as failure is very possible but they can also be very unique places to work with great benefits, a focus on innovation, and great opportunities to learn.The main challenge for this problem is dealing with an imbalance dataset where one class is overrepresented, but under/oversampling cannot be used as a technique to balance the data. In order to address this, an ensemble-based technique that can combine the results of a high precision anomaly detection algorithm with a good classifier should be used.

4. Duration

The time span of Internship is 10 weeks. Starting from the first week to the last week that is 10th week is represented in the gantt-chart below which include everything to be done in an entire intern time period.



TASK NAME	START DATE	END DATE	START ON DAY*	DURATION* (WORK DAYS)	PERCENT COMPLETE
Problem Anaysis	4/1	4/10	0	10	100%
Data Collection	4/11	4/17	10	7	80%
Data Preprocessing	4/18	4/28	17	11	60%
Exploratory Data Analysis	4/29	5/10	28	12	40%
Data Modeling	5/11	5/30	40	20	10%
Data Visualization	5/31	6/3	60	4	5%
Model Deployment	6/4	6/15	64	12	2%

Figure 1: Gantt-Table

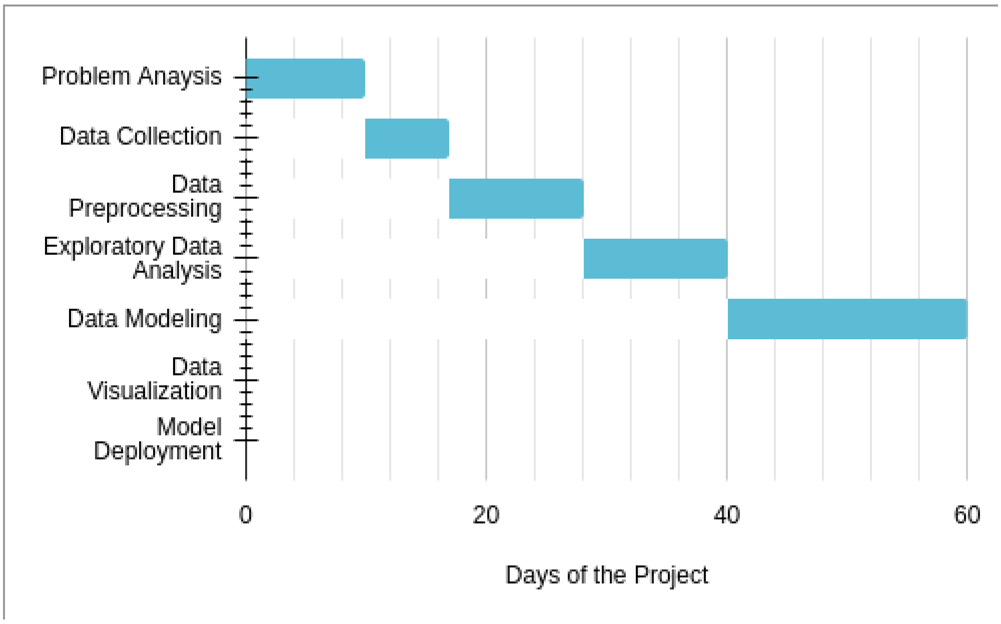


Figure 2: Gantt-Chart

5. Reason for Internship

5.1. Reason

Through my college teacher Mr. Pratik Gautam. He personally suggest me an Company named as Technocolabs PVT LTD owned by Mr. Yasin Shah. Mr. Pratik suggested this company to me based on my interest and my matching skills to an organisation.

5.2. Learning Objectives

- Explore career alternatives before graduation.
- Integration of theory and practice.
- Assess interests and abilities in the field of study.
- Learn to appreciate work and its function in the economy.
- Develop the work habits and attitudes necessary for business success.
- Develop communication, interpersonal and other critical skills during the job interview process.
- Create a work experience record.

5.3. Objectives of the Project

This information system will be developed using the SDLC (Systems Development Life Cycle), which has four stages: planning, analysis, design, and implementation. It is primarily intended to forecast results. The major goal is to create a model that can anticipate the startup's financial health. However, a large section of the research focuses on project success or failure probabilities over time, taking into account temporally dynamic data like the amount raised, funding rounds, amount funded and active days.There will be two models one will predict whether the startup is active or not and another model will predict the funding amount it will get on the basis of the company code category and other infos of the startup.
Models will be built by using following steps :

- **Understanding the business problem (and define success):**
  The first phase of any machine learning project is developing an understanding of the business requirements.we need to know what problem we're trying to solve before attempting to solve . The goal is to convert this knowledge into a suitable problem definition for the machine learning project and devise a preliminary plan for achieving the project's objectives. There are a lot of questions to be answered during the first step, answering or even attempting to answer them will greatly increase the chances of overall project success.Setting specific, quantifiable goals will help realize measurable ROI from the machine learning project instead of simply implementing it as a proof of concept that'll be tossed aside later. The goals should be related to the business objectives and not just to machine learning. While machine learning-specific measures -- such as precision, accuracy, recall and mean squared error -- can be included in the metrics, more specific, business-relevant key performance indicators (KPIs) are better.

- **Understanding and identifying data:**
  A machine learning model is built by learning and generalising from training data, then applying that acquired knowledge to new data it has never seen before to make predictions and fulfil its purpose. Lack of data prevents us from building the model, and access to data isn't enough. Useful data needs to be cleaned and in a good shape. Identifying data needs and determining whether the data is in proper shape for the machine learning project. The focus should be on data identification, initial collection, requirements, quality identification, insights and potentially interesting aspects that are worth further investigation. During this phase of the project, it's also important to know if any differences exist between real-world data and training data as well as test data and training data, and what approach we need to take to validate and evaluate the model for performance.

- **Collecting and preparing data**
  Once data is appropriately identified,we need to shape that data so it can be used to train your model. The focus is on data-centric activities necessary to construct the data set to be used for modelling operations. Data preparation tasks include data collection, cleansing, aggregation, augmentation, labelling, normalisation and transformation as well as any other activities for structured, unstructured and semi-structured data. Data preparation and cleansing tasks can take a substantial amount of time. Surveys of machine learning developers and data scientists show that the data collection and preparation steps can take up to 80% of a

machine learning project's time. As the saying goes, "garbage in, garbage out." Since machine learning models need to learn from data, the amount of time spent on prepping and cleansing is well worth it.

- **Determining the model's features and training it**

  Once the data is in usable shape and knowing the problem we're trying to solve, it's finally time to move to the step to do: Training the model to learn from the good quality data that we've prepared by applying a range of techniques and algorithms. This phase requires model technique selection and application, model training, model hyperparameter setting and adjustment, model validation, ensemble model development and testing, algorithm selection, and model optimization.

- **Evaluating the model's performance and establish benchmarks**

  Evaluation includes model metric evaluation, confusion matrix calculations, KPIs(Key Performance Indicators), model performance metrics, model quality measurements and a *final* determination of whether the model can meet the established business goals. Model evaluation can be considered the quality assurance of machine learning. Adequately evaluating model performance against metrics and requirements determines how the model will work in the real world.

- **Putting the model in operation and making sure it works well**

  Model deployment is done to make the model operational and to make it sure that it works well. Model operationalization might include deployment scenarios in a cloud environment, at the edge, in an on-premises or closed environment, or within a closed, controlled group. Among operationalization considerations are model versioning and iteration, model deployment, model monitoring and model staging in development and production environments. Depending on the requirements, model operationalization can range from simply generating a report to a more complex, multi-endpoint deployment.

- **Iterating and adjusting the model**

  When it comes to implementing technologies, it's often said that the formula for success is to start small, think big and iterate often. So repeating the process and making improvements in time for the next iteration. Business requirements change. Technology capabilities change. Real-world data changes in unexpected ways. All of which might create new requirements for deploying the model onto different endpoints or in new systems.

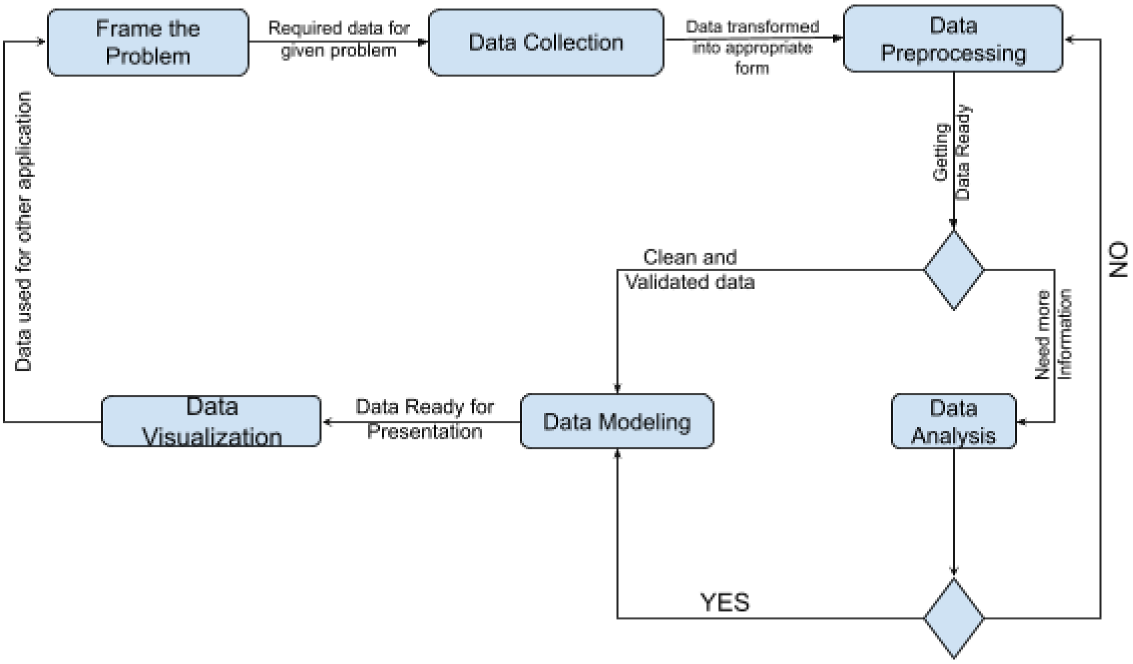


Figure 3: System Flow Diagram

5.4. Internship Assignment

I will be assigned to do the modelling part. In the modelling part I have to choose the best ML algorithm and implement it to the cleaned and validated data.

5.5. Research Question

Research on the targeted audience will be done by analysing the customers feedback and interest. Research on the competitors will be done and comparison of both the businesses will be done.

5.6. Methodological Approach

There are multiple methodological approach to be done in the intern project which are given below:

- Problem Analysis
- Data Analysis
- Algorithm Research
- Implementation of different Algorithm
- Choosing Algorithm
- Hyperparameter tuning
- Algorithm Optimization
- Code optimization

Bibliography

William L. Hosch (2022, April 7). Machine Learning . Retrieved from https://www.britannica.com/technology/neural-network