1.

Thank you for coming to my presentation. I currently work as a data scientist in Laureate Education which is a company doing for profit education. The clients/customers of the company are the students.

But we don’t deal with students directly. I’m in the business intelligence department. Our clients are various business units of the company.

We do simple things such as creating reports and visualization, and also more complicated things as data analysis, building forecasting models and optimization models.

Today I’ll show some applications of statistics in the education business.

2.

Our clients often want to try new approaches to improve business. For example, new software, new ways of contacting students etc. And they want to know whether the new approaches are better. In this regard, it’s a little like clinical trials. But there are differences. First it’s not as crucial as testing new medicine. Although money and resource are involved, but it’s not about life and death. Many times it’s hard to do randomized controlled trials because the business partners don’t want to do so. So a lot of times we are dealing with observational data.

3.

Today I’m going to focus on two types problems. The first one is the thing that I just mentioned, that how to evaluate new approaches.

The second type of problem usually occurs in this scenario. The clients know that a new method is good.

For example, their agents calling students get good feedback. But the problem is that they don’t have the man power to call every relevant student. So they want us to score or rank the students so that they can prioritize using the scores. This typically involves building a statistical learning model to forecast certain probabilities.

4.

Our typical work flow is as this:

First we meet with business contact who has domain knowledge. We try to understand the request, formulate the problem and learn data source. Then we collect data and do some cleaning. Usually we use SAS and SQL for this task. The next step is data analysis, typically we use R. Finally if a statistical learning model is needed, we usually use R or Python to build such a model.

5.

Now I’m going go over the first case study. The problem came to us last year. A business unit wanted to improve the second term retention of new students. Here the second term retention indicates whether a new student is enrolled in his second term after completing the first term. Simply put, they wanted to keep more students enrolled. They came up with a new approach and applied it to the new students started in June of 2017. And the result is quite positive compared with 2016. Hence their request to us is whether it’s really effective since it costs resource to put the approach into production level.

6.

This and the next few slides show the data. Firstly even though the problem was real, the data are not, because I’m not allowed to use real student data outside of company. The treatment group consists of 455 new students who enrolled in 2017 and the control group are 594 students from a year before, also started in June. Here tc stands for treatment or control, and ret stands for whether the student was retained in the 2nd term or not.

7.

Here are more variables. For gender, 1 stands for male and 0 for female. And prog is program, and here we only consider two programs. Balance indicates whether there is a balance in their account.

8.

Here honesty actually is a button on the website when they signed in for a module. In some sense, it indicates how serious they were. And payment plan, generally there are two types, 1 installment and 3 installment, hence is a binary variable here. Assignment indicates whether they have completed the module assignment when the approach is applied. And forum is the number of forum posts that they have posted up to that point.

9.

Finally there are the regions they are from.

10.

Of course the most difficult part is to find all confounders. After discussion with business contact and considering the available data, we will only consider the previous listed variables, in other words we will ignore other potential unobserved confounders.

What business people usually do is just to compare the proportions of retention of the two groups. But because this was an observational study, we had to worry about whether the covariates are balanced across the two groups. Usually we check the balance by several ways. Naturally first we check some plots for a feel.

11.

This a simple bar chart to compare the proportions of programs across the two groups. The pink bar stand for the education program EDD while the green ones are for Management. They are clearly unbalanced between treatment and control. If there are very few covariates, we can certainly condition on program first. For our case, we proceed without literally conditioning on program.

12.

Of course a statistical test can be more helpful for more rigorously checking the balance. Here we can use the z-test (with continuity correction) to check.

The p-value is very small, so it seems reasonable to conclude the data are not balanced with respect to the covariate prog.

13 -- 16.

Similarly we can check the balance of other covariates. Here is payment plan which is also kind of unbalanced. But the proportion of students who have clicked the honesty declaration button seem to well balanced.

17.

Another popular way that we often use to check balance is to use the standardized mean difference. Which by definition is just the difference in means (or proportions for binary variables) divided by pooled standard deviation.

The rule of thumb is that if the SMD is < 0.1 then we have good balance for the corresponding covariate.

If it’s between 0.1 and 0.2, then it’s OK, not too bad. Greater than 0.2 indicates serious problem.

18.

In R, it’s quite easy to calculate SMDs using the tableone package. This table lists the result for the original dataset. The second column gives the mean of each covariate on the control group. The numbers inside the parentheses are the corresponding standard deviation. Here we can see that program is highly unbalanced, payment plan is kind of unbalance, while the honesty indicator is well balanced.

Since there exists unbalance, we need to do some matching. We could try directly matching on all covariates, but sometimes it’s hard to adjust so many variables. Typically we use the so called propensity score matching.

19.

The propensity score is the probability of receiving treatment, rather than control, given the covariates.

It is balancing score, in other words, if we match on the propensity score, we should achieve balance of all the covariates. Of course, in reality for observational studies It’ hard to get the exact propensity score, so usually we just estimate it. Typically a logistic regression model is used, like the R script here shows: glm with binomial family gives us logistic regression, and we are regressing treatment/control to the rest of the covariates. The model is trained on all the data and then applied to the same data to get the estimation of propensity scores. Although logistic regression model is most frequently used, some other more advanced models can also be used.

20.

Once the propensity scores are estimated, but before matching, it is useful to look for overlap by comparing the distribution of propensity score for treated and control subjects.

As would be expected, the propensity scores are on average slightly higher in the treatment group. We can see that there is a good degree of overlap, where we can find individuals in both treatment groups for any propensity scores between 0.2 and 0.7. This is important, because the essential principle of propensity score analysis is that if we find two individuals, one in each treatment group, we can imagine that those two individuals were 'randomly' assigned to each group in the sense of either allocation being equally likely.

21.

Once we confirmed the positivity assumption is reasonable, we can do the propensity score matching. The goal is to find a subset of data so that within this subset the treatment and the control groups are balanced in propensity score and hence balanced in all covariates. We can match on the propensity scores directly, but a lot of times we match on the logit of propensity scores because logit transformation spread the scores out across the real line and makes matching easier.

Matching itself has a lot of variation. We can do greedy matching which is quick and dirty, or do optimal matching which is more computational intensive. And we have to choose the distance measure and caliper values. The essential goal is to keep sufficient subjects and make the covariates balance across the treatment and control groups.

These can be done easily in R, usually we use packages Match, MatchIt and OptMatch.

Propensity score matching has been used a lot since its invention. Recently there is discussion about its validity. There are papers with title “Why we should not use propensity matching” and “Why we should use propensity matching”. In this case, it serves our purpose.

The right hand side shows the R script. The resulting subjects were put into the matched dataset.

After the matching it’s important to check the balance again. Because we actually matched on estimated propensity score instead of the real propensity score, so it’s possible that the balance of covariates have not been achieved. Hence a check is a must. Also we would like to keep a big portion of treatment subjects in the matched data.

22 -- 27

Here we can see that while previously unbalanced, prog is now balanced in the matched set. So is payment plan. Honesty declaration stays balanced as before matching.

28.

Here is tableone of the standardized mean difference after matching. About 91.6% treatment subjects are kept in the matched set which is pretty good. By checking the SMDs we can see that other than the account balance, every covariate is very well balanced in the two groups. And even the account balance is not so bad. So at this point, we are satisfied with the matching result and proceed to do outcome analysis.

29.

Here we simply performed a proportion z-test and found that there is no strong evidence to support the claim that the new approach improves second retention rate. It is kind disappointing to our business partner, but that is what we reported to them. This concludes the first case study.

30.

In the second case study, our client is a team in the finance department. Their job is to contact students who are behind in payment and push them to pay. I don’t really know how do they do their job, apparently they are quite good. Their problem is that there are many students owning money and they are very stretched in man power if they want to call them all.

Their request to us is to help them prioritize the students so that they can work more efficiently. More concretely they want us to rank their students by the probability of not making payment in the next 3 months.

31.

As usual, after discussing with business contacts, our first job is to collect historical data. Now again because of company policy, although we developed model using real data, what I presented here are not based on real data.

There are a little more than 10 thousands records in total. These records belong to two classes depending on the risk variable. If risk = 1, then the student didn’t pay within 3 months, otherwise the student paid within 3 months. In addition to the target variable risk, there are 26 raw features. Our job is to build a model that can calculate the conditional probability of risk = 1 given concrete values of these features.

Finally as usual we randomly split the data into 70 % training and 30 % testing

32 – 33.

Here are some sample of features. …

Once we collect sufficient data, the first step is always to perform exploratory data analysis (EDA).

34.

The main purpose of EDA is to get to know the data. We want to have a rough understanding about which features maybe useful. We also want to identify potential issues such as missing values and outliers and decide how to deal with them. Maybe they will cause trouble for model building, maybe they won’t, we will have to investigate. Finally it’s quite common to transform some raw features to make them easier to model and understand, for example, log transformation for skewed and wide distributions, bin transformation to convert continuous variable into a discrete one. Using age as an example, many times we are just interested in knowing someone is in his 30’s instead of the exact years.

35.

Typically we start with summary statistics. It’s quite easy to do that in R. Just one command “summary” as showing here. This variable is a special index used by the business so I won’t discuss its meaning here.

We can spot some issues immediately. The 3rd quantile is just about 100, while the maximum is nearly 2000. So it’s likely that we have a few very big outliers. Also notice there are over 2000 missing values.

It’s about 20% and kind of alarming. In general, when the number of missing value is not too big, we usually do some kind of imputation of to fill in the blank. On a side note, some software can handle missing values internally. One example is XGBoost which stands for extreme gradient boosting tree. It is a great open source tool that can do quick and good regression and classification. It has python and R interface and has won many data science competitions.

In addition to summary statistics, visualization is also often used.

36.

I often use the so called double density plot which is a kind of continuous histogram stratified by the target variable. Here is one for cumulative GPA. When checking out a density plot usually we focus more on the overall shape of the curve than the actual value on the y-axis. So here the green curve correspond to risk equals 1 while the red one are for risk 0. It tells us something agreeing with intuition. That higher gpa often associates with less risk. So GPA could be a useful feature in the modeling phase.

37.

It’s a good practice to play with models built using one variable. They can help us to identify promising features. We can evaluate them as usual using metrics such as accuracy, recall, specificity, sensitivity, and an interesting one as the area under the roc curve.

38.

Here is the ROC curve for the model based on program. It is created this way. The model will output a probability, or a number between 0 and 1. We can choose the threshold. For example, if we choose 0.5 as the threshold, then we will classify the subject as 1 if the output probability > 0.5, otherwise classify it as 0. Now if we want to increase the power of the classifier, or the true positive rate, we can reduce the threshold so that subjects become easier to be classified as 1. On the other hand, if we want reduce the type 1 error or false positive rate, we can increase the threshold hence subjects become harder to be classified as 1.

When we change the threshold from 0 to 1 we get this curve. Here the x-axis is the false positive rate or type 1 error, y-axis is the true positive rate or power. The area under this curve is an indicator of the goodness of the model. Roughly speaking, bigger area corresponds to bigger true positive rate and smaller false positive rate. The best case is like this which leads to area equals 1.

Here the AUC =.

39.

This is the AUC of cumulative credits. Intuitively the area under this curve should be bigger that the previous one. The calculated value is 0.7.

40.

Once we have cleaned the data and have a reasonable sense of them, we move onto the model building phase. We have played with lots of learning models. This book is a classical reference for statistical learning models. It’s very technical written by professors from Stanford, but a little bit outdated since it’s published in 2008. Logistic regression is a very old model. We often try it out as a bench mark. But usually it performs worse than these more advanced models. Random Forest, Extra Trees and Gradient boosting are all popular state-of-art methods based on collection of trees. We can also try deep learning or neural network model and some ensemble or stacking methods to group a bunch of classifiers together.

41.

It usually takes a lot of time to do model selection, variable selection and parameter tuning. The main strategy is just cross validation. We train a model on some data, then test it on unseen data. We can do it multiple times using something like 10-fold cross validation and average the performance.

Each type of model often has several parameters that need to be tuned. Usually we combine cross validation with grid search.

Our final model is the extra tree model.

42.

As I mentioned before, there are many ways to evaluate the performance of a learning model such as accuracy, recall, sensitivity, specificity, AUC etc. In this case, our business partner wanted us to use the two decile plots to show the validity of the model.

Here is the first plot. It is created in the following way. First we order all students by their predicted risk scores in decreasing order. Then we put them in 10 equal size deciles. So decile 1 represents the students with the highest risk while decile 10 are of those with least risk. Now because these are historical data so we do know whether the students paid their due with 3 months or not. The y-axis shows the exact percentage of the number of students who didn’t pay in each decile. Ideally we should see the height of the bars going down as the decile number increases.

43.

This is the second decile plot. Now we put students in different bins based on the exactly value of their risk probabilities. For example, if the risk probability of a student is between 0 and 0.1 then we put him in bin 1; if the probability is say, 055, then we put him in the bin 0.5 to 0.6.

Here the height of each bar represents the number of students in each decile, the numbers are shown on the left y-axis.. So a lot of high risk students. The right y-axis shows the percentage of risk 1 students in each decile.