No:

Week 4

Managing ML projects successfully

Introduction

* key considerations for successfully managing ML projects
    1) Business value.
    2) Data strategy.
    3) Governance.
    4) Team. (expertise)
    5) Culture (mindset)

~~Business value~~ Business Value

* Let's assessing the business value for ML
    1) What is the current gap? How is the gap impacting the organization? (current status)

    2) What would happen if we did nothing? (priority)

    3) How would solving this problem improve or benefit the business, customers, and/or people in genaral? (benefit analysis

    4) How would you classify the project: quick win, long-term development, or full trasformation? (value-add)

    5) What is the estimated cost of this project? (cost)

    6) Do we have any existing budget, expertise, and/or leadership support? (resources and buy-in)

~~Data strategy~~ Data Strategy

* Referes to the data you have and data you will need and method you will use to collect and prepare the required data.

Ex: Let's assume we run a boat company. You make good boats, but now you want to build better boats. How would you learn how to build a ~~b~~ better boat?

* type of water: lakes or ocean?
* Boat transport: trucks or docks?
* Cutomer: middle class or wealthy?
* Type of boat: power boats or kayaks?
* who's not being served
* who are suppliers?

Data Strategy (pillars 1-7)

Towers

Towers

* Pillars of ~~the~~ a successfull data strategy are:

1) Design systems so that you will have more data next year.

2) Break down the silos

3) Transition from data lakes to data warehouse.

4) Learn about your data

5) Integrate pilots into your tools.

6) Run ML models on real-time data to extract the most value

7) Collect more data

No:

Design system so that you will have more data next year

* Suppose that, a television network has desided to personalize its customers television experienced by offering an on-demand digital channel. And, They want to build a system, that recommends tv shows and movies to its viewers.

* In above case, the television network can start with their existing data, and use traditional business intelligence to identify ways to make a movie or TV recommandation.

Ex: age of viewers, gender, income bracket, location, genres watched

Break down the data silos

* To manage your projects successfully, you will need to ensure that your data is organized in such a way that you can query it together and create ML training dataset.

* When breaking down data silos, you need some data strategies like.

  - you need new systems
  - you might need shift the company culture to change the perceived value for data.
  - make sure all your data is digital. Hard copies are inherently siloed.

Transition from data lakes to data warehouse

* Even if you have cultural data sharing your data may be stored in different databases, and access to each database may be restricted

Ex: If an ML engineering needs ~~to~~ access to a production environment on a different team, they may not have it. Even if they have given their own environment so they can create a copy of a specific data set, another team might be concern that the ML Engineer will consume their compute resources

solution :- separate compute and storage. In other words, build a data warehouse. BigQuery is a example for that.

Learn about your data

* Before you start building an ML model, you have to understand your data.
* There are few ways to analyze data
    1) Descriptive analytics.
    2) Dashboards.
    3) Machine learning APIs.
    4) Testing ML on data warehouse.

Descriptive analytics :- Refers to data or content analysis, which is usually done manually to answer what happened or what is happening.

Ex:- pie charts
bar charts
line graphs tables
genarated narratives

Dashboard :- visualizations of your organisations data that can be accessed centrally and provide insights into how the overall function or business is performing

* Ex:- Data studio
  Tableau
  Looker

I)
II)

ML APIs :- use ~~technolo~~ technologies like vision API, natural language API, auto amount, text, and auto amount vision to enrich your use of unstructured data.

I)
II)

ML on DW :- You can now do ML directly in the data warehouse you can handle large amount of ML model in DW.

Integrate pilots into your tools

* If you building an ML model to estimate bicycle demand, build a piloting strategy for the navigation app that your suppliers use. You can have all users beta tester amount model features, or you can choose a subset to provide your feedback on the performance of the amount model.

Run ML models on real-time data to extract the most value

* You might choose to pilot projects with last month's data, but ML models gain real value when they are used in real-time and running on fresh data. To achive this, you will need to build an IT infrastructure, either on-preises or using public cloud service to get data in real-time.

Date: __/__/____

Collect more data

* Suppose you own a ski resort and want to predict the number of ticket sales you get per day for the next ski season.

I) What data would you need for prediction?
II) How would you collect it?

I) Previous ticket sales and nearby event information
II) From an online forecasting service.

Additionally
- I) weather forecast
  - Current snow levels
  - nearby events
  - Hotel prices
- II) previous ticket sales from database
  - An automation or manual snow level measuring system
  - Nearby events and hotel prices information.

* There are other new ways to collecting data
  1) Develop an IoT strategy.
  2) Build partnership around data.

No:

Data Governance

* In the content of ML projects, governance refers to the ongoing practice of applying rules for protecting and controling access to your data

* When you are implimanting ML use cases, it is important to balance data access within your company against a security implications of that axis.

* The most secure option is to know how to use the data, but if you want to acquire insights contained in the raw data-set, you should consider ways to limit access to sensitive data.

Ex: Assume, you train an ML model that uses customer feedback on a product and protect the privacy of the people who submitted the feedback. The problem is information such as delivery address and purchase history is critically important for training the ML model. After the data is provided to the data science team, they will need to query it for data exploration purposes, so it's important to protect your sensitive data fields before making it available.

* There are Three goals for ML and privacy
  1) Identify sensitive data.
  2) Protect sensitive data.
  3) Create public governance documentation.

Identify sensitive data

* Sensitive data can be appear in several forms.
  1) Specific columns in structured datasets.
     ex: set of columns containing a user's first name, last name, and mailing address.
  2) Unstructured text-based datasets (Patterned text) that follows known patterns
     ex: credit card numbers in chat transcripts
  3) Free form unstructured data.
     ex: text reports, audio, vidio, and images
  4) Combinations of fields.

Protect sensitive data

1) Remove it.
2) Mask it.
3) Coarsen it.

Mask : when you can't remove sensitive data fields, it might be possible for you to train effective models, put the data in a masked format.

Ex: I come to bury Caesar ← without mask
I come to bury 6d086db ← after mask
named Caesar has been replaced by encrypted value

Coarsen : Lowering the level of precision of a piece of data.
Ex: Using zip code instead of town name

: ways of coarsening data are GPS location, zip codes, Numeric quntities, and IP addresses

Create public governance documentation

* If our plan is to store data in the cloud, you will need data governance principles when you use public service.

* There are some factors to consider when managing data in the cloud for ML
    1) Securing data
    2) Regulations and compliance
        eg:- CCPA, the European Union's, GDPR
    3) Visibility and control

Team (expertise)

* Having core expertise on the team to carry out a project end to end.

* The three most important data science roles are
    1) Professional dat enginners to build pipelines that routinely ingest and transform data.

    2) ML engineers who build predictive models using ~~caltu~~ curated data.

    3) Data analysts to collect, curate, and explore data opportunities further. They create dashboards.