1. Definitions:
   1. Label engineering
   2. Sample design
   3. Correction term of average y after log transformation.
   4. NP-hard
   5. LASSO
   6. tuning parameter
2. What’re the steps of prediction?
3. What’re the parts of sample design?
4. How does LASSO reach Pareto Optimality?
5. What does Larger λ values mean?
6. definitions
   1. functional form of y
   2. define the data for analysis
   3. A picture containing text

      Description automatically generated
   4. Non-deterministic polynomial acceptable. As p increase, trying out all options becomes prohibitively complicated and computationally intractable.
   5. LASSO (the acronym of Least Absolute Shrinkage and Selection Operator) is a method to select variables to include in a linear regression to produce good predictions and avoid overfitting.
7. as below:

Diagram

Description automatically generated

1. filtering and spotting errors
2. LASSO modifies the way regression coefficients are estimated by adding a penalty term for too many coefficients. The way its penalty works makes LASSO assign zero coefficients to variables whose inclusion does not improve the fit of the regression much.
3. Larger λ values lead to more aggressive selection and thus fewer variables left in the regression.

The analyst that is you will doesn't have to do anything in terms of variable selection but I just wanted to make sure that I add that that you can sell it you can say well I'm just going to use this 10 variables in my card but you don't you don't you don't have to and there is no aggregation here the third one that that's the most that's the most kind of divided so anyone who picked the first one why did you pick the first one and not the third one Not average decision rules only the prediction anyone who did the opposite any why did you pick why did you pick the third one anyone who did that I think the third one maybe because of the beginning description of the data sets by bootstrapping it is true but but the Scott was saying that's the that's the point that you cannot average decision rules so there really what you average is I don't think but I feel like remember it's not the same data set we can building on which is what the best too but the first one does not say that you're using the same data set for me that was the assumption that it made yeah yes that's OK let me just close this so there was a question one of you asked to mattie and I just wanted to repeat that I think it was a good question it was basically asking whether you can kind of combine lost soul and court if there is a way to kind of he was lost so to pick some you know some subset of variables and then combine it with corn and so the answer or use lost so after cart to kind of cut back the tree and the answer to that is that you typically don't do that right because you use pruning as a way to cut back the tree or you use the stop stopping rule to make sure that the tree is small and at last so it's just not not really an option to use to use that and there there are like really advanced ways to do something that includes many many different methods but that's really out of the scope of this course so basically you can you can combine them that's the answer so I just wanted to repeat it for the whole class OK so let's continue where we stopped last time so I will spend first like it many minutes or so to to finish the previous set of lectures and then we move on to the next one so this is where we stopped and i i talked you know we talked about random forest and and and current and i told you that there And then other ways called boosting I'm I'm going to be you know not going to spend too much time on it but I just want to make sure to mention it and kind of give you a little bit of the intuition I found boosting to be first often slightly bit more complicated that random forest and most importantly there are many different so why less random forest is a pretty well established algorithm and most libraries in R or Python or fairly similar there are many different boosting solutions and there is some similarity and I will talk about that but there is a lot more you can you can read about this if you are if you are interested so the idea is that we are still going to aggregate trees so this idea that we're going to have multiple trees and we're going to aggregate them for our prediction to stabilize the variance that's going to be the same but basically the way we build trees an aggregate is going to be different OK no nothing happens Anne so it's eating it's boosting is an alternative to random forest right both are based on regression trees but it's different is how you combine them on sambol method is another way of combining right it's it's another way of saying that we're going to combine different different trees So what you have seen in bagging you know which is the basis of random forest that India was that we try to build independent trees and we hope that by building independent trees and random forest kind of adds to begging by making it even more independent by building independent trees we are going to have a stable prediction and have a model that is that is that may be biased because of the decorrelation and all that but at the same time it will have a lower variance and that's going to that's going to reduce the error overall now what boosting does is we're going to build trees and we're going to start building a small tree and build another tree on top of it and then we can combine all the trees that we have built to make a prediction so the trees are not going to be independent trees but they're going to be built on top of each other so the idea of boosting is to grow trees sequentially and we're going to use the information right how a tree performs to help us build a better tree next time and and we are in particular the way it's done is that we are trying to extract some information from any tree in the prediction that comes from it and try to improve in a way to improve the part that didn't work well right did that they try to predict observations that that the first we didn't managed to predict where to make it to make it better and so every new tree is built to capture some patterns that the previous one was unable to capture and improve the improve the tree on that regard sorry in sharing your screen because I don't see anything now I have to tell you that you are absolutely right and you should see this It was that was where we started right saying that boosting is just going to be an alternative ensemble method right and and that's the difference is that from independent trees we're going to be a trees on top of it top of each other and like that's the first like informative slide I don't know I sometimes you know I sometimes think about that students should shouldn't show slides at all just speak maybe it's better I don't know anyway right so the the idea is that boosting so you know I'm not going to go into much details about boosting because we don't we don't have time and you can hear about it and maybe other courses and read about it if you want but the basic idea of books thing is that we're building 3 sequentially and at every iteration we're trying to make a new tree that is better able to capture patterns that the previous one didn't fit on observation better that were poorly predicted and still we are not in the business of trying to build the best tree but instead we are building sequentially a bunch of trees and then they're going to aggregate in every time you build a tree the prediction will not only come from from that new tree but all the previous ones it's going to always combine the previous tree and the new tree that European so every time you are you're kind of building a set of a set of trees and using that to to make a prediction and and once again there is going to be a stopping rule that determines when to stop building trees on top of each other right so the key thing to remember when you think about boosting is that unlike random for us you are not building the set of independent trees you are building a lot of trees but you are building on top of each other by every iteration trying to improve on the previous one and at every step the prediction itself is going to be is going to come as an average from the past trees that you have built I guess the that's the idea of of boosting and so the finally menu are making the prediction and you have a stopping rule going to take all previous values and make the final prediction so the outcome of boosting is basically a set of prediction coming from a set of trees that you have built and then you aggregate just as you did previously the main idea is again that you have that you don't have independent trees you have trees dettore improved version of each other and again I think the interesting aspect is that rather than using the kind of the best tree the last three that you have built for prediction it turns out that it's better to use over the previous trees for predictions and the idea is similar that it kind of stabilizes variance right so the new thing compared to random for us is that trees are gradually built and that you are not anymore pushing for these trees to be independent from each other they can be because they are built on top of each other now I mentioned that boosting has a lot of different libraries and there is 1 one so we you know one version of boosting called gradient boosting machines maybe that's one of the the earlier earlier once and all these different kind of libraries are based on the same core idea that you are building trees on top of each other but how you do that what kind of you know what kind of search algorithms you use to find a better fit that's the one that's different and typically boosting have more tuning parameters than random forest you have to determine the complexity of trees and number of trees also how you combine trees to form your predictions how large each tree should be so there are more a little bit more tuning parameters than random forest so if you are even you have a large data set it will turn out to be a slower a slower library than than random for us because of the need to find a good tuning parameters there are there are less kind of some rules for boosting than it is for foreign or forest and then there are you know there are different libraries and XG boost and light GBM and many many others that are that are alternative OK so that's that's basically the idea I just wanted to give you a quick run through so you know how it works and and typically when you look at cases kaggle or any other prediction exercise where you compare different algorithms it turns out that most often random forest and one boosting library are the ones that are doing that are doing well so maybe some of you are eager to ask can you combine boosting and random forest is anyone eager to know that yes so good because I can tell you that that you can write you can combine different algorithms this I mean when you talk about ensemble methods in a broader context it means basically combining different methods you can run a random forest have a set of predictions you can run a GBM set of another predictions and take the average as your model right so your model could be widened I random for us why don't I do GBM and I take the average right it takes more time right so remember last time you had time to go for a coffee have some snack and do 10 pushups to do all this will take much more time but you can absolutely do that and there is some evidence but when you combine multiple methods it's even an data set is large enough it's even helpful because it will even further kind of reduce the variance I said the answer to this is yes you can the second point subpoint is that it's going to take a lot of time the third is especially when the data set is large and you have a lot of variables it can actually be helpful and again very often these also blue models are slightly doing better but the experience is that GBM and Randall for us are typically pretty close in performance and when you combine them you may get slightly better but it's not going to be a game changer account so this is this is actually the case for R beta and you can see that card alone is the 1st followed by OS followed by either complicated or less or loss so and then basically random forest and GBM or you know very close to each other the and you know they are they are slightly different you can also see that actually tuning by hand or using some algorithm to tune in this case it doesn't make a large difference so the big jump is really when you move from OS or lost soul to any kind of machine learning any kind of ensemble model that's the big big difference and better you pick random forest in georgi BM or whether you spend a lot of time tuning it's typically very little now this is for 50,000 observations and like 100 features if you have 2 million observations and 500 variables probably there will be a larger difference right but but in what I've read and seen the big the big jump is really using any kind of ensemble methods and random ferestre GBM or edge boosts they are going to give you fairly similar fits and there is no rule like in terms of the ordering here there is no theory or rule that would tell you how it should look right it is possible that linear relationships are so strong and the data generating process is so simple that OS is great and everything as he's just going to be very very close there are cases if you have a lot of interactions that matter a lot of nonlinearities a lot of noisy variables that should be dropped then then run them faster GBM can substantially outperform orders typically when you have a lot of a lot of variables and variables selection is important relative to the number of rows that that's where you know these three based algorithms are especially useful yeah there is there is there's nothing really here so typically you know what you do when you do machine learning in practice is basically you know you're going to try out a bunch of different models and pick the one that works the best an you know that's that's pretty much and you know you're going to try out stuff and we're going to try out different feature engineering options and and then that's it and you know that the advantage of all else is that you can interpret it at the same time the advantage of random forest or or boosting is that you just have a small code you run it and then that's it there's nothing to nothing to tweak about you know nothing to decide so you know our view is that that that there are a lot of machine learning models out there and we're going to cover them in data science courses random forest and boosting tend to work pretty well so This is why we picked them and importantly you know they're based on this idea of building a bunch of trees and then aggregating them and I think that's that's that's useful too it's useful to to know so you know if you kind of summarize what we've seen yesterday and earlier and earlier today the key advantages of machine learning is really this idea that they are tend to perform better than the simple regressions and that they're easy to use right you don't have to worry about thinking about interactions and modeling nonlinearities and they're very easy to use it it takes more time but but they're they're really easy to use in a lot of things are pretty pretty automatic right so the key problem or the key difficulty with machine learning is that it's a black box and when I say black box I mean we don't have coefficients that we can interpret direct right in as a consequence but we have seen yesterday not yesterday but last week that you can gain insight by looking at you know variable importance or or looking at some partial correlations but what you cannot do is analysis like So what would happen like interventions what would happen if there was a tax on dogs and cats like if if there is a coefficient on text on dogs and cats right what is the what is the relationship between allowing dogs and cats and price you can kind of look at you know So what if I introduce a text or what if I prohibit cats and dogs how would prices change and that's much harder if you have or end up first and and this is my very often when you have when you have a regulator like in a bank they will want to have models that have coefficients so in finance for example you cannot have a black box model right there in most countries there is a regulation that if a client comes to you and So what is the reason my score is 435 and not less we will need to able to say well this is my model and everybody who is you know five years order a male and this and that will have this core and you can't just say so can you please hold on for a second I need 2 hours to run my model and then show it to you 500 fees that's not not Patrick so that's like a big big advantage and and finally you know just let me have some comments so there are random forests are certainly slower and one option you can use is H2O have you heard of H tool has anyone showed you so you will certainly see that in data science too it's a fantastic application please don't don't Google it now it just means that you're doing it now so please don't don't look at it later it has a fantastic API and you can you can run your models in the cloud and so that's one advice I can give you when you have a very large data set that you can use H2O and run other cloud applications the second is so if you have a very large data set like 10 million observation there is some evidence that it's better to take a small random sample 100,000 two 100,000 and one percent 5% something like that build a complicated model and use that model in production or use try out that model later on rather than use a simpler model on the whole data set so think about you know because you know you may have not enough computational power to run a random forest and tendean observations right in in one option is rather than trying to run or less is to take a small sample run a complicated model on that and use that rather than a simpler model trained on a larger sample it's not super straightforward but there is some evidence that this is actually a better way I I always suggest to have something simple that you understand always running away less that you know what's going on with maybe just some variables to have like feel that the data makes sense and the results are sensible like a sanity check and the third is don't spend too much time on fine tuning hyperparameters they they in my experience they had relatively little compared to the time it can take so the final bit in this part is is thinking about outside validity and and and causality and the role of causality in prediction so if you remember 3 weeks ago this is how the class started that I asked you about like how should we think about causality the role of causality in prediction and so there is some discussion and you can argue that whatever helps prediction is fine and doesn't have to be a closer relationship but certainly and I think this was also raised by one of you back then that certainly causality helps to make sure that we have you know variables there are likely to be stable and I think it is it is important to understand what are the reasons behind your behind your key predictors like why do we think that they actually have a causal effect on price because if you can think about some of them that there is causality you have a better chance that the model will work you know next time you run it right so having some idea of how the world works when you have prediction helps you to have a model where you have some faith that whatever you found makes sense why that people do react to larger apartments over these amenities right that you capture have a positive value and so on OK any questions I mean you know so when you when you think about linear regressions and machine learning right I think the idea important thing is that there is some tradeoffs and I think this idea of that yes machine learning model like random random for us will do better but it's harder to interpret that should be in your in the back of your mind and very often to have a simple model that maybe aversed performance in terms of fit but it's simple easy to explain is actually a very good thing so very often when you when you talk to clients and I actually I think I mentioned that I did some consulting jobs a couple of years ago to figure out what people want like but business is 1 many do prediction and very often what they want is they want to have a simple model that they understand like the key variables over less you know something very simple where you can convince them that your your data set and the whole set up makes sense and then have something complicated with bit as good a fit as possible so very often you want to Present two models something simple that you start with and explained and then say well based on that I built like this super complicated model that is very good introduction OK right so this was this was this was the the process we had numerical target and you know we built different ways is in different ways to build predictors we did cross validation we had linear regression card random for instance and boosting and and so you can you can see a comparison of these models and there is there is nothing new here it's just try to put together like what what are the parts that you do by hand what are the parts that algorithms to and I think you know this could be useful to look back to and the speed is basically on my laptop and you can see that running 1 one model or less lost soul card is pretty fast but once you start building a lot and combining them that can take a lot of time and importantly you know there are different libraries like you know for random for us for boosting and you will see improvements in time and accuracy Anne so you know maybe maybe now they are better or faster there you know they're they're better built more optimized and stuff like that so that's fine but in terms of you know ballpark figures I think they're they're realistic So what happens when you know we mentioned this a few times So what happens when you have a very very large data center right and typically when you have a very large data set millions of observations ten millions hundred millions often that allows you to build models that can capture small and relatively unimportant patterns and a lot of them so this advantage of automatic ways to find patterns or particularly useful when you have a lot of information because these models are kind of you know crawling through different cut offs and trying out many things and capturing information that would be hard to do by hand right so big advantage of big data is you can you can really leverage the power of machine learning right and then there are you know I mentioned in the beginning we're going to cover like you know regression trees than random forest and a bit of boosting but there are many other support vector machines neural net deep learning to cover that in data science so there are many other ways how to how to do this the plan here last weekend and this week is basically to drill in on a few and then you're going to cover a lot more right so so big data is helpful really in the fact that the difference between a simple model and the complicated model is more pronounced when you have a lot of hidden small information in a very large data set but that's how big data that's what's that's what's different in in big data however you know what's what's oh another thing that's different right is that when we were talking about overfitting right one part of overfitting was that your sample your data center maybe a small sample of some very large data set right so you need to have you know you you need to avoid overfitting because that particular sample in particular data so sorry is just a sample of a very large data set now sometimes big data means you have the whole population right there is a there is some sensor in the sensors collecting data and you have all the data that is collected by the sensor and in that case this part of overfitting is not an issue because you have the whole population right so big data could be helpful to reduce this fear of overfitting because you have the whole data set however external validity like this thing that I'm that I'm pushing you every every week this is still true no matter how large your data because the data that you have even if it's millions billions of observations it still referring to a particular state of the world it's it's it's for the past it's 41 cities for one grid it's for one set of machines and when you want to use it in the future I think this may change right sometimes they don't write sometimes you have the same sensors for the same machines nothing changes but maybe the weather is going to be different maybe it's going to be used on different machines right maybe it's going to be done for different people different cities etc right so this idea that when you use your model for prediction the live data regardless of how large your data external validity is still an issue if you remember you know I think I already mentioned a few things that if you should only remember that but if you remember this is going to be a very useful lesson from this course and that big data is very helpful to reduce capturing some noise that comes from small samples but the idea of external validity is there regardless of the size of the data any any questions so I mean I tried to kind of with this course I try to stand in the middle in the sense that I am not saying that machine learning is just some useless stuff that that is only useful in for Google I think it could be useful for anything you do at the same time machine learning is just curve fitting is just building models trying to find patterns in the data regressions are still useful to understand what's going on an external validity whatever you do should be in the back of your mind yeah so for example we can do random forest via the package as well as wires on random forest package in same week and pay for extra extra boost algorithms so something that you're looking to as to which package should be used to run algorithm so AM will so I looked a bit more on on random forest an I have to tell you that they are more similar to each other I think they way output is organized is is slightly better so the character uses this Ranger package as opposed to random forest Ranger is a bit more modern and how algorithms are done so I like that I think carrot is fantastic package and and I think I mentioned that somewhere in the slides or in the book that you should read the documentation for carrot it's absolutely in I learned a lot from it so I like carrot a lot because I think it's just a fantastic super smart people and I kind of trust them in terms of boosting I think XG boost and light light GBM are now slightly better done Anne then when I was writing the book and typically I think people use that but there will be new or anywhere libraries and solutions there may be differences so we're just we're just doing the the hard impatient you know we're we're translating the codes from R to Python And we're just doing the this part and started to look at the libraries and you know they are slightly different and I don't know it's a it's it's and in so it's an important question and I think he will so this is something that came up last year like why do we pick certain libraries and I often say because you know told me I shall and so I now I ask you know to talk about more Anne I was kind of convinced how extremely good the the like the description of characters that was the main motivation you can also ask mattie so Matt is working with both Python And R and you can also ask him about his experience thank you four so what's going to happen now is we're going to have another small poll or quiz on logic and then I'm going to start talking about classification so you should see a new pull in it's going to be 2 questions both are single choice can you see Fun OK so it's fairly most of you picked the right answer OK so I made this point I ask you also to look into the bits on logic because that's going to be that's going to be important can you see my slides best thanks OK so we're going to talk about two things two things that are related probability prediction and classification so compared to what you've seen in the in before the main difference today is going to be that the target variable is going to be a binary is going to be 0 or one yes or no and Anne basically it means that that the outcome is kind of a quality variable it's better at adapter default or not or an email is spam or not or it could be something like a game football game result being win lose and draw and so in this class I'm only going to consider cases where there are two classes 01 yes or no default or not but this can be extended on multiclass cases and maybe you will see in other other courses examples for that but I want to kind of drill on the one that is most widely used A01 case invite me in in when we think think about probability right we're going to talk about the frequency of yes or the frequency of one and remember this concept that what when you think of the probability of some event happening or not that is translated to frequency in a data set I said the probability of something as as a theoretical concept or the way you think about it when you look at the data set the way to think about this is that is the pavilions the frequency of yeses or ones in and so importantly we are going to have two different actions and first action is going to be predicting probability and that's going to be you know what is the chance what is the probability at that are will default that's going to be our case study and once we have probability we're going to think about how can we turn that predicted probability into 01 that's called classification so classification is the act of turning predicted probabilities into classes like one end and zero finding again that there could be multiple classes it it could be possible that you predict the probability and that's how probability of different different values and you put them into different classes but today I'm only going to focus on the case then why is binary today is going to be you know that the process is going to be pretty probability and you have seen that before India too and This is why I ask you to review that because there is not there isn't going to be that much novelty there the predicted probability will be between zero and one that's going to be the probability of an event happening and then we're going to take that probability and and very often that's going to be the end of the exercise so very often you just want to be have an idea how frequently you know how can I predict that that a certain client chance of leaving my service it very often I don't necessarily need a 01 prediction I just want to have an idea so death client for that client what is the chance of me losing losing that right and very often that's it and sometimes you're going to continue from that and classify into zeros in one or loss function is going to be the brier score which is which is basically mean squared error so you have reviewed it but I just want to make one important point to remember that we are going to use logit and wireless in Chapter 11 when or DA2 the key issue was we wanted to understand how certain variables are correlated with the binary outcome or probabilities and in those cases linear probability models worked well because they were simple you could interpret coefficients and we didn't really care about prediction now when you care about prediction and you want to have a predicted probability between zero and one you wanna use a model that generates probabilities between zero and one and logic is one such model can we use probit for prediction he absolutely can do we know which one is better it's an empirical question we don't right it turns out that in most of the cases it doesn't matter and logic has some advantages because it is linked to some nice theories and we can talk about it if you want in the in the break and therefore budget is slightly more preferred but in terms of empirics you know whichever you want as long as predicted probabilities are between zero and one OK so what's new when we have binary target really that we predict the probability and not a value we don't predict the price but we predict the probability so the predicted object is going to be between zero and one so that's one thing that is new with the binary target and the second thing that is new is that you may want to have and need to make classification to assign zero or one right assigning zero or one or classifying 01 that's the same thing run in and so we're going to talk about how to do it and that's going to be the key new thing today so once you know we will still you know what's not different is that we still going to try to have the best fit we still going to care about external validity we still going to have our usual varies of overfitting and we are going to have you know regression models probability models logic rather than awareness and then we're going to see how to do it with card and and random forest so when you you know when you do probability prediction right basically we wanna have and model that is able to predict probabilities an and and often again that's it and sometimes when you want to classify right when your end question is predicting zero or one classification you still going to need to predict probabilities right you're still going to need some model that assigns probabilities 1st and then use those probabilities 2 to classify now when you build probability models this idea that you are building a bunch of different models right you're gonna have a measure of fit just as before and you're going to cross validate and try to find the best model for predicting probabilities that's unchanged and so the basic process of trying to come up with different models and then picking picking the model that is best able to predict probability that's not going to be different yes we are going to use logic rather than OS and the target is zero or one but we're still looking at prediction error we still going to have you know I measure fit and the loss function and and build models with that and this idea that you build lodge it's based on domain knowledge just as you build or less that's unchanged can you use law so yes you can it's called logic law soon is slightly more complicated the how it works is more complicated than last so so I'm not going to talk about it but you can use that right so there is a lost soul for logic but the way you remember last so in essence was simple you had a sum of squared minus some penalty right for logic the process is more complicated because logic itself is not a linear model so the whole process is more complicated but the idea that you can use logic and that it does similar in the sense that it will reduce some coefficients to zero and also reduce some predictor coefficient tord 0 and that's that's the same and then we can use Randall for us and going to talk about that later today we're going to pick a model with cross validation and and our loss function is going to be the real score check there are other loss functions but the beer's quest is good it slightly was in it's the same as pyramus E could there be other loss functions absolutely and why does air Missy is pretty much widely used when the target is a number when the target is binary there are other there's a lot of other loss functions that are actually used I'm not going to talk about them today but the ones that are mentioned in Chapter 11 log loss and stuff like that they are actually used in practice we looked at the ranking of models by a variety of loss functions like the most values ones and there is very little difference it doesn't seem to matter much it may matter in some datasets it doesn't seem to be in practice a huge difference so let me stop for a moment to let you breathe and ask a question if there is any but I didn't mention to is that today is going to be slightly different because I'm going to go through theory first and I'm going to discuss the case that is separately because I want to talk a little bit more about the case study like the whole how how you design make a whole prediction process I wanna talk a little bit more about that not just the classification so I will first cover the theory and then once he's done beginner start looking at the case that he had maybe some of it will split will spill to next week will see but that's the plan I first just focus on theory so This is why you're not going to see for awhile like blue tables OK so until now nothing really new what's going to be and still nothing new for a little while 'cause you know this idea of making errors and false negative and false positive that that was not new right so this kind of a table is something that you have seen is is how you can make prediction and how it compares to the actual values and making true false true positive and false negative just bear in mind I'm going to use these TNF NTP FP later on so two negative false negative false negative false positive to positive and in D and you can see that there are two cases when they make an error but we can make a false false negative in the false positive predicting you know if we take the example that you will see in the case study right when we are trying to predict if firm defaults or not or the firm stays in business or not right the errors that you can make is predicting the firm to stay when in fact it exits or the other way around I said there are two ways to be right predicting stay in the firm stays and then also two ways to be wrong so we're going to use the frequency of of making errors to think about how well a classification works and importantly I'm going to introduce and this is where novel theme begins I'm going to introduce three measures also how well the classification works so the first one is accuracy I think this is the most straightforward one but accuracy is just the share ask you can see my hand right Right so that the accuracy is just I mean I think that's the most straightforward one right it's just how also called hit rate is how frequently you are here or here compared to here right the next one the next 2 one sensitivity and specificity the first one is the proportion of two positives amongst all positives and the second one is the proportion of two negatives among all negatives I have to tell you that when I teach this class I remember which one is which and when this class is over I immediately forget maybe I will forget by Wednesday so if you don't remember I don't I don't mind as long as you know that one is the proportion of true positives among all actual positives and the other is the proportion of two negatives among all actual negatives right but there cause sensitivity and specificity an this kind of language again comes from I think engineering and computer science and whatnot I'm sorry do you count these ratios after you predict our values and you compare it to the to the real values right so you have no we're going to talk about how to how to make the predictions now or next or soon but suppose you already this is a good point right suppose you already have your predictions and you know having their predictions or helping you to be in one of the four you know one of these four blocks right so you make a prediction yes or no and you know in your in your training data if it's yes or no and you compare thank you good point so we're going to have these three measures of classification and sometimes accuracy is what you care about sometimes all you care about is how accurately you're able to do that but very often that's not that's not going to be enough and so the first important insight I wanna come way when you think about classification is this is something you will see a bit more later on is that there is a tradeoff between making false positive and false negative errors right you may be in a way very strict and classify cases that are pretty certain to be a defaulting firm or a fraud but if you are very strict that that's going to mean that that you are making a lot of errors when you are classifying something to be an error or fraud and it's not the other way around if you are very relaxed in classifying cases into fraud or error you're gonna make the other type of error so again this is something you will see more and more but it's it's I just wanted to flag that that's going to be our key first insight that you have to kind of pick and choose what kind of errors you are more comfortable with because there is a tradeoff and and the reason that specific city and sensitivity are useful because you can express this trade off between false positive and false negatives in the way to do that as we are going to use a curve it's called the receiver operating correct correct ristic curve surprise this comes from engineering and so the Roc curve is going to be a graphic it's going to be a graph as you will see it's going to look like this I'm going to tell you what it is but it's going to look like something like this is going to be a curve it's going to be showing a tradeoff between false positive and and two positive rate so specificity and sensitivity and So what we're going to do is we're going to say so let's have different thresholds let's use these thresholds of probability to classify into zeros and one and let's see what kind of errors we're making as they classify into zeros and ones right so the threshold is going to be a number it's going to be somewhere between zero and one make me pretty probabilities be going to have a threshold and we gonna say if the predicted probability is greater than that threshold is going to say it's one if it's below that threshold we're going to be we're going to say 0 and that threshold we're going to move this threshold and we're going to come and for every threshold we're going to compute that table that you've seen and by the way it's at this table is called confusion matrix or confusion table or classification table as I've told you everything has multiple names right so classification table confusion table confusion matrix same thing basically these two by two matrix where you put two negatives false negatives false positives and true positives right so you predict you make the prediction you have a threshold if the predicted value is above the threshold you going to classify to one if it's built over that you're going to classify it to 0 you gonna put you know you're an investment fund this question comes in it after that is when you put your your observations into these two by two matrix you can calculate sensitivity and you can calculate specificity using you know those four numbers and that's going to give you some you know some perspective about about positives and false positives and false negatives right and what the Roc curve does is going to show you this relationship between two types of errors for different kinds of thresholds when and that's going to illustrate this tradeoff between the two types of errors when in remember every threshold is a number above which you classify your observations to 1 below which classified to 0 one such threshold could be 0.5 if you predict something to be more likely than not the classified to one if you predict something less likely than not classified to 0 this is 0.5 is 1 possible threshold another possible threshold is the sample mean off wise like the average likelihood that default happens in our case that you will see that 20% of firms will default so 0.2 could be another threshold saying if their probability is higher than what you see in the data to be going to classify it too yes if it's below being a classified to 0 if it's not feeling classified to 0 right and then there are many other possible threshold maybe you want to be really certain and just want to classify cases to default media to very strong chance and so the threshold could be 0.8 or it could be very very low right and So what the Roc curve does it allows you to make all these comparisons and graphically displayed on it OK so let me show you this graph and then we can discuss any questions you may have so this is a very pretty and colorful graph and I'm very proud of it because it looks very nice and here is what it tells you each dot is a threshold each dollar is a classification threshold and this is 0.8 this is 0.6 this is 0.4 this is 0.1 transfer each point here means that if your threshold is 0.2 this is where this is the value of sensitivity and this is the value of specificity that you can expect so this tells you the tradeoff between false positive rate and true positive rate this test is a tradeoff between the two kinds of errors that you are making what you can see right this is like the case for tradeoff that you can achieve different values of sensitivity but the cost of that will be an increasing false positive rate but that's what it tells you that there is a tradeoff and as you move along this curve you can pick different combinations we're going to have different combinations of these two types of errors but there is a tradeoff you cannot have you cannot have it all so how do you think this curve would look if there is no tradeoff would it be linear yes it would be a it would be aligned that's funny when you hold on but do you mean linear or flat roughly yeah linear but maybe just a call not linear no not linear linear is still there is a trade off right as long as this has you know that so this if there is a line here or even better there is a line here right here that means no tradeoff because no matter how I change false positive rate I always have a certain value of of two positive rate or here or or or or a vertical line like this right that means that it doesn't matter the other error that I make as long as there is some relationship it means whether it's linear or or like like this danum then there is a tradeoff any practice this can I mean it's typically looks like a curvy thing it doesn't have to be it can be rugged you know it doesn't have to be nice but it illustrates a tradeoff Brendan is as you move along the threshold you kind of have different you're making a different combination of errors right so as we move along since so neither when you look at the Roc curve the threshold or the thresholds or not on either of the axes because in the axis you see the false positive rate and the true positive rate sensitivity and the one minus specificity this is what you see on the axis the thresholds are you know combination of values our dots in this in this in this graph and you kind of can move along the thresholds and you move along the Roc curve so you know for one prediction Model 1 probability prediction model you can generate one Roc curve by changing the threshold above or below you classified 2 zero and one how do we choose the threshold for classic I'm going to talk about choosing the threshold in the minimum in a minute and I see in a minute maybe bit more but that's going to be the next one and so you know this is the first one this is the Kirby scene and you can kind of you know create a curve that is just continuous basically you know looking at very small bits of the curve rested these two graphs or the same thing the left graph illustrates the idea that each point is a threshold and that you are moving from right top to left button as you increase the threshold as it becomes harder and harder to classify something into one you are moving from here to here so that's the advantage of this curve the adventure of this curve is it shows you the whole curve like nicely and it can you know if you look at it you can see it's kind of rugged it's not smooth so we can so when we compare 2 predictions for each of them we can generate a curve right so how does this curve you think relat to the performance of the model my answer is it doesn't another honest phrase it is in how the steeper the slope in the beginning the better the model the steeper despite like here yes in the beginning so if the true positive rate increases faster than the false positive rate that is true can you be a bit more general so you are right this deeper it is but like how will so if it's steep how will this affect the whole curve so the bigger the area or the area under the curve the better the model is right but why so the you know that it is the same thing the question is the surface is very sticky here like even if there is another curve here is it better or worse and if yes why do not work here if if we if the perfect model would be the whole square so the area would be the whole square so the perfect model would be something like this and then here right yeah sorry why because that would mean that the true positive rate is 1 no matter the false positive rate so it always predicts all sorry no matter the treshold it always predicts perfectly and this is absolutely right right so the point is that the perfect model the model where we always get it right is something like regardless of the false positive rate we always get the true positives perfectly so it's basically a line here it jumps and then he's here but that's that's the benchmark OK so the closer or curve this guy here the closer our curve to that in a theoretical maximum the better needs another way to assess the performance of our predictive model is remembered one predict one probability predictive Model 1 Roc curve right so we can compare predictive models by looking at door curves and looking at how close this is to this theoretical bit So what is what is the 45 degree line here what is what is the meaning of this line this is if we predicted randomly like 50% we say yes 50% we say no so you are right and why is this corresponds to a random so one way to predict right would be just completely randomly picking Anne you know picking picking values not using any model just complete complete randomness so why is this why the 45 degree line is kind of this is this is when we are right completely right this is when it's complete randomness right obviously our curve is going to be somewhere in between but why the why is the 45 degree line correspond to night random renting the nest It's because the chances are the same to actually make it right or make it wrong so your model was completely random that sense yes it's it's completely random in the sense that that whenever you are likely to make a false positive you also make you're also likely to make a false negative so there is the model doesn't help you to classify that can you be here can you have a you have a model that is here like this right here below the 45 degree line can you be here can you have a model that is versed than just flipping coins if you take the inverse of your current prediction the rocker I guess yes but that would make no sense no it doesn't make sense but you can write so it's possible to have a curve that is below the 45 degree line and that exactly means sometimes you screw up your labels and you come up with a model like that and then you know where I just pulled up my labels or you know you have a model that that for some reason predicts exactly the opposite so I for a short period of time before I did my PhD I've worked for an investment bank and there was a guy who can who was the analyst on Russia and so people on the traders really listen to this guy and what he predicted and that they did the exact opposite I see in that sense dead guy was useful for prediction because it was better than random fun the the prediction curve from that guys below here right so you just need to flip it and do the opposite it was very funny I think the night when I learned that people pay attention to him a lot even though he's wrong alot because they realize that he's more likely to be wrong than none so there is information in what he's predicting I think that was very funny OK so we can be below the 45 degree line in and as you mentioned so one important metric that we can that we can bring out of this is the area under the Roc curve so the area under the Roc curve is going to be a number between zero and one which is going to be when you integrate this curve out it is going to be the the size of the green thing compared to one right so it's a 1 by 1 table or one by one graph and basically the area is just going to be a number between zero and one and when it's 0.5 it's basically the size under this linear line right so that's just going to be 0.5 and anything above that means that you had your model have a predictive power and again can the UCB build oh point 5 it can if the opposite is more likely to happen I said the area under the curve is going to be a number between zero and one end but what you care about is how larger it is compared to 0.5 right because 0.5 is randomness so you want to say how more confident I am in my prediction this is just repeating what I what I just said so when you think about selecting models and you want to pick a probability model we have seen one metric one loss function to do that right basically use mean squared error and pick the one that has the smallest mean squared error the smallest prior score and now I showed you another one so you build a predictive model you make predictions you classify but you classify by looking at all potential thresholds calculate the Roc curve calculate the area under the Roc curve which which encapsulates how well your model is able to perform under this tradeoff between the two errors in peak the model that has the highest AUC you can use both methods to pick a probability model in practice EUC is more frequently used it has some nice properties and I will come back to them and it turns out again that whether you picked by AOC or picked by air Messi is not something that matters greatly but I think drawing up a Roc curve and calculating AC is a very good way to understanding how your probability model works any question in terms of calibrating the threshold for the model for implementing the model are we going for to be as far away from the random prediction line as possible in terms of choosing a threshold value notes I'm gonna I'm gonna I'm gonna talk about how to set a threshold I will probably do that after the break OK alright but that's definitely going to talk about it but I mean when you you know when you yeah so we're so there there is there's a lot to think about how you set the threshold and it's not just going to be how far you are from randomness I mean the a we see so when you know I mentioned in the beginning that often you have two different objectives the first is to have a probability prediction model when that's the case you want to have a prediction probability prediction model then AUC and how far you are from random is a great way to pick your your best model and also great way to describe it right so when you want to say so how how my model performs a we'll see has has you know has a meaningful interpretation and therefore AUC is a good way too a good way to talk about your model as well as picture model she but once you want to classify and actually assign zeros and ones you will always need a threshold right that's that's going to be the in the second insight today is that without a threshold you cannot classify no threshold no classification any questions if not I think it's a good way to break good point to break so see you how fast can someone please remind me to restart recording OK so

Session 2:

Bury them or Victor or band who will set the threshold it can it can be both you mean us and an algorithm yeah I mean we can decide on a on a preset threshold and you can I guess there is a way to find an optimal threshold mean what does optimal mean maybe to somehow minimize errors so misclassification errors by the way we build and rebuild the probability model right it is basically by middle sitting and estimating a logic model is model fitting and and speaking your model by mean squared error is by prediction errors so we've already done done that we have the best model in that sense I think it depends on how sensitive we are to through positive sorry false positives and false negatives So what do you sensitive I mean that if we really want to avoid making a kind of error like we really don't wanna do false positives and then we might want to go with threshold each in which we don't make that many false positives in then I think so that that's an important insight in the sense that you are arguing that it's us who makes the threshold basically they make the threshold by kind of deciding about what kind of errors we are more inclined to tolerate my job but that's what you are kind of yes it if it if it matters then yes if all we care about this prediction then maybe we go with you know whatever results in highest accuracy or but but importantly so you are right but whenever we only care about probability prediction then this then then we stop here OK so when we care about probability prediction we're not making any classifications and we stop here that's it so OK so basically you know that the selecting the threshold is really there in their different ways there is what's called majority voting which is 50% just more or less likely or frequency in in the data now it turns out that neither is a good way to think about threshold selection because the good way to think about threshold selection is closer to what drumbore was suggesting basically weighing up what kind of losses we incur alright so when you say what are we more likely to tolerate let me rephrase that what kind of losses right to be encouraged by making a false positive or a false negative prediction right it's kind of alluring to the same idea what are you more likely to tolerate translated translated into monetary or loss function language and your loss function is basically telling you what kind of losses monetary or any kind of harmonized losses or associated with false positive or false negative sometimes these costs or very different rights you are really when you are trying to decide if a cell is cancerous or not right the cost of saying that a cell is OK while it's not a probably higher than falsely accusing the cell of being cancerous because then you do another test and then it's going to turn out to be OK but missing out that someone has cancer is really costly right so there aren't circumstances when the cost of false positive and false negative are very different and again this idea of cost is the same as kind of how well you tolerate or which one you prefer which kind of area preferred to make and the loss function is away to express this cost and compare them and so you know what author or someone else was saying regarding minimizing we are going to minimize something and that minimization is going to be expected loss what is expected loss the probability of making an error times the value associated with that error so the time the probability makes false negative times the last year you suffer as the probability of false positive times the loss weather in that sense it's really this relative cost that will turn out to matter and how more costly it is to mix false negative result of false positive So what we're going to do we have a loss function that is the loss function for the classification it's it's is a way to find the optimal classification threshold I said this loss function is not the loss function we had before for probability prediction this is the loss function that takes into account the cost associated with with the errors that we make right so the stuff that you guys mentioned that we going to minimize something and that it's going to be related to tolerance that's both true and are captured by by this simple object now how are we going to find the best threshold then it it turns out that there is two way to do that there is a formula which is based on some assumptions and then there is an algorithm that is always true so so first the algorithm so the algorithm does the following thing it is minimizing the aim is to minimize expected loss this guy here and what the algorithm does it looks through all possible thresholds and peace and picks the best option minimizing expected loss so it it looks at you know basically those points it compares it moves along the Roc curve it does not compare Roc curves when you compare models you compare rockers when you are looking for the best threshold you move along the Roc curve that is associated with bond predictive model yes no yes right so you move along those dots and This is why I'm very happy with that graph because it hurts him to kind of dimension that you are for each threshold we're moving along these thoughts and you compare all of them but you not compare them in terms of model fit you compare minimizing loss and you pick the one that is the best and there is some \*\*\*\*\*\*\*\* language here that it's not minimizing expected laws but it's probability cost sensitive youden index whatever it turns out that they are the same it's often there is some magic language and it turns out that everything is the same and you can see that in our appendix it took us quite a while to realize that they are it seems like the same thing but you can prove that they are exactly the same thing and and so that's what the algorithm does now there is a formula to do that that formula says that the threshold is just this ratio but it's just the relative loss associated with false positives it turns out that this formula is based on two assumptions assumption number one that your data set is large enough right it it typically means that these thresholds we'll approximate what is found by the algorithm as you increase sample size and the 2nd is that our model is a good model in and they are they are in parentheses because there is like the definition of that is complicated but as long as you think you have a good good enough model and your data set is large enough then the thresholds are fine so the pro of using this threshold is there easy to use and most of the time it's closed you will see that in the case that is pretty close Anne the con is that they are not the best cut off their clothes does this formula is close to the best color it's not the best one and the smaller the data or the poor or your model the less likely that this threshold is very close to the actual one OK so this threshold is an approximation under some assumptions of what the minimization algorithm gives you and under a things are nice then they're they're close so when when you wanna select a model but in and basically you have a loss function then then you can then you can do that directly based on classification right so you have different models and you can calculate probabilities find the optimal threshold use their threshold to classify and calculate expected loss and pick the model that has the threshold that leads you to the smallest loss right so one way to select a model but we have seen before was by AUC AUC is agnostic about the threshold so if you don't have a threshold you select the model by AC or mean squared error whichever you want but based on the probability prediction performance then if you have a loss function you can have your different types of model Model 1 model 2 model 3 my 3 logics say for each of these three models you can search for the best you know search for the threshold that gives you the smallest expected loss compare and pick the model that has the threshold that gives you the smallest expected loss across these three models that's another way of model selection OK but I mean the key insight in this bit is that if you want to classify you need a threshold to have a threshold you need a loss function if you don't have a loss function you ain't gonna have a threshold if you ain't gonna have a threshold you're not going to be able to classify if I have a symmetric loss function as in I don't care more .5 if you have a if you if you have a symmetric loss function which means that the loss associated with both be two types of error equal this guy is going to be 0.5 and that's going to be a threshold right so it's not like like in that case when I don't know I'm betting for example and I don't care how am I wrong in what is I didn't get in gambling for football games for example then I thought I can still come up Carol calibrate my threshold based on how it's going to affect my predictive accuracy strictly has to be 0 so if you are payoff is such that if you guess it right it's $1 and if you get it wrong it's zero and then that's it then your loss is equal and if your loss is equal then then the correct threshold is 0.5 but you do have a loss function is just a super simple one right because it's possible is that for some reason the loss function of getting game results right or wrong is not you know I mean you can if there are odds right if there are odds so if you know Manchester City plays Burnley then getting getting the the predicting city to win or not getting the result right is probably you know you can you can be more if it's certain results and it's not 5050 and you know it could be that your loss function is different so the key you know the key message that I have is that you the analyst so why let's meet the threshold you know the threshold itself comes from either a formula or an algorithm but the input to the threshold comes always from the analyst it's not the algorithm is not the computer it is it's not even all by Victor it's you but it's always the analyst who decides what the loss function is indent you know that decision will determine the loss the threshold either by algorithm or by formula but the input is always always comes from the analyst and it can be that it's 5050 doesn't matter and that's a perfectly OK threshold but it's a conscientious decision right yeah OK so let's talk how to do cards and random forest for classification so you can build a classification tree in that's something where we predict classes zeros and ones the way you build the classification tree in the sense that it's still recursive binary splitting splitting it still saying top top down greedy algorithm that's unchanged what's going to be different is that prediction will not be the that the mean of the values is going to be the share of wise in a bin and you know be going to kind of leverage this idea that probability and frequency or the same thing and and this idea that there is a threshold is going to is going to matter so it turns out that when you have a classification tree the measure of fit is called not impurity the idea is that when you build a tree and we will see that in the case study when you build a tree and you think about the quality of the prediction no the impurity means that when you look at your terminal notes when no based on which you make the prediction if you were able to collect all the ones and all the zeros across across nodes you are prediction is going to be cleaner in other words when you are able to have in one node only zeros and then other nodes only ones that's going to mean that on that node you really able to predict zeros or ones pretty well and so one of the measures debt that is used as a measure of this impurity is called the gini index it will turn out to be the case that the gini index is nothing different than the mean squared error to be more precise they lead exactly to the same result and you can check them in the appendix if you are inclined to read derivations suit in this lingo the loss function is called gini index of node impurity rather than mean squared error but it will turn out to be the same So there is so so we have this this this this loss function when they when they build when we build the trees and then you know we can build trees and we can we can build random forest and we can use random forest instead of logic models to do the classification so everything that we have done with the continuous case we can we can do here any turns out just this was the case for random for US versus what is that when you compare random for us and lodge it's random forest will turn out to be a better and be able to create a better probability prediction model just as the case was with what I said right now for us is going to be slower can use boosting for binary by yes you can absolutely OK but I'm just going to focus on random forest in this segment so here is how this threshold business comes into play and that's actually not trivial and it took us a lot of time to understand that and hopefully we do understand it now so here is what what you can do you can there are two ways once again there are two ways two options how to do classification that random forest one thing is you can build what is called a probability forest and use threshold search with the algorithm and the other is called classification forest and then use the threshold formula I'm going to I'm going to say a bit more about these two options but basically the idea is that you can hide there have should not be predicted probabilities is outcome of the random forest and treat that in a way as it is why was continuous so the outcome are predicted probabilities and then you can use this search algorithm that that that I showed you before or you can make you can you can tell the algorithm what is my threshold and then the algorithm will have classification zeros and ones as an outcoming so when you have probability for us then this threshold search algorithm you are predicting probabilities and then use them to find the threshold or use the formula to classify and you are going to aggregate probability predictions and and used predicted probability that that are these averages and then use kind of classification simply applying the optimal classification threshold to the predicted probabilities right so the outcome here that comes out our probabilities ours are numbers not 0 ones and then you separately use search algorithm to find the best threshold then you have the classification for us when you need the formula because you're going to add that formula into the algorithm and the algorithm that itself used that formula and carry out the classification in itself so when you're interested in using random forest for predicting probabilities then this is the way use random for us to have to turn out predicted probabilities when you wanna have classification this is the right approach to have classification through threshold formula right and then you are interested in classification we can use both we can use this which is only good for classification or we can use the probability forest which can be used either S and output either as a final product predicted probabilities or combined with algorithm and searching for threshold and making classifications I mean this I think seems pretty apps abstract and when you know I go back to the case study I think it will be clear and then you will look at the code and that will be that will be even even clearer and you know what we found is that when you are into classification and your data is large enough and everything is good enough then then these two approaches predicting probabilities 1st and then using the the threshold search algorithm or doing it in one go using classification forest and the formula the results that come out of this are pretty close they are not the same I either are OK to use and it turns out that in the end they are not very different because the optimal threshold is pretty close to the by optimal threshold by algorithm is pretty close to the one by formula the key point and then I will stop for questions the key point I want to make here again is that if you want to classify you will need a threshold and you will need a loss function for that now it turns out dead people who created the random forest for classification had a default loss function built in and the default is 0.5 they gave a fancy name called majority voting and all this seems pretty convincing however this is not right right because when the loss from false negative equals the loss from false positive that loss function could be true right as Bruno was saying it could be true but he doesn't have to be and this default is only valid if you have such a loss function that means that the default setting of classification random forest should not be used right because it's a special case it could be right it could be wrong but do not use something just because it's the default setting can you use 0.5 of course you can but you don't have to you can set the threshold in the classification for us you can you can have any value you want Ouch I want to say about the case when we have FPNFN equal to each other I think it was when it was random right Nope yeah it's not it it it yes when when the frequency of them are equal to 1 equal for all thresholds my dad was random but here we are talking about the loss the loss associated with different errors that was not on the graph before it says it's the loss when you pick you know the threshold is is the outcome that you know is the result that you pick by comparing losses so is the rate that were the same but not the last the loss is something that you defined again if you remember one thing from today it should be that a loss function is something that the analyst define and the loss function will drive classification and without loss function you're unable to classify So what do you do if you don't have a loss function? So in some cases right there is a direct loss function coming from your business case it could be 0.5 it could be something else So what do you do when you don't have a loss function say a reasonable estimate that's exactly right so I don't know like you know loss function I'm not sure but I feel that making false positive is more costly so why you know so I'm just going to say well here is this here is this formula so I think the loss of false positive I would say you know it's one third or fourth negative so it's going to be one 1 + 3 so it's going to be 0.25 or one to two or one you know I I feel it's a large difference so it's going to be 1 to 10 one 12:50 right so when you don't know you make one up but by making up you kind of pin down I think about the problem whether your threshold is 0.2 or 0.25 or zero point 15 that's not a huge deal but whether it's 0.2 zero point 5 zero point 85 that's actually a huge differences will see when we have the classification for us and then there is an observation and let's say that three trees classified as zero and that one tree classifieds it as one then in the end we would say that it's zero right or how to be aggregate them we look at so we will collect all these outcomes and then so we're going to have eventually a set of zeros and ones right and that's going to be our predicted this is how it's kind of different so we're going to have a set of zeros and ones and then we're going to have the threshold telling us you know which one to pick whether it's zero or one so there is no you know you don't use majority voting anywhere each of the trees will give you a zero or a one you have 500 trees so have you have 500 zeros and ones and use the threshold to decide if it's a zero or a one also dental threshold is the in the classification reclassification station presented that the threshold drives higher pick and that's the difference between these two models and again right now as I was just explaining this the I mean there is the only way you will see it is when you look at the code that's when you will understand the difference I think for now the point is that there are these two ways to go about it both procedures will need to have the threshold search in it so that you can classify either by 4 formula or by search that's not the main thing the main thing is that you need to you know ensure that the threshold is there 'cause if he if it doesn't then it would just take the 0.5 and that's not the right threshold because we know if we cannot classify without the threshold no matter how smart the algorithm is without a threshold in the loss function there is no way to classify professor may I ask something yes so to my understanding on a concave function but to sacrifice to gain I do specificity or sensitivity to get rid of this false positive errors but the model itself build dictates how much to sacrifice I guess so without selecting the model how are we going to decide the threshold is essentially going to change so this is I think it's I think with your alluring tool is that eventually our approach of selecting the thresholds and everything would kind of influence which model which probability prediction model we start with it's true it turns out that you can you can kind of combined threshold selection and model selection is very complicated and there are only certain certain cases when when it's fairly easy there is some reference in at the end of the textbook for this but you can check Anne I mean there is another way to think about it if you had a simple model so for those of you who are have some econ background when you think about optimization that there is some there is a curve right and then you which is some some indifference curve and then you have a budget constraint right and you are interested in how that budget constrained and the indifference curve kind of gets you an optimal with the Roc curve and the kind of budget constrained translated to the false negative false positive rate you can kind of have something similar in mind but it's like it but because our models are really complicated it's not as simple as that but you know you can just forget it actually it's really in brackets maybe it's just confusing at this stage but you aren't your point about maybe our probability model selection should be influenced by the second stage it's it's true but it's not you you first pick a probability model and you use that to classify so random forest turns out to work well for prediction when the target is binary and you may always use probability prediction you can use the classification for us when you have a classification problem but again the important point is that when you use when you do classification you should have an explicit loss function and that's clear when you do logic because it's really 2 steps but you can just run classification random forest and you will never know that it has a default of 0.5 so you need to pay attention OK so the last point before we go to to the to the case study is a kind of a technical technical note Ends on the assumption that you observe a decent amount of zeros and a decent amount of ones logic models random forest classification everything only works only work if you observe both zeros and ones not in some data sense it turns out that either zero or one is actually pretty rare this is called class imbalance or class imbalance means that either the zeros or the ones are very rare in the data set that you have there are some datasets where this is really typical so data set about fraud Troy you and transaction so imagine we have a list of credit card transactions you want to find fraud Fortunately fraudulent transactions or you know 0.01% of all transactions or something like that really rare when you look at sport injuries it's going to be below 1% now what is rare I mean I don't know it really depends on the size of the data set typically when you have below 1% and then it could be a very if you have billions of observations then the rare is something less so the larger the data set that you know the more problematic case can be solved so why do you think by the way this is the case why is the case that in very large datasets class imbalance is less of a problem so what's different when you have a billion observations and because even if really really small percentage is still a relatively large number of observations which we can analyze not relatively but but but you're right right so relative compared to small case is true but right so when you have a billion observation even something that is you know 1 zero point 1% it's still going to have a lot of roles where you observe it right when you have 1000 observations and something is zero point 1% it's like 1 out of 1000 when you have 10,000 it's end when you have 100,000 is still 100 I said it's pretty infrequent and we know that anything that is infrequent is really unstable you feel me I say anything anything that is rare is a problem wrestle class imbalance is a problem for not very large datasets and and it's a problem because all the models that you have seen assume somewhere deep down assume that you have both classes observed decently so it turns out that when you have a large problem of class imbalance and the methods that you have seen or not very good at handling them and that's true for both predicting probabilities as well as classification but imagine you have to classify with the model or predict probabilities when in your data does the Y equal 1 is like zero point 1% but imagine being the model that says it's zero always which is not super useful easy but you have to kind of beat that model so This is why class imbalance could be could be a problem so that the consequences it's it's that the models will not really you know we will not really workout well and cross validation and everything is just not going to work well because there are so few observations or wise that that it's possible that involved forward you have 10 of them in another one you have only three right when you have small numbers anything is kind of possible and measures of fit that you have will not really be able to to select across models simply because you have to always beat let's just say it's 0 I said the consequences of class imbalance is poor model performance and that the set up for model fitting and selection or not ideal OK that's that these are the these are the consequences So what can we do and there are in there are two things that you can do the first is to acknowledge that we have a problem like every solving every problem is start with acknowledging that we have a problem so the first insight is when you see that in your data there is a very strong class imbalance one of the classes are very infrequent you kind of need to be ready that your model will poorly perform maybe in as much as being completely useless right so when you have to predict and classify then you have to predict probabilities and classify zeros and ones the first thing you should do is look at the frequency of Y but how how frequently are the two classes are there that should always be a first thing because often the case is that I just don't have enough observation to do this 'cause one of the classes are really infrequent OK no suppose you have large enough data so that you can do something and then you need an action and the action is to re balance the sample for the purpose to make the models work better and this is not the class to explain why this is the case but you can you can read about it you can read about it if you are into kind of math and complicated models but for now you just need to believe me that the predictive models that we use both logic and random forest or or performing slightly better when you re balance the data So what does rebalancing mean you either reduce zeros or increase the frequencies of ones so either downsampling or upsampling oversampling right so when you have 99% zeros and 1% ones you can either increase the ones which means basically just resampling them and increasing the frequency randomly or you can take the you can take the 99% of the zeros drop 90% of them and therefore increase the relative shahrouz once and then there are smart Argo algorithm that kind of combines both one of them is called smoot and that's somehow very popular my experience is that what whichever you do seems to improve right which you do does not seem to matter that much but doing one of them is kind of helpful I find out sampling the most transparent way because it kind of tells you that although I think I have 10,000 observations in reality I have ten 10,000 at any kind of helps you to focus on model building and all that teacher but doing this downsampling there is no way to actually create some bias into your data I mean you're doing it randomly right so you're not creating buyers because you're randomly sampling from the ones and in downsampling you're randomly deleting from the zeros OK but still so I'm like not touching it at 1000 months I mean like it can it goes randomly in the terms of the where the event is not happening but not when the event is happening but now when they even understand happening in this case so I know I understood that we are like it's random from from when we are reducing the distance from 99,000 two 9000 but still is there no way for you doing this you crazy I mean you are you are right in the sense that we are going information right that cost and that could that could bring in instead bias we don't know in expectation it doesn't but it can but you're doing this because we know that the probably is some kind of a balance otherwise they would just not work well so we're kind of destroying some information or oversampling we are kind of putting in something weird but we're doing that in a way too to improve the performance of the probability predicting model yeah but the question was more related to the fact that you randomly drop some observations from the from the from the event itself and you're actually didn't of course in order to increase the rate but you didn't decrease nothing from the from the one day event does not happen so you simply didn't touch in the 1000 like in that sense you know like to drop a little bit from the auto one as well and in the instance you don't wanna lose of their rare events I think I mean I can you know I can see that this could bring in some biases I can I can imagine that reweighting something weirdly so I can see your point that this can be kind of adding bias I'm not 100% sure that this is happening but I I can see that's not impossible but I think but the problem is that all these models that are about probability prediction requires to have some kind of balance in one way I mean either way is kind of a little bit of cheating Anne in do you know this this syntactic algorithm this is smart algorithm is kind of trying to incorporate it and do it more more efficiently than just you know coming up with something from your head of 10% or 1% or something like that but I think the main I mean the main takeaway is that when you have a rare event it's very hard to predict that no yeah you can try you can try stuff they will help a little bit I mean these re balancing stuff right they are not a major major major major major major improvement there are some improvement alright they're not like OK that's going to solve up our problems they are helpful a little bit so that our logic models are you know the properties that we know about them or are there but I mean whichever you do right you you kind of destroying or the data or redoing the data or you just ask in the sense that you only drop values from 1 phone case and not the other so I'm not saying like yeah I mean like England down from from 99 K2 9000 and and the other one is still 1000 I'm just not saying like you could do have dropped like for example 100 observations so that maybe you could I think these so I mean these are the three most frequent ways to do it either destroy from one of them or oversample the rare one or use an algorithm that kind of does both Anne so to summarize before we go to the case study we are in the business of doing two things predicting probabilities and classification but we have an outcome that is a binary target variable some event happening or not and probability prediction is going to build the model that predict the frequency they expect frequency of this happening in our data when our aim is probability prediction and very often this is the case then you want to build the best model and you pick the model by mean squared error or AUC and then you stop or if you want to actually classify right then then need some threshold and you need a loss function to get that threshold and you need to find this optimal threshold either by an algorithm or by a formula select see OK so in the next 40 minutes or so I'm going to start and see how far we go discuss about the case study this is the most maybe the most complicated and complex case study that that we have and it's complicated because it's it's pretty I mean there are there are many things happening so the bed and this is actually based on the consulting job that I've done with my cost or like a couple of years ago so this is really inspired it's a version of what we actually did in the discussions or kind of versions of what we actually not exactly the same but I kind of version of so here is the business case and the reason I I want to take this case that is separately is because this is the one I want to have some discussion about how we use these models in business is my understanding is that that you are interested in business that's what you are that's why you are here so very often when you have you know banks and business partners they may be interested in the financial stability of their of their partners so very often there are cases in business where you want to predict that the partners that you have suppliers customers are going to be around in the future rain and and there are services and companies that do that and be work for one of them but how likely it is that your partner stays in business or or exits and so our task will be thinking about this probability prediction and often the probability may be enough and then maybe thinking about classification as well and and so we are our aim is going to be predicting corporate exit or default I'm going to be a little bit more specific what I mean by exit and I like this case today because this was exactly the setup that that the client just said we need to predict default and we don't really understand anything beyond that we have data um and you tell us how to think about probabilities and classification and all that and want to communicate what shall we communicate to our or our clients is that they didn't specify what exit is and importantly they just said so we want to tell when their company is not around So what do you mean by not around well when they are you know when they defaulted or when they stopped operating or but like for how long and you know there are many questions and so when I when I talked about two weeks ago kind of defining your target or or labor label design This is why I did that because often the business does not tell you exactly what is your target and you have to come up with that right so the data use comes from this note which is a European Anne I think it's based in sunbreaker in Scandinavia need firm level information about headquarters it's a very complicated data we created a panel at this so basically we observed firms over multiple years right so it's a firm times year panel data which we created I'm not I'm not sure maybe I already shared the road data but you can imagine that the road data is really long and complicated and it's relational database right you have firms that you observe year by year these are the financial information you have management information where you observe if the firm is under safe or in management between in a period same with ownership there are headquarter which is unchanged overtime it's a snapshot right so there's a lot of model of linking of datasets and the end product is a panel data cancel one role in this data is a company ID in the year OK that's going to be our starting data and what we'll do is we'll focus on a cross section of 2012 and we want to predict a firms that are in operation in 2012 or still there in a couple of years so the first thing is is label right or or target engineering defining what our target in this exercise will be right and and again there is no exit in the data right and again this is very frequent in real life that nobody tells you what exactly the target should be they give a concept and you have to kind of carve it out so the way we think about it is that the firm is in operation in year T but is not in business in T + 2 so we create a binary target which we call exit which is 1 if the firm exited within two years that could mean is there a question no sorry I just didn't have voice so I had to reconnect and I had someone so but you can hear me right cannot find OK super so in target it's a binary code exit right it's one if the firm exited within two years so that means it was still in operation in T it may have exited in T + 1 or in T + 2 right that by 2 + 2 is out and zero otherwise it's a very broad definition right it could include defaults or forced exits it could mean orderly closure like you have been producing chips in you decided I don't want to produce chips anymore and I closed my company it could mean acquisitions I used to have a company but I saw it too multinational so the company does not exist anymore it's a very broad definition would I have created a different one arguing that well acquisition is not a problem I could have if I have more information about what happened to the firm then maybe I can but this was not our case or it was just modestly our case the data was very noisy on the details so we decided to keep it simple but clearly you can have a different definition there are some cases when the firm does not exist in T + 2 but it exists again in key plus 3 what do you do well you make a decision right there is it is it something that is reasonable maybe you want to keep it maybe you want to drop it you know you don't know but but you have to look for these weird cases and make a decision so are you with 2012 we kept in the data firms that were operational in 2012 we also kept the new firm so firms that were established in 2012 you could have dropped them we also thought that well I think you know we can have a decent chance to predict exit off like small medium sized firms but not like really really really small ones because they may be just you know zombie firms or they may be just like sleeping consultancy of a person and above a certain threshold they are just too large and maybe we don't observe enough so we basically kept our sample between 10,000,000 and 1000 euros I didn't say which country it is it's a medium sized European unamed country the data comes from a medium sized medium size unamed punches which I cannot say which one EU member and and so we're going to end up with about 20,000 observations an in that sample we're going to have a 20% default rate what's the default is still or exit rate to be more precise so exit is still much less likely than staying alive but we don't we have no problem of balance right so 20% is perfect so we have to make quite a few decisions about which variables to keep you know what are they what are the most important variables that we want to use in our model which one which are the ones that wanna clean certainly we're interested in balance sheet information we are interested in ownership we are interested in industry classification but there is a lot of information about the size of of management or or rhetoric has the board of directors or not or you know the nationality of the owners so we just kept simple and and used for in or not but you could you could have done to credential more but the 20,000 observations you know there is only so much you can you can get so the key predictors when you look at the features that you will see the key predictors of firm exit will be the size sales and sales growth the management whether there is foreign manager or female young number of managers in which region in the country the company is located in industry how long the firm has been in operation and a bunch of other financial variables from the balance sheet and profit and loss how did you find these very posts it is dumb this have you tried it I hear you very poorly I don't hear you well can you come closer to the mic yes so users I don't hear you well can you repeat your question beside that can you hear me now now is better yeah OK so based on what have you decided that these these were the key predictors skills yes this is a good question so this was domain knowledge there is a there is a large literature on firm defaults so we read a bunch of papers and these were the variables that seem to be the most important ones that we have but still it's going to be a very large set of variables because we're going to try out a lot of financial variables and a lot of them will make no difference so it was it was basically what we have in our data and we kind of simplified so we have the management as young and old rather than the actual age so we made some feature engineering design to have like a model that is not very large but the decision came from domain knowledge and in our case that was reading papers about firm firm defaults and especially because you know I mean these matter more for for for logics so when we have front door for us we can just have much more but as I said before it's typically good to use domain knowledge to have a relatively small model and to check if your numbers make sense and so functional form and everything is going to be is going to be he was very quick financial data before have you seen balance sheet and profit and loss earnings data I see one person really and and and that's it jamba is nothing so let's save it for one firm or for like no no for for you know a data set made of financial data not for bond firm for thousands of firms so when but even maybe for one frame but basically so one thing that that you will see is that there is a lot of weird ship going on so there is a lot of things that should not be the case there's a lot of errors there's a lot of weird like not many but like some so feature engineering in our case is going to be is going to be important so let me let me tell you some of the decisions that we made and you know this is going to be related to your second designer so you can make completely different decisions right some of the decisions are again based on on what they have read but some of them are fairly arbitrary so we look at growth rates and believe that that's important and use one year growth rate of sales but you could have a longer period you could have growth rate of net revenues you can have you know whatever whatever you want ownership and management you know we try to keep it simple because there is a lot of missing observations so we want to have something that is well covered in the data and sometimes we you know simplified created young and old CEO used you know there are there are age of CEO CEO that for below 15 that were age that was above like 110 right so there there were weird stuff the age of the firm I mean there are firms that were established in 2097 and then there were firms that were established in 14 something so both seem pretty unrealistic for this country and so we made some you know some changes as you will see there are too many industry categories so there was one question I think last week or two weeks ago about what happens when you have a categorical variable with too many values industry classification can have a lot of values So what we did is we combine into a few but the aggregated them foreign ownership So what is a foreign firm we had to pick a threshold and we did but you can you can do different things you can have the share for in foreign ownership anything like that functional forms sometimes we took logs as you will see sometimes we took polynomials basically we look a lot of scatter plot lowest and make decisions when you look at the code you will see a bunch of graphs that didn't make it into the book but kind of go through some of the variables and give insight of why we picks in certain functional forms never a bunch of cases when we had to make some cleaning inputing values replacing missing with zeros when it makes sense the key thing and that you may remember from day one so when you make cleaning step and the binary variable a flag that kind of picks up that there was something changed and add that into your model because maybe that cleaning is brings in a buyer so you want to control that with the flag that that that that that that must that may have come in India one or not you can you can look it up so there is 1 1 technique that I want to introduce that could be useful in many many cases so this is log sales growth so like says growth in percentage something like that or close to that and the probability of default but you can remember a scatter plot of a binary Y is basically dots on one and Z1 and zero right so this is exactly what so the green dots are the scatter plot and the blue are the lowest this is annual growth in sales and the probability of default so when you see a graph So what what do you say what what I mean sorry Garber isn't this the difference in growth year over year so the acceleration would be sorry you're right you're right the graph should should just be no it's the the difference of scenery is good it's the it's the you're right it's the difference in log sales you're right that it's a difference in log sales you're right it's a difference in locks here sorry it's a difference in lock sales is what it is so it's a it's a it's a growth rate an approximation of growth rate and not very good one because some of the numbers are right are high but think about it there's a growth for annual growth rate you're right thanks make a mental Note 2 2 make a mental and I hope I will remember your right so after this correction So what do you what does this graph make you think I mean that's your life in the next 80 years or 60 years or whatever years you're doing this before you become management looking at graphs and trying to figure out what they mean and then you make into management and then you make other people do it growth and growth slows down it tends to lead to higher default rates I know it doesn't seem but companies whose growth rate slows down in can expect to default on average with a higher rate or higher percentage for higher chance I mean it's a nonlinear Adam right so it's it's I mean it's it's true for some segments but I mean you know it's it's there's ups and downs I mean clearly there it goes up and then it goes down then it goes up then goes down maybe companies that have a stable growth small stable growth over years have to interpret this mother chance to I don't hear you I don't know if others but I don't hear it 'cause I don't hear you I don't hear you sorry yeah it did but I mean look at the whole whole thing like just just white So what does this weird shape tells you what could be the reason behind this weird shape maybe the reason is that changing the sales really have something to do with the being default I mean imagine So what would be you know what does the value 5 means it means that the the difference in lock cells is 5 so it's like 8 times something like that 10 times you can see large number you cashed out as a business owner you took the money around that is that is possible but what else is possible the same is very large minus number it means that your sales has completely collapsed In the end you see this U shaped relationship so why what does it mean how should we think about this so was there anyone telling you stuff about extreme values and errors and stuff like that so hopefully there was a couple of months ago but imagine I mean these numbers on the two edge of this distribution are probably defined by a very few observations I mean there are very few firms that would say I have my sales growth 10 times or I have you know my sales in in 110th in a year in a year the year before exit right I mean I mean there these are very rare cases so we're fitting a curve on the very few observations here and here as well could this be overfitting fitting a curve on stuff that is very rare and so one reason this is the case is that simply there is just very few observations here and here like there's 20,000 observations altogether and you know if you look at around this it's like 100 and this is another 100 and then most of the firms are really here that's the one thing you can do is you can look at only the part that is between one and a half and minus one and a half I mean it still is large change right it means almost doubling your almost tripling your sales overtime or you know coming to us losing a lot of your sales this is 99% of the distribution this is what we really care about just wanna make sure that my hands are in the camera window and this is what we care button and now we can interpret this part because this is you know this is the part that we kind of trust I said once you cut out the extremes the rest kind of makes sense right if you are not falling but growing the default the probability of default is coming down but at some point if you are growing very fast it starts to kind of creep up makes sense so it has a weird shape the vast majority of the distribution is OK right this is exactly just cutting out like this bit my guess is exactly that nothing else the zooming on that part that we believe so here is how you deal with situations like this so when on the extremes there is something weird going on and then you work with financial data or many other datasets that have want it is this is this is going to be a case that we have some weird stuff going on on the end of the distributions in domain noise domain knowledge is important so let me tell you about the technique how to deal with this it's called winsorizing winsorizing is the following idea you have you have the variable like lock cells and you look at it and you and you see there is some weird stuff going on in any way you think well if a firm grows by by five times maybe you know the data is just not reliable or it's collapsing by 1/5 the data is just not reliable so I just I just want to simplify the edges to have a better pattern and not let these extreme values drive my results so but winsorizing does is you you focus on a single variable log sales lock cells is between minus 10 and whatever plus 10 you make two thresholds in our case it was one and a half and minus 1 1/2 whatever is in between you keep as is whatever is greater than 1 1/2 you really you change the value to 1 1/2 from whatever it was and the same from the lower values so whatever it was below minus 1 1/2 you replace it with one and you replace it with minus 1 1/2 so this is how the the original and the winsorized data will look like for the minus one and a half and plus one and a half region it's the 45 degree line so it's exactly the same but the edges are cut off but you don't drop any observations you just change their values it's changed at values so that to reduce noise the noise that comes from these extreme weird stuff you don't want to throw these observations right beer valid observations you just think that for whatever reason the numbers are overstated or understated and there is some noise and there is some weird thing going on So what you do is you replace their values and you add a flag for this part and this part so you add two more variables to capture that you made a change here and you made the change there it's called winsorizing and you can do that by like 1% in 99% I don't like that but you can typically look at the distribution use some domain knowledge and determine the threshold and the idea is that you are it's a way to deal with extreme values that are likely to be errors it's a conservative way because you're not dropping them we're just replacing the values it's a practical thing it's not like game changer but we use this but we had to use it in this case study because especially when you when when you use it so do you think this is this is equally useful winsorization for logic and random forest where it's more useful for one or the other was so there's 10 more minutes so try to 15 I think for the logic no need since logic's bottom and top edge is still flat and by default and maybe for endevours force I have no idea any other view I would say the opposite that the logic function like under no outside the logic function there is a linear probability model and if you have extreme values it's going to is going to end up giving you back 100% if you have extra values in this variable and for the random forest you end up with classifications and if you use classification then that doesn't seem to matter too much but OK any other view I sent my friend as since you're not dropping the values in random forest they will still be there before the split might happen still at the same place 'cause you have the same amount of observations on one side versus the other while in logic and profits you by affecting the outcome variable you would affect the formula or at least some part of it yeah I mean it's in logic right it's true that you know there is this link function but the link function is about the prediction Bryson but not what's inside inside as was saying if there is still a linear probability there is not a linear probability but the linear model there is just a link function that that kind of you know the predicted values are not the not the model itself and the model itself can be heavily affected by extreme values just as the case in OS right when you have a linear model in the extreme values they can kind of you know make all kinds of weird stuff happening so for logic it's actually pretty important to do all this for end of forest right we expect the trees to potentially find these weird observations and make cuts and get kind of rid of them or cut make an early cut and say well these are very large weird values I don't know right so we expect that the trees like any feature engineering right and it's a feature in Geneva functional for any functional form stuff that you do is more important for linear models or like regression model than it is for for for random forest and there's this is no no difference here so we have all these variables and and you can see a list of model features that we have bunch of variables that you have a bunch of winsorized financial variables flags that come from winsorization or you know for imputation any other stuff so we're going to have a bunch of these these variables these are the financial variables and then we have more from HR from we have some variables capturing data quality we have some interactions and you know they are going to lead into logics that we estimate so once again we are in the place where we build a bunch of models how we do that well you know we consider the few variables that from the literature it will say are the most important and then we keep adding more and more and more maybe build 5 logic models and then we have logic loss so IBM and the logic lost so when the logic seems to be in our case not super different I mean they have you know we have 20,000 observations 150 predictors are not huge right and and again I mean this is super arbitrary So what models you build you want to build more and more complex models so that you can compare you you know you let your your domain knowledge drive which ones you build you try out a bunch of them and then that's it and then use at the end you lost so to try to use your most complicated model and simplified and see whichever works best again try to have a simple model that is easier to explain and then have more and more complicated ones hoping for a better performance so we have 19,000 observations 15,000 will be our work said we're going to use almost 404 thousand in the holdout said and we only going to use it for diagnostics right so we're going to do the cross validation and everything in the works it just as we did for continuous case and basically we have different models with a bunch of variables and coefficients and and then the estimated cross validated mean squared error raeann and be used at 2 speaker model in in this case it turns out that these models are not very different and it turns out that you that once you have this model then in terms of cross validation error you don't it doesn't matter that much add again a little bit but you lose a little bit it's not it's not super different in this case is I can to pick you know when you have miss graders are close to each other I pick the similar ones but you can pick whichever you want there it doesn't seem to make a huge huge difference and we put it there so this is what we're going to be going to use so we picked picked a model by mean squared error score and for classification we need threshold so let me say a few words and then then then we go and then it's over today and you're going to see a bit more in the seminar and next time we're going to finish it next week so this is actually the graph that I showed you before in the lecture of course they come from from this case study and this is the Roc curve that is based on this case study and we are and we're kind of moving through the thresholds between 0.05 and zero point zero 75 by steps of 0.05 so this is what you can this is what you can see here and again these both of these curves are from the case study so they are generated by by this and and this is the Roc curve that that corresponds to one model right remember 1 product one probability prediction Model 1 Roc curve and so this is AUC calculated for all these models for a mean squared error remember we want the smallest one for AUC one the largest one the one that is close to 1 and as you can see the ranking is actually pretty close I said both both of these are saying that model 4 is the best and in particular kind of the signal is stronger for a we see we like it more as a model selection tool is more widely used in practice so that confirms our choice on mean squared error it doesn't have to be OK so AUC and mean square does not have to pick exactly the same model they're going to be models that are close but it doesn't have to be the same residue are in the door in Indiana in this prediction business or in statistics in in a broader term there are things that turn out to be exactly the same and there are things that are very close but not the same this is a case for stuff when when two metrics are pretty close but not the same and they don't have to be the same OK so before before we go let me just tell you So what happens when we have two different thresholds right and that was that was something that came up like an hour ago so how we get the confusion matrix or the classification table same thing you will need a threshold and we need to have done you know the classification so these are the two these are the two thresholds one is when the threshold is 0.5 and one is when it's 0.2 so the one is like 5050 and the other is the sample mean So what is what do you see in this table that's going to be the last thing we do before rebreak what do you see here what do you learn from from from this why am I showing it to you this late so what's the message of this this table looking at two different thresholds this very causes a very different prediction result right so this is 1 important takeaway right I mentioned this earlier dificiles matter it's not better it's 0.2 or zero point 15 but if it's 0.2 or 0.5 or zero point it matters and you can see an example how it matters so the first observation is it matters then can you guys tell me more about how it matters so in the in the first case we made more errors in a way that we said that the company will stay in the market but it actually existed but we made less errors predicting that it will exit but actually it stayed various for lower threshold these are The inverse of each other or they're not the same so it's really what we want to achieve at the end so do we want to predict with more certainty that someone build stay on the market or not or better we are more interested in who's gonna default or not this is absolutely raining in in you know if we go back to the graph you know 0.2 is is around here then in 0.5 is around here I said this is how we move on the Roc curve with these different thresholds and those thresholds then you know these are the these are the inputs to calculate where we are and you can see as you said that the numbers are different right so the different the errors that we make right the areas that we make are different but the actual stay in the actual exit here these are of course the two same numbers because they come from the data but the total predictions are different right here we predict a lot of firms to say and only very few to execute predict a lot of exits right and that means that that you know the ratio of getting it right for these different different parts or just just very different so this is how thresholds matters and This is why we want to pick a threshold by by loss function very good so let me stop here and we're going to continue going to also see it starting on