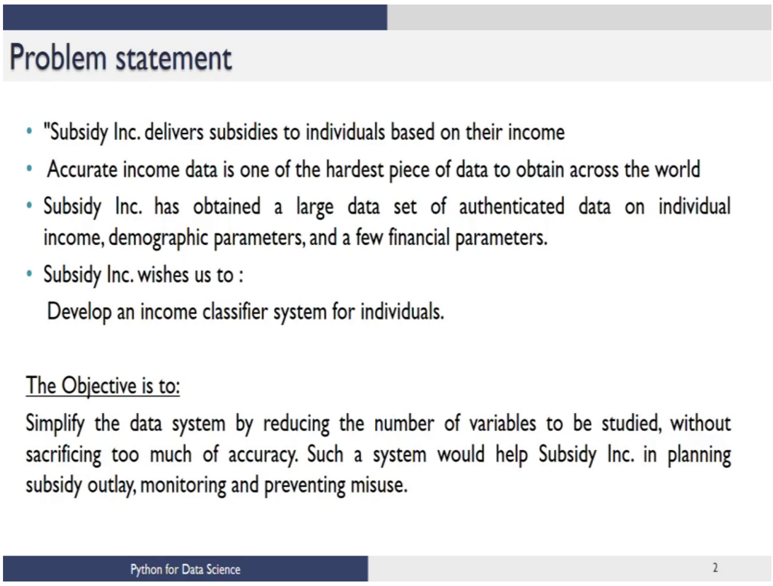
Python for Data Science
Prof. Ragunathan Rengasamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 22
Introduction to Classification Case Study

Welcome to the case study descriptions for this course. I hope till now you have practiced and learned enough Python, to be able to do programming comfortably. Our aim now is to basically help you understand, how you could use Python as a tool in solving actual data analytics or data science problems.

And we are going to give you two case studies that we will teach; and one of them is on classification problem, and the other one is a function approximation or a regression problem. So, what I will do, is in this lecture I will explain or Introduce this Classification Problem that you are going to solve. And then we will go through the data and actual solution to the problem in another lecture.

(Refer Slide Time: 01:15)



Problem statement

- "Subsidy Inc. delivers subsidies to individuals based on their income
- Accurate income data is one of the hardest piece of data to obtain across the world
- Subsidy Inc. has obtained a large data set of authenticated data on individual income, demographic parameters, and a few financial parameters.
- Subsidy Inc. wishes us to :
  Develop an income classifier system for individuals.

The Objective is to:

Simplify the data system by reducing the number of variables to be studied, without sacrificing too much of accuracy. Such a system would help Subsidy Inc. in planning subsidy outlay, monitoring and preventing misuse.

Python for Data Science 2

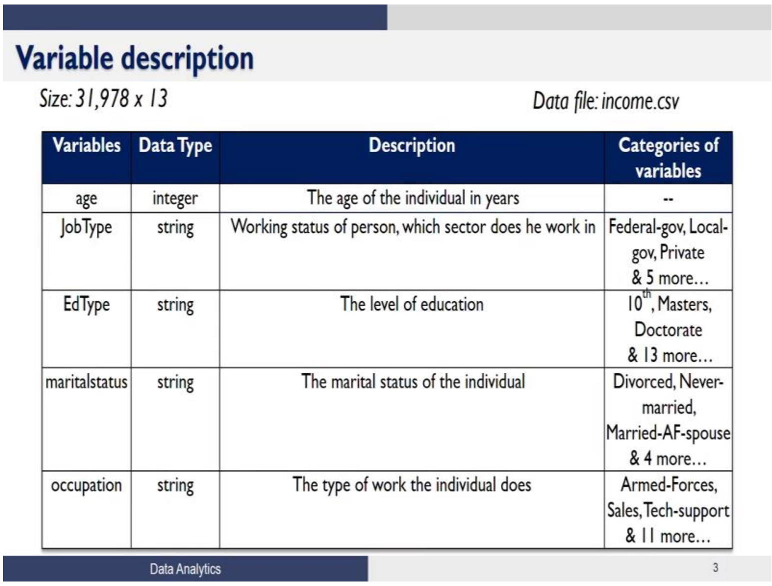
So, the problem statement is as follows; let us say there is a company called Subsidy Inc. which delivers subsidies to individuals based on their income. So, if your income levels are different, you get different subsidies and so on. However and whenever someone new comes in, it is very difficult to get information about personal income. So, that is one of the most difficult pieces of data to get. So, here is an idea in terms of using an already existing database, where we have various attributes of these people and wherever personal income has been disclosed we also have that as a data.

So, what we are trying to do here is, we are trying to see, if we can somehow classify an income level based on the attributes that we have for individuals. So, basically this becomes then a classification problem, where we are going to say the individual has an income level beyond a certain value or below a certain value. In this case we have this income level as 50000; and what we want to do is we want to classify individuals with less than 50000 personal income and greater than 50000 personal incomes.

So, this is basically a simple classification technique. And you one might ask, if you were able to do this how can you use it; if you have multiple customers that come in, then we can kind of identify the proportion of customers who are likely to be having an income less than 50000 and the percentage of people more than 50000; and that could allow us to be able to plan an outlay of resources and so on. So, that is one use case that you can think of.

Another use case is really, if someone actually discloses an income and if it is an outlier in terms of, let us say someone says they are earning less than 50000; but our classifier says with all these attributes it is very very likely the income is greater than 50000, then it might allow us to look at those particular cases in more detail, so that there is no misuse of these schemes.

(Refer Slide Time: 03:43)



Variable description

Size: 31,978 x 13

Data file: income.csv

Variables	Data Type	Description	Categories of variables
age	integer	The age of the individual in years	--
JobType	string	Working status of person, which sector does he work in	Federal-gov, Local-gov, Private & 5 more…
EdType	string	The level of education	10th, Masters, Doctorate & 13 more…
maritalstatus	string	The marital status of the individual	Divorced, Never-married, Married-AF-spouse & 4 more…
occupation	string	The type of work the individual does	Armed-Forces, Sales, Tech-support & 11 more…

Data Analytics 3

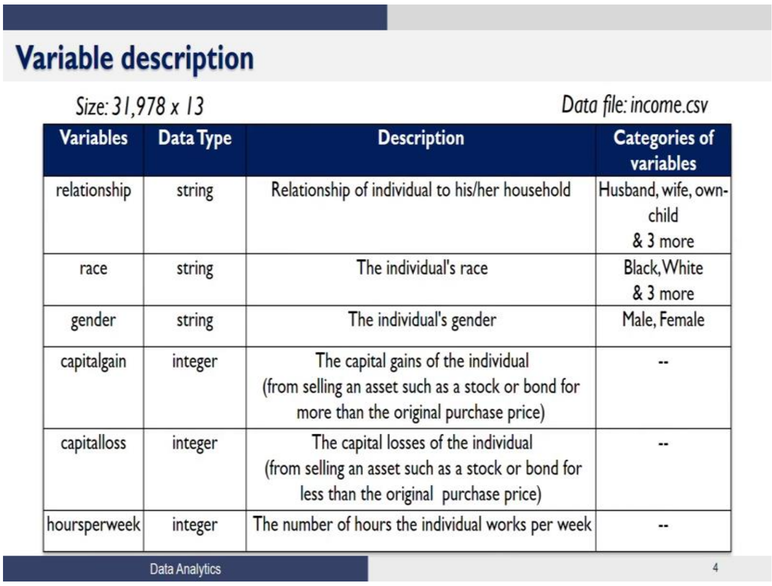
So, whenever we have a problem, data science problem all of this starts with data. So, we first need to understand the data that we have. So, in this case you will be given a data set, which was about 31978 samples; that is 31978 individuals for whom this data is available including their personal income. And this is the data set that is going to be used in developing this classifier. Now I talked about features for this individual. So, we are going to look at multiple features. In this case we have 12 features that we are going to look at, and one variable which is personal income which makes the total number of variables 13.

So, if you want to think of this data in a matrix form, then we have about a matrix of the size 31978 times 13. Now let me quickly go through the variable definitions and what type of variable they are and their description. And this will help you to kind of visualize this data once you actually start working on this case study with the instructors. So, one variable is JobType. So, this is a string and this basically tells us about the working status of the person is he in and works which sector he or she works in, it is Federal government, Local government, Private and 5 more such identifiers for JobType.

Education type is again string, which tells you the level of education and the multiple possibilities or 10th, Masters, Doctorate and 13 more. Marital status is again a string, which tells you the marital status of the individual which could be Divorced, Never married and so on. Occupation is again another string and the type of work that the individual does it, could be Armed Forces, Sales, Tech support and 11 more categories in this case. And as we talked about this you can start getting an idea of why these variables are relevant for this problem; and you can really see that all of these variables are relevant, because the pay depends on the sector that you work in, the level of education, the type of work you do and so on.

So, as we go along you will see that there is this notion of picking relevant data that is useful for this problem.

(Refer Slide Time: 06:27)



Variable description

Size: 31,978 x 13

Data file: income.csv

Variables	Data Type	Description	Categories of variables
relationship	string	Relationship of individual to his/her household	Husband, wife, own-child & 3 more
race	string	The individual's race	Black, White & 3 more
gender	string	The individual's gender	Male, Female
capitalgain	integer	The capital gains of the individual (from selling an asset such as a stock or bond for more than the original purchase price)	--
capitalloss	integer	The capital losses of the individual (from selling an asset such as a stock or bond for less than the original purchase price)	--
hoursperweek	integer	The number of hours the individual works per week	--

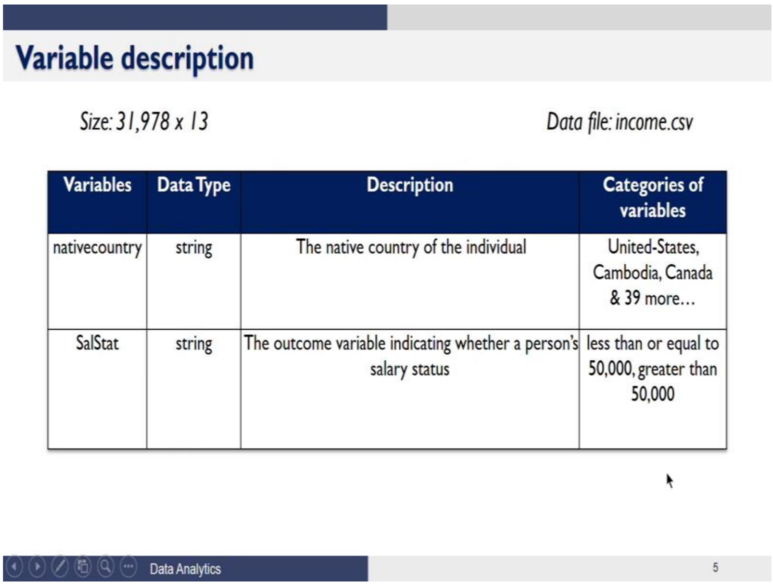
Data Analytics

4

So, moving on, there is a variable called relationship which tells you the relationship of an individual to his or her household, so this could have value such as husband, wife and so on. The race of the individual black white and 3 more; gender of the individual male, female. And then there are other features that talk to the real financial aspects of the individual.

So, there is one variable called capital gain which is the capitalgain of the individual which is a rounded off integer in this case; capitalloss another integer again a rounded off integer which basically as the name suggests is a capital loss. And the hoursperweek that the individual works again it is rounded off and the type is as integer.

(Refer Slide Time: 07:25)



Variable description

Size: 31,978 x 13 *Data file: income.csv*

Variables	Data Type	Description	Categories of variables
nativecountry	string	The native country of the individual	United-States, Cambodia, Canada & 39 more…
SalStat	string	The outcome variable indicating whether a person's salary status	less than or equal to 50,000, greater than 50,000

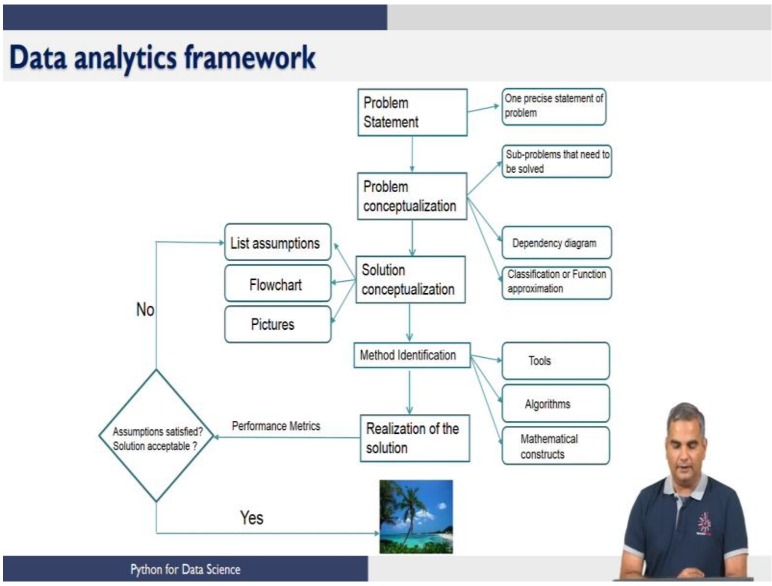
Data Analytics 5

And also there is a variable which says which nativecountry that this person comes from, which could be United States, Cambodia, Canada and so on. And then finally, the output variable which is what we are trying to build a classifier for which is salary status, which is outcome variable, which is basically in this case just split into 2 categories less than or equal to 50000 and greater than 50000; those are just the two categories.

So, basically what we have here is a binary classification problem; where there are two categories less than or equal to 50000 and greater than 50000. And based on the feature set that we have, we have to basically build a classifier that will be able to tell us once we get a new customer. And we put the features of that new customer into the system; get an information about salary status. Now, notice that some of this input features themselves might be difficult to get such as capital gains and capital loss and so on.

Nonetheless what we will try to do is, we will try to see what best in terms of a classifier we can build and then one of the ideas as I will point out later is also to reduce the number of input variables and themselves. So, instead of 12, let us say we build a classifier with these 12 variables and then we drop a few and then see the performance does not change much. But from a data collection effort in the future it might be much easier to get information about certain things and not other things, then we can see whether we can use those of ideas to actually build a classifier that is practical and useful in the future.

(Refer Slide Time: 09:13)



Data analytics framework

Problem Statement

One precise statement of problem

Problem conceptualization

Sub-problems that need to be solved

Dependency diagram

Classification or Function approximation

Solution conceptualization

List assumptions

Flowchart

Pictures

Method Identification

Tools

Algorithms

Mathematical constructs

Realization of the solution

Performance Metrics

Assumptions satisfied? Solution acceptable ?

No

Yes

Python for Data Science

Now, whenever you do a data science problem, you basically have to think about it in some flow process. This flow process I have explained in the other courses that I have taught; but just quickly as an overview in this course on python for data science. I just want to go through this, so that we can see how we kind of map the solution that we build in this case to this flow chart.

The very first thing is to basically say, what the process statement or the problem is, right. So, now, in this case we have already done this, but if someone just came and gave you this data; and then says you know we are collecting a lot of data and this is the data that we have, is there something worthwhile or useful we can do with this data. Then one problem statement might be to say, look can I figure out the salary status of the person based on just the features in terms of that individual several attributes.

And that is a precise problem statement that you can post here; of course, with this data itself you can post other problem statements, but here we are going to stick to this. And typically when it is a very large data science case study, clearly this is not, this is in terms of data still 30000 odd data is not small by any means; but it is still not large by any means either. More importantly this is a very simple binary classification problem.

In many cases when you are presented with a problem, it is really not obvious exactly what data science problem you need to solve. So, in those cases, you have to think about some notion of problem conceptualization. Breaking down a very loosely worded problem statement into multiple smaller problem statements; and then under identifying these smaller problem statements as what type problems they are either classification or function approximation problems. And then once you are able to solve the smaller problems, how do you put them in some logical arrangement of solution. So, that you can solve the larger loosely frame problem.

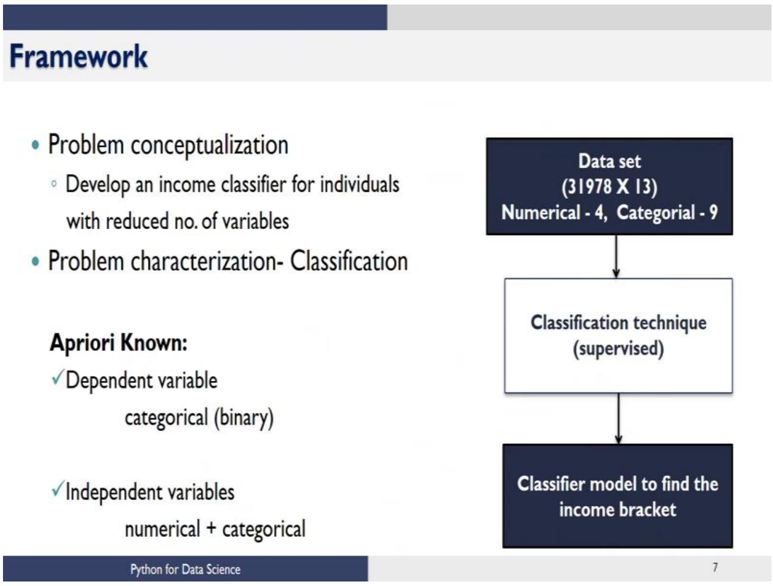
So, that is actually where the intellectual exercise of data science comes in. But in this case, in this first course this is a simple problem you directly understand that this problem can be conceptualized as a binary classification problem, so there is nothing much more to do here in this case. Then once you have a binary classification problem, you know there are several tools that exist some of which we have taught to you.

So, how do you use a certain tool or how do you pick one tool out of these many tools that are available. Again there is a very logical way in which you can do this, you can make some assumptions about the underlying data distributions and model types and so on. And then choose a tool that is relevant for those assumptions and then see whether the tool works and all that. That is little more sophisticated data science thinking.

Again in this course what we are going to do is, we are going to simply look at a couple of tools and then run those tools on this data set; and then basically we are going to judge which tool is better based on just outcomes in terms of the prediction error or the confusion matrix and so on. So, once you do that method identification, then basically you have to realize that solution in some platform of choice.

As I mentioned before the two most popular platforms of choice are python and R. And in this case we are going to use python to solve this problem. And, once you solve this problem and you are happy with the results; then that is the end of this whole data analytics solution, strategy or methodology.

(Refer Slide Time: 13:30)



Framework

- Problem conceptualization
  - Develop an income classifier for individuals with reduced no. of variables
- Problem characterization- Classification

Apriori Known:

✓Dependent variable
categorical (binary)

✓Independent variables
numerical + categorical

Python for Data Science

7

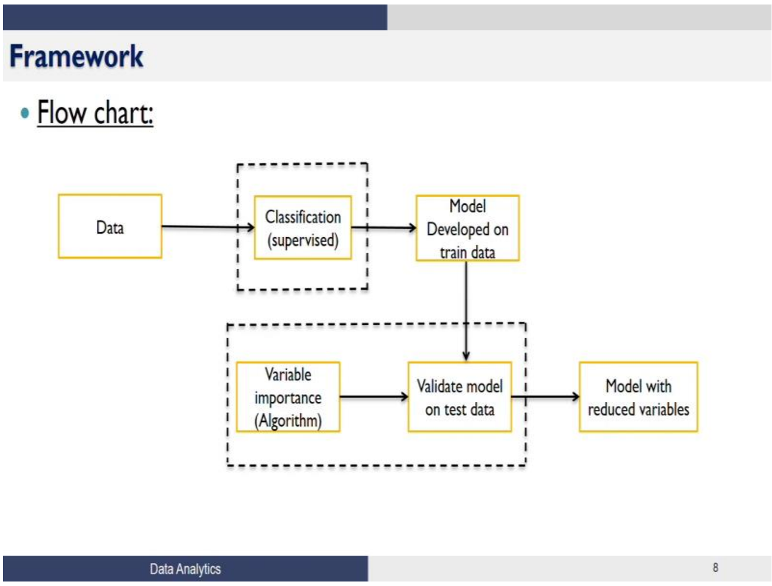
So, now, let us look at this. So, as we said before, problem conceptualization is pretty simple; here we are developing an income classifier for individuals. We have also added another small aspect to this problem which is not really obvious unless you start thinking about it carefully; like I said before you could say I will use all the features and build the model, but we would also like to see if we can reduce the number of features and get comparable performance of the classifier.

And as I mentioned before the reason is, then it makes this solution, practical to be used in the future. So, for example, if you are totally tied to knowing everything in the input future and without those you cannot find a solution, then you might not be able to use it as effectively. Whereas if you were able to reduce the number of feature, so that there are a few things that you need to know to be able to make a prediction, those kinds of systems are much more useful.

So, that is another additional wrinkle I would say, that we have introduced into this problem. So, again we had the data set of 13 columns; one of them is an outcome variable, the other 12 are independent or input variables from our viewpoint. And of those 13, 4 are numerical and 9 are categorical; and this is a classification problem. In fact, this is a binary classification problem and this is also a supervised classification problem; because in the data set that we are going to use to build a classification model, for every data sample we know the outcome variable value.

So, basically we are going to supervise our learning algorithm. So, that it learns the given characteristics of the data. So, Apriori known all the dependent variable, dependent variable which is categorical binary we have already talked about it and also the independent variables which are basically a combination of numerical and categorical variables.

(Refer Slide Time: 15:47)



Framework

- Flow chart:

Data

Classification (supervised)

Model Developed on train data

Variable importance (Algorithm)

Validate model on test data

Model with reduced variables

Data Analytics

8

So, if you think about putting all of this into a flow chart based on the extra wrinkle that we threw into the problem in trying to reduce the number of input variables that we need to solve the problem. So, we start with data, then we build a classification model supervised classification model; and then once we build the model we validate the model on test data. So, this is another important idea in all of data science and data analytics; if you use all the data that is given to you and build a model, then the model is very likely to do well because it has seen all the data.

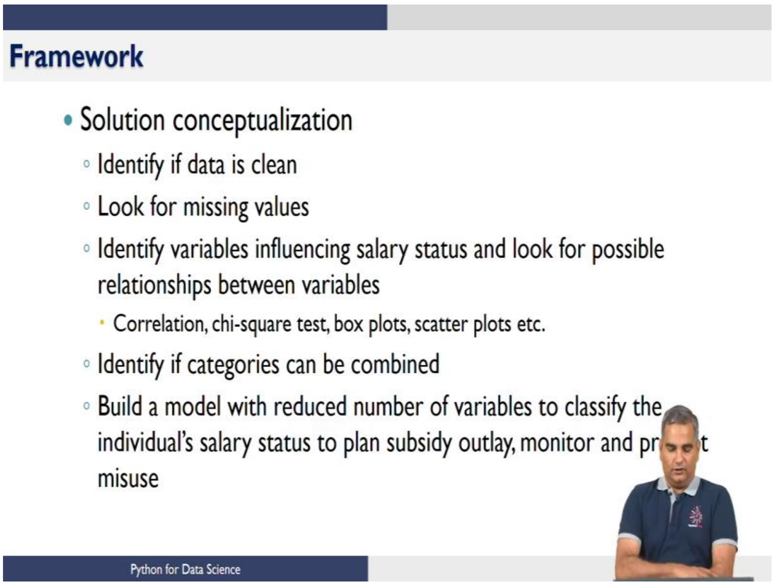
So, basically all of data science or data analytics is useful, only when we can use that really in predictions as we see more samples or more data that comes in. So, we want to have some notion of how well our algorithm is likely to do when I get new data. So, to identify that what we do is, we basically split the data into what we call as train and test data; and when we build a model, we basically build a model using the train data, and then we test this model based on the test data we have left out within the data that we had.

Now, if our model does quite well on the test data which is not seen before; then we can be comfortable that this model is likely to work in the future also. So, that is what we have here as validate model on test data. And once we have this validated model where we are happy with the performance on the test data, then we can do something called a variable importance step; which is, if I have these 12 variables that I am using to predict the salary status, is there some way in which I can look at these variables and order them in terms of importance.

So, what do we mean by importance? So, if I were to ask you, pick one variable which is the most important for predicting this, what would that variable be right; then if I ask you pick two variables which are important for predicting the salary level as less than 50000 or 50000; what would those two variables be from these 12 variables input variables? So, if you keep asking this question, at some point once we have a certain set of variables which you can use to figure out, whether the salary level is less than equal to 50000 or greater than 50000.

Then adding more variables, if it is not really improving the performance tremendously, particularly on the test set; then what you do is, you basically say you know maybe 5 or 6 or 4 or whatever the number turns out to be those many number of variables are good enough to predict the outcome variable. Again as I mentioned before, what this does is that, it makes it usable in future.

(Refer Slide Time: 18:59)



Framework

- Solution conceptualization
  - Identify if data is clean
  - Look for missing values
  - Identify variables influencing salary status and look for possible relationships between variables
    - Correlation, chi-square test, box plots, scatter plots etc.
  - Identify if categories can be combined
  - Build a model with reduced number of variables to classify the individual's salary status to plan subsidy outlay, monitor and pr t misuse

Python for Data Science

So, this is what we will do in terms of the steps and you will see python code and python exercises, which does each one of these steps and you will be taught this based on the other things that you have been taught in python. The first step is to identify is data is clean. What that means, is that if there are certain categories for each of these variables, all the data are they in the same category.

So, for example, if you are expecting to see a string, do you see a string or in some case some other number and so on. So, basically some kind of consistency check in terms of your understanding of the data and does the data confirm to your understanding is the first step. And when you do this test of whether data is clean or not, there might be cases where there are certain data points that are missing.

So, for example, if there 12 features that we are talking about, not all individuals might have disclosed all the 12 features. So, there might be rows where certain columns might be missing. So, for example, some people might not have disclosed capital gains or capital loss and so on. So, those are rows with missing values, those are samples that are not complete in themselves. So, how do you handle this, how do you look for missing values; there are two things you can do, one is to remove all the samples which have incomplete data.

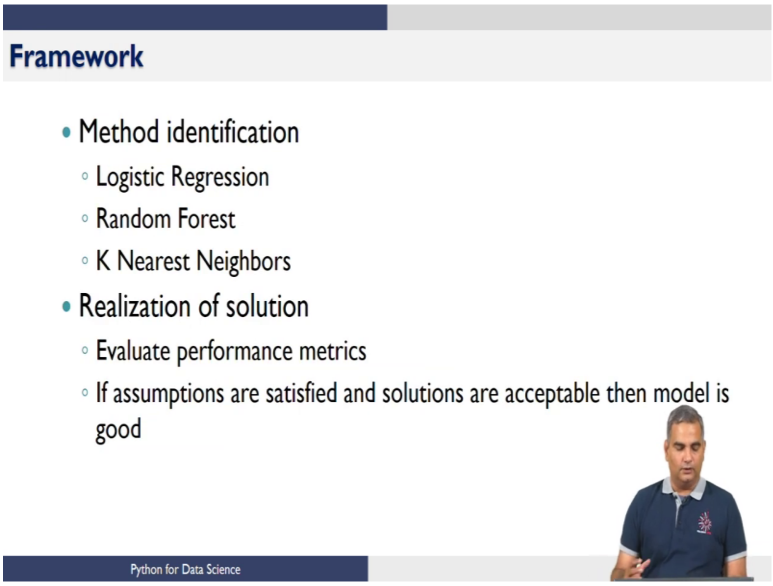
The upside is you are not going to introduce any artificiality into this problem; but the downside is you will lose a lot of data points. And it is important, because when we have 12 features, let us say there is only one feature that is missing; if you throw that data, you are also throwing the 11 other feature information with that sample. So, you lose some information through this loss of data. The other approach is to say somehow I am going to fill that missing value, and there are multiple ways of filling missing values.

Then the upside is we retain the whole data set, but the downside is we introduce some artificiality and which we do not know whether it is right or wrong. Because whatever procedure that you come up with to fill your missing data, it is based on some assumptions you make. And if those assumptions are not truly satisfied or they are violated, you might be imputing or putting in values which might not be close to the true value. So, that is the downside. So, we have to think about how we do this and basically depending on what percentage of data is missing and so on we make some judgments.

And basically, then we will start looking at some descriptive statistics to think about; how to identify variables, which impacts salary status tremendously and are there relationships internally between these input variables and so on. And if there are these relationships between different variables; so for example, if a particular category variable when it takes a value, another category variable takes some value and they are correlated all the time. Let us say if this is taking some value, I can predict the other categorical variable will take some other value; if you have situation like that, then using both the categorical variables in your classification is not very useful at all, because it is redundant information.

In those cases can be combined in this category, so that we do better; better could mean better in terms of performance or better in terms of future use. And finally, we want to build a model with this reduced number of variables, to classify the individual salary status; and basic idea or the use case for this is, you can plan subsidy outlay, monitor and prevent misuse.

(Refer Slide Time: 23:06)



Framework

- Method identification
  - Logistic Regression
  - Random Forest
  - K Nearest Neighbors
- Realization of solution
  - Evaluate performance metrics
  - If assumptions are satisfied and solutions are acceptable then model is good

Python for Data Science

So, as I said before, there are very rational ways of choosing techniques for different problems, but; that means, that you have to have some information, some knowledge about the underlying data and so on. If you are going to have really not much information, you are just going to try different things; one standard thing that everyone does nowadays, is to take a number of these techniques and then apply all of them on the same problem, and simply pick the technique that does the best, right. So, that is a very practical and utilitarian viewpoint of data science or data analytics algorithms. And it works in many cases of course, the upside is you do not have to really know too much about what the algorithm does even; you just need to know if I have a binary classification problem, there is this laundry list of techniques that I can use. So, if you have understanding at that level, then you can start using it.

The downside is subtle and in most cases it might not be a problem, but in some cases it can be a serious problem; which is that, if the data is you know biased in terms of how the sampling has been done and so on, and it is not truly representative of what is happening. Then sometimes you might have a technique which works really well with this data, but when new data comes in it might do very poorly. So, always it is a good idea to understand the data and then kind of pick a technique; but in this case what we are going to do is, we are going to pick all of these techniques and then run them.

And once you run those techniques, this is where python comes in; you whatever you have learned in your coding and so on, you will use that to realize the solution. And basically you will evaluate performance metrics. So, in this case how well it does on the test data, is a good performance metric to check. And typically if we had made some assumptions about the data, then you will kind of tie in this with the assumptions that you made and then see whether everything checks out.

But in this case, if you are just picking a list of techniques and then trying them on; all you are going to look at is, how well is it doing on my test data and then simply pick the best method in terms of the results on the test data. So, this is the introduction to the problem. This will be followed by a lecture on actually a solution strategy with python code, in terms of loading the data, doing all of this. So, that you get the feel of doing a complete data science project in this course.

Thank you.