UVA | SCHOOL of DATA SCIENCE

Project Discovery

DS 6011: Capstone Part I / Capstone Prep
School of Data Science
University of Virginia

Last updated: September 8, 2023

Background

UVA SDS Faculty and Staff vet projects for capstone suitability

What follows are some illustrative projects for student awareness

(these may be from leader of data science / engineering team)

Background, contd.

Of particular importance:

1) Is this a suitable data science project? why/why not

2) Ask questions to understand the vision / goals / data / current state

3) Defining scope / understanding timeline

Project 1

We have a massive database of license plate information (10 terabytes)

We want to use generative AI to search the database for license plates by state

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Project 1 - Answer

We have a massive database of license plate information (10 terabytes)

We want to use generative AI to search the database for license plates by state

This can be solved with a database query

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Project 2

We have a large sample of patient data.
Half of the patients were given a new medication.
The other half were given a placebo.
We need to understand if the medication resulted in a significant difference in outcome.

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Project 2 - Answer

We have a large sample of patient data.
Half of the patients were given a new medication.
The other half were given a placebo.
We need to understand if the medication resulted in a significant difference in outcome.

This can be solved with a t-test from statistics

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Project 3

Automated detection of objects of interest in images.

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Project 3 - Answer

Automated detection of objects of interest in images.

This can be a suitable DS project using deep learning
Want to understand what will be detected, data available

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Project 4

Using AI to predict turnover and increase retention (decrease churn)

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Project 4 - Answer

Using AI to predict turnover and increase retention (decrease churn)

This can be a good use case for data science:
Binary classification problem using ML

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Project 5

Using Generative AI to build a math tutor

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Project 5 - Answer

Using Generative AI to build a math tutor

This is a data science problem...

But it can be a large project. Need to scope appropriately.

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Project 6a

Understanding how an athlete can increase her ranking among peers.
The ranking is a weighted combination of known, measurable attributes.
The weights are known.

Example:
Rank[i] = (0.6 * X1[i]) + (0.3 * X2[i]) + (0.1 * X3[i])

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Project 6a - Answer

Understanding how an athlete can increase her ranking among peers.
The ranking is a weighted combination of known, measurable attributes.
The weights are known.

Example:
Rank[i] = (0.6 * X1[i]) + (0.3 * X2[i]) + (0.1 * X3[i])

Given weights and attributes, relationship X → Y is straightforward.
This is not a suitable data science project.

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Project 6b

Understanding how an athlete can increase her ranking among peers.
The ranking is a weighted combination of unknown attributes.

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Project 6b

Understanding how an athlete can increase her ranking among peers.
The ranking is a weighted combination of unknown attributes.

In this problem variation, this may be a suitable project
The scientist would need to look for attributes, test them, ...

1) Is this a suitable data science project? why/why not
2) Ask questions to understand the vision / goals / data / current state
3) Defining scope / understanding timeline

Takeaways

What makes for a suitable project?

Among many things,

- Clear problem statement
- Realistic goals (metrics, time frame)
- Sponsor Commitment
- Available data, labeled (when needed)
- Data that may be useful for prediction (when needed)
- Free of personally identifiable information (PII / PHI)
- Modeling (in most cases)
- Analytics
- Wrangling data