NEW HORIZON COLLEGE OF ENGINEERING
DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

ALUMNI TALK -ODD SEM (2024)

Subject: Introduction to Data Science, Building blocks and beyond

Expert name: Mr. Aditya Srivastava

Audience: 3rd Semester Students **Date:05-01-2024(10:00AM-11:00 AM)**

The department of Information Science and Engineering has conducted Alumni talk on the topic "Introduction to Data Science, Building blocks and beyond " for the 3rd semester students on the 5th of January, 2024 under the supervision of ISE I Head of the Department, Dr.Vandana C P, offline. The expert speaker, 'Aditya Srivastava' was invited to conduct the same.

- The Speaker, Aditya Srivastava is currently working as a Senior Application Analyst developer. who examines the puzzle of motivation starting with a fact that social scientists know but most managers don't.

<u>VARIOUS SESSION S THROUGHOUT PROGRAM:</u>

Alumni Talk by Aditya Srivastava:

TOPICS COVERED:

- **About Data Science**
- **Building Blocks of Data Science**
- **Beyond the Basics**
- **Data Science Life-cycle**
- **Tools/Products**

About Data Science:

Data Science is a multidisciplinary field that combines techniques from statistics, mathematics, computer science, and domain-specific knowledge to extract valuable insights and knowledge from data.

It involves collecting, cleaning, analyzing, and interpreting data to make informed decisions and solve complex problems.

The field has evolved rapidly in recent years, driven by the increasing availability of data and advancements in computing power.

- **Building Blocks of Data Science:**

1. Data Collection: Data is collected from various sources, such as machine sensors (temperature, pressure, vibration, etc.), maintenance logs, and production data. The data is gathered in real-time and stored in a central data storage system, like a data lake or a data warehouse.

2. Data Integration: Data from different sources must be integrated and transformed into a consistent format for further analysis. Data engineers create ETL (Extract, Transform, Load) pipelines to clean, normalise, and aggregate the data. This involves handling missing data, removing duplicates, and converting units, among other tasks.

3. Data Storage: The transformed and clean data is stored in a structured database or data warehouse optimised for analytical processing. This storage system must be scalable, reliable, and cost-effective to accommodate the growing volume of data generated by the manufacturing plant.

4. Feature Engineering: Data engineers work with domain experts and data scientists to identify the most relevant features for predicting machine failure. These features may include rolling averages of sensor readings, time since the last maintenance, or other derived metrics. The feature engineering process involves creating new variables or transforming existing ones to better capture the underlying patterns in the data.

5. Data Modelling: Data scientists develop machine learning models to predict machine failure based on the processed and feature-engineered data. The models are trained and tested using historical data, with the goal of accurately identifying patterns that indicate an impending failure.

6. Model Deployment: The trained predictive models are deployed into a production environment, where they can be used to monitor the real-time data streaming from the machines. If the model predicts a high likelihood of failure for a specific machine, maintenance staff can be alerted to perform preventive maintenance, avoiding costly downtime and improving overall efficiency.

7. Monitoring and Maintenance: Data engineers continuously monitor the performance of the ETL pipelines, storage systems, and predictive models to ensure they are functioning optimally. They may also need to update or retrain the models as new data is collected, to account for changes in the manufacturing process or equipment.

Based on the business requirements, the analysis needed are:

- **Exploratory analysis** is the process of analyzing the dataset to summarize or get an overview of it. It is often done with visual methods using libraries like matplotlib, d3.js, and applications like a tableau.
- **Predictive analysis** is the major branch of data science where models are created using existing data to make predictions on future or unknown data.
- **A prescriptive analysis** is like an extension of predictive analysis in the sense that it not only predicts what will happen, it also suggests decision options to change the outcome.
- **IPA analysis** — Interpretative phenomenological analysis (IPA) is an approach which deals with psychological research.

Tools/Products

- **Visualization** — For exploratory analysis, the tableau is a popular tool to create interactive data visualizations. D3.js is an open source library that is used to create visualizations inside web pages.
- **Programming Languages** — Python, R are the most used languages by data scientists. Python is useful to create end-to-end product as it can be used to create websites. R is preferred for research purposes.
- For dealing with large amounts of data, open source big data tools like spark, hive, hadoop are useful.

Data Science Life-cycle

Business Requirement

The first step is to define the objective by discussing with customers or stakeholders to identify the business problems and define the target metric for the project.

Collecting the data

The next step is to acquire the relevant data by direct sources like analytics or from third party sources if necessary. High-quality data is an important requirement for a data science project.

Understanding the data

Before training a model, it is important to explore the data first. Most of the data in production have missing values and errors, they should be dealt with domain knowledge and available algorithms. The data may also be normalized and transformed for better model training.

Creating a model

Out of all the columns available in the dataset, choosing the relevant columns is an important task, this is called feature engineering. It needs exploration of data and domain expertise to decide on the features to use for training the model.

Based on the problem statement of the project, there are different types of models available to choose from. The models can be compared with each other by metrics like accuracy.

- **QUESTION ANSWER SESSION**

  - The speaker answered questions-based on:

  1. How to build career on Data Science

  2.What are the building blocks of data engineering?

The outcome of this program is that the students were provided good knowledge about the future scope of Data Science and Building blocks beyond it.

Class Teachers:

1.Ms.Shubhi Srivastava

2.Ms. Sri Harshini

3.Mr.Kiran Kumar Bonthu

Alumni Talk Coordinator: Mrs. Krishnaveni A HOD-ISE