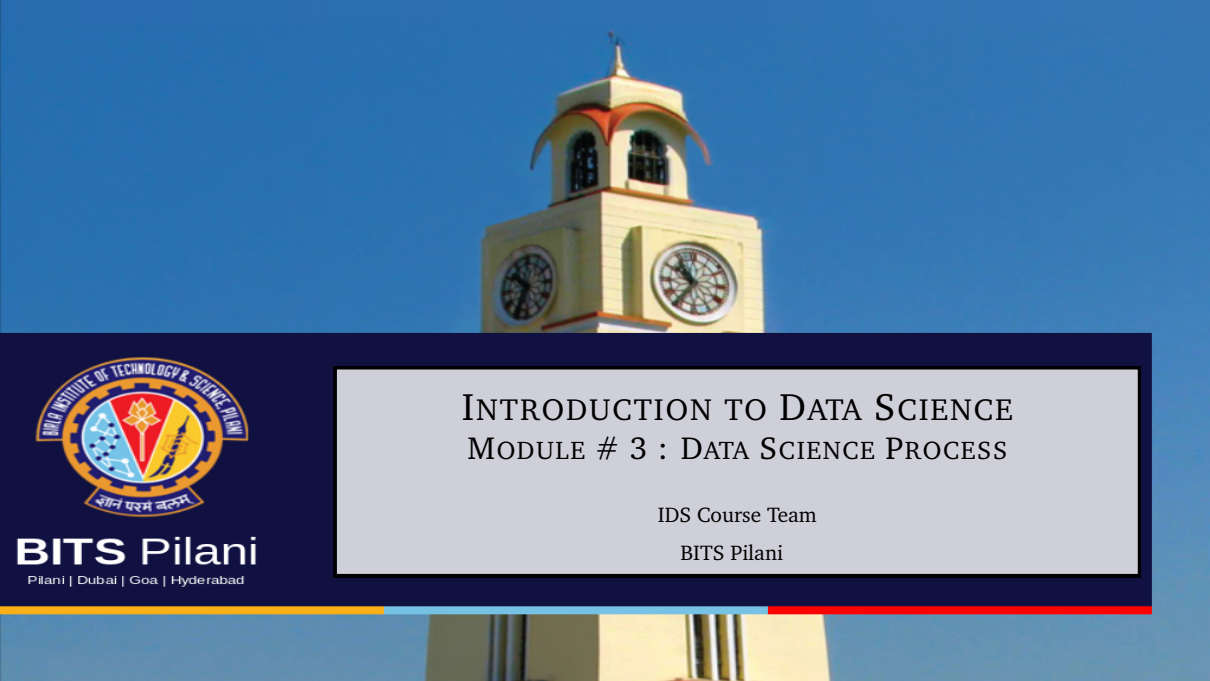

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science
Module # 3 : Data Science Process

IDS Course Team

BITS Pilani

The instructor is gratefully acknowledging the authors who made their course materials freely available online.

Table of Contents

Questionnaire to Prepare Proposal

- What is the business problem we are trying to solve? Write an exact definition.
- Identify the type of the problem.
- Are we addressing a specific problem or a problem specific to a team? Is it a generic problem across all business? (help to create certain frameworks or accelerators)
- Who are the targeted audience?
- How do you evaluate your solution outcome? Are there any evaluation metrics available?
- What is the acceptance criteria for the solution? (for e.g. for a classification task accuracy should be above 65%)

Questionnaire to Prepare Proposal

- Business Understanding
  - What is the business problem we are trying to solve?
  - Write an exact definition.
    - Is it a prediction problem?
      → e.g. predicting company's profit in next quarter.
    - Are we doing a segmentation?
      → e.g. a customer segmentation for targeted marketing.
    - Are we going to recommend something say a product to the user?
    - Is it anomaly detection or a fraud detection problem?
    - Is it an optimization problem?
      → e.g. optimizing revenue of a company.

1. Prediction

- Classification
  - Given a new individual observation, predicts which class it belongs to.
    - e.g. whether a credit card customer will default or not given his data like credit card balance, income etc.
- Regression/Time Series Forecasting
  - Given a new individual observation, estimates the value of a particular variable specific to that individual.
    - e.g. predicting the revenue for the next quarter

1. Prediction... contd...

- Scoring or Class Probability Estimation
  - Related to the classification problem
  - Instead of class prediction, predict a score representing the probability or likelihood that the individual belongs to the class.
    - ⋆ e.g. Evaluate individual customers and produce a score to indicate that how likely they react to an offer.
- Survival Analysis/Churn Analysis
  - Analysis of data where outcome is the time duration until the occurrence of an event of interest.
    - ⋆ e.g. Customer life time with a provider

2. Segmentation / Clustering

- Clustering attempts to group individuals based on similarity.
  - e.g. Segment the customers to High spenders and Low spenders based on their buying pattern and other data.

3. Recommendation / Similarity Matching

- Similarity matching attempts to find similar individuals based on the data known about them. This is useful in recommendation problem setting.
  - e.g. Finding people similar to you who have purchased or liked similar products, recommending a movie to a user based on his preferences and similar users' interests.

4. Anomaly Detection

Profiling / Fraud Detection / Look-alike Modeling

- Profiling attempts to characterize the typical behavior of an individual or a group.
  - e.g. Profiling the driver's behavior for deciding his insurance premium.
  - e.g. If we know the credit card usage pattern of a user, then we can alert if there's any suspicious activity occurs.

5. Causal Modelling / Root Cause Analysis

- Casual modeling helps to understand the casual relationship between events or what events/actions influence other.
  - e.g. What are the possible root causes for an anomaly detected?
  - e.g. Whether the advertisements influenced consumer's decision to purchase or not?

6. Market Basket Analysis

Co-occurrence Grouping / Association Rule Discovery / Frequent Item set Mining

- Find the association between the entities based on the transactions involving them.
  - e.g. What items are purchased together by consumers at a supermarket.

7. Data Reduction

- Replace a large data with a smaller set of data that contain most of the important information in the large dataset.
- Involves loss of information.
  - e.g. Massive data sets of customers dining preferences may be reduced to much smaller data set revealing their cuisine preferences.
  - e.g. A large time series sensor data at a second interval may be reduced to hourly data or to a smaller data set with only changed values

8. Optimization

- Optimization is a mathematical technique for finding the minimum and maximum of an objective function subject to a set of constraints.
  - e.g. Optimizing the profit within a given cost.

Questions to be asked based on Task

- Prediction
  - Do we know what variable (target) to be predicted?
  - Is that target variable defined precisely?
  - What values or ranges of values that this variable can take?
  - Will modelling this target variable address all the problems defined in the scope or only a sub problem?
- Clustering
  - Do we know the end objective? i.e. Is an EDA (Exploratory Data Analysis) path clearly defined to see where our analysis is going?

Solution Approach

- Is the proposed analytical solution formulated appropriately to solve the business problem OR is it an approximation?
- Will the proposed solution address all the problems defined in the scope or only a sub problem?
- What will be the benefits of the proposed solution? Benefit vs. Cost.
- What will be the specific end objectives to be met by the proposed solution?
- What should be the anticipated outcomes by the proposed solution?

Data Preparation

- What are the important variables that you think we should collect?
- Are these variables readily available? Or is there an additional effort needed to collect these variables?
- What are the types of data?
  - e.g. Sensor data, ERP, e-commerce and SAP CRM data are structured (OLTP), Social networking data is unstructured.
- Where are the locations of data in the system?
  - e.g. Product master and sales transaction data in ERP SQL RDBMS database, OLAP data in SQL server for BI reporting, Text data for customer review and sentiment from Tweets and FB posts etc.
- Where are the data coming from?
  - e.g. data from sensor, sales data from ERP, online store

Data Preparation ... contd ...

- Who are the current consumers of the data?
  - e.g. Visualization tools, BI application etc.
- What are the methods to acquire data?
  - e.g. Sensor data are ingested to data lake. ERP, e-commerce, and SAP CRM are inside organization's data center and proper access control needs to be granted to access the data. Social networking data are retrieved from streaming API as a nightly job and are stored in a NoSQL database etc.
- What are the integration points?
  - e.g. IT team needs to provide database access and needs to build API services to access certain data.
- Will it be practical to get all the relevant variables and load it to our workspace?

Data Preparation ... contd ...

- What are the problems in acquiring the data?
  - e.g. Sensor data are archived and deleted after x days. Request needs to be raised to store the data and to archive the data to make enough sample data for analyses and modelling.
  - Social networking data may not be available for a longer term. All relevant data are captured by existing systems, and request needs to be raised and approved for accessing data from servers.
- For the prediction problems, is sufficient amount of labelled examples available? Or is there a cost involved in getting these values?
  - e.g. a field survey may be needed to collect the response from a customer to see the likelihood of joining a new plan.
- Are the training data drawn from a similar population on which the model to be applied? If not, are the selection biases noted? What are the plans to compensate?

Modelling

- Is the choice of model appropriate for the business problem? Is it in line with our prior knowledge of the problem?
  - Classification, ranking, clustering, etc.
- Does the modelling technique meet all the other requirements (functional and non-functional) of the problem?
- Should various modelling techniques be tried and compared using appropriate evaluation metrics?
- Check the amount of data required, generalization performance (i.e. how our model would be using another sample), learning time

Evaluation

- Is there a plan for domain expert validation?
- If so, will the model be in a form that they can understand?
- Is there an evaluation metric set up by the business? (e.g. For a classification problem, there should be less than x% of False Positives). Is that appropriate for the business problem?
- Is there a hold-out data (i.e. data used for validation) available?
- Against what baseline or benchmark the results are compared? (e.g. for a classification problem if there is no baseline given by the client we can compare against a random classifier or a majority class classifier)
- Are the business costs and benefits considered into account?

Evaluation

- For a classification problem, is there a threshold defined (for e.g. different thresholds can give different implications in terms of benefits like reducing the threshold to a 0.70 can reduce the False Positives)
- For a regression problem, how will we evaluate the quality of prediction in the business context?
- For a clustering problem, how the clustering is interpreted in the context of the business problem?
- How will we measure the business impact of the final model? How will we justify the project expense against the benefits?

Existing Systems / Requirements

- Do you have any related requirements? If yes, what are those? e.g. calculating price elasticity or demand forecasting where the main problem is dynamic pricing.
- What are the existing/related systems within the capability that capture/use related information? For e.g. A time series model is already being used for demand forecasting using the same data
- What are the gaps?
- Who are the stakeholders?
- Who will be affected by this implementation?

Assumptions / Dependencies / Challenges

- Note down the assumptions; things like availability of necessary data, access to the infrastructure, licenses etc.
- Any Licenses/Commercials needed in case of proprietary solutions?
- Note down the dependencies: things like dependency on setting up and access to the infrastructure/tools, on access rights etc.
- Are there any other dependencies?
- Do you see any other problems/challenges?

Implementation

- Does the client have a technology preference?
- Does the client have limited / unlimited infrastructure?

A Guide to Designing a Data Science Project

- To get started, brainstorm possible ideas that might interest you.
- Write a proposal along the CRISP-DM Standards.
- Planning
  - Keep a timeline with a To Do, In Progress, Completed and Parking section.
- Track the progress
  - Keep track of how much progress you are making on your metrics.
  - Maintain a code repo for a code review.
- Know when to stop
  - Identify an minimum viable product (MVP) to help you know when to stop.

- Data Science for Business by Tom Fawcett and Foster Provost, O'Reilly
- https://www.linkedin.com/pulse/ask-questions-while-preparing-proposal-data-science-project-menon
- http://www.acheronanalytics.com/acheron-blog/a-guide-to-designing-a-data-science-project

THANK YOU