LLO 8200: Key Issues for Data Analysis Projects

Youtube Generated subtitles

We're going to start with the
textbook and think about this first
chapter and some key problems with data
science projects and we want to start
here because we want to get projects in
general in your projects for the class
off on the right foot and it's important
to think about the
the ways in which we can have a
successful and unsuccessful projects
right at the beginning so we can
understand
uh exactly how to define a project in a
way that's likely to maximize its
success let's get started talking about
that
so the first question that they ask in
the book and this is an excellent
question is why is the problem important
lots of times we have this tendency to
do data analysis for data analysis sake
like where there might be like uh you
know there's a data science department
and they you know they're going to do
data science and they're going to give
us analyzes but
the the question always to answer at the
beginning of any project is why is the
problem important like what makes it
worth getting into the data in this
issue so
some really good reasons for doing this
is that a lack of understanding inhibits
decision making that uh we may be
concerned for example about the
retention of assistant principals in a
school district
and we don't know what the cause might
be is it that their salaries are too low
is that they're very interested in
career advancement and moving on to
principal positions are they unhappy
with the working conditions so what
decision to make right what to change
about the situation for these
individuals cannot be answered until
more information is collected
the other really good reason is that we
would like to know what might happen
next
um what is the expected number of
customers for this business
in the spring of next year knowing that
that will affect how the the kind of

inventory that we would have on hand
so
um any kind of prediction problem that's
going to lead to different decisions
about what to do
um
the other another good reason is that
you may be aware of an overall problem
uh for an organization uh for example
you know there may have been a survey
where it reveals that morale appears to
be alarmingly low in the organization it
might not be everywhere so understanding
where something is happening can really
affect decision making you may engage in
a you know an overall effort to improve
morale when it turns out it's
concentrated entirely in the sales
division
okay
what are some bad reasons why might we
not want to pursue a given problem the
two that I see all the time is one that
methodology just like hey um you know
neural networks are cool let's run a
neural network and get some kind of
result
happens all the time and the results get
ignored right or this nobody really
knows what to do with them or what to
make of them
um so this that does unfortunately tend
to be driven by analysts
um they get kind of really interested in
a particular methodology
um some kind of cool thing that's out
there but it's never really connected
with an important problem
um there was uh there was a big Trend in
using something called Data envelopment
analysis to
um to understand the efficiency of
organizations how efficiently they used
inputs
um tons of work on this there's been so
little decision making based on it that
I can see it just doesn't tend to get
used very often
um the other thing is just getting
focused on some kind of deliberate
deliverable that's going to be similar
to what it's perceived that you the
competitors have
um this is a little bit different in
that it's not driven by the analyst but
instead it's driven by people who are
not analysts but are like aware that you
know uh what's going on in other

organizations so they'll say something like I understand that our competitors are using AI let's use AI as well instead of saying here's a given problem can we work on a predictive or descriptive Solution that's going to help us understand that particular problem better

so just saying our you know our competitors our neighbors have this cool thing we should have it too I really really want to avoid that as a reason for undertaking a data science project okay

so the next thing to focus on is who does the problem affect

um and that will really help us to get to actionable

results from a data analysis project and the key thing here is start with this question and not the data many many times even somebody acknowledges the problem is important and not thinking through this problem they'll just kind of say oh okay well if there's an employee retention issue we will take all of the data on all of our employees and look at the factors that might affect employee retention it's like well who's what's really going on here is it and if the issue is with assistant principles like let's say

um we've got this issue with assistant principles let's focus on that and think about the patterns in assistant principle retention in a descriptive study in the way that we were talking about

um so in if you imagine if we come out with an overall report about employee retention say like employee retention in the school district who's going to read it and what are they going to do it's like well nobody's really going to think it's about them right principals aren't going to think it's about them because they it's like employee retention is too broad it includes teachers and staff and other administrators

um is the superintendent going to use it they might think oh this is something for principals like who's going to end up using it but if it is hey we've got an issue with assistant principals and we think that uh the people most likely to be able to affect that are going to be the principles that they're working with let's gather information in a way

that they can use it and in a way that we think might change their work that they could affect the working conditions of the assistant principals in their schools so really really important to start with this question for data analysis projects who's going to use it um what's going to change about their work

great example of this comes from a research organization that's right here in Vanderbilt the Tennessee education research Alliance the state of Tennessee undertook a big effort to change how they do teacher evaluation now the Tennessee education research and alliances is a storehouse of all of the data on students teachers schools in Tennessee and what they did was say okay we've undertaken This research on or that we've taken this change and how teachers are evaluated what happened afterwards do we think that this teacher evaluation is actually affecting student performance

and what they found indeed was that after this evaluation form you can see that Tennessee school districts their performance has gone up to be much closer to the performance of districts around the country very similar school districts around the country they're now Tennessee school districts are performing really close to where they are a noticeable Improvement that you can notice between 2010 and 2013.

so what does this mean it means well yeah actually that the the um the evaluation of teachers the new evaluation system

um really does seem to be related to uh the student performance so this is something that is very worthwhile to continue using it doesn't mean it's perfect it doesn't mean it shouldn't be changed but a chain you know this this change was made we can see student performance changing afterwards we've got this actionable information now that says yeah don't give up on this evaluation process because we can see real differences for students

okay

so the next question to ask is what if we don't have the right data and there's a couple of classic pitfalls here that I want to warn you against right at the beginning

um so
and we'll think about this in terms of
the dependent variable the outcome ammo
and then the independent variable if
we're looking at the dependent variable
the most common problem is sampling on
the outcome
and what that means is we only have
information about one part of the
dependent variable
really really common to look only at
successes so you could let's say you're
interested in employee retention but you
have no information on the employees
that left you only have the information
on the employees who stayed
right so you say oh what are the factors
associated with employees staying and
you say oh the employees that stayed
look like this it doesn't tell us
anything we have like that is not useful
information because we don't know if the
employees who stayed if the factors that
we've got for them how are they
systematically different than the
employees who left
um and you know in general all of best
practices research share this uh flaw
that if you're only looking at in
individuals or firms or organizations
that are doing really well
um you can't contrast them with
organizations or firms or whatever that
are doing poorly it's supposed to be
like back in the day it was supposed to
be a good management practice for people
to walk around uh that the manager
should walk around and talk to other
employees
um and that got implemented really
widely and it turned out that it made no
difference whatsoever right that when it
was systematically studied it looked
like at successful firms managers were
walking around a lot but when they it
was implemented widely the managers
could walk around all that they wanted
that was not the defining difference
between successful and unsuccessful
firms so we want to be really careful
and skeptical of best practices research
we want variation in the dependent
variable both failures and successes
high scores and low scores
a wide range of outcomes so we can get a
good sense of what's associated with all
of the different outcomes not just the
good ones

when we're thinking about independent variables or the predictors something that you know an input into our model what are some classic issues that we come up with one is that we just don't have any variation that a variable should vary
um you can you can write that down that seems important
yeah we should have variation
um one of my students once wanted to do a study on the um what effects uh kindergartners attendance
um and it's an important problem the more days that a kindergartner attends school it looks like that's really going to affect their academic performance
um you know we to the extent that we can get you know kids in schools
um that does seem to be a correlate of overall performance particularly for young children
and she was concerned that maybe student attitudes how they felt about school might affect their attendance so uh we had data on this and
she looked at the independent variables the independent variable was the student attitude about school so they asked these five-year-olds right it's kindergartners how do you feel about school how do you like your teacher how do you like your classmates all those things here's the thing
five-year-olds are happy they're generally positive and so there was almost no variation in the independent variable they all said oh yes I like my teacher I like my school I like my friends all of those kinds of things
so we need variation in the independent variable it's a key part of the data to have for example and if you're looking at an employee evaluation system if every employee is maxed out in terms of their evaluation you're not going to have any variation to work with as an input
the other thing a big issue with independent variables that people roll into using an independent variable with an assumption of causality you think oh I already know that the reason that assistant principles don't stay is that they're not paid enough
um starting right there you like we don't know that like we would need to figure that out from other means

um but uh and one of the the risks of just kind of like rolling in and thinking that we already know what the the causal variable is is if we look at some patterns and we find no relationship
um we can see right away we've got an issue but I've seen it lots of times with projects where people pre-assumed that they knew what the the causal independent variable was
and then the last issue here is measurement measurement measurement um
the there's lots of times there's big gaps between our concept and the way that we're measuring it we're going to talk about this in class
we want to be really careful and thoughtful about how we describe and label variables from the beginning for instance once in a study the author had a rather lovely description of Social Capital a very in-depth description of social capital for high school students and they went through the literature on it and had like I said just a really rich nuanced description of social capital and then they got to their measurement and I said they said our measurement of social capital is high school GPA
is social capital somewhat reflected in high school GPA of course is GPA by itself a sufficient numerical representation of an individual Social Capital absolutely not
so that study did not have the right data to support this concept of social capital so we want to make sure that our concepts are linked to measurements in a way that worked really well
okay an example of sampling on the dependent variable this author is not me it is the other William Doyle who studies education uh William Doyle's a education commentator who has spent a lot of time talking about the Virtues Of The Finnish school system the why Finland has the best schools
however the research is based only looking at Finland right Finland does have very high test scores for its students
um and they have some unique practices that they they utilize in their schools
it's not at all clear that the unique practices used in Finland would

translate anywhere else or even that the the things that we think are unique about Finland are what contribute to its outcomes so this is an example of kind of best practices thinking where we just say oh well let's look at what Finland has high test scores let's look at what Finland does what we need to do is contrast what's happening in high performing
countries with what's happening in lower performing countries there may be you know Finland has eight hour school days there might be lots of countries that have eight hour school days that are not doing nearly as well as Finland so we always want to be very very careful about
this use of sampling on the dependent variable basically we never want to sample on the dependent variable forget being careful just don't do it
okay
another key question to ask is when is the project over how what's the definition of done and we want to start at the beginning we want to say like what is our definition of done
um the number one rule here is we're it's not we're going to keep going until we find what we want all right that we're just going to like analyze and analyze and analyze until we get something that says aha like I told you so um it was low salaries that was causing lows assistant
principle retention we don't know that until maybe like but we can't start with our conclusions we can't just assume that the data analysis project is going to show us what we want I see this all the time people say oh I want to do a data analysis project that's going to prove that
um different search engine optimization would result in more customer engagement it's like wait like let's you can't define the project that way
okay
one very good way to structure a data analysis project is to start with a number that one is looking for so let's say A you know a
college or university wants more applicants I want to increase the applicant pool
if you start if the data analysis project is how do we get more applicants

it's very difficult to know what the definition of done is for that project but if the the project is to what extent did our placement of ads increase the number of applicants so then you say all right we're going to place the ads in these Outlets we will see what happens to the the applicants afterwards and then we can come up with a number after we place these ads the number of applicants increased by five percent so we've got a number we're looking for and we can answer that question to what extent did our placement of ads increase the number of applicants the num the answer would be by five percent all right that's a good definition of done okay

um the uh another um way that we get issues with data science projects is this um uh thinking about um a uh proving a pet hypothesis right that we just expect it's similar to what I was talking about before that we just think oh um uh the you know if you if a particularly of an organizational leader has commissioned a data analysis project to prove something that they suspect to be true those kind of projects fail all the time because the project reveals that the pet hypothesis does not have evidence to support it but people are reluctant to report that back to the leader so it's very important for leaders and decision makers not to design a project to prove a pet hypothesis

um the links between the results and the decisions are not defined right that it's we have if you say okay here's a decision we need to make um here's a data analysis project that we will use to support that decision great if it's just like we're going to do a data analysis project and hope it will inform um something generally that those projects either never finished or are the results of the projects are never utilized

the last is that um the the user is not the in like thought of as the intended audience for a data analysis project this is gets back to kind of the Cool Tools problem

cool tool
uh projects kind of end up here so if
you say I want to get to a point where I
have like the best possible estimate of
the relationship between teacher salary
and teacher retention
um
and you know like using the fanciest
method and so on but it's going to be
very very difficult to explain to
somebody as opposed to
I would like decision makers to know the
rate of retention by teachers by
experience all right so that you if you
start with the user who is it that's
going to use this and what is the
information they'll have in their hands
at the end
um so and then you know in the user
experience they talk a lot about the
user experience literature they talk
about user stories
um a user wants to do this how are they
going to do it that's actually a really
good way to think about data analysis
projects as well
one interesting thing here is this study
from KPMG about how analytics are used
by CEOs and remarkably a lot of CEOs
doubt their data they doubt the
Integrity of the data
um they doubt the ability that it's you
know it's going to be useful for the
decisions that they need to make that's
this I've just found this you know
really alarming more than half of the
the CEOs surveyed said that this was the
case
it seems to me that the engagement from
the beginning that this is something not
to learn about the CEO at the end of a
project but at the beginning what would
it be about this project that would
convince the decision maker that this is
useful information actionable
information and then design the project
around the decision maker that seems to
me to be the key
okay so in thinking about your own data
analysis projects I'd like you
to to start off with these kinds of
questions so as we think through the
different elements in this chapter how
would a data analysis project that you
put together avoid these various
outcomes that we've said we want to
avoid what would be the ways that you
could design a data analysis project

that is responsive to users is going to result in actionable information to for decision makers

English (auto-generated)

All

Listenable

Recently uploaded

Watched