International Journal of Modern Electronics and Communication Engineering (IJMECE) ISSN: 2321-2152
Volume No.-7, Issue No.-3, May, 2019

Implementing an End to End Solution for Data Science Project Cycle-A Complete Roadmap for Data Aspirants

Ekta Maini
Department of Computer Engineering
Dayananda Sagar University
Bengaluru, India
E-mail: ekta.marwaha@gmail.com

Bondu Venkateshwarlu
Department of Computer Engineering
Dayananda Sagar University
Bengaluru, India
E-mail: bonduvenkat-cse@dsu.edu.in

Abstract: — A huge amount of meaningful information and knowledge can be mined from the raw data using data science. The capability to transform data into meaningful insights is the primary key to gain competitive edge for any organization. Data Science is considered as the most attractive job for 21st century. The job of a data scientist begins with collection of raw data and munging on it iteratively till meaningful insights are obtained. Though there are a large number of technical papers available on data science, yet it has been observed that there is a lack of enough literature for a data science aspirant in developing an end to end solution for a data science project. The aim of this paper is to provide a standardized way to enable a budding data scientist to unravel meaningful insights and data driven evidences that can benefit organizations in a significant way and provide a competitive edge. The phases of a data science project cycle i.e. data extraction, data pre-processing, building predictive models and exposing these models as APIs to real time integration have been discussed in this paper. This paper also emphasizes on a few challenges faced by a data scientist and the best practices to deal with them. Thus, this paper shall help a beginner to kick start his journey as budding data scientist.

Keywords: Data science project cycle, Data extraction, Exploratory Data Analysis, Predictive model, API

I. INTRODUCTION

Massive amount of data is being generated across the world. Nearly 2.3 trillion gigabytes of data are being generated daily and it is estimated that 40 trillion gigabytes would be generated by the year 2020[1].But this data is useless if we cannot transform it into meaningful information. Data Science is the set of fundamentals which help to generate valuable insights from the raw data. This can help an organization to have a better competitive edge That is the reason why Data Science is one of the most exciting field these days. According to Harvard Business Review, data scientist is one of the sexist jobs in 21st century.Fig.1.1 illustrates the data science project cycle.

A. Ask an interesting question

Asking an interesting question is both an art as well as a science. One should have a clear idea about the goal of the organization as well as the limitations of the data. Identifying the issue is the hardest challenge faced by a data scientist. What do we wish to predict? What is the meaning of success for the organization? What should be the parameters to evaluate the success? All such queries should be answered in this phase of project cycle.

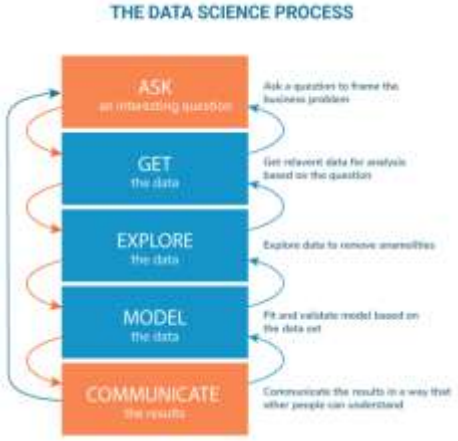

Figure1.The Data Science Process

B. Data Extraction

Data can be collected from a variety of sources. It can be extracted from files or can be scraped from the websites. It can be downloaded from company's database or may need to be taken from some forms/surveys. Data may also be derived from some device through the available ports. In this paper, we discuss how the data can be extracted from the databases and web scraping.

There are a variety of databases that exist in the market. SQL Server, MySQL and SQLite. Python can be an effective language to study data science. The following steps are executed to fetch data from a database.

a. Import the package (import package name)

e.g. sqlite3 is the package for SQLite database while pymysql is used for MySQL Database and pymssql for Microsoft SQL Server database.

b. Make a connection to the database using connection string

(connection=package_name. connect(con_str))

c. Create a cursor to read the data from database.

Cursor=package_name. cursor ()

d. Execute the query

Cursor. Execute(query)

e. Fetch the query results and iterate over these.

cursor. fetchall ()

f. Close the connection.

Connection. Close ()

Sometimes, the data can be scraped from the web also. However, it is always advisable to make sure that we have the permission to take data before attempting to scrape the data from a website. This shall prevent falling trapped in some legal issues. Python libraries 'Request' and 'Beautiful Soup 'are quite useful to fetch data by web scraping. These can be used to extract body, text, title, h1 tag etc. using 'find' function.

From where to get the public datasets?

It is recommended to work on a variety of datasets to understand the intrigue details of different problems. Many datasets can be obtained from US government open data initiative[2]. Large datasets can be obtained from Amazon AWS public datasets. UCI machine learning repository also hosts large number of data sets. Github repository is also a good collection of datasets. It is advisable to work on many datasets to ensure better learning. In this paper, a 'Titanic disaster dataset 'and 'Iris dataset' have been taken from Kaggle website.

C. Explore and process the data

Raw data is not suitable to build the predictive models. It is necessary to process the data to get valuable insights. It is the most time-consuming phase. Data scientists face many challenges in cleaning the data. Fig 2 shows that exploring and processing the data is the most time-consuming task of a data scientist. As it is obvious from this figure, Data cleaning and organizing is the most time-consuming activity. Data can be available from a variety of sources. A complicated problem needs an intense model with more crucial model parameters which in turn means more data requirement. It is very challenging to find quality data to train such models. The path of converting raw data into processed data is iterative in nature as illustrated in the Fig.3.

Exploratory data analysis is done to explore the data using basic statistics concepts. This helps to identify any missing values or outliers in the data. Data munging is carried out to look out for the potential issues and take care of these issues. missing values may either be neglected or may be replaced by the mean or median.

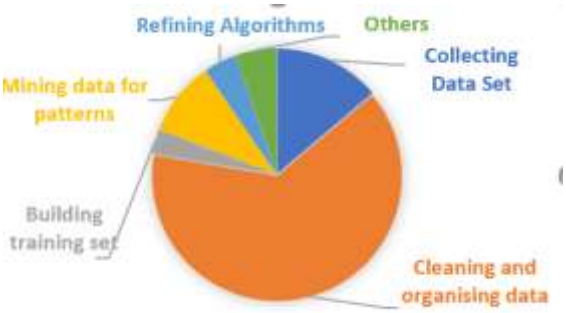

Figure2 Tasks performed by a data scientist

Feature engineering refers to process of generating more features that can be used for modelling. Finally, advanced visualizations are carried out to gain more insights into the data. These visualizations serve as an important component for final presentation. All these steps can be repeated iteratively till a desired value is obtained.

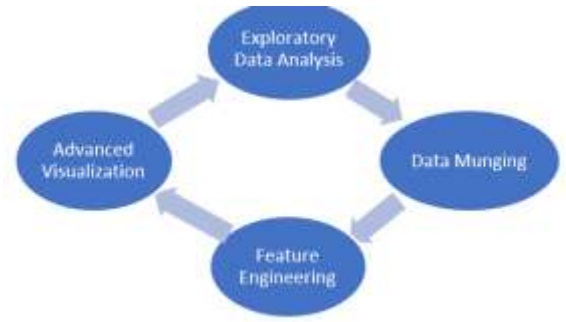

Figure3. Iterative steps in data exploration and processing

EDA comprises of the components as shown in Fig 4.

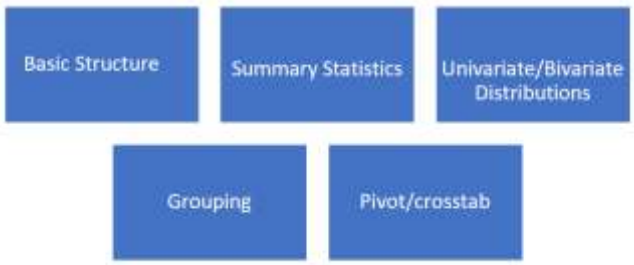

Figure4 Components of Exploratory Data Analysis

Basic structure is related to identification of total number of observations in the data set, number of columns/features in the data science and exploring the data types of columns.

Basic statistics deals with the measurement of measures of central tendency (mean, median, mode) and dispersion (standard deviation, variance, percentiles) for numerical features .In case of categorical variables ,it deals with the total count, unique count and proportions.

Distributions like histograms, KDE and scatterplot provide better insights to the attributes of the data. Grouping/Pivot/crosstab are used to identify how one attribute varies with respect to the other in a data set.

Data Mungıng is done by the data scientists to deal with problem of missing values and outliers. The missing values could either be replaced with mean, median or the mode.

Feature engineering is a process to transform raw data into better representative features to create better predictive models. With a domain expertise, better features can be created. Domain knowledge and technical expertise make feature engineering an art.

Almost all machine learning algorithms expect the parameters to be in numerical form Categorical attributes are converted into numerical ones using categorical feature encoding[3]. Binary encoding can be used to encode the gender information males can be made 0 while females as 1. But for multilevel variables, level encoding is used. e.g. if there are three levels low, medium or high then these can be encoded as 0,1,2. Besides these, one hot encoding is usually used to convert the categorical features into numerical ones.

D. Building the predictive model

Machine learning concepts are applied to learn from the examples and generate patterns which can be used to build the predictive models. The complete dataset is split into train and test data set. There are a variety of machine learning algorithms like Linear regression, logistic regression, decision trees, random forest, boosting, bagging etc. which are applied to training dataset and build the predictive model[4].

The performance of the model is evaluated on the test data set. The performance metrics can be accuracy, recall, precision in a confusion matrix. It is however considered to be a best practice to create a baseline model before applying the machine learning algorithms. Scikit Learn library is a powerful tool to develop machine learning algorithms.

The model can be fine-tuned to avoid the issues of underfitting and overfitting using the technique of Regularization.

E. Communicate the Results

The results obtained need to be communicated to the concerned people so that they can get better understanding and can thus take suitable actions from these insights .Using matplotlib and seaborn library, wonderful graphs can be made.

Once the model is ready, it can be saved on disc so that it ca be used whenever it is desired. This prevents the need of retraining the model every time and the model can also be shared without sharing the data or the code. Pickle library can be used for model persistence. Fig 5 show the model persistence. Flask library is used to create API in Python.

The task of machine learning API is to return model predictions when input data is given to it. Client send the input data wrapped in HTTP request for which the prediction needs to be done. API hosted on the server extracts the input data from the request object and then uses the persisted model to make predictions on the input data. These predictions are sent back to the client wrapped in HTML response.

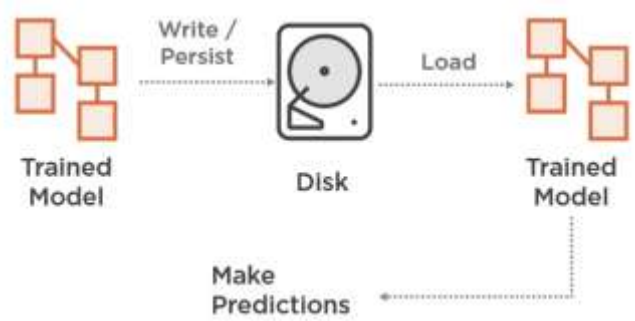

Figure5 Model Persistence

II. CASE STUDY EXAMPLES

A. Titanic Disaster Dataset

This data set is taken from Kaggle datasets. Titanic ship sank it in its maiden voyage in 1912.The task is to apply the tools of machine learning to predict which passengers survived the tragedy.

B. Iris Dataset

Originally published at UCI Machine Learning Repository: Iris Data Set, is small dataset from 1936 is often used for testing out machine learning algorithms and visualizations (for example, Scatter Plot). Each row of the table represents an iris flower, including its species and dimensions of its botanical parts, sepal and petal, in centimeters. The task is to identify the class of the flower.

III. IMPLEMENTATION TOOL-PYTHON

Though there a variety of tools and languages that can be used to study machine learning, yet it is advisable to use Python for the beginners. Python has many libraries and frameworks that make coding easy. NumPy, used for scientific computation, SciPy for advanced computation, and Scikit-learn for data mining and data analysis, are among the most popular libraries.

Python is an open-source programming language and is supported by a lot of resources and high-quality documentation. It also boasts a large and active community of developers willing to provide advice and assistance through all stages of the development process.

Python's simple syntax and readability promote rapid testing of complex algorithms and make the language accessible to non-programmers. It also reduces the cognitive overhead on developers, freeing up their mental resources so that they can

concentrate on problem-solving and achieving project goals. Finally, the simple syntax makes it easier to collaborate or transfer projects between developers.

IV. MACHINE LEARNING ALGORITHMS-A REVIEW

A variety of machine learning algorithms are used for prediction. A few of them are linear regression, logistic regression, decision trees, boosting, bagging and ensembling. Scikit learn library from python supports the implementation of these algorithms with ease. In this paper, Logistic regression has been used to predict the results for the datasets (after thorough data cleaning and data preprocessing)[5].

Logistic Regression is one of the basic and popular algorithms to solve a classification problem. Logistic regression predicts the probability of an outcome. The prediction is based on the use of one or several predictors (numerical and categorical)

The logistic function is a Sigmoid function, which takes any real value between zero and one. Fig 6 illustrates how the probability of the outcome varies between 0 and 1.

Fig. 6 Sigmoid function for logistic regression

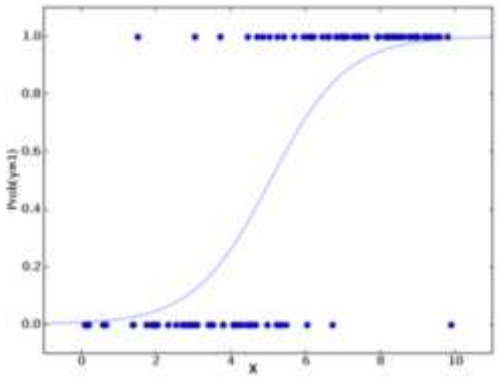

Figure6 Sigmoid function for logistic regression

V. RESULTS AND DISCUSSIONS

The 'Titanic' dataset was extracted from the Kaggle repository while the 'Iris' dataset was extracted from UCI machine library. These are csv files. Using the pandas and numpy features of Python, these datasets are downloaded into Jupyter notebooks.

The first step is to gather the information about the data set. All the attributes were explored. Fig 7a and 7b illustrate the brief ideas about Titanic dataset and Iris dataset, respectively.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1309 entries, 1 to 1309
Data columns (total 11 columns):
Age          1046 non-null float64
Cabin        295 non-null object
Embarked     1307 non-null object
Fare         1308 non-null float64
Name         1309 non-null object
Parch        1309 non-null int64
Pclass       1309 non-null int64
Sex          1309 non-null object
SibSp        1309 non-null int64
Survived     1309 non-null int64
Ticket       1309 non-null object
dtypes: float64(2), int64(4), object(5)
memory usage: 122.7+ KB
```

Figure7a Attributes of Titanic dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
Id               150 non-null int64
SepalLengthCm    150 non-null float64
SepalWidthCm     150 non-null float64
PetalLengthCm    150 non-null float64
PetalWidthCm     150 non-null float64
Species          150 non-null object
dtypes: float64(4), int64(1), object(1)
memory usage: 7.1+ KB
```

Figure7b Attributes of Iris dataset

It is clear from the above figures that Titanic dataset has 1309 observations with 11 columns each while Iris dataset has 150 observations with 6 columns each. There are no missing values in Iris dataset while it is observed that there are a few missing values in columns like 'Age', 'Fare', 'Embarked' etc.

Data scientists need to explore the data in depth to gain a better understanding of the subject and need to decide what needs to be done to tackle the issue of missing values etc. The statistical information of both the datasets is represented in Fig8a and 8b.

	Age	Fare	Parch	Pclass	SibSp	Survived
count	1046.000000	1308.000000	1309.000000	1309.000000	1309.000000	1309.000000
mean	29.881138	33.295479	0.385027	2.294882	0.498854	22.614209
std	14.413493	51.758668	0.865560	0.837836	1.041658	32.471067
min	0.170000	0.000000	0.000000	1.000000	0.000000	0.000000
25%	21.000000	7.895800	0.000000	2.000000	0.000000	0.000000
50%	28.000000	14.454200	0.000000	3.000000	0.000000	1.000000
75%	39.000000	31.275000	0.000000	3.000000	1.000000	70.000000
max	80.000000	512.329200	9.000000	3.000000	8.000000	70.000000

Figure8 a Statistical measures of 'Titanic dataset'

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.054000	3.758667	1.198667
std	43.445368	0.828066	0.433594	1.764420	0.763161
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	38.250000	5.100000	2.800000	1.600000	0.300000
50%	75.500000	5.800000	3.000000	4.350000	1.300000
75%	112.750000	6.400000	3.300000	5.100000	1.800000
max	150.000000	7.900000	4.400000	6.900000	2.500000

Figure8 b Statistical measures of 'Iris' dataset

It is very easy to plot a variety of graphs in Python. Fig 9 shows a bar graph of the number of passengers travelling in first, second and third class in Titanic. It clearly reflects that the majority of people were travelling in third class. Similar graphs were plotted for the other attributes too. Histograms are also of great interest to the data scientists as these reflect insights into the frequency table of an attribute.Fig.10 shows the histogram of 'age' in Titanic dataset. It reflects how many people (travelling in Titanic) are in a particular age group. Such information attracts the data scientists for better insights

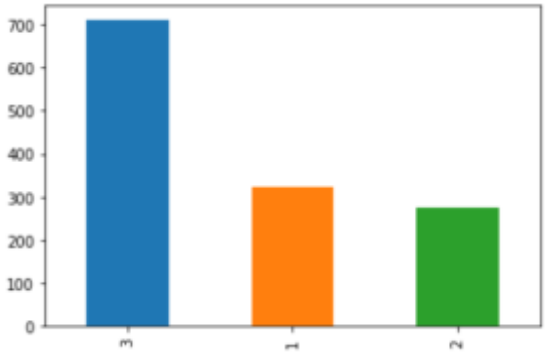

Figure9. Number of passengers in different classes in Titanic

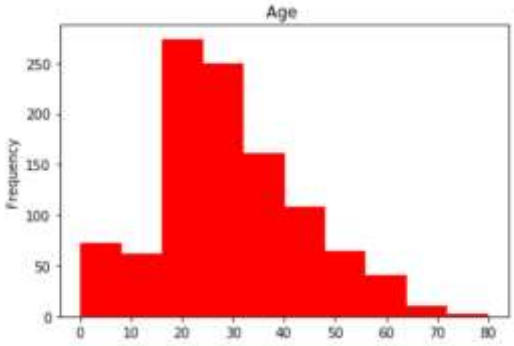

Figure10. Histogram of 'age' attribute in Titanic dataset

Similarly, histograms of all the attributes in 'Iris' dataset are illustrated in Fig 11

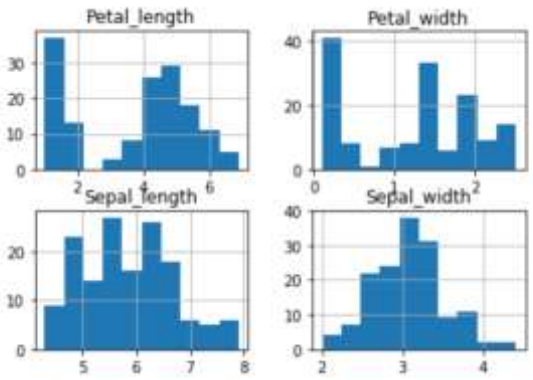

Figure11 Histograms of Iris dataset

The data scientists should be very keen in understanding the relationship patterns between the various attributes.A scatterplot is a good way to show the change of an attribute value with respect to the other.Fig 12 illustrate the scatter plots for Titanic and Iris dataset respectively.

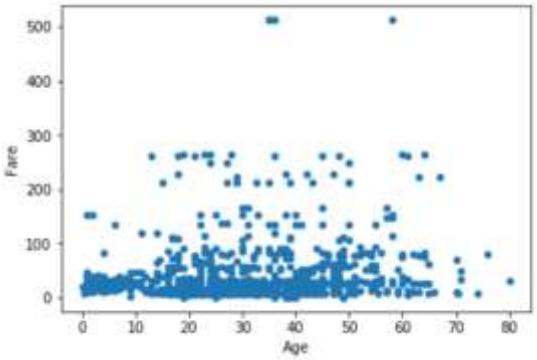

Figure12 Fare' vs 'Age' in Titanic dataset

Missing values were replaced with the appropriate median values. One hot encoding was performed for categorical variable in Titanic dataset. Since there were no categorical attributes and no missing values in Iris dataset, these steps were not performed for this dataset. Scikit learn is a powerful library to develop predictive models. The datasets were split into training and test datasets. The training dataset used to train the model while the accuracy was checked on the test dataset. Logistic Regression was used to predict the survival and the class in Titanic and Iris dataset, respectively. The accuracy was observed to be 83%in Titanic dataset while it was as high as 96% in Iris dataset. Further, the model persistence was carried out and APIs were hosted.

VI. CONCLUSION

Data Science is the most attractive field in this era. A huge amount of data is being generated daily. Data science can effectively be used to generate valuable insights which can be used by the organizations to take suitable decisions which in turn can help to gain competitive edge. This paper was written to enable the data aspirants to develop an end to end solution for a data science project. The paper also discusses the few challenges faced by data scientists and how to overcome those with a few best practices. This paper shall serve as a step by step guide for the beginners who wish to unravel the mysteries of data science.

REFERENCES

[1] Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong," Big Data for Health" IEEE Journal of biomedical and health informatics, vol. 19, no. 4, july 2015

[2] Yoo , Patricia Alafaireet , Miroslav Marinov ,Keila Pena-Hernandez , Rajitha Gopidi ,Fu Chang , Lei Hua," Data Mining in Healthcare and Biomedicine: A Survey of the Literature" J Med Syst (2012) 36:2431–2448

[3] Anuradha J," A Brief introduction on Big Data 5Vs characteristics and Hadoop Technology" Procedia Computer Science 48(2015)319-324, Elsevier

[4] Mathew, P.S., Pillai, A.S.: Big data challenges and solutions in healthcare: A survey. In: Innovations in Bio-Inspired Computing and Applications. Springer (2016) 543–55

[5] Ekta Maini,Bondu Venkateshwarlu,Arbind Gupta," Applying Machine Learning Algorithms to Develop a Universal Cardiovascular Disease Prediction System "International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI),2018

`AUTHOR'S BIOGRAPHIES

Ekta Maini holds B.Tech (with honors)degree from Kurukshetra University. She held first rank in ME from Panjab University. She is pursuing her PhD degree from Dayananda Sagar University. Her areas of interest include machine learning, data mining and artificial intelligence. She has authored many papers in reputed journals.

Dr. Bondu Venkateshwarlu received his Ph.D. degree from Department of Computer Science and Systems Engineering at the same university in 2016. He has authored several papers in international conference proceedings and refereed journals. His current research interests include Data Mining, Soft Computing Techniques, Software Engineering and Image processing.