FOUNDATIONAL OF VARIOUS DATA SCIENCE FACET AND FUNDAMENTAL ELUCIDATING OF ITS ARTISANS ILK

Kartikay Laddha, student B.Tech Data Science (Business Analytics), NMIMS University, Mumbai, India, laddhakartikay8@rediffmail.com

JayKrishna Joshi, Asst. Professor, NMIMS University, Mumbai, India, jaykrishna88@gmail.com

Abstract - This paper gives a very rudimentary idea about how "DATA" can be effectuate and the immense information can be gained from it which could be engineered by all types of Businesses. It also gives information on the three exciting realms with simple exemplar. Also, rendition of how a Business Organisation has various hierarchy and designated posts/positions allotted with respect to data science unit along with the software contrivance requirement is highlighted.

Keywords – Data Science, Deep Learning, Artificial Intelligence.

I. INTRODUCTION

Data Science is a Dynamic field. It is currently the most innovative and buzziest term used all around the globe. Tons of information about it is discussed on numerous platforms day in and day out, sometimes with a lot of complex technical explanation. However, data science actually can be explained in a very simple way (one out of the many ways). It is a set of methodologies for taking in thousands of forms of data that are available in many possible ways and using them to draw meaningful conclusions that is data gets transformed into information. Data is omnipresent. Every like on social media, click on digital platform, email browsing, credit card sweep or tweet is a new piece of data that can be used to describe the present or personality as well as predict the future in the most possible efficient way. Data science is being accepted and implemented extensively in every organisation and institution – from the lowest to the largest because of the immense enhancement it can provide to them. Depending on the nature of the company, different tools to perform various tasks associated with data science are being used. As mentioned earlier, the concept of data science can be used and is rather used for strengthening the establishment.

Some of the elementary things which are done includes:-

- Data is used to describe the current state of an organization process, this could be accomplished with dashboards or alerts, simplifying time intensive reporting process with new data technology.
- Data Analysis helps detect anomalous events. Using past data (and applying various techniques on them) detection of new events (that is unexpected) can be obtained.
- Data interpretation can also diagnose the causes of absurd events and behaviours rather than determining correlations between small numbers of events. Data science techniques helps in understanding the complex systems with many possible causes.
- Data anatomy can also help in predicting the future events, using innovative techniques to take various causes to account and predict potential outcomes. Further it is used to evaluate the probability of prediction mathematically, to clarify the level of uncertainty.

Let us discuss a simple quotidian workflow example. In order to buy a car, suppose we visit a car dealership outlet and fill out the information form. This analogue information is entered into the computer and now it is available in digital form. This digital data is combined with other data from hundreds of car dealership store spread all across into one single central database. Various data cleaning techniques are applied and clean data is now available on that single database spot. Authorised personal can use this information and can link it to many open source information. The email address that we provided when we bought that car can be used to tie our car purchase data with our data from social media, or web browsing. This can manifest a complete picture about every individual customer who purchased a car, their ages, their likes, dislikes, friends and family. This information can be used to predict copious things like what other purchases we are likely to make or how best they can sell us the insurance with offers etc. Hence it is said that Data is generated everywhere and its incredibly valuable for business if used with correct dexterity.

Data science, generally can be divided into three steps for any project (workflow):-

- Firstly, collecting data from many sources (Data Collection), such as customer surveys, web traffic results, emails between sales representatives and potential clients and financial transactions.

DOI : 10.35291/2454-9150.2020.0397

- Next, exploring and visualising the data (Exploration and Visualisation), this could involve the building of dashboards, to track how our data changes over time or perform comparisons between two sets of data and then.
- Final stage is of predictions (Experimentation and Predictions) with the help of that data for example, this could involve building a system that segments customers or classifies pictures of different types of cars.

II. APPLICATIONS OF DATA SCIENCE

Let's understand the three exhilarating areas of data science:

- Traditional Machine Learning

Suppose we are working at the fraud detection department in a large bank, then one of our primary aim would be to use data for determining the probability whether the current occurring transactions are fake or not. To begin, we start by gathering information about each purchase, such as the amount, date, location, purchase type and card holder's address. We will need many examples of transactions including this information as well as a label that tells us whether each transaction is valid or fraudulent, luckily, all this information in available in the our database - all these records are called training data and are used to build an algorithm.

Each time a new transaction occurs we will give our algorithm information like amount or date and it will answer the original question what is the probability that this transaction is fraudulent.

What do we need for machine learning to work?

I. A data science problem begins with a well-defined question.
   - What is the probability that this transaction is fraudulent?

II. Next, we need some data to analyse, a set of example data.
   - We had months of old credit card data transactions in associated meta data that had already been identified, as either "Fraudulent" or "Valid".

III. Finally, we need additional data every time we want to make new predictions,
   - We need to have same kind of information on every new purchase, so that we could label it as fraudulent or valid. New credit card transactions

- Internet of Things (IOT)

Now suppose we are trying to build a smartwatch for monitoring physical activity, we want to be able to auto detect different activities such as walking or running. Now the smart watch is equipped with a special sensor called an accelerometer, that monitors motion in three dimensions, the data generated by this sensor is the basis of our machine learning problem. We can ask several volunteers to wear our watch while running or walking to help record the data, we can then develop an algorithm that reorganises accelerometer data as representing one of those two states - walking or running.

Our smart watch is part of a fast growing field called the internet of things (IOT), which is often combined with data science. IOT refers to gadgets that are not standard computers but still have the ability to transmit data. This includes:

- Smart Watches
- Internet connected home security systems
- Electronic toll collection system
- Building energy management systems, and much more

IOT data is great resource for data science projects.

Learning

A key task for self-driving cars is identifying when an image contains humans, the dataset for this problem would be so that, we would be expressing the picture as a matrix of numbers where each number represents a pixel. However this approach would probably fail if we fed the matrix into a traditionally machine learning model as there is simply too much of input data. Hence we need a more advanced algorithm that is known as deep learning.

In deep learning multiple layers of mini algorithms called neurons work together to draw complex conclusions. This technique requires much – much more training data then a traditional machine learning model and is also able to learn relationships that traditional models cannot. Deep learning is used to solve data intensive problems such as image classifications or language learning/understanding.

III. DATA COLLECTION AND STORAGE

Data Sources and Risks

We know data is everywhere and almost every business process can generate tonnes of amounts of data.

Some of the common sources of data are:

- Web events
- Customer data
- Logistics data
- Financial transactions, etc.

It's possible that a company is already collecting all of these information and it is best to ask a data engineer about what is collected and what is not. And to emphasise the importance of starting the centralised data collections process sooner rather than later.

- Web Data

DOI : 10.35291/2454-9150.2020.0397

Whenever a user visits a web page or clicks on a link it can be helpful to track this information in order to calculate conversion rates or monitor the popularity the different pieces of content. At a minimum a business would like to collect the name of the event which could be the URL for the page visited or an identifier for an element that was clicked, secondly the timestamp of the event and an identifier for the user that performs the action. Suppose we have a customer who visits our company website and likes our product. We might choose to track the user's name, the timestamp and the object the user clicked on. It is important to understand that the user's name is Personally Identifiable Information (PII).

PII includes a person's name, location, E-mail address and any other piece of information that can be used to tie a web event back to a real human. PII should be treated with extreme sensitivity and caution. One of the easiest way to protect the user's identity is to split the information into two separate entries. We can assign a user_id against the user's name and then store that information in the user's table. We may then identify the user's event using the user_id. We now call the data in the events table pseudonymized because the user cannot be identified with that table alone. But the user can be identifies if we combine data from the event's table and the user's table.

To protect the user, we want to make sure that the access to the user's table is restricted to only folks who needs to know Jane's identity such as senior customer service representatives or members of the legal team. We will also want to periodically audit, who has access this data and how they have used it to ensure that user's data is respected. The best way to protect user's privacy is to destroy the information in the user's table after assigning a user_id to the user. Without the user's table the event's table is fully anonymised data. For many analysis purposes, anonymised data is sufficient. We need to know that the user is a unique individual but we don't need to know. Its name or any. Other PII.

An example for protecting the privacy can be seen as GDPR stands for General Data Protection Regulation and applies to all data inside the European Union. The purpose of this is to give individuals control over their personal data. Among other things GDPR regulates how long data can be stored, mandates appropriate anonymization and requires data collection. To be disclosed and consent to be obtained. It's always best to consult a lawyer when dealing with any data inside the EU to ensure that we comply with GDPR.

- Solicit Data

The data which we obtain by asking the customers for their opinion known as Solicit data. Solicit. Data can be used to create marketing Collateral such as a post about what percent of our customers are satisfied with our product. It can also be used to reduce the risk in decision making process, such as when we survey users to gauge interest in our new products. Finally it can be used to monitor quality, such as when a company asks it's users to rate their products. Common types of solicited data includes, Surveys, Customer reviews, In-app questionnaire, and focus groups.

Solicited data can be:

- Qualitative such as conversations and open-ended questions, this data is very subjective and requires a lot of analysis. In general collection of small scale qualitative data is used in generating hypothesis.
  For example – a focus group might provide some ideas about what features we might want to build.

- Quantitative such as multiple choice questions or rating scales, this data can be easily summaries in a graph or chart. Larger scale quantitative collection is needed to validate the above mentioned hypothesis.
  For example – we can ask the users to rank a list of features from most desirable to least desirable.

It is important for us to know that solicited data generally tells us our user's data preferences. Stated preference is when someone tells us what they want, or believe somewhat hypothetical and subjective. When a user actually takes an action such as purchasing decisions, we learn their revealed preference. We hope that our users stated preferences are good indicators of their revealed preference but this isn't always the case. Many people have the stated preference of going to the gym frequently, however many of the same people's revealed preference is only to go occasionally. Some people have an entire business model based on the expected difference between people stated and revealed preference for exercise.

Now that we know some types of solicited data and know the pitfalls of revealed over stated preference, let us see how to ask questions to gather information:

- We should try to be as specific as possible while asking questions, this specificity must apply to both the wording of the question and the potential answer choices that we give.
- We should avoid loaded language, especially if it might bias response towards a particular choice.
- Whenever possible calibrate the survey to possible known quantities.
- Finally, it can be tempting to ask as many questions as possible. We must fight that instinct and in fact ensure that every question we ask will help us make a decisive action.

Collecting Additional Data

Collecting internal data which is useful for some data science projects is only one part of the puzzle, often we need to gather data from external sources as well. There are many

DOI : 10.35291/2454-9150.2020.0397

ways in which we can collect the additional for an organisation. A few common methods include:

- Data APIs

API stands for Application Programming Interface, it's an easy way. Of requesting data from a third party over the internet. Many companies have APIs to let the team access the data. Some notable APIs include:

- Twitter
- Wikipedia
- Yahoo! Finance
- Google maps, etc.

- Public Records

Public Records are another way of gathering additional data. In the USA data.gov has health, education and commerce data available for free download. In the EU data.europa.eu has similar data. These can be great sources for understanding population trends, or gathering location and economic trends.

- Mechanical Turk

Depending on what kind of training set is needed, Mechanical turk can also be a great option. Mechanical turk means asking humans to complete a task that we eventually plan on computerising. In an image processing example this would mean labelling a handful of pictures to create a training set for image reorganisation. Rather than asking one person to label thousands of images, we may recruit thousands of people to label a few images. To ensure quality we might ask two or three people to review the same image and take the most common answer. Many platforms exists to help built a mechanical turk problem and recruit helpers such as AWS MTurk. This isn't just for image reorganisation, we can also use this to label customer reviews as positive or negative. Extract text from a form. Or highlight key word inn a sentence.

IV. Data Storage and Retrieval

Let us see some of the efficient ways to store the data which an organisation collects. A business generally creates far more data that could even be stored on a single computer. In order to make sure that all the data is saved and easy to access, we will want to store it across many different computers. A company might have its own kind of storage computers called a cluster, or a server on premises. Alternatively a company may pay another company to store data for it. This is referred to as cloud storage. Common cloud storage providers include Microsoft Azure, Amazon Web Services and Google Cloud. These services provide more than just data storage, they can also help an organisation with data analytics, machine learning and deep learning.

Different types of data requires different data storage solutions.

- Some data is Unstructured, like Emails, Text, video and audio files, web pages and social media this type of data is often stored in a kind of database called as Document Database.

- More commonly, data can be expressed as tables of information like one can find in a spreadsheet. A database that stores information in a tabular form is called a Relational Database.

Both of these kinds of data storages can be found on the cloud storage providers mentioned above. Once data has been stored in a document database or relational database, we will need to access it. At a basic level we will want to be able to request a specific piece of data such as all of the images that were created on a particular date, or all of the customer addresses in Bombay. In addition we might also want to analysis such as summing, counting or averaging data. Each. Kind of database has its own query language, Document Database mainly use NoSQL while Relational Database use SQL. SQL stands for Structured Query Language and NoSQL stands for Not Only SQL.

Storing a company's data is like building a library. First, we need to decide where to build our library that corresponds to choosing a cloud. Either an On-Premises cluster or one of the cloud provider. Next we need to decide what types of shelves to install to keep the books. They types of shelves will depend on the types of books. This is analogous to choosing between a documented database for unstructured data or relational database for tabular data. Just like a library might have multiple types of shelves, we might need to store some data in the document database and the other in relational database. Finally, we will need a system for referencing and checking out books. The way we locate and retrieve each book depends on how that book is stored. Similarly, we need a query language to speak to the database. For document databases we use NoSQL and for relational databases we use SQL.

V. ANALYSIS AND VISUALISATION

Dashboards

We have already seen that how data engineers collect and store data in a database. Now we will dive deep into how data analysts visualise and explore that data using Dashboards. A Dashboard is a set of matrix usually in the form of graphs that update on a schedule. Some dashboards can update in real time but others update daily or weekly. Let us see some of the common dashboards elements:

- One of the most common dashboard element is a Time Series that tracks a value over time. This type of plot shows both a current value of a quantity and enough historical data that gives a context to that value.

DOI : 10.35291/2454-9150.2020.0397

- Another common dashboard element is a Categorical Comparison using a Bar Chart. Whereas time series provides a historical basis for comparison, a bar chart compares different groups during the same time period.
- Generally, we avoid adding tables to dashboards because graphics are easier to read. Display text are an exception. For example – displaying a small number of customer comments is a great way of adding some qualitative data to an otherwise quantitative picture.

There are many ways of creating dashboards, some of them are built with the Spreadsheet tools such as Excel or Google Sheets, and others are built using specialised business intelligence or BI tools such as Tableau, Power BI or Looker. For something really customised some analysts use programming languages or customised tools such as R Shiny or d3.js. all of these tools are great ways to get fast and accurate dashboards. It's best practiced to ensure that everyone in the organisation uses the same one so that there is no confusion about where to go about dashboard information.

Before issuing a request, it's better to be sure that the new dashboard is actually the best solution to the problem. Dashboards are required when we need to access the information many times, or information needs to be updated frequently and the information we need will always be the same. Once we are sure that we need a dashboard, we need to make a request as specific as possible. Do we need a single number of comparison over time, what timeframe is relevant to us, we need to specify our use case, and this can help a data analyst choose a type of dashboard that is best for us.

Ad hoc Analysis

As we know dashboards help data analysts fill a frequent need for the same data. Now we may consider how does an analyst feel for one off request for data. An ad-hoc request is a request for data or an analysis that does not needed to be repeated on a weekly or daily basis. Many different departments make ad-hoc requests from data analysis team, like product might want an end of quarter report on the success of an initiative. Finance might need a list of users who haven't made payments in the past month. Engineering might ask for the total number of clicks on a particular button. Making a good ad-hoc request includes:

- Being specific about the stuff which can be easily defined.
- It should also include the context, providing the context can help analyst. Spot any additional data that might be helpful.
- Finally, a good request should also include a priority level and due date.

This will help the analytics team handle any ad-hoc requests that they receive. For a team manager ad-hoc requests are tricky, because they are unpredictable, and they can steal time away from scheduled work. A good strategy for handling. Ad-hoc requests is:

- By using a Tickering system such as Trello, JIRA or Asana. A Tickering system allows internal customers like product, finance or engineering to submit requests for ad-hoc analysis.
- Ticketing systems can help ensure that the requests are specific and precise and also include a due date along with priority level.

The data team whether isolated, embedded or hybrid can then assign the tickets to an appropriate analysts. Once the ticketing system is in place, an analyst can track the frequency and duration of ad-hoc requests and improve scheduling for future quarters.

A/B Testing

A/B Testing is a type of experiment for de-risking choices between two options, such as changes to a website, addition of new features or wording of email subjects. Suppose we need to decide the title for a blog through an A/B test, then we randomly divide the audience into two groups. Each group sees a different title, eventually we will pick the better title and make it permanent. There are four steps to running an A/B test:

- Picking a metric to track
- Calculating sample size
- Running the experiment
- Checking for significance

First we pick a measurable outcome to track, here we will examine the percent of people who click on the link with the title of the article. Next we will be deciding how long to run the experiment. We will run the experiment until we reach a sample size large enough to be certain that any difference we observe is not due to random chance. The necessary sample size depends on a baseline metric, In this case our baseline metric is how often people generally click on our links to one of our blogs. If this rate is close to 50% then we will need a smaller sample size, if the rate is much larger or much smaller, which is typical for something like clicking a link then we will need a larger sample size. The sample size also depends on how sensitive our tests needs to be. A test sensitivity tells us how smaller the change in our metric we are able to detect. Larger sample sizes allows us to detect smaller changes. One may think that we always need the highest possible of sensitivity, but we actually want to optimise for an amount of sensitivity that is meaningful for our business problem.

For example – if the first link is clicked on by 7% of the viewers, and the second title is clicked on by 7.01% of viewers we don't actually care about that difference, it doesn't affect our profits by enough. Generally we care

DOI : 10.35291/2454-9150.2020.0397

about a relative increase between 10% and 20% of the baseline metric. We now run our experiment until we reach a calculated sample size. Stopping the experiment too early, or running it for too long can lead to biased results. Once we have reached the target sample size, we check our metric. We see some difference between the two titles, but how do we know that difference is meaningful? We check by performing a test of statistical significance. If the results are significant we can be reasonably sure that the difference observed is not due to random chance but to an actual difference in preference. But what if the results aren't significant! If there are any difference in click through rates between the two titles, they are smaller than the threshold we choose while determining the sensitivity. Running our test longer won't help, it will let us detect the smaller difference but we have already decided that those smaller differences are irrelevant to our business problem. It is important to remember that there still might be a difference in click through rates between the two article titles but that difference is not significant to our business problem.

VI. PREDICTION

Supervised Machine Learning

Machine learning is a set of methods for making predictions based on existing data. Supervised machine learning is a subset of machine learning methods, where the existing data has a specific structure it has labels and features. Some problems that can be solves via supervised machine learning includes:

- Recommendation systems
- Email subject optimisation
- Churn prediction

Suppose we have a subscription business and we want to predict whether a customer would likely to stay subscribed or churn. First, we will need some training data. This would be historical data from our customers. Some of these customers would have maintained their subscription while others will have churn. We eventually want to be able to predict a label for each customer either churn or subscribed. We will need features to make this prediction. Features are different pieces of information about each customer that might affect our label. For example, perhaps the age, gender, date of last purchase, date of last visit, household income, and profession will help in predicting the cancellations.

The magic of machine learning is that we can analyse many features all at once. We can use these features and labels to train a model and make predictions on new data. Suppose we have a customer that may or may not churn, we can collect future data on this customer based on the features as discussed above and then we can feed this data to the training model that we built and then that trained model will give us the prediction. If the customer is. Not in danger of churning we can count their revenue for another month. If they are in danger of churning, we can reach out to them with a special promotion or customer support to keep them subscribed.

Supervised machine learning make predictions based on data which has features and labels. Labels are the quantity that we want to predict, in our example whether or not the customer has churned. Features are the data that might help in predicting the labels, such as age, household income, profession etc. Once we have features and labels we can train the model and use it to make predictions on new data. To be assured that the collected historical data is useful to us while we build the training model it is always a good practice to not feed all of the collected data to that model. This withheld data is called as the test set and can be used to evaluate the goodness of the trained model. In our example we can ask the model to predict whether a set of customers would churn and then measure how often the predictions were accurate.

Model Evaluation is also an important aspect in prediction, hence it is important to note how often the model incorrectly predicted that the customer would churn and how often it incorrectly predicted that the customer would not churn. Checking both outcomes is particularly important for rare events.

Clustering

Clustering is a set of machine learning algorithms that divide the data into categories called clusters. Clustering can help us see patterns in messy datasets. Machine learning scientists use clustering to divide customers into segments, images into categories and behaviours into typical and anomalous (Anomaly detection). Clustering is a part of a broader category in machine learning, called unsupervised learning. Unsupervised learning differs from supervised learning in the structure of the training data. While supervised learning uses data with features and labels, unsupervised learning uses data with only features this makes unsupervised learning and clustering particularly appealing. We can use it even when we don't know much about our dataset.

Let us see how clustering helps in customer segmentation, customer segmentation is a process of dividing a pool of customers into different groups with common attributes. We can use these segments to device targeted advertising campaigns or to explain otherwise confusing results by analysing the behaviour of individual segments rather than just looking at the customers as a whole. First we need to brainstorm a list of features that will accurately describe our customers. Let us consider that we are working for an airline and our customers are travellers. Important features might include:

- The number of flights taken in the past year
- The percent of those flights that were international
- How far in advance they typically buy tickets

DOI : 10.35291/2454-9150.2020.0397

- Lastly, what percent of the tickets were business class

Some clustering algorithms need us to define how many clusters we want to create, the number of clusters we ask for greatly affects how the algorithm will segment our data. Having a strong hypothesis about our data helps us get better results from the clustering algorithm. For our airline example we might expect to have business travellers, family travellers and adventures.

Clustering is an unsupervised machine learning method that divides an unlabelled dataset into categories. In order to perform clustering:

- We must first select relevant features of our dataset
- Next, we select number of clusters based on the hypothesis about our data.
- Finally, we use the results of our clustering to solve business problems such as advertising or price selling.

We can even use clustering as a way of breaking up a larger machine learning problem, rather than modelling all of our data once, we can model different models for each of the clusters to get better models.

Special Topics in Machine Learning

Machine Learning has two special fields, Time Series prediction and Natural Language Processing.

- Time Series Prediction:

Time Series Forecasting. Refers to any type of supervised learning where time is an important feature. A good time series. Forecast will account for recent behaviour as well as weekly, monthly and yearly trends. Time series forecasting can help us catch the periodic events known as Seasonality. Seasonality can happen on any time scale. For example – television viewership are lower on Friday nights because many folks wish to go out rather than stay back and watch TV, so this is a weekly trend. Certain spending can spike at the end of the month, when people receive a pay check, this is a monthly trend and ice cream sales are lower in the winters because people don't like to eat cold food. When it is cold outside, this is an annual trend.

- Natural Language Processing(NLP):

NLP refers to any machine learning problem where the data set is text, possible input includes customer reviews, tweets, medical records or email subjects. Understanding text is difficult to define and more difficult to do in practice, but NLP can accomplish many simpler tasks such as classifying sentiments of customer reviews or clustering medical records with similar pathology. Successful NLP depends on having a specific question and creating a good set of features from the input text. Previously the features for a machine learning problems have been numbers or categories, what we do when the data is text. A simple option is to count the number of times important words appear in each piece of text. Although word counts are commonly used in NLP, yet there are a few obvious limitations:

- First, word count don't take into account for negation.
- Word counts don't help us consider synonyms, we would like to consider that each synonym refers to the same thing.

One solution to these problems is Word Embedding. It is a special way of creating features that group together similar words, it will create similar features for various synonyms for the same word. It has another interesting property, their mathematical representation of words that obey intuitive rules. For example- in word embedding if we take the features for king subtract the features for man and add the features for a woman, we get a set of features that are very close to those of queen.

Now we can say that time series forecasting is a special area of machine learning where time is an important feature, it helps us account for periodic trends in our data called seasonality. Another special area is NLP, which uses text input data, two important ways of truing text into features are word count which is simple but imprecise and word embedding which are difficult to implement but can be more precise.

Deep Learning and Explainable Artificial Intelligence:

Deep Learning also sometimes called as Neural Networks or Neural nets is a special type of machine learning that can solve more complex problems, it requires much-much more data than traditional machine learning. It is best used in cases with inputs are less structured, such as large amount of texts or images. One of the main drawback of deep learning is lack of explaining ability, although deep learning can make very accurate predictions, it's not always clear why the model is making a specific prediction. Methods that allow us to understand the factors that lead to each predictions are also known as Explainable AI.

Let us consider a typical example for explainable AI, suppose we are investigating customer cancellations using traditional machine learning model. Our trained model can tell us two things:

- First, it can predict whether or not a given customer is likely to churn.
- Second, it can tell us which features were important in making this decision, this is the explainable part. This additional explaining ability can provide important insights.

For example, we might learn that certain demographics are much more likely to cancel their subscriptions, our

DOI : 10.35291/2454-9150.2020.0397

marketing and customer support teams can change their strategies to outreach these and address this deficiency. Contrasted example with a typical deep learning problem, suppose we want to recognise hand written letters, we don't really care why a particular image is classifies as an 'A' as long as the predictions are highly accurate. Deep learning is a perfect solution to this problem because we don't care about explaining ability and we probably have a large image based set of training data.

Before we choose deep learning a solution to the problem, we should ask ourselves some questions; First, does our training data have many features or is it difficult to understand it as a simple array of features(is the training data complex). Data such as images or text are particularly suitable for deep learning. Secondly, do we have a very large amount of data. Deep learning requires much more data than traditional machine learning. If we don't have millions of examples then deep learning might not work for our company. Finally, does the model need to be predictive or explanatory. Deep learning is great for predictive modelling but can leave us perplex why each predictions were made, simpler models might have less predictive power but can be better when the clarity is essential.

VII. BUILDING A DATA SCIENCE TEAM

Building and structuring of a good data team is very essential to meet the business need of an organisation. It is not very surprising to state that data science isn't a single field. It's actually three different jobs

- Data Engineer

Data Engineers control the flow of information as information architects, they help in building specialised data storage systems and the infrastructure to ensure that the data is easy to obtain and process which they do by maintain the data access.

Most data engineers are very familiar with SQL, which they use to store and manage big and large quantities of data. They also use some of the programming languages such as Java, Scala or Python for processing data and automating data related tasks.

- Data Analyst

Data Analysts describe the present view data, they do this by creating dashboards, Hypothesis Testing and data visualisation. They often have some background in statistics or computer science but tend to have less engineering experience than data engineers and have less math experience than machine learning scientist. Data Analysts use spreadsheets (Excel or google sheets) to perform simple analysis on small quantities of data (simple storage and analysis). They use SQL (the same language used by data engineers), for large scale analysis. While data engineers build and configure SQL storage solutions, data analysts use existing databases to consume and summarise data. Analysts also use Business Intelligence or BI Tolls such as Tableau, Power BI or Looker for creating dashboards and sharing information and their analysis.

- Machine Learning Scientist

Machine learning is perhaps the buzziest part of data science, it's used to predict and extrapolate what is likely to be true from what we already know. These scientists use training data to classify larger unrulier data, for example machine learning can help us tell how much money a stock may be worth in the next week, can help predicting which image contains a car by image processing or what sentiments are expressed using a tweet by automated text analysis or sentiment analysis.

Machine learning scientist either use Python or R programming languages for creating predictive models. These both are great programming languages for data science and a candidate who knows one language can likely read code in the other language. This is to be noted that programming languages aren't as difficult to learn as spoken languages. If someone knows how to speak Hindi, it might take them years to learn to speak Spanish. Programming languages are more similar to power tools. If we know how to use a power drill, we may not necessarily know how to use an electric saw, but we may probably learn with a little training or help.

Now that we have seen that each position uses a slightly different set of tools to achieve their goals.

Data Engineer	**Data Analyst**	**Machine Learning Scientist**
Store and maintain data	Visualise and describe data	Model and predict with data
SQL + Java/Scala/Python	SQL + BI Tools + Spreadsheets	Python/R

- Data Science team Structure

Once a business organisation hires some data professionals, there are three main ways a data team can be structured.

Isolated

An isolated type of data team can contain one or multiple kinds of data employees without any other team like engineer or product. This is a great structure for training new team members in quickly changing each project each member is working on.

Embedded

Alternatively it can helpful to use an embedded model. Where each data employee is part of a squad which also contains engineers and product managers. This models lets each data employee gain experience on a specific business project, making them a valuable expert.

DOI : 10.35291/2454-9150.2020.0397

Hybrid

Now the hybrid model seems similar to the embedded model, but with an additional sync for all data employees across all squads. This additional layer of organisation allows, for uniform data processes and career development, regardless of which project an employee is assigned to.

VIII. Conclusion

As discussed in the paper, most of the Business Institutes use the promulgated process of collecting, cleaning, pre-processing, model-fitting, evaluation and deployment (if required) to deal with data and come up with de rigueur concomitant.

References

[1] EDISON Data Science Framework: A Foundation for Building Data Science Profession for Research and Industry, Yuri Demchenko ; Adam Belloum ; Wouter Los ; Tomasz Wiktorski ; Andrea Manieri ; Holger Brocks ; Jana Becker ; Dominic Heutelbeck ; Matthias Hemmje ; Steve Brewer,2016 IEEE International Conference on Cloud Computing Technology and Science.

[2] The ambiguity of data science team roles and the need for a data science workforce framework, Jeffrey S. Saltz ; Nancy W. Grady, 2017 IEEE International Conference on Big Data (Big Data).

[3] Open sourcing education for Data Engineering and Data Science David E Drummond, 2016 IEEE Frontiers in Education Conference (FIE).

[4] Data Science in Open-Access Research on-Line Resources Dmytro Lande ; Valentyna Andrushchenko ; Iryna Balagura, 2018 IEEE Second International Conference on Data Stream Mining &amp; Processing (DSMP).

[5] Towards a Data-Centric Research and Development Roadmap for Large-Scale Science User Facilities, E. Wes Bethel, 2017 IEEE 13th International Conference on e-Science (e-Science).

DOI : 10.35291/2454-9150.2020.0397