

CHAPTER 1: INTRODUCTION

1.1 Introduction

The 10 week internship at DLytica Academy Pvt.Ltd ,from March 2 to May ,2025 , provided me a practical platform to apply Data science concepts learned during my academic studies. The project focused on building the data –driven solutions using Python, SQL, machine learning techniques, and advanced tools like Apache Superset to address real-world problems.

This report outlines my work during the eight semester internship toward the BSc. CSIT degree at Tribhuvan University. It describes my contributions, technical learnings and professional growth during this internship.

1.2 Problem Statement

Traditional data analysis methods oftens rely on manual processes and lack scalability for handling large datasets.Today’s Businesses need efficient tools to process ,a nalyze and visualize data to inforomed desicisions . The internship projects aims to address these challenges by quering, developing Machine learning models and interactive visualization.

1.3 Objectives

The primary objectives of this internship project are as follows:

- a) To develop python abilities for data analysis and visualization.
- b) To write SQL queries for efficient data retrivel and analysis.
- c) To build and validate machine learning models using scikit-learn.
- d) To explore advance tools like Apache superset for interactive visualization.

1.4 Scope and Limitation

1.4.1 Scope

- Development of data preprocessing using Python and libraries like pandas ,Numpy and matplotlib.
- Implementation of SQL queries for data analysis using PostfresSQL.

- Training and validating the machine learning models with scikit-learn.
- Visualization of data insights using seaborn and Apache Superset.

1.4.2 Limitation

Although the internship project aims to implement several interesting features, there are some limitations to be considered:

- Limited to prototype –level machine learning models without production – grade optimization.
- Focused on small datasets due to processing capacity of the machine used.
- Basic visualization tools were used, advanced BI tools like tableau were not explored.
- Time constraints limited the depth of model tuning and feature engineering.

1.5 Report Organization

This report is organized into four chapters:

- Chapter 1 introduces the project context, problem, objectives, scope, and report layout.
- Chapter 2 provides details about DLYtica Academy , its structure, and the intern department, along with a brief literature review of ride-sharing systems.
- Chapter 3 documents my roles, weekly activities, project description, and detailed tasks performed.
- Chapter 4 concludes with a summary of accomplishments, learning outcomes, and reflections.

CHAPTER 2: ORGANIZATION DETAILS AND LITERATURE REVIEW

2.1 Introduction of Organization

Dlytica Academy Pvt Ltd, is a based specialized firm in data-driven solutions and training programs . The company focuses on Data Science, machine learning , and software development ,serving industries like education, e-commerce, finance and healthcare . With a commitment to innovation and skill development , Dlytica provides a nurturing environment for interns to gain hands-on experience.

Table 1: Contact details of the organization

Location:	Balkumari Lalitpur, Nepal
Opening Hours:	9:30 AM to 6:00 PM
Work days:	Sunday to Friday
Contact No.:	+977-9765978458
Email Address:	academy@dlytica.com
Website:	info@dlytica.com

2.2 Organizational Hierarchy

Dlytica Academy follows a flat organizational structure to foster collabaoration and agility.

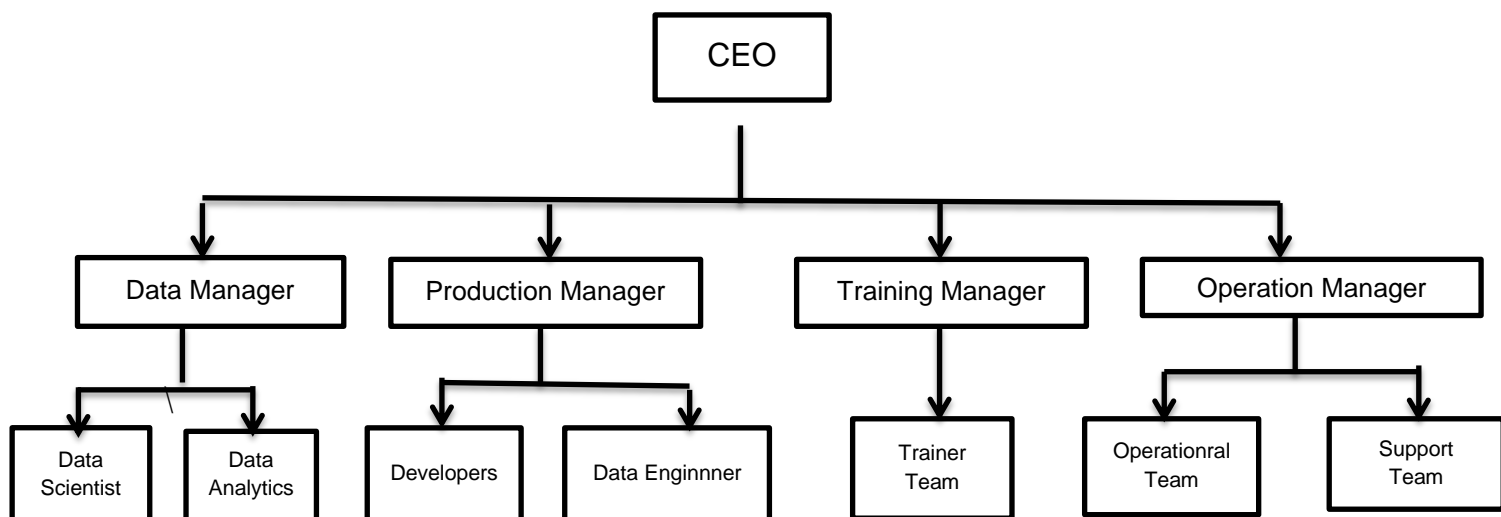


Figure 1: Organization Hierarchy of the company

2.3 Organizational Structure

The company emphasizes cross-functional collaboration, with interns participating in real projects under the guidance of experienced professionals.

- Data science and Analytics Team
- Software Development Team
- Training and Mentorship Team
- Operations and support

Interns generally work within the Data Science and Analytics Team, working closely with mentors.

2.4 Intern Department Description

Interns are placed under the supervision of an experienced team lead who provides mentorship throughout the internship. Activities include:

- Hands-on coding and data-analysis tasks.
- Participation in weekly sprint meetings
- Mentorship sessions for technical and professional growth.
- Feedback reviews to track progress and to find areas for improvement.

LITRETURE REVIEW

The field of Data Science has rapidly evolved, driven by advancements in computational power, data storage, and algorithmic techniques. This literature review explores key concepts, tools, and methodologies relevant to the internship projects, focusing on machine learning, SQL, visualization, and introductory AI concepts applied to customer segmentation and loan approval prediction.

McKinney (2022) emphasizes the role of Python as a cornerstone for Data Science due to its robust libraries like pandas, NumPy, and scikit-learn. Pandas facilitates efficient data manipulation, enabling tasks like filtering and grouping, which were critical for preprocessing customer and loan datasets during the internship. Scikit-learn provides accessible implementations of machine learning algorithms, such as Logistic Regression and Decision Trees, which were used in the Capstone Projects for their simplicity and interpretability (Brownlee, 2020).

Silberschatz et al. (2020) highlight the importance of SQL in managing relational databases like PostgreSQL. SQL's ability to handle complex queries, including joins, subqueries, and window functions, supports scalable data analysis. The internship leveraged SQL to extract and analyze customer and loan data, aligning with Silberschatz's assertion that optimized queries enhance performance in data-driven applications.

Advanced machine learning algorithms, such as Random Forest and Support Vector Machines (SVM), have gained prominence for their robustness in handling complex datasets. Brownlee (2020) notes that Random Forest, an ensemble method, improves accuracy over single Decision Trees by reducing overfitting, making it suitable for customer segmentation tasks. Similarly, SVM's ability to handle non-linear relationships through kernel functions makes it effective for loan approval prediction, where applicant data may exhibit complex patterns (Hastie et al., 2009).

Visualization tools like Apache Superset play a crucial role in communicating insights to stakeholders. The Apache Superset Documentation (2024) describes its capability to create interactive dashboards, which was instrumental in presenting model predictions and trends during the internship. Visualizations such as heatmaps and scatter plots helped interpret customer segments and loan approval factors, aligning with Tufte's (2001) principles of clear and effective data presentation.

The exploration of AI concepts, particularly neural networks, represents a frontier in Data Science. Goodfellow et al. (2016) explain that neural networks excel in modeling non-linear relationships in large datasets, offering potential for future applications in financial risk assessment and customer behavior analysis. While the internship introduced these concepts at a foundational level, they provided a theoretical framework for understanding advanced predictive modeling.

In the context of the internship, prior studies on customer segmentation and loan approval prediction offer relevant insights. Gupta and Sharma (2020) discuss the application of machine learning in customer segmentation, noting that Logistic Regression and Random Forest models effectively identify high-value customers, supporting marketing strategies. Similarly, Li and Zhang (2019) highlight the use of Decision Trees and SVM in loan approval systems, emphasizing their ability to assess credit risk based on applicant features like income and credit history.

This literature review underscores the relevance of the tools and techniques used during the internship. Python and SQL provided a foundation for data processing and analysis, while machine learning algorithms and visualization tools enabled practical solutions for business problems. The introduction to AI concepts opened avenues for future exploration, aligning with the evolving demands of Data Science in industry applications.

CHAPTER 3: INTERNSHIP ACTIVITIES

3.1 Roles and Responsibilities

As a Data Science Intern, my responsibilities included:

- Writing and optimizing SQL queries using PostgreSQL for data analysis.
- Developing Python Scripts using numpy, pandas ,matplotlib and seaborn for data processing and visualization.
- Building and validating machine learning models with scikit learn.
- Intergrating SQL and Python for business analytics projects,
- Creating Interactive dashboard using ApacheSuperset.
- Documenting the procedures, finding insights and presenting to team.

3.2 Weekly Logs

Weekly log Summarize work activities for the week. It Includes daily activities. Explain how interns made the most of their internship: skills learned & used, responsibilities given, accomplishments, achievements. Include observations about the workplace environment to get comfortable in working environment.

Weekly Internship Log

Name of the Student: Sujan Mishra

Date: 2025/03/09

Project/ Job Title: Data Science Internship

Week Number: 1

Responsibilities

Attending the internship program.
Setuping the working environment in laptop.

Activities

Attended internship orientation session and understood the company and it's environment.
Installed Python ,PostgreSQL and Virtual studio code for IDE.
Reviewd the basic Python and SQL concepts.

Observations

Company culture is collaborative and supportive.
Training sessions are very informative.
Team members are experienced and willing to help.

Plan for Next Week

Learning Python Libraries like Numpy , pandas and exploring SQL basics .

Mentor Feedback

Good Start, get famalize with tools.

Weekly Internship Log

Name of the Student: Sujan Mishra

Date: 2025/03/16

Project/ Job Title: Data Science Internship

Week Number: 2

Responsibilities

Learning python libraries like pandas ,Numpy and matplotlib lib.
Learning Basic Sql Query Writing.

Activities

Installed and configured Python environment with libraries.
Practice data manipulation with pandas like grouping and filtering.
Wrote basic SQL SELECT queries to retrieve data.

Observations

Hands-on practice is enhancing understanding of technologies.
Configuring environment was easy and straightforward with provided document and guidance.

Plan for Next Week

Explore Advance SQL like JOIN, subqueries.
Get familiarize with matplotlib for visualization

Mentor Feedback

Good in learning, focus on efficiency.

Weekly Internship Log

Name of the Student:Sujan Mishra

Date: 2025/03/23

Project/ Job Title: Data Science Internship

Week Number: 3

Responsibilities

Writing SQL queries with join and subqueries
Create Visualization With Matplotlib.

Activities

Implemented different kinds of joins like
INNER,OUTER,FULL,LEFT to combine tables.
Used subqueries to filter the data by applying condition.
Plotted Bar charts and histograms using matplotlib.pyplot.

Observations

Joins were faster than subqueries, and work effectively on large datasets.
Visualization requires knowledge for selecting the chart for better Clarity.

Plan for Next Week

Learn Window function in SQL
Explore seaborn for better visualization.

Mentor Feedback

Good in querying, recommended practicing window functions.

Weekly Internship Log

Name of the Student:Sujan Mishra

Date: 2025/03/30

Project/ Job Title: Data Science Internship

Week Number: 4

Responsibilities

Implementing windows function and using sea born for visualization.

Activities

Used windows function like RANK ,DENSERANK,ROW NUMBER to analyze order ranking.
Heatmaps and pair plot are created for coreealtion analysis.
Optimized the SQL by replacing the subqueries with joins.

Observations

Window function simplified ranking and aggregation tasks.
Seaborn offer more pleasing visuals and can be more customized than matplotlib.

Plan for Next Week

Begin Machine Learning with sckit-learn for project.
Work on data preprocessing for ML models.

Mentor Feedback

Very Good. Now,shift focus on ML.

Weekly Internship Log

Name of the Student:Sujan Mishra

Date: 2025/04/06

Project/ Job Title: Data Science Internship

Week Number: 5

Responsibilities

Preprocess the data for Machine learning.
Train a Logistic Regression for the Project 1.

Activities

Claened and normalized the dataset using pandas for Project.
Trained a Logistic Regression Model with Scikit-learn.
Evalated model performance using cross-validation and accuracy Metrics.

Observations

Data preprocessing was critical and most required step for model performance and cross validation help to ensure robust model evaluation.

Plan for Next Week

Start project 2 with decision tree model
IntegrateSQL with python for data extraction.

Mentors' Approval

Signature:

.....

Bidhya Pokhrel

Supervisors' Verification

Signature:

.....

Bishal Shrestha

Mentor Feedback

Good, recommended exploring feature engineering.

Weekly Internship Log

Name of the Student: Sujan Mishra

Date: 2025/04/13

Project/ Job Title: Data Science Internship

Week Number: 6

Responsibilities

Integrate SQL and Python for data workflows.
Work on Project 2 with Decision Tree model.

Activities

Connected PostgreSQL with Python using psycopg2 for data extraction.
Developed SQL queries for analysis in project2.
Trained Decision Tree model to predict the loan based on customer data.

Observations

Decision tree were easier to interpret but can easily be overfitted.
SQL-Python integration makes the data workflow easier.

Plan for Next Week

Begin exploring Apache Superset for visualization and tune models.

Mentor Feedback

Strong skills, focus on validating model with best efficiency.

Weekly Internship Log

Name of the Student:Sujan Mishra

Date: 2025/04/20

Project/ Job Title: Data Science Internship

Week Number: 7

Responsibilities

Set up Apache Superset for Visualization.
Finalize ML models.

Activities

Apache Superset was installed and configured, and successfully connected to the PostgreSQL .
Created initial dashboard for the project.
Logistic regression and decision tree were tuned using hyperparameters.

Observations

Hyperparameter tuning improved model accuracy slightly.
Superset's interface was intuitive for building dashboards.

Plan for Next Week

Develop dashboard for project 2.
Prepare project documentation.

Mentor Feedback

Good progress, focus on dashboard design.

Weekly Internship Log

Name of the Student: Sujan Mishra

Date: 2025/04/27

Project/ Job Title: Data Science Internship

Week Number: 8

Responsibilities

Create dashboard for project 2 in Apache Superset.
Document project workflows.

Activities

Designed interactive dashboards in Superset .
Visualized model prediction using barplots and heatmaps.
Documented workflow for both projects in notebook .

Observations

Superset Dashboard made data insights accessible to non-technical users which is very crucial in businesses.
Clear documentation helps to make things clear to mentores.

Plan for Next Week

Explore advance machine learning algorithms for capstone projects
Begin learning basic AI concepts.

Mentor Feedback

Good dashboard, work on presentation and visualization polishing.

Weekly Internship Log

Name of the Student:Sujan Mishra

Date: 2025/05/04

Project/ Job Title: Data Science Internship

Week Number: 9

Responsibilities

Explore advance machine learning algorithms.
Learn AI concepts to enhance knowledge base.

Activities

Studied Random Forest and Support Vector Machine(SVM) algorithms.
Applied Random Forest to project 1 to improve model accuracy.
Explored basic AI concepts like neural network fundamentals and it's application.

Observations

Random Forest improved the model performance but required more computational resources.
Neural network concepts were complex to understand, but it 's crucial for future.

Plan for Next Week

Apply SVM to project 2 nad finalize superset and prepare for presentation.

Mentor Feedback

Good to get the AI knowledge beforehand, focus on practical implications.

Weekly Internship Log

Name of the Student:Sujan Mishra

Date: 2025/05/11

Project/ Job Title: Data Science Internship

Week Number: 10

Responsibilities

Apply SVM in project 2 and document it well.
Finalize the interactive visualization in Apache Superset.

Activities

Trained SVM model for project 2 and comparing with the Decision Tree Model.
Finalized Superset dashboards for both projects.
Documented both projects so that it will help others to understand as well.
Presented final projects and reflection to the team.

Observations

SVM provided better accuracy for certain cases but was sensitive to parameter tuning.
Superset dashboards effectively communicated complex model insights to stakeholders.
AI concepts widen my knowledge , as it seems to be foundation of the

Plan for Next Week

Internship Completed

Mentor Feedback

Excellent progress in developing algorithms and visualizations and
Well prepared for datascience roles.

3.3 Description of the Project Involved During Internship

The internship involved two capstone projects focused on real-world business scenarios using data science techniques.

1. Project 1: customer segmentation:
 - Objective: Segment customers based on purchasing behavior using machine models.
 - Tasks: Extracted customer and order data from PostgreSQL using SQL queries (joins, window functions), preprocessed data with pandas, trained and validated a Logistic regression model, later applied random forest model to improve accuracy. Data is visualized in Apache Superset.
 - Outcome: Identified high-value customer which help in target marketing in businesses and improved accuracy from Random Forest model.
2. Project 2: Loan Approval Prediction
 - Objective: Predict the loan approval either it is safe or not based on the customer data using machine learning models.
 - Task: Developed SQL queries for loan application analysis, processed data with Numpy and pandas, trained a Decision Tree model, and later applied Support Vector Machine (SVM) model for Comparison. Created interactive dashboard using heatmaps, scatterplots, boxplots to visualize in Apache Superset.
 - Outcome: Provide insights which help in loan approval, and with SVM, there is the rise of accuracy in prediction.

Both projects demonstrated the practical application of Data Science in business analytics, integrating SQL, Python, advanced machine learning algorithms and visualization tools.

3.4 Task/Activities Performed

Some of the specific tasks I handled:

- Setup development environment and tools that are python, PostgreSQL, VS-code.
- Wrote SQL queries with joins and window functions for customer and loan data analysis.
- Developed python scripts for data preprocessing and visualization using numpy, pandas, matplotlib, seaborn.
- Trained and validated ML models: Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, used in both models.
- Created interactive dashboards for both project using Apache Superset.
- Explored Basic Introductory AI and its application, its influence in the present time.
- Documented Workflows and presented findings to the team.

Table 3: Test Cases

SN	Feature	Steps	Expected Result	Remark
1	SQL JOIN	SQL query with INNER JOIN for Customer order.	Retrived all matching records.	Pass
2	SQL SUBQUERY	Subquery for above average order total	Filter customer correctly	Pass
3	LOGISTIC REGRESSION	Use of Logistic regression model, measure accuracy in project 1	Accuracy above 85%	Achieved 90%
4	RANK FUNCTON	Window function (Rank) for customer ranking	Correct ranking of customers	Pass
5	DECISION TREE	Use of decision tree model, measure accuracy in project 2	Accuracy above 80%	Achieved 82%
6	RANDOM FOREST	Use of random forest model, measure accuracy in project 1	Accuracy above 87%	Achieved 89%
7	SIMPLE VECTOR MACHINE	Use of simple vector machine(SVM) model, measure accuracy in project 2	Accuracy above 82%	Achieved 84%

CHAPTER 4: CONCLUSION AND LEARNING OUTCOMES

Conclusion

The 10-week internship at Dlytica Academy Pvt.Ltd , from March2 to May11, 2025, was a transformative experience that deepened my understanding of Data Science .Starting with basic theoretical knowledge, I progressed to handling complex SQL queries build model to predict loan and cutomers using different machine learning models like Logistic Regression, Random Forest,Decision Tree and Simple vector Machine commonly known as SVM, and creating interactive dashboards with Apache Superset.The two capstone Projects allowed me to integrate SQL,Python for seamless data flow in the model, and advanced algorithms to solve real-world business problems-customer segementation and loan approval prediction.Exploring introductory AI concepts, like neural network, has made my mind that to learn about it more, as it enhance future Data Science to address more complex predication particularly in disease detection, finiancial risk assessment with promotes the businesses as well.

Beyond technical skills,I developed essentials Beyond technical skills, I developed essential soft skills, including time management, technical communication, and teamwork. The collaborative environment at DLytica fostered my growth, and regular feedback from my mentor helped me refine my work. This internship not only enhanced my technical proficiency but also built my confidence to pursue a career in Data Science.

Learning Outcomes

- **Technical skills:** Gained better knowledge and proficiency in python and it's libraries like numpy, pandas, seaborn, scikit-learn.And also in SQL like subqueries,joins>window functions, with PostgreSQL and Apache Superset. Learned advanced ML algorithms and get to know about AI and it's applications.
- **Analytical Skill:**Understood the importance of data preprocessing ,query optimization ,model validation, and visualization in Data Science Projects.
- **Professional Skills Improved Presentation and documentation skill,built ability to work in collaborative team.**
- **Networking:** Built connections with professionals at Dlytica ,good bonding with the fellow trainees.
- **Self-Motivation:**Learned to do thing by own, explore models and learn new concepts.

This internship has contributed me with the skills and confidence to contribute to Data Science Projects in professional setting. The experience highlighted areas for further improvement, such as deep learning, advanced model tuning, and AI exploration which I want to pursue in the future.

4.1 References

- [1] J. Brownlee, Machine Learning Mastery with Python, Machine Learning Mastery, 2020
- [2] W. McKinney, Python for Data Analysis, O'Reilly Media, 2022
- [3] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016

