Data Science Project Lifecycle

KHANDOKAR FAISAL ISLAM

Table of contents

Data Science:

- Data science is the discipline of information technology that deals with large volumes of data.
- It uses modern tools and techniques to derive insights from both structured and unstructured data.
- Data science applies several analysis techniques including data mining, machine learning, mathematics, statistic, and predictive analysis [1].

Machine learning:

Machine learning is an area of study that enables the computer system to learn automatically without explicit programming. It also enables the machine to learn from experience [2]. It concentrates on the development of computer programs that can access data and learn from it. It is an application of artificial intelligence. There are four types of machine learning [3].

- Supervised Machine Learning
- Unsupervised Machine Learning
- Semi-supervised Machine Learning
- Reinforcement Machine Learning

Knowing the project:

It is very essential to understand the project and business requirements before starting the data science project. The data scientist must have a clear concept about the business or organization they want to improve by applying data science steps. Must define the aim, objectives, and artifacts that help to complete the project. The data scientist must identify the variables that need to predict [1].

Data collection:

The primary step of the Data Science project is collecting data relevant to the business goal. It is not necessary to be a data scientist, anyone who has a clear understanding of various data sets and can identify the sources can perform this task [4]. Popular sources for collecting data are as follows:

- Kaggle, UCL machine learning repository, WHO(world health organization) repository, UN data, Google public data explorer, and so on [5].
- Using different API or application program interface, it is a popular method to collect data using a set of protocols [5].
- Government website like Data.gov, data.gov.bd, U.S Census Bureau and, etc

Ethical and Legal Requirements

- Data should not be acquired without the consent of the respective individual or organization.
- Make sure dataset collected for the data science project suitable for project goals.
- While collecting data it is important to follow the laws and regulations set by GDPR(the European general data protection regulation)
- ISO 27001 standard should be maintained to protect data.
- A data scientist must aware of the outcomes that can impact others. **Biased prediction should be avoided.**
- **The dataset and the prediction model used in the project must be valuable for the business and individual. Make sure it won't create problem to others**

Data Cleansing

- Collected data always comes in raw form like txt, CSV, JSON therefore data cleansing is required to read, explore, and visualize the data.
- It is the most time-consuming part of data analysis.
- Data cleansing helps to remove duplicate values or unwanted data and it also fixes structural errors [6].
- Unwanted outliers and missing or null values were also detected and removed during data cleansing.
- categorical data also converts to numeric through data cleansing.
- Data cleansing helps to produce more accuracy while applying the model [7].
- Libraries like Pandas, Numpy helped data cleansing in python.
- Packages like reader, tidyr, tidyverse helped data cleansing in R.

Data Exploration

- Data exploration is another important step that comes just after data cleansing. It involves summarizing the main characteristic of the dataset [8].
- It provides a clear understanding of the structure, values, distributions of the data set.
- Through data exploration, the relationships within different variables are usually checked.
- **Pandas library used in python to explore the dataset. Readr, Dplyr package import in R to explore the data.**

Visualization:

- Data visualization is one of the crucial stages in data science. It provides a clear idea about the dataset using the graph, chart, histogram, and heatmap [9].
- Through visualization, it is easy to identify trends, relation, patterns, and the outlier in the datasets
- Data visualization can be performed in both python and R using different libraries and packages
- Matplotlib and seaborn usually used to visualize in python and ggplot 2 used in R

Modeling:

- To predict useful insights and uncover trends from the dataset it is essential to apply a machine learning model [10].
- There are various machine learning models available developed by the data scientist. Different models can be used for a different purpose.
- Data scientists create good models by train and testing using different algorithms and compare their performance.
- Popular machine learning algorithms are Linear Regression, Logistic Regression, Random forest, support vector machine, neural network, and so on [11].
- During Molding four types of error such as training set error, testing set error, validation set error, and train dev set error analyzed
- To evaluate the prediction of the model RMSE(root mean square) is used. RMSE contains both precision and accuracy.

Train and testing model

- Data scientists usually split the data set into three parts. 60% used for training model, 20% used for validation, and 20% for testing the model
- In python dataset splitting performed using train_test_split() imported from scikit-learn libraries. Dplyr, caret, CaTools library used to split the dataset [16].
- A training model is a procedure to train the machine learning algorithm using the dataset. It reveals the relationship between features and target variables [12].
- Validation is generally applied to see the relationship within the known outcomes for the target variable and dataset features.
- Test data is the collection of observations used to evaluate the performance of the machine learning model in the real-world using some measures.
- Skitlearn library is used to perform train, validation, and testing in python.

Accuracy check:

There are four ways machine learning model accuracy has checked:

- Accuracy: The effectiveness of the machine learning model is measured by accuracy. It is the parameter that is used to predict the correctness of the model.
- Precision: Precision is a parameter that decides the right number of positive predictions.
- Recall: The recall is the number of valid results that should be returned divided by the number of results.
- F1 score: It is an accuracy parameter for the model and used to assess binary classification systems. It classifies instances into ‘positive and negative.
- Confusion matrix: summarizes the results of a classification problem prediction. The number of accurate and incorrect forecasts is summed up and divided into count values by class. It is the key to the matrix of uncertainty [13].

Deployment

Once the machine learning model is ready and performed well, it can be deployed in various platforms such as online websites, spreadsheets, dashboards, cloud platforms like Azure, google cloud using API [14] .

Re-Evaluate

The accuracy of the model will not remain the same forever. Thus, it is essential to Re-Evaluate the model. By adding a new dataset and retraining and testing Machine learning model can be improved [15].

Thank You

References

- [1]"Data Science Life Cycle - Data Science Project Management", *Data Science Project Management*, 2021. [Online]. Available: https://www.datascience-pm.com/domino-data-science-lifecycle/. [Accessed: 04- Jan- 2021].
- [2]K. Hao, "What is machine learning?", *MIT Technology Review*, 2018. [Online]. Available: https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/. [Accessed: 25- Dec- 2020].
- [3]Abdi, Asad. (2016). Three types of Machine Learning Algorithms. 10.13140/RG.2.2.26209.10088.
- [4]"How to Collect Data for Your Analysis", *Medium*, 2021. [Online]. Available: https://towardsdatascience.com/how-to-collect-data-for-your-analysis-a8bc58043e64. [Accessed: 04- Jan- 2021].
- [5]"These Are The Best Free Open Data Sources Anyone Can Use", *freeCodeCamp.org*, 2021. [Online]. Available: https://www.freecodecamp.org/news/https-medium-freecodecamp-org-best-free-open-data-sources-anyone-can-use-a65b514b0f2d/. [Accessed: 04- Jan- 2021].
- [6]"8 Ways to Clean Data Using Data Cleaning Techniques", *Digital Vidya*, 2021. [Online]. Available: https://www.digitalvidya.com/blog/data-cleaning-techniques/. [Accessed: 04- Jan- 2021].
- [7]"Data cleaning: The benefits and steps to creating and using clean data", *Tableau*, 2021. [Online]. Available: https://www.tableau.com/learn/articles/what-is-data-cleaning. [Accessed: 04- Jan- 2021].
- [8]"Data Exploration - A Complete Introduction | OmniSci", *Omnisci.com*, 2021. [Online]. Available: https://www.omnisci.com/learn/data-exploration. [Accessed: 04- Jan- 2021].

References

- [9]"What is data visualization and why is it important?", *SearchBusinessAnalytics*, 2021. [Online]. Available: https://searchbusinessanalytics.techtarget.com/definition/data-visualization. [Accessed: 04- Jan- 2021].
- [10]*Datasciencegraduateprograms.com*, 2021. [Online]. Available: https://www.datasciencegraduateprograms.com/data-modeling. [Accessed: 04- Jan- 2021].
- [11]C. Codes), "Commonly Used Machine Learning Algorithms | Data Science", *Analytics Vidhya*, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/. [Accessed: 04- Jan- 2021].
- [12]"Training and Test Sets: Splitting Data | Machine Learning Crash Course", *Google Developers*, 2021. [Online]. Available: https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data. [Accessed: 05- Jan- 2021].
- [13]"Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures - Exsilio Blog", *Exsilio Blog*, 2021. [Online]. Available: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/. [Accessed: 05- Jan- 2021].
- [14]"Building and Deploying a Data Science Project in Two weeks", *Medium*, 2021. [Online]. Available: https://towardsdatascience.com/building-and-deploying-a-data-science-project-in-two-weeks-3c63f0acdab1. [Accessed: 05- Jan- 2021].
- [15] Data science project evaluation , 2021. [Online]. Available: https://www.kdnuggets.com/2017/09/evaluating-data-science-projects-case-study-critique. [Accessed: 05- Jan- 2021].
- [16]D. Singh and S. R, "Splitting and Combining Data with R | Pluralsight", *Pluralsight.com*, 2021. [Online]. Available: https://www.pluralsight.com/guides/splitting-combining-data-r. [Accessed: 05- Jan- 2021].
- [17]A. Smith, "7 Fundamental Steps to Complete a Data Analytics Project", *Blog.dataiku.com*, 2021. [Online]. Available: https://blog.dataiku.com/2019/07/04/fundamental-steps-data-project-success. [Accessed: 05- Jan- 2021].